Multicolinealidad

1 Planteamiento

Una de las hipótesis del modelo de regresión lineal múltiple establece que no existe relación lineal exacta entre los regresores, o, en otras palabras, establece que no existe *multicolinealidad perfecta* en el modelo. Esta hipótesis es necesaria para el cálculo del vector de estimadores mínimo cuadráticos, ya que en caso contrario la matriz **X'X** será no singular. La multicolinealidad perfecta no se suele presentar en la práctica, salvo que se diseñe mal el modelo como veremos en el epígrafe siguiente. En cambio, sí es frecuente que entre los regresores exista una relación aproximadamente lineal, en cuyo caso los estimadores que se obtengan serán en general poco precisos, aunque siguen conservando la propiedad de lineales, insesgados y óptimos. En otras palabras, la relación entre regresores hace que sea difícil cuantificar con precisión el efecto que cada regresor ejerce sobre el regresando, lo que determina que las varianzas de los estimadores sean elevadas. Cuando se presenta una relación aproximadamente lineal entre los regresores, se dice que existe *multicolinealidad no perfecta*. Es importante señalar que el problema de multicolinealidad, en mayor o menor grado, se plantea porque no existe información suficiente para conseguir una estimación precisa de los parámetros del modelo.

El problema de la multicolinealidad hace referencia, en concreto, a la existencia de relaciones aproximadamente lineales entre los regresores del modelo, cuando los estimadores obtenidos y la precisión de éstos se ven seriamente afectados.

Para analizar este problema, vamos a examinar la varianza de un estimador. En el modelo de regresión lineal múltiple, el estimador de la varianza de un coeficiente cualquiera – por ejemplo, de $\hat{\beta}_j$ - se puede formular de la siguiente forma:

$$\widehat{\text{var}(\hat{\beta}_j)} = \frac{\hat{\sigma}^2}{T(1-R_j^2)S_j^2} \qquad (1)$$

donde

R_j^2 es el coeficiente de determinación obtenido al efectuar la regresión de X_j sobre el resto de los regresores del modelo.

S_j^2 es la varianza muestral del regresor X_j

Como se deduce de la expresión anterior, el estimador de la varianza viene afectado por los siguientes factores:

a) Cuanto mayor es $\hat{\sigma}^2$, es decir, cuanto mayor es la dispersión de los datos en modelo ajustado, mayor será la varianza del estimador (Figura 1).

b) Al aumentar el tamaño de la muestra se reduce la varianza del estimador.

c) Cuanto menor sea la varianza muestral del regresor, es decir, cuanto menor sea la variabilidad muestral del regresor, mayor será la varianza del correspondiente coeficiente. (Figura 2)

d) Cuanto mayor sea R_j^2, cuanto mayor sea la correlación del regresor con el resto de los regresores mayor será la varianza de $\hat{\beta}_j$.

FIGURA 1. Influencia de σ^2 sobre el estimador de la varianza.

