Computational Statistics

Diagnosis and quantification of the non-essential collinearity

--Manuscript Draft--

Manuscript Number:	COST-D-18-00382R2
Full Title:	Diagnosis and quantification of the non-essential collinearity
Article Type:	Original Paper
Keywords:	multicollinearity, multiple linear regression, non-essential multicollinearity, centered variables
Corresponding Author:	Catalina García Universidad de Granada SPAIN
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universidad de Granada
Corresponding Author's Secondary Institution:	
First Author:	ROMAN SALMERON
First Author Secondary Information:	
Order of Authors:	ROMAN SALMERON
	AINARA RODRIGUEZ
	Catalina García
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	Marquandt and Snee (1975), Marquandt (1980) and Snee and Marquardt (1984) refer to non-essential multicollinearity as that caused by the relation with the independent term. Although it is clear that the solution is to center the independent variables in the regression model, it is unclear when this kind of collinearity exists. The goal of this study is to diagnose the non-essential collinearity parting from a simple linear model. The collinearity indices kj , traditionally misinterpreted as Variance Inflation Factors (VIF), are reinterpreted in this paper where they will be used to distinguish and quantify the essential and non-essential collinearity. The results can be immediately extended to the multiple linear model. The study also has some recommendations for statistical software such as SPSS, Stata, GRETL or R for improving the diagnosis of non-essential collinearity.

Noname manuscript No.
(will be inserted by the editor)

Diagnosis and quantification of the non-essential collinearity

Román Salmerón Gómez · Ainara Rodríguez Sánchez · Catalina García García

Received: date / Accepted: date

Abstract Marquandt and Snee (1975), Marquandt (1980) and Snee and Marquardt (1984) refer to non-essential multicollinearity as that caused by the relation with the independent term. Although it is clear that the solution is to center the independent variables in the regression model, it is unclear when this kind of collinearity exists. The goal of this study is to diagnose the non-essential collinearity parting from a simple linear model. The collinearity indices k_j, traditionally misinterpreted as Variance Inflation Factors (VIF), are reinterpreted in this paper where they will be used to distinguish and quantify the essential and non-essential collinearity. The results can be immediately extended to the multiple linear model. The study also has some recommendations for statistical software such as SPSS, Stata, GRETL or R for improving the diagnosis of non-essential collinearity.
Keywords: multicollinearity, multiple linear regression, non-essential multicollinearity, centered variables

R. Salmerón
Department of Quantitative methods for economics and business, Campus Universitario de la Cartuja 18071 Granada (Spain), University of Granada
E-mail: romansg@ugr.es

A. Rodríguez
Phd. student at University of Granada, Campus Universitario de La Cartuja, 18071 Granada (Spain)

C. García
Department of Quantitative methods for economics and business, Campus Universitario de la Cartuja 18071 Granada (Spain), University of Granada
E-mail: cbgarcia@ugr.es

1 INTRODUCTION

General linear regression (GLR) models are widely applied to analyze the relation between a dependent variable ($\mathbf{Y}$) and a set of regressors ($\mathbf{X}_1, \ldots, \mathbf{X}_p$, $p \geq 1$). This relation allows us to quantify the value of the dependent variable based on the values of the regressors. The model is defined for n observations and p regressors, as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{1}$$

where $\mathbf{u}$ is the random disturbance (which is supposed to be spherical), $\mathbf{X}_{n\times p} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$ is a matrix of observations of the regressors, $\mathbf{Y}_{n\times 1}$ is a vector of the observations of the dependent variable and $\boldsymbol{\beta}_{p\times 1} = (\beta_1, \ldots, \beta_p)^t$ is a vector of coefficients regressors.

Multicollinearity exists in the data when there is a strong linear relationship between the regressors. Depending on the degree, collinearity is said to be exact or approximate. Following Stock and Watson (2012), we say there is exact collinearity if one of the regressors is a perfect linear function of the remaining regressors (or one of them). In this case, an estimation is not possible. On the other hand, approximate multicollinearity arises when one of the regressors is highly, but not perfectly correlated with the other regressors. Johnston and Dinardo (2001) indicate that regressors are often close to linear dependence, in which case the ordinary least squares (OLS) method can be applied, although the estimators may present very high standard errors.

The concept of multicollinearity is relatively clear. However, the definition of an independent or explanatory variable as a synonym of regressor is not as clear because some authors consider the constant term to be an independent variable, while others do not.

According to Johnston and Dinardo (2001), in the model given by (1), *the equation identifies* $p-1$ *explanatory variables or regressors (*$\mathbf{X_2}, \mathbf{X_3}. \ldots, \mathbf{X_p}$*) that influence the dependent variable. The vector* $\mathbf{X_1}$ *is a column of ones, allowing the existence of an intercept term in the equation.* This seems to indicate that the constant term is not considered as an independent variable. This reasoning is supported by other authors, such as Wooldridge (2009), who indicates that the linear regression model (1) can be expressed as follows:

$$\mathbf{Y} = \beta_1 + \beta_2\mathbf{X}_2 + \beta_3\mathbf{X}_3 + \cdots + \beta_p\mathbf{X}_p + \mathbf{u}, \tag{2}$$

where there are $p-1$ independent variables and a constant term. Stock and Watson (2012) presented a similar interpretation of model (2). Therefore, these authors seem to corroborate the idea that the constant term is not considered as an independent variable.

However, for example, Uriel et al. (1997) propose that to homogenize the treatment of the regressors in model (1), the intercept term should be multiplied by the regressor $\mathbf{X_1}$, which always takes values equal to one. Novales (1993) and Gujarati (2003) interpret the regression model as incorporating a constant term that accompanies a first explanatory variable $\mathbf{X}_1$, the value of which is always one ($X_{1t} = 1$, $t = 1, 2, 3, \ldots, n$).

To clarify the notation, it is considered that the intercept is an independent variable and the linear regression model is defined as expression (2) considering n observations, p independent/explanatory variables or regressors (as synonyms) and $\mathbf{X}_{1t} = 1$, for $t = 1, \ldots, n$, representing the intercept. With this consideration, two kinds of near multicollinearity can be distinguished:

– Non-essential: near-linear relationship between the intercept with at least one of the remaining independent variables.
– Essential: near-linear relationship between at least two independent variables excluding the intercept.

This distinction is relevant, as statistical packages, widely used in econometrics, such as GRETL or R (package *car*) do not allow the measurement of the degree of multicollinearity existing in the simple linear regression (SLR, model (2) for p=2). That is, they deny the existence of non-essential multicollinearity. In contrast, this study begins from the SLR to diagnose the occurrence of non-essential multicollinearity in multiple linear regression. Thus, considering that in the SLR model could be presented collinearity, the next step is to apply measures to diagnose its presence. This study analyzes the application of the Variance Inflation Factor (VIF) and the Condition Number (CN) to a SLR.

On the other hand, Novales (1993) states that collinearity in an SLR means that the explanatory variable $\mathbf{X}_2$ is approximately constant. At the same time, Christensen (2018) stated that *Gunst (1984) pointed out that collinearity of a single variable with the intercept is easily diagnosed by a small coefficient of variation.* Thus, we also examine how small the variance (or coefficient of variation) of the independent variable must be to cause worrying collinearity. Finally, the collinearity detected in a SLR, can only be non-essential. Thus, the method of detecting collinearity in the SLR will also detect non-essential collinearity in a GLR.

