rvif: a Decision Rule to Detect Troubling Statistical Multicollinearity Based on Redefined VIF

by Román Salmerón-Gómez and Catalina B. García-García

Abstract Multicollinearity is relevant in many different fields where linear regression models are applied since its presence may affect the analysis of ordinary least squares estimators not only numerically but also from a statistical point of view, which is the focus of this paper. Thus, it is known that collinearity can lead to incoherence in the statistical significance of the coefficients of the independent variables and in the global significance of the model. In this paper, the thresholds of the Redefined Variance Inflation Factor (RVIF) are reinterpreted and presented as a statistical test with a region of non-rejection (which depends on a significance level) to diagnose the existence of a degree of worrying multicollinearity that affects the linear regression model from a statistical point of view. The proposed methodology is implemented in the rvif package of R and its application is illustrated with different real data examples previously applied in the scientific literature.

1 Introduction

It is well known that linear relationships between the independent variables of a multiple linear regression model (multicollinearity) can affect the analysis of the model estimated by Ordinary Least Squares (OLS), either by causing unstable estimates of the coefficients of these variables or by rejecting individual significance tests of these coefficients (see, for example, Farrar and Glauber (1967), Gunst and Mason (1977), Gujarati (2003), Silvey (1969), Willan and Watts (1978) or Wooldrigde (2020)). However, the measures traditionally applied to detect multicollinearity may conclude that multicollinearity exists even if it does not lead to the negative effects mentioned above (see Subsection Effect of sample size.. for more details), when, in fact, the best solution in this case may be not to treat the multicollinearity (see O'Brien (2007)).

Focusing on the possible effect of multicollinearity on the individual significance tests of the coefficients of the independent variables (tendency not to reject the null hypothesis), this paper proposes an alternative procedure that focuses on checking whether the detected multicollinearity affects the statistical analysis of the model. For this disruptive approach, a methodology is necessary that indicates whether multicollinearity affects the statistical analysis of the model. The introduction of such methodology is the main objective of this paper. The paper also shows the use of the **rvif** package of R (R Core Team (2025)) in which this procedure is implemented.

To this end, we start from the Variance Inflation Factor (VIF). The VIF is obtained from the coefficient of determination of the auxiliary regression of each independent variable of linear regression model as a function of the other independent variables. Thus, there is a VIF for each independent variable except for the intercept, for which it is not possible to calculate a coefficient of determination for the corresponding auxiliary regression. Consequently, the VIF is able to diagnose the degree of essential approximate multicollinearity (strong linear relationship between the independent variables except the intercept) existing in the model but is not able to detect the non-essential one (strong relationship between the intercept and at least one of the independent variables). For more information on multicollinearity of essential and non-essential type, see Marquardt and Snee (1975) and Salmerón-Gómez et al. (2019).

However, the fact that the VIF detects a worrying level of multicollinearity does not always translate into a negative impact on the statistical analysis. This lack of specificity is due to the fact that other factors, such as sample size and the variance of the random

disturbance, can lead to high values of the VIF but not increase the variance of the OLS estimators (see O'Brien (2007)). The explanation for this phenomenon hinges on the fact that, in the orthogonal variable reference model, which is traditionally considered as the reference, the linear relationships are assumed to be eliminated, while other factors, such as the variance of the random disturbance, maintain the same values.

Then, to avoid these inconsistencies, Salmerón et al. (2025) propose a QR decomposition in the matrix of independent variables of the model in order to obtain an orthonormal matrix. By redefining the reference point, the variance inflation factor is also redefined, resulting in a new detection measure that analyzes the change in the VIF and the rest of relevant factors of the model, thereby overcoming the problems associated with the traditional VIF, as described by O'Brien (2007) among others. The intercept is also included in the detection (contrary to what happens with the traditional VIF), it is therefore able to detect both essential and non-essential multicollinearity. This new measure presented by Salmerón et al. (2025) is called Redefined Variance Inflation Factor (RVIF).

In this paper, the RVIF is associated with a statistical test for detecting troubling multicollinearity, this test is given by a region of non-rejection that depends on a significance level. Note that most of the measures used to diagnose multicollinearity are merely indicators with rules of thumb rather than statistical tests per se. To the best of our knowledge, the only existing statistical test for diagnosing multicollinearity was presented by Farrar and Glauber (1967) and has received strong criticism (see, for example, Haitovsky (1969), Kumar (1975), Wichers (1975) and O'Hagan and McCabe (1975)). Thus, for example, Haitovsky (1969) indicates that the Farrar and Glauber statistic indicates that the variables are not orthogonal to each other; it tells us nothing more. In this sense, O'Hagan and McCabe (1975) indicates that such a test simply indicates whether the null hypothesis of orthogonality is rejected by giving no information on the value of the matrix of correlations determinant above which the multicollinearity problem becomes intolerable. Therefore, the non-rejection region presented in this paper should be a relevant contribution to the field of econometrics insofar as it would fill an existing gap in the scientific literature.

The paper is structured as follows: Sections Preliminares and A first attempt of... provide preliminary information to introduce the methodology used to establish the non-rejection region described in Section A non-rejection region.... Section rvif package presents the package **rvif** of R (R Core Team (2025)) and shows its main commands by replicating the results given in Salmerón et al. (2025) and in the previous sections of this paper. Finally, Section Conclusions summarizes the main contributions of this paper.

2 Preliminaries

This section identifies some inconsistencies in the definition of the VIF and how these are reflected in the individual significance tests of the linear regression model. It also shows how these inconsistencies are overcome in the proposal presented by Salmerón et al. (2025) and how this proposal can lead to a decision rule to determine whether the degree of multicollinearity is troubling, i.e., whether it affects the statistical analysis (individual significance tests) of the model.

2.1 The original model

The multiple linear regression model with n observations and k independent variables can be expressed as:

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times k} \cdot \boldsymbol{\beta}_{k\times 1} + \mathbf{u}_{n\times 1}, \tag{1}$$

where the first column of $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_2 \dots \mathbf{X}_i \dots \mathbf{X}_k]$ is composed of ones representing the intercept and $\mathbf{u}$ represents the random disturbance assumed to be centered and spherical. That is, $E[\mathbf{u}_{n\times 1}] = \mathbf{0}_{n\times 1}$ and $var(\mathbf{u}_{n\times 1}) = \sigma^2 \cdot \mathbf{I}_{n\times n}$, where $\mathbf{0}$ is a vector of zeros, σ^2 is the variance of the random disturbance and $\mathbf{I}$ is the identity matrix.

Given the original model (1), the VIF is defined as the ratio between the variance of the estimator in this model, $var\left(\widehat{\beta}_i\right)$, and the variance of the estimator of a hypothetical reference model, that is, a hypothetical model in which orthogonality among the independent variables is assumed, $var\left(\widehat{\beta}_{i,o}\right)$. This is to say:

$$var\left(\widehat{\beta}_i\right) = \frac{\sigma^2}{n \cdot var(\mathbf{X}_i)} \cdot \frac{1}{1 - R_i^2} = var\left(\widehat{\beta}_{i,o}\right) \cdot VIF(i), \quad i = 2, \ldots, k, \tag{2}$$

$$\frac{var\left(\widehat{\beta}_i\right)}{var\left(\widehat{\beta}_{i,o}\right)} = VIF(i), \quad i = 2, \ldots, k, \tag{3}$$

where $\mathbf{X}_i$ is the independent variable i of the model (1) and R_i^2 the coefficient of determination of the following auxiliary regression:

$$\mathbf{X}_i = \mathbf{X}_{-i} \cdot \boldsymbol{\alpha} + \mathbf{v},$$

where $\mathbf{X}_{-i}$ is the result of eliminating $\mathbf{X}_i$ from the matrix $\mathbf{X}$.

As observed in the expression (2), a high VIF leads to a high variance. Then, since the experimental value for the individual significance test is given by:

$$t_i = \left| \frac{\widehat{\beta}_i}{\sqrt{\frac{\widehat{\sigma}^2}{n \cdot var(\mathbf{X}_i)} \cdot VIF(i)}} \right|, \quad i = 2, \ldots, k, \tag{4}$$

a high VIF will lead to a low experimental statistic (t_i), provoking the tendency not to reject the null hypothesis, i.e. the experimental statistic will be lower than the theoretical statistic (given by $t_{n-k}(1 - \alpha/2)$, where α is the significance level).

However, this statement is full of simplifications. By following O'Brien (2007), and as can be easily observed in the expression (4), other factors, such as the estimation of the random disturbance and the size of the sample, can counterbalance the high value of the VIF to yield a low value for the experimental statistic. That is to say, it is possible to obtain VIF values greater than 10 (the threshold traditionally established as troubling, see Marquardt (1970) for example) that do not necessarily imply high estimated variance on account of a large sample size or a low value for the estimated variance of the random disturbance. This explains, as noted in the introduction, why not all models with a high value for the VIF present effects on the statistical analysis of the model.

