mathematics

MDPI

Article

Detection of Near-Nulticollinearity through Centered and Noncentered Regression

Román Salmerón Gómez [1], Catalina García García [1,*] and José García Pérez [2]

[1] Department of Quantitative Methods for Economics and Business, University of Granada, 18010 Granada, Spain; romansg@ugr.es
[2] Department of Economy and Company, University of Almería, 04120 Almería, Spain; jgarcia@ual.es
* Correspondence: cbgarcia@ugr.es

Received: 1 May 2020; Accepted: 4 June 2020; Published: 7 June 2020

Abstract: This paper analyzes the diagnostic of near-multicollinearity in a multiple linear regression from auxiliary centered (with intercept) and noncentered (without intercept) regressions. From these auxiliary regressions, the centered and noncentered variance inflation factors (VIFs) are calculated. An expression is also presented that relates both of them. In addition, this paper analyzes why the VIF is not able to detect the relation between the intercept and the rest of the independent variables of an econometric model. At the same time, an analysis is also provided to determine how the auxiliary regression applied to calculate the VIF can be useful to detect this kind of multicollinearity.

Keywords: centered model; noncentered model; intercept; essential multicollinearity; nonessential multicollinearity

MSC: 62JXX; 62J20; 60PXX

1. Introduction

Consider the following multiple linear model with n observations and k regressors:

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

where $\mathbf{y}$ is a vector with the observations of the dependent variable, $\mathbf{X}$ is a matrix containing the observations of regressors and $\mathbf{u}$ is a vector representing a random disturbance (that is assumed to be spherical). Generally, the first column of matrix $\mathbf{X}$ is composed of ones to denote that the model contains an intercept. Thus, $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_2 \dots \mathbf{X}_k]$ where $\mathbf{1}_{n\times 1} = (1\ 1 \dots 1)^t$. This model is considered to be centered.

When this model presents worrying near-multicollinearity (hereinafter, multicollinearity), that is, when the linear relation between the regressors affects the numerical and/or statistical analysis of the model, the usual approach is to transform the regressors (see, for example, Belsley [1], Marquardt [2] or, more recently, Velilla [3]). Due to the transformations (centering, typification or standardization) implying the elimination of the intercept in the model, the transformed models are considered to be noncentered. Note that even after transforming the data, it is possible to recover the original model (centered) from the estimations of the transformed model (noncentered model). However, in this paper, we refer to the centered and noncentered model depending on whether the intercept is initially included or not. Thus, it is considered that the model is centered if $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_2 \dots \mathbf{X}_k]$ and noncentered if $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2 \dots \mathbf{X}_k]$, given that $\mathbf{X}_j \neq \mathbf{1}$ with $j = 1, \dots, k$.

From the intercept is also possible to distinguish between essential and nonessential multicollinearity:

Nonessential: A near-linear relation between the intercept and at least one of the rest independent variables.

Essential: A near-linear relation between at least two of the independent variables (excluding the intercept).

A first idea of these definitions was provided by Cohen et al. [4]: Nonessential ill-conditioning results simply from the scaling of the variables, whereas essential ill-conditioning results from substantive relationships among the variables. While in some papers the idea of distinguishing between essential and nonessential collinearity is attributed to Marquardt [5], it is possible to find this concept in Marquardt and Snee [6]. These terms have been widely used not only for linear models but also, for example, for moderated models with interactions and/or with a quadratic term. However, these concepts have been analyzed fundamentally from the point of view of the solution of collinearity. Thus, as Marquardt and Snee [6] stated: In a linear model, centering removes the correlation between the constant term and all linear terms.

The variance inflation factor is one of the most applied measures to detect multicollinearity. Following O'Brien [7], commonly a VIF of 10 or even one as low as 4 have been used as rules of thumbs to indicate excessive or serious collinearity. Salmerón et al. [8] show that the VIF does not detect the nonessential multicollinearity, while this kind of multicollinearity is detected by the index of Stewart [9] (see Salmerón Gómez et al. [10]). This index has been misunderstood in the literature since its presentation by Stewart, who wrongly identified it with the VIF. Even Marquardt [11] when published a comment of the paper of Stewart [9] stated: Stewart collinearity indices are simply the square roots of the corresponding variance inflation factor. It is not clear to me whether giving a new name to the square of a VIF is a help or a hindrance to understanding. There is a long and precisely analogous history of using the term "standard error" for the square root of the corresponding "variances". Given the continuing necessity for dealing with statistical quantities on both the scale of the observable and the scale of the observable squared, there may be a place for a new term. Clearly, the essential intellectual content is identical for both terms.

However, in Salmerón Gómez et al. [12] it is shown that the VIF and the index of Stewart are not the same measure. This paper analyzes in what cases use one measure or another, focusing on the initial distinction between centered and noncentered models. Thus, the algebraic contextualization provided by Salmerón Gómez et al. [12] will be complemented from an econometric point of view. This question was also presented by Jensen and Ramirez [13], striving to commit to a clarification of the misuse given to the VIF over decades since its first use, who insinuated: To choose a model, with or without intercept, is substantive, is specific to each experimental paradigm and is beyond the scope of the present study. It was also stated that: This differs between centered and uncentered diagnostics.

This paper, focused on the differences between essential and nonessential multicollinearity in relation to its diagnostic, analyzes the behaviour of the VIF depending on whether model (1) initially includes the intercept or not. For this analysis, it will be considered that the auxiliary regression used for its calculation is centered or not since as stated by Grob [14] (p. 304): Instead of using the classical coefficient of determination in the definition of VIF, one may also apply the centered coefficient of determination. As a matter of fact, the latter definition is more common. We may call VIF uncentered or centered, depending on whether the classical or centered coefficient of determination is used. From the above considerations, a centered VIF only makes sense when the matrix $\mathbf{X}$ contains ones as a column. Additionally, although initially in the centered version of model (1) it is possible to find these two kinds of multicollinearity, and in the noncentered version, it is only possible to find essential multicollinearity, this paper shows that this statement is subject to some nuances.

On the other hand, throughout the paper the following statement of Cook [15] will be illustrated: As a matter of fact, the centered VIF requires an intercept in the model but at the same time denies the status of the intercept as an independent "variable" being possibly related to collinearity effects. Furthermore, another statement was provided by Belsley [16] (p. 29): The centered VIF has no ability to discover collinearity involving the intercept. Thus, the second part of the paper analyzes why the

centered VIF is unable to detect the nonessential multicollinearity and, for this, the centered coefficient of determination of the centered auxiliary regression to calculate the centered VIF is analyzed. This analysis will be applied to propose a methodology to detect the nonessential multicollinearity from the centered auxiliary regression.

The structure of the paper is as follows: Section 2 presents the detection of multicollinearity in noncentered models from the noncentered auxiliary regressions, Section 3 analyzes the effects of high values of the noncentered VIF on the statistical analysis of the model and Section 4 presents the detection of multicollinearity in centered models from the centered auxiliary regressions. Section 5 illustrates the contribution of the paper with two empirical applications. Finally, Section 6 summarizes the main conclusions.

