The *RVIF* package of **R** to detect multicollinearity

El paquete *RVIF* de **R** para detectar multicolinealidad

R. Salmerón (romansg@ugr.es) **C.B. García** (cbgarcia@ugr.es)
Department of Quantitative Methods for Economics and Business
University of Granada
J. García (jgarcia@ual.es)
Department of Economics and Business, University of Almería

XXXVI International Conference ASEPELT
Inequality, poverty exclusion and policy implications for a better world

Évora, July 5-7, 2023

Índice

Introduction

Econometrics

Econometrics is a branch of Economics that provides a basis for refining or refuting theoretical knowledge and provides signs and magnitudes of the relationships of variables that are being analyzed.

With this objective, the coefficients of the general linear model have to be estimated:

$$\mathbf{Y} = \beta_1 \cdot \mathbf{X}_1 + \beta_2 \cdot \mathbf{X}_2 + \cdots + \beta_i \cdot \mathbf{X}_i + \cdots + \beta_p \cdot \mathbf{X}_p + \mathbf{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

Applying Ordinary Least Squares (OLS) we obtain the following normal equations system:

$$\left(\mathbf{X}^t\mathbf{X}\right)\beta = \mathbf{X}^t\mathbf{Y}.$$

To obtain a unique solution, there must exist $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}.$$

If the variables that form $\mathbf{X}$ are linearly dependent, then does not exist $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$ (perfect multicollinearity).

Introduction

What if they are almost linearly independent? (approximate multicollinearity):

- Small changes in the data can lead to large variations in the estimates of the parameters.
- Tendency not to reject that the coefficients of the regressors are zero due to "inflated" estimated standard deviations of the coefficients.
- High determination coefficient and, consequently, a tendency to reject that all the coefficients are zero simultaneously.
- Non-compliance with the *ceteris paribus* assumption.

Summing up, the estimation by OLS can present instability, leading to erroneous conclusions that could even call into question the validity of the analysis.

Types of Multicollinearity

Marquardt and Snee (1975) [1] distinguish two types of multicollinearity:

Nonessential multicollinearity: relationship between the constant term and one of the independent variables (it is known the solution is to center the variables).

Essential multicollinearity: relationship between at least two independent variables (without considering the constant term).

Detection of multicollinearity

Measures for the detection of approximate multicollinearity

Matrix of simple linear correlations and its determinant, coefficient of variation, the variance inflation factor, the Condition Number (CN) or the Stewart Index.

Variance inflation factor

One of the most applied measure to detect multicollinearity in a general lineal model is the variance inflation factor (VIF) given by:

$$FIV(i) = \frac{1}{1 - R_i^2}, \quad i = 2, \ldots, p,$$

where R_i^2 is the coefficient of determination of the regression of $\mathbf{X}_i$ as a function of the rest of independent variables $\mathbf{X}_{-i}$.
Values higher than 10 indicate troubling multicollinearity

Note The VIF does not analyzed the relationship between $\mathbf{X}_2 \ldots \mathbf{X}_p$, with the constant term, $\mathbf{X}_1 = \mathbf{1}$ (Salmerón, R., García, C.B. and García, J. (2018) [2]). Consequently, the VIF is not an appropriate measure to detect non-essential collinearity.

Variance inflation factor (VIF)

How does the above expression arise?

Para $i = 2, \ldots, p$:

$$\text{Initial model:} \quad \widehat{\text{var}}\left(\widehat{\beta}_i\right) = \frac{\widehat{\sigma}^2}{n \cdot var(\mathbf{X}_i)} \cdot \frac{1}{1 - R_i^2},$$

$$\text{Orthogonal model:} \quad \widehat{\text{var}}\left(\widehat{\beta}_{i,o}\right) = \frac{\widehat{\sigma}^2}{n \cdot var(\mathbf{X}_i)},$$

$$\text{Comparation between both models:} \quad \frac{\widehat{\text{var}}\left(\widehat{\beta}_i\right)}{\widehat{\text{var}}\left(\widehat{\beta}_{i,o}\right)} = \frac{1}{1 - R_i^2} = FIV(i).$$

The VIF measures how much the variance of $\widehat{\beta}_i$ increases with respect to the orthogonal case where it is assumed that $R_i^2 = 0$.

Note Is it reasonable to think that in the orthogonal case the value of R_i^2 changes but $var(\mathbf{X}_i)$ does not change? What happens with the estimate of σ?

Variance inflation factor (VIF)

If there were orthogonality...

$$\mathbf{X}^t\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^t\mathbf{X}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{X}_2^t\mathbf{X}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{X}_p^t\mathbf{X}_p \end{pmatrix},$$

then:

$$\widehat{\text{var}}\left(\widehat{\beta}_{i,o}\right) = \frac{\widehat{\sigma}^2}{\mathbf{X}_i^t\mathbf{X}_i} \neq \frac{\widehat{\sigma}^2}{n \cdot var(\mathbf{X}_i)},$$

due to $\mathbf{X}_i^t\mathbf{X}_i = n \cdot \left(var(\mathbf{X}_i) + \overline{\mathbf{X}}_i^2\right)$.

Thus, the initial approach is valid if $\overline{\mathbf{X}}_i = 0$ for $i = 2, \ldots, p$ (for $i = 1$ is impossible) and if the estimation of σ coincides in the original and orthogonal model.

An alternative orthogonal model

Performing a QR decomposition for the matrix $\mathbf{X}$ it is obtained that $\mathbf{X} = \mathbf{X}_o \cdot \mathbf{P}$ where $\mathbf{X}_o$ is an orthonormal matrix of the same dimensions as $\mathbf{X}$ and $\mathbf{P}$ is an upper triangular matrix.
Orthogonal model: $\mathbf{Y} = \mathbf{X}_o \cdot \boldsymbol{\beta}_o + \mathbf{W}$.

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

In short, $\widehat{\text{var}}\left(\widehat{\beta}_{i,o}\right) = \widehat{\sigma}^2$.

In this case, due to $\widehat{\text{var}}\left(\widehat{\beta}_i\right) = \frac{\widehat{\sigma}^2}{n \cdot var(\mathbf{X}_i)} \cdot \frac{1}{1-R_i^2}$, then:

$$\frac{\widehat{\text{var}}\left(\widehat{\beta}_i\right)}{\widehat{\text{var}}\left(\widehat{\beta}_{i,o}\right)} = \frac{1}{n \cdot var(\mathbf{X}_i)} \cdot \frac{1}{1-R_i^2} = \frac{FIV(i)}{n \cdot var(\mathbf{X}_i)}, \quad i = 1, \ldots, p.$$

Redefinition of the VIF

Taking into account that $\frac{FIV(i)}{n \cdot var(\mathbf{X}_i)} = \frac{1}{SCR_i}$ where SCR_i is the sum of squares of the residuals from the regression of $\mathbf{X}_i$ as a function of the rest of independent variables, $\mathbf{X}_{-i}$...

Redefined variance inflation factor

The variance inflation factor is redefined as:

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

For unit-length data, $\mathbf{X}_i^t \mathbf{X}_i = 1$, RVIF coincide with the Stewart index
(This transformation is performed to calculate the Condition Number)

- Since $FIVR(i) > 0$, then $a_i = \mathbf{X}_i^t \mathbf{X}_{-i} \cdot \left(\mathbf{X}_{-i}^t \mathbf{X}_{-i}\right)^{-1} \cdot \mathbf{X}_{-i}^t \mathbf{X}_i \leq \mathbf{X}_i^t \mathbf{X}_i = 1$.
- If $\mathbf{X}_i$ is orthogonal to $\mathbf{X}_{-i}$, then $a_i = 0$.
- $a_i \in [0, 1]$ is interpreted as the percentage of multicollinearity due to $\mathbf{X}_i$.
- $FIVR(i) \geq 1$ for $i = 1, \ldots, p$.