> Example 1. Thus, for example, García et al. (2019) considered an extension of the interest rate model presented by Wooldrigde (2020), where $k = 3$, in which all the independent variables have associated coefficients significantly different from zero, presenting a VIF equal to 71.516, much higher than the threshold normally established as worrying. In other words, in this case, a high VIF does not mean that the individual significance tests are affected. This situation is probably due to the fact that in this case 131 observations are available, i.e. the expression (4) can be expressed as:
>
> $$t_i = \left| \frac{\widehat{\beta}_i}{\sqrt{\frac{\widehat{\sigma}^2}{131 \cdot var(\mathbf{X}_i)} \cdot 71.516}} \right| = \left| \frac{\widehat{\beta}_i}{\sqrt{0.546 \cdot \frac{\widehat{\sigma}^2}{var(\mathbf{X}_i)}}} \right|, \quad i = 2, 3.$$
>
> Note that in this case a high value of n compensates for the high value of VIF. In addition, the value of n will also cause $\widehat{\sigma}^2$ to decrease, since $\widehat{\sigma}^2 = \frac{\mathbf{e}^t \mathbf{e}}{n-k}$, where $\mathbf{e}$ are the residuals of the original model (1).

The Subsection Effect of sample size.. provides an example that illustrates in more detail the effect of sample size on the statistical analysis of the model. ◊

On the other hand, considering the hypothetical orthogonal model, the value of the experimental statistic of the individual significance test, whose null hypothesis is $\beta_i = 0$ in face of the alternative hypothesis $\beta_i \neq 0$ with $i = 2, \ldots, k$, is given by:

$$t_i^o = \left| \frac{\widehat{\beta}_i}{\sqrt{\frac{\widehat{\sigma}^2}{n \cdot var(\mathbf{X}_i)}}} \right|, \quad i = 2, \ldots, k, \tag{5}$$

where the estimated variance of the estimator has been diminished due to the VIF always being greater than or equal to 1, and consequently, $t_i^o \geq t_i$. However, it has been assumed that the same estimates for the independent variable coefficients and random disturbance variance are obtained in the orthogonal and original models, which does not seem to be a plausible supposition (see Salmerón et al. (2025) Section 2.1 for more details).

2.2 An orthonormal reference model

In Salmerón et al. (2025) the following QR decomposition of the matrix $\mathbf{X}_{n \times k}$ of the model (1) is proposed: $\mathbf{X} = \mathbf{X}_o \cdot \mathbf{P}$, where $\mathbf{X}_o$ is an orthonormal matrix of the same dimensions as $\mathbf{X}$ and $\mathbf{P}$ is a higher-order triangular matrix of dimensions $k \times k$. Then, the following hypothetical orthonormal reference model:

$$\mathbf{y} = \mathbf{X}_o \cdot \boldsymbol{\beta}_o + \mathbf{w}, \tag{6}$$

verifies that:

$$\widehat{\boldsymbol{\beta}} = \mathbf{P}^{-1} \cdot \widehat{\boldsymbol{\beta}}_o, \ \mathbf{e} = \mathbf{e}_o, \ var\left(\widehat{\boldsymbol{\beta}}_o\right) = \sigma^2 \cdot \mathbf{I},$$

where $\mathbf{e}_o$ are the residuals of the orthonormal reference model (6). Note that since $\mathbf{e} = \mathbf{e}_o$, the estimate of σ^2 is the same in the original model (1) and in the orthonormal reference model (6). Moreover, since the dependent variable is the same in both models, the coefficient of determination and the experimental value of the global significance test are the same in both cases.

From these values, taking into account the expressions (2) and (3), it is evident that the ratio between the variance of the estimator in the original model (1) and the variance of the estimator of the orthonormal reference model (6) is:

$$\frac{var\left(\widehat{\beta}_i\right)}{var\left(\widehat{\beta}_{i,o}\right)} = \frac{VIF(i)}{n \cdot var(\mathbf{X}_i)}, \quad i = 2, \ldots, k.$$

Consequently, Salmerón et al. (2025) defined the redefined VIF (RVIF) for $i = 1, \ldots, k$ as:

$$RVIF(i) = \frac{VIF(i)}{n \cdot var(\mathbf{X}_i)} = \frac{\mathbf{X}_i^t \mathbf{X}_i}{\mathbf{X}_i^t \mathbf{X}_i - \mathbf{X}_i^t \mathbf{X}_{-i} \cdot \left(\mathbf{X}_{-i}^t \mathbf{X}_{-i}\right)^{-1} \cdot \mathbf{X}_{-i}^t \mathbf{X}_i}, \tag{7}$$

which shows, among other questions, that it is defined for $i = 1, 2, \ldots, k$. That is, in contrast to the VIF, the RVIF can be calculated for the intercept of the linear regression model.

Other considerations to be taken into account are the following:

- If the data are expressed in unit length, same transformation used to calculate the Condition Number (CN), then:

$$RVIF(i) = \frac{1}{1 - \mathbf{X}_i^t \mathbf{X}_{-i} \cdot \left(\mathbf{X}_{-i}^t \mathbf{X}_{-i}\right)^{-1} \cdot \mathbf{X}_{-i}^t \mathbf{X}_i}, \quad i = 1, \ldots, k.$$

- In this case (data expressed in unit length), when $\mathbf{X}_i$ is orthogonal to $\mathbf{X}_{-i}$, it is verified that $\mathbf{X}_i^t\mathbf{X}_{-i} = \mathbf{0}$ and, consequently $RVIF(i) = 1$ for $i = 1, \ldots, k$. That is, the RVIF is always greater than or equal to 1 and its minimum value is indicative of the absence of multicollinearity.

- Denoted by $a_i = \mathbf{X}_i^t\mathbf{X}_i \cdot \left(\mathbf{X}_{-i}^t\mathbf{X}_{-i}\right)^{-1} \cdot \mathbf{X}_{-i}^t\mathbf{X}_i$, it is verified that $RVIF(i) = \frac{1}{1-a_i}$ where a_i can be interpreted as the percentage of approximate multicollinearity due to variable $\mathbf{X}_i$. Note the similarity of this expression to that of the VIF: $VIF(i) = \frac{1}{1-R_i^2}$ (see equation (2)).

- Finally, from a simulation for $k = 3$, Salmerón et al. (2025) show that if $a_i > 0.826$, then the degree of multicollinearity is worrying. In any case this value should be refined by considering higher values of k.

On the other hand, given the orthonormal reference model (6), the value for the experimental statistic of the individual significance test with the null hypothesis $\beta_{i,o} = 0$ (given the alternative hypothesis $\beta_{i,o} \neq 0$, for $i = 1, \ldots, k$) is:

$$t_i^o = \left|\frac{\widehat{\beta}_{i,o}}{\widehat{\sigma}}\right| = \left|\frac{\mathbf{p}_i \cdot \widehat{\boldsymbol{\beta}}}{\widehat{\sigma}}\right|, \tag{8}$$

where $\mathbf{p}_i$ is the i row of the matrix $\mathbf{P}$.

By comparing this expression with the one given in (5), it is observed that, as expected, not only the denominator but also the numerator has changed. Thus, in addition to the VIF, the rest of the elements in expression (4) have also changed. Consequently, if the null hypothesis is rejected in the original model, it is not assured that the same will occur in the orthonormal reference model. For this reason, it is possible to consider that the orthonormal model proposed as the reference model in Salmerón et al. (2025) is more plausible than the one traditionally applied.

2.3 Possible scenarios in the individual significance tests

To determine whether the tendency not to reject the null hypothesis in the individual significance test is caused by a troubling approximate multicollinearity that inflates the variance of the estimator, or whether it is caused by variables not being statistically significantly related, the following situations are distinguished with a significance level α:

a. If the null hypothesis is initially rejected in the original model (1), $t_i > t_{n-k}(1-\alpha/2)$, the following results can be obtained for the orthonormal model:

a.1. the null hypothesis is rejected, $t_i^o > t_{n-k}(1-\alpha/2)$; then, the results are consistent.
a.2. the null hypothesis is not rejected, $t_i^o < t_{n-k}(1-\alpha/2)$; this could be an inconsistency.

b. If the null hypothesis is not initially rejected in the original model (1), $t_i < t_{n-k}(1-\alpha/2)$, the following results may occur for the orthonormal model:

b.1 the null hypothesis is rejected, $t_i^o > t_{n-k}(1-\alpha/2)$; then, it is possible to conclude that the degree of multicollinearity affects the statistical analysis of the model, provoking not rejecting the null hypothesis in the original model.

b.2 the null hypothesis is also not rejected, $t_i^o < t_{n-k}(1-\alpha/2)$; then, the results are consistent.