The remainder of this paper is structured as follows. Section 2 analyzes the application of the VIF and the CN to a SLR. As alternative measures, and following Novales (1993) and Gunst (1984), section 3 examines how small the variance or coefficient of variation of $\mathbf{X_2}$ must be to cause serious collinearity. This analysis is interesting because the calculation of these measures is simpler than that of other diagnosis measures. Section 4 analyzes the indices presented by Stewart (1987) noting that traditionally they have been wrongly identified as variance inflation factors and showing the application to quantify the essential and non-essential collinearity. The application to the multiple linear regression is shown in section 5 using an empirical example in the field of finance. In this case, the variance is used as a risk measure. Finally, section 6 concludes the paper.

2 COLLINEARITY DIAGNOSTICS

Given model (1), where X_{1t} is equal to 1 for $t = 1, \ldots, n$, the most popular measures used to diagnose collinearity are presented below:

– The Variance Inflation Factor (VIF):

$$VIF(i) = \frac{var(\hat{\beta}_i)}{var(\hat{\beta}_i^0)} = \frac{1}{1 - R_i^2}, \quad i = 2, \ldots, p, \tag{3}$$

where $\hat{\boldsymbol{\beta}}$ is the OLS estimator of model (1), $\hat{\boldsymbol{\beta}}^0$ is the OLS estimator of model (1) supposing that the independent variables are orthogonal and R_i^2 is the coefficient of determination of the following auxiliary regression:

$$\mathbf{X}_i = \mathbf{X}_{-i}\boldsymbol{\delta} + \mathbf{w}, \tag{4}$$

where $\mathbf{X}_{-i}$ is equal to matrix $\mathbf{X}$, after eliminating variable $\mathbf{X}_i$, for $i = 2, \ldots, p$. Because $0 \leq R_i^2 \leq 1$, it is verified that $VIF(i) \geq 1,\ i = 2, \ldots, p$. VIF values greater than 10 indicate that the linear regression model presents a significant degree of collinearity. As pointed out by Curto and Pinto (2011), the real impact on variance can be overestimated by the traditional VIF. To solve this, Curto and Pinto (2011) presented a corrected version known as the corrected VIF (CVIF) that was posteriorly improved by Salmerón et al. (2017). In spite of these alternatives, the VIF remains the most widespread collinearity measure.

– The Condition Number (CN):

$$K(X) = \sqrt{\frac{\xi_{max}}{\xi_{min}}}, \tag{5}$$

where ξ_{max} and ξ_{min} are the maximum and minimum eigenvalues of matrix $\mathbf{X^tX}$, respectively. Note that before calculating the eigenvalues, the matrix $\mathbf{X}$ has to be transformed to have columns of unit length. **Values of CN between 20 and 30 indicate moderate collinearity, whereas values higher than 30 indicate serious collinearity, Belsley (1982).**

The following subsections applies these measures to a SLR model.

2.1 VIF in a SLR

The auxiliary regression (4) of an SLR given by:

$$\mathbf{Y} = \beta_1 + \beta_2 \mathbf{X}_2 + \mathbf{u}, \tag{6}$$

is defined as follows:

$$\mathbf{X}_2 = \alpha + \mathbf{w}, \tag{7}$$

where the OLS estimator is given by:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{1}^t\mathbf{1})^{-1}\mathbf{1}^t\mathbf{X}_2 = \frac{1}{n} \cdot \sum_{t=1}^{n} X_{2t} = \overline{\mathbf{X}}_2, \tag{8}$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones. In this case, $\mathbf{e}_t = \mathbf{X}_{2t} - \overline{\mathbf{X}}_2$ for $t = 1, \ldots, n$. Then:

$$SSR_{aux} = \mathbf{e}^t\mathbf{e} = \sum_{t=1}^{n} \left(X_{2t} - \overline{\mathbf{X}}_2\right)^2 = n \cdot Var(\mathbf{X}_2),$$

$$SST_{aux} = \sum_{t=1}^{n} \left(X_{2t} - \overline{\mathbf{X}}_2\right)^2 = SSR_{aux},$$

where $Var(\mathbf{X}_2)$ is the sample variance of $\mathbf{X}_2$ and SSR_{aux} and SST_{aux} are, respectively, the sum of squares of the residuals and total of the regression (7).

Thus, it is evident that the coefficient of determination of the auxiliary regression will always be equal to zero:

$$R^2_{aux} = 1 - \frac{SSR_{aux}}{SST_{aux}} = 1 - \frac{n \cdot Var(\mathbf{X}_2)}{n \cdot Var(\mathbf{X}_2)} = 0. \tag{9}$$

Consequently, it makes no sense to use the VIF to diagnose the possible existence of collinearity in an SLR.

Example 1 For a SLR given by (6) and supposing that $\beta_1 = 1$ and $\beta_2 = 0.5$, a dependent variable $\mathbf{Y}$ is calculated considering $\mathbf{u} \sim N(0,1)$ and $\mathbf{X}_2 = (3.1, 2.9, 3, 3.1)^t$:

$$1 \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + 0.5 \cdot \begin{pmatrix} 3.10 \\ 2.90 \\ 3 \\ 3.10 \end{pmatrix} + \begin{pmatrix} -1.53 \\ -0.64 \\ 1.66 \\ 0.08 \end{pmatrix} = \begin{pmatrix} 1.02 \\ 1.81 \\ 4.16 \\ 2.47 \end{pmatrix}.$$

The OLS estimation of SLR model is $\hat{\boldsymbol{\beta}} = (9.97, -2.51)^t$. If we perturb $\mathbf{X}_2$ to $\mathbf{X}_2^* = (3.1, 2.9, 3.12, 3.1)^t$ is obtained $\hat{\boldsymbol{\beta}}^* = (-11.47, 4.53)^t$.

It is observed significant differences between the OLS estimations that also differ substantially from the **true parameters**. This is a symptom of serious collinearity: large variation in the estimations for slight changes in the sample. Calculating the VIF using various software packages, it is concluded that GRETL does not allow us to calculate the VIF in a linear regression model. The same occurs in R when trying to obtain the VIF in an SLR: a) using the library "car" generates the message: "Error in vif.default(lm(y x2)): model contains fewer than 2 terms," and b) incoherent results are obtained when using the library "fmsb" since the VIF will always be equal to one in an SLR and, however, this library provides results of:

$$VIF(1) = 1.375, \quad VIF(2) = 1.000372,$$

for the two data sets, respectively. We are not aware of other libraries that calculate the VIF in R. Other software such as Stata and SPSS provide an option to calculate the VIF, always obtaining a value of one. Thus, the duality mentioned in the introduction when considering the **constant term** as an explanatory variable is also reflected in popular statistical software packages applied to analyze linear regression models.

In addition, given the results obtained from Stata and SPSS, it is possible to conclude that there is no collinearity, because the VIF is equal to 1. Thus, although such software enables the diagnosis of collinearity in the SLR, they can lead to an erroneous conclusion when the VIF is applied, to say that there is not a degree of worrying multicollinearity when it does exist. Thus, for example, the Condition Numbers for $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_2]$ and $\mathbf{X}^* = [\mathbf{1}\ \mathbf{X}_2^*]$ are, respectively, 72.5407 and 68.1926. Then, using this measure (that will be analyzed in the following subsection), we conclude that the model presents collinearity because the values are higher than 30.

