Limitations in Detecting Multicollinearity due to Scaling Issues in the mcvis Package

by Roman Salmeron Gomez, Catalina B. Garcia Garcia, Ainara Rodriguez Sanchez, and Claudia Garcia Garcia

Abstract Transformation of the observed data is a very common practice when a troubling degree of near multicollinearity is detected in a linear regression model. However, it is important to take into account that these transformations may affect the detection of this problem, so they should not be performed systematically. In this paper we analyze the transformation of the data when applying the R package mcvis, showing that it only detects essential near multicollinearity when the *studentise* transformation is performed.

1 Introduction

Given the model $\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{u}$ for n observations and p independent variables where $\mathbf{Y}$ is a vector that contains the observations of the dependent variables, $\mathbf{X} = [\mathbf{1} \ \mathbf{X}_2 \dots \mathbf{X}_p]$ is a matrix whose columns contain the observations of the independent variables (where the first column is a vector of ones representing the intercept) and $\mathbf{u}$ represents the spherical random disturbance, the existence of linear relationships between the independent variables of a model is known as multicollinearity. It is well-known that a high degree of multicollinearity can affect the analysis of a linear regression model. In this case, it is said that the multicollinearity is troubling (Novales 1988; Ramanathan 2002; Wooldridge 2008; Gujarati 2010). It is also interesting to note the distinction made, for example, by Marquardt (1980) or Snee and Marquardt (1984), between essential (near-linear relationship between at least two independent variables excluding the intercept) and non-essential multicollinearity (near-linear relationship between the intercept and at least one of the remaining independent variables).

Note that the detection process is key to determining which tool is best suited to mitigation of the problem: for example, ridge regression of Hoerl and Kennard (1970) and Hoerl and Kennard (1970); LASSO of Tibshirani (1996) or elastic net of Zou and Hastie (2005), among others.

The most commonly applied measures to detect whether the degree of multicollinearity is troubling are the following:

- The Variation Inflation Factor (VIF) that is obtained with the following expression $VIF(j) = \frac{1}{1-R_j^2}$, $j = 2, \dots, p$, where R_j^2 is the coefficient of determination of the auxiliary regression $\mathbf{X}_j = \mathbf{X}_{-j} \cdot \boldsymbol{\alpha} + \mathbf{w}$ with $\mathbf{X}_{-j}$ being the matrix obtained when the independent variable $\mathbf{X}_j$ is eliminated from matrix $\mathbf{X}$. It is considered that values of VIF higher than 10 indicate that the degree of multicollinearity is troubling (see, for example, Marquardt (1970) or O'Brien (2007)). R. Salmerón, García, and García (2018) and Román Salmerón, Rodríguez, and García (2020) showed that the VIF is not an appropriate measure to detect linear relations between the intercept and other explanatory variables (non-essential collinearity).
- The Condition Number (CN) that is obtained with the following expression $CN(\mathbf{X}) = \sqrt{\frac{\mu_{max}}{\mu_{min}}}$ where μ_{max} and μ_{min} are the maximum and minimum eigenvalue of matrix $\mathbf{X}^t\mathbf{X}$. To obtain the eigenvalues, the matrix $\mathbf{X}$ has to be previously transformed in order to ensure that all its columns present unit length (see R. Salmerón, García, and García (2018) for more details about this transformation). Values lower than 20 imply light collinearity, between 20 and 30 moderate collinearity, while values higher than 30 imply strong collinearity (see, for example, D. A. Belsley, Kuh, and Welsch (1980) and D. Belsley (1991)). Another alternative is to calculate the CN with and without the intercept with the goal of analyzing the contribution of the intercept.

Another set of measures to detect the existence of troubling multicollinearity are the matrix of simple linear correlations between the independent variables, $\mathbf{R} = (cor(X_l, X_m))_{l,m=2,\dots,p}$ and its determinant, $|\mathbf{R}|$. C. García, Salmerón, and García (2019) show that values for the coefficient of simple correlation between the independent variables higher than $\sqrt{0.9}$ and determinant lower than $0.1013 + 0.00008626 \cdot n - 0.01384 \cdot p$ indicate a troubling degree of multicollinearity (see Román Salmerón, García, and García (2021a) or Román Salmerón, García, and García (2021b) for more details). The first value differs strongly from the threshold normally proposed equal to 0.7 to indicate a problem of near collinearity (see, for example, Halkos and Tsilika (2018)).

Also useful to use the coefficient of variation (CV), values less than 0.1002506 indicate the existence of troubling multicollinearity (see Román Salmerón, Rodríguez, and García (2020) for more details).

J. García et al. (2016) and R. Salmerón, García, and García (2020) showed that the VIF is invariant to origin and scale changes, which is the same as saying that model $\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{u}$ and the model $\mathbf{Y} = \mathbf{x} \cdot \boldsymbol{\beta} + \mathbf{u}$ present the same VIF, where $\mathbf{x} = [\mathbf{x}_1 \; \mathbf{x}_2 \ldots \mathbf{x}_p]$ with $\mathbf{x}_i = \frac{\mathbf{X}_i - a_i}{b_i}$ for $a_i \in \mathbb{R}$, $b_i > 0$ and $i = 1, \ldots, p$. Note that if $a_i = \overline{\mathbf{X}}_i$, $\mathbf{x}_1$ is a vector of zeros, i.e. the intercept disappears from the model. Instead, R. Salmerón et al. (2018) showed that the CN is not invariant to origin and scale changes, meaning that the two previous models present different CNs. This fact implies that models $\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{u}$ and $\mathbf{Y} = \mathbf{x} \cdot \boldsymbol{\beta} + \mathbf{u}$ present different eigenvalues.

Consequently, transforming the data in a linear regression model may affect the detection of the multicollinearity problem depending on the diagnostic measure. Furthermore, note that this sensitivity to scaling is due to the fact that there are certain transformations (such as data centering) that mitigate the multicollinearity problem, so that the dependence or otherwise on scaling simply highlights the capacity/incapacity of each measure to detect this reduction of the degree of near multicollinearity.

Therefore, when transforming the data in a linear regression model and analyzing whether the degree of multicollinearity is of concern or not, it is necessary to be clear whether the measure used to detect it is affected by the transformation and whether it is capable of detecting the two types of near multicollinearity mentioned (essential and non-essential). Thus, in this paper we analyze the MC index recently presented in Lin, Wang, and Mueller (2020).

First, this paper will briefly review the MC index. In order to show that the MC index depends on the transformation of the data and its inability to detect non-essential multicollinearity, we present two simulations with a troubling degree of essential and non-essential multicollinearity, respectively, and a third simulation where the degree of multicollinearity is not troubling. For all these cases, we calculate the different measures to detect multicollinearity commented on the introduction together with the MC index. Two empirical applications recently applied in the scientific literature are also presented. After a discussion of the results we propose a scatter plot between the VIF and CV to detect which variables are the cause of the troubling degree of multicollinearity and the kind of multicollinearity (essential or non-essential) existing in the model. Finally, the main conclusions of the paper are summarized.

