Detección de Multicolinealidad en el Análisis de Datos en las Ciencias Sociales

R. Salmerón (romansg@ugr.es) **C.B. García** (cbgarcia@ugr.es)
Departamento de Métodos Cuantitativos para la Economía y la Empresa
Universidad de Granada
María Becerra Moreno (mariabecerra@correo.ugr.es)
Alumna de Beca de Colaboración en el Grado en Economía
Universidad de Granada

XL Congreso Nacional de Estadística e Investigación Operativa y las XIV Jornadas de Estadística Pública

Elche, 7-10 noviembre 2023

Índice

Introducción

Econometría

La 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.

Por tanto, es importante diagnósticarla y tratarla de forma adecuada.

Tipos de multicolinealidad y su detección

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).

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.

Todas las medidas (excepto coeficiente de variación) empeoran al añadir variables independientes al modelo lineal, tengan relación o no con las ya existentes.

Detección de multicolinealidad: FIV

Factor de Inflación de la Varianza (FIV)

Una de las medidas más usadas para detectar el grado de multicolinealidad existente en el modelo lineal general 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, $\mathbf{X}_{-i}$:

$$\mathbf{X}_i = \alpha_1 \mathbf{X}_1 + \cdots + \alpha_{i-1}\mathbf{X}_{i-1} + \alpha_{i+1}\mathbf{X}_{i+1} + \cdots + \alpha_p \mathbf{X}_p + \mathbf{v} = \mathbf{X}_{-i}\boldsymbol{\alpha} + \mathbf{v}.$$

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) [2]). Por tanto, no detecta la multicolinealidad no esencial.

FIV y número de variables independientes (independientes)

Se realizada una simulación en la que se generan datos para 250 variables $\mathbf{M}_i \sim N(\mu, \sigma)$ con $\mu, \sigma \in \{2, 3, 4, 5\}$ y para distintos tamaños de muestra, concretamente, se considera que $n \in \{25, 50, 75, 100, 125, 150, 175, 200\}$.
A continuación, para $\gamma \in \{0, 0.25, 0.5, 0.75, 0.8, 0.85, 0.9, 0.95\}$ fijo, se plantean de forma progresiva las siguientes regresiones:

$$\begin{aligned}
\text{(Modelo 1, } p = 3) \ \mathbf{Y} &= \beta_1 + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \mathbf{u}, \\
\text{(Modelo 2, } p = 4) \ \mathbf{Y} &= \beta_1 + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \beta_4 \mathbf{X}_4 + \mathbf{u}, \\
\text{(Modelo 3, } p = 5) \ \mathbf{Y} &= \beta_1 + \beta_2 \mathbf{X}_2 + \beta_3 \mathbf{X}_3 + \beta_4 \mathbf{X}_4 + \beta_5 \mathbf{X}_5 + \mathbf{u}, \\
&\vdots
\end{aligned}$$

donde las variables que forman la matriz $\mathbf{X}$ se generan como sigue:

$$\mathbf{X}_i = \sqrt{1 - \gamma^2} \cdot \mathbf{M}_i + \gamma \cdot \mathbf{M}_p, \ \ i = 2, \ldots, p.$$

Esta forma de simular datos para la matriz $\mathbf{X}$ ha sido previamente usada en [3], [4], [5], [6] o [7] y tiene como objetivo que la correlación entre cualesquiera dos variables independientes sea igual a γ^2.
Para cada modelo (1, 2, 3, etc) se calcula el máximo FIV. ¿Qué se observa?

FIV y valor de p (independientes $\gamma = 0$)