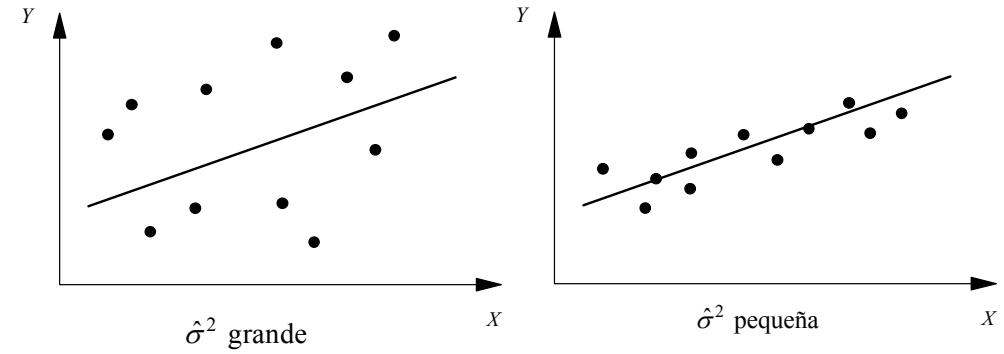
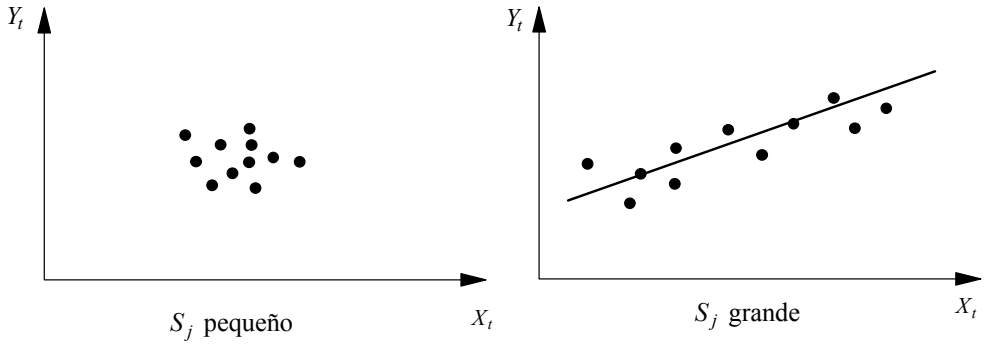


FIGURA 2. Influencia de S_j sobre el estimador de la varianza.



De los cuatro factores señalados es el factor d) el que se refiere a la multicolinealidad. Cuando se presenta multicolinealidad de una cierta gravedad, es decir cuando uno o más de los R_j^2 se aproximan a 1, se presentan los siguientes problemas al realizar inferencias con el modelo:

a) Las varianzas de los estimadores son muy grandes.

b) Se puede aceptar con frecuencia la hipótesis nula de que un parámetro es cero, aun cuando la correspondiente variable sea relevante.

c) Los coeficientes estimados serán muy sensibles ante pequeños cambios en los datos.

2 Detección

Como la multicolinealidad es un problema muestral, ya que va asociada a la configuración concreta de la matriz **X**, no existen contrastes estadísticos, propiamente dichos, que sean aplicables para su detección. En cambio, se han desarrollado numerosas reglas prácticas que tratan de determinar en qué medida la multicolinealidad afecta gravemente a la estimación y contraste de un modelo. Estas reglas no son siempre fiables, siendo en algunos casos muy discutibles. A continuación se van a exponer dos procedimientos (el factor de agrandamiento de la varianza y el número de condición) que son los que gozan de mayor soporte - especialmente el segundo - en la literatura econométrica actual.

Factor de agrandamiento de la varianza

En un modelo de regresión múltiple, si el regresor *j*-ésimo fuera ortogonal con respecto a los demás regresores (es decir, si la correlación con el resto de los regresores fuera nula), la fórmula para la varianza quedaría reducida a

$$\widehat{\text{var}(\beta_j^*)} = \frac{\hat{\sigma}^2}{TS_j^2} \tag{2}$$

El cociente entre (1) y (2) es precisamente el factor de agrandamiento de la varianza (*FAV*), cuya expresión será

$$FAV(\hat{\beta}_j) = \frac{1}{1-R_j^2} \tag{3}$$

Así, pues, el $FAV(\hat{\beta}_j)$ es la razón entre la varianza observada y la que habría sido en caso de que X_j estuviera incorrelacionada con el resto de regresores del modelo. Dicho de otra forma, el *FAV* muestra en qué medida se «agranda» la varianza del estimador como consecuencia de la no ortogonalidad de los regresores. Algunos autores consideran que existe un problema grave de multicolinealidad cuando el FAV de algún coeficiente es mayor de 10, es decir, cuando el R_j^2>0,90. En algunos programas de ordenador (el SPSS, por ejemplo) se define el término de tolerancia como la diferencia entre 1 y el R_j^2. Análogamente con el criterio aplicado a las otras medidas, se puede decir que existe un problema de multicolinealidad cuando la *tolerancia*<0,10.