2 Background: MC index

The MC index presented in Lin, Wang, and Mueller (2020) is based on the existing relation between the VIFs ant the inverse of the eigenvalues of the matrix $\mathbf{Z}^t\mathbf{Z}$ where $\mathbf{Z}$ represents the standardized matrix of $\mathbf{X}$. This is to say, is the matrix $\mathbf{x}$ mentioned in the introduction when for all i it is obtained that $a_i = \overline{\mathbf{X}}_i$ and $b_i = \sqrt{n \cdot var\left(\mathbf{X}_i\right)}$ where $var\left(\mathbf{X}_i\right)$ represents the variance of $\mathbf{X}_i$. More precisely, taking into account the fact that in the main diagonal of $\left(\mathbf{Z}^t\mathbf{Z}\right)^{-1}$ we find the VIFs (for standardized data), it is possible to establish the following relation:

$$\begin{pmatrix} VIF(2) \\ \vdots \\ VIF(p) \end{pmatrix} = \mathbf{A} \cdot \begin{pmatrix} \frac{1}{\mu_2} \\ \vdots \\ \frac{1}{\mu_p} \end{pmatrix}, \tag{1}$$

where $\mathbf{A}$ is a matrix that depends on the eigenvalues of $\mathbf{Z}^t\mathbf{Z}$ and μ_p is the maximum eigenvalue of this matrix.

From this relationship, resampling and obtaining the regression of $1/\mu_p$ as a function of the VIFs, Lin, Wang, and Mueller (2020) proposed the use of the t-statistics to conclude which variable contributes the most to this relationship, thus identifying which variables are responsible for the degree of approximate multicollinearity in the model. These authors defined the MC index as *an index from zero to one, larger values indicating greater contribution of the variable i in explaining the observed severity of multicollinearity*.

Taking into account that the calculation of the MC index is based on the relation established between the VIFs and the inverse of the smallest eigenvalue, it seems logical to consider that the transformation of the data may affect the calculation of this measure. Thus, it is possible to conclude:

- If the MC index is calculated with a transformation of the data that leads to the elimination of the intercept, the non-essential multicollinearity will be ignored.
- Due to the fact that MC index is based on the VIF, it should inherit its inability to detect the non-essential multicollinearity.

In conclusion, regardless of whether the data is transformed or not, the MC index is not capable of detecting non-essential multicollinearity. It is expected that it will show its usefulness in the case of essential multicollinearity.

Therefore, when Lin, Wang, and Mueller (2020) commented that *there are different views on what centering technique is most appropriate in regression [. . .] To facilitate this diversity of opinion, in the software implementation of mcvis, we allow the option of passing matrices with different centering techniques as input. The role of scaling is not the focus of our work as our framework does not rely on any specific scaling method*, far from facilitating the use of their proposal, they consider scenarios for which the MC index is not designed, since in their theoretical development and step 2 of their method, the standardization of the data is performed. However, as shown in this paper, the MC index is capable of detecting multicollinearity of the essential type only when used with its default option *studentise*.

3 Simulations

In this section, different versions of the matrix $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_2\ \mathbf{X}_3\ \mathbf{X}_4]$ will be simulated. The results on the correlation matrix, its determinant, condition number (with and without intercept), variance inflation factor and coefficient of variation are obtained using the **multiColl** package (see Román Salmerón, García, and García (2021a) and Román Salmerón, García, and García (2021b) for more details).

In all cases, are calculated the values for the MC index for each set of simulated data considering the two alternative transformation for the data: *Euclidean* (centered by mean and divided by Euclidean length) and *studentise* (centered by mean and divided by standard deviation). In each case (simulation and kind of transformation), the calculation of the MC index was performed 100 times.

Simulation 1

In this case, 100 observations are generated according to $\mathbf{X}_i \sim N(10, 100)$, $i = 2, 3$, and $\mathbf{X}_4 = \mathbf{X}_3 - \mathbf{p}$ where $\mathbf{p} \sim N(1, 0.5)$. The goal of this simulation is to ensure that the variables $\mathbf{X}_3$ and $\mathbf{X}_4$ will be highly correlated (essential multicollinearity). This fact is confirmed when taking into account the following results in relation to the correlation matrix, correlation matrix's determinant, the CN with and without the intercept (with its corresponding increasing), the VIFs and the coefficient of variation of the different variables.

```
  RdetR(X_S1)

#> $`Correlation matrix`
#>             X2_S1      X3_S1       X4_S1
#> X2_S1  1.00000000 -0.0875466 -0.09110579
#> X3_S1 -0.08754660  1.0000000  0.99881845
#> X4_S1 -0.09110579  0.9988184  1.00000000
#>
#> $`Correlation matrix's determinant`
#> [1] 0.002330192

  CNs(X_S1)

#> $`Condition Number without intercept`
#> [1] 38.67204
#>
#> $`Condition Number with intercept`
#> [1] 66.94135
#>
#> $`Increase (in percentage)`
#> [1] 42.22997

  VIF(X_S1)

#>      X2_S1      X3_S1      X4_S1
#>   1.013525 425.587184 425.860062

  CVs(X_S1)

#> [1] 1.239795 1.052896 1.166335
```

Table 1: MC index for the independent variables in Simulation 1. Three random iterations, the average value of the 100 times and the standard deviation for Euclidean and studentise transformations. The relationship between the second and third variables is detected when studentise transformation is performed.

	X2	X3	X4
Euclidean - Random 1	0.2846628	0.3670736	0.3482636
Euclidean - Random 2	0.1466484	0.4444270	0.4089246
Euclidean - Random 3	0.4026253	0.3140131	0.2833616
Euclidean - Average	0.2505292	0.3899482	0.3595226
Euclidean - Standard Deviation	0.1075164	0.0550333	0.0526817
Studentise - Random 1	0.0000338	0.4942761	0.5056901
Studentise - Random 2	0.0001294	0.4901233	0.5097473
Studentise - Random 3	0.0000307	0.4950536	0.5049157
Studentise - Average	0.0000519	0.4944290	0.5055191
Studentise - Standard Deviation	0.0000264	0.0023510	0.0023528

Table 1 shows three random iterations, the average value of the 100 times and the standard deviation. As expected in the case of essential multicollinearity, from the average values of Simulation 1 it is noted that (specially with the transformation *studentise*) the MC index correctly identified that the variables $\mathbf{X}_3$ and $\mathbf{X}_4$ are causing the troubling degree of essential multicollinearity. However, it is noted that in some cases the intercept or the variable $\mathbf{X}_2$ are identified as relevant in the existing linear relations when the *Euclidean* transformation is performed. This behavior is not observed when the *studentise* transformation is performed. This fact seems to indicate that the MC index depends on the transformation with the *studentise* transformation being the most appropriate.