Finally, note that the variance of the variables $\mathbf{X}_2$ and $\mathbf{X}_2^*$ is 0.006875 and 0.008075, respectively. This supports the idea that a small variance in the explanatory variable can lead to collinearity in the SLR. ♢

2.2 CN in a SLR

The CN for model (6) is calculated as follows. First, the matrix $\mathbf{X}$ must be transformed to obtain a matrix, $\widetilde{\mathbf{X}}$, with columns of unit length:

$$\mathbf{X} = \begin{pmatrix} 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{2n} \end{pmatrix}, \quad \widetilde{\mathbf{X}} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{X_{21}}{\sqrt{\sum_{t=1}^{n} X_{2t}^2}} \\ \vdots & \vdots \\ \frac{1}{\sqrt{n}} & \frac{X_{2n}}{\sqrt{\sum_{t=1}^{n} X_{2t}^2}} \end{pmatrix}.$$

Then, the eigenvalues of the following matrix must be calculated:

$$\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}} = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix},$$

where:

$$a = \frac{\sum_{t=1}^{n} X_{2t}}{\sqrt{n\sum_{t=1}^{n} X_{2t}^2}} = \frac{\overline{\mathbf{X}}_2}{\sqrt{Var(\mathbf{X}_2) + \overline{\mathbf{X}}_2^2}}. \tag{10}$$

Note that $|a| \leq 1$ because if $|a| > 1$, we have that $Var(\mathbf{X}_2) < 0$, which is not possible. In addition, $a = \pm 1$ if $Var(\mathbf{X}_2) = 0$ (in this case, $\mathbf{X}_2$ **is an** n **dimensional constant vector of** $\bar{\mathbf{X}}_2 \in R$**)**. This expression can be rewritten as:

$$a = \frac{|\mathbf{1}^t \cdot \mathbf{X}_2|}{\sqrt{n} \cdot \| \mathbf{X}_2 \|} = \sqrt{\left(\frac{\mathbf{X}_2}{\| \mathbf{X}_2 \|}\right)^t \left(\frac{\mathbf{1} \cdot \mathbf{1}^t}{n}\right) \cdot \left(\frac{\mathbf{X}_2}{\| \mathbf{X}_2 \|}\right)},$$

which is just the square root of the orthogonal projection of the unit vector $\frac{\mathbf{X}_2}{\|\mathbf{X}_2\|}$ onto the line spanned by the column $\mathbf{1}$, that is, the natural way of measuring closeness of $\mathbf{X}_2$ to $\mathbf{1}$. However, the section continues with expression (10) in order to exploit its relation with the condition number (see section 3).

Thus, due to the eigenvalues are $1 \pm a$, if it is considered that $0 \leq a \leq 1$, consequently $1 + a$ is the maximum eigenvalue, and $1 - a$ is the minimum eigenvalue. When $a < 0$, it is evident that the maximum eigenvalue is $1 - a$ and the minimum eigenvalue is $1 + a$. Thus, by considering $b = -a$ the results obtained for a follow automatically. Note that the sign of a is given by $\sum_{t=1}^{n} X_{2t}$ or, equivalently, by $\overline{\mathbf{X}}_2$.

Then, the CN is given by:

$$CN = \sqrt{\frac{1+a}{1-a}}, \quad a > 0. \tag{11}$$

From this expression, it is possible to conclude that the CN is an increasing function in a, because $\frac{1+a}{1-a}$ is an increasing function in a: the derivative is always positive $\left(\frac{2}{(1-a)^2} > 0\right)$ and the square root is a monotonic increasing function. Additionally, it is verified that $\lim_{a\to 0} CN = 1$ and $\lim_{a\to 1} CN = +\infty$. Figure 1 shows expression (11) for values of a in the interval $[0,\ 0.999)$.

On the other hand, from expression (10), is possible to conclude that a can be close to 1 (and, consequently, the CN will be very large) if $Var(\overline{\mathbf{X}}_2)$ is close to zero or a can be close to 0 (and, consequently, the CN will be close to 1) if:

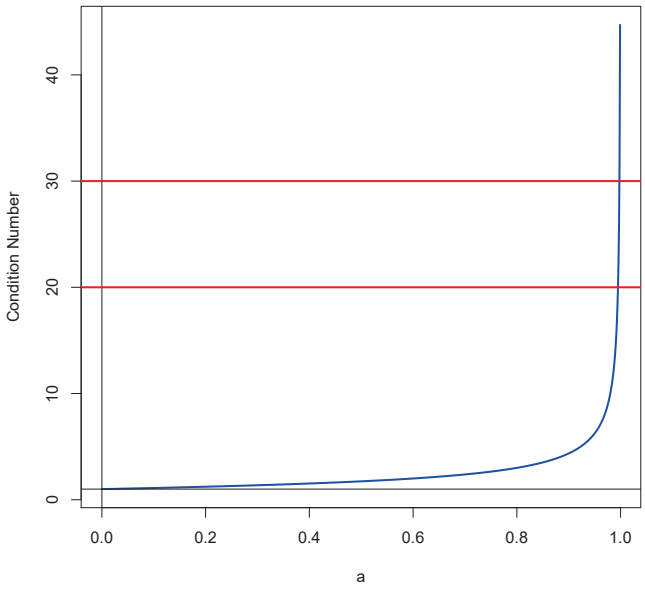

Fig. 1 Representation of the CN in a SLR as a function of a

– The variance of the variable $\mathbf{X}_2$ is sufficiently large in relation to its mean. Thus, a large variance is associated with an absence of collinearity.
– The mean of the variable $\mathbf{X}_2$ is close to zero. In this case, the explanatory variable is orthogonal to the **constant term**, because it is verified that:

$$0 = \mathbf{1}^t \mathbf{X}_2 = \sum_{t=1}^{n} X_{2t}.$$

Therefore, $a = 0$ and $CN = 1$ (from this result is evident that the solution is to center the problematic variable). Thus, it is possible to obtain values of a close to zero, even with a small variance (for further detail, see Example 2).
– The size of the sample, n, is large. Here, the CN decreases as n increases. Salmerón and Blanco (2016) analyze the relation between collinearity and a reduction in the sample size.

These questions are displayed in Figure 2 where the CN is represented for different values of the mean and variance of $\mathbf{X}_2$ following expressions (10) and (11). For large values of the variance, the CN tends to its minimum value, while for small values of the variance the CN decreases as the mean tends to zero.