In conclusion, when option b.1 is given, the null hypothesis of the individual significance test is not rejected when the linear relationships are considered (original model) but is rejected when the linear relationships are not considered (orthonormal model). Consequently, it is possible to conclude that the linear relationships affect the statistical analysis

Table 1: Data set presented previously by Wissell

t	D	C	I	CP
1996	3.805	4.770	4.879	808.23
1997	3.946	4.778	5.051	798.03
1998	4.058	4.935	5.362	806.12
1999	4.191	5.100	5.559	865.65
2000	4.359	5.291	5.843	997.30
2001	4.545	5.434	6.152	1140.70
2002	4.815	5.619	6.521	1253.40
2003	5.129	5.832	6.915	1324.80
2004	5.615	6.126	7.423	1420.50
2005	6.225	6.439	7.802	1532.10
2006	6.786	6.739	8.430	1717.50
2007	7.494	6.910	8.724	1867.20
2008	8.399	7.099	8.882	1974.10
2009	9.395	7.295	9.164	2078.00
2010	10.680	7.561	9.727	2191.30
2011	12.071	7.804	10.301	2284.90
2012	13.448	8.044	10.983	2387.50

of the model. The possible inconsistency discussed in option a.2 is analyzed in detail in Appendix Inconsistency, concluding that it will rarely occur in cases where a high degree of multicollinearity is assumed. The other two scenarios provide consistent situations.

3 A first attempt to obtain a non-rejection region associated with a statistical test to detect multicollinearity

3.1 From the traditional orthogonal model

Considering the expressions (4) and (5), it is verified that $t_i^o = t_i \cdot \sqrt{VIF(i)}$. Consequently, in the orthogonal case, with a significance level α, the null hypothesis $\beta_{i,o} = 0$ is rejected if $t_i^o > t_{n-k}(1-\alpha/2)$ for $i = 2, \ldots, k$. That is, if:

$$VIF(i) > \left(\frac{t_{n-k}(1-\alpha/2)}{t_i}\right)^2 = c_1(i), \quad i = 2, \ldots, k. \tag{9}$$

Thus, if the VIF associated with the variable i is greater than the upper bound $c_1(i)$, then it can be concluded that the estimator of the coefficient of that variable is significantly different from zero in the hypothetical case where the variables are orthogonal. In addition, if the null hypothesis is not rejected in the initial model, the reason for the failure to reject could be due to the degree of multicollinearity that affects the statistical analysis of the model.

Finally, note that since the interesting cases are those where the null hypothesis is not initially rejected, $t_i < t_{n-k}(1-\alpha/2)$, the upper bound $c_1(i)$ will always be greater than one.

> Example 2. Table 1 shows a dataset (previously presented by Wissel (2009)) with the following variables: outstanding mortgage debt (**D**, trillions of dollars), personal consumption (**C**, trillions of dollars), personal income (**I**, trillions of dollars) and outstanding consumer credit (**CP**, trillions of dollars) for the years 1996 to 2012.

Table 2: OLS estimation for the Wissel model

	Estimator	Standard Error	Experimental t	p-value
Intercept	5.469	13.017	0.420	0.681
Personal consumption	-4.252	5.135	-0.828	0.422
Personal income	3.120	2.036	1.533	0.149
Outstanding consumer credit	0.003	0.006	0.500	0.626
(Obs, Sigma Est., Coef. Det., F exp.)	17.000	0.870	0.923	52.305

Table 3: OLS estimation for part of the Wissel model

	Estimator	Standard Error	Experimental t	p-value
Intercept	-9.594	1.351	-7.102	0.000
Personal consumption	2.629	0.214	12.285	0.000
(Obs, Sigma Est., Coef. Det., F exp.)	17.000	0.890	0.910	150.925

Table 2 shows the OLS estimation of the model explaining the outstanding mortgage debt as a function of the rest of the variables. That is:

$$\mathbf{D} = \beta_1 + \beta_2 \cdot \mathbf{C} + \beta_3 \cdot \mathbf{I} + \beta_4 \cdot \mathbf{CP} + \mathbf{u}.$$

Note that the estimates for the coefficients of personal consumption, personal income and outstanding consumer credit are not significantly different from zero (a significance level of 5% is considered throughout the paper), while the model is considered to be globally valid (experimental value, F exp., higher than theoretical value).

In addition, the estimated coefficient for the variable personal consumption, which is not significantly different from zero, has the opposite sign to the simple correlation coefficient between this variable and outstanding mortgage debt, 0.953. Thus, in the simple linear regression between both variables (see Table 3), the estimated coefficient of the variable personal consumption is positive and significantly different from zero. However, adding a second variable (see Tables 4 and 5) none of the coefficients are individually significantly different from zero although both models are globally significant. This is traditionally understood as a symptom of statistically troubling multicollinearity.

By using expression (9) in order to confirm this problem, it is verified that $c_1(2) = 6.807$, $c_1(3) = 1.985$ and $c_1(4) = 18.743$, taking into account that $t_{13}(0.975) = 2.160$. Since the VIFs are equal to 589.754, 281.886 and 189.487, respectively, it is concluded that the individual significance tests for the three cases are affected by the degree of multicollinearity existing in the model. ◊

3.2 From the alternative orthonormal model (6)

In the Subsection An orthonormal reference model the individual significance test from the expression (8) is redefined. Thus, the null hypothesis $\beta_{i,o} = 0$ will be rejected, with a

Table 4: OLS estimation for part of the Wissel model

	Estimator	Standard Error	Experimental t	p-value
Intercept	-0.117	6.476	-0.018	0.986
Personal consumption	-2.343	3.335	-0.703	0.494
Personal income	2.856	1.912	1.494	0.158
(Obs, Sigma Est., Coef. Det., F exp.)	17.000	0.823	0.922	82.770

Table 5: OLS estimation for part of the Wissel model

	Estimator	Standard Error	Experimental t	p-value
Intercept	-8.640	9.638	-0.896	0.385
Personal consumption	2.335	2.943	0.793	0.441
Outstanding consumer credit	0.001	0.006	0.100	0.922
(Obs, Sigma Est., Coef. Det., F exp.)	17.000	0.953	0.910	70.487

Table 6: OLS estimation for the orthonormal Wissel model

	Estimator	Standard Error	Experimental t	p-value
Intercept	-27.882	0.932	-29.901	0.000
Personal consumption	11.592	0.932	12.432	0.000
Personal income	-1.355	0.932	-1.453	0.170
Outstanding consumer credit	0.466	0.932	0.500	0.626
(Obs, Sigma Est., Coef. Det., F exp.)	17.000	0.870	0.923	52.305

significance level α, if the following condition is verified:

$$t_i^o > t_{n-k}(1-\alpha/2), \quad i = 2, \ldots, k.$$

Taking into account the expressions (4) and (8), this is equivalent to:

$$VIF(i) > \left(\frac{t_{n-k}(1-\alpha/2)}{\widehat{\beta}_{i,o}} \right)^2 \cdot \widehat{var}\left(\widehat{\beta}_i\right) \cdot n \cdot var(\mathbf{X}_i) = c_2(i). \tag{10}$$

Thus, if the $VIF(i)$ is greater than $c_2(i)$, the null hypothesis is rejected in the respective individual significance tests in the orthonormal model (with $i = 2, \ldots, k$). Then, if the null hypothesis is not rejected in the original model and it is verified that $VIF(i) > c_2(i)$, it can be concluded that the multicollinearity existing in the model affects its statistical analysis. In summary, a lower bound for the VIF is established to indicate when the approximate multicollinearity is troubling in a way that can be reinterpreted and presented as a region of non-rejection of a statistical test.

Example 3. Continuing with the dataset presented by Wissel (2009), Table 6 shows the results of the OLS estimation of the orthonormal model obtained from the original model.

When these results are compared with those in Table 2, the following conclusions can be obtained:

- Except for the outstanding consumer credit variable, whose standard deviation has increased, the standard deviation has decreased in all cases.
- The absolute values of the experimental statistics of the individual significance tests associated with the intercept and the personal consumption variable have increased, while the experimental statistic of the personal income variable has decreased, and the experimental statistic of the outstanding consumer credit variable remains the same. These facts show that the change from the original model to the orthonormal model does not guarantee an increase in the absolute value of the experimental statistic.
- The estimation of the coefficient of the personal consumption variable is not significantly different from zero in the original model, but it is in the orthogonal model. Thus, it is concluded that multicollinearity affects the statistical analysis of the model. Note that there is also a change in the sign of the estimate, although the purpose of the orthogonal model is not to

obtain estimates for the coefficients, but rather to provide a reference point against which to measure how much the variances are inflated. Note that an orthonormal model is an idealized construction that may lack a proper interpretation in practice.