Simulation 2

In this case, 100 observations are generated according to $\mathbf{X}_i \sim N(10, 100)$, $i = 2, 3$, and $\mathbf{X}_4 \sim N(10, 0.0001)$. The goal of this simulation is to ensure that the variable $\mathbf{X}_4$ will be highly correlated to the intercept (non-essential multicollinearity). This fact is confirmed from the following results taking into account that Román Salmerón, Rodríguez, and García (2020) showed that a value of the CV lower than 0.1002506 indicates a troubling degree of non-essential multicollinearity.

```
RdetR(X_S2)
```

```
#> $`Correlation matrix`
#>             X2_S2       X3_S2      X4_S2
#> X2_S2  1.0000000 -0.08754660 0.06676070
#> X3_S2 -0.0875466  1.00000000 0.09445547
#> X4_S2  0.0667607  0.09445547 1.00000000
#>
#> $`Correlation matrix's determinant`
#> [1] 0.9778526
```

```
CNs(X_S2)
```

```
#> $`Condition Number without intercept`
#> [1] 2.999836
#>
#> $`Condition Number with intercept`
#> [1] 2430.189
#>
#> $`Increase (in percentage)`
#> [1] 99.87656
```

```
VIF(X_S2)
```

```
#>    X2_S2    X3_S2    X4_S2
#> 1.013525 1.018091 1.014811
```

Table 2: MC index for the independent variables in Simulation 2. Three random iterations, the average value of the 100 times and the standard deviation for Euclidean and studentise transformations. The relationship between the four variable and the intercept is not detected.

	X2	X3	X4
Euclidean - Random 1	0.2671991	0.2097102	0.5230907
Euclidean - Random 2	0.1649092	0.5119198	0.3231711
Euclidean - Random 3	0.2126011	0.2678652	0.5195337
Euclidean - Average	0.1678676	0.3335266	0.4986058
Euclidean - Standard Deviation	0.0725673	0.0845660	0.1029678
Studentise - Random 1	0.3307490	0.3342851	0.3349658
Studentise - Random 2	0.3646487	0.2995420	0.3358093
Studentise - Random 3	0.3107573	0.3481032	0.3411395
Studentise - Average	0.3541923	0.3150319	0.3307758
Studentise - Standard Deviation	0.0201173	0.0203717	0.0187872

```
CVs(X_S2)

#> [1] 1.239794695 1.052896496 0.001022819
```

Table 2 presents three random iterations, the average value of the 100 times and the standard deviation for Simulation 2 for *Euclidean* and *studentise* transformations. Before commenting the results of Simulation 2, it is important to take into account the fact that with transformations that imply the elimination of the intercept it will not be possible to detect the non-essential multicollinearity. Note that in some occasions, when *Euclidean* transformation is performed, it is concluded that $\mathbf{X}_3$ and $\mathbf{X}_4$ are the most relevant while, when *studentise* transformation is performed, all variables seem to present the same relevance. In the first case, a higher stability is observed by considering the average values, although the conclusion is that there is a relation between $\mathbf{X}_3$ and $\mathbf{X}_4$ when the relationship is between the intercept and $\mathbf{X}_4$.

Simulation 3

Finally, in this case 100 observations are generated according to $\mathbf{X}_i \sim N(10, 100)$, $i = 2, 3, 4$. The goal of this simulation is to ensure that the degree of multicollinearity (essential and non-essential) will be not troubling. This fact is confirmed when taking into account the following results.

```
RdetR(X_S3)

#> $`Correlation matrix`
#>            X2_S3       X3_S3      X4_S3
#> X2_S3  1.0000000 -0.08754660 0.06676070
#> X3_S3 -0.0875466  1.00000000 0.09445547
#> X4_S3  0.0667607  0.09445547 1.00000000
#>
#> $`Correlation matrix's determinant`
#> [1] 0.9778526

CNs(X_S3)

#> $`Condition Number without intercept`
#> [1] 2.07584
#>
#> $`Condition Number with intercept`
#> [1] 3.60862
#>
#> $`Increase (in percentage)`
#> [1] 42.47552

VIF(X_S3)

#>    X2_S3    X3_S3    X4_S3
#> 1.013525 1.018091 1.014811
```

Table 3: MC index for the independent variables in Simulation 3. Three random iterations, the average value of the 100 times and the standard deviation for Euclidean and studentise transformations. The not troubling relationship between the variables is not detected.

	X2	X3	X4
Euclidean - Random 1	0.1318422	0.3832372	0.4849206
Euclidean - Random 2	0.1679453	0.4315253	0.4005294
Euclidean - Random 3	0.0728076	0.5693939	0.3577986
Euclidean - Average	0.1083812	0.4107279	0.4808910
Euclidean - Standard Deviation	0.0410296	0.0661189	0.0628212
Studentise - Random 1	0.3586837	0.3154482	0.3258681
Studentise - Random 2	0.3805084	0.3205751	0.2989166
Studentise - Random 3	0.3651891	0.3069920	0.3278189
Studentise - Average	0.3541923	0.3150319	0.3307758
Studentise - Standard Deviation	0.0201173	0.0203717	0.0187872

```
CVs(X_S3)
```

```
#> [1] 1.239795 1.052896 1.019006
```

Table 3 presents three random iterations, the average value of the 100 times and the standard deviation for Simulation 3 for *Euclidean* and *studentise* transformations. Simulation 3 shows different situations depending on the transformation: when the *Euclidean* transformation is performed, the variable $\mathbf{X}_3$ is also identified apart from variable $\mathbf{X}_4$; with the *studentise* transformation, all the variables seem to be relevant.

Interpretation of the obtained results

From the above results, it is concluded that the MC index applied individually is not able to detect if the degree of multicollinearity is troubling. This conclusion is in line with the comment presented by Lin, Wang, and Mueller (2020) where it is stated that *those classical collinearity measures are used together with mcvis for the better learning of how one or more variables display dominant behavior in explaining multicollinearity*. That is to say, it is recommended to use measures such as the VIF and the CN to detect whether the degree of multicollinearity is troubling and, if it is, then use the MC index to detect which variables are more relevant.

We should reiterate the fact that the results of the MC index depend on the transformation performed with the data.

Finally, it is worthy of note that the lowest dispersion is obtained when the *studentise* transformation is performed, which indicate that with this transformation a higher stability exists in the results obtained with the 100 iterations performed.

4 Examples

In this section we will analyze two examples applied recently to illustrate the multicollinearity problem. The first one focuses on the existence of non-essential approximate multicollinearity while the second one is focused on essential multicollinearity.

Example 1

Román Salmerón, Rodríguez, and García (2020) analyzed the Euribor as a function of the harmonized index of consumer prices (**HICP**), the balance of payments to net current account (**BC**) and the government deficit to net non-financial accounts (**GD**).

The following determinant of the matrix of correlations of the independent variables, the VIFs, condition number without intercept and with intercept and the coefficients of variation indicate that the degree of essential near multicollinearity is not troubling while the non-essential type (due to the variable **HIPC**) is.