2. Auxiliary Noncentered Regressions

This section presents the calculation of the VIF uncentered, VIFnc, considering that the auxiliary regression is noncentered, that is, it has no intercept. First, the method regarding how to calculate the coefficient of determination for noncentered models is presented.

2.1. Noncentered Coefficient of Determination

Given the linear regression of Equation (1) with or without the intercept, the following decomposition for the sum of squares is verified:

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \widehat{y}_i^2 + \sum_{i=1}^{n} e_i^2, \tag{2}$$

where $\widehat{\mathbf{y}}$ represents the estimation of the dependent variable of the model that is fit by employing ordinary least squares (OLS) and $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$ are the residuals obtained from that fit. In this case, the coefficient of determination is obtained by the following expression:

$$R_{nc}^2 = \frac{\sum\limits_{i=1}^{n} \widehat{y}_i^2}{\sum\limits_{i=1}^{n} y_i^2} = 1 - \frac{\sum\limits_{i=1}^{n} e_i^2}{\sum\limits_{i=1}^{n} y_i^2}. \tag{3}$$

Comparing the decomposition of the sums of squares given by (2) with the traditionally applied method to calculate the coefficient of determination in models with the intercept, as in model (1):

$$\sum_{i=1}^{n} (y_i - \overline{\mathbf{y}})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{\mathbf{y}})^2 + \sum_{i=1}^{n} e_i^2, \tag{4}$$

it is noted that both coincide if the dependent variable has zero mean. If the mean is different from zero, both models present the same residual sum of squares but different explained and total sum of squares.

Thus, these models lead to the same value for the coefficient of determination (and, as a consequence, for the VIF) only if the dependent variable presents a mean equal to zero.

2.2. Noncentered Variance Inflation Factor

The VIFnc is obtained from the expression:

$$VIFnc(j) = \frac{1}{1 - R_{nc}^2(j)}, \quad j = 1, \ldots, k, \tag{5}$$

where $R^2_{nc}(j)$ is the coefficient of determination, calculated by following (3), of the noncentered auxiliary regression:

$$\mathbf{X}_j = \mathbf{X}_{-j}\delta + \mathbf{w}, \tag{6}$$

where $\mathbf{X}_{-j}$ is equal to the matrix $\mathbf{X}$ after eliminating the variable $\mathbf{X}_j$, for $j = 1, \ldots, k$, and it does not have a vector of ones representing the intercept.

In this case:

- $\sum_{i=1}^{n} X_{ij}^2 = \mathbf{X}_j^t \mathbf{X}_j$, and
- $\sum_{i=1}^{n} \widehat{X}_{ij}^2 = \widehat{\mathbf{X}}_j^t \widehat{\mathbf{X}}_j = \mathbf{X}_j^t \mathbf{X}_{-j} \cdot \left(\mathbf{X}_{-j}^t \mathbf{X}_{-j}\right)^{-1} \cdot \mathbf{X}_{-j}^t \mathbf{X}_j$ due to $\widehat{\mathbf{X}}_j = \mathbf{X}_{-j} \cdot \left(\mathbf{X}_{-j}^t \mathbf{X}_{-j}\right)^{-1} \cdot \mathbf{X}_{-j}^t \mathbf{X}_j$.

Then:

$$\begin{aligned} R^2_{nc}(j) &= \frac{\mathbf{X}_j^t \mathbf{X}_{-j} \cdot \left(\mathbf{X}_{-j}^t \mathbf{X}_{-j}\right)^{-1} \cdot \mathbf{X}_{-j}^t \mathbf{X}_j}{\mathbf{X}_j^t \mathbf{X}_j}, \\ 1 - R^2_{nc}(j) &= \frac{\mathbf{X}_j^t \mathbf{X}_j - \mathbf{X}_j^t \mathbf{X}_{-j} \cdot \left(\mathbf{X}_{-j}^t \mathbf{X}_{-j}\right)^{-1} \cdot \mathbf{X}_{-j}^t \mathbf{X}_j}{\mathbf{X}_j^t \mathbf{X}_j}, \\ VIFnc(j) &= \frac{\mathbf{X}_j^t \mathbf{X}_j}{\mathbf{X}_j^t \mathbf{X}_j - \mathbf{X}_j^t \mathbf{X}_{-j} \cdot \left(\mathbf{X}_{-j}^t \mathbf{X}_{-j}\right)^{-1} \cdot \mathbf{X}_{-j}^t \mathbf{X}_j}. \end{aligned} \tag{7}$$

Thus, the VIFnc coincides with the expression given by Stewart [9] for the VIF and is denoted as k_j^2, that is, $VIFnc(j) = k_j^2$.

However, recently, Salmerón Gómez et al. [12] showed that the index presented by Stewart has been misleadingly identified as the VIF, verifying the following relation between both measures:

$$k_j^2 = VIF(j) + n \cdot \frac{\overline{\mathbf{X}}_j^2}{RSS_j}, \quad j = 2, \ldots, k, \tag{8}$$

where $\overline{\mathbf{X}}_j$ is the mean of the $j-$variable of $\mathbf{X}$. This expression is also shown by Salmerón Gómez et al. [10], where it is used to quantify the proportion of essential and nonessential multicollinearity existing in a concrete independent variable.

Note that the expression:

$$VIFnc(j) = VIF(j) + n \cdot \frac{\overline{\mathbf{X}}_j^2}{RSS_j}, \tag{9}$$

is obtained by Chennamaneni et al. [17] (expression (6) page 174), although it is also limited to the particular case of the moderated regression $\mathbf{Y} = \alpha_0 \cdot \mathbf{1} + \alpha_1 \cdot \mathbf{U} + \alpha_2 \cdot \mathbf{V} + \alpha_3 \cdot \mathbf{U} \times \mathbf{V} + \boldsymbol{\nu}$ where $\mathbf{U}$ and $\mathbf{V}$ are ratio-scaled explanatory variables in n-dimensional data vectors. Indeed, these authors proposed a new measure to detect multicollinearity in moderated regression models that is derived from the noncentered coefficient of determination. However, this use of the noncentered coefficient of determination lacks of the statistical contextualization provided by this paper

Finally, from expression (9), it is shown that the VIFnc and the VIF only coincide if the associated variable has zero mean, analogously to what happens in the decomposition of the sum of squares. Note that this expression also clarifies why Stewart's collinearity indices diminish when the variables are centered, which the author attributed to errors in regression variables: This phenomenon is a consequence of the fact that our definition of collinearity index compels us to work with relative errors.