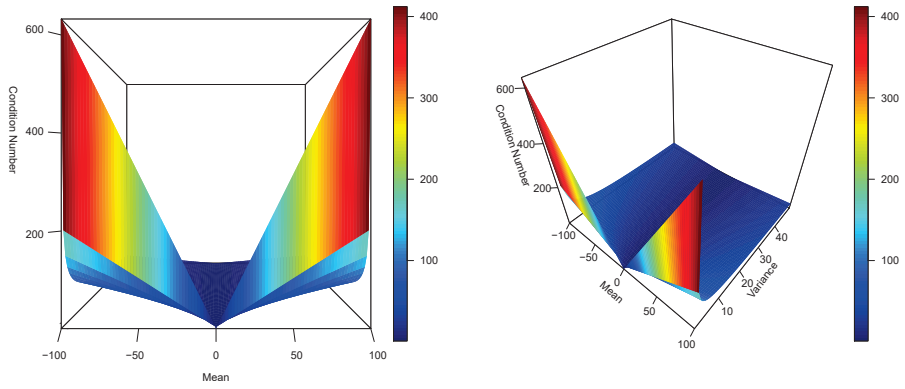

Fig. 2 Representation of the CN in a SLR as a function of the mean and the variance of $\mathbf{X}_2$ considering that $\overline{\mathbf{X}}_2 \in [-20, 20]$ and $Var(\mathbf{X}_2) \in [0.1, 10]$

Example 2 Suppose the following two data sets for the model (6):

$$\mathbf{X}_2 = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \\ -0.2 \end{pmatrix}, \quad \mathbf{X}_2^* = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.3 \\ -0.2 \end{pmatrix}.$$

Here, the variances are small (0.016875 and 0.031875, respectively), but the values of the CN are close to one (1.211 and 1.505, respectively). Therefore, these values of the CN are not coherent?. In this case, the means are so close to zero (0.025 and -0.075, respectively) that the CN is showing the quasi-orthogonality of $\mathbf{X}_2$ and $\mathbf{X}_2^*$ with the **constant term**. ♢

3 HOW SMALL SHOULD THE VARIANCE BE FOR MULTICOLLINEARITY TO EXIST?

Note in Example 2 that a small variance alone is not necessarily indicative of serious multicollinearity. Novales (1993) warned that *this result is generally valid*.

Thus, to determine how small the variance of the explanatory variable $\mathbf{X}_2$ must be for serious multicollinearity to exist in model (6), we need to control the possible orthogonality between this variable and the **constant term**. Since quasi-orthogonality is captured by the mean of $\mathbf{X}_2$, we require both descriptive statistics. As previously commented, this idea was presented by Gunst (1984) who stated *That the "nonconstant" predictor variables are essentially constant is apparent from their coefficients of variation [...] a coefficient of variation this small calls for immediate investigation of collinearity [...]*. However, the author did not provide any information about how small has to be the coefficient of variation.

From expression (11), if the CN is higher than h then it is verified that:

$$a > \frac{h^2-1}{h^2+1},$$

while that:

$$a > k \Leftrightarrow Var(\mathbf{X}_2) < \frac{1-k^2}{k^2} \cdot \overline{\mathbf{X}}_2^2.$$

Then, if $k = \frac{h^2-1}{h^2+1}$:

$$CN > h \Leftrightarrow Var(\mathbf{X}_2) < \frac{4h^2}{(h^2-1)^2} \cdot \overline{\mathbf{X}}_2^2.$$

In this case, taking into account the thresholds usually applied for the CN, it is possible to conclude for $h = 20$ there is moderated collinearity in SLR if:

$$Var(\mathbf{X}_2) < 0.01005019 \cdot \overline{\mathbf{X}}_2^2 \Leftrightarrow \frac{Var(\mathbf{X}_2)}{\overline{\mathbf{X}}_2^2} < 0.01005019. \tag{12}$$

and for $h = 30$ there is high collinearity in SLR if:

$$Var(\mathbf{X}_2) < 0.004454337 \cdot \overline{\mathbf{X}}_2^2 \Leftrightarrow \frac{Var(\mathbf{X}_2)}{\overline{\mathbf{X}}_2^2} < 0.004454337. \tag{13}$$

Thus, by taking into account the possible orthogonality between the explanatory variable and the **constant term**, we have determined how small the variance must be in model (6) for serious multicollinearity to exist. Indeed, taking into account that $\frac{Var(\mathbf{X}_2)}{\overline{\mathbf{X}}_2^2}$ is the square of the coefficient of variation of $\mathbf{X}_2$, $CV(\mathbf{X}_2)$, it is obtained that the rules given by (12) and (13) are, respectively, equivalents to:

$$CV(\mathbf{X}_2) < 0.1002506, \quad CV(\mathbf{X}_2) < 0.06674082. \tag{14}$$

The advantage is that these measures are calculated practically in all the statistical and econometric software available to the researcher.

At the same time, this result complements the suggestion made by Gunst (1984) since he did not provide a bound to determine how small has to be the coefficient of variation of a variable to avoid the existence of worrying collinearity. Indeed, this contribution fills a gap in the scientific literature since it provides an answer to the comment recently presented by Velilla (2018): *Although the collinearity in this problem may be perhaps explained by the contribution of the intercept, it is not clear how to determine exactly the reasons why.*

The implications for the GLR are immediate. It is clear that the VIF is unable to detect non-essential multicollinearity while the CN is able to do so (example 1). The same occurs in the case of the GLR (see the following Remark 1). However, value of the CN above the established thresholds do not allow establish if the degree of worrying collinearity is essential or non-essential. This is essential, as the treatment of both types of collinearity is different. However, applying the rules obtained in this section using the mean and the variance of each variable will allow distinguishing the type of collinearity and the causing variables. Thus, it would reveal the variables that need to be centered to solve the detected problem.

Remark 1 For $n \in \{15, 20, 25, \ldots, 200\}$, $\mu_1, \mu_2 \in \{-15, -14.25, \ldots, 14.25, 15\}$ and $\sigma^2 \in \{0.1, 0.2, \ldots, 10\}$, we have the matrix $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_2\ \mathbf{X}_3]$, where $\mathbf{X}_i \sim N(\mu_i, \sigma^2)$ for $i = 2, 3$, and $\mathbf{1}$ is a vector of ones with adequate dimensions.

Of 6387800 simulations, we introduce only those where $\sigma^2 < 0.01 \cdot \mu_1^2$ or $\sigma^2 < 0.01 \cdot \mu_2^2$ (i.e., 727776 (11.39%)). For these cases, the VIF and CN obtained presents the following characteristics:

	Minimum	Mean	Maximum
CN	12.33433	34.18785	209.961
VIF	1	1.016467	5.103372

Here, the VIF is unable to detect the degree of existing multicollinearity. However, the CN does do so because it takes into account the role of the **constant term**. Therefore, depending on the variance and mean of the explanatory variables, the previous decision rules can be used successfully, as in the case of the multiple linear regression ($p \geq 2$).

On the other hand, Berk (1977) shows that the CN is an upper bound of the maximum VIF. That is, the CN captures certain information about multicollinearity that the VIF does not. Our results support this observation. More specifically, the CN captures the relationship with the intercept while the VIF ignores it completely. □

Finally, although the goal of this paper is the detection and quantification of non-essential multicollinearity, it could be interesting to analyze the treatment of the centering variable and how can it be interpreted in relation to the coefficient of variation.

Thus, it is worth to remembering the following: *The decision to center or not to center the data in linear least squares depends solely on the substantive meaning of the data. We can give hundreds of examples of data for which centering (or subtracting some meaningful nonzero value) is the only sensible decision to make. We can give just as many examples for which it is sensible not to center*, Wood (1984).

