Regresión Alzada y Multicolinealidad NO Esencial

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)
Dpto. de Economía y Empresa
Universidad de Almería

ASEPELT, Jaén 6-9 de Octubre de 2021

XXXIV Congreso Internacional de Economía Aplicada

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:

$$\mathbf{Y} = \beta_1 + \beta_2 \cdot \mathbf{X}_2 + \cdots + \beta_i \cdot \mathbf{X}_i + \cdots + \beta_p \cdot \mathbf{X}_p + \mathbf{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \quad (1)$$

Aplicando Mínimos Cuadrados Ordinarios (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 $(\mathbf{X}'\mathbf{X})^{-1}$:

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

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

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, elastic-net, etc).
- Mínimos Cuadrados Restringidos (MCR).

Introducción

Las causas que producen multicolinealidad en un modelo son diversas.

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).

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

Salmerón, R. y otros (2020) [3]

Valores de las variables independientes del modelo (1) del coeficiente de variación inferiores a 0.1002506 indican que el grado de multicolinealidad aproximada del tipo no esencial es preocupante.

Introducción

Mitigación de la multicolinealidad no esencial

Este tipo de multicolinealidad se elimina centrando la variable (media cero $=$ coeficiente de variación infinito).

Sin embargo, el centrado de variables no mitiga el grado de multicolinealidad esencial que pudiera existir.

¿La regresión alzada mitiga ambos tipos de multicolinealidad?

La esencial sí (ver Roldán y otros (2020) [2] y Salmerón y otros (2020) [4]).
La otra...

Regresión Alzada

La regresión alzada consiste en estimar por MCO el modelo $\mathbf{Y} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{v}$ donde:

$$\widetilde{\mathbf{X}} = \left[\mathbf{1}\ \mathbf{X}_2\ \ \ldots\ \ \widetilde{\mathbf{X}}_i\ \ \ldots\ \ \mathbf{X}_p\right], \quad \widetilde{\mathbf{X}}_i = \mathbf{X}_i + \lambda_i \mathbf{e}_i, \quad \lambda_i \geq 0, \quad i \geq 2,$$

con $\mathbf{e}_i$ los residuos de la regresión auxiliar:

$$\mathbf{X}_i = \alpha_1 + \alpha_2 \cdot \mathbf{X}_2 + \cdots + \alpha_{i-1} \cdot \mathbf{X}_{i-1} + \alpha_{i+1} \cdot \mathbf{X}_{i+1} + \cdots + \alpha_p \cdot \mathbf{X}_p + \mathbf{w} = \mathbf{X}_{-i}\boldsymbol{\alpha} + \mathbf{w}. \quad (2)$$

Como $\mathbf{e}_i$ es ortogonal a $\mathbf{X}_{-i}$, al añadirlo a $\mathbf{X}_i$ se obtiene que $\widetilde{\mathbf{X}}_i$ se *separe geométricamente* de $\mathbf{X}_{-i}$.

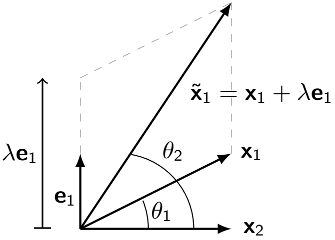


Figura : Ejemplo ilustrativo para dos variables estandarizadas

Coeficiente de Variación en la Regresión Alzada

Como en $\mathbf{X}_{-i}$ está el término constante, se presupone que la multicolinealidad del tipo no esencial se mitiga.

En efecto, teniendo en cuenta que $CV(\widetilde{\mathbf{X}}_i) = \frac{\sqrt{var(\widetilde{\mathbf{X}}_i)}}{\overline{\widetilde{\mathbf{X}}}_i}$ y que:

- $var(\widetilde{\mathbf{X}}_i) = var(\mathbf{X}_i + \lambda_i \cdot \mathbf{e}_i) = var(\mathbf{X}_i) + \lambda_i^2 \cdot var(\mathbf{e}_i) + 2 \cdot \lambda_i \cdot cov(\mathbf{X}_i, \mathbf{e}_i) = var(\mathbf{X}_i) + (\lambda_i^2 + 2 \cdot \lambda_i) \cdot var(\mathbf{e}_i)$ ya que:
$$cov(\mathbf{X}_i, \mathbf{e}_i) = \frac{1}{n} \cdot \mathbf{X}_i^t \mathbf{e}_i = \frac{1}{n} \cdot (\mathbf{e}_i^t \mathbf{e}_i + \widehat{\mathbf{X}}_i^t \mathbf{e}_i) = \frac{1}{n} \cdot \mathbf{e}_i^t \mathbf{e}_i,$$
donde $\widehat{\mathbf{X}}_i$ es la estimación de $\mathbf{X}_i$ obtenida a partir de la regresión auxiliar (2) y, por tanto, se verifica que $\widehat{\mathbf{X}}_i^t \mathbf{e}_i = 0$.
- $\overline{\widetilde{\mathbf{X}}}_i = \overline{\mathbf{X}}_i + \lambda_i \cdot \overline{\mathbf{e}}_i = \overline{\mathbf{X}}_i$.

