Omisión de variables en modelos de regresión con alta multicolinealidad

Correspondencia: Javier Llorca. Cátedra de Medicina Preventiva y Salud Pública.
Facultad de Medicina. Avda. Cardenal Herrera Oria s/n. 39011 Santander.
E-mail: llorcaj@medi.unican.es

Recibido: 9 de marzo de 1999
Aceptado: 9 de abril de 1999

La multicolinealidad (correlación elevada entre dos o más covariables) es un problema frecuente en los modelos de regresión. Sáez y Barceló han propuesto un criterio sencillo para decidir si una covariable debe ser eliminada del modelo, en función de cómo varía el error cuadrático medio de las otras covariables correlacionadas con ella.[1] En resumen, ellos comparan los dos siguientes modelos:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \quad (I)$$
$$y_i = \beta_1^* x_{1i} + u_i \quad (II)$$

El problema básico es que si x_1 y x_2 tienen un coeficiente de correlación lineal alto, entonces los estimadores de β_1 y β_2 tienen grandes varianzas mientras que el estimador de β_1^* es sesgado. La solución de Sáez y Barceló indica que el error cuadrático medio del estimador de β_1^* será menor que el de β_1 (y, por lo tanto, el modelo II preferible al modelo I) si el valor estimado de *t* para β_2 es menor que 1. Este sencillo resultado es sorprendente porque indica que la decisión de introducir una variable en un modelo múltiple depende sólo de la distribución de su parámetro y es independiente del sesgo introducido por β_1^*, del coeficiente de correlación entre x_1 y x_2 (salvo en lo que éste afecta a la varianza de β_2), y del error aleatorio en cada modelo.

La solución de Sáez y Barceló contiene un error: en la demostración se asume que en los modelos I y II los errores son idénticos (u_i con distribución normal de media 0 y varianza σ^2). Esta asunción es demostrablemente falsa salvo que el coeficiente de correlación entre x_1 y x_2 sea exactamente 1 (en valor absoluto), en cuyo caso la distinción entre x_1 y x_2 es irrelevante.

El problema se plantea correctamente comparando los modelos Ia y IIa:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \quad (Ia)$$
$$y_i = \beta_1^* x_{1i} + v_i \quad (IIa)$$

donde u_i y v_i se distribuyen normalmente con media 0 y varianzas respectivas σ_u^2 y σ_v^2, sobre las cuales no se realiza ninguna asunción previa. En el anexo se presenta una sencilla demostración de que la varianza del error en el modelo IIa es mayor que en el modelo Ia. Entonces, siguiendo los mismos pasos que Sáez y Barceló, se llega a la conclusión de que el error cuadrático medio del estimador de β_1^* es menor que el de β_1 sí y solo si el cuadrado de la *t* de β_2 es menor que $[1 - (1-r_{12}^2)\sigma_v^2/\sigma_u^2]/r_{12}^2$, donde r_{12} es el coeficiente de correlación entre x_1 y x_2. Este resultado es más complejo y depende de la varianza de los errores en los dos modelos y de la correlación entre las dos variables explicativas.

El problema de la multicolinealidad es más complicado que el cálculo del error cuadrático medio porque, como señalan Box et al. el concepto de modelo óptimo es más complejo que la mera reducción de las varianzas o de los errores cuadráticos.[2] Para decidir si β_2 debe mantenerse en el modelo habrá que tener en cuenta al menos los siguientes criterios: 1) la magnitud del sesgo introducido en β_1^* (por ejemplo, con el criterio de porcentaje de cambio en la estimación[3,4]), 2) el aumento de varianza residual (y, por lo tanto, el menor valor del estadístico *F*) en el modelo IIa[5], 3) la menor varianza de β_1^* respecto a β_1 (aspecto señalado por Sáez y Barceló) y 4) por supuesto, el error cuadrático medio. Es difícil que pueda encontrarse un sólo estadístico que resuma los cuatro criterios señalados.

J. Llorca
Cátedra de Medicina-Preventiva y Salud Pública.
Facultad de Medicina, Santander.

Bibliografía

1. Sáez M, Barceló MA. Un criterio para omitir variables superfluas en modelos de regresión. Gac Sanit 1998; 12: 281-3.
2. Box GEP, Hunter WG, Hunter JS. Estadística para investigadores. Barcelona: Reverte; 1998. p. 484-6.
3. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. Am J Epidemiol 1989; 129: 125-37.
4. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. Am J Epidemiol 1993; 138:923-36.
5. De la Peña Sánchez de Rivera D. Estadística, modelos y métodos, volumen 2: Modelos lineales y series temporales. Madrid: Alianza Universidad; 1989.

