Multicolinealidad aproximada en el modelo de regresión lineal simple e implicaciones en la regresión múltiple

Román Salmerón (romansg@ugr.es)
Catalina García (cbgarcia@ugr.es)
Ainara Rodríguez (ainararodriguez@correo.ugr.es)

Universidad de Granada
Departamento de Métodos Cuantitativos para la Economía y la Empresa

SEIO, 29, 30 y 31 de Mayo y 1 de Junio de 2018

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:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \cdots + \beta_p X_p + u = \mathbf{X}\boldsymbol{\beta} + u. \qquad (1)$$

Aplicando MCO se llega al sistema de ecuaciones normales:

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

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

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

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

Introducción

¿Qué ocurre si son casi linealmente independientes? (multicolinealidad aproximada): resultados inestables y contradictorios.

- 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.
- Coeficiente de determinación alto y, en consecuencia, tendencia a rechazar que todos los coeficientes son cero de forma simultánea.

Posibles soluciones:

- Mejora del diseño muestral, aumento del tamaño de la muestra o usar información a priori.
- Eliminar variables que se consideran problemáticas.
- Centrar variables.
- Usar métodos de estimación alternativos a MCO (como es la regresión cresta, alzada, LASSO, etc).

Introducción

Las causas que producen multicolinealidad en un modelo son diversas.

Según Spanos y McGuirk (2002)

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 Marquandt y Snee (1975)

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).

Introducción

Luego, se podrían distinguir los siguientes cuatro casos:

Multicolinealidad	Sistemática	Errática
No esencial	1	2
Esencial	3	4

En este trabajo nos centraremos en multicolinealidad no esencial (primera fila). Más concretamente, en su detección (partiendo del modelo lineal simple), ya que el tratamiento queda claro (centrar las variables).

Introducción

Luego, se podrían distinguir los siguientes cuatro casos:

Multicolinealidad	Sistemática	Errática
No esencial	1	2
Esencial	3	4

En este trabajo nos centraremos en multicolinealidad no esencial (primera fila). Más concretamente, en su detección (partiendo del modelo lineal simple), ya que el tratamiento queda claro (centrar las variables).

Multicolinealidad

Se entiende que en el modelo de regresión lineal general hay multicolinealidad aproximada cuando existen dos o más variables exógenas relacionadas linealmente de forma aproximada.

¿Es posible que en el modelo lineal simple exista multicolinealidad?

$$Y = \beta_0 + \beta_1 X_1 + u. \tag{2}$$

Si hay, tiene que ser del tipo no esencial y, en principio, debida a que $Var(X_1) \simeq 0$.

Índice

Detección: MLG y FIV

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:

$$FIV(i) = \frac{1}{1 - R_i^2}, \quad i = 2, \ldots, p, \tag{3}$$

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, $\mathbf{1}$ (Salmerón, R., García, C.B. y García, J. (2018)). Por tanto, no detecta la multicolinealidad no esencial.

Detección: MLG y NC

Número de Condición

Otra medida muy extendida es el Número de Condición (NC), el cual viene dado por:

$$NC = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}, \tag{4}$$

donde λ_{max} y λ_{min} son, respectivamente, los autovalores máximo y mínimo de $\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$ donde:

$$\widetilde{\mathbf{X}} = [\widetilde{\mathbf{1}}\ \widetilde{\mathbf{X}}_2 \dots \widetilde{\mathbf{X}}_p],$$

$$\widetilde{\mathbf{1}} = \frac{\mathbf{1}}{\sqrt{n}}, \quad \widetilde{\mathbf{X}}_i = \frac{\mathbf{X}_i}{\sqrt{\sum_{j=1}^{n} X_{ji}^2}}, \ i = 2, \dots, p.$$

Si esta medida es superior a 20 se supone que el grado de multicolinealidad presente en el modelo es moderado y si es superior a 30 preocupante.

¡Ojo! El NC tiene en cuenta la relación de las variables exógenas del modelo, $\mathbf{X}_2 \dots \mathbf{X}_p$, con la constante, $\mathbf{1}$.

Otras medidas: determinante de la matriz de correlaciones próximo a cero (también ignora la relación con la constante).

