Un criterio para omitir variables superfluas en modelos de regresión

Introducción

Un problema muy frecuente en la investigación aplicada es el de la presencia de multicolinealidad, es decir de una elevada correlación entre dos o más covariables. La multicolinealidad no modifica las propiedades de los estimadores. En este sentido los estimadores continúan siendo *insesgados,* o lo que es lo mismo están bien calculados, y son *óptimos,* o de otro modo son de varianza mínima entre todos los estimadores insesgados. La multicolinealidad, sin embargo, se manifiesta en un aumento de las varianzas, y por tanto de los errores estándar, de los estimadores de los parámetros. Así, puede ocurrir que los errores estándar aumenten tanto que se reduzca el valor del estadístico de significación individual (t de Student o test de Wald, por ejemplo) a valores por debajo de su nivel de significación ($p > 0,05$). De hecho puede ocurrir que una variable relevante, es decir, con estimador del parámetro asociado a la misma estadísticamente significativo, pueda no parecerlo. Como señalamos más arriba, es importante destacar que aún con elevada multicolinealidad y consecuentemente con errores estándar de los parámetros elevados, los estimadores de los parámetros continúan siendo insesgados. El problema resultará de la omisión de una variable que, aún siéndolo, no parezca relevante. Puesto que efectivamente lo es, su omisión provocará que los estimadores de los parámetros y de los errores estándar estén sesgados. Si la variable es superflua, sin embargo, podría omitirse sin mayor problema. Es más se reduciría la multicolinealidad que su presencia pudiese provocar, lo que aumentaría la eficiencia de las estimaciones. Así pues, el dilema se presenta a la hora de decidir si omitimos una variable con valor p (del estadístico de significación individual del parámetro asociado) mayor que 0,05. A priori no sabemos si la variable es superflua o, por el contrario, es relevante; pero la elevada multicolinealidad ha enmascarado su significación.

Mostramos un criterio que permita indicar si podemos o no omitir una variable con parámetro no estadísticamente significativo (véase Anexo). Únicamente es posible omitir una variable con parámetro no significativo si el valor del estadístico de significación individual, t de Student o test de Wald, es menor que la unidad (en valor absoluto en el caso de la t de Student) o equivalentemente $p > 0,3175$. Si el valor de tales estadísticos fuese mayor que uno (en valor absoluto en el caso de la t de Student) podría ocurrir que la variable correspondiente fuese relevante pero que existiese elevada multicolinealidad que enmascarase la significación. En este caso, su omisión sesgaría las estimaciones y los errores estándar, invalidando cualquier inferencia que pudiese realizarse.

Ilustración

Mostramos a continuación una ilustración con datos reales. Se pretende analizar la relación entre la mortalidad por todas las causas, excepto externas (CIE-9:001-799) y la contaminación atmosférica en la ciudad de Barcelona. Se disponen de datos diarios para el período comprendido ente 1991 y 1995. La mortalidad analizada corresponde a los residentes en Barcelona fallecidos en la ciudad. Los contaminantes de interés son humos negros y dióxido de nitrógeno, ambos en niveles promedios de 24 h y en $\mu g/m^3$. Se controlan posibles confusores de la relación, tales como la tendencia, la estacionalidad y los efectos de calendario presentes en la variable dependiente (mortalidad); variables meteorológicas (temperatura y humedad, en promedios de 24 h); y la ocurrencia de epidemias de gripe. Cuando se introducen conjuntamente ambos contaminantes como variables explicativas en la regresión, ninguno de los parámetros asociados a los mismos resulta estadísticamente significativo. En concreto el parámetro asociado a humos negros es igual a 0,005822 (error estándar igual a 0,003723) con una t de Student igual a 1,564; y el asociado a dióxido de nitrógeno igual a 0,004376 (error estándar igual a 0,003476) con una t de Student igual a 1,259. Estos resultados podrían hacer pensar que dichos contaminantes no están asociados con la mortalidad total. Lo que ocurre, sin embargo, es que la elevada correlación entre contaminantes enmascara su significación. En este sentido, cuando dichas variables se introducen separadamente en dos modelos de regresión distintos, los parámetros resultan ser estadísticamente significativos (t de Student para humos negros igual a 2,224 y para dióxido de nitrógeno igual a 2,206). Además, tanto los errores estándar como, sobretodo, los estimadores de los parámetros son muy diferentes a los del modelo de regresión que los incluye conjuntamente: humos negros