ANEXO. Demostración de que la varianza residual es mayor en el modelo IIa que en el Ia

La alta correlación lineal entre x1 y x2 significa que puede establecerse un modelo lineal que las relaciona:

$x_{2i} = (r_{12}\sigma_1 / \sigma_2)\, x_{1i} + w_i = (\Sigma x_{1i}x_{2i} / \Sigma x_{2i}^2)x_{1i} + w_i$ (III)

donde σ_1 y σ_2 son las desviaciones típicas de x_1 y x_2, respectivamente, y el error wi tiene distribución normal con media 0 y varianza σ_w^2.

Sustituyendo III en Ia se obtiene:

$y_i = \beta_1 x_{1i} + \beta_2 [(\Sigma x_{1i}x_{2i} / \Sigma x_{2i}^2)x_{1i} + w_i] + u_i =$

$= [\beta_1 + \beta_2 (\Sigma x_{1i}x_{2i} / \Sigma x_{2i}^2)]x_{1i} + \beta_2 w_i + u_i =$

$= \beta_1^* x_{1i} + v_i$ (IIa)

donde $\beta_1^* = \beta_1 + \beta_2 (\Sigma x_{1i}x_{2i} / \Sigma x_{2i}^2)$, por lo tanto el sesgo introducido por β_1^* es $\beta_2 (\Sigma x_{1i}x_{2i} / \Sigma x_{2i}^2)$ como señalan Sáez y Barceló; y $v_i = \beta_2 w_i + u_i$ es normal de media 0 y varianza $\sigma_v^2 = \beta_2\sigma_w^2 + \sigma_u^2 \geq \sigma_u^2$. La igualdad se produce sólo en el caso $\sigma_w^2 = 0$, es decir, cuando x_2 está completamente determinada por x_1 ($r_{12} = 1$)

Respuesta

Correspondencia: Marc Sáez. Departament d'Economia.
Universitat de Girona. Campus de Montivili. 17071 Girona.

Recibido: 21 de abril de 1999
Aceptado: 21 de abril de 1999

Hemos leído con sumo interés los comentarios de Llorca[1] a nuestra carta, en la que propusimos un criterio para omitir variables superfluas en modelos de regresión[2]. Como toda réplica que se valga, estamos en parte de acuerdo y en parte en desacuerdo con los contenidos en estos comentarios.

Como bien menciona Llorca, nuestro propósito fue el de proponer un criterio *sencillo*, en el sentido de no considerar incumplimientos o violaciones de las hipótesis básicas del modelo, lo que sin duda complicaría en exceso cualquier discusión. Es bien sabido que la elevada multicolinealidad *no* representa un incumplimiento de *ninguna* de las hipótesis básicas del modelo. En este sentido a veces se ha dicho que la multicolinealidad no es un problema sino una molestia que, por desgracia, siempre se debe soportar. Llorca, plantea modelos en los que se incumple al menos una de las principales hipótesis básicas. Nos referimos en concreto a la existencia de regresores estocásticos.

Así, en su modelo Ia:

$y_i = \beta(_1 x_{1i} + \beta_2 x_{2i} + u_i$ (Ia)

la variable x_2 es estocástica, por cuanto tal y como la formula contiene un error aleatorio:

$x_{2i} = (\Sigma x_{1i}x_{2i}/\Sigma x_{2i}^2)x_{1i} + w_i$ (III)

donde el error w_i tiene una distribución normal con media 0 y varianza σ_w^2.

Pero es que además, los errores u_i y w_i podrían correlacionados. Como consecuencia, el estimador de β_1 no *sólo* estaría sesgado en (IIa), resultado de la omisión de una variable relevante, *sino también* en (Ia), a causa del incumplimiento de la citada hipótesis básica. Así, Llorca compararía estimadores sesgados por lo que, creemos, no puede deducir ninguna consecuencia relevante.

Por otra parte, y de nuevo parafraseando a Llorca, sus comentarios *contienen un error*. En la última frase del anexo señala que 'la igualdad se produce sólo en el caso σ_w^2=0, es decir, cuando x_2 está completamente determinada por x_1 (r_{12}=1)'. Volviendo a (III) es evidente que aún siendo σ_w^2=0, el coeficiente de correlación entre ambos regresores puede ser diferente de la unidad. De hecho cuando σ_w^2=0 (y por tanto σ_v^2=σ_u^2) la variable explicativa x_2 será determinista, el caso que, precisamente, comentamos nosotros.

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

Bibliografía

1. Llorca J. Omisión de variables en modelos de regresión con alta multicolinealidad. Gac Sanit 1999;13(3):243-4.

2. Sáez M y Barceló MA. Un criterio para omitir variables superfluas en modelos de regresión. Gac Sanit 1998;12:281-3.