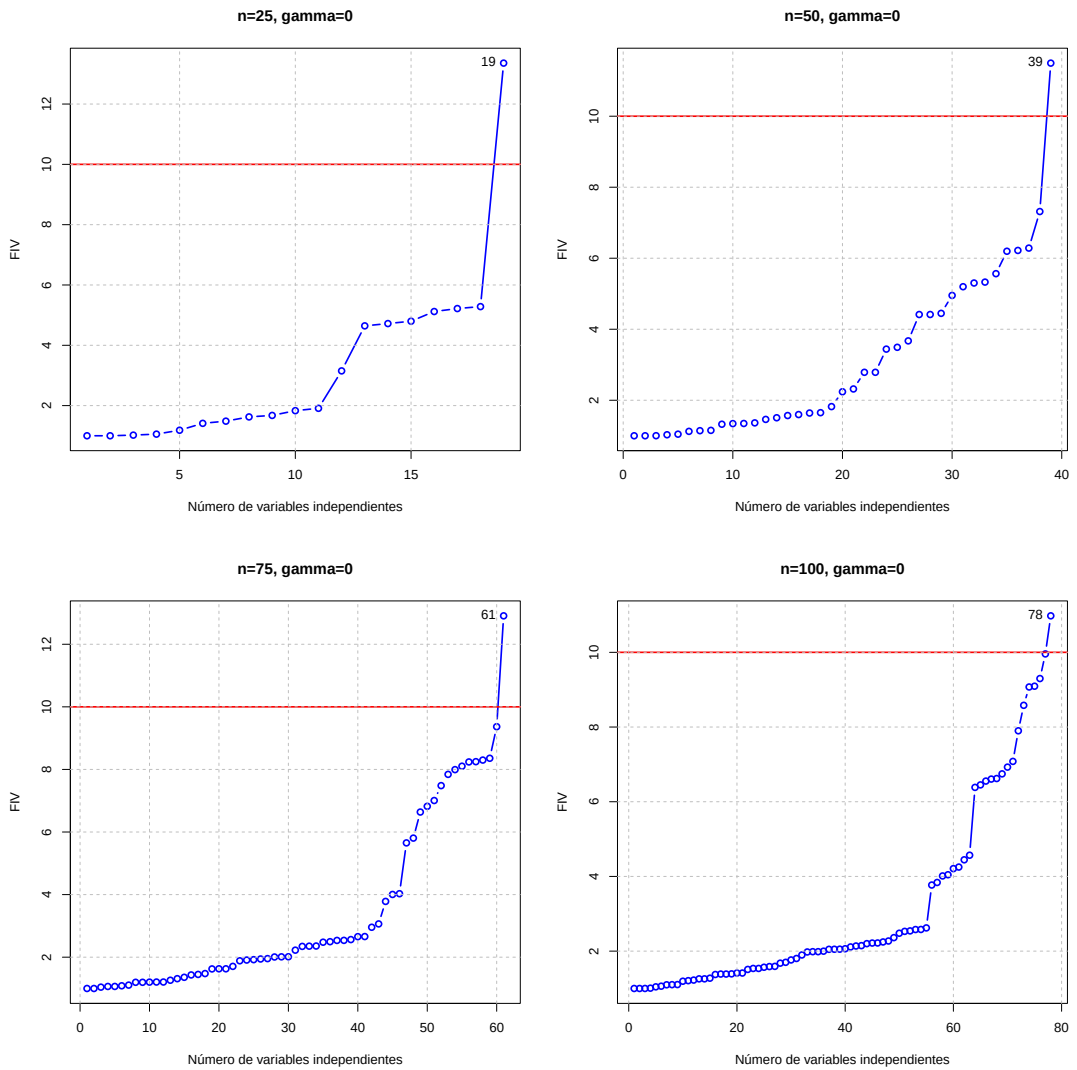

Figura : $n = 25(p = 19), 50(p = 39), 75(p = 61), 100(p = 78)$

FIV y valor de p (cierto grado lineal $\gamma = 0.5$)