- The values corresponding to the estimated variance for the random disturbance, the coefficient of determination and the experimental statistic (F exp.) for the global significance test remain the same.

On the other hand, considering the VIF of the independent variables except for the intercept (589.754, 281.886 and 189.487) and their corresponding bounds (17.809, 623.127 and 3545.167) obtained from the expression (10), only the variable of personal consumption verifies that the VIF is higher than the corresponding bound. These results are different from those obtained in Example 2, where the traditional orthogonal model was taken as a reference.

Finally, Tables 2 and 6 show that the experimental values of the statistic t of the variable outstanding consumer credit are the same in the original and orthonormal models. ◊

The last fact highlighted at the end of the previous example is not a coincidence, but a consequence of the QR decomposition, see Appendix Test of.... Therefore, in this case, the conclusion of the individual significance test will be the same in the original and in the orthonormal model, i.e. we will always be in scenarios a.1 or b.2.

Thus, this behavior establishes a situation where it is required to select the variable fixed in the last position. Some criteria to select the most appropriate variable for this placement could be:

- To fix the variable that is considered less relevant to the model.
- To fix a variable whose associated coefficient is significantly different from zero, since this case would not be of interest for the definition of multicollinearity given in the paper. Note that the interest will be related to a coefficient considered as zero in the original model and significantly different from zero in the orthonormal one.

These options are explored in the Subsection Choice of the variable to be fix....

4 A non-rejection region associated with a statistical test to detect multicollinearity

Salmerón et al. (2025) show that high values of RVIF are associated with a high degree of multicollinearity. The question, however, is how high RVIFs have to be to reflect troubling multicollinearity.

Taking into account the expressions (7) and (10), it is possible to conclude that multicollinearity is affecting the statistical analysis of the model if it can be verified that:

$$RVIF(i) > \left(\frac{t_{n-k}(1-\alpha/2)}{\widehat{\beta}_{i,o}} \right)^2 \cdot \widehat{var}\left(\widehat{\beta}_i\right) = c_3(i), \tag{11}$$

for any $i = 1, \ldots, k$. Note that the intercept is included in this proposal, in contrast to the previous section, in which it was not included.

By following O'Brien (2007) and taking into account that the estimation of the expression (2) can be expressed as:

$$\widehat{var}\left(\widehat{\beta}_i\right) = \widehat{\sigma}^2 \cdot RVIF(i) = \frac{\mathbf{e}^t\mathbf{e}}{n-k} \cdot RVIF(i),$$

Table 7: Data set presented previously by Klein and Goldberger

Consumption	Wage income	Non-farm income	Farm income
62.8	43.41	17.10	3.96
65.0	46.44	18.65	5.48
63.9	44.35	17.09	4.37
67.5	47.82	19.28	4.51
71.3	51.02	23.24	4.88
76.6	58.71	28.11	6.37
86.3	87.69	30.29	8.96
95.7	76.73	28.26	9.76
98.3	75.91	27.91	9.31
100.3	77.62	32.30	9.85
103.2	78.01	31.39	7.21
108.9	83.57	35.61	7.39
108.5	90.59	37.58	7.98
111.4	95.47	35.17	7.42

Table 8: OLS estimation for the Klein and Goldberger model

	Estimator	Standard Error	Experimental t	p-value
Intercept	18.702	6.845	2.732	0.021
Wage income	0.380	0.312	1.218	0.251
Non-farm income	1.419	0.720	1.969	0.077
Farm income	0.533	1.400	0.381	0.711
(Obs, Sigma Est., Coef. Det., F exp.)	14.000	36.725	0.919	37.678

there are other factors that counterbalance a high value of RVIF, thereby avoiding high estimated variances for the estimated coefficients. These factors are the sum of the squared residuals (SSR= $\mathbf{e}^t\mathbf{e}$) of the model (1) and n. Thus, an appropriate specification of the econometric model (i.e., one that implies a good fit and, consequently, a small SSR) and a large sample size can compensate for high RVIF values. However, contrary to what happens for the VIF in the traditional case, these factors are taken into account in the threshold $c_3(i)$, as established in the expression (11) in $\widehat{var}\left(\widehat{\beta}_i\right)$.

Example 4. This contribution can be illustrated with the data set previously presented by Klein and Goldberger (1955), which includes variables for consumption, **C**, wage incomes, **I**, non-farm incomes, **InA**, and farm incomes, **IA**, in United States from 1936 to 1952, as shown in Table 7 (data from 1942 to 1944 are not available because they were war years).

Table 8 shows the OLS estimations of the model explaining consumption as a function of the rest of the variables. Note that there is some incoherence between the individual significance values of the variables and the global significance of the model.

The RVIFs are calculated, yielding 1.275, 0.002, 0.014 and 0.053, respectively. The associated bounds, $c_3(i)$, are also calculated, yielding 0.002, 0.0001, 0.018 and 1.826, respectively.

Since the coefficient of the wage income variable is not significantly different from zero, and because it is verified that $0.002 > 0.0001$, from (11) it is concluded that the degree of multicollinearity existing in the model is affecting its statistical analysis. ◊

Table 9: Theorem results of the Wissel model

	RVIFs	c0	c3	Scenario	Affects
Intercept	194.866090	7.371069	1.017198	b.1	Yes
Personal consumption	30.326281	4.456018	0.915790	b.1	Yes
Personal income	4.765888	2.399341	10.535976	b.2	No
Outstanding consumer credit	0.000038	0.000002	0.000715	b.2	No

Table 10: Theorem results of the Klein and Goldberger model

	RVIFs	c0	c3	Scenario	Affects
Intercept	1.275948	1.918383	0.002189	a.1	No
Wage income	0.002653	0.000793	0.000121	b.1	Yes
Non-farm income	0.014131	0.011037	0.018739	b.2	No
Farm income	0.053355	0.001558	1.826589	b.2	No

4.1 From the RVIF

Considering that in the original model (1) the null hypothesis $\beta_i = 0$ of the individual significance test is not rejected if:

$$RVIF(i) > \left(\frac{\widehat{\beta}_i}{\widehat{\sigma} \cdot t_{n-k}(1-\alpha/2)} \right)^2 = c_0(i), \quad i = 1, \ldots, k,$$

while in the orthonormal model, the null hypothesis is rejected if $RVIF(i) > c_3(i)$, the following theorem can be established:

> Theorem. Given the multiple linear regression model (1), the degree of multicollinearity affects its statistical analysis (with a level of significance of α%) if there is a variable i, with $i = 1, \ldots, k$, that verifies $RVIF(i) > \max\{c_0(i), c_3(i)\}$.

Note that Salmerón et al. (2025) indicate that the RVIF must be calculated with unit length data (as any other transformation removes the intercept from the analysis), however, for the correct application of this theorem the original data must be used as no transformation has been considered in this paper.

> Example 5. Tables 9 and 10 present the results of applying the theorem to the Wissel (2009) and Klein and Goldberger (1955) models, respectively. Note that in both cases, there is a variable i that verifies that $RVIF(i) > \max\{c_0(i), c_3(i)\}$, and consequently, we can conclude that the degree of approximate multicollinearity is affecting the statistical analysis in both models (with a level of significance of 5%). ♢

5 The rvif package

The results developed in Salmerón et al. (2025) and in this paper have been implemented in the **rvif** package of R (R Core Team (2025)). The following shows how to replicate the results presented in both papers from the existing commands `rvifs` and `multicollinearity` in **rvif**. For this reason, the code executed is shown below.

In addition, the following issues will be addressed:

- Discussion on the effect of sample size in detecting the influence of multicollinearity on the statistical analysis of the model.

- Discussion on the choice of the variable to be fixed as the last one before the orthonormalization.

The code used in these two Subsections is available at https://github.com/rnoremlas/RVIF/tree/main/rvif%20package. It is also interesting to consult the package vignette using the command `browseVignettes("rvif")`, as well as its web page with `browseURL(system.file("docs/index.html", package = "rvif"))` or https://www.ugr.es/local/romansg/rvif/index.html.

5.1 Detection of multicollinearity with RVIF: does the degree of multicollinearity affect the statistical analysis of the model?

In Salmerón et al. (2025) a series of examples are presented to illustrate the usefulness of RVIF to detect the degree of approximate multicollinearity in a multiple linear regression model. Results presented by Salmerón et al. (2025) will be reproduced by using the command `rvifs` of **rvif** package and complemented with the contribution developed in the present work by using the command `multicollinearity` of the same package. In order to facilitate the reading of the paper, this information is available in Appendix Examples of....