In addition, given a variable $\mathbf{Z}$ with coefficient of variation $CV(\mathbf{Z}) = \frac{\sqrt{Var(\mathbf{Z})}}{|\overline{\mathbf{Z}}|}$, its transformation $\mathbf{z} = \frac{\mathbf{Z}-a}{b}$, being $a \in \mathbb{R}$, $b > 0$, presents the following coefficient of variation $CV(\mathbf{z}) = \frac{\sqrt{Var(\mathbf{Z})}}{|\overline{\mathbf{Z}}-a|}$. Consequently, although a change of scale affects the magnitude of the estimates, it can not mitigate the non-essential collinearity due to the fact that the coefficient of variation is invariant. However, a change of origin does not modify the estimates although it does mitigate the non-essential collinearity.

Within the change of origin, the transformation $a = \overline{\mathbf{Z}}$ is the most efficient in statistical terms since it leads to $CV(\mathbf{z}) \to +\infty$. At the same time, other changes of origin could exist that are less efficient than centering but that allow the increase of the coefficient of variance above the thresholds given by the expressions (14), mitigating the non-essential collinearity and with a better interpretation. These results support the comment by Belsley (1984) about structural interpretability: *Even variates like price, weight, profits or acceleration, which seem to have natural origins of zero, might be structuraly interpretable with respect to different origins in certain situations. The Dow-Jones average, for example, for a long time had a psychological plateau of 800 (nowadays its is 1200). If one were modeling such aspescts of stock market behavior it is quite possible that large and small relative changes should be assessed with respect to the deviation of the Dow-Jones from this plateau, and not from zero.*

4 QUANTIFICATION OF NON-ESSENTIAL COLLINEARITY: Stewart indices

Appendix A presents the index of Stewart for any matrix and shows its application to measure the linear relation between the columns of that matrix. Next, the measure will be contextualized in the multiple and simple linear regressions.

4.1 Stewart indices in a multiple linear regression

Applying this index in a model similar to (1) where $\mathbf{X}_1 = \mathbf{1}$, expression (21) is given by:

$$k_i^2 = \frac{\mathbf{X}_i^t \mathbf{X}_i}{\mathbf{X}_i^t \mathbf{X}_i - \mathbf{X}_i^t \mathbf{X}_{-i} \cdot \left(\mathbf{X}_{-i}^t \mathbf{X}_{-i}\right)^{-1} \cdot \mathbf{X}_{-i}^t \mathbf{X}_i} = \begin{cases} \frac{1}{1 - \frac{1}{n} \overline{\mathbf{X}}_{-i} \cdot \left(\mathbf{X}_{-i}^t \mathbf{X}_{-i}\right)^{-1} \cdot \overline{\mathbf{X}}_{-i}^t} & , i = 1 \\ \frac{\mathbf{X}_i^t \mathbf{X}_i}{SSR_i} & , i = 2, \ldots, p, \end{cases}, \tag{15}$$

where $\overline{\mathbf{X}}_{-i}$ is a file vector composed of the sum of the elements of variables $\mathbf{X}_{-i}$ and SSR_i is the sum of squared residuals of auxiliary regression (4).

When $i = 1$, the orthogonality between the constant term and the rest of independent variables are measured, it is to say, the non essential collinearity. From (15) it is obtained that $k_1^2 \geq 1$ since $\left(\mathbf{X}_{-1}^t \mathbf{X}_{-1}\right)^{-1}$ is semidefined positive and $k_1^2 = 1$ if $\overline{\mathbf{X}}_{-1} = \mathbf{0}$, it is to say, if all variables of $\mathbf{X}_{-1}$ are centered. Note that in this case, non-essential collinearity is not present as $\mathbf{1}$ is orthogonal to $\mathbf{X}_{-1}$, which is captured by k_1^2 with its minimum value.

On the other hand, if $i = 2, \ldots, p$ if $\mathbf{X}_i^t \mathbf{X}_{-i} = \mathbf{0}$ it is obtained that $\overline{\mathbf{X}}_i = 0$, since $\mathbf{1}$ is present in $\mathbf{X}_i$. It is to say, in this case orthogonality implies no correlation. Also, the expression given by (15) can be expressed as:

$$k_i^2 = \frac{SST_i}{SSR_i} + n \cdot \frac{\overline{\mathbf{X}}_i^2}{SSR_i} = VIF(i) + n \cdot \frac{\overline{\mathbf{X}}_i^2}{SSR_i}, \tag{16}$$

where $VIF(i)$ is defined as (3), since $\mathbf{X}_i^t \mathbf{X}_i = n \cdot \left(Var(\mathbf{X}_i) + \overline{\mathbf{X}}_i^2\right) = SST_i + n \cdot \overline{\mathbf{X}}_i^2$ where SST_i is the total sum of squares of auxiliary regression (4). Finally, $k_i^2 \geq 1$ since $VIF(i) \geq 1$ and $n \cdot \frac{\overline{\mathbf{X}}_i^2}{SSR_i} \geq 0$.

From expression (16) it is possible to state that the indices of collinearity presented by Stewart (1987) have been traditionally wrongly identified as VIFs since in Stewart's original paper he states: *Since our collinearity indices (or rather their squares) are already present in the statistics literature as variance inflation factors [...]*. From (16) is evident that this analogy only is verified when the variable $\mathbf{X}_i$ for which it is calculated has zero mean. It is possible to find papers where the Stewart index has been mistakenly identified with the VIF, such as: Jensen and Ramírez (2013) and, more recently, Velilla (2018). This second paper is very interesting since it treats to identify also the collinearity caused by the intercept. The anomaly in the following definition of the VIF:

$$VIF(\hat{\beta}_j) = \frac{||\mathbf{x}_j||_2^2}{s_j^2} \cdot VIF(\hat{\alpha}_j),$$

is highlighted by Christensen (2018) in his comment about the paper presented by Velilla but without clarifying what is the cause. In this sense, by following the notation of Velilla (2018) where $VIF(\hat{\alpha}_j) = \frac{1}{1-R_i^2} = \frac{SCT_j}{SCR_j}$ and taking into account that $||\mathbf{x}_j||_2^2 = n \cdot \left(var(\mathbf{X}_j) + \overline{\mathbf{X}}_j^2\right)$ and $s_j^2 = n \cdot var(\mathbf{X}_j)$ it is obtained that:

$$VIF(\hat{\beta}_j) = \frac{n \cdot \left(var(\mathbf{X}_j) + \overline{\mathbf{X}}_j^2\right)}{n \cdot var(\mathbf{X}_j)} \cdot \frac{SCT_j}{SCR_j} = VIF(\hat{\alpha}_j) + \frac{n \cdot \overline{\mathbf{X}}_j^2}{SCR_j}.$$

Note that this expression coincides with expression (16) of our contribution.
As this study shows, the VIF is unable to capture the non-essential collinearity; therefore, this measure must be exclusively associated with essential collinearity. Consequently, the second term of (16) should be identified with non-essential collinearity. This is supported by the fact that this second term is equal to zero when the analyzed variable is centered, which is the solution to the non-essential collinearity. From this last association, the ratio:

$$\frac{VIF(i)}{k_i^2} = \frac{1}{1 + n \cdot \frac{\overline{\mathbf{X}}_i^2}{SSR_i}},$$

is the proportion of essential collinearity in $\mathbf{X}_i$ and the ratio:

$$\frac{n \cdot \frac{\overline{\mathbf{X}}_i^2}{SSR_i}}{k_i^2} = \frac{1}{\frac{SST_i}{n \cdot \overline{\mathbf{X}}_i^2} + 1},$$

is the proportion of non-essential collinearity in $\mathbf{X}_i$, $i = 2, \ldots, p$. The proportion of essential collinearity existing in $\mathbf{X}_i$ is worrying if $VIF(i) > 10$, while in the case of non-essential collinearity, it is necessary to follow the rules given by (12), (13) or (14).