Example 1

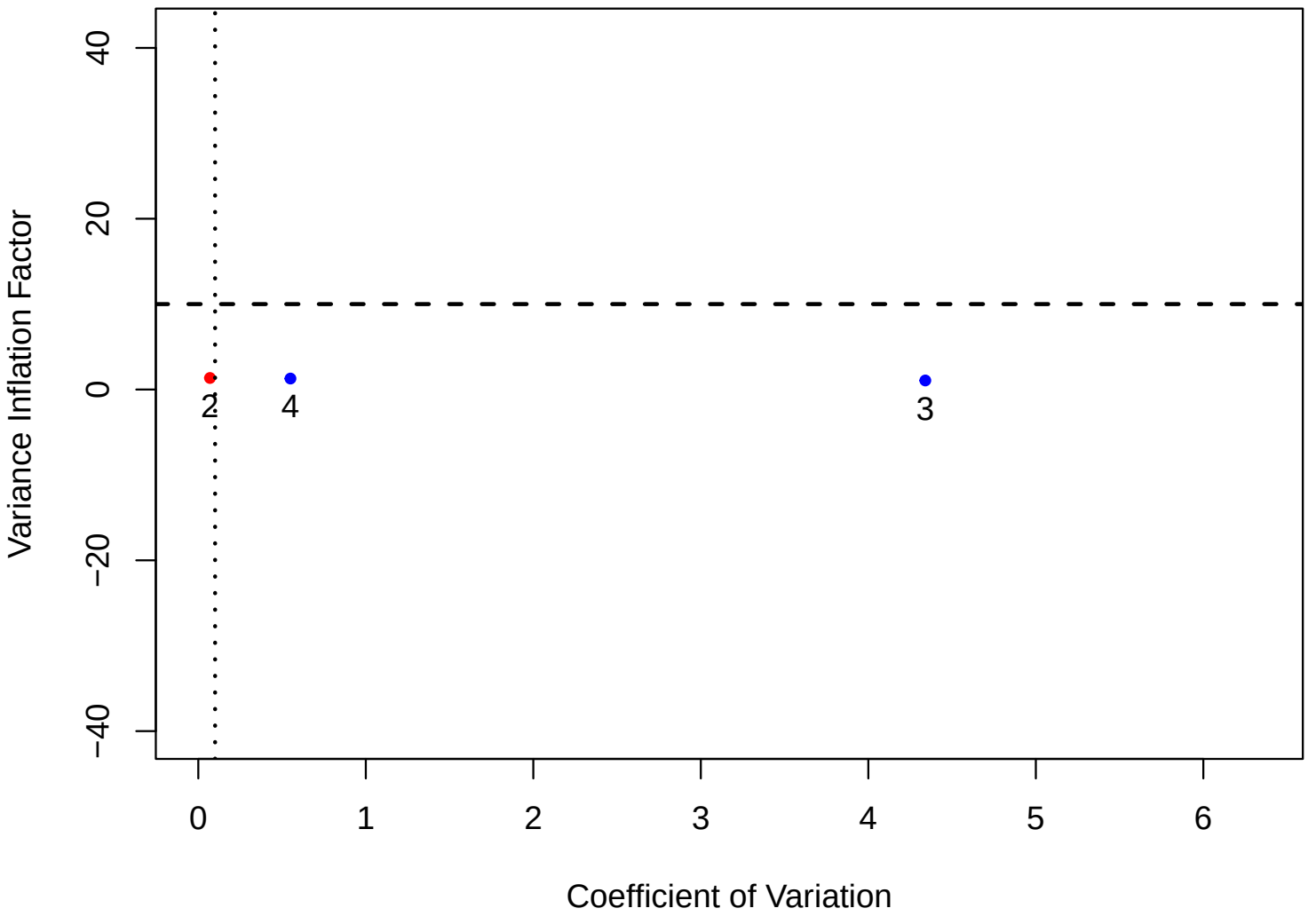
Salmerón, R., Rodríguez, A. and García, C.B. (2019) [3]

In this model the Euribor (**E**, %) is analyzed as a function of Harmonized index of consumer prices (**HICP**, %), the balance of payments to net current account (**BC**, millions of euros) and the government deficit to net nonfinancial accounts (**GD**, millions of euros).
In this model it is established that **HICP** is related to the constant due to the coefficient of variation is equal to 0.06957876.

```
E = c(3.63, 3.90, 3.45,..., 0.51, 0.54)
HIPC = c(92.92, 93.85, 93.93,..., 117.55, 117.34)
BC = c(17211, 2724, 17232,..., 56376, 48981)
GD = c(-51384.0, -49567.1, -52128.4,...,-77738.8, -73003.3)
data1 = cbind(rep(1, length(E)), HIPC, BC, GD)
```

```
>   RVIF(data1, l_u = T)
                  RVIF       %
Intercept   250.294157 99.6005
Variable 2  280.136873 99.6430
Variable 3    1.114787 10.2967
Variable 4    5.525440 81.9019
```

Example 1


Variance Inflation Factor
40
20
0
−20
−40
2
4
3
0
1
2
3
4
5
6
Coefficient of Variation

Ejemplo 2

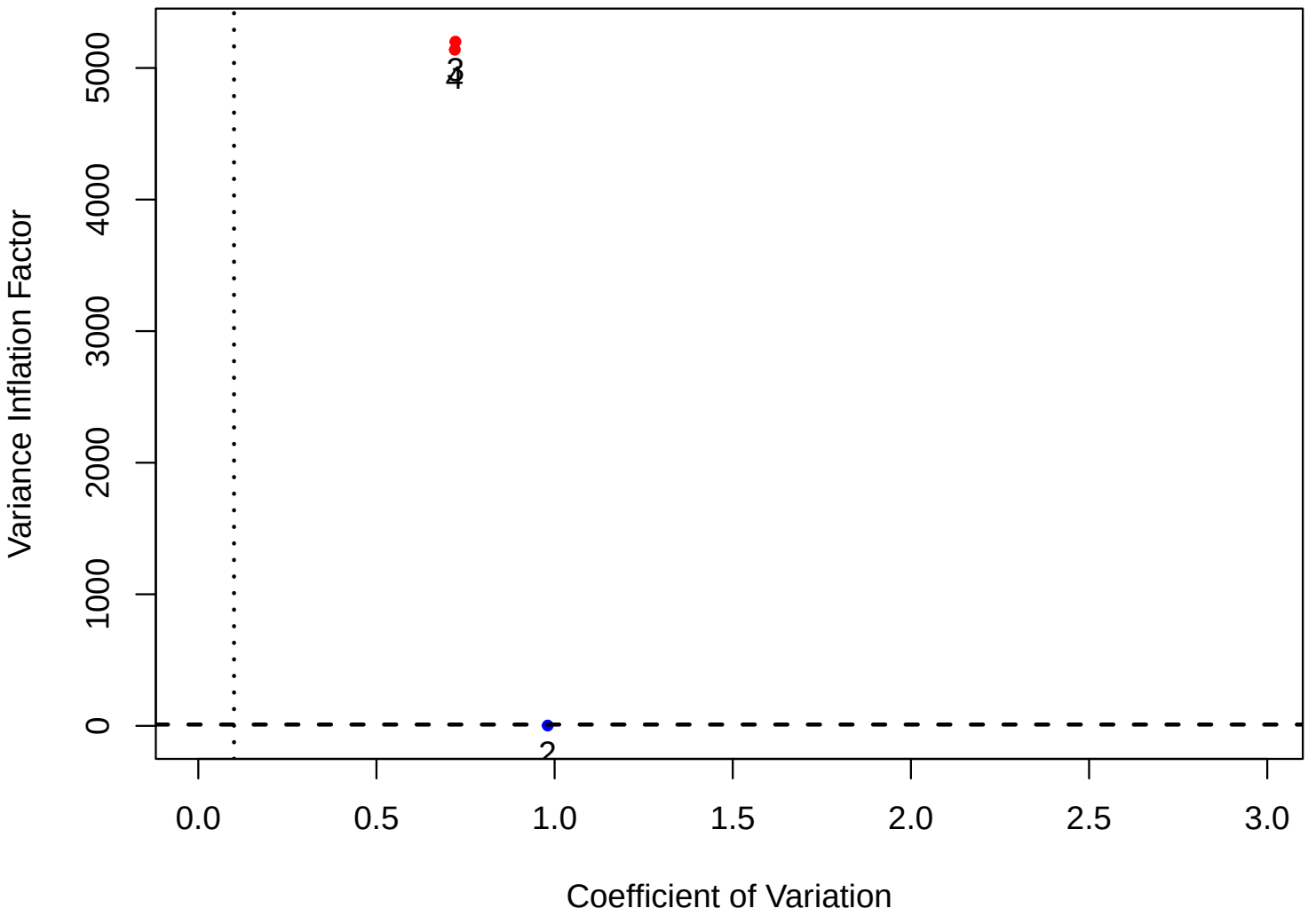
Salmerón, R., Rodríguez, A., García, C.B. and García, J. (2020) [4]