On the other hand, the following shows how to use the above commands to obtain the results shown in Table 9 of this paper:

```
y_W = Wissel[,2]
X_W = Wissel[,3:6]
multicollinearity(y_W, X_W)
```

```
#>          RVIFs           c0           c3 Scenario Affects
#> 1 1.948661e+02 7.371069e+00 1.017198e+00      b.1     Yes
#> 2 3.032628e+01 4.456018e+00 9.157898e-01      b.1     Yes
#> 3 4.765888e+00 2.399341e+00 1.053598e+01      b.2      No
#> 4 3.821626e-05 2.042640e-06 7.149977e-04      b.2      No
```

It is noted that the first two arguments of the `multicollinearity` command are, respectively, the dependent variable of the linear model and the design matrix containing the independent variables (intercept included as the first column).

While the results in Table 10 can be obtained using this code:

```
y_KG = KG[,1]
cte = rep(1, length(y))
X_KG = cbind(cte, KG[,2:4])
multicollinearity(y_KG, X_KG)
```

```
#>         RVIFs           c0           c3 Scenario Affects
#> 1 1.275947615 1.9183829079 0.0021892653      a.1      No
#> 2 0.002652862 0.0007931658 0.0001206694      b.1     Yes
#> 3 0.014130621 0.0110372472 0.0187393601      b.2      No
#> 4 0.053354814 0.0015584988 1.8265885762      b.2      No
```

As is known, in both cases it is concluded that the degree of multicollinearity in the model affects its statistical analysis.

The `multicollinearity` command is used by default with a significance level of 5% for the application of the Theorem set in Subsection From the RVIF. Note that if the significance level is changed to 1% (third argument of the `multicollinearity` command), in the Klein and Goldberger model it is obtained that the individual significance test of the intercept is also affected by the degree of existing multicollinearity:

```
multicollinearity(y_W, X_W, alpha = 0.01)
```

```
#>          RVIFs           c0           c3 Scenario Affects
#> 1 1.948661e+02 3.791375e+00  1.977602791      b.1     Yes
#> 2 3.032628e+01 2.291992e+00  1.780449066      b.1     Yes
#> 3 4.765888e+00 1.234122e+00 20.483705068      b.2      No
#> 4 3.821626e-05 1.050650e-06  0.001390076      b.2      No

multicollinearity(y_KG, X_KG, alpha = 0.01)

#>         RVIFs           c0           c3 Scenario Affects
#> 1 1.275947615 0.9482013897 0.0044292796      b.1     Yes
#> 2 0.002652862 0.0003920390 0.0002441361      b.1     Yes
#> 3 0.014130621 0.0054553932 0.0379131147      b.2      No
#> 4 0.053354814 0.0007703211 3.6955190555      b.2      No
```

It can be seen that the values of c_0 and c_3 change depending on the significance level used.

5.2 Effect of the sample size on the detection of the influence of multicollinearity on the statistical analysis of the model

The introduction has highlighted the idea that the measures traditionally used to detect whether the degree of multicollinearity is of concern may indicate that it is troubling while the model analysis is not affected by it. Example 1 shows that this may be due, among other factors, to the size of the sample.

To explore this issue in more detail, below is given an example where traditional measures of multicollinearity detection indicate that the existing multicollinearity is troubling while the statistical analysis of the model is not affected when the sample size is high. In particular, observations are simulated for $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_2\ \mathbf{X}_3\ \mathbf{X}_4\ \mathbf{X}_5\ \mathbf{X}_6]$ where:

$$\mathbf{X}_2 \sim N(5, 0.1^2), \quad \mathbf{X}_3 \sim N(5, 10^2), \quad \mathbf{X}_4 = \mathbf{X}_3 + \mathbf{p}$$

$$\mathbf{X}_5 \sim N(-1, 3^2), \quad \mathbf{X}_6 \sim N(15, 2.5^2),$$

where $\mathbf{p} \sim N(5, 0.5^2)$ and considering three different sample sizes: $n = 3000$ (Simulation 1), $n = 100$ (Simulation 2) and $n = 30$ (Simulation 3). In all cases the dependent variable is generated according to:

$$\mathbf{y} = 4 + 5 \cdot \mathbf{X}_2 - 9 \cdot \mathbf{X}_3 - 2 \cdot \mathbf{X}_4 + 2 \cdot \mathbf{X}_5 + 7 \cdot \mathbf{X}_6 + \mathbf{u},$$

where $\mathbf{u} \sim N(0, 2^2)$.

To set the results, a seed has been established using the command *set.seed(2024)*.

With this generation it is intended that the variable $\mathbf{X}_2$ is linearly related to the intercept as well as $\mathbf{X}_3$ to $\mathbf{X}_4$. This is supported by the results shown in Table 11, which have been obtained using the **multiColl** package of R (R Core Team (2025)) using the commands `CV`, `VIF` and `CN`.

The results imply the same conclusions in all three simulations:

- There is a worrying degree of non-essential multicollinearity in the model relating the intercept to the variable $\mathbf{X}_2$ since its coefficient of variation (CV) is lower than 0.1002506.
- There is a worrying degree of essential multicollinearity in the model relating the variables $\mathbf{X}_3$ and $\mathbf{X}_4$ since the associated Variance Inflation Factors (VIF) are greater than 10.

However, does the degree of multicollinearity detected really affect the statistical analysis of the model? According to the results shown in Tables 12 to 14 this is not always the case:

Table 11: CVs, VIFs and CN for data of Simulations 1, 2 and 3

	Simulation 1	Simulation 2	Simulation 3
X2 CV	0.020	0.019	0.025
X3 CV	2.010	1.827	3.326
X4 CV	1.004	0.968	1.434
X5 CV	3.138	1.948	2.413
X6 CV	0.167	0.176	0.194
X2 VIF	1.003	1.053	1.167
X3 VIF	388.669	373.092	926.768
X4 VIF	388.696	373.280	929.916
X5 VIF	1.001	1.014	1.043
X6 VIF	1.003	1.066	1.254
CN	148.247	162.707	123.025

Table 12: Theorem results of the Simulation 1 model

	RVIFs	c0	c3	Scenario	Affects
Intercept	0.934369	1.916912	0.000001	a.1	No
X2	0.034899	1.359909	0.000168	a.1	No
X3	0.001299	5.339519	0.000000	a.1	No
X4	0.001296	0.230992	0.000004	a.1	No
X5	0.000036	0.257015	0.000000	a.1	No
X6	0.000053	3.160352	0.000000	a.1	No

- In Simulation 1, when $n = 3000$, the degree of multicollinearity in the model does not affect the statistical analysis of the model; scenario a.1 is always verified, i.e., both in the model proposed and in the orthonormal model, the null hypothesis is rejected in the individual significance tests.
- In Simulation 2, when $n = 100$, the degree of multicollinearity in the model affects the statistical analysis of the model only in the individual significance of the intercept; in all other cases scenario a.1 is verified again.
  - As will be seen below, the fact that the individual significance of the variable $\mathbf{X}_2$ is not affected may be due to the number of observations in the data set. But it may also be because multicollinearity of the nonessential type affects only the intercept estimate. Thus, for example, in Salmerón et al. (2019) it is shown (see Table 2 of Example 2) that solving this type of approximate multicollinearity (by centering the variables that cause it) only modifies the estimate of the intercept and its standard deviation, with the estimates of the rest of the independent variables remaining unchanged.
- In Simulation 3, when $n = 30$, the degree of multicollinearity in the model affects the statistical analysis of the model in the individual significance of the intercept, in $\mathbf{X}_2$ and in $\mathbf{X}_4$.
  - In this case, as discussed, the reduction in sample size does not prevent the individual significance of $\mathbf{X}_2$ from being affected.

In conclusion, as O'Brien (2007) indicates, it can be seen that the increase in sample size prevents the statistical analysis of the model from being affected by the degree of existing multicollinearity, even though the values of the measures traditionally used to detect this problem indicate that it is troubling. To reach this conclusion, the use of the RVIF proposed by Salmerón et al. (2025) and the theorem developed in this paper is decisive.

Table 13: Theorem results of the Simulation 2 model

	RVIFs	c0	c3	Scenario	Affects
Intercept	32.965272	0.228580	0.001580	b.1	Yes
X2	1.179581	1.678248	0.014061	a.1	No
X3	0.037287	5.662562	0.000001	a.1	No
X4	0.036687	0.113376	0.000353	a.1	No
X5	0.001269	0.252728	0.000006	a.1	No
X6	0.001601	3.060976	0.000001	a.1	No

Table 14: Theorem results of the Simulation 3 model

	RVIFs	c0	c3	Scenario	Affects
Intercept	70.990340	46.793605	0.008667	b.1	Yes
X2	2.524792	0.000570	0.005083	b.1	Yes
X3	0.187896	3.892727	0.000007	a.1	No
X4	0.187317	0.168758	0.005113	b.1	Yes
X5	0.003863	0.169923	0.000325	a.1	No
X6	0.005193	2.139108	0.000013	a.1	No

5.3 Selection of the variable to be set as the last before orthonormalization

Since there are as many QR decompositions as there are possible rearrangements of the independent variables, it is convenient to test different options to determine whether the degree of multicollinearity in the regression model affects its statistical analysis.

A first possibility is to try all possible reorderings considering that the intercept must always be in first place. Thus, in the Example 2 of Salmerón et al. (2025) (see Appendix Examples of... for more details) it is considered that $\mathbf{X} = [\mathbf{1}\ \mathbf{K}\ \mathbf{W}]$ (see Table 15), but it could also be considered that $\mathbf{X} = [\mathbf{1}\ \mathbf{W}\ \mathbf{K}]$ (see Table 16).

Note that in these tables the values for each variable of RVIF and c_0 are always the same, but those of c_3 change depending on the position of each variable within the design matrix.

It is observed that in one of the two possibilities considered, the individual significance of the labor variable is affected by the degree of existing multicollinearity.

Therefore, to state that the statistical analysis of the multiple linear regression model is not affected by the multicollinearity present in the model, it is necessary to check all the possible QR decompositions and to determine in all cases that the statistical analysis is not affected. However, to determine that the statistical analysis of the model is affected by the presence of multicollinearity, it is sufficient to find one of the possible rearrangements in which the situation b.1 occurs.

Another possibility is to set in the last position of $\mathbf{X}$ a particular variable following a specific criterion. Thus, for example, in Example 3 of Salmerón et al. (2025) (see Appendix Examples of... for more details) it is verified that the variable FA has a coefficient significantly different from zero. Fixing this variable in third place since the individual significance will not be modified yields the results shown in Table 17.

Table 15: Theorem results of the Example 2 of Salmerón et al. (2025)

	RVIFs	c0	c3	Scenario	Affects
Intercept	6388.887975	88495.933700	1.649518	a.1	No
Capital	4.136993	207.628058	0.050431	a.1	No
Work	37.336378	9.445619	147.582132	b.2	No

Table 16: Theorem results of the Example 2 of Salmerón et al. (2025) (reordination 2)

	RVIFs	c0	c3	Scenario	Affects
Intercept	6388.882446	88495.933700	1.649518	a.1	No
Work	37.336378	9.445619	1.163201	b.1	Yes
Capital	4.136993	207.628058	0.082430	a.1	No

Table 17: Theorem results of the Example 3 of Salmerón et al. (2025) reordination

	RVIFs	c0	c3	Scenario	Affects
OI	1.696454e-12	9.594942e-13	1.775244e-13	b.1	Yes
S	1.718535e-12	1.100437e-12	1.012113e-12	b.1	Yes
FA	1.829200e-16	2.307700e-16	1.449800e-16	a.1	No

It can be seen that in this case the degree of multicollinearity in the model affects the individual significance of the OI and S variables.

6 Conclusions

In this paper, following Salmerón et al. (2025), we propose an alternative orthogonal model that leads to a lower bound for the RVIF, indicating whether the degree of multicollinearity present in the model affects its statistical analysis. These thresholds serve as complements to the results presented by O'Brien (2007), who stated that the estimated variances depend on other factors that can counterbalance a high value of the VIF, for example, the size of the sample or the estimated variance of the independent variables. Thus, the thresholds presented for the RVIF also depend on these factors meeting a threshold associated with each independent variable (including the intercept). Note that these thresholds will indicate whether the degree of multicollinearity affects the statistical analysis.

As these thresholds are derived from the individual significance tests of the model, it is possible to reinterpret them as a statistical test to determine whether the degree of multicollinearity in the linear regression model affects its statistical analysis. This analytic tool allows researchers to conclude whether the degree of multicollinearity is statistically troubling and whether it needs to be treated. We consider this to be a relevant contribution since, to the best of our knowledge, the only existing example of such a measure, presented by Farrar and Glauber (1967), has been strongly criticized (in addition to the limitations highlighted in the introduction, it should be noted that it completely ignores approximate non-essential multicollinearity since the correlation matrix does not include information on the intercept); consequently, this new statistical test with a non-rejection region will fill a gap in the scientific literature.

On the other hand, note that the position of each of the variables in the matrix $\mathbf{X}$ uniquely determines the reference orthonormal model $\mathbf{X}_o$. It is to say, there are as many reference models given by the proposed QR decomposition as there are possible rearrangements of the variables within the matrix $\mathbf{X}$.

In this sense, as has been shown, in order to affirm that the statistical analysis of the model is not affected by the degree of multicollinearity existing in the model (with the degree of significance used in the application of the proposed theorem), it is necessary to state that in all the possible rearrangements of $\mathbf{X}$ it is concluded that scenario b.1 does not occur. On the other hand, when there is a rearrangement in which this scenario appears, it can be stated (to the degree of significance used when applying the proposed theorem) that the degree of existing multicollinearity affects the statistical analysis of the model.

Finally, as a future line of work, it would be interesting to complete the analysis presented here by studying when the degree of multicollinearity in the model affects its numerical analysis.

7 Acknowledgments

This work has been supported by project PP2019-EI-02 of the University of Granada (Spain) and by project A-SEJ-496-UGR20 of the Andalusian Government's Counseling of Economic Transformation, Industry, Knowledge and Universities (Spain).

8 Appendix

8.1 Inconsistency in hypothesis tests: situation a.2

From a numerical point of view it is possible to reject $H_0 : \beta_i = 0$ while $H_0 : \beta_{i,o} = 0$ is not rejected, which implies that $t_i^o < t_{n-k}(1-\alpha/2) < t_i$. Or, in other words, $t_i/t_i^o > 1$.

However, from expression (8) it is obtained that $\widehat{\sigma} = |\widehat{\beta}_{i,o}|/t_i^o$. By substituting $\widehat{\sigma}$ in expression (4), taking into account expression (7), it is obtained that

$$\frac{t_i}{t_i^o} = \frac{|\widehat{\beta}_i|}{|\widehat{\beta}_{i,o}|} \cdot \frac{1}{\sqrt{RVIF(i)}}.$$

From this expression it can be concluded that in situations with high collinearity, $RVIF(i) \to +\infty$, the ratio t_i/t_i^o will tend to zero, and the condition $t_i/t_i^o > 1$ will rarely occur. That is to say, the inconsistency in situation a.2, commented on in the preliminaries of the paper, will not appear.

On the other hand, if the variable i is orthogonal to the rest of independent variables, it is verified that $\widehat{\beta}_{i,o} = \widehat{\beta}_i$ since $p_i = (0 \dots \underbrace{1}_{(i)} \dots 0)$. At the same time, $RVIF(i) = \frac{1}{SST_i}$ where SST denotes the sum of total squares. If there is orthonormality, as proposed in this paper, $SST_i = 1$ and, as consequence, it is verified that $t_i = t_i^o$. Thus, the individual significance test for the original data and for the orthonormal data are the same.

8.2 Test of individual significance of coefficient k

Taking into account that it is verified that $\boldsymbol{\beta}_o = \mathbf{P}\boldsymbol{\beta}$ where:

$$\boldsymbol{\beta}_o = \begin{pmatrix} \beta_{1,o} \\ \beta_{2,o} \\ \vdots \\ \beta_{k,o} \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1k} \\ 0 & p_{22} & \dots & p_{2k} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & p_{kk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix},$$

it is obtained that $\beta_{k,o} = p_{kk}\beta_k$. Then, the null hypothesis $H_0 : \beta_{k,o} = 0$ is equivalent to $H_0 : \beta_k = 0$. Due to this fact, Tables 2 and 6 showed an expectable behaviour. However, this behaviour will be analyzed with more detail.

The experimental value to be considered to take a decision in the test with null hypothesis $H_0 : \beta_{k,o} = 0$ and alternative hypothesis $H_1 : \beta_{k,o} \neq 0$ is given by the following expression:

$$t_k^o = \left| \frac{\widehat{\beta}_{k,o}}{\sqrt{var\left(\widehat{\beta}_{k,o}\right)}} \right|.$$

Taking into account that $\widehat{\boldsymbol{\beta}}_o = \mathbf{P}\widehat{\boldsymbol{\beta}}$ and $var\left(\widehat{\boldsymbol{\beta}}_o\right) = \mathbf{P}var\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{P}^t$, it is verified that $\widehat{\beta}_{k,o} = p_{kk}\widehat{\beta}_k$ and $var\left(\widehat{\beta}_{k,o}\right) = p_{kk}^2 var\left(\widehat{\beta}_k\right)$. Then:

$$t_k^o = \left|\frac{p_{kk}\widehat{\beta}_k}{p_{kk}\sqrt{var\left(\widehat{\beta}_k\right)}}\right| = \left|\frac{\widehat{\beta}_k}{\sqrt{var\left(\widehat{\beta}_k\right)}}\right| = t_k,$$

where t_k is the experimental value to take a decision in the test with null hypothesis $H_0 : \beta_k = 0$ and alternative hypothesis $H_1 : \beta_k \neq 0$.

