CAPíTULO 5

Errores de especificación

Estrictamente hablando, un error de especificación es el incumplimiento de cualquiera de los supuestos básicos del modelo lineal general. En un sentido más laxo, esta expresión hace referencia al supuesto implícito de que la matriz de diseño $\mathbf{X}$ está correctamente especificada; dicho de otro modo, que las variables explicativas incluidas en la matriz $\mathbf{X}$ son las verdaderamente relevantes en la explicación de $\mathbf{y}$. En una aplicación práctica, al diseñar la matriz $\mathbf{X}$ podemos cometer dos tipos de errores especificación: omisión de variables relevantes (infraespecificación) e inclusión de variables irrelevantes (sobreespecificación). En este capítulo, se estudian las consecuencias que se derivan de estos dos errores de especificación sobre las propiedades estadísticas del estimador de mínimos cuadrados ordinarios.

5.1. Omisión de variables relevantes

Supongamos que el modelo correcto para explicar la variable dependiente $\mathbf{y}$ es, en forma particionada,

$$\begin{array}{ccccccc} \mathbf{y} & = & \mathbf{X}_r\boldsymbol{\beta}_r & + & \mathbf{X}_s\boldsymbol{\beta}_s & + & \mathbf{u} \\ (n\times 1) & & (n\times r)(r\times 1) & & (n\times s)(s\times 1) & & (n\times 1) \end{array}$$

pero por error especificamos el modelo incorrecto

$$\mathbf{y} = \mathbf{X}_r\boldsymbol{\beta}_r + \epