ERROR DE ESPECIFICACION EN MINIMOS CUADRADOS GENERALIZADO *

POTLURI RAO **

En el análisis de regresión mútilple se especifica una relación causal lineal entre una variable dependiente y un conjunto de variables independientes. La estimación de este tipo de ecuación posee la propiedad estadística deseable de mínima varianza entre los estimadores lineales e insesgados, solamente cuando los errores en el modelo estimado son serialmente independientes y homocedásticos.

Los textos, al discutir las consecuencias de los errores autocorrelacionados en las propiedades de los estimadores, simplemente "suponen" que los errores están correlacionados serialmente, en vez de explicar porqué lo están. En algunos casos, como por ejemplo los modelos de desfasajes distribuidos donde las ecuaciones a estimar son derivadas de otra ecuación, la transformación puede introducir correlación serial. En otros casos, a menudo resulta difícil racionalizar la existencia de correlación serial en los errores.

Desafortunadamente los econometristas interpretan la correlación serial en los *residuos* como causada siempre por la correlación serial en los *errores*. Es cierto que si los errores en el modelo verdadero están correlacionados serialmente los residuos exhibirán correlación serial [1], pero también es cierto que aunque los errores sean serialmente independientes, estimando un modelo mal especificado con una variable omitida que esté serialmente correlacionada puede introducir correlación serial en los residuos.

Frecuentemente se cree que sin importar qué es lo que causó la correlación serial en los residuos, utilizando información de la

* El autor agradece a los profesores Víctor J. ELIAS y Raúl P. MENTZ por sus útiles comentarios en la preparación de este trabajo.

** El autor es profesor visitante en el Instituto de Investigaciones Económicas de la Facultad de Ciencias Económicas de la Universidad Nacional de Tucumán (Argentina).

1 Ver Potluri RAO y Z. GRILICHES: "Small Sample Properties of Several Two-Stage Regression Methods in the Context of Auto-Correlated Errors", *Journal of the American Statistical Association*, 64, March 1969, 253-72.

correlación serial de los mismos necesariamente "mejora" la estimación. Este es un pensamiento erróneo. Se demostró en otra parte[2] que aún cuando la ecuación estimada es la verdadera y el error está correlacionado serialmente, utilizando una estimación del parámetro que sea diferente del parámetro verdadero desconocido, puede resultar en estimaciones que son menos eficientes que las de mínimos cuadrados simples. En este trabajo analizaremos las consecuencias de utilizar el método de mínimos cuadrados generalizado cuando la ecuación a estimar tiene un error de especificación consistente en una variable omitida[3]

Consideramos el modelo siguiente:

$$(1) \quad y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \epsilon_t$$

$$(2) \quad x_{1t} = \lambda x_{1t-1} + v_t \qquad |\lambda| < 1$$

$$(3) \quad x_{2t} = \mu x_{2t-1} + u_t \qquad |\mu| < 1$$

$$(4) \quad \epsilon_t = \rho \epsilon_{t-1} + w_t$$

Todas las variables están expresadas como desvíos de su media, por lo que el término constante está implícito. Dado que los resultados dependen crucialmente en la forma que están generadas las variables independientes, nosotros supondremos un proceso simple, que es conveniente analíticamente y que es una buena representación de muchas de las variables económicas en estudios de series de tiempo. Supondremos que los errores (u, v, w) son serialmente independientes y homocedásticos.

Para calcular el estimador de mínimos cuadrados generalizado como se lo presenta en los textos, deberíamos conocer el valor del parámetro ρ, pero en la investigación aplicada sólo podemos tener una estimación de él, digamos ρ^*. Llamaremos estimadores por mínimos cuadrados generalizado a los obtenidos al usar ρ^* en vez del verdadero valor del parámetro ρ.

[2] Potluri RAO and R. L. MILLER, *Applied Econometrics*, Wadsworth Publishing Company, Belmont, 1971, pp. 67-75.

[3] Las consecuencias de variables omitidas en el método de mínimos cuadrados simple están discutidas en detalle en: Potluri RAO: "Some notes on Misspecification in Multiple Regressions", *American Statistician*, por aparecer.

En lugar de estimar el modelo verdadero dado por la ecuación (1), el investigador estima el modelo siguiente, que está especificado erróneamente [4].

$$(5) \quad y_t = \beta_1 x_{1t} + \eta_t$$

La ecuación (5) tiene un error de especificación dado que la variable x_2 ha sido omitida. Los errores en la ecuación a estimar (η's) pueden estar correlacionados serialmente aunque los errores (ϵ's) no lo estén ($\rho = 0$).

La ecuación mal especificada puede ser estimada de dos maneras distintas: por mínimos cuadrados simple y por el generalizado utilizando un valor ρ^* como el parámetro de la correlación serial en los errores.