8.3 Examples of Salmerón et al. (2025)

Example 1 of Salmerón et al. (2025): Detection of traditional nonessential multicollinearity. Using data from a financial model in which the Euribor (E) is analyzed from the Harmonized Index of Consumer Prices (HICP), the balance of payments to net current account (BC) and the government deficit to net nonfinancial accounts (GD), we illustrate the detection of approximate multicollinearity of the non-essential type, i.e. where the intercept is related to one of the remaining independent variables (for details see Marquardt and Snee (1975)). For more information on this data set use *help(euribor)*.

Note that Salmerón-Gómez et al. (2019) establishes that an independent variable with a coefficient of variation less than 0.1002506 indicates that this variable is responsible for a non-essential multicollinearity problem.

Thus, first of all, the approximate multicollinearity detection is performed using the measures traditionally applied for this purpose: the Variance Inflation Factor (VIF) and the Condition Number (CN). Values higher than 10 for the VIF (see, for example, Marquardt (1970)) and 30 for the CN (see, for example, Belsley (1991) or Belsley et al. (1980)), imply that the degree of existing multicollinearity is troubling. Moreover, according to Salmerón-Gómez et al. (2019), the VIF is only able to detect essential multicollinearity (relationship between independent variables excluding the intercept, see Marquardt and Snee (1975)), while the CN detects both essential and non-essential multicollinearity.

Therefore, the values calculated below (using the `VIF`, `CN` and `CVs` commands from the **multiColl** package, see Salmerón et al. (2021) and Salmerón et al. (2022) for more details on this package) indicate that the degree of approximate multicollinearity existing in the model of the essential type is not troubling, while that of the non-essential type is troubling due to the relationship of HIPC with the intercept.

```
E = euribor[,1]
data1 = euribor[,-1]

VIF(data1)

#>     HIPC       BC       GD
#> 1.349666 1.058593 1.283815

CN(data1)

#> [1] 39.35375

CVs(data1)

#> [1] 0.06957876 4.34031035 0.55015508
```

This assumption is confirmed by calculating the RVIF values, which point to a strong relationship between the second variable and the intercept:

```
rvifs(data1, ul = T, intercept = T)

#>                  RVIF       %
#> Intercept  250.294157 99.6005
#> Variable 2 280.136873 99.6430
#> Variable 3   1.114787 10.2967
#> Variable 4   5.525440 81.9019
```

The output of the `rvifs` command provides the values of the Redefined Variance Inflation Factor (RVIF) and the percentage of multicollinearity due to each variable (denoted as a_i in the An orthonormal... section).

In this case, three of the four arguments available in the `rvifs` command are used:

- The first of these refers to the design matrix containing the independent variables (the intercept, if any, being the first column).
- The second argument, *ul*, indicates that the data is to be transformed into unit length. This transformation makes it possible to establish that the RVIF is always greater than or equal to 1, having as a reference a minimum value that indicates the absence of worrying multicollinearity.
- The third argument, *intercept*, indicates whether there is an intercept in the design matrix.

Note that these results can also be obtained after using the `lm` and `model.matrix` commands as follows:

```
reg_E = lm(euribor[,1]~as.matrix(euribor[,-c(1,2)]))
rvifs(model.matrix(reg_E))

#>                  RVIF       %
#> Intercept  250.294157 99.6005
#> Variable 2 280.136873 99.6430
#> Variable 3   1.114787 10.2967
#> Variable 4   5.525440 81.9019
```

Finally, the application of the Theorem established in Subsection From the RVIF detects that the individual inference of the second variable (HIPC) is affected by the degree of multicollinearity existing in the model. These results are obtained using the `multicollinearity` command from the **rvif** package:

```
multicollinearity(E, data1)

#>          RVIFs           c0           c3 Scenario Affects
#> 1 5.325408e+00 1.575871e+01 2.166907e-02      a.1      No
#> 2 5.357830e-04 3.219456e-06 4.249359e-05      b.1     Yes
#> 3 5.109564e-11 1.098649e-09 2.586237e-12      a.1      No
#> 4 1.631439e-11 3.216522e-10 8.274760e-13      a.1      No
```

Therefore, it can be established that the existing multicollinearity affects the statistical analysis of the Euribor model.

Example 2 of Salmerón et al. (2025): Detection of generalized nonessential multicollinearity. Using data from a Cobb-Douglas production function in which the production (P) is analyzed from the capital (K) and the work (W), we illustrate the detection of approximate multicollinearity of the generalized non-essential type, i.e., that in which at least two independent variables with very little variability (excluding the intercept) are related to each

other (for more details, see Salmerón et al. (2020)). For more information on this dataset use *help(CDpf)*.

Using the `rvifs` command, it can be determined that both capital and labor are linearly related to each other with high RVIF values below the threshold established as a worrying value:

```
P = CDpf[,1]
data2 = CDpf[,2:4]

rvifs(data2, ul = T)

#>                 RVIF       %
#> Intercept  178888.71 99.9994
#> Variable 2  38071.44 99.9974
#> Variable 3 255219.74 99.9996
```

However, the application of the Theorem established in Subsection From the RVIF does not detect that the degree of multicollinearity in the model affects the statistical analysis of the model:

```
multicollinearity(P, data2)

#>         RVIFs           c0           c3 Scenario Affects
#> 1 6388.887975 88495.933700   1.64951764      a.1      No
#> 2    4.136993   207.628058   0.05043083      a.1      No
#> 3   37.336378     9.445619 147.58213164      b.2      No
```

Now then, if we rearrange the design matrix $\mathbf{X}$ we obtain that:

```
data2 = CDpf[,c(2,4,3)]
multicollinearity(P, data2)

#>         RVIFs           c0         c3 Scenario Affects
#> 1 6388.882446 88495.933700 1.64951764      a.1      No
#> 2   37.336378     9.445619 1.16320125      b.1     Yes
#> 3    4.136993   207.628058 0.08242979      a.1      No
```

Therefore, it can be established that the existing multicollinearity does affect the statistical analysis of the Cobb-Douglas production function model.

Example 3 of Salmerón et al. (2025): Detection of essential multicollinearity. Using data from a model in which the number of employees of Spanish companies (NE) is analyzed from the fixed assets (FA), operating income (OI) and sales (S), we illustrate the detection of approximate multicollinearity of the essential type, i.e., that in which at least two independent variables (excluding the intercept) are related to each other (for more details, see Marquardt and Snee (1975)). For more information on this dataset use *help(employees)*.

In this case, the `rvifs` command shows that variables three and four (OI and S) have a high VIF value, so they are highly linearly related:

```
NE = employees[,1]
data3 = employees[,2:5]

rvifs(data3, ul = T)

#>                    RVIF       %
#> Intercept      2.984146 66.4896
#> Variable 2     5.011397 80.0455
#> Variable 3 15186.744870 99.9934
#> Variable 4 15052.679178 99.9934
```

Note that if in *rvifs(data3, ul = T)* the unit length transformation is avoided, which is done in the `multicollinearity` command, the RVIF cannot be calculated since the system is computationally singular. For this reason, the intercept is eliminated below since it has been shown above that it does not play a relevant role in the linear relationships of the model.

Finally, the application of the Theorem established in Subsection From the RVIF detects that the individual inference of the third variable (OI) is affected by the degree of multicollinearity existing in the model:

```
multicollinearity(NE, data3[,-1])
```

```
#>          RVIFs           c0           c3 Scenario Affects
#> 1 1.829154e-16 2.307712e-16 4.679301e-17      a.1      No
#> 2 1.696454e-12 9.594942e-13 2.129511e-13      b.1     Yes
#> 3 1.718535e-12 1.100437e-12 2.683809e-12      b.2      No
```

Therefore, it can be established that the existing multicollinearity affects the statistical analysis of the model of the number of employees in Spanish companies.

Example 4 of Salmerón et al. (2025): The special case of simple linear model. The simple linear regression model is an interesting case because it has a single independent variable and the intercept. Since the intercept is not properly considered as an independent variable of the model in many cases (see introduction of Salmerón-Gómez et al. (2019) for more details), different software (including R, R Core Team (2025)) do not consider that there can be worrisome multicollinearity in this type of model.