Detección: MLS y NC

Sólo el NC detectaría multicolinealidad en el MLS

$$NC = \sqrt{\frac{1+a}{1-a}}, \quad a = \frac{\overline{\mathbf{X}}_1}{\sqrt{Var(\mathbf{X}_1) + \overline{\mathbf{X}}_1^2}} > 0 \ (b = -a > 0). \tag{5}$$

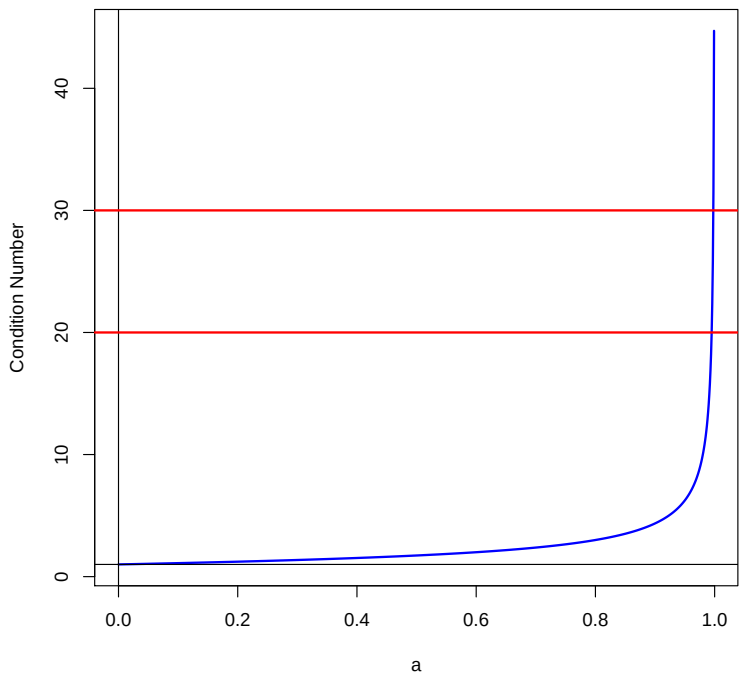

Figura : Representación del CN en el MLS en función de $a \in [0, 0{,}999]$

Detección: MLS y NC

¿Qué hace que $a = \frac{\overline{\mathbf{X}}_1}{\sqrt{Var(\mathbf{X}_1)+\overline{\mathbf{X}}_1^2}}$ sea próximo a 0 o a 1?