Figure 1 shows a tour displayed with a scatterplot by using the **tourr** package. This package allows tours of multivariate data, see Wickham et al. (2011) for more details. From the tour on all

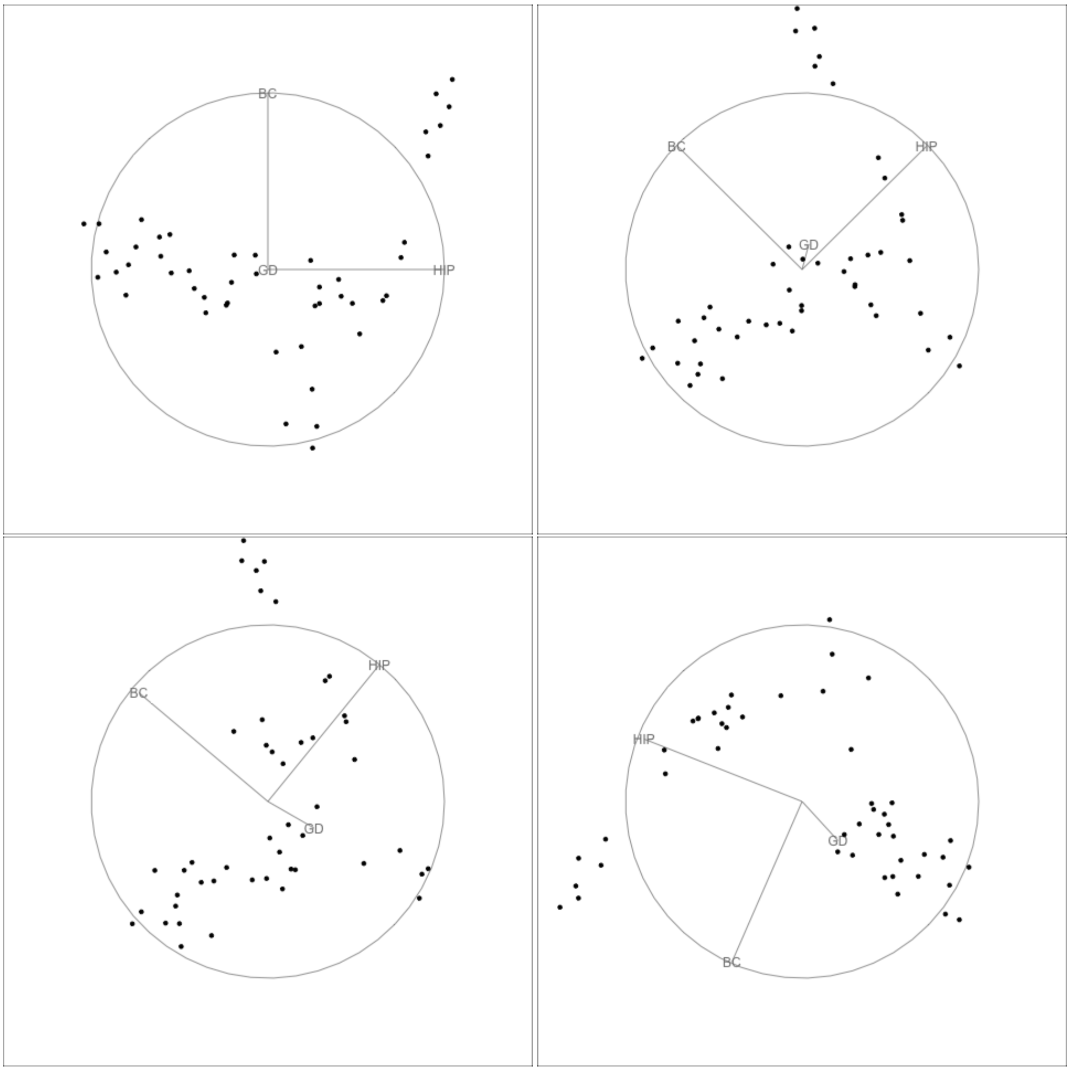

Figure 1: Representation of example 1 using the R tourr package. No linear relationship is observed between the explanatory variables in any of the four plots. See html version for an interactive 2-D tour.

the explanatory variables (it runs for 3.47 minutes in the html version), no linear relation is observed between the explanatory variables. Note that this package does not allow us to work with the intercept.

In relation to the application of the mcvis package in this example, it can be observed that if the *Euclidean* transformation is performed the method concludes that all the variables seem to be relevant:

```
#>      HIPC   BC   GD
#> tau3 0.29 0.39 0.32
```

For *studentise* transformation it concludes the establishing of a relationship between **HIPC** and **GD**:

```
#>      HIPC   BC   GD
#> tau3 0.49 0.13 0.38
```

Clearly, these conclusions are not consistent. Moreover, by eliminating the intercept when transforming the data, it is not feasible to detect the non-essential multicollinearity.

Example 2

James et al. (2013) illustrated multicollinearity with a data set which contains information about the dependent variable **balance** (average credit card debt for a number of individuals) and, among others, the following quantitative independent variables: **income**, **limit** (credit limit), **rating** (credit rating),