To illustrate this situation, Salmerón et al. (2025) randomly generates observations for the following two simple linear regression models $\mathbf{y}_1 = \beta_1 + \beta_2\mathbf{V} + \mathbf{u}_1$ and $\mathbf{y}_2 = \alpha_1 + \alpha_2\mathbf{Z} + \mathbf{u}_1$, according to the following code:

```
set.seed(2022)
obs = 50
cte4 = rep(1, obs)
V = rnorm(obs, 10, 10)
y1 = 3 + 4*V + rnorm(obs, 0, 2)
Z = rnorm(obs, 10, 0.1)
y2 = 3 + 4*Z + rnorm(obs, 0, 2)

data4.1 = cbind(cte4, V)
data4.2 = cbind(cte4, Z)
```

For more information on these data sets use *help(SLM1)* and *help(SLM2)*.

As mentioned above, the R package (R Core Team (2025)) denies the existence of multicollinearity in this type of model. Thus, for example, when using the `vif` command of the **car** package on *reg=lm(y1~V)* the following message is obtained: *Error in vif.default(reg): model contains fewer than 2 terms*.

Undoubtedly, this message is coherent with the fact that, as mentioned above, the VIF is not capable of detecting non-essential multicollinearity (which is the only multicollinearity that exists in this type of model). However, the error message provided may lead a non-specialized user to consider that the multicollinearity problem does not exist in this type of model. These issues are addressed in more depth in Salmerón et al. (2022).

On the other hand, the calculation of the RVIF in the first model shows that the degree of multicollinearity is not troubling, since it presents very low values:

```
rvifs(data4.1, ul = T)
```

```
#>                RVIF       %
#> Intercept  2.015249 50.3783
#> Variable 2 2.015249 50.3783
```

While in the second model they are very high, indicating a problem of non-essential multicollinearity:

```
rvifs(data4.2, ul = T)

#>                RVIF       %
#> Intercept  9390.044 99.9894
#> Variable 2 9390.044 99.9894
```

By using the `multicollinearity` command, it is found that the individual inference of the intercept of the second model is affected by the degree of multicollinearity in the model:

```
multicollinearity(y1, data4.1)

#>          RVIFs        c0           c3 Scenario Affects
#> 1 0.0403049717 0.6454323 1.045802e-05      a.1      No
#> 2 0.0002675731 0.8383436 8.540101e-08      a.1      No

multicollinearity(y2, data4.2)

#>        RVIFs         c0         c3 Scenario Affects
#> 1 187.800878 21.4798003 0.03277691      b.1     Yes
#> 2   1.879296  0.3687652 9.57724567      b.2      No
```

Therefore, it can be established that the multicollinearity existing in the first simple linear regression model does not affect the statistical analysis of the model, while in the second one it does.

References

D. Belsley. A guide to using the collinearity diagnostics. *Computational Science in Economics and Manegement*, 4:33–50, 1991. doi: 10.1007/BF00426854. URL https://doi.org/10.1007/BF00426854. [p209]

D. Belsley, E. Kuh, and R. Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley and Sons, 1980. URL https://onlinelibrary.wiley.com/doi/book/10.1002/0471725153. [p209]

D. E. Farrar and R. R. Glauber. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 49(1):92–107, 1967. doi: 10.2307/1937887. URL https://doi.org/10.2307/1937887. [p192, 193, 207]

C. B. García, R. Salmerón, C. García-García, and J. García. Residualization: justification, properties and application. *Journal of Applied Statistics*, 47(11):1990–2010, 2019. doi: 10.1111/insr.12575. URL https://doi.org/10.1111/insr.12575. [p194]

D. Gujarati. *Basic Econometrics*. McGraw-Hill (fourth edition), 2003. URL https://highered.mheducation.com/sites/0072335424/. [p192]

R. Gunst and R. Mason. Advantages of examining multicollinearities in regression analysis. *Biometrics*, 33(1):249–260, 1977. doi: 10.2307/2529320. URL https://doi.org/10.2307/2529320. [p192]

Y. Haitovsky. Multicollinearity in regression analysis: Comment. *The Review of Economics and Statistics*, 51(4):486–489, 1969. doi: 10.2307/1926450. URL https://doi.org/10.2307/1926450. [p193]

L. R. Klein and A. S. Goldberger. *An Economic Model of the United States 1929-1952*. Amsterdan: North-Holland Publishing Company, 1955. doi: 10.2307/2227976. URL https://doi.org/10.2307/2227976. [p201, 202]

T. K. Kumar. Multicollinearity in regression analysis. *The Review of Economics and Statistics*, 57 (3):365–366, 1975. doi: 10.2307/1923925. URL https://doi.org/10.2307/1923925. [p193]

D. Marquardt. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*, 12(3):591–612, 1970. doi: 10.2307/1267205. URL https://doi.org/10.2307/1267205. [p194, 209]

D. Marquardt and R. Snee. Ridge regression in practice. *The American Statistician*, 29(1):3–20, 1975. doi: 10.2307/2683673. URL https://doi.org/10.2307/2683673. [p192, 209, 211]

R. O'Brien. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41(5):673–690, 2007. URL https://link.springer.com/article/10.1007/s11135-006-9018-6. [p192, 193, 194, 200, 205, 207]

J. O'Hagan and B. McCabe. Tests for the severity of multicolinearity in regression analysis: A comment. *The Review of Economics and Statistics*, 57(3):368–370, 1975. doi: 10.2307/1923927. URL https://doi.org/10.2307/1923927. [p193]

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 4.5.1 edition, 2025. URL https://www.R-project.org/. [p192, 193, 202, 204, 212]

R. Salmerón, C. García, and J. García. Comment on "A note on collinearity diagnostics and centering" by Velilla (2018). *The American Statistician*, 74(1):68–71, 2019. doi: 10.1080/00031305.2019.1635527. URL https://doi.org/10.1080/00031305.2019.1635527. [p205]

R. Salmerón, C. García, and J. García. Detection of near-multicollinearity through centered and noncentered regression. *Mathematics*, 8(6):931, 2020. doi: 10.3390/math8060931. URL https://doi.org/10.3390/math8060931. [p211]

R. Salmerón, C. García, and J. García. A guide to using the R package multicoll for detecting multicollinearity. *Computational Economics*, 57:529–536, 2021. doi: 10.1007/s10614-019-09967-y. URL https://doi.org/10.1007/s10614-019-09967-y. [p209]

R. Salmerón, C. García, and J. García. The multicoll package versus other existing packages in R to detect multicollinearity. *Computational Economics*, 60:439–450, 2022. doi: 10.1007/s10614-021-10154-1. URL https://doi.org/10.1007/s10614-021-10154-1. [p209, 212]

R. Salmerón, C. B. García, and J. García. A redefined variance inflation factor: overcoming the limitations of the variance inflation factor. *Computational Economics*, 65:337–363, 2025. doi: 10.1007/s10614-024-10575-8. URL https://doi.org/10.1007/s10614-024-10575-8. [p193, 195, 196, 200, 202, 203, 205, 206, 207, 209, 210, 211, 212]

R. Salmerón-Gómez, A. Rodríguez-Sánchez, and C. García-García. Diagnosis and quantification of the non-essential collinearity. *Computational Statistics*, 35:647–666, 2019. doi: 10.1007/s00180-019-00922-x. URL https://doi.org/10.1007/s00180-019-00922-x. [p192, 209, 212]

S. Silvey. Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31(3):539–552, 1969. URL https://www.jstor.org/stable/2984357. [p192]

C. R. Wichers. The detection of multicollinearity: A comment. *The Review of Economics and Statistics*, 57(3):366–368, 1975. doi: 10.2307/1923926. URL https://doi.org/10.2307/1923926. [p193]

A. Willan and D. Watts. Meaningful multicollinearity measures. *Technometrics*, 20(4):407–412, 1978. doi: 10.1080/00401706.1978.10489694. URL https://doi.org/10.1080/00401706.1978.10489694. [p192]

J. Wissel. *A new biased estimator for multivariate regression models with highly collinear variables*. 2009. URL https://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/docId/2949. [p197, 199, 202]

J. Wooldrigde. *Introductory Econometrics. A Modern Approach*. South-Western, CENGAGE Learning (7th edition), 2020. URL https://www.cengage.uk/c/introductory-econometrics-a-modern-approach-7e-wooldridge/9781337558860PF/. [p192, 194]

Román Salmerón-Gómez
University of Granada
Department of Quantitative Methods for Economics and Business
Campus Universitario de La Cartuja, Universidad de Granada. 18071 Granada (España)
https://www.ugr.es/~romansg/web/index.html
ORCiD: 0000-0003-2589-4058
romansg@ugr.es

Catalina B. García-García
University of Granada
Department of Quantitative Methods for Economics and Business
Campus Universitario de La Cartuja, Universidad de Granada. 18071 Granada (España)
https://metodoscuantitativos.ugr.es/informacion/directorio-personal/catalina-garcia-garcia
ORCiD: 0000-0003-1622-3877
cbgarcia@ugr.es