El procedimiento de mínimos cuadrados simple estima a β_1 como:

$$(6) \quad \hat{\beta}_1 = \frac{\Sigma x_{1t}\, y_t}{\Sigma x_{1t}^2}$$

Para obtener el estimador de mínimos cuadrados generalizado el investigador primero transforma los datos originales utilizando el valor ρ^* como:

$$(7) \quad y_t^* = y_t - \rho^* y_{t-1}$$
$$x_{1t}^* = x_{1t} - \rho^* x_{1t-1}$$

y luego calcula la regresión de y_t^* en x_t^* para obtener la estimación de mínimos cuadrados generalizado de β_1 como: [5]

$$(8) \quad \beta_1^* = \frac{\Sigma x_{1t}^*\, y_t^*}{\Sigma x_{1t}^{*2}}$$

Estamos interesados en estudiar las propiedades estadísticas de las dos estimaciones diferentes de β_1 con el objeto de poder elegir la que tenga las propiedades deseables.

Con el objeto de obviar una confusión corriente en la interpretación de las propiedades de las estimaciones que acontecen en modelos mal especificados, discutimos primeramente el significado

4 Malas especificaciones de este tipo son muy comunes en investigaciones empíricas. Se omiten variables ya sea porque el investigador no lo advierte en la ecuación o no se dispone de datos para dichas variables.

5 Por conveniencias analíticas ignoramos la primera observación correspondiente a y_1^*, x_1^*.

de la distribución de los estimadores. Realizamos un experimento en el que se selecciona un conjunto de T valores de x_1 y x_2 (generados por los procesos (2) y (3) respectivamente). Nuestro experimento consiste en varias pruebas. En todas las pruebas utilizamos el mismo conjunto de valores de x_1 y x_2, pero en cada prueba se genera un nuevo conjunto de errores (ϵ's) por el proceso (4). Los valores de y en cada prueba satisfacen la ecuación (1).

Aunque las variables independientes sean las mismas en todas las pruebas, la variable dependiente no va a ser la misma dado que extraemos un conjunto nuevo de errores en cada prueba. La estimación de los parámetros en cada prueba va a ser diferente dado que y cambia en cada prueba. Nosotros estamos interesados en conocer la sensitividad de las estimaciones con respecto a los errores. Por ello repetimos las pruebas un número "infinito" de veces, extrayendo los errores en todas las pruebas del mismo proceso (4) y vemos cómo cambia la estimación de los parámetros β. La información sobre la sensitividad de las estimaciones puede ser convenientemente resumida calculando la distribución de probabilidad de las estimaciones del parámetro obtenidas de todas las pruebas. Estamos interesados en particular en su media y su varianza.

Consideramos ahora la media de la distribución de la estimación por mínimos cuadrados simple de β_1 a partir de la ecuación errónea (5), cuando la relación verdadera está dada por la ecuación (1). Dado que la ecuación (1) es la verdadera podemos reescribir la estimación de mínimos cuadrados simple dada por la ecuación (6) como:

$$(9)\quad \hat{\beta}_1 = \frac{\Sigma x_{1t}\, y_t}{\Sigma x_{1t}^2} = \frac{\Sigma x_{1t}\,(\beta_1 x_{1t} + \beta_2 x_{2t} + \epsilon_t)}{\Sigma x_{1t}^2} =$$

$$= \beta_1 + \beta_2 \frac{\Sigma x_{1t}\, x_{2t}}{\Sigma x_{1t}^2} + \frac{\Sigma x_{1t}\, \epsilon_t}{\Sigma x_{1t}^2}$$

Por conveniencia escribimos $\dfrac{\Sigma x_{1t}\, x_{2t}}{\Sigma x_{1t}^2}$ como b_{21}, por lo que la ecuación (9) puede reescribirse como

$$(10)\quad \hat{\beta}_1 = \beta_1 + \beta_2\, b_{21} + \frac{\Sigma x_{1t}\, \epsilon_t}{\Sigma x_{1t}^2}$$

El valor medio de la distribución de $\hat{\beta}_1$ es:

$$(11)\quad E(\hat{\beta}_1) = \beta_1 + \beta_2\, b_{21}$$

dado que b_{21} es una constante y no cambia de prueba a prueba y $E(\epsilon_t) = 0$.

El estimador de mínimos cuadrados simple está sesgado. El sesgo no depende del parámetro de correlación serial de los errores en el modelo verdadero. Para comparar este resultado con el de mínimos cuadrados generalizado consideremos ahora la media de la distribución de β^*_1.

Como los valores de y en cada prueba están dados por el modelo verdadero (1), podemos reescribir la estimación de mínimos cuadrados generalizado β^*_1 de la ecuación (8) como:

$$(12)\quad \beta^*_1 = \frac{\Sigma x^*_{1t}\, y^*_t}{\Sigma x^{*2}_{1t}} = \frac{\Sigma\,(x_{1t} - \rho^* x_{1t-1})\,(y_t - \rho^* y_{t-1})}{\Sigma\,(x_{1t} - \rho^* x_{1t-1})^2} =$$

$$= \frac{\Sigma\,(x_{1t} - \rho^* x_{1t-1})\,[\beta_1\,(x_{1t} - \rho^* x_{1t-1}) + \beta_2\,(x_{2t} - \rho^* x_{2t-1}) + (\epsilon_t - \rho^* \epsilon_{t-1})\,]}{\Sigma\,(x_{1t} - \rho^* x_{1t-1})^2}$$

$$= \beta_1 + \beta_2 \frac{\Sigma\,(x_{1t} - \rho^* x_{1t-1})\,(x_{2t} - \rho^* x_{2t-1})}{\Sigma\,(x_{1t} - \rho^* x_{1t-1})^2} +$$

$$+ \frac{\Sigma\,(x_{1t} - \rho^* x_{1t-1})\,(\epsilon_t - \rho^* \epsilon_{t-1})}{\Sigma\,(x_{1t} - \rho^* x_{1t-1})^2}$$

El numerador del término medio en la expresión de arriba puede ser calculado como:

$$(13)\quad \Sigma\,(x_{1t} - \rho^* x_{1t-1})\,(x_{2t} - \rho^* x_{2t-1}) = \Sigma\, x_{1t}\, x_{2t} - \rho^* \Sigma\, x_{2t}\, x_{1t-1} -$$

$$- \rho^* \Sigma\, x_{1t}\, x_{2t-1} + \rho^{*2} \Sigma\, x_{1t-1}\, x_{2t-1}$$

$$= \Sigma\, x_{1t}\, x_{2t} - \rho^* \Sigma (\mu x_{2t-1} + u_t)\, x_{1t-1} - \rho^* \Sigma\,(\lambda x_{1t-1} + v_t)$$

$$x_{2t-1} + \rho^{*2} \Sigma x_{1t-1}\, x_{2t-1}$$

Utilizando la aproximación $\Sigma u_t\, x_{1t-1} = \Sigma v_t\, x_{2t-1} = 0$ y

$\Sigma x_{1t}\, x_{2t} = \Sigma x_{1t-1}\, x_{2t-1}$, tendremos [6]

$$(14)\quad \Sigma\,(x_{1t} - \rho^* x_{1t-1})\,(x_{2t} - \rho^* x_{2t-1}) =$$

$$= \Sigma x_{1t}\, x_{2t}\,[1 + \rho^{*2} - \rho^* \lambda - \rho^* \mu]$$

[6] Nótese que $\Sigma x_{1t}\, x_{2t}$ y $\Sigma x_{1t-1}\, x_{2t-1}$ tienen el mismo número de términos.

Igualmente el denominador del término medio de la ecuación (12) se puede expresar como:

$$(15)\quad \Sigma\,(x_{1t}-\rho^{*}\,x_{1t-1})^{2} = \Sigma x_{1t}^{2} + \rho^{*2}\,\Sigma x_{1t-1}^{2} - 2\,\rho^{*}\,\Sigma x_{1t}\,x_{1t-1} =$$
$$= \Sigma x_{1t}^{2} + \rho^{*2}\,\Sigma x_{1t-1}^{2} - 2\,\rho^{*}\,\Sigma(\lambda x_{1t-1} + v_{t})\,x_{1t-1} =$$
$$= \Sigma x_{1t}^{2}\,(1 + \rho^{*2} - 2\,\rho^{*}\lambda)$$

utilizando la aproximación $\Sigma x_{1t}^{2} = \Sigma x_{1t-1}^{2}$.

Con las expresiones (14) y (15) la ecuación (12) puede reescribirse como:

$$(16)\quad \beta^{*}_{1} = \beta_{1} + \beta_{2}\,\frac{\Sigma x_{1t}\,x_{2t}}{\Sigma x_{1t}^{2}}\cdot\frac{1+\rho^{*2}-\rho^{*}\lambda-\rho^{*}\mu}{1+\rho^{*2}-2\rho^{*}\lambda} +$$
$$+\,\frac{\Sigma\,(x_{1t}-\rho^{*}\,x_{1t-1})\;(\epsilon_{t}-\rho^{*}\,\epsilon_{t-1})}{\Sigma\,(x_{1t}-\rho^{*}\,x_{1t-1})^{2}}$$

El valor medio de la distribución de β^{*}_{1} será:

$$(17)\quad E\,(\beta^{*}_{1}) = \beta_{1} + \beta_{2}\,b_{21}\left\{\frac{1+\rho^{*2}-\rho^{*}\,\lambda-\rho^{*}\,\mu}{1+\rho^{*2}-2\,\rho^{*}\,\lambda}\right\}$$

El estimador de mínimos cuadrados generalizado también es sesgado y también su sesgo no depende del parámetro de correlación serial (ρ) de los errores verdaderos. Debe hacerse notar que ρ^{*} es un número que el investigador utiliza para computar el estimador de mínimos cuadrados generalizado y no es necesariamente igual a ρ.