Example 1. *Considering $k = 4$ in model (1), we use the noncentered coefficient of determination, R^2_{nc}, to calculate the noncentered variance inflation factor, $VIFnc$. For it, we consider the values displayed in Table 1.*

Note that variables $\mathbf{y}$, $\mathbf{X}_2$ *and* $\mathbf{X}_3$ *were originally used by Belsley [1] and we have added a new variable,* $\mathbf{X}_4$*, that has been randomly generated (from a normal distribution with a mean equal to 4 and a variance equal to 16) to obtain a variable that is linearly independent with respect to the rest.*

Table 1. Data set applied by Belsley [1].

y	1	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$
2.69385	1	0.996926	1.00006	8.883976
2.69402	1	0.997091	0.998779	6.432483
2.70052	1	0.9973	1.00068	−1.612356
2.68559	1	0.997813	1.00242	1.781762
2.7072	1	0.997898	1.00065	2.16682
2.6955	1	0.99814	1.0005	4.045509
2.70417	1	0.998556	0.999596	4.858077
2.69699	1	0.998737	1.00262	4.9045
2.69327	1	0.999414	1.00321	8.631162
2.68999	1	0.999678	1.0013	−0.4976853
2.70003	1	0.999926	0.997579	6.828907
2.702	1	0.999995	0.998597	8.999921
2.70938	1	1.00063	0.995316	7.080689
2.70094	1	1.00095	0.995966	1.193665
2.70536	1	1.00118	0.997125	1.483312
2.70754	1	1.00177	0.998951	−1.053813
2.69519	1	1.00231	1.00102	−0.5860236
2.7017	1	1.00306	1.00186	−1.371546
2.70451	1	1.00394	1.00353	−2.445995
2.69532	1	1.00469	1.00021	5.731981

In these data, the existence of nonessential multicollinearity is intuited. This fact is confirmed by the small values of the coefficient of variation (CV) in two of the independent variables and the following conclusions obtained from the value of the condition indices and the proportions of the variance (see, for example, Belsley et al. [18] and Belsley [16] for more details) shown in Table 2:

- *Variables* $\mathbf{X}_2$ *and* $\mathbf{X}_3$ *present a CV lower than* 0.06674082 *and than* 0.1002506 *that were presented by Salmerón Gómez et al. [10] as thresholds to indicate that a variable may be related to the constant and the model will present strong and moderate nonessential multicollinearity, respectively.*
- *The second index is associated with a high proportion of the variance with the variable* $\mathbf{X}_4$*, although it is not worrisome since it does not present a high value.*
- *The third index presents a value higher than the established thresholds (20 for moderate multicollinearity and 30 for strong multicollinearity), and it is also associated with high proportions in the variables* $\mathbf{X}_2$ *and* $\mathbf{X}_3$*.*
- *The last index identified as the condition number is clearly related to the intercept, and at the same time, it includes the relation between* $\mathbf{X}_2$ *and* $\mathbf{X}_3$ *as previously commented.*
- *Finally, the condition number, 1614.829, is higher than the threshold traditionally established as indicative of worrisome multicollinearity.*

Table 2. Diagnostic of collinearity of Belsley–Kuh–Welsch and coefficient of variation of the considered variables.

		Proportion of the Variance			
Eigenvalue	**Index of Condition**	**1**	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$
3.517205	1.000	0	0	0	0.022
0.4827886	2.699	0	0	0	0.784
4.978345×10^{-6}	840.536	0	0.423	0.475	0.003
1.348791×10^{-6}	1614.829	1	0.577	0.525	0.191
Coefficients of variation			0.002	0.002	1.141

Now, other models are proposed apart from the initial model for $k = 4$*:*

- *Model 0 (**Mod0**):* $\mathbf{y} = \beta_1 \cdot \mathbf{1} + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \cdot \mathbf{X}_3 + \beta_4 \cdot \mathbf{X}_4 + \mathbf{u}$.
- *Model 1 (**Mod1**):* $\mathbf{y} = \beta_1 \cdot \mathbf{1} + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \cdot \mathbf{X}_3 + \mathbf{u}$.
- *Model 2 (**Mod2**):* $\mathbf{y} = \beta_1 \cdot \mathbf{1} + \beta_2 \cdot \mathbf{X}_2 + \beta_4 \cdot \mathbf{X}_4 + \mathbf{u}$.
- *Model 3 (**Mod3**):* $\mathbf{y} = \beta_1 \cdot \mathbf{1} + \beta_3 \cdot \mathbf{X}_3 + \beta_4 \cdot \mathbf{X}_4 + \mathbf{u}$.

*Table 3 presents the VIF and the VIFnc of these models. Note that by using the original variables applied by Belsley (**Mod1**), the traditional VIF (from the centered model, see Theil [19]) provides a value equal to 1 (its minimum possible value), while the VIFnc is equal to 100,032.1. If the additional variable* $\mathbf{X}_4$ *is included (**Mod0**), the traditional VIFs are also close to one while the noncentered VIFs present values higher than 100,000. The conclusion is that the VIF is not detecting the existence of nonessential multicollinearity (see Salmerón et al. [8]) while the VIFnc "does detect it". However, since the calculation of VIFnc excludes the constant term, the detected relation refers to the one between* $\mathbf{X}_2$ *and* $\mathbf{X}_3$*, and not to the relation between* $\mathbf{X}_2$ *and/or* $\mathbf{X}_3$ *with the intercept.*

*This fact is supported by the values obtained for the VIF and VIFnc of the second and fourth variables (**Mod2**) and for the third and fourth variables (**Mod3**).*

Table 3. Variance inflation factor (VIF) and VIF uncentered (VIFnc) of models proposed from Belsley [1] dataset.

		$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$
Mod0	VIF	1.155	1.084	1.239
	VIFnc	100,453.8	100,490.6	1.737
Mod1	VIF	1	1	
	VIFnc	100,032.1	100,032.1	
Mod2	VIF	1.143		1.143
	VIFnc	1.765		1.765
Mod3	VIF		1.072	1.072
	VIFnc		1.766	1.766

2.3. What Kind of Multicollinearity Detects the VIFnc?

The results of Example 1 for **Mod0** suggest a new definition of nonessential multicollinearity as the relation between at least two variables with little variability. Thus, the particular case when one of these variables is the intercept leads to the definition initially given by Marquardt and Snee [6]. Then, the initial idea that in a noncentered model, is not possible to find nonessential collinearity is of a nuanced nature.

By following Salmerón et al. [8] and Salmerón Gómez et al. [10], it can be concluded that the VIF only detects the essential multicollinearity and, with these results, the VIFnc detects the nonessential multicollinearity but in its generalized definition since the intercept is eliminated in the corresponding auxiliary regression.

This fact is contradictory to the fact that the VIFnc coincides with the index of Stewart, see expression (7), since this measure is able to detect the nonessential multicollinearity (see Salmerón Gómez et al. [10]). This is because the VIFnc could be fooled, including the constant as an independent variable in a model without the intercept, that is:

$$\mathbf{y} = \beta_1 \cdot \mathbf{X}_1 + \beta_2 \cdot \mathbf{X}_2 + \cdots + \beta_k \cdot \mathbf{X}_k + \mathbf{u},$$

where $\mathbf{X}_1$ is a column of ones but is not considered as the intercept.

Example 2. *Now, we part from model 1 in the Belsley example but include the constant as an independent variable in a model without the intercept (**Mod4**) and two additional models (**Mod5** and **Mod6**):*

- *Model 4 (**Mod4**):* $\mathbf{y} = \beta_1 \cdot \mathbf{X}_1 + \beta_2 \cdot \mathbf{X}_2 + \beta_3 \cdot \mathbf{X}_3 + \mathbf{u}$.
- *Model 5 (**Mod5**):* $\mathbf{y} = \beta_1 \cdot \mathbf{X}_1 + \beta_2 \cdot \mathbf{X}_2 + \mathbf{u}$.