4.2 Stewart index in a simple linear regression

The Stewart index for model (6) is calculated from:

$$\mathbf{X} = \begin{pmatrix} 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{2n} \end{pmatrix}, \quad \mathbf{X}^t\mathbf{X} = \begin{pmatrix} n & n \cdot \overline{\mathbf{X}}_2 \\ n \cdot \overline{\mathbf{X}}_2 & \sum_{t=1}^{n} X_{2t}^2 \end{pmatrix},$$

$$\left(\mathbf{X}^t\mathbf{X}\right)^{-1} = \frac{1}{n^2 \cdot Var(\mathbf{X}_2)} \cdot \begin{pmatrix} \sum_{t=1}^{n} X_{2t}^2 & -n \cdot \overline{\mathbf{X}}_2 \\ -n \cdot \overline{\mathbf{X}}_2 & n \end{pmatrix},$$

then

$$\begin{aligned} k_1^2 = k_2^2 &= \frac{n \cdot \sum_{t=1}^{n} X_{2t}^2}{n^2 \cdot Var(\mathbf{X}_2)} = \frac{n^2 \cdot \left(Var(\mathbf{X}_2) + \overline{\mathbf{X}}_2^2\right)}{n^2 \cdot Var(\mathbf{X}_2)} \\ &= 1 + \frac{\overline{\mathbf{X}}_2^2}{Var(\mathbf{X}_2)}. \end{aligned} \tag{17}$$

Note that:

- The two possible indices coincide, $k_1^2 = k_2^2 = k^2$, since both should detect the unique possible collinearity: the non-essential collinearity.
- The non-essential collinearity is diagnosed from the mean and the variance of $\mathbf{X}_2$ as it is shown in section 3. Indeed, beginning from rule given by (12) the collinearity existing in model (6) will be worrying if

$$k^2 > 1 + \frac{1}{0.01005019} = 100.5006,$$

and by following rule (13), if:

$$k^2 > 1 + \frac{1}{0.004454337} = 225.5003.$$

- Note again that ratio $\frac{\overline{\mathbf{X}}_2^2}{Var(\mathbf{X}_2)}$ coincide with the inverse of $CV(\mathbf{X}_2)^2$, where CV denotes the coefficient of variation. It is to say, the high values of $\frac{\overline{\mathbf{X}}_2^2}{Var(\mathbf{X}_2)}$ and, consequently, of k^2, are associated with a lower coefficient of variation (greater homogeneity in the values of $\mathbf{X}_2$).
- The expression given by (17) is a particular case of the one given by (16) for $p = 2$ since a) as shown the VIF is always equal to 1 in the SLR and b) in this case, see subsection 2.1, $SCR_{aux} = \sum_{t=1}^{n}(X_{2t} - \overline{\mathbf{X}}_2)^2$ and, then, $\frac{1}{n}SCR_{aux} = Var(\mathbf{X}_2)$.

Finally, taking into account that $k^2 = \frac{Var(\mathbf{X}_2)+\overline{\mathbf{X}}_2^2}{Var(\mathbf{X}_2)}$ and $a = \frac{\overline{\mathbf{X}}_2}{\sqrt{Var(\mathbf{X}_2)+\overline{\mathbf{X}}_2^2}}$ it is obtained that:

$$a = \frac{\overline{\mathbf{X}}_2}{\sqrt{k^2 \cdot Var(\mathbf{X}_2)}},$$

and then the following relation can be established:

$$k^2 = \frac{\overline{\mathbf{X}}_2^2}{Var(\mathbf{X}_2)} \cdot \left(\frac{CN^2+1}{CN^2-1}\right)^2.$$

Remark 2 For $n \in \{15, 20, 25, \ldots, 200\}$, $\mu \in \{1, 6, 11, \ldots, 41, 46\}$ and

$\sigma^2 \in \{0.0001, 0.0002, \ldots, 0.0008, 0.001, 0.002, \ldots, 0.008, 0.01, 0.02, \ldots, 0.1, 0.15, \ldots, 0.5\}$,

we have matrix $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_2]$, where $\mathbf{X}_2 \sim N(\mu, \sigma^2)$ and $\mathbf{1}$ is a vector of ones with adequate dimensions. To calculate the CN, this matrix is transformed to obtain unit length columns (see subsection 2.2).

As this calculation is repeated 10 times, 106400 simulations are obtained. For these cases, the Stewart index and CN present the following characteristics:

	Minimum	Mean	Maximum
Stewart index	1.641497	647929.2	$5.236914 \cdot 10^7$
CN	2.082144	780.32797	14473.30414

In this case, both measures are able to detect the relation with the constant term. Indeed, a quadratic relation is observed between both indices from the representation of the simulated values of the Stewart index and the CN (see Figure 3). □

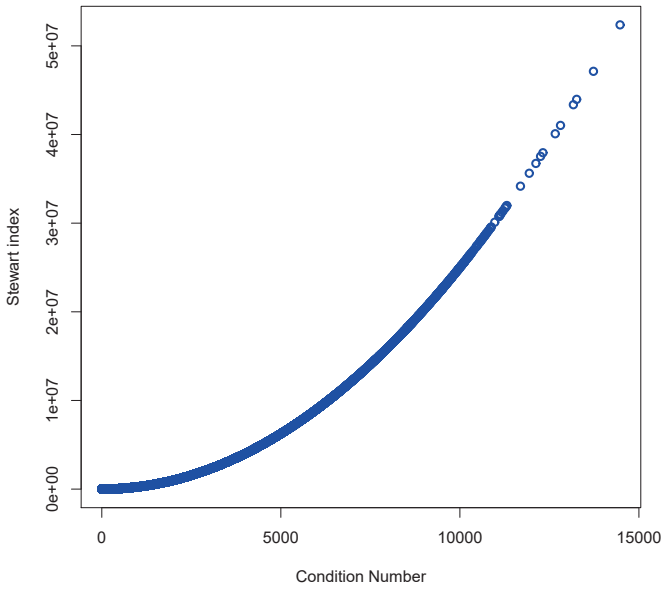

Fig. 3 Relation between the CN and the Stewart index

5 EMPIRICAL APPLICATION IN FINANCE

In order to illustrate the contribution of this study, this section presents an analysis of the following financial model of Euribor (**100%**):

$$\mathbf{Euribor} = \beta_1 + \beta_2\mathbf{HICP} + \beta_3\mathbf{BC} + \beta_4\mathbf{GD} + \mathbf{u}, \tag{18}$$

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

The data set is taken from the database of Eurostat (2018) and EMMI (2018). These temporal series are composed by 47 Eurozone observations for the period January 2002 to July 2013 (quarterly and seasonally adjusted data).

