

DIAGNÓSTICO EN MODELOS LINEALES: COLINEALIDAD

**Martín Rodríguez, J.
Díaz Ceno, M. S.
Tardáguila García, P.**

INTRODUCCIÓN

En muchas ciencias, ocurre con frecuencia que las variables consideradas en el análisis, no son independientes. Por lo tanto los riesgos al hacer estimaciones son incalculables, y un porcentaje de variaciones explicadas muy alto por el modelo de regresión puede ser perfectamente compatible con un modelo sin ningún poder predictivo. Esta problemática se conoce con el nombre de colinealidad.

Se dice que hay colinealidad cuando existe relación lineal entre las regresoras y diremos que la colinealidad está ausente cuando las regresoras son ortogonales. En el caso de colinealidad extrema, es decir, si al menos dos regresoras están perfectamente relacionadas, los coeficientes de regresión mínimo cuadráticos no están definidos.

El problema surge cuando se da una colinealidad no perfecta, ya que entonces los estimadores de los coeficientes de regresión se hacen inestables, pudiendo —incluso— aparecer con signo contrario al que cabría esperar.

SINTOMATOLOGÍA

El primer paso para poder actuar frente a la colinealidad, es tomar conciencia de su posible existencia. Hay una serie de síntomas o indicios que pueden presentarse cuando se da el problema de la colinealidad. Entre otros citaremos los siguientes:

1. El valor absoluto de la correlación empírica entre dos variables regresoras varía entre 0 y 1 (en el caso de que no exista colinealidad o que ésta sea total, respectivamente). Por ello, si al analizar la matriz de correlaciones, se detecta que un subconjunto de dichas variables está altamente correlacionado, será un síntoma a tener en cuenta.

2. Si las pruebas de nulidad de los coeficientes de regresión, conducen a eliminar del modelo variables que el investigador, basándose en su experiencia, considera relevantes.

3. Si el signo de un coeficiente de regresión es opuesto al que cabría esperar.

4. Si las varianzas de los estimadores de los coeficientes de regresión tienen valores anormalmente grandes, disminuyendo drásticamente al eliminar una o varias variables regresoras del modelo.

5. Encontrar un coeficiente de correlación múltiple entre cada regresora y las demás muy elevado.

6. Intervalos de confianza grandes para los coeficientes de regresión que representan a variables importantes en el modelo.

De todas formas, puede haber colinealidad sin que estos síntomas se hagan patentes. Solamente la diagonalización de la matriz de correlaciones y el examen de los últimos valores propios proporcionarán una información precisa.

Existen varios coeficientes que no indican el grado de colinealidad entre las variables:

El j -ésimo valor de la diagonal de $(X'X)^{-1}$ es precisamente $1/(1-R_j^2)$ siendo R_j^2 el cuadrado del coeficiente de correlación múltiple para la variable regresora X_j con el resto de las variables. A este término $1/(1-R_j^2)$ se le denomina **Factor de Inflación de la Varianza (VIF)** y es la cantidad que aumenta el error estándar del estimador j -ésimo por efecto de la correlación entre X_j y el resto de las variables regresoras.

En condiciones óptimas (ausencia de colinealidad $VIF_j = 1$ (ya que $R_j^2 = 0$)). Conforme aumente el problema de colinealidad el valor VIF se va haciendo cada vez mayor de modo que el correspondiente estimador para la j -ésima variable se va haciendo cada vez más inestable.

(THEIL, 1971). Por lo tanto, un VIF grande nos indica que el coeficiente de regresión asociado se encuentra afectado por el problema de colinealidad.

3. Además, la relación entre los valores propios nos sirve como indicador del grado de colinealidad existente en nuestros datos. De este modo, la raíz cuadrada de la razón existente entre el primer autovalor y el último (mayor y menor respectivamente):

$$K = \sqrt{\frac{\lambda_1}{\lambda_n}}$$

se denomina «Condition number», y es un índice de la inestabilidad global de los coeficientes de regresión mínimo cuadráticos (BELSLEY, KUH & WELSCH, 1980).

4. Estos mismos autores definen el «condition index»

$$K_j = \sqrt{\frac{\lambda_1}{\lambda_j}}$$

En este apartado trabajaremos sobre un estudio de simulación que nos permita poner de manifiesto cómo en presencia de colinealidad, los estimadores clásicos de Gauss-Marcov nos dan estimaciones sesgadas e inestables que no son interpretables. Asimismo, se pretende poner de manifiesto la cautela con la que debe trabajarse al utilizar los métodos de regresión paso a paso, tan profusamente utilizado por los investigadores en todos los ámbitos científicos.

MODELO ETABLECIDO «A PRIORI»

Sean X_1, X_2, X_3, X_4 variables cuyos valores son obtenidos con ayuda de un generador de números aleatorios, Y Tomamos X_4 fr manera que sea combinación lineal de otras tres; es decir

$$X_4 = 1250 + 6.5X_2 - 20.7 X_3 + \epsilon$$

La variable dependiente se elige deliberadamente según el siguiente modelo:

$$Y = 1350 - 3X_1 + 12X_2 - 20 X_3 + 15X_4 + 25X_5 - 13X_6 + \epsilon$$

SINTOMATOLOGÍA

Estimación de los coeficientes de regresión

La matriz de correlaciones $X'X$ entre las variables independientes es la que aparece a continuación (ver tabla

	1	2	3	4	5	6
1	1.000	0.057	0.130	-0.113	0.548	0.152
2	0.057	1.000	0.231	0.063	0.031	-0.264
3	0.130	0.231	1.000	-0.956	0.010	-0.238
4	-0.113	0.063	-0.956	1.000	0.004	0.165
5	0.548	0.031	0.010	0.004	1.000	-0.245
6	0.152	-0.264	-0.238	0.165	-0.245	1.000

Tabla 1. Matriz de correlaciones entre las variables

Vemos como el coeficiente de correlación entre las variables X_4 y X_3 es próximo a 1, lo cual es ya un primer indicio sobre la posible existencia de colinealidad.

