Detección de la multicolinealidad a partir del Factor de Inflación de la Varianza Redefinido

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

3R4EDI
Tercer Congreso R para Empresa, Docencia e Investigación

Almagro, 19-20 octubre 2023

Índice

Introducción

Econometría

Econometría es una rama de la Economía que proporciona una base para refinar o refutar conocimiento teórico y conseguir signos y magnitudes de las relaciones de variables que se desean analizar.

Con tal objetivo se han de estimar los coeficientes del modelo lineal general:

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

Aplicando Mínimos Cuadrados Ordinarios (MCO) se llega al sistema de ecuaciones normales:

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

Para que tenga solución única debe existir $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$:

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

Si las variables que forman $\mathbf{X}$ son linealmente dependientes, entonces no existe $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$ (multicolinealidad perfecta).

Introducción

¿Qué ocurre si son casi linealmente independientes? (multicolinealidad aproximada):

- Pequeños cambios en los datos pueden suponer cambios sustanciales en las estimaciones de los coeficientes de los regresores.
- Tendencia a no rechazar que los coeficientes de los regresores son cero debido a desviaciones típicas estimadas de los coeficientes "infladas".
- Coeficiente de determinación alto y, en consecuencia, tendencia a rechazar que todos los coeficientes son cero de forma simultánea.
- Incumplimiento del *ceteris paribus*.

En definitiva, posibilidad de obtener resultados inestables y/o contradictorios.

Tipos de Multicolinealidad y su Detección

Las causas que producen multicolinealidad en un modelo son diversas:

Mientras que Marquardt y Snee (1975) [1]

Multicolinealidad no esencial: relación lineal de las variables exógenas con la constante (es sabido que se solventa centrando las variables).

Multicolinelidad esencial: relación lineal entre las variables exógenas (excluida la constante).

Al igual que la batería de herramientas para detectarlas:

Herramientas para la detección de la multicolienalidad aproximada

Detectan sólo la no esencial: coeficiente de variación.

Detectan sólo la esencial: matriz de correlaciones lineales simples y su determinante, factor de inflación de la varianza.

Detectan ambas: número de condición o índice de Stewart.

Factor de Inflación de la Varianza

Definición del FIV

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

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

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

$$\text{comparación entre ambos modelos:} \quad \frac{\widehat{var}\left(\widehat{\beta}_i\right)}{\widehat{var}\left(\widehat{\beta}_{i,o}\right)} = \frac{1}{1 - R_i^2} = FIV(i).$$

El FIV mide cuánto aumenta la varianza de $\widehat{\beta}_i$ con respecto al caso ortogonal donde se supone que $R_i^2 = 1$.

¡Ojo! ¿Es razonable pensar que en el caso ortogonal cambia el valor de R_i^2 pero $var(\mathbf{X}_i)$ no? ¿Qué ocurrirá con la estimación de σ?

Un modelo ortogonal alternativo

A partir de una descomposición QR para la matriz $\mathbf{X}$ se tiene que $\mathbf{X} = \mathbf{X}_o \cdot \mathbf{P}$ donde $\mathbf{X}_o$ es una matriz ortonormal de las mismas dimensiones de $\mathbf{X}$ y $\mathbf{P}$ es una matriz triangular superior.

Modelo ortogonal: $\mathbf{Y} = \mathbf{X}_o \cdot \boldsymbol{\beta}_o + \mathbf{W}$.

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

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

En tal caso, como $\widehat{\text{var}} \left(\widehat{\beta}_i \right) = \frac{\widehat{\sigma}^2}{n \cdot var(\mathbf{X}_i)} \cdot \frac{1}{1-R_i^2}$, entonces:

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

Redefinición del Factor de Inflación de la Varianza

Teniendo en cuenta que $\frac{FIV(i)}{n \cdot var(\mathbf{X}_i)} = \frac{1}{SCR_i}$ donde SCR_i es la suma de cuadrados de los residuos de la regresión de $\mathbf{X}_i$ sobre el resto de variables independientes, $\mathbf{X}_{-i}$:

Factor de Inflación de la Varianza Redefinido

Se redefine el FIV como:

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

Para datos de longitud unidad, $\mathbf{X}_i^t \mathbf{X}_i = 1$, FIVR coincide con el índice de Stewart.
(esta transformación se usa para calcular el Número de Condición)

- Como $FIVR(i) = \frac{1}{1-a_i} > 0$, entonces $a_i = \mathbf{X}_i^t \mathbf{X}_{-i} \cdot \left(\mathbf{X}_{-i}^t \mathbf{X}_{-i} \right)^{-1} \cdot \mathbf{X}_{-i}^t \mathbf{X}_i \leq \mathbf{X}_i^t \mathbf{X}_i = 1$.
- Si $\mathbf{X}_i$ es ortogonal a $\mathbf{X}_{-i}$, entonces $a_i = 0$.
- $a_i \in [0, 1]$ se interpreta como porcentaje de variabilidad debida a $\mathbf{X}_i$.
- $FIVR(i) \geq 1$ para $i = 1, \dots, p$.

¿Umbrales para a_i?

Se realizada una simulación en la que se generan datos para un modelo de regresión limeal múltiple en el que $p = 3$ y donde las variables que forman la matriz $\mathbf{X}$ se generan como sigue:

$$\mathbf{X}_i = \sqrt{1-\gamma^2} \cdot \mathbf{M}_i + \gamma \cdot \mathbf{M}_2, \ \ i = 1, 2$$

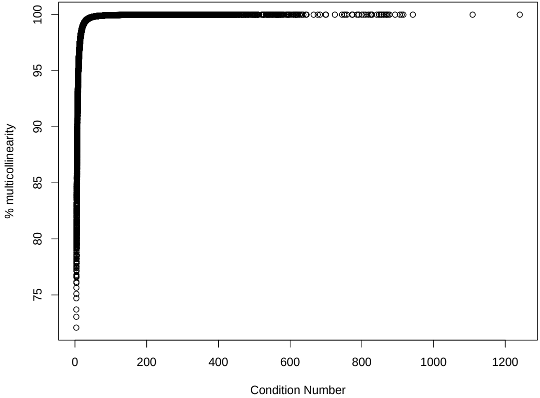
donde $\gamma \in \{0, 0.1, 0.2, \dots, 0.8, 0.9, 0.95, 0.96, 0.97, 0.98, 0.99\}$, $\mathbf{M}_i \sim N(1, \sigma)$ y $\sigma \in \{0.001, 0.002, \dots, 0.009, 0.01, 0.02, \dots, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

La simulación anterior se realiza para distintos tamaños de muestra, concretamente, se considera que $n \in \{15, 20, 25, \dots, 190, 195, 200\}$.

Por tanto, se simulan 13680 modelos en los que se calcula el máximo FIVR, el máximo porcentaje de multicolinealidad y el número de condición.

¿Umbrales para a_i?

$$\widehat{\%} = 24.29787 \cdot log(NC), \ R^2 = 92.97\,\% \rightarrow \text{si } NC > 30, \text{ entonces } \% > 82.64185$$



Ejemplo 1

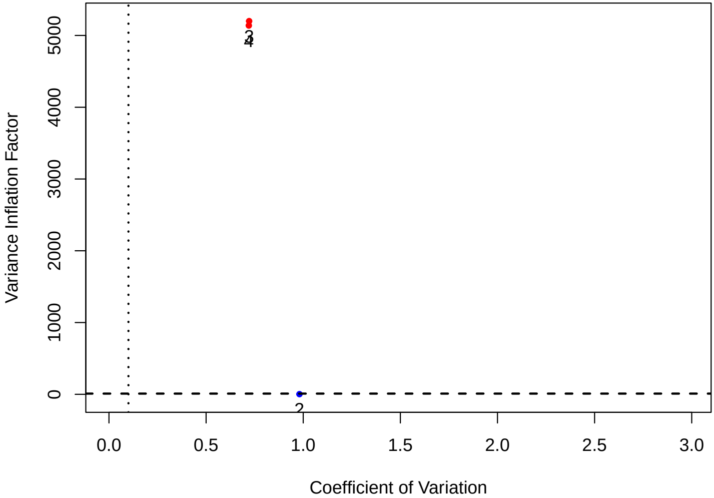
Salmerón, R., Rodríguez, A., García, C.B. y García, J. (2020) [3]