Table (1) displays the estimations, VIF and CN of model (18). Given these results it is observed that VIFs to the three explanatory variables are smaller than 10, however CN is bigger than 20 presenting a multicollinearity problem. The results of CN and VIFs are contradictory which suggests the existence of a possible non essential collinearity problem (relathionship with the intercept) since VIF like it is said before does not take under consideration the constant term.

Table 1 Estimations of model (18). Estimated standard deviations are shown in parentheses.

	Estimation	VIF	$Var(\mathbf{X}_i) < 0.01 \cdot \overline{\mathbf{X}}_i^2$
Intercept	4.376 (1.258)		
HICP	-0.002 (0.013)	1.351	$33.072 < 0.01 \cdot 54.762^2$
BC	-3.647 $\cdot 10^{-5}$ (3.897 $\cdot 10^{-6}$)	1.059	$105.219 \not< 0.01 \cdot 450388464^2$
GD	1.971 $\cdot 10^{-5}$ (2.202 $\cdot 10^{-6}$)	1.285	$4837.298 \not< 0.01 \cdot 1710699512^2$
R^2	0.820		
F_{exp}	70.68		
$\hat{\sigma}^2$	0.297		
CN	33.072		

Table 2 Stewart's index for model (18)

	k_i^2	$VIF(i)$	$n \cdot \frac{\overline{\mathbf{X}}_i^2}{SSR_i}$	% essential	% non-essential
Intercept	250.281			0 %	100 %
HIPC	280.175	1.351	278.824	0.482 %	99.518 %
BC	1.115	1.059	0.056	94.959 %	5.041 %
GD	5.529	1.285	4.244	23.234 %	76.765 %

To mitigate this problem, is not necessary to apply other alternative estimation methods, there is only to center the independent variable that presents this problem. But what independent variable is it?, for this is used the expression of variance exposed previously. In Table (1) is showed like the variable **HICP** presents a possible problem of non essential multicollinearity as the condition of variance is verified. In addition, if the index of collinearity given by Stewart (1987) (see Table 2) is calculated a worrying non-essential collinearity is detected, as k_1^2 presents a high value. Furthermore, it is possible to identify the independent variable **HICP** as the cause, and that 99.518% of the collinearity caused by this variable is non-essential.

Therefore, this variable is centered and consecutively the new model (19) is estimated:

$$\mathbf{Euribor} = \beta_1 + \beta_2 \mathbf{HICP}^* + \beta_3 \mathbf{BC} + \beta_4 \mathbf{GD} + \mathbf{u}, \tag{19}$$

where $\mathbf{HICP}^*$ is the centered Harmonized Index of Consumer Prices. In this case, we conclude that non essential collinearity has been mitigated, since the new CN is smaller than 20 (see Table 3) and VIFs remain constant (see Table 4), due to these are not affected for origin or scale changes (see García et al. (2016)). From Table 4 it is possible to conclude that for **HIPC** the non-essential collinearity has been eliminated and only essential collinearity remains.

On the other hand, about the estimation (see Table 3) is obtained that the estimator and the estimated variance of the constant term has changed. The variance has diminished considerably, mitigating one of the symptoms of collinearity. However, the estimator and estimated variance of unchanged variables remain the same in both models. The model is globally significant since the experimental statistic (F_{exp}) allows to reject the null hypothesis. Also, the coefficient of determination (R^2), the global significance (F_{exp}), and the estimated variance ($\hat{\sigma}^2$), are the same in both models.

Table 3 Estimations of model (19). Estimated standard deviations are shown in parentheses.

	Estimation
Intercept	4.158 (0.184)
HICP*	-0.002 (0.013)
BC	-3.647 $\cdot 10^{-5}$ (3.897 $\cdot 10^{-6}$)
GD	1.971 $\cdot 10^{-5}$ (2.202 $\cdot 10^{-6}$)
R^2	0.820
F_{exp}	70.68
$\hat{\sigma}^2$	0.297
CN	5.329

Table 4 Stewart's index for model (19)

	k_i^2	$VIF(i)$	$n \cdot \frac{\overline{\mathbf{X}}_i^2}{SSR_i}$	% essential	% non-essential
Intercept	5.344			0 %	100 %
HIPC	1.351	1.351	0	100 %	0 %
BC	1.115	1.059	0.056	94.959 %	5.041%
GD	5.529	1.285	4.244	23.234 %	76.765 %

Besides, in a financial prediction model, a lower variance means lower risk and a better prediction, because the standard deviation and volatility are lower. From this point of view, we require that the financial variable has the lowest possible variance. However, as discussed above, a lower variance of the independent variable may mean greater non essential multicollinearity in an GLR model. Then, the existence of worrying non-essential collinearity may be relatively common in financial econometric models.

6 CONCLUSION

The SLR given by expression (6) is relevant, despite being the simplest case of a linear regression, because it allows us to detect, for example, the impact of variations in the explanatory variable $\mathbf{X}_2$ on the dependent variable (see Novales (2010), pp. 95–96).

One of the problems that can affect the estimation of this impact is the existence of severe multicollinearity. In this case, it is well known that unstable estimators can be obtained, leading to unexpected signs. However, as discussed in the introduction, there is some ambiguity about whether this problem exists in a SLR. **For example, the software GRETL does not allow the application of common tools to determine the seriousness of existing multicollinearity in these types of models while in R is possible to do, for example, with the recently published package *multiColl*, Salmeron et al. (2019).**

Following Novales (1993), we considered that it is possible that a SLR presents serious collinearity owing to a small variance of the explanatory variable $\mathbf{X}_2$. Thus, one of the goals of this study was to determine how small this variance needs to be for collinearity becomes a serious problem. Using the CN, we obtained an expression that links the variance of $\mathbf{X}_2$ with its mean, and indicates when the collinearity in the SLR becomes problematic. This contribution

reinforces the idea given by Gunst (1984) that the coefficient of variation may serve as a measure to detect the non-essential multicollinearity.

Then, we have shown that the VIF is unable to detect the presence of collinearity in the SLR because its value is always equal to one, regardless of the data set. This is because the VIF ignores the **constant term** (this question was disregarded in papers such as Jensen and Ramírez (2013) and Velilla (2018)). Thus, specialized packages, such as Stata or SPSS, should not provide this measure for a SLR because it could lead to misleading conclusions.

Although this study focuses on SLR, the results are easily extendible to general regressions. In the case of multiple linear regressions, collinearity becomes a problem when one of the variables verifies the condition obtained from its mean and variance. In this case, the non-essential collinearity detected can be mitigated by centering or another change of origin considered more appropriate for interpretations purposes) the independent variables and eliminating the relation with the constant term. Note that this kind of collinearity will not be detected by the VIF because it ignores the **constant term** and the CN does not clarify if the detected collinearity is essential or non-essential. Consequently, it is not possible to know if the collinearity will be mitigated by centering the independent variable. However, with the rules proposed in this study, it is possible to know if the model presents non-essential collinearity, which variables provoke it, and consequently, the variables that have to be centered. This solution will mitigate the non-essential collinearity without requiring the application of other alternative methodologies such as ridge, raise, Lasso, etc.