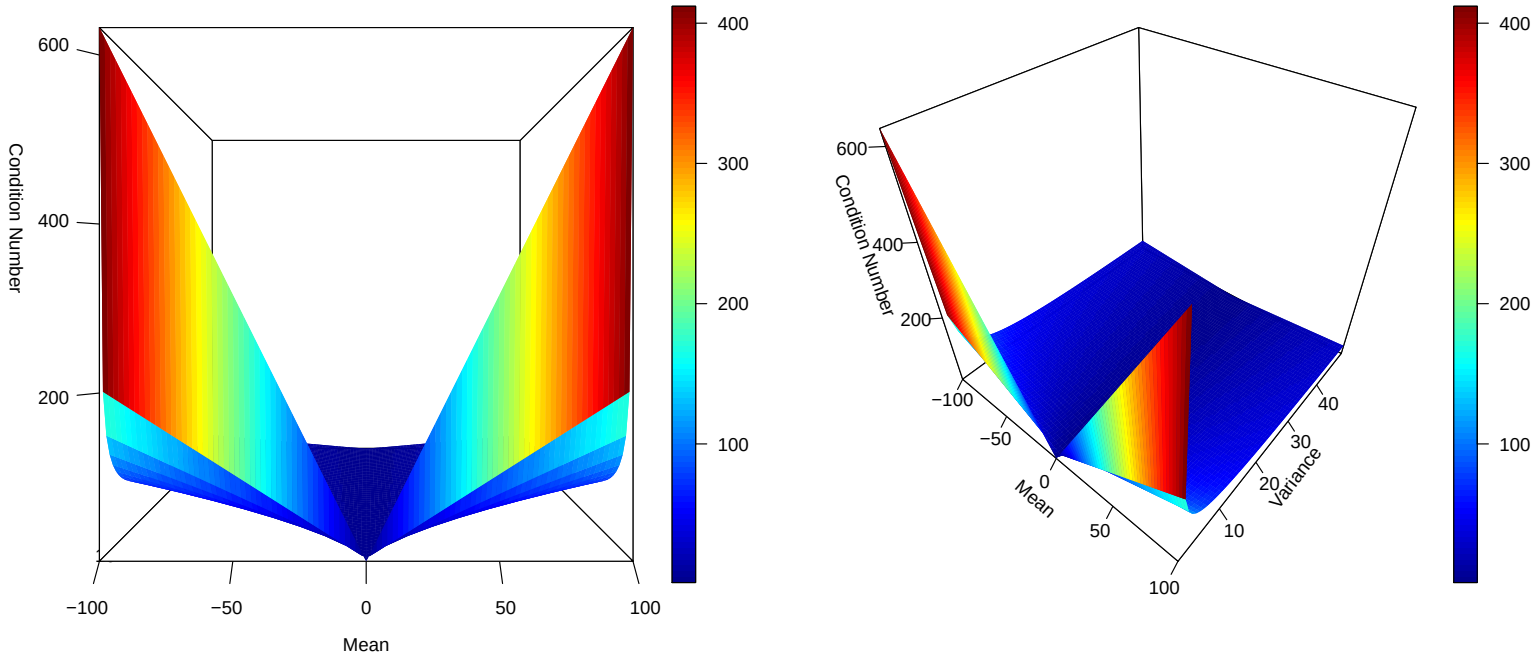

Figura : Simulación del CN en función de la media ($\mathbf{X}_1 \in [-100, 100]$) y varianza ($Var(\mathbf{X}_1) \in [0{,}1, 50]$)

Detección: MLS y NC

A partir de la expresión (5) se tiene que:

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

mientras que:

$$a > k \Leftrightarrow Var(\mathbf{X}_1) < \frac{1 - k^2}{k^2} \cdot \overline{\mathbf{X}}_1^2.$$

Luego, si $k = \frac{h^2-1}{h^2+1}$ se tiene que:

$$NC > h \Leftrightarrow Var(\mathbf{X}_1) < \frac{4h^2}{(h^2 - 1)^2} \cdot \overline{\mathbf{X}}_1^2.$$

Detección: MLS y NC

A partir de la expresión (5) se tiene que:

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

mientras que:

$$a > k \Leftrightarrow Var(\mathbf{X}_1) < \frac{1 - k^2}{k^2} \cdot \overline{\mathbf{X}}_1^2.$$

Luego, si $k = \frac{h^2-1}{h^2+1}$ se tiene que:

$$NC > h \Leftrightarrow Var(\mathbf{X}_1) < \frac{4h^2}{(h^2 - 1)^2} \cdot \overline{\mathbf{X}}_1^2.$$

Para $h = 20$:

$$Var(\mathbf{X}_1) < 0{,}01005019 \cdot \overline{\mathbf{X}}_1^2 \Leftrightarrow \frac{Var(\mathbf{X}_1)}{\overline{\mathbf{X}}_1^2} < 0{,}01005019 \Leftrightarrow CV(\mathbf{X}_1) < 0{,}1002506.$$

Índice

Índice de Stewart

Dada $\mathbf{A}_{n\times p} = [\mathbf{A}_i, \mathbf{A}_{-i}]$, Stewart (1987) definió el siguiente índice para detectar relación entre $\mathbf{A}_i$ y el resto de columnas de $\mathbf{A}$ para $i = 1, \ldots, p$:

$$k_i^2 = \frac{|\mathbf{A}_{-i}^t \mathbf{A}_{-i}| \cdot \mathbf{A}_i^t \mathbf{A}_i}{|\mathbf{A}^t \mathbf{A}|} = \frac{\mathbf{A}_i^t \mathbf{A}_i}{\mathbf{A}_i^t \mathbf{A}_i - \mathbf{A}_i^t \mathbf{A}_{-i} \cdot \left(\mathbf{A}_{-i}^t \mathbf{A}_{-i}\right)^{-1} \cdot \mathbf{A}_{-i}^t \mathbf{A}_i}. \tag{6}$$

Si $\mathbf{A}_i^t \mathbf{A}_{-i} = \mathbf{0}$ (ortogonalidad), entonces $k_i^2 = 1$.