This paper analyzed the relationship between the number of employees (**NE**) and the fixed assets (**FA**, en euros), operating income (**OI**, en euros) and sales (**S**, en euros) for 15 Spanish companies for the year 2016.
In this model it is determined a high relationship between **OI** and **S**.

```
NE = c(2637, 15954, 162503,..., 15122, 13881)
FA = c(44153, 9389509, 17374000,..., 26787667, 6681800)
OI = c(38903, 4293386, 23703000,..., 4916125, 4472900)
S = c(38867, 4231043, 23649000,..., 4758244, 4472900)
data2 = cbind(rep(1, length(NE)), FA, OI, S)
```

```
>   RVIF(data2, l_u = T)
                    RVIF       %
Intercept       2.984146 66.4896
Variable 2      5.011397 80.0455
Variable 3  15186.744870 99.9934
Variable 4  15052.679178 99.9934
```

Example 2


Variance Inflation Factor
0
1000
2000
3000
4000
5000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Coefficient of Variation
3
4
2

Main results

Conclusions

- According to Johnston (1984) [5] "the orthogonal case does not mean that it is an achievable goal but it is used as a reference point to measure the relative increase in the sample variance of the estimators".
- However, although it is not an achievable goal, it must be credible.
- We propose an alternative orthogonal reference that leads to a redefinition of the Variance Inflation Factor.
- The RVIF package allows:
  a) detecting the approximate multicollinearity of essential and non-essential type,
  b) identify the variables that cause it and
  c) quantify the percentage of multicollinearity due to a particular independent variable.

Bibliography (in order of appearance)

- Marquardt, D. W. and R. Snee (1975). "Ridge regression in practice". The American Statistician 29 (1), pp. 3–20.
- Salmerón, R., García, C.B. y García, J. (2018). "Variance Inflation Factor and Condition Number in multiple linear regression". Journal of Statistical Computation and Simulation, 88 (12), pp. 2365–2384.
- Salmerón, R., Rodríguez, A. y García, C.B. (2019). "Diagnosis and quantification of the non-essential collinearity". Computational Statistics, 35, pp. 647–666.
- Salmerón, R., Rodríguez, A., García, C.B. y García, J. (2020). "The VIF and MSE in raise regression". Mathematics, 8 (4), pp. 605.
- Johnston, J. (1972). "Econometric Methods". McGraw-Hill.

Acknowledgments

This work has been funded by:

- Project PP2019-EI-02 of the University of Granada (Spain) entitled "Redefinition of the Variance Inflation Factor and its thresholds" and
- Project A-SEJ-496-UGR20 of the Andalusian Government's Counseling of Economic Transformation, Industry, Knowledge and Universities (Spain) entitled "Disruption in the multicollinearity problem: a new vision for its diagnosis and treatment".

The *RVIF* package of **R** to detect multicollinearity

El paquete *RVIF* de **R** para detectar multicolinealidad

R. Salmerón (romansg@ugr.es) **C.B. García** (cbgarcia@ugr.es)
Department of Quantitative Methods for Economics and Business
University of Granada
J. García (jgarcia@ual.es)
Department of Economics and Business, University of Almería

XXXVI International Conference ASEPELT
Inequality, poverty exclusion and policy implications for a better world

Évora, July 5-7, 2023