Tabla 1. Resultados de la simulación

		Porcentaje de sesgo	Error Cuadrático Medio relativo
N = 50	$r_{x1,x2} = 0,1$	0,9745	58,3490
	$r_{x1,x2} = 0,5$	0,8663	100,7298
	$r_{x1,x2} = 0,9$	0,7855	166,8892
N = 250	$r_{x1,x2} = 0,1$	0,9738	66,6024
	$r_{x1,x2} = 0,5$	0,8754	267,5742
	$r_{x1,x2} = 0,9$	0,7753	717,9064
N = 1.000	$r_{x1,x2} = 0,1$	0,9747	92,9293
	$r_{x1,x2} = 0,5$	0,8750	918,2112
	$r_{x1,x2} = 0,9$	0,7772	2701,5817

Así por ejemplo, la tabla debe leerse como sigue: en una muestra de tamaño igual a 250 observaciones y con una correlación moderada entre las variables explicativas (r = 0,5), omitir una variable explicativa aparentemente no relevante (aún siéndolo) causa que el estimador del parámetro de la variable que permanece, sea un 12,46% ((1-0,8754)*100) inferior a su verdadero valor y que el ECM del mismo sea 267,57 veces mayor que en el modelo correcto.

0,007403 (error estándar 0,003329); dióxido de nitrógeno 0,006884 (error estándar 0,003121). En definitiva se trata de variables relevantes que no pueden ser omitidas en ningún caso del modelo de regresión. De hecho, y siguiendo nuestro criterio, no podríamos omitirlas, por cuanto en ambos casos el valor del estadístico t de Student en el modelo que los incluye conjuntamente es mayor que la unidad ($p < 0,3175$).

Finalmente hemos realizado una pequeña simulación en la que consideramos un modelo de regresión con dos variables explicativas relevantes. Variamos el tamaño de la muestra (50, 250 y 1.000 observaciones) y la correlación entre las dos variables explicativas (reducida, r = 0,1; moderada, r = 0,5; y elevada colinealidad, r = 0,9). Realizamos 1.000 replicaciones en cada uno de los casos. Una de las variables explicativas es marginalmente significativa, en el sentido que cuando se introduce conjuntamente con la otra, el valor de su estadístico t de Student es menor que 1,96 ($p > 0,05$) aunque sin embargo es mayor que la unidad ($p < 0,3175$). Con-

Anexo 1

Consideremos el siguiente modelo de regresión con dos variables explicativas:

$$Y_i = \beta o + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$$

el cual puede ser expresado en desviaciones ($y_i = Y_i - \overline{Y}$; $x_{1i} = X_{1i} - \overline{X}_1$, $x_{2i} = X_{2i} - \overline{X}_2$):

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

Supongamos que existe una importante multicolinealidad, $r_{12} \approx 1$, siendo r_{12} el coeficiente de correlación simple entre las dos variables explicativas.

Representemos a los coeficientes resultantes de la aplicación de Mínimos Cuadrados Ordinarios (MCO) al modelo de regresión por β_1 y β_2, estimadores de β_1 y β_2, respectivamente. Por las propiedades del modelo MCO sabemos que estos estimadores son insesgados y que sus varianzas muestrales son:

$$\text{var} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \sigma^2 (X'X)^{-1} = \frac{\sigma^2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2} \begin{bmatrix} \Sigma_{x2}^2 & -\Sigma_{x1x2} \\ -\Sigma x_1 x_2 & \Sigma x_1^2 \end{bmatrix}$$

por lo que:

$$\text{var}\,(b1) = \frac{\sigma^2 \Sigma x_2^2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2}$$

multiplicando y dividiendo por Σx_2^2,

$$\text{var}\,(b_1) = \frac{\sigma^2}{\Sigma x_1^2 (1 - r_{12}^2)} \qquad \text{var}\,(b_2) = \frac{\sigma^2}{\Sigma x_2^2 (1 - r_{12}^2)}$$

Está claro que cuando r_{12}^2 se acerca a la unidad, es decir existe una elevada multicolinealidad, las dos varianzas muestrales aumentan considerablemente. Supongamos que este hecho provoca que, aun siendo variables explicativas relevantes, ninguno de sus estimadores sea estadísticamente significativo al 95% ($p > 0,05$).