Índice de Stewart

Dada $\mathbf{A}_{n\times p} = [\mathbf{A}_i, \mathbf{A}_{-i}]$, Stewart (1987) definió el siguiente índice para detectar relación entre $\mathbf{A}_i$ y el resto de columnas de $\mathbf{A}$ para $i = 1, \ldots, p$:

$$k_i^2 = \frac{|\mathbf{A}_{-i}^t \mathbf{A}_{-i}| \cdot \mathbf{A}_i^t \mathbf{A}_i}{|\mathbf{A}^t \mathbf{A}|} = \frac{\mathbf{A}_i^t \mathbf{A}_i}{\mathbf{A}_i^t \mathbf{A}_i - \mathbf{A}_i^t \mathbf{A}_{-i} \cdot \left(\mathbf{A}_{-i}^t \mathbf{A}_{-i}\right)^{-1} \cdot \mathbf{A}_{-i}^t \mathbf{A}_i}. \tag{6}$$

Si $\mathbf{A}_i^t \mathbf{A}_{-i} = \mathbf{0}$ (ortogonalidad), entonces $k_i^2 = 1$.
Aplicando este índice al MLG dado en (1), es decir, $\mathbf{A} = \mathbf{X}$ con $\mathbf{X}_1 = \mathbf{1}$:

$$\begin{aligned} k_i^2 &= \frac{\mathbf{X}_i^t \mathbf{X}_i}{\mathbf{X}_i^t \mathbf{X}_i - \mathbf{X}_i^t \mathbf{X}_{-i} \cdot \left(\mathbf{X}_{-i}^t \mathbf{X}_{-i}\right)^{-1} \cdot \mathbf{X}_{-i}^t \mathbf{X}_i} \\ &= \begin{cases} \frac{1}{1 - \frac{1}{n} \overline{\mathbf{x}}_{-i} \cdot \left(\mathbf{x}_{-i}^t \mathbf{x}_{-i}\right)^{-1} \cdot \overline{\mathbf{x}}_{-i}^t} & , i = 1 \\ \frac{\mathbf{x}_i^t \mathbf{x}_i}{SSR_i} = FIV(i) + n \cdot \frac{\overline{\mathbf{x}}_i^2}{SSR_i} & , i = 2, \ldots, p, \end{cases} \end{aligned} \tag{7}$$

donde $\overline{\mathbf{X}}_{-i}$ es un vector formado por la suma de los elementos de las variables $\mathbf{X}_{-i}$ y SSR_i es la suma de los cuadrados de los residuos de la regresión auxiliar usada al calcular el FIV.

Índice de Stewart

Entonces:

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

es la proporción de multicolinealidad esencial existente en $\mathbf{X}_i$ y:

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

es la proporción de multicolinealidad no esencial existente en $\mathbf{X}_i$, $i = 2, \ldots, p$.

Índice

Ejemplo

$$\textbf{Euribor} = \beta_1 + \beta_2\textbf{IP} + \beta_3\textbf{BP} + \beta_3\textbf{DG} + \textbf{u},$$

	Estimation	FIV	$Var(\mathbf{X}_i) < 0{,}01 \cdot \overline{\mathbf{X}}_i^2$
Intercept	4.376 (1.258)		
IPC	-0.002064 (0.013)	1.3506	$33{,}07189 < 0{,}01 \cdot 54{,}76202^2$
BP	$-3.647 \cdot 10^{-5}$ $(3.897 \cdot 10^{-6})$	1.0587	$105{,}2187 \not< 0{,}01 \cdot 450388464^2$
DG	$1.971 \cdot 10^{-5}$ $(2.202 \cdot 10^{-6})$	1.2846	$4837{,}298 \not< 0{,}01 \cdot 1710699512^2$
R^2	0.8196		
F_{exp}	70.68		
$\hat{\sigma}^2$	0.29722		
CN	33.07189		