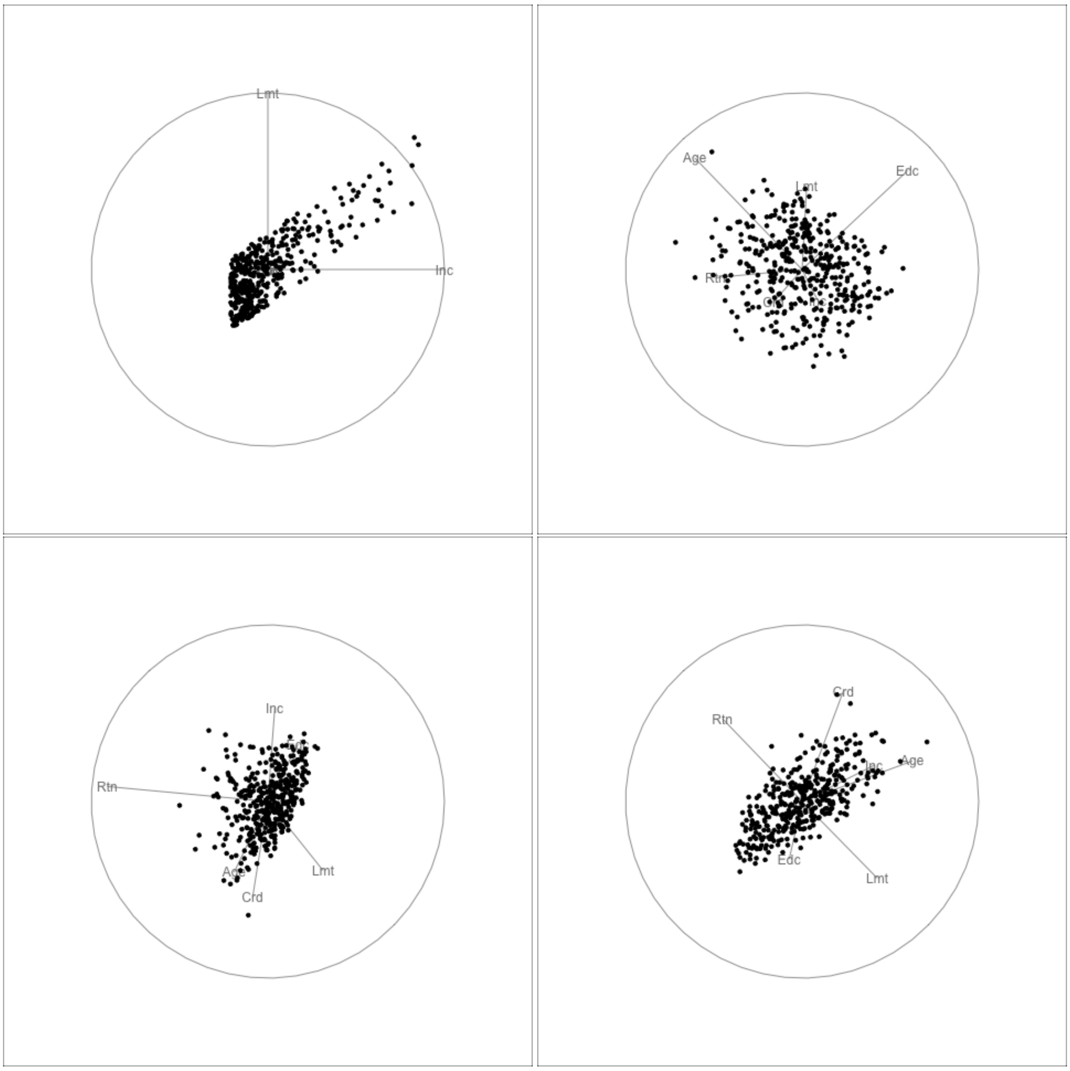

Figure 2: Representation of example 2 using the R tourr package. From the 3.47 minutes tour on all the explanatory variables, a certain linear relationship between independent variables is observed specially in plots one (top and left) and four (bottom and right). See html version for an interactive 2-D tour.

cards (number of credit cards), **age**, **education** (years of education). Data is available in the **ISLR** package (see James et al. (2021) for more details).

The following determinant of the matrix of correlations of the independent variables, the simple correlation between **limit** and **rating** (0.99687974), the VIFs, condition number without intercept and with intercept and CVs indicate that the degree of approximate multicollinearity of the non-essential type is not troubling while that of the essential type (due to the relationship between variables **limit** and **rating**) is troubling.

Figure 2 displays its tour by using again **tourr** package. Multicollinearity was checked using a tour on all the explanatory variables (it runs for 3.47 minutes in html version). In this case a certain linear relationship is observed, although it is difficult to determine which variables are related.

In relation to the application of the **mcvis** package in this example, it can be observed that if the *Euclidean* transformation is performed the variable with the highest value is **income**, with the variable **rating** being the second-lowest in value.

```
#>      Income Limit Rating Cards  Age Education
#> tau6   0.62  0.14   0.05  0.05 0.11      0.03
```

Finally, when the *studentise* transformation is applied, the method clearly indicates that variables **limit** and **rating** are the ones responsible for the multicollinearity problem.

```
#>      Income Limit Rating Cards Age Education
#> tau6      0  0.49   0.51     0   0         0
```

5 Discussion

The results shown in the previous sections indicate that, on the one hand, the calculation of the MC index depends on the transformation performed and, on the other hand, that to apply this index with guarantees, the *studentise* transformation is the most appropriate.

It was also showed that the MC index "inherits' ' the same limitations indicated by Lin, Wang, and Mueller (2020) when they state that *collinearity indices such as the variance inflation factor and the condition number have limitations and may not be effective in some applications*. In the case of VIF, these limitations are well summarized by Lin, Wang, and Mueller (2020): *The VIF can show how variables are correlated to each other, but as shown, low correlation does not guarantee low level of collinearity*. Note that the example applied by the authors in Section 1 (similar to the one presented in R. Salmerón, García, and García (2018) and a reduced version of the one presented by David A. Belsley (1984)) is a clear case in which the multicollinearity is non-essential. As was commented earlier, the VIF ((R. Salmerón, García, and García 2018) and (Román Salmerón, Rodríguez, and García 2020)) and the MC index are not able to detect this kind of multicollinearity, for this reason they are only recommended for detecting essential multicollinearity.

Another limitation of the VIF (and also of the MC index) worthy of note is that it is not adequate for calculating with dummy variables. Using these kinds of variables, the VIF is obtained from a coefficient of determination of an auxiliary regression whose dependent variable is a dummy variable. Although it is possible to apply ordinary least squares in this kind of regression, it is well known that it can present some problems: for example, the coefficient of determination is not representative since it measures the linear relation but the relation between the dummy variable and the rest of independent variables is not linear. For this reason, these kinds of regressions are estimated with non-linear models such as the logit/probit. Thus, we consider that if it is not adequate for calculating the VIF associated to a dummy variable, these kinds of variables should be avoided in the calculation of the MC index.

Finally, Simulation 3 shows that MC index is not able to detect whether the degree of essential multicollinearity is troubling. For this reason, it should only be used once this situation is determined by other measures (as set out in the introduction) and in order to determine which variables cause it.

6 Solution

The distinction between essential and non-essential multicollinearity and the limitations of each measure for detecting the different kinds of multicollinearity, can be very useful for detecting whether there is a troubling degree of multicollinearity, what kind of multicollinearity it is and which variables are causing the multicollinearity. Thus, taking into account the fact that the VIF is useful for detecting essential multicollinearity and the CV is useful for detecting non-essential multicollinearity, the scatter plot of both measures can provide interesting information in a joint way.

We present the scatter plot for the values of the VIF and the CV for the three simulations previously performed. Note that the figures include the lines corresponding to the established thresholds for each measure: red dashed vertical line for 0.1002506 (CV) and red dotted horizontal line for 10 (VIF). These lines determine four regions that can be interpreted as follows: A, existence of troubling non-essential and non-troubling essential multicollinearity; B, existence of troubling essential and non-essential multicollinearity; C, existence of non-troubling non-essential and troubling essential multicollinearity; D: non-troubling degree of existing multicollinearity (essential and non-essential).

Considering this classification, in Simulation 1 (Figure 3) it is noted that there is a troubling degree of essential multicollinearity due to the variables $\mathbf{X}_3$ and $\mathbf{X}_4$, while in Simulation 2 (Figure 4) it is noted that there is a troubling degree of non-essential multicollinearity due to the variable $\mathbf{X}_4$ and in Simulation 3 (Figure 5) it is noted that the degree of multicollinearity is not troubling.

Analogously, we present Figures 6 and 7 for Examples 1 and 2, respectively. For Example 1, it is concluded that the only type of troubling multicollinearity is non-essential due to the second variable (**HIPC**). On the other hand, for Example 2 it is concluded that the only type of troubling multicollinearity is that essentially due to the relationship existing between the third and fourth variables (**limit** and **rating**).