El problema que tiene el *FAV* (o el R_j^2 o la tolerancia) es que no suministra ninguna información que pueda utilizarse para corregir el problema.

Número de condición

Este procedimiento de detección de la multicolinealidad es el más adecuado entre los actualmente disponibles, según afirman Judge *et at*. (1985)

Fue planteado inicialmente por Rachudel (1971) y desarrollado posteriormente por Belsley *et al*.(1980), y Belsley (1982).

El número de condición, $\kappa(X)$, es igual a la raíz cuadrada de la razón entre la raíz característica más grande ($\lambda_{\max}$) y la raíz característica más pequeña ($\lambda_{\min}$) de la matriz $\mathbf{X'X}$, es decir,

$$\kappa(X) = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \tag{4}$$

Como la matriz $\mathbf{X'X}$ es de dimensión $k \times k$ se obtienen k raíces características, pudiéndose calcular para cada una de ellas un índice de condición definido de la siguiente forma:

$$ic(\lambda_i) = \sqrt{\frac{\lambda_{\max}}{\lambda_i}} \tag{5}$$

El número de condición mide la sensibilidad de las estimaciones mínimo-cuadráticas ante pequeños cambios en los datos. De acuerdo con los estudios realizados por Belsley y otros (op. cit.), y Belsley (op. cit.), tanto con datos observados como con datos simulados, el problema de la multicolinealidad es grave cuando el número de condición toma un valor entre 20 y 30. Naturalmente, si este indicador superase el valor de 30, el problema sería ya manifiestamente grave. Estos valores vienen generalmente referidos a regresores medidos con escala de longitud unidad (es decir, con los regresores divididos por la raíz cuadrada de la suma de los valores de las observaciones), pero no centrados. Parece que no es conveniente centrar los datos (es decir, restarles sus correspondientes medias), ya que esta operación oscurece cualquier dependencia lineal que implique al término independiente.

Una información de interés para identificar el origen de la multicolinealidad es la proporción que tiene cada raíz característica en cada uno de los regresores, según veremos más adelante.

CASO 1 Analizando la multicolinealidad en el caso del absentismo laboral

En el cuadro 1 se han recogido los resultados de la regresión del programa SPSS en la que la variable endógena es ABSEN y los regresores son, aparte del término independiente o constante, factores que hipotéticamente podrían explicar el absentismo. En concreto, las variables explicativas son EDAD, ANTIGUE y SALARIO (Véase el caso 3.7 de *Econometría Aplicada*).

CUADRO 1. Resultados de la regresión en el caso 1.

	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error típ.	Beta		
(Constante)	14,4133	1,6030		8,9913	0,0000
EDAD	-0,0960	0,0478	-0,3453	-2,0060	0,0510
ANTIGUE	-0,0776	0,0672	-0,2098	-1,1544	0,2546
SALARIO	-0,0364	0,0073	-0,4661	-4,9658	0,0000
Variable dependiente: ABSEN					

Las dos primeras columnas del cuadro 1 se refieren a conceptos con los que ya estamos familiarizados: coeficientes no estandarizados beta (es decir, los obtenidos directamente al aplicar al modelo MC) y los correspondientes errores típicos (o desviaciones típicas). La tercera columna presenta los *coeficientes estandarizados beta*, cuyo significado vamos a examinar a continuación. Previamente señalaremos que la interpretación de los coeficientes no estandarizados es la misma que se dio al coeficiente de la variable explicativa en la regresión lineal simple, aunque aquí al interpretar cada coeficiente habría que añadir la expresión "manteniéndose constantes las demás variables" al tratarse de un modelo de regresión múltiple. Si a uno le preguntaran cuál es la variable explicativa que tiene mayor influencia (en valor absoluto) sobre la variable endógena podría estar tentado de responder que la EDAD, ya que su coeficiente (-0,0960) es el mayor en valor absoluto. Sin embargo, hay que tener en cuenta que el valor que toman los coeficientes viene condicionado por las escalas en que vienen medidas las variables del modelo. Los coeficientes estandarizados beta no están afectados por este problema y se calculan según la siguiente fórmula:

$$\hat{\beta}_j^{ES} = \hat{\beta}_j \frac{S_{X_j}}{S_Y} \tag{6}$$

donde S_{X_j} y S_Y son las desviaciones típicas muestrales de las variables X_j e Y respectivamente.

De acuerdo con (6) el coeficiente estandarizado $\hat{\beta}_j^{ES}$ refleja el incremento en la variable Y (medido en desviaciones típicas de Y) producido por un incremento de una desviación típica de la variable X_j.

Como puede verse en la columna de coeficientes estandarizados del cuadro 8.4 no es la EDAD sino el SALARIO la variable con mayor influencia en el absentismo. Las dos últimas columnas de este cuadro se examinarán más adelante al tratar los contrastes de significación.

Veamos ahora si la relación estimada está o no afectada por el problema de la multicolinealidad. Si en el programa SPSS se solicitan, en *Estadísticos*, los *Diagnósticos de colinealidad*, la salida del programa ofrece la información del cuadro 3, y, además, en la tabla de los coeficientes aparecen dos nuevas columnas que son las que se muestran en el cuadro 2.

En cuadro 2 se da información de la tolerancia de cada variable y del FIV (factor de inflación de la varianza) y al que nosotros hemos denominado FAV. Según estos estadísticos la multicolinealidad no parece afectar al SALARIO pero si tiene un cierto grado de importancia en las variables EDAD y ANTIGUE. En el cuadro 3 aparecen las raíces características (o autovalores) de la matriz $\mathbf{X'X}$, los índices de condición, y la participación de cada variable en las varianzas asociadas a las distintas raíces características. Como las raíces están ordenadas de mayor a menor, el índice de condición de la última fila es el número de condición, que es el que tiene mayor interés para el analista. El número de condición (15,48) es relativamente elevado aunque sin alcanzar el límite de 20 marcado por Belsley. Dejando aparte el término independiente, podemos observar que la multicolinealidad afecta a las variables EDAD y ANTIGUE, que son las que tienen mayor proporción de varianza asociada al número de condición. Podría anticiparse este resultado ya que los empleados con más años de antigüedad en la empresa serán en general también los que tengan mayor edad. Cuando el número de variables en la regresión sea elevado es conveniente analizar no solo el número de condición sino también todos los índices de condición que tomen valores elevados.

CUADRO 2. Tolerancia y FIV

	Estadísticos de colinealidad	
	Tolerancia	FIV
(Constante)		
EDAD	0,2346	4,2634
ANTIGUE	0,2104	4,7532
SALARIO	0,7891	1,2673

CUADRO 3. Diagnósticos de colinealidad

Dimensión	Autovalor	Indice de condición	Proporciones de la varianza			
			(Constante)	EDAD	ANTIGUE	SALARIO
1	3,6529	1,0000	0,0026	0,0021	0,0055	0,0051
2	0,2741	3,6509	0,0355	0,0009	0,2019	0,0296
3	0,0578	7,9492	0,0957	0,1237	0,0287	0,7295
4	0,0152	15,4807	0,8663	0,8734	0,7639	0,2358
Variable dependiente: ABSEN						

3 Soluciones

En principio, el problema de la multicolinealidad está relacionado con deficiencias en la información muestral. El diseño muestral no experimental es, a menudo, el responsable de estas deficiencias. Sin embargo, la aproximación cuantitativa a los conceptos teóricos puede ser inadecuada, haciendo que en el término de perturbación se absorban errores de especificación. Veamos a continuación algunas de las soluciones propuestas para resolver el problema de la multicolinealidad.

