Redefinición del Factor de Inflación de la Varianza

R. Salmerón (romansg@ugr.es) **C.B. García** (cbgarcia@ugr.es)
Dpto. de Métodos Cuantitativos para la Economía y la Empresa
Universidad de Granada
J. García (jgarcia@ual.es)
Departamento de Economía y Empresa, Universidad de Almería

XXXIX Congreso Nacional de Estadística e Investigación Operativa
XIII Jornadas de Estadística Pública

SEIO, Granada 7 al 10 de junio de 2022

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

Las causas que producen multicolinealidad en un modelo son diversas.

Según Spanos y McGuirk (2002) [1]

Multicolinealidad sistemática: debida a un problema estructural, es decir, a la alta correlación lineal de las variables exógenas consideradas.

Multicolinelidad errática: debido a un problema puramente numérico, es decir, a un mal condicionamiento de los datos considerados.

Mientras que Marquardt y Snee (1975) [2]

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).

En este caso nos centraremos en la multicolinealidad aproximada del tipo esencial/no-esencial.

Detección de Multicolinealidad

Herramientas para la detección de la multicolienalidad aproximada

Matriz de correlaciones lineales simples y su determinante, coeficiente de variación, factor de inflación de la varianza, número de condición o índice de Stewart.

Factor de Inflación de la Varianza

Una de las medidas más usadas para detectar el grado de multicolinealidad existente en el MLG es el Factor de Inflación de la Varianza (FIV) dado por:

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

donde R_i^2 es el coeficiente de determinación de la regresión de $\mathbf{X}_i$ sobre el resto de variables independientes, $\mathbf{X}_{-i}$.
Si esta medida es superior a 10 se supone que el grado de multicolinealidad presente en el modelo es preocupante.

¡Ojo! El VIF no tiene en cuenta la relación de las variables exógenas del modelo, $\mathbf{X}_2 \ldots \mathbf{X}_p$, con la constante, $\mathbf{X}_1 = \mathbf{1}$ (Salmerón, R., García, C.B. y García, J. (2018) [3]). Por tanto, no detecta la multicolinealidad no esencial.

Factor de Inflación de la Varianza

¿Cómo surge la expresión anterior?

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

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

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

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

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?

Factor de Inflación de la Varianza

Si hubiese ortogonalidad...

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

luego:

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

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

Por tanto, el planteamiento inicial es válido si $\overline{\mathbf{X}}_i = 0$ para $i = 2, \ldots, p$ (para $i = 1$ es imposible) y si la estimación de σ coincide en el modelo original y ortogonal.

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:

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

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:

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

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 $FIVR(i) > 0$, 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 $i = 1, \ldots, p$.

Ejemplo 1

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 la variable **HICP** está relacionada con la constante al tener un coeficiente de variación igual a 0.06957876.

Variable	FIV	FIVR	%
Constante		250.294157	99.6005
HIPC	1.349666	280.136873	99.6430
BC	1.058593	1.114787	10.2967
GD	1.283815	5.52544	81.9019

Ejemplo 2

Salmerón, R., Rodríguez, A., García, C.B. y García, J. (2020) [5]

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**.

Variable	FIV	FIVR	%
Constante		2.984146	66.4896
FA	2.45664	5.011397	80.0455
OI	5200.3153	15186.744870	99.9934
S	5138.53548	15052.679178	99.9934

¿Umbrales?

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

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

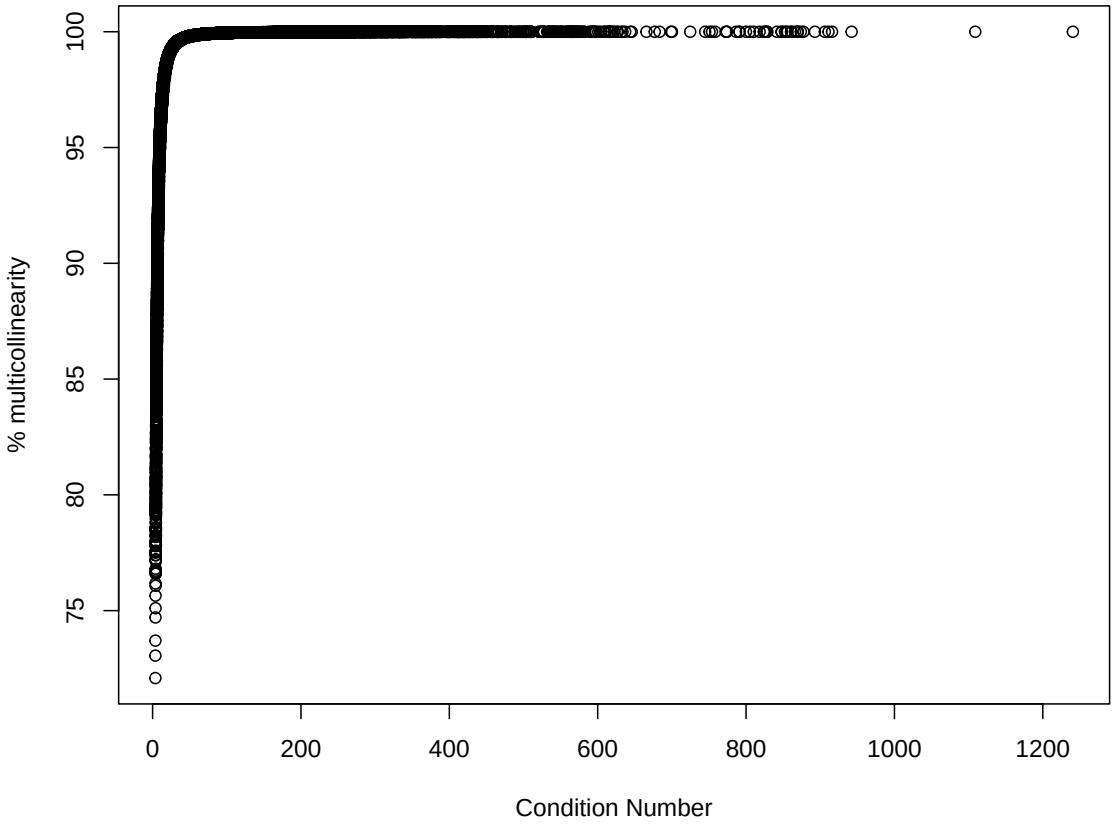
Por tanto, se simulan 13680 modelos en los que se calcula el máximo FIVR, el máximo porcentaje de multicolinealidad y el número de condición.

Esta forma de simular datos ha sido previamente usada en [6], [7], [8], [9], [10], [11], [12] o [13].

¿Umbrales?

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

el umbral funciona en los ejemplos considerados



Principales resultados

Conclusiones

- Según Johnston (1984) [14] "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 a) detectar la multicolinealidad aproximada de tipo esencial y no esencial y b) 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)

- Spanos, A. y McGuirk, A. (2002). "The problem of near-multicollinearity revisited: erratic vs systematic volatily". Journal of Econometrics, 108 (2), pp. 365–393.
- 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.

Bibliografía (por orden de aparición)

- McDonald, G.C. y Galarneau, D.I. (1975). "A Monte Carlo evaluation of some ridge type estimators". Journal of American Statistical Association, 70, pp. 407–416.
- Wichern, D.W. y Churchill, G.A. (1978). "A comparison of ridge estimators". Technometrics, 20, pp. 301–311.
- Gibbons, D.G. (1981). "A simulation study of some ridge estimators". Journal of American Statistical Association, 76, pp. 131–139.
- Kibria, B. (2003). "Performance of some new ridge regression estimators". Communications in Statistics - Simulation and Computation, 32 (2), pp. 419–435.
- Salmerón, R., García, J., López, M.M. y García, C.B. (2016). "Collinearity diagnostic in ridge estimation through the variance inflation factor". Journal of Applied Statistics, 43 (10), pp. 1831–1849.

Bibliografía (por orden de aparición)

- Salmerón, R. and García, C. y García, J. (2018). "Variance inflation factor and condition number in multiple linear regression". Journal of Statistical Computation and Simulation, 88, pp. 2365–2384.
- Rodríguez, A., Salmerón, R. y García, C. (2021). "Obtaining a threshold for the stewart index and its extension to ridge regression". Computational Statistics, 36, pp. 1011–1029.
- Rodríguez, A., Salmerón, R. y García, C. (2022). "The coefficient of determination in the ridge regression". Communications in Statistics - Simulation and Computation, 51 (1), pp. 201–219.
- Johnston, J. (1972). "Econometric Methods". McGraw-Hill.

Agradecimientos

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" y
- por el proyecto FEDER A-SEJ-496-UGR20 de la Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades titulado "Disrupción en el problema de multicolinealidad: una nueva visión para su diagnóstico y tratamiento".

Redefinición del Factor de Inflación de la Varianza

R. Salmerón (romansg@ugr.es) **C.B. García** (cbgarcia@ugr.es)
Dpto. de Métodos Cuantitativos para la Economía y la Empresa
Universidad de Granada
J. García (jgarcia@ual.es)
Departamento de Economía y Empresa, Universidad de Almería

XXXIX Congreso Nacional de Estadística e Investigación Operativa
XIII Jornadas de Estadística Pública

SEIO, Granada 7 al 10 de junio de 2022