Se verifica que:

$$CV(\widetilde{\mathbf{X}}_i) = \frac{\sqrt{var(\mathbf{X}_i) + (\lambda_i^2 + 2 \cdot \lambda_i) \cdot var(\mathbf{e}_i)}}{\overline{\mathbf{X}}_i}, \quad i \geq 2.$$

En tal caso, como $(\lambda_i^2 + 2 \cdot \lambda_i) \cdot var(\mathbf{e}_i) \geq 0$, entonces $CV(\widetilde{\mathbf{X}}_i) \geq CV(\mathbf{X}_i) = \frac{\sqrt{var(\mathbf{X}_i)}}{\overline{\mathbf{X}}_i}$ para $\lambda_i \geq 0$. Además, claramente, creciente en λ_i.

Ejemplo

$$\mathbf{Y} = 1 + 2\mathbf{X}_2 - 3\mathbf{X}_3 - 4\mathbf{X}_4 + \mathbf{u}, \quad \mathbf{u} \sim N(0,2)$$

$$N = 100, \quad \mathbf{X}_2 \sim N(4, 0.01), \quad \mathbf{X}_3 \sim N(-4, 4), \quad \mathbf{X}_4 = \mathbf{X}_3 + \mathbf{p};\ \mathbf{p} \sim N(0, 0.3),$$

$$\mathbf{R} = \begin{pmatrix} 1 & -0.0875466 & -0.08189602 \\ -0.08754660 & 1 & 0.99740733 \\ -0.08189602 & 0.9974073 & 1 \end{pmatrix}$$

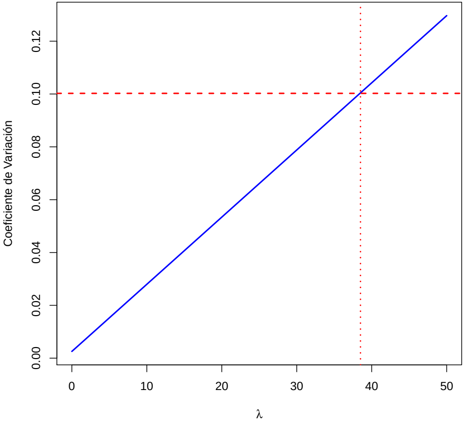
Herramienta	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$
CVs	0.002559791	1.049570454	1.059849221
FIVs	1.013525	194.400829	194.213444

$$NC \text{ (sin cte)} = 43.58349, \quad NC \text{ (con cte)} = 1021.771$$

En el modelo hay multicolinealidad aproximada preocupante del tipo no esencial (variable $\mathbf{X}_2$) y esencial (entre $\mathbf{X}_3$ y $\mathbf{X}_4$).

Ejemplo

Primer λ_2 que hace que el coeficiente de variación de $\widetilde{\mathbf{X}}_2 = \mathbf{X}_2 + \lambda_2 \mathbf{e}_2$ es mayor que 0.1002506 es 38.5 ($\mathbf{e}_2$ son los residuos de la regresión auxiliar $\mathbf{X}_2 = \alpha_1 + \alpha_3 \mathbf{X}_3 + \alpha_4 \mathbf{X}_4 + \mathbf{w}$)



Ejemplo

Variable	MCO	Alzado ($\lambda_2 = 38.5$)	Centrado
cte	70.128 (73.268)	**10.133** (1.873)	**8.561** (0.259)
$\mathbf{X}_2$	-15.399 (18.329)	-0.39 (0.464)	-15.399 (18.329)
$\mathbf{X}_3$	**-2.449** (0.618)	**-2.408** (0.616)	**-2.449** (0.618)
$\mathbf{X}_4$	**-4.608** (0.612)	**-4.646** (0.61)	**-4.608** (0.612)
R^2	0.996	0.996	0.996
$\widehat{\sigma}^2$	3.470769	3.470769	3.470769
$F_{3,96}$	8541	8541	8541

- Se mantienen todas las carecterísticas globales iniciales intactas (los residuos, número de observaciones y variables son los mismos en todos los modelos).
- Disminución importante en la varianza estimada (especialmente en el alzado).

Conclusiones y futuras líneas

Conclusiones y futuro

- Si en el modelo sólo existe multicolinealidad aproximada preocupante del tipo no esencial, el problema se resuelve centrando la variable o variables que lo producen.
- Si además existe también la del tipo esencial, el centrado de variables no la mitiga. Por tanto:
  - o bien se trata la del tipo no esencial y posteriormente la no esencial,
  - o bien se usa desde el inicio alguna técnica que mitigue ambos tipos de multicolinealidad.
- La regresión alzada mitiga tanto la multicolinealidad aproximada de tipo esencial como no esencial.

Working Paper: Salmerón, R., García, C.B. y García, J. (2021). "The Raise Regression: Justification, properties and application". arXiv: https://arxiv.org/abs/2104.14423.

Bibliografía

- Marquardt, D. W. and R. Snee (1975). “Ridge regression in practice”. The American Statistician 29 (1), pp. 3–20.
- Roldán, A.F., García, C.B. y Salmerón, R. (2020). “Analysis of the condition number in the raise regression”. Communications in Statistics - Theory and Methods. DOI: 10.1080/03610926.2020.1740737.
- Salmerón, R., A. Rodríguez-Sánchez, and C. García-García (2020). “Diagnosis and quantification of the non-essential collinearity”. Computational Statistics, 35, 647–666.
- Salmerón, R., Rodríguez, A., García, C.B. y García, J. (2020). “The VIF and MSE in Raise Regression”. Mathematics, 8 (4), 605–633.

Regresión Alzada y Multicolinealidad NO Esencial

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)
Dpto. de Economía y Empresa
Universidad de Almería

ASEPELT, Jaén 6-9 de Octubre de 2021

XXXIV Congreso Internacional de Economía Aplicada