- *Model 6 (**Mod6**):* $\mathbf{y} = \beta_1 \cdot \mathbf{X}_1 + \beta_3 \cdot \mathbf{X}_3 + \mathbf{u}$.

Table 4 presents the VIFnc obtained from expression (5) in Models 4–6. Results indicate that, considering the centered model and calculating the coefficient of determination of the auxiliary regressions as if the model were noncentered, it is possible to detect the nonessential multicollinearity. Thus, the contradiction indicated at the beginning of this subsection is saved.

Table 4. VIFnc of Models 4–6 including the constant as an independent variable in a model without the intercept.

	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$
Mod4	400,031.4	199,921.7	200,158.3
Mod5	199,921.7	199,921.7	
Mod6		200,158.3	200,158.3

3. Effects of the Vifnc on the Statistical Analysis of the Model

Given the model (1), the expression obtained for the variance of the estimator is given by:

$$var(\widehat{\beta}_j) = \frac{\sigma^2}{RSS_j}, \quad j = 1, \ldots, k, \tag{10}$$

where RSS_j is the residual sum of squares of the auxiliary regression of the $j-$independent variable as a function of the rest of the independent variables (see expression (6)).

From expression (10), and considering that expression (7) can be rewritten as:

$$VIFnc(j) = \frac{\mathbf{X}_j^t \mathbf{X}_j}{RSS_j},$$

it is possible to obtain:

$$var(\widehat{\beta}_j) = \frac{\sigma^2}{RSS_j} = \frac{\sigma^2}{\mathbf{X}_j^t \mathbf{X}_j} \cdot VIFnc(j), \quad j = 1, \ldots, k. \tag{11}$$

Establishing a model as a reference is required to conclude whether the variance has been inflated (see, for example, Cook [20]). Thus, if the variables in $\mathbf{X}$ are orthogonal, it is verified that $\mathbf{X}^t\mathbf{X} = diag(d_1, \ldots, d_k)$ where $d_j = \mathbf{X}_j^t \mathbf{X}_j$. In this case, $\left(\mathbf{X}^t\mathbf{X}\right)^{-1} = diag(1/d_1, \ldots, 1/d_k)$, and consequently, the variance of the estimated coefficients in the hypothetical orthogonal case is given by the following expression:

$$var(\widehat{\beta}_{j,o}) = \frac{\sigma^2}{\mathbf{X}_j^t \mathbf{X}_j}, \quad j = 1, \ldots, k. \tag{12}$$

In this case:

$$\frac{var(\widehat{\beta}_j)}{var(\widehat{\beta}_{j,o})} = VIFnc(j), \quad j = 1, \ldots, k,$$

and it is then possible to state that the VIFnc is a factor that inflates the variance.

As consequence, high values of $VIFnc(j)$ imply high values of $var(\widehat{\beta}_j)$ and a tendency not to reject the null hypothesis in the individual significance test of model (1). Thus, the statistical analysis of the model will be affected.

Note from expression (11) that this negative effect can be offset by low values of the estimation of σ^2, that is, low values of the residual sum of squares of model (1) or high values of the number of observations, n. This is similar to what happen to the VIF (see O'Brien [7] for more details).

4. Auxiliary Centered Regressions

The use of the coefficient of determination of the auxiliary regression (6) where matrix $\mathbf{X}_{-j}$ contains a column of ones that represents the intercept is a very common approach to detect the linear relations between the independent variables of the model (1). This is motivated due to the higher relation between $\mathbf{X}_j$ and the rest of the independent variables, that is, the higher the multicollinearity is, the higher the value of that coefficient of determination.

However, since the coefficient of determination ignores the role of the intercept, this measure is unable to detect the nonessential linear relations. The question is evident: Does another measure exist related to the auxiliary regression that allows detection of the nonessential multicollinearity?

4.1. Case When There Is Only Nonessential Multicollinearity

Example 3. *Suppose that 100 observations are simulated for variables* $\mathbf{X}$, $\mathbf{Z}$ *and* $\mathbf{W}$ *from normal distributions with a mean of 5, 4 and -4 and a standard deviation of 0.01, 4 and 0.01, respectively. Note that* $\mathbf{X}$ *and* $\mathbf{W}$ *present light variability and, for this reason, it is expected that the model presents nonessential multicollinearity.*

Then, $\mathbf{y} = \mathbf{1} + \mathbf{X} + \mathbf{Z} - \mathbf{W} + \mathbf{v}$ *is generated by simulating* $\mathbf{v}$ *as a normal distribution with a mean equal to 0 and a standard deviation equal to 2.*

The second column of Table 5 presents the results obtained after the estimation by ordinary least squares (OLS) of model $\mathbf{y} = \beta_1 \cdot \mathbf{1} + \beta_2 \cdot \mathbf{X} + \beta_3 \cdot \mathbf{Z} + \beta_4 \cdot \mathbf{W} + \mathbf{u}$. *Note that the estimations of the coefficients of the model differ substantially from the real values used to generate* $\mathbf{y}$*, except for the coefficient of the variable* $\mathbf{Z}$ *(this situation illustrates the fact that if the interest is to estimate the effect of variable* $\mathbf{Z}$ *on* $\mathbf{y}$*, the analysis will not be influenced by the linear relations between the rest of the independent variables), which is the variable free of multicollinearity (indeed, it is the unique coefficient significantly different from zero, with a 5% significance—the value used by default in this paper).*

Table 5. Estimation by ordinary least squares (OLS) of the first simulated model and its corresponding auxiliary regressions (estimated standard deviation in parenthesis and coefficients significantly different from zero in bold).

Dependent Variable	$\widehat{\mathbf{y}}$	$\widehat{\mathbf{X}}$	$\widehat{\mathbf{Z}}$	$\widehat{\mathbf{W}}$
Intercept	173.135	**4.969**	−27.63	**−3.953**
	(123.419)	(0.369)	(240.08)	(0.557)
$\mathbf{X}$	−38.308		−17.05	−0.009
	(20.035)		(38.94)	(0.111)
$\mathbf{Z}$	0.939	−0.0001		−0.0002
	(0.052)	(0.0002)		(0.0002)
$\mathbf{W}$	−7.173	−0.007	−29.34	
	(18.2309)	(0.092)	(35.34)	
R^2	0.7773	0.001	0.008	0.007
VIF		1.001	1.008	1.007

This table also shows the results obtained from the estimations of the centered auxiliary regressions. Note that the coefficients of determination are very small, and consequently, the associated VIFs do not detect the degree of multicollinearity. However, note that in the auxiliary regressions corresponding to variables $\mathbf{X}$ *and* $\mathbf{W}$*:*

- *The estimation of the coefficient of the intercept almost coincides with the mean from which each variable was generated, 5 and −4, and, at the same time, the coefficients of the rest of the independent variables are almost zero.*
- *The estimations of the coefficients of the intercept are the unique ones that are significantly different from zero.*

Thus, note that the auxiliary regressions are capturing the existence of nonessential multicollinearity. The problem is that it is not transferred to its coefficient of determination but to another characteristic.