Omitiremos incorrectamente la variable relevante x_2 (error de especificación). El modelo estimado es pues:

$$y_i = \beta_1 x_{1i} + u_i$$

El estimador de β_1 es igual a:

$$b_1^* = \frac{\Sigma y x_{1i}}{\Sigma x_{1i}^2}$$

Sustituyendo 'y' por su *verdadero* valor:

$$b_1^* = \frac{\Sigma(\beta_1 x_{1i} + \beta_{2x2i} + u_i) x_{1i}}{\Sigma x_{1i}^2} = \frac{\beta_1 \Sigma x_{1i}^2 + \beta_2 \Sigma x_{1i} x_{2i} + \Sigma x_{1i} u_i}{\Sigma x_{1i}^2} = \beta_1 + \beta_2 \frac{\Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2} + \frac{\Sigma x_{1i} u_i}{\Sigma x_{1i}^2}$$

Tomando esperanzas:

$$E(b_1^*) = \beta_1 + \beta_2 \frac{\Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2} + \frac{\Sigma x_{1i} E(u_i)}{\Sigma x_{1i}^2} = \beta_1 + \beta_2 \frac{\Sigma x_{1i} x_{2i}}{\Sigma x_{i1}^2}$$

por cuanto β son parámetros y las variables explicativas no son variables aleatorias.

Así pues, b_1^* es un estimador sesgado de β_1, a no ser que X_1 y X_2 sean ortogonales, de forma que $\Sigma x_{1i} x_{2i} = 0$.

La varianza de b_1^*, sin embargo,

$$\text{Var}\,(b_1^*) = \frac{\sigma^2}{\Sigma x_1^2}$$

es menor que la varianza de b_1 en el *verdadero* modelo. Por tanto, existe la posibilidad de elección entre sesgo y varianza, al menos teóricamente. La cuestión crucial es bajo qué condiciones b_1^* en el modelo estimado puede tener un Error Cuadrático Medio (ECM) menor que el de b_1 en el verdadero modelo, en cuyo caso podrá omitirse la variable X_2.
Como se sabe, ECM = varianza muestral + (sesgo)^2.
Así pues,

$$\text{ECM}\,(b_1^*) = \frac{\sigma^2}{\Sigma x_1^2} + \beta_2 \left(\frac{\Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2} \right)$$

y

$$\text{ECM}\,(b_1) = \frac{\sigma^2}{\Sigma x_1^2 (1 - r_{12}^2)}$$

puesto que b_1 es insesgado.
Tras algunas operaciones algebráicas:

$$\frac{\text{ECM}\,(b_1^*)}{\text{ECM}\,(b_1)} = \frac{\frac{\sigma^2}{\Sigma x_{1i}^2} + \beta_2^2 \left(\frac{\Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2} \right)^2}{\frac{\sigma^2}{\Sigma x_{1i}^2 (1 - r_{12}^2)}} = \frac{\frac{\sigma^2}{\Sigma x_{1i}^2} + \beta_2^2 \left(\frac{\Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2} \right)^2}{\frac{\sigma^2}{\Sigma x_{1i}^2 - r_{12}^2 \Sigma x_{1i}^2}} = 1 + r_{12}^2 (\tau^2 - 1)$$

en la que,

$$\tau^2 = \frac{\beta_2^2}{\sigma^2 / \Sigma x_{2i}^2 (1 - r_{12}^2)} = \frac{\beta_2^2}{\text{Var}\,(\beta_2)} = (t\beta_2)^2$$

Si $\tau^2 < 1$, ECM (b_1^*) < ECM (b_1).

Por tanto, únicamente será plausible eliminar a X_2 de la regresión si el valor *estimado de t* (en valor absoluto) es numéricamente menor que 1.

tradiciendo nuestro criterio omitimos dicha variable por cuanto aparentemente no es relevante. Mostramos el sesgo y la ineficiencia, en términos del Error Cuadrático Medio (ECM), del estimador del parámetro de la variable que permanece en la regresión. Nótese que el sesgo al omitir la variable aparentemente no relevante (siéndolo efectivamente) aumenta conforme aumenta la multicolinealidad. Pero además, la ineficiencia en la estimación (la cual aumenta al incrementarse el tamaño muestral) es muy importante incluso con correlaciones reducidas. Resumiendo, omitir una variable explicativa aparentemente no relevante (siéndolo efectivamente) sesga los parámetros y los errores estándar, aumentando además la ineficiencia. Se desprende de la simulación que no debe omitirse una variable si su estadístico de significación individual tiene un valor mayor que la unidad (en valor absoluto en el caso de la t de Student), o equivalentemente $p < 0,3175$.