Thus, if statistical software such as Stata, SPSS, GRETL or R incorporate this relation between the mean and the variance of each independent variable in a multiple linear regression, it will allow for the diagnosis of non-essential collinearity and indicate the variables that must be centered to mitigate it. Note that **this** relation can be expressed as a function of the coefficient of variation that is already calculated by such software, but not in the context presented in this study. In any case, note that it seems not appropriate the application of a unique measure to diagnose the existence of worrying collinearity. The use of the CV proposed in this paper can be complemented by other kinds of measures existing in the literature.

Finally, it is shown that the collinearity indices presented by Stewart (1987) can be used to determine the percentage of essential and non-essential collinearity of each independent variable, which cannot be determined by the VIF or the CN. Traditionally, these indices have been misinterpreted as VIFs.

A Stewart indices

Given matrix $\mathbf{A}$ with dimensions $n \times p$ partitioned as $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_i, \ldots, \mathbf{A}_p] = [\mathbf{A}_i, \mathbf{A}_{-i}]$ where $|\mathbf{A}|$ is the determinant of $\mathbf{A}$ and $\mathbf{A}_{-i}$ is equal to $\mathbf{A}$ after eliminating column i, Stewart (1987) defined the following index to measure the relation between $\mathbf{A}_i$ and the rest of the columns of $\mathbf{A}$:

$$k_i^2 = \frac{|\mathbf{A}_{-i}^t \mathbf{A}_{-i}| \cdot \mathbf{A}_i^t \mathbf{A}_i}{|\mathbf{A}^t \mathbf{A}|}, \quad i = 1, \ldots, p. \tag{20}$$

Since $|\mathbf{A}^t \mathbf{A}| = |\mathbf{A}_{-i}^t \mathbf{A}_{-i}| \cdot |\mathbf{A}_i^t \mathbf{A}_i - \mathbf{A}_i^t \mathbf{A}_{-i} \cdot \left(\mathbf{A}_{-i}^t \mathbf{A}_{-i}\right)^{-1} \cdot \mathbf{A}_{-i}^t \mathbf{A}_i|$, is clear that:

$$k_i^2 = \frac{\mathbf{A}_i^t \mathbf{A}_i}{\mathbf{A}_i^t \mathbf{A}_i - \mathbf{A}_i^t \mathbf{A}_{-i} \cdot \left(\mathbf{A}_{-i}^t \mathbf{A}_{-i}\right)^{-1} \cdot \mathbf{A}_{-i}^t \mathbf{A}_i}, \quad i = 1, \ldots, p. \tag{21}$$

Then, it is verified that:

$$k_i^2 = 1, \quad \text{if } \mathbf{A}_i^t \mathbf{A}_{-i} = \mathbf{0},$$
$$k_i^2 \neq 1, \quad \text{if } \mathbf{A}_i^t \mathbf{A}_{-i} \neq \mathbf{0},$$

where $\mathbf{0}$ is a vector composed of zeros with appropriate dimensions. In addition, when $i = 1, \ldots, p$, it is verified that:

- $k_i^2 > 1$ if $\mathbf{A}_{-i}^t \mathbf{A}_{-i}$ is positive defined.
- $k_i^2 < 1$ if $\mathbf{A}_{-i}^t \mathbf{A}_{-i}$ is negative defined.

Thus, this index can capture the orthogonality between $\mathbf{A}_i$ and the rest of the columns of matrix $\mathbf{A}$. However, note that orthogonality does not imply that there is no correlation:

$$\mathbf{A}_i^t \mathbf{A}_j = \sum_{k=1}^{n} A_{ik} A_{jk} = 0 \nRightarrow corr(\mathbf{A}_i, \mathbf{A}_j) = -\frac{\overline{\mathbf{A}}_i \cdot \overline{\mathbf{A}}_i}{\sqrt{\sum_{k=1}^{n} \left(A_{ik} - \overline{\mathbf{A}}_i\right)^2} \cdot \sqrt{\sum_{k=1}^{n} \left(A_{jk} - \overline{\mathbf{A}}_j\right)^2}} = 0,$$

for $i, j = 1, \ldots, p, \ i \neq j$, unless the columns have zero mean.

References

Belsley DA (1982) Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. Journal of Econometrics 20(2):211–253

Belsley DA (1984) Demeaning conditioning diagnostics through centering. The American Statistician 38(2):73–77

Berk KN (1977) Tolerance and condition in regression computations. J Amer Statist Assoc 72:863–866

Christensen R (2018) Comment on a note on collinearity diagnostics and centering. The American Statistician 72(1):114–117

Curto JD, Pinto JC (2011) The corrected vif (cvif). Journal of Applied Statistics 38(7):1499–1507

EMMI (2018) European money markets institute. URL: https://wwwemmi-benchmarkseu pp Checked: 01–02–2018

Eurostat (2018) European commission. URL: http://eceuropaeu/eurostat/web pp Checked: 01–02–2018

García J, Salmerón R, García C, López M (2016) Standardization of variables and collinearity diagnostic in ridge regression. International Statistical Review 84(2):245–266

Gujarati D (2003) Basic Econometrics (4ª edición). New York: McGraw-Hill

Gunst RF (1984) Toward a balanced assessment of collinearity diagnostics. The American Statistician 38:79–82

Jensen D, Ramírez D (2013) Revision: variance inflation in regression. Advances in decision sciences Article ID 671204:15 pages

Johnston JD, Dinardo J (2001) Métodos de econometría. Barcelona: Ed. Vicens Vives

Marquandt DW (1980) You should standardize the predictor variables in your regression models. J Amer Statist Assoc 75(369):87–91

Marquandt DW, Snee R (1975) Ridge regression in practice. Amer Statist 29(1):3–20

Novales A (1993) Econometría (2ª edición). Madrid: Ed. McGraw-Hil

Novales A (2010) Análisis de regresión. URL: https://wwwucmes/data/cont/docs/518-2013-11-13-Analisis%20de%20Regresionpdf pp Checked: 16–10–2017

Salmerón R, Blanco V (2016) El problema de un tamaño muestral pequeño en la regresión lineal: Micronumerosidad. Rect@ 17(2):167–177

Salmerón R, García J, García C, Martín ML (2017) A note about the corrected vif. Statistical Papers 58(3):929–945

Salmeron R, Garcia C, Garcia J (2019) multiColl: Collinearity Detection in a Multiple Linear Regression Model. URL https://CRAN.R-project.org/package=multiColl, r package version 1.0

Snee RD, Marquardt DW (1984) Collinearity diagnostics depend on the domain of prediction, the model, and the data. Amer Statist 38(2):83–87

Stewart G (1987) Collinearity and least squares regression. Statist Sci 2(1):68–100
Stock J, Watson M (2012) Introducción a la Econometría (3ª Edición). Madrid: Ed. Pearson
Uriel E, Periró A, Contreras D, Moltó M (1997) Econometría: El Modelo Lineal. Madrid: Ed. Alfa Centauro
Velilla S (2018) A note on collinearity diagnostics and centering. The American Statistician 72(2):140–146
Wood F (1984) Comment on effect of centering on collinearity and interpretation of the constant. The American Statistician 38(2):88–90
Wooldridge J (2009) Introductory Econometrics: A modern approach. Canada: South-Western Cengage Learning