From this finding, it is possible to propose a way to detect the nonessential multicollinearity from the centered auxiliary regression traditionally applied to calculate the VIF:

Condition 1 (C1): Quantify the contribution of the estimation of the intercept to the total sum of the estimations of the coefficients of model (6), that is, calculate:

$$\frac{|\delta_1|}{\sum_{j=1}^{k-1}|\delta_j|}\cdot 100\%.$$

Condition 2 (C2): Calculate the number of independent variables with coefficients significantly different from zero and quantify the contribution of the intercept.

A Montecarlo simulation is presented considering the model (1) where $k = 3$ and the variable $\mathbf{X}_2$ has been generated as a normal distribution with mean $\mu_2 \in \mathbf{A}$ and variance $\sigma_2^2 \in \mathbf{B}$, the variable $\mathbf{X}_3$ has been generated as normal distribution with mean $\mu_3 \in \mathbf{A}$ and variance $\sigma_3^2 \in \mathbf{C}$ being $\mathbf{A} = \{0, 1, 2, 3, 4, 5, 10, 15, 20\}$, $\mathbf{B} = \{0.00001, 0.0001, 0.001, 0.1, \mathbf{C}\}$ and $\mathbf{C} = \{1, 2, 3, 4, 5, 10, 15, 20\}$. The results are presented in Table 6. Taking into account that the sample size has varied within the set $\{15, 20, 25, \ldots, 140, 145, 150\}$, 235872 iterations have been performed.

Table 6. Values of condition **C1** depending on the coefficient of variation (CV).

	P_5	P_{95}	Mean	Typical Deviation
$CV < 0.06674082$	99.402%	99.999%	99.512%	3.786%
$CV > 0.06674082$	52.678%	99.876%	89.941%	16.837%
$CV < 0.1002506$	95.485%	99.999%	98.741%	6.352%
$CV > 0.1002506$	51.434%	99.842%	89.462%	17.114%

Considering the thresholds established by Salmerón Gómez et al. [10], 90% of the simulations present values for condition **C1** between 99.402% and 99.999% if $CV < 0.06674082$ and between 95.485% and 99.999% if $CV < 0.1002506$. Thus, we can consider that values of condition **C1** higher than 95.485% will indicate that the auxiliary centered regressions are detecting the presence of nonessential multicollinearity.

Table 7 shows that a high value is obtained for the condition **C1**, even if any estimated coefficient is significantly different from zero (**C2** = NA).

Thus, the previous threshold, 95.485%, will be considered as valid if it is accompanied by a high value in the second condition.

Table 7. Values of condition **C1** depending on condition **C2**.

	C2	**NA**	**50%**	**100%**
C1	P_5	39.251%	67.861%	89.514%
	P_{95}	98.751%	99.984%	99.997%
	Mean	81.378%	91.524%	96.965%
	Typical Deviation	19.622%	13.598%	9.972%

Example 4. *Applying these criteria to the data of the Example 1 for* ***Mod1****, it is obtained that:*

- *In the auxiliary regression* $\mathbf{X}_2 = \delta_1 \cdot \mathbf{1} + \delta_3 \cdot \mathbf{X}_3 + \mathbf{w}$*, the estimation of the intercept is equal to 99.988% of the total, and the individual significance of the intercept corresponds to 100% of the significant estimated coefficients.*
- *In the auxiliary regression* $\mathbf{X}_3 = \delta_1 \cdot \mathbf{1} + \delta_2 \cdot \mathbf{X}_2 + \mathbf{w}$*, the estimation of the intercept is equal to 99.988% of the total, and the individual significance of the intercept corresponds to 100% of the significant estimated coefficients.*

Thus, the symptoms shown in the previous simulation also appear, and consequently, in both situations, the nonessential multicollinearity will be detected.

Replicating both situations where the VIFnc was not able to detect the nonessential multicollinearity, it is obtained that:

- *For **Mod2** it is obtained that:*
  - *In the auxiliary regression $\mathbf{X}_2 = \delta_1 \cdot \mathbf{1} + \delta_4 \cdot \mathbf{X}_4 + \mathbf{w}$, the estimation of the intercept is equal to the 99.978% of the total, and the individual significance of the intercept corresponds to 100% of the significant estimated coefficients.*
  - *In the auxiliary regression $\mathbf{X}_4 = \delta_1 \cdot \mathbf{1} + \delta_2 \cdot \mathbf{X}_2 + \mathbf{w}$, the estimation of the intercept is equal to 50.138% of the total, and none of the estimated coefficients are significantly different from zero.*
- *For **Mod3** it is obtained that:*
  - *In the auxiliary regression $\mathbf{X}_3 = \delta_1 \cdot \mathbf{1} + \delta_4 \cdot \mathbf{X}_4 + \mathbf{w}$, the estimation of the intercept is equal to 99.984% of the total, and the individual significance of the intercept corresponds to 100% of the significant estimated coefficients.*
  - *In the auxiliary regression $\mathbf{X}_4 = \delta_1 \cdot \mathbf{1} + \delta_3 \cdot \mathbf{X}_3 + \mathbf{w}$, the estimation of the intercept is equal to 50.187% of the total, and none of the estimated coefficients are significantly different from zero.*

Once again, it was shown that with this procedure, it is possible to detect the nonessential multicollinearity and the variables that are causing it.

4.2. Relevance of a Variable in a Regression Model

Note that the conditions **C1** and **C2** are focused on measuring the relevance of one of the variables, in this case, the intercept, within the multiple linear regression model. It is interesting to analyze the behavior of other measures with this same goal as, for example, the index ι_j of Stewart [9]. Given model (1), Stewart defined the relevance of the $j-$variable as the number:

$$\iota_j = \frac{|\beta_j| \cdot ||\mathbf{X}_j||}{||\mathbf{y}||}, \quad j = 1, \ldots, p,$$

where $||\cdot||$ is the usual Euclidean norm. Stewart considered that a variable with a relevance higher than 0.5 should not be ignored.

Example 5. *Table 8 presents the calculation of ι_j for situations shown in Example 1. Note that in all cases, the intercept will be considered relevant, even when the variable $\mathbf{X}_4$ is analyzed as a function of $\mathbf{X}_2$ or $\mathbf{X}_3$, despite that it was previously shown that the intercept was not relevant in these situations (at least in relation to nonessential multicollinearity).*

Table 8. Calculation of ι_j for situations **Mod1**, **Mod2** and **Mod3** shown in Example 1.