M. Sáez
Departament d'Economia
Universitat de Girona
M. A. Barceló
Departament d'Economia
Universitat de Girona

Bibliografía

1. Feldstein MS. Multicolinearity and the mean square error of alternative estimation. Econometrica 1973;41:337-46.

2. Greene WH. Econometric Analysis. Englewood Clifs, New Jersey: Prentice-Hall; 1993.p.215-3.

3. Johnston J. Métodos de Econometría. Barcelona: Vicens-Vives; 1987.p.304-7.

Cambios en la incidencia de linfomas en la población de Tarragona 1984-1992

Sr. Director:

En los últimos 10 años se han publicado diversos trabajos procedentes de países industrializados en los que se describe un aumento de las tasas de incidencia de los linfomas no Hodgkin (LNH)[1]. Aunque este aumento coincide en el tiempo con la epidemia de SIDA, la contribución del VIH parece ser pequeña y se limitaría a un cierto grupo de edad[2]. La interpretación de esta tendencia temporal se ve obstaculizada por los cambios en la clasificación de los linfomas y por el progreso en las técnicas de diagnóstico que, sin duda, han modificado el reconocimiento de determinados grupos de linfomas. Utilizando datos procedentes del Registro de Cáncer de Tarragona para el período 1984-92 se estudia la tendencia temporal de las neoplasias linfoides re-organizando los códigos de la ICD-O 9590-9850 en 10 categorías adaptadas de la clasificación REAL según consenso de un grupo de patólogos europeos[3]. Para estimar las tasas de incidencia, se estimaron los denominadores para cada año utilizando información censal de los años 1981, 1986 y 1991 e interpolaciones anuales. Se presentan las tasas de incidencia ajustadas por edad tomando como grupo de referencia la estructura de edad de la población mundial[4]. Para estimar la tendencia lineal de las tasas de incidencia en el período de estudio se realizó una regresión de Poisson. Durante el período de estudio se identificaron 472 casos con el diagnóstico de linfoma. La edad promedio fue de 56,6 años. En la tabla 1 se muestran el número de casos diagnosticados en las distintas categorías y las tasas bi-anuales ajustadas por edad. A su vez se presenta el porcentaje de cambio anual de las tasas de incidencia para el período de estudio. Los linfomas aumentaron globalmente durante el período un promedio de 3,7% anual afectando tanto a hombres como mujeres. De las categorías estudiadas, el subgrupo de linfomas foliculares y el de los linfomas extranodales aumentaron significativamente. El aumento de los linfomas extranodales no se evidenció en ninguna localización particular. A lo largo de todo el período de estudio, el linfoma gástrico fue la localización más frecuente. El aumento de las tasas de incidencia de todos los linfomas afecta a todos los grupos de edad pero de forma más marcada en el grupo de edades más avanzadas. El aumento global de las neoplasias linfoides en Tarragona se puede atribuir a un crecimiento de los linfomas extranodales de células B. El aumento de un 22,2% de los linfomas foliculares debe interpretarse considerando que la tasa de incidencia es baja, siendo al inicio del período de 0,4 casos por 100.000 y al final del período de 1 por 100.000. La enfermedad de Hodgkin, al igual que en otros países industrializados se mantiene estable durante el período. Esta evolución de la incidencia es consistente con el aumento de la mortalidad para los LNH y con el descenso de mortalidad para la Enfermedad de Hodgkin observado en Tarragona durante este período[5]. La interpretación de las tendencias temporales basadas en períodos relativamente cortos tiene como inconveniente que los cambios observados pueden ser debidos a fluctuaciones naturales en la población o a artefactos debido al cambio de sistemas diagnósticos. Sin embargo, la confluencia en distintos países industrializados de la misma observación y su persistencia a lo largo de los años parece indicar