En este trabajo se analiza la relación entre el número de empleados (**NE**) de 15 compañias españolas en función de los activos fijos (**FA**, en euros), el ingreso operativo (**OI**, en euros) y las ventas (**S**, en euros).
En este caso se determina que existe una alta relacion lineal entre **OI** y **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)
data1 = cbind(rep(1, length(NE)), FA, OI, S)

library(rvif)
RVIF(data1, 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
```

Ejemplo 1



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

Ejemplo 2

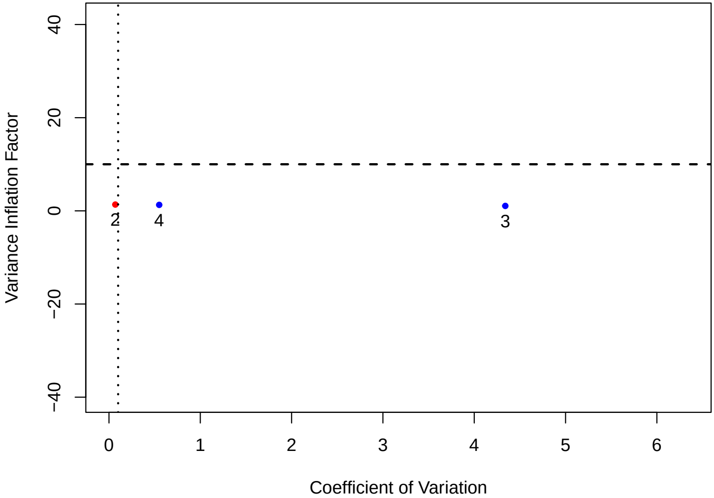
Salmerón, R., Rodríguez, A. y García, C.B. (2019) [4]

Consideran un modelo financiero en el que el euribor (**E**, %) es analizado a partir del índice de precios de consumo harmonizado (**HICP**, %), la balanza de pagos a la cuenta corriente neta (**BC**, millones de euros) y el déficit público a cuentas no financieras netas (**GD**, millones de euros).
En este modelo se establece que **HICP** está relacionada con la constante.

```
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)
data2 = cbind(rep(1, length(E)), HIPC, BC, GD)

RVIF(data2, l_u = T)                    # library(rvif)
                 RVIF       %
Intercept  250.294157 99.6005
Variable 2 280.136873 99.6430
Variable 3   1.114787 10.2967
Variable 4   5.525440 81.9019
```

Ejemplo 2



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

Principales resultados

Conclusiones

- Según Johnston (1984) [5] "el caso ortogonal no significa que sea una meta realizable pero se usa como punto de referencia para medir el aumento relativo de la varianza muestral de los estimadores".
- Ahora bien, aunque no sea una meta realizable si debe ser creible.
- Proponemos una referencia ortogonal alternativa que conduce a una redefinición del Factor de Inflación de la Varianza.
- Esta definición permite, por ejemplo, usando el paquete *RVIF* de **R**:
  a) detectar la multicolinealidad aproximada preocupante de tipo esencial y no esencial,
  b) qué variable la provoca y
  c) cuantificar el porcentaje de multicolinealidad debida a una variable independiente concreta.
- Se establece un umbral para determinar que el porcentaje de multicolinealidad cuantificado es preocupante.

Bibliografía (por orden de aparición) y Agradecimientos

- 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., García, C.B. y García, J. (2020). "The VIF and MSE in raise regression". Mathematics, 8 (4), pp. 605.
- Salmerón, R., Rodríguez, A. y García, C.B. (2019). "Diagnosis and quantification of the non-essential collinearity". Computational Statistics, 35, pp. 647–666.
- Johnston, J. (1972). "Econometric Methods". McGraw-Hill.

Este trabajo ha sido financiado por el proyecto PP2019-EI-02 de la Universidad de Granada titulado "Redefinición del factor de inflación de la varianza y de sus umbrales".

Detección de la multicolinealidad a partir del Factor de Inflación de la Varianza Redefinido

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

3R4EDI
Tercer Congreso R para Empresa, Docencia e Investigación

Almagro, 19-20 octubre 2023