	Auxiliary Regression	ι_1	ι_2
Mod1	$\mathbf{X}_2 = \delta_1 \cdot \mathbf{1} + \delta_3 \cdot \mathbf{X}_3 + \mathbf{w}$	0.999	0.0001
	$\mathbf{X}_3 = \delta_1 \cdot \mathbf{1} + \delta_2 \cdot \mathbf{X}_2 + \mathbf{w}$	0.999	0.0001
Mod2	$\mathbf{X}_2 = \delta_1 \cdot \mathbf{1} + \delta_4 \cdot \mathbf{X}_4 + \mathbf{w}$	1.0006	0.001
	$\mathbf{X}_4 = \delta_1 \cdot \mathbf{1} + \delta_2 \cdot \mathbf{X}_2 + \mathbf{w}$	119.715	119.056
Mod3	$\mathbf{X}_3 = \delta_1 \cdot \mathbf{1} + \delta_4 \cdot \mathbf{X}_4 + \mathbf{w}$	1.0005	0.0007
	$\mathbf{X}_4 = \delta_1 \cdot \mathbf{1} + \delta_3 \cdot \mathbf{X}_3 + \mathbf{w}$	88.346	87.687

*Thus, the application of ι_j seems not to be appropriate contrarily to what happens with conditions **C1** and **C2**.*

4.3. Case When There Is Generalized Nonessential Multicollinearity

Example 6. *Suppose that the previous simulation is repeated, except for the generation of the variable $\mathbf{Z}$, which, in this case, is considered to be given by $Z_i = 2 \cdot X_i - a_i$, for $i = 1, \ldots, 100$, where a_i is generated from a normal distribution with a mean equal to 2 and a standard deviation equal to 0.01.*

Table 9 presents the results of the estimation by OLS of the model $\mathbf{y} = \beta_1 \cdot \mathbf{1} + \beta_2 \cdot \mathbf{X} + \beta_3 \cdot \mathbf{Z} + \beta_4 \cdot \mathbf{W} + \mathbf{u}$ *and its possible auxiliary regressions.*

In this case, none of the coefficients are significantly different from zero and the coefficients are very far from the real values used in the simulation.

Table 9. Estimation by OLS of the second simulated model and its corresponding auxiliary regressions (estimated standard deviation in parenthesis and coefficients significantly different from zero in bold).

Dependent Variable	$\widehat{\mathbf{y}}$	$\widehat{\mathbf{X}}$	$\widehat{\mathbf{Z}}$	$\widehat{\mathbf{W}}$
Constant	−233.37	**1.977**	**−2.638**	**−4.959**
	(167.33)	(0.2203)	(0.673)	(0.715)
X	12.02		**2.213**	−0.059
	(56.98)		(0.102)	(0.298)
Z	8.89	**0.374**		0.156
	(23.44)	(0.017)		(0.121)
W	−29.96	−0.006	0.107	
	(19.41)	(0.034)	(0.107)	
R^2	0.034	0.838	0.841	0.073
VIF		6.172	6.289	1.078

In relation to the auxiliary regression, it is possible to conclude that:

- *When the dependent variable is* **X**, *the coefficients that are significantly different from zero are the ones of the intercept and the variable* **Z**. *At the same time, the estimation of the coefficient of the intercept differs from the mean from which the variable* **X** *was generated. In this case, the contribution of the estimation of the intercept is equal to 83.837% of the total and represents 50% of the coefficients significantly different from zero.*
- *When the dependent variable is* **Z**, *the coefficients significantly different from zero are the ones of the intercept and the variable* **X**. *In this case, the contribution of the estimation of the intercept is equal to 53.196% of the total and represents 50% of the coefficients significantly different from zero.*
- *When the dependent variable is* **W**, *the signs shown in the previous section are maintained. In this case, the contribution of the intercept is equal to 95.829% of the total and represents 100% of the coefficients significantly different from zero.*
- *Finally, although it will require a deeper analysis, the last results indicate that the estimated coefficient that is significantly different from zero in the auxiliary regression represents the variables responsible for the existing linear relation (intercept included).*

Note that the existence of generalized nonessential multicollinearity distorts the symptoms previously detected. Thus, the fact that in a centered auxiliary regression, the contribution (in absolute terms) of the estimation of the intercept to the total sum (in absolute value) of all estimations will be close to 100%, and the estimation of the intercept will be uniquely significantly different from zero, are indications of nonessential multicollinearity. However, it is possible that these symptoms are not manifested but there exists worrisome nonessential multicollinearity. Thus, these conditions are sufficient but not required.

However, in situations shown in Example 6 where conditions **C1** and **C2** are not verified, the VIFnc will be equal to 1109,259.3, 758,927.7 and 100,912.7. Thus, note that these results complement the results presented in the previous section in relation to the VIFnc. Thus, VIFnc detects generalized nonessential multicollinearity while conditions **C1** and **C2** detect the traditional nonessential multicollinearity given by Marquardt and Snee [6].

5. Empirical Applications

In order to illustrate the contribution of this study, this section presents two empirical applications with financial and economic real data. Note that in a financial prediction model, a financial variable

with low variance means low risk and a better prediction, because the standard deviation and volatility are lower. However, as discussed above, a lower variance of the independent variable may mean greater nonessential multicollinearity in a GLR model. Thus, the existence of worrisome nonessential collinearity may be relatively common in financial econometric models and this idea can be extended in general to economic applications. Note that the objective is to diagnose the type of multicollinearity existing in the model and indicate the most appropriate treatment (without applying it).

5.1. Financial Empirical Application

The following model of Euribor (**100%**) is specified from the data set composed by 47 Eurozone observations for the period January 2002 to July 2013 (quarterly and seasonally adjusted data) and previously applied by Salmerón Gómez et al. [10]:

$$\textbf{Euribor} = \beta_1 + \beta_2 \cdot \textbf{HICP} + \beta_3 \cdot \textbf{BC} + \textbf{u}, \tag{13}$$

where **HICP** is the Harmonized Index of Consumer Prices (**100%**), **BC** is the Balance of Payments to net current account (millions of euros) and **u** is a random disturbance (centered, homoscedastic, and uncorrelated).

Table 10 presents the analysis of model (13) and its corresponding auxiliary regressions. The values of the VIFs which are very close to one will indicate that there is not essential multicollinearity. The correlation coefficient between **HICP** and **BC** is 0.231 and the determinant of the correlation matrix is 0.946. Both values indicate that there is no essential multicollinearity, see García García et al. [21] and Salmerón Gómez et al. [22].

However, the condition number is higher than 30 indicating a strong multicollinearity associated, see conditions **C1** and **C2**, with variable **HICP**. The values of conditions **C1** and **C2** are conclusive in the case of variable **HICP**. In the case of variable **BC**, although condition **C1** presents a high value, none of the coefficients of the auxiliary regression is significatively different from zero (condition **C2**). By following the simulation presented in subsection, this indicate that the variable **BC** is not related to the intercept. This conclusion is in line with the value of the coefficient of variation of variable **HICP** that is lower than 0.1002506, the threshold established by Salmerón Gómez et al. [10] for moderate nonessential multicollinearity.

Table 11 presents the calculation of the VIFnc. Note that it is not detecting the non-essential multicollinearity. As previously commented, the VIFnc only detects the essential and the generalized nonessential multicollinearity. This table also presents the VIFnc calculated in a model without intercept but including the constant as an independent variable (see Section 2.3). In this case, the VIFnc is able to detect the nonessential multicollinearity between the intercept and the variable **HIPC**.

In conclusion, this model will present nonessential multicollinearity caused by the variable **HICP**. This problem can be mitigated by centering that variable (see, for example, Marquardt and Snee [6] and Salmerón Gómez et al. [10]).

Table 10. Estimations by OLS of model (13) and its corresponding auxiliary regressions (estimated standard deviation in parenthesis and coefficients significantly different from zero in bold).