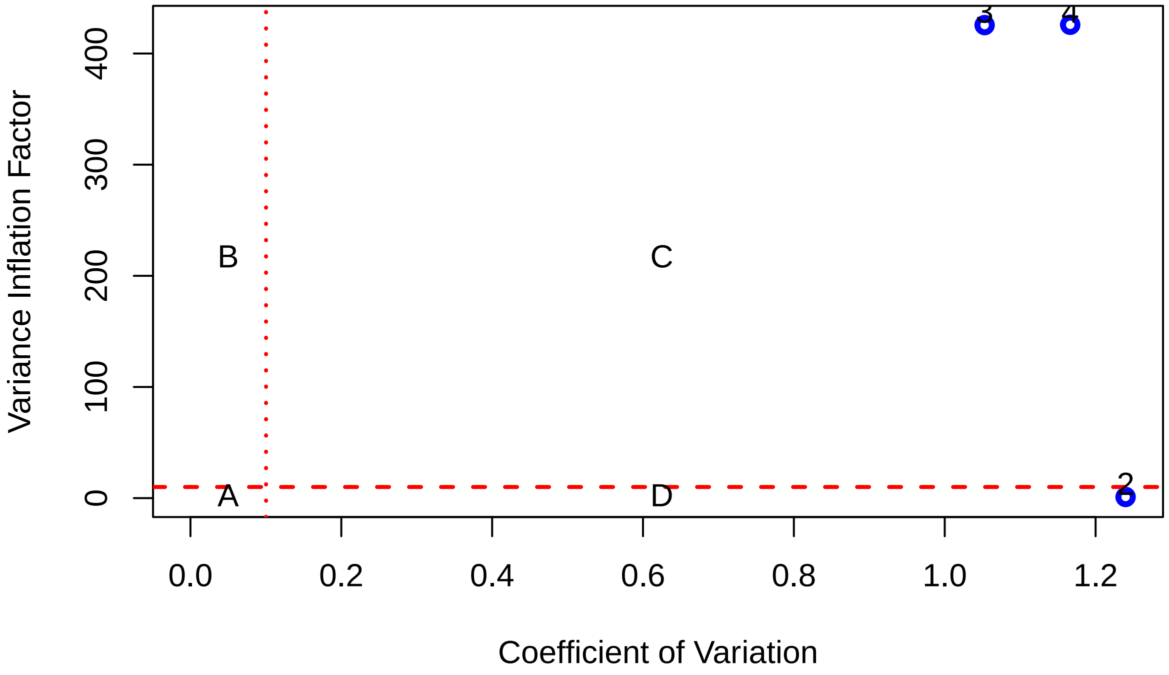

Figure 3: Representation of the VIFs and CVs of explanatory variables in Simulation 1. A troubling degree of essential multicollinearity due to variables third and four is detected by considering thresholds of 0.1002506 (CV) and 10 (VIF) highlighted with red lines. Note that VIFs of variables three and four are greater than 10 and the CV of variable 2 is lower than 0.1002506.

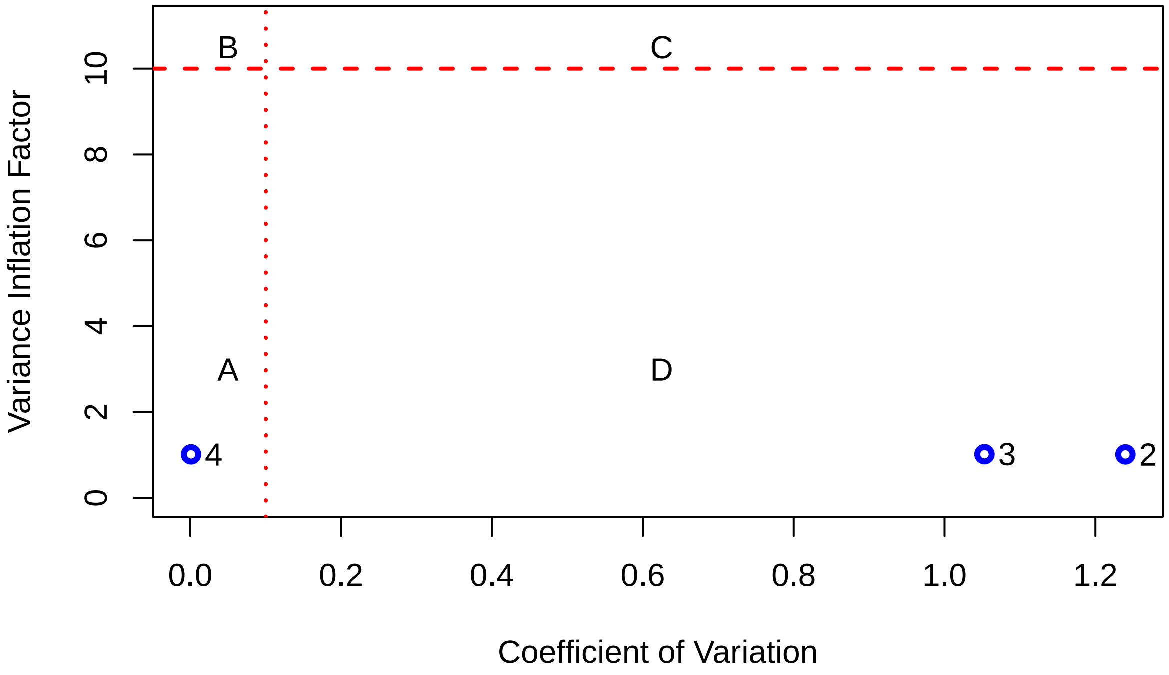

Figure 4: Representation of the VIFs and CVs of explanatory variables in Simulation 2. A troubling degree of non-essential multicollinearity due to variable four is detected by considering thresholds of 0.1002506 (CV) and 10 (VIF) highlighted with red lines. Note that the VIFs of variables three and four are less than 10 and the CV greater than 0.1002506; while the CV of second variable is less than 0.1002506.