Los estimadores mínimo-cuadráticos en el modelo de regresión son los de la tabla 2:

Número	Coefficiente	Error estándar	Estadístico t
Const	21789.6569		
1	-2.7580	1.8872	-1.4614
2	13488.5913	32.0335	4.0851
3	-1982.4843	102.6947	-3.8303
4	1380.3377	4.9657	0.6113
5	23.4743	1.4740	15.9233
6	-13.1239	1.7799	-8.4971

Tabla 2. Parámetros del modelo de regresión

Los errores estándar para las variables 2, 3 y 4 son muy grandes lo cual es también un síntoma de una potencial colinealidad.

Resumen del análisis

Varianza residual: 69992.3431
% de variaciones controladas 99.91 %

Obsérvese cómo a pesar de que el porcentaje de variaciones explicadas es 99.91% los valores de los estimadores de algunos de los coeficientes de regresión difieren sensiblemente de los verdaderos coeficientes (ver tabla 1). siendo incluso en alguno de los casos de signo contrario al que debería (lo que ocurre con el de la variable 4). lo cual es también un síntoma del posible problema de colinealidad.

Para hacer un efectivo diagnóstico del problemaa, deberemos conocer:

1. Si está presente una colinealidad importante («condition number»).
2. Qué coeficientes de regresión están afectados por la mismaa (factores de inflación de la varianza).
3. Qué regresoras está involucradas en la cuasi-dependenciaa (contribución de cada componente al factor de inflación).

Cálculo de los valores propios de la matriz de correlaciones.

	1	2	3	4	5	6
Valor propio	2.0922	1.3341	1.0419	0.9634	0.5682	0.0001

Tabla 3. Valores propios de la matriz de correlaciones entre las regresoras.

Vemos como el último valor propio es muy próximo a cero, lo cual nos indica ya que deberemos estar alerta por un posible problema de colinealidad, pues nos está indicando que la matriz $X'X$ es casi singular.

Estudio de los vectores propios de la matriz de correlaciones entre las regresoras

Analizaremos ahora la matriz de vectores propios de las regresoras, pues deberemos localizar cuáles son las variables con coeficientes grandes en componentes cor-

	1	2	3	4	5	6
1	0.1198	-0.2757	0.8602	-0.0337	-0.4107	0.0024
2	0.1838	0.4696	0.3232	0.6869	0.3579	-0.2033
3	0.6743	-0.1377	0.0567	0.0593	0.1697	-0.2777
4	-0.6366	0.2813	0.1517	0.1427	-0.0644	0.7600
5	0.0854	0.4813	0.3082	-0.7084	0.4054	-0.0009
6	-0.2905	-0.6113	0.1856	0.0375	0.7113	-0.0009

Tabla 4. Matriz de vectores propios para las regresoras

La tabla anterior (tablaa pone de manifiesto que las variables X_3 y X_4 son las que están implicadas en la colinealidad. (Vemos como esta afirmación coincide con la

construcción del modelo, además el siguiente coeficiente más grande se corresponde con la variable X_2).

Cálculo del «Condition Index», del «Condition number» y del VIF

El valor para el «condition number» es 135.21 lo cual evidencia la inestabilidad global de los coeficientes mínimo-cuadráticos (recordemos que se considera peligro-

	1	2	3	4	5	6
cond.index	1	1.2523	1.4171	1.4737	1.9188	135.22
VIF	1.1229	362.7	4287.1	4090.7	1.0849	1.2528

Tabla 5. «Condition index» y VIF

El valor para el index correspondiente a la variable indica una vez más que una colinealidad importante está presente.

Los VIF para las variables 2, 3 y 4 son muy grandes; valdrían 1 en el caso de ser ortogonales. Nos están indicando que, efectivamente, son los coeficientes para dichas variables los que se ven afectados por el problema de colinealidad. La misma información se obtiene estudiando el incremento en el error estándar de cada regresora.

CONCLUSIONES

Según hemos podido comprobar los estimadores mínimo cuadráticos son inestables y pierden, por tanto, su poder predictivo, poniendo de manifiesto la importancia de llevar a cabo un estudio sobre la posible colinealidad

BIBLIOGRAFÍA

- BELSLEY, D. A.; KUH, E & WELSCH, E. (1980): *Regression diagnostics: Identifying influential data and sources of collinearity*. New York, Wiley.
- CARBONELL, E Y COLS. (1983): *Regresión lineal: Un enfoque conceptual y práctico*. I.N.I.A.
- GALINDO, M. P. Y CUADRAS, C. M. (1986): *Una extensión del método Biplot y su relación con otras técnicas*. Publicaciones de Bioestadística y Biomatemática. Univ. Barcelona.
- GALINDO, M. P. (1987): *El método Biplot: una alternativa más en el diagnóstico de colinealidad*. XVI Reunión Nacional de Estadística, Investigación Operativa e Informática.
- THEIL, H. (1971) *Principles of Econometrics*. New York, Wiley.