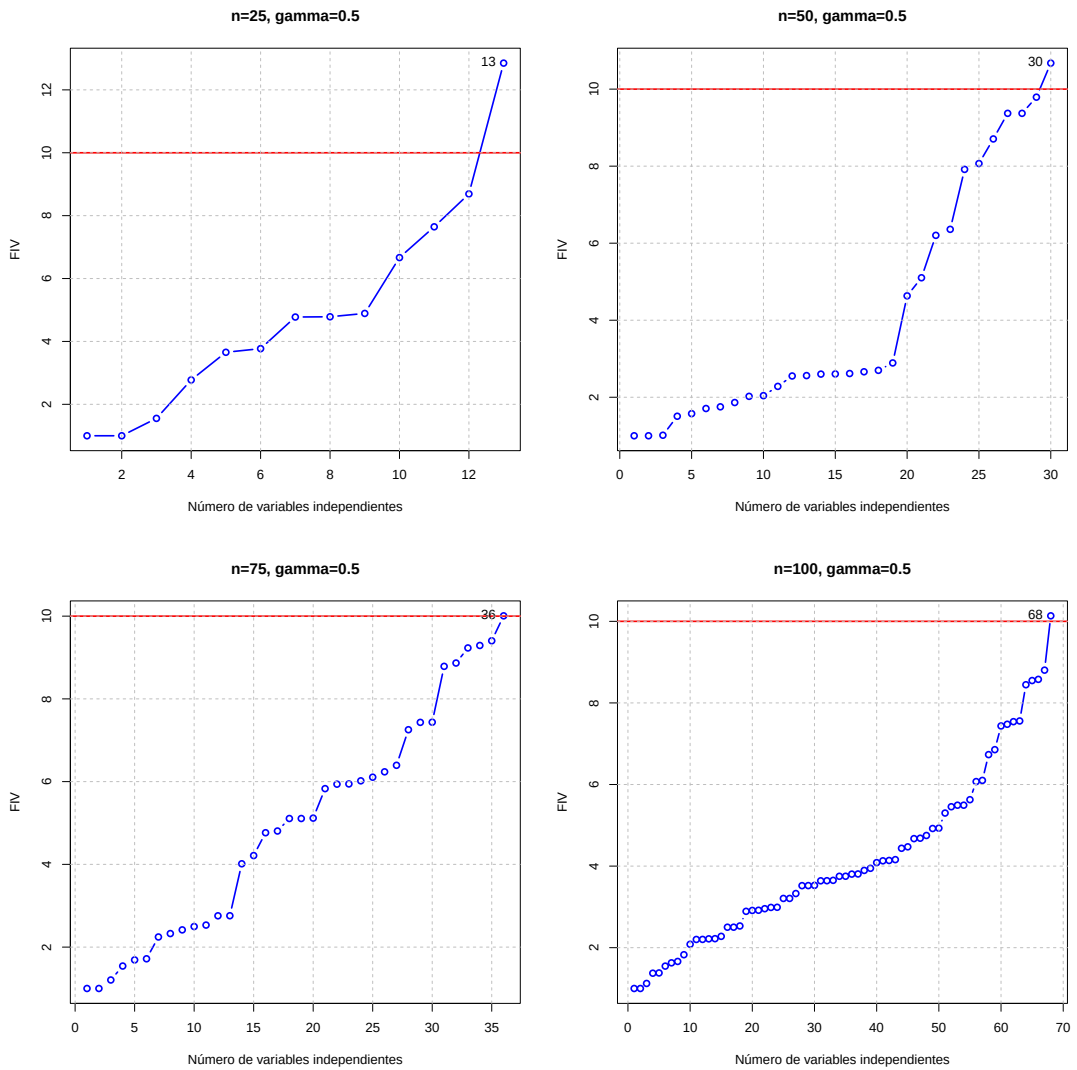

Figura : $n = 25(p = 13), 50(p = 30), 75(p = 36), 100(p = 68)$

FIV y valor de p

Valor de p que hace que el máximo FIV sea superior a 10

n - γ	0	0.25	0.5	0.75	0.8	0.85	0.9	0.95
25	19	19	13	4	4	3	3	3
50	39	39	30	18	9	5	4	3
75	61	59	36	7	7	4	3	3
100	78	79	68	19	10	4	4	3
125	105	106	98	70	49	12	7	4
150	117	116	90	13	8	8	4	3
175	142	143	132	74	28	6	5	3
200	165	161	146	40	11	6	3	3

A mayor valor de γ menor valor de p (esperable).
A mayor valor de n mayor valor de p (esperable).

Dos opciones para tener en cuenta este comportamiento

O bien se modifica el umbral de 10 (aumentándolo) o bien se corrige el FIV (disminuyéndolo).

Corrección del FIV

Propuesta

Usar en la regresión auxiliar $\mathbf{X}_i = \mathbf{X}_{-i}\boldsymbol{\alpha} + \mathbf{v}$, el coeficiente de determinación corregido:

$$\overline{R}_i^2 = 1 - \frac{n-1}{n-(p-1)} \cdot (1 - R_i^2),$$

en lugar del coeficiente de determinación, R_i^2.

Consecuencia: Factor de Inflación de la Varianza Corregido (FIVc)

$$FIVc(i) = \frac{1}{1-\overline{R}_i^2} = \frac{n-p+1}{n-1} \cdot FIV(i), \quad i = 2, \ldots, p. \tag{1}$$

Propiedades del FIVc

Propiedades del FIVc

- En el modelo de regresión lineal simple ($p = 2$): $FIVc(i) = FIV(i)$ ($FIV(i) = 1$).
- Si $p > 2$, entonces $c(n,p) = \frac{n-p+1}{n-1} < 1$. En conseciencia: $FIVc(i) < FIV(i)$ para $i = 2, \ldots, p$. ¡¡¡Precisamente lo buscado!!!
- $\lim_{n \to +\infty} FIVc(i) = FIV(i)$ para $i = 2, \ldots, p$.
- Si FIV toma su mínimo valor (ortogonalidad), esto es, $FIV(i) = 1$, para $i = 2, \ldots, p$, entonces $FIVc(i) = \frac{n-p+1}{n-1} = c(n,p)$. Por tanto, $FIVc(i) \geq c(n,p)$ para $i = 2, \ldots, p$.
- $c(n,p)$ se puede interpretar como el pocentaje de corrección que se ejecuta sobre el FIV.
- El FIVc sólo detecta la multicolinealidad aproximada de tipo esencial.

Propiedades del FIVc

Valor de $c(n,p)$ para $n \in \{25, 50, 75, 100, 125, 150, 175, 200, 225, 250\}$ y $p \in \{4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$

n-p	4	5	6	7	8	9	10	11	12	13	14	15
25	0.917	0.875	0.833	0.792	0.750	0.708	0.667	0.625	0.583	0.542	0.500	0.458
50	0.959	0.939	0.918	0.898	0.878	0.857	0.837	0.816	0.796	0.776	0.755	0.735
75	0.973	0.959	0.946	0.932	0.919	0.905	0.892	0.878	0.865	0.851	0.838	0.824
100	0.980	0.970	0.960	0.949	0.939	0.929	0.919	0.909	0.899	0.889	0.879	0.869
125	0.984	0.976	0.968	0.960	0.952	0.944	0.935	0.927	0.919	0.911	0.903	0.895
150	0.987	0.980	0.973	0.966	0.960	0.953	0.946	0.940	0.933	0.926	0.919	0.913
175	0.989	0.983	0.977	0.971	0.966	0.960	0.954	0.948	0.943	0.937	0.931	0.925
200	0.990	0.985	0.980	0.975	0.970	0.965	0.960	0.955	0.950	0.945	0.940	0.935
225	0.991	0.987	0.982	0.978	0.973	0.969	0.964	0.960	0.955	0.951	0.946	0.942
250	0.992	0.988	0.984	0.980	0.976	0.972	0.968	0.964	0.960	0.956	0.952	0.948

- Disminuye por filas: a mayor valor de p menor valor de $c(n,p)$ (esperable/deseable).
- Aumenta por columnas: a mayor valor de n mayor valor de $c(n,p)$ (esperable/deseable).

FIVc y valor de p (independientes $\gamma = 0$)

Repitiendo la simulación anterior...