	Euribor	**HICP**	**BC**
Intercept	**8.442**	**104.8**	−64,955
	(1.963)	(1.09)	(43,868)
HICP	**−0.054**		663.3
	(0.018)		(415.9)
BC	$\mathbf{-3.493 \times 10^{-5}}$	8.065×10^{-5}	
	(6.513×10^{-6})	(5.057×10^{-5})	
R^2	0.517	0.053	0.053
VIF		1.055	1.055
CN	30.246		
Condition 1 (**C1**)		99.999%	98.98%
Condition 2 (**C2**)		100%	NA
Coefficients of variation		0.069	4.3403

Table 11. VIFnc of auxiliary regressions associated to model (13).

	$\mathbf{X_1}$	**HICP**	**BC**
VIFnc		1.0609	1.0609
VIFnc	217.672	219.291	1.112

5.2. Economic Empirical Application

From French economy data from Chatterjee and Hadi [23], also analyzed by Malinvaud [24], Zhang and Liu [25] and Kibria and Lukman [26], among others, the following model is analyzed:

$$\mathbf{I} = \beta_1 + \beta_2 \cdot \mathbf{DP} + \beta_3 \cdot \mathbf{SF} + \beta_4 \cdot \mathbf{DC} + \mathbf{u}, \tag{14}$$

for years 1949 through 1966 where imports (**I**), domestic production (**DP**), stock formation (**SF**) and domestic consumption (**DC**), all are measured in billions of French francs and **u** is a random disturbance (centered, homoscedastic, and uncorrelated).

Table 12 presents the analysis of model (14) and its corresponding auxiliary regressions. The values of the VIFs of variables **DP** and **DC** indicate strong essential multicollinearity. The condition number is higher than 30 also indicating a strong multicollinearity.

Note that the values of condition **C1** for variables **DP** and **DC** are lower than threshold shown in the simulation. Only the variable **SF** presents a higher value but, in this case, condition **C2** indicates that none of the estimated coefficients of the auxiliary regression are significatively different from zero. This conclusion is in line with the coefficients of variation that are higher than the threshold established by Salmerón Gómez et al. [10] indicating that there is no nonessential multicollinearity.

Table 13 presents the calculation of the VIFnc. Note that it is detecting the essential multicollinearity. This table also presents the VIFnc calculated in a model without intercept but including the constant as an independent variable. In this case, the VIFnc is also detecting the essential multicollinearity between the variables **DP** and **DC**. From thresholds established by Salmerón Gómez et al. [10] for simple linear regression ($k = 2$), the value 60.0706 will not be worrisome and, consequently, the nonessential multicollinearity will not be worrisome.

Table 12. Estimations by OLS of Model (14) and its corresponding auxiliary regressions (estimated standard deviation in parenthesis and coefficients significantly different from zero in bold).

	I	**DP**	**SF**	**DC**
Intercept	−**19.725**	−**18.052**	2.635	**12.141**
	(4.125)	(3.28)	(3.234)	(2.026)
DP	0.032		0.025	**0.654**
	(0.186)		(0.149)	(0.007)
SF	0.414	0.075		−0.038
	(0.322)	(0.444)		(0.291)
DC	0.242	**1.525**	−0.029	
	(0.285)	(0.018)	(0.228)	
R^2	0.973	0.997	0.047	0.997
VIF		333.333	1.049	333.333
CN	247.331			
Condition 1 (**C1**)		91.85%	97.94%	94.6%
Condition 2 (**C2**)		50%	NA	50%
Coefficients of variation		0.267	0.473	0.248

Table 13. VIFnc of auxiliary regressions associated to Model (14).

	$\mathbf{X}_1$	**DP**	**SF**	**DC**
VIFnc		2457.002	5.753	2512.562
VIFnc	60.0706	7424.705	6.008	8522.1308

To conclude, this model presents essential multicollinearity caused by the variables **DP** and **DC**. In this case, the problem will be mitigated by applying estimation methods other than OLS such as ridge regression (see, for example, Hoerl and Kennard [27], Hoerl et al. [28], Marquardt [29]), LASSO regression (see Tibshirani [30]), raise regression (see, for example, García et al. [31], Salmerón et al. [32], García and Ramírez [33], Salmerón et al. [34]), residualization (see, for example, York [35], García et al. [36]) or the elastic net regularization (see Zou and Hastie [37]).

6. Conclusions

The distinction between essential and nonessential multicollinearity and its diagnosis has not been not been adequately treated in either the scientific literature or in statistical software and this lack of information has led to mistakes in some relevant papers, for example Velilla [3] or Jensen and Ramirez [13]. This paper analyzes the detection of essential and nonessential multicollinearity from auxiliary centered and noncentered regressions, obtaining two complementary measures between them that are able to detect both kinds of multicollinearity. The relevance of the results is that they are obtained within an econometric context, encompassing the distinction between centered and noncentered models that is not only accomplished from a numerical perspective, as was the case presented, for example, in Salmerón Gómez et al. [12] or Salmerón Gómez et al. [10]. An undoubtedly interesting point of view of this situation is the one presented by Spanos [38] that stated: It is argued that many confusions in the collinearity literature arise from erroneously attributing symptoms of statistical misspecification to the presence of collinearity when the latter is misdiagnosed using unreliable statistical measures. That is, the distinction related to the econometric model provides confidence to the measures of detection and avoids the problems commented by Spanos.

From a computational point of view, this debate clarifies what is calculated when the VIF is obtained for centered and noncentered models. It also clarifies, see Section 2.3, what type of multicollinearity is detected (and why) when the uncentered VIF is calculated in a centered model. At the same time, a definition of nonessential multicollinearity is presented that generalizes the definition given by Marquardt and Snee [6]. Note that this generalization can be understood as a particular kind of essential multicollinearity:

A near-linear relation between two independent variables with light variability. However, it is shown that this kind of multicollinearity is not detected by the VIF, and for this reason, we consider it more appropriate to include it within the nonessential multicollinearity.

In relation to the application of the VIFnc, this paper shows that the VIFnc detects the essential and the generalized nonessential multicollinearity and even the traditional nonessential multicollinearity if it is calculated in a regression without the intercept but including the constant as an independent variable. Note that the VIF, although widely applied in many different fields, only detects the essential multicollinearity. This paper has also analyzed why the VIF is unable to detect the nonessential multicollinearity, and two conditions are presented as sufficient (but not required) to establish the existence of nonessential multicollinearity. Since these conditions, **C1** and **C2**, are based on the relevance of the intercept within the centered auxiliary regression to calculate the VIF, this scenario was compared to the measure proposed by Stewart [9], ι_j, to measure the relative importance of a variable within a multiple linear regression. It is shown that conditions **C1** and **C2** are preferable to the calculation of ι_j.

To summarize:

- A centered model can present essential, generalized nonessential and traditional nonessential collinearity (given by Marquardt and Snee [6]) while in a noncentered model only it is only possible to find the essential and the generalized nonessential collinearity.
- The VIF only detects the essential collinearity, the VIFnc detects the generalized nonessential and essential collinearity and the conditions **C1** and **C2** the traditional nonessential collinearity.
- When there is generalized nonessential collinearity it is understood that there is also traditional nonessential collinearity, but this is not detected by the conditions **C1** and **C2**. Thus, in this case it is necessary to use other alternative measures as the coefficient of variation of the condition number.