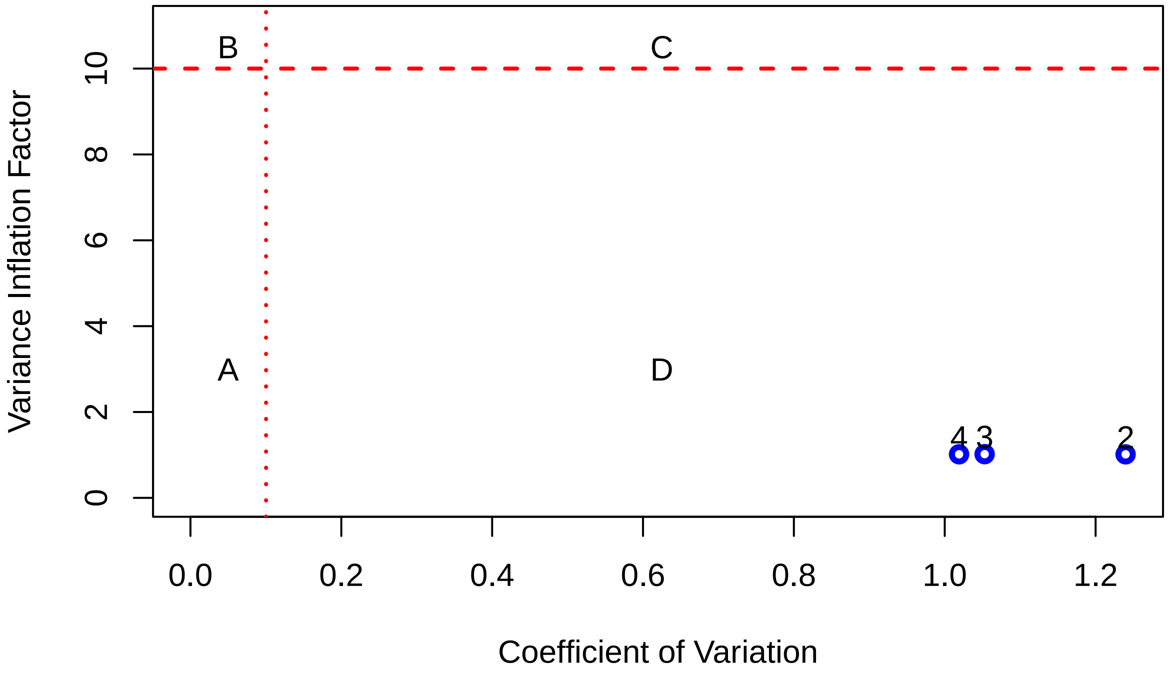

Figure 5: Representation of the VIFs and CVs of explanatory variables in Simulation 3. The degree of multicollinearity is not troubling by considering thresholds of 0.1002506 (CV) and 10 (VIF) highlighted with red lines. Note that the VIFs of all variables are lower than 10 and the CVs greater than 0.1002506.

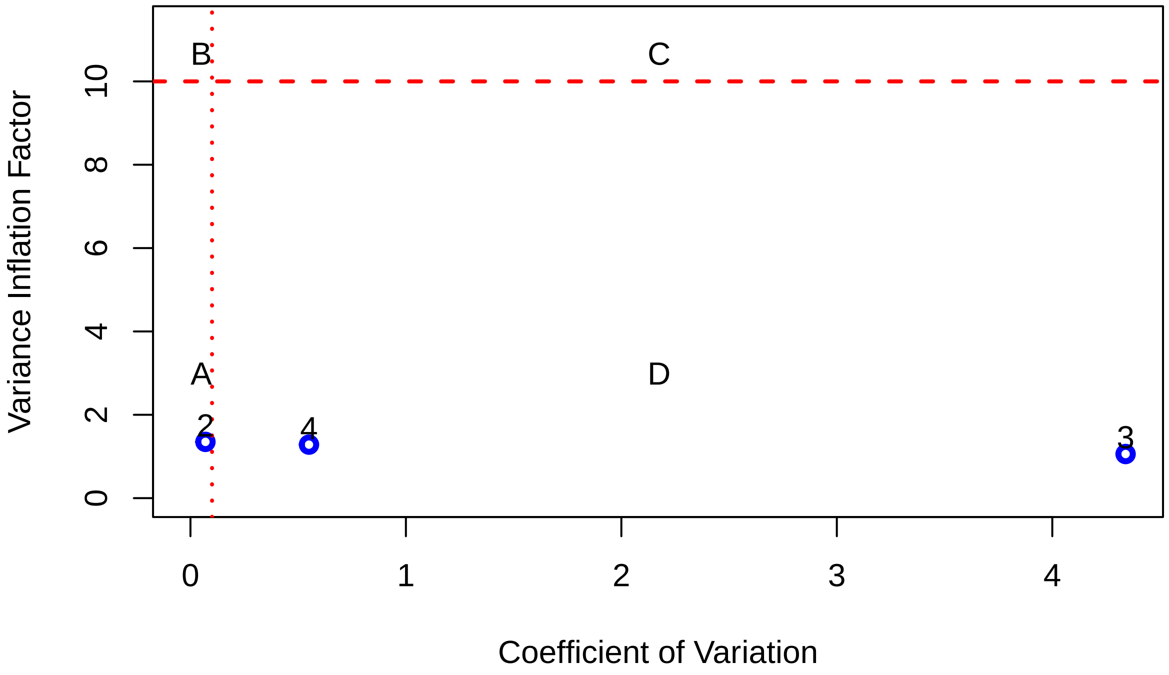

Figure 6: Representation of the VIFs and CVs of the variables of example 1. A troubling degree of non-essential multicollinearity generated by variable 2 (HIPC) is detected by considering thresholds of 0.1002506 (CV) and 10 (VIF) highlighted with red lines. Note that the VIFs of variables three and four are less than 10 and the CV greater than 0.1002506; while the CV of second variable is less than 0.1002506.

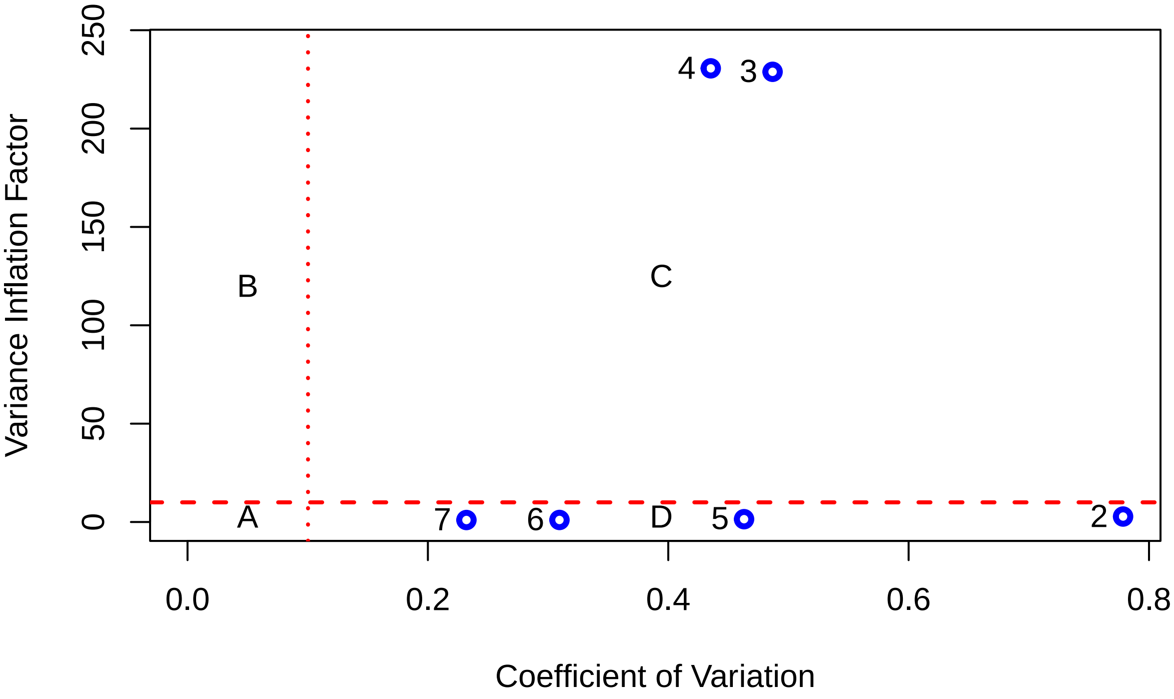

Figure 7: Representation of the VIFs and CVs of the variables of example 2. A troubling degree of essential multicollinearity generated by variables three and four (limit and rating) is detected by considering thresholds of 0.1002506 (CV) and 10 (VIF) highlighted with red lines. Note that the VIFs of variables three and four are greater than 10 while the VIFs and CVs of the rest of variables are, respectively, lower than 10 and greater than 0.1002506.