FIVc en celeste

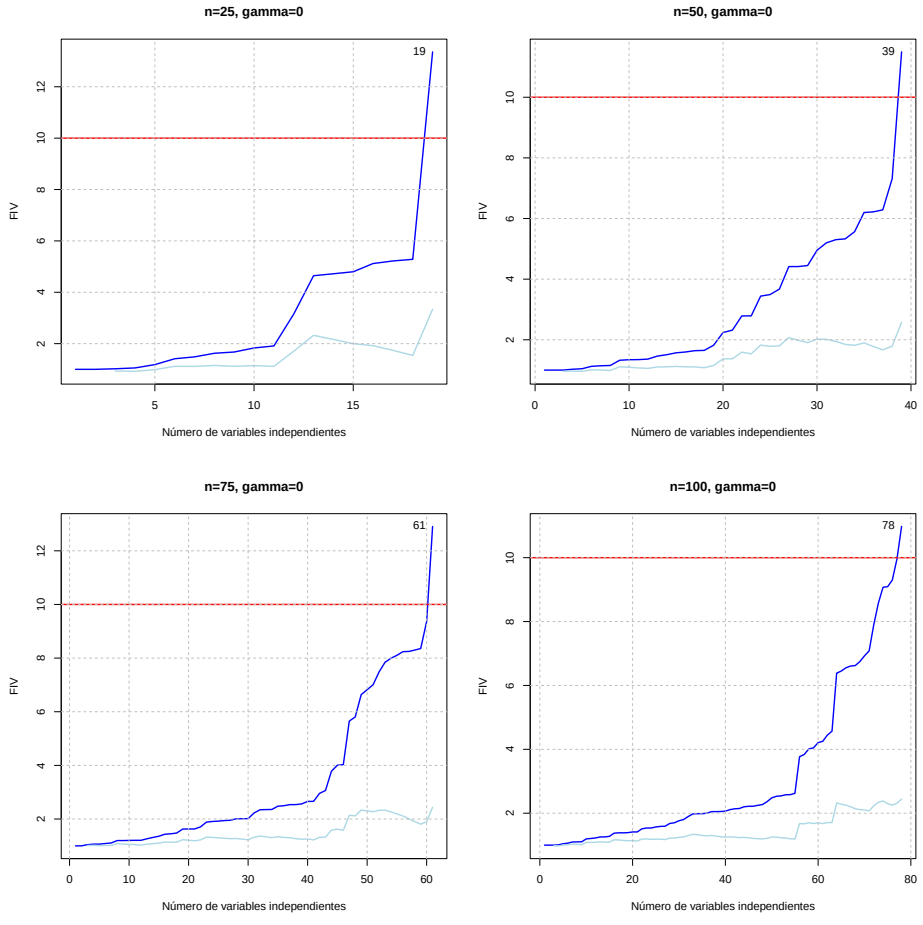

Figura : $n = 25(p = 19), 50(p = 39), 75(p = 61), 100(p = 78)$

FIVc y valor de p (cierta grado lineal $\gamma = 0.5$)