Eliminación de variables

La multicolinealidad puede atenuarse si se eliminan los regresores que son más afectados por la multicolinealidad. El problema que plantea esta

solución es que los estimadores del nuevo modelo serán sesgados en el caso de que el modelo original fuera el correcto. Sobre esta cuestión conviene hacer la siguiente reflexión.

El investigador está interesado en que un estimador sea preciso (es decir, que no tenga sesgo o que este sea muy pequeño) y con una varianza reducida. El error cuadrático medio (*ECM*) recoge ambos tipos de factores. Así para el estimador $\hat{\beta}_j$, el *ECM* se define de la siguiente manera:

$$ECM(\hat{\beta}_j) = \left[sesgo(\hat{\beta}_j) \right]^2 + Var(\hat{\beta}_j) \tag{7}$$

Si un regresor es eliminado del modelo, el estimador de un regresor que se mantiene (por ejemplo, $\hat{\beta}_j$) será sesgado, pero, sin embargo, su *ECM* puede ser menor que el correspondiente al modelo original, debido a que la omisión de una variable puede hacer disminuir suficientemente la varianza del estimador. En resumen, aunque la eliminación de una variable no es una práctica que en principio sea aconsejable, en ciertas circunstancias puede tener su justificación cuando contribuye a disminuir el *ECM*.

Aumento del tamaño de la muestra

Teniendo en cuenta que un cierto grado de multicolinealidad acarrea problemas cuando aumenta ostensiblemente la varianza muestral de los estimadores, las soluciones deben ir encaminadas a reducir esta varianza.

Existen dos vías: por un lado, se puede aumentar la variabilidad a lo largo de la muestra de los regresores colineales introduciendo observaciones adicionales. Esta solución no siempre es viable, puesto que los datos utilizados en las contrastaciones empíricas proceden generalmente de fuentes estadísticas diversas, interviniendo en contadas ocasiones el investigador en la recogida de información.

Por otro lado, cuando se trate de diseños experimentales, se podrá incrementar directamente la variabilidad de los regresores sin necesidad de incrementar el tamaño de la muestra.

Finalmente, conviene no olvidar que el término de perturbación no debe contener ningún factor que sea realmente relevante para la explicación de las variaciones del regresando, con el fin de reducir todo lo posible la varianza del término de perturbación.

Utilización de información extramuestral

Otra posibilidad es la utilización de información extramuestral, bien estableciendo restricciones sobre los parámetros del modelo, bien aprovechando estimadores procedentes de otros estudios.

El establecimiento de restricciones sobre los parámetros del modelo reduce el número de parámetros a estimar y, por tanto, palia las posibles deficiencias de la información muestral. En cualquier caso, para que estas

restricciones sean útiles deben estar inspiradas en el propio modelo teórico o, al menos, tener un significado económico.

En general, un inconveniente de esta forma de proceder es que el significado atribuible al estimador obtenido con datos de corte transversal es muy diferente del obtenido con datos temporales. A veces, estos estimadores pueden resultar realmente «extraños» o ajenos al objeto de estudio. Por otra parte, al estimar las varianzas de los estimadores obtenidos en la segunda regresión hay que tener en cuenta la estimación previa.

Utilización de ratios

Si en lugar del regresando y de los regresores del modelo original se utilizan *ratios* con respecto al regresor que tenga mayor colinealidad, puede hacer que la correlación entre los regresores del modelo disminuya. Una solución de este tipo resulta muy atractiva, por su sencillez de aplicación. Sin embargo, las transformaciones de las variables originales del modelo utilizando *ratios* pueden provocar otro tipo de problemas. Suponiendo admisibles las hipótesis básicas con respecto a las perturbaciones originales del modelo, esta transformación modificaría implícitamente las propiedades del modelo, de tal manera que las perturbaciones del modelo transformado utilizando *ratios* ya no serían perturbaciones homoscedásticas, sino heteroscedásticas.