To conclude, in order to detect the kind of multicollinearity and its degree, the greatest number of measures must be used (variance inflation factors, condition number, correlation matrix and its determinant, coefficient of variation, conditions **C1** and **C2**, etc.) as in Section 5, and it is inefficient to limit oneself to the management of only a few. Similarly, it is necessary to know what kind of multicollinearity is capable of detecting each one of them.

Finally, the following will be interesting as future lines of inquiry:

- to establish the threshold for the VIFnc,
- to extend the Montecarlo simulation of Section 4.1 for models with $k > 3$ regressors,
- a deeper analysis to conclude if the variable responsible for the existing linear relation can be identified as the one whose estimated coefficient is significantly different from zero in the auxiliary regression (see Example 6) and
- the development of a specific package in R Core Team [39] to perform the calculation of VIFnc and conditions **C1** and **C2**.

Author Contributions: Conceptualization, R.S.G. and C.G.G.; methodology, R.S.G.; software, R.S.G.; validation, R.S.G., C.G.G. and J.G.P.; formal analysis, R.S.G. and C.G.G.; investigation, R.S.G., C.G.G. and J.G.P.; resources, R.S.G., C.G.G. and J.G.P.; writing—original draft preparation, R.S.G. and C.G.G. ; writing—review and editing, R.S.G. and C.G.G.; supervision, J.G.P.; project administration, R.S.G.; funding acquisition, J.G.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by University of Almería.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Belsley, D.A. Demeaning conditioning diagnostics through centering. *Am. Stat.* **1984**, *38*, 73–77.
2. Marquardt, D.W. A critique of some ridge regression methods: Comment. *J. Am. Stat. Assoc.* **1980**, *75*, 87–91.
3. Velilla, S. A note on collinearity diagnostics and centering. *Am. Stat.* **2018**, *72*, 140–146.

4. Cohen, P.; West, S.G.; Aiken, L.S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*; Psychology Press: London, UK, 2014.
5. Marquardt, D. You should standardize the predictor variables in your regression models. Discussion of: A critique of some ridge regression methods. *J. Am. Stat. Assoc.* **1980**, *75*, 87–91.
6. Marquardt, D.W.; Snee, R.D. Ridge regression in practice. *Am. Stat.* **1975**, *29*, 3–20.
7. O'Brien, R. A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **2007**, *41*, 673–690.
8. Salmerón, R.; García, C.; García, J. Variance Inflation Factor and Condition Number in multiple linear regression. *J. Stat. Comput. Simul.* **2018**, *88*, 2365–2384.
9. Stewart, G. Collinearity and least squares regression. *Stat. Sci.* **1987**, *2*, 68–84.
10. Salmerón Gómez, R.; Rodríguez, A.; García García, C. Diagnosis and quantification of the non-essential collinearity. *Comput. Stat.* **2020**, *35*, 647–666.
11. Marquardt, D.W. [Collinearity and Least Squares Regression]: Comment. *Stat. Sci.* **1987**, *2*, 84–85.
12. Salmerón Gómez, R.; García García, C.; García Pérez, J. Comment on A Note on Collinearity Diagnostics and Centering by Velilla (2018). *Am. Stat.* **2019**, 114–117. doi:10.1080/00031305.2017.1392896.
13. Jensen, D.R.; Ramirez, D.E. Revision: Variance inflation in regression. *Adv. Decis. Sci.* **2013**, *2013*, 671204.
14. Grob, J. *Linear Regression*; Springer: Berlin, Germany, 2003.
15. Cook, R. Variance Inflations Factors. *R News Newsl. R Proj.* **2003**, *3*, 13–15.
16. Belsley, D.A. A guide to using the collinearity diagnostics. *Comput. Sci. Econ. Manag.* **1991**, *4*, 33–50.
17. Chennamaneni, P.R.; Echambadi, R.; Hess, J.D.; Syam, N. Diagnosing harmful collinearity in moderated regressions: A roadmap. *Int. J. Res. Mark.* **2016**, *33*, 172–182.
18. Belsley, D.A.; Kuh, E.; Welsch, R.E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; Wiley: New York, NY, USA, 1980.
19. Theil, H. *Principles of Econometrics*; Wiley: New York, NY, USA, 1971; Volume 4.
20. Cook, R. [Demeaning Conditioning Diagnostics through Centering]: Comment. *Am. Stat.* **1984**, *38*, 78–79.
21. García García, C.; Salmerón Gómez, R.; García García, C. Choice of the ridge factor from the correlation matrix determinant. *J. Stat. Comput. Simul.* **2019**, *89*, 211–231.
22. Salmerón Gómez, R.; García García, C.; García García, J. A Guide to Using the R Package "multiColl" for Detecting Multicollinearity. *Comput. Econ.* **2020**. doi:10.1007/s10614-019-09967-y.
23. Chatterjee, S.; Hadi, A.S. *Regression Analysis by Example*; John Wiley & Sons: Hoboken, NY, USA, 2015.
24. Malinvaud, E. *Statistical Methods of Econometrics*; North Holland: New York, NY, USA, 1980.
25. Zhang, W.; Liu, L. A New Class of Biased Estimate in the Linear Regression Model. *J. Wuhan Univ. Nat. Sci. Ed.* **2006**, *52*, 281.
26. Kibria, B.; Lukman, A.F. A New Ridge-Type Estimator for the Linear Regression Model: Simulations and Applications. *Scientifica* **2020**, *2020*, 9758378 .
27. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67.
28. Hoerl, A.; Kannard, R.; Baldwin, K. Ridge regression: Some simulations. *Commun. Stat. Theory Methods* **1975**, *4*, 105–123.
29. Marquardt, D. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics* **1970**, *12*, 591–612.
30. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288.
31. García, C.G.; Pérez, J.G.; Liria, J.S. The raise method. An alternative procedure to estimate the parameters in presence of collinearity. *Qual. Quant.* **2011**, *45*, 403–423.
32. Salmerón, R.; García, C.; García, J.; López, M.d.M. The raise estimator estimation, inference, and properties. *Commun. Stat. Theory Methods* **2017**, *46*, 6446–6462.
33. García, J.; Ramírez, D. The successive raising estimator and its relation with the ridge estimator. *Commun. Stat. Theory Methods* **2017**, *46*, 11123–11142.
34. Salmerón, R.; Rodríguez, A.; García, C.; García, J. The VIF and MSE in Raise Regression. *Mathematics* **2020**, *8*, 605–633.
35. York, R. Residualization is not the answer: Rethinking how to address multicollinearity. *Soc. Sci. Res.* **2012**, *41*, 1379–1386.
36. García, C.; Salmerón, R.; García, C.; García, J. Residualization: Justification, properties and application. *J. Appl. Stat.* **2017**. doi:10.1080/02664763.2019.1701638.

37. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.* **2005**, *67*, 301–320.
38. Spanos, A. Near-collinearity in linear regression revisited: The numerical vs. the statistical perspective. *Commun. Stat. Theory Methods* **2019**, 48, 5492–5516.
39. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.