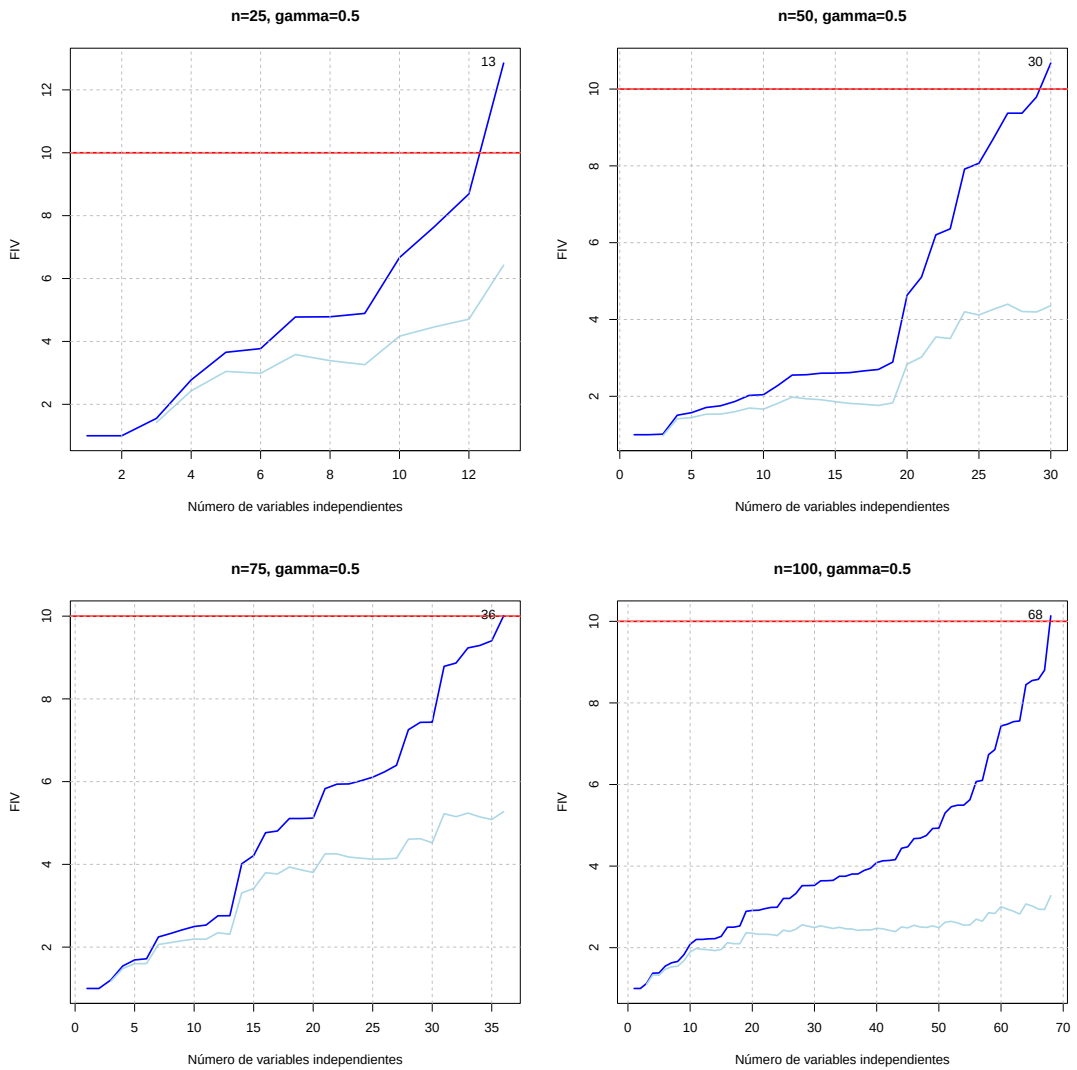

Figura : $n = 25(p = 13), 50(p = 30), 75(p = 36), 100(p = 68)$

FIVc y valor de p

Valor de p que hace que el máximo FIVc (FIV entre paréntesis) sea superior a 10

n - γ	0	0.25	0.5	0.75	0.8	0.85	0.9	0.95
25	NE (19)	NE (19)	22 (13)	4 (4)	4 (4)	3 (3)	3 (3)	3 (3)
50	NE (39)	49 (39)	NE (30)	20 (18)	11 (9)	5 (5)	4 (4)	3 (3)
75	NE (61)	70 (59)	61 (36)	9 (7)	7 (7)	4 (4)	3 (3)	3 (3)
100	99 (78)	95 (79)	95 (68)	28 (19)	10 (10)	4 (4)	4 (4)	3 (3)
125	NE (105)	NE (106)	NE (98)	NE (70)	119 (49)	17 (12)	7 (7)	4 (4)
150	147 (117)	NE (116)	136 (90)	13 (13)	8 (8)	8 (8)	4 (4)	3 (3)
175	NE (142)	174 (143)	174 (132)	156 (74)	93 (28)	7 (6)	5 (5)	3 (3)
200	195 (165)	195 (161)	193 (146)	112 (40)	12 (11)	6 (6)	3 (3)	3 (3)

NE = no existe p tal que $p < n$ y $FIVc > 10$

- Aumenta el número de variables independientes necesarias para superar el umbral de 10 establecido.
- Hay ocasiones (cuando la relación lineal es baja) en que este aumento implica que para el número de observaciones fijado, no se supera el umbral de 10 teniendo en cuenta que se ha de verificar que $n > p$.
- En otras el valor de p se aproxima mucho a n.
- Cuando $\gamma \geq 0.85$ los valores de p necesarios para superar el umbral prácticamente coinciden.

Principales resultados

Conclusiones

- Se observa que el valor del Factor de Inflación de la Varianza (FIV) aumenta conforme aumenta el número de variables independientes del modelo independientemente de si éstas están relacionadas linealmente con las ya existentes.
- Se plantea penalizar la inclusión de variables en el cáclulo del FIV usando el coeficiente de determinación corregido en lugar del coeficiente de determinación.
- Se obtiene así el Factor de Inflación de la Varianza corregido (FIVc), el cual:
  a) es menor que el FIV,
  b) disminuye conforme aumenta el número de variables independientes y
  c) aumenta conforme aumenta el tamaño de la muestra.
- Se mantiene para el FIVc el umbral de 10 usado para el FIV.
- Puesto que el FIVc es una ponderación del FIV, tampoco detecta la multicolinealidad de tipo no esencial.

Bibliografía (por orden de aparición)

- 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.
- 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.

Bibliografía (por orden de aparición) y Agradecimientos

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.

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 Multicolinealidad en el Análisis de Datos en las Ciencias Sociales

R. Salmerón (romansg@ugr.es) **C.B. García** (cbgarcia@ugr.es)
Departamento de Métodos Cuantitativos para la Economía y la Empresa
Universidad de Granada
María Becerra Moreno (mariabecerra@correo.ugr.es)
Alumna de Beca de Colaboración en el Grado en Economía
Universidad de Granada

XL Congreso Nacional de Estadística e Investigación Operativa y las XIV Jornadas de Estadística Pública

Elche, 7-10 noviembre 2023