Ejemplo

$$\textbf{Euribor} = \beta_1 + \beta_2\textbf{IP} + \beta_3\textbf{BP} + \beta_3\textbf{DG} + \textbf{u},$$

	Estimation	FIV	$Var(\mathbf{X}_i) < 0{,}01 \cdot \overline{\mathbf{X}}_i^2$
Intercept	4.376 (1.258)		
IPC	-0.002064 (0.013)	1.3506	$33{,}07189 < 0{,}01 \cdot 54{,}76202^2$
BP	$-3.647 \cdot 10^{-5}$ $(3.897 \cdot 10^{-6})$	1.0587	$105{,}2187 \not< 0{,}01 \cdot 450388464^2$
DG	$1.971 \cdot 10^{-5}$ $(2.202 \cdot 10^{-6})$	1.2846	$4837{,}298 \not< 0{,}01 \cdot 1710699512^2$
R^2	0.8196		
F_{exp}	70.68		
$\hat{\sigma}^2$	0.29722		
CN	33.07189		

	k_i^2	$FIV(i)$	$n \cdot \frac{\overline{\mathbf{X}}_i^2}{SSR_i}$	% esencial	% no esencial
Intercept	250.2812			100 %	0 %
Índice de Precios Consumo (IPC)	280.1747	1.3506	278.824	0.482 %	99.5179 %
Balance de Pagos (BP)	1.1149	1.0587	0.0562	94.959 %	5.0407 %
Déficit del Gobierno (DG)	5.52907	1.2846	4.2444	23.234 %	76.765 %

Ejemplo

	Estimation
Intercept	4.158 (0.1838)
IPC*	-0.002064 (0.01262)
BP	$-3.647 \cdot 10^{-5}$ $(3.897 \cdot 10^{-6})$
DG	$1.971 \cdot 10^{-5}$ $(2.202 \cdot 10^{-6})$
R^2	0.8196
F_{exp}	70.68
$\hat{\sigma}^2$	0.29722
CN	5.32927

	k_i^2	$FIV(i)$	$n \cdot \frac{\overline{\mathbf{x}}_i^2}{SSR_i}$	% esencial	% no esencial
Intercept	5.344101			0 %	100 %
IPC	1.3506	1.3506	0	100 %	0 %
BP	1.1149	1.0587	0.0562	94.959 %	5.0407 %
DG	5.52907	1.2846	4.2444	23.234 %	76.765 %

La multicolinealidad que queda (esencial y no esencial) no es preocupante.

Índice

Conclusiones

Conclusiones

- La multicolinealidad no esencial se sabe cómo tratarla, pero (dentro de nuestro conocimiento) no cómo detectarla.
- A partir del MLS se obtiene una regla en función del coeficiente de variación que permite establecer si una variable exógena del MLG está relacionada linealmente con la constante.
- El índice de Stewart permite determinar qué porcentaje de multicolinealidad esencial y no esencial hay en cada variable exógena.
- Identificando las variables exógenas en las que existe este tipo de multicolinealidad, evitamos centrar variables de forma innecesaria o aplicar otros métodos más complejos como la regresión cresta, LASSO, etc.

Índice

Bibliografía

- Marquandt, 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. DOI: 10.1080/00949655.2018.1463376.
- Spanos, A. y McGuirk, A. (2002). "The problem of near-multicollinearity revisited: erratic vs systematic volatily". Journal of Econometrics, 108 (2), pp. 365–393.
- Stewart, G. (1987). "Collinearity and least squares regression". Statististical Science, 2 (1), 68–100.

Multicolinealidad aproximada en el modelo de regresión lineal simple e implicaciones en la regresión múltiple

Román Salmerón (romansg@ugr.es)
Catalina García (cbgarcia@ugr.es)
Ainara Rodríguez (ainararodriguez@correo.ugr.es)

Universidad de Granada
Departamento de Métodos Cuantitativos para la Economía y la Empresa

SEIO, 29, 30 y 31 de Mayo y 1 de Junio de 2018