7 Summary

This paper analyses the limitations that may arise in detecting the problem of troubling multicollinearity in a linear regression model due to transformations performed on the data. The discussion is used to illustrate that the MC index presented by Lin, Wang, and Mueller (2020) depends on the way that the data are transformed.

It was shown that when the *studentise* transformation is performed, the measure is stable for *measuring what variable contributes the most to the linear relationship, and thus identifying what variables best explain the collinearity in the original data* (Lin, Wang, and Mueller 2020), as long as the multicollinearity is essential. This kind of transformation was taken as default by the authors and this paper contributes with a formal justification. Note that the MC index only provides interesting information if the troubling multicollinearity is previously detected using other measures (such as the VIF or the CN).

Summarizing, the achievement of the goal intended by the **mcvis** package is conditioned by the following limitations:

- It is not able to detect non-essential multicollinearity.
- Other measures (as the VIF or CN) should be applied previously to determine whether the degree of essential multicollinearity is troubling.
- The *studentise* transformation should be applied.
- It is not applicable for dummy variables.

Finally, it is proposed to use a scatter plot between the VIFs and the CVs, on the one hand, to detect whether the degree of multicollinearity (essential or not essential) is troubling and, on the other hand, to detect which variables are causing the multicollinearity. This would overcome the limitations of the **mcvis** package discussed above while achieving its overall purpose.

8 Acknowledgments

This work has been supported by project PP2019-EI-02 of the University of Granada (Spain), by project A-SEJ-496-UGR20 of the Andalusian Government's Counseling of Economic Transformation, Industry, Knowledge and Universities (Spain) and by project I+D+i PID2019-107767GA-I0 financed by MCIN/AEI/10.13039/501100011033.

References

Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.

Belsley, David. 1991. "A Guide to Using the Collinearity Diagnostics." *Computational Science in Economics and Manegement* 4: 33–50.

Belsley, David A. 1984. "Demeaning Conditioning Diagnostics Through Centering." *The American Statistician* 38 (2): 73–77.

García, C., R. Salmerón, and C. B. García. 2019. "Choice of the Ridge Factor from the Correlation Matrix Determinant." *Journal of Statistical Computation and Simulation* 89 (2): 211–31.

García, J., R. Salmerón, C. García, and M. López. 2016. "Standardization of Variables and Collinearity Diagnostic in Ridge Regression." *International Statistical Review* 84: 245–66.

Gujarati, D. 2010. *Basic Econometrics*. 8th ed. McGraw Hill.

Halkos, G., and K. Tsilika. 2018. "Programming Correlation Criteria with Free Cas Software." *Computational Economics* 52 (1): 299–311.

Hoerl, A. E., and R. W. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12 (1): 55–67.

James, Gareth, Daniela Witten, Trevor Hastie, and Rob Tibshirani. 2021. *ISLR: Data for an Introduction to Statistical Learning with Applications in r*. https://CRAN.R-project.org/package=ISLR.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Application in r*. Springer.

Lin, C., K. Wang, and S. Mueller. 2020. "MCVIS: A New Framework for Collinearity Discovery, Diagnostic and Visualization." *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2020.1779729.

Marquardt, D. W. 1970. "Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation." *Technometrics* 12 (3): 591–612.

———. 1980. "You Should Standardize the Predictor Variables in Your Regression Models." *J. Amer. Statist. Assoc.* 75 (369): 87–91.

Novales, A. 1988. *Econometría*. Madrid: McGraw-Hill.

O'Brien, R. M. 2007. "A caution regarding rules of thumb for variance inflation factors." *Quality & Quantity* 41: 673–90.

Ramanathan, Ramu. 2002. "Introductory Econometrics with Applications."

Salmerón, R., C. García, and J. García. 2018. "Variance Inflation Factor and Condition Number in Multiple Linear Regression." *Journal of Statistical Computation and Simulation* 88: 2365–84.

———. 2020. "Comment on 'a Note on Collinearity Diagnostics and Centering' by Velilla (2018)." *The American Statistician* 74 (1): 68–71.

Salmerón, R., J. García, C. García, and M. López. 2018. "Transformation of Variables and the Condition Number in Ridge Estimation." *Computational Statistics* 33: 1497–1524.

Salmerón, Román, Catalina García, and José García. 2021a. "A Guide to Using the r Package multiColl for Detecting Multicollinearity." *Computational Economics* 57: 529–36. https://doi.org/10.1007/s10614-019-09967-y.

———. 2021b. "The multiColl Package Versus Other Existing Packages in r to Detect Multicollinearity." *Computational Economics*. https://doi.org/10.1007/s10614-021-10154-1.

Salmerón, Román, Ainara Rodríguez, and Catalina García. 2020. "Diagnosis and Quantification of the Non-Essential Collinearity." *Computational Statistics* 35: 647–66.

Snee, R. D., and D. W. Marquardt. 1984. "Comment: Collinearity diagnostics depend on the domain of prediction, the model, and the data." *The American Statistician* 38 (2): 83–87.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–88.

Wickham, Hadley, Dianne Cook, Heike Hofmann, and Andreas Buja. 2011. "Tourr: A r Package for Exploring Multivariate Data with Projections." *Journal of Statistical Software* 40 (2): 1–18. https://doi.org/10.18637/jss.v040.i02.

Wooldridge, J. M. 2008. *Introducción a la econometría. Un enfoque moderno*. Second. Madrid: Thomson Paraninfo.

Zou, H., and T. Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67: 301–20.

Roman Salmeron Gomez
University of Granada
Department of Quantitative Methods for Economics and Busines
Poligono La Cartuja sn, 18071, Granada, Spain.
http://metodoscuantitativos.ugr.es/pages/web/romansg
ORCiD: 0000-0003-2589-4058
romansg@ugr.es

Catalina B. Garcia Garcia
University of Granada
Department of Quantitative Methods for Economics and Busines
Poligono La Cartuja sn, 18071, Granada, Spain.
http://metodoscuantitativos.ugr.es/pages/web/cbgarcia
ORCiD: 0000-0003-1622-3877
cbgarcia@ugr.es

Ainara Rodriguez Sanchez
U.N.E.D.
Department of Applied Economics and Economic History
Madrid, Spain.
http://metodoscuantitativos.ugr.es/pages/web/cbgarcia
ORCiD: 0000-0003-1622-3877
arsanchez@cee.uned.es

Claudia Garcia Garcia
Complutense University of Madrid
Department of Applied Economics, Structure and History
Madrid, Spain.
http://metodoscuantitativos.ugr.es/pages/web/cbgarcia
ORCiD: 0000-0003-1622-3877
clgarc13@ucm.es