Sin embargo el sesgo en el estimador de mínimos cuadrados generalizado depende de los parámetros de la correlación serial de las variables independientes (λ y μ).

Ambos estimadores son sesgados. Una comparación de las ecuaciones (11) y (17) revela la correspondencia siguiente entre los parámetros y el sesgo, para valores positivos de ρ^{*}:

$$(18)\quad \lambda = \mu \quad \text{Sesgo}\,(\hat{\beta}_{1}) = \text{Sesgo}\,(\beta^{*}_{1})$$
$$\lambda > \mu \quad |\text{Sesgo}\,(\hat{\beta}_{1})| < |\text{Sesgo}\,(\beta^{*}_{1})|$$
$$\lambda < \mu \quad |\text{Sesgo}\,(\hat{\beta}_{1})| > |\text{Sesgo}\,(\beta^{*}_{1})|$$

Esto es, el estimador de mínimos cuadrados generalizado tendrá menos sesgo que el de mínimos cuadrados simple solamente cuando

la variable omitida tenga mayor correlación serial que la variable incluida. Si la variable omitida es serialmente no correlacionada, el estimador de mínimos cuadrados generalizado tiene un mayor sesgo que el de mínimos cuadrados simple, al margen que los errores verdaderos sean o no correlacionados serialmente.

Consideremos ahora las varianzas de ambos estimadores:

$$(19)\quad V(\hat{\beta}_1) = E\left[\hat{\beta}_1 - E(\hat{\beta}_1)\right]^2 = E\left\{\frac{\Sigma x_{1t}\ \epsilon_t}{\Sigma x_{1t}^2}\right\}^2$$

y

$$(20)\quad V(\beta_1^*) = E\left[\beta_1^* - E(\beta_1^*)\right]^2 =$$

$$= E\left\{\frac{\Sigma (x_{1t} - \rho^* x_{1t-1})\ (\epsilon_t - \rho^*\ \epsilon_{t-1})}{\Sigma\ (x_{1t} - \rho^*\ x_{1t-1})^2}\right\}^2$$

Estas expresiones para las varianzas son idénticas a las obtenidas a partir del modelo:

$$(21)\qquad \begin{aligned} y_t &= \beta_1 x_{1t} + \epsilon_t \\ x_{1t} &= \lambda x_{1t-1} + v_t \\ \epsilon_t &= \rho\epsilon_{t-1} + w_t \end{aligned}$$

analizadas en otra parte[7]. Desde que la eficiencia relativa de mínimos cuadrados simple con respecto al generalizado de la ecuación (21) ha sido estudiada con detalle no reproduciremos esos resultados aquí.

Sin embargo debemos puntualizar que esas varianzas no depende de cómo está generada la variable omitida y de si la ecuación estimada es errónea o no debido a una variable omitida. La variable omitida contribuye sólo al sesgo de las distribuciones de las estimaciones.

7 Ver Potluri RAO y R. L. MILLER, op. cit.

ERROR DE ESPECIFICACION EN MINIMOS CUADRADOS GENERALIZADO

Resumen

Cuando en un modelo de regresión existe un error de especificación debido a una variable excluida, las estimaciones mediante el método de mínimos cuadrados generalizado (MCG) son sesgadas y el sesgo depende en forma crucial de la forma en que se generan las variables incluidas y excluidas, y no de la forma en que se genera el verdadero error. El sesgo en las estimaciones MCG es mayor que en las obtenidas por mínimos cuadrados simple, excepto cuando la variable excluída tiene una autocorrelación mayor que la variable incluída. La eficiencia relativa de MCG con respecto a mínimos cuadrados simple no depende de la variable omitida.

SPECIFICATION BIAS IN THE GENERALIZED LEAST SQUARES

Summary

When a linear regression model is misspecified by a left out variable the Generalized Least Squares (GLS) estimates are biased, and the bias depends crucially on how the excluded and the included variables are generated and not on how the true error terms are generated. The bias in GLS estimates is larger than that of ordinary least squares estimates except when the left out variable has higher autocorrelation than the included variable. The relative efficiency of GLS with respect to ordinary Least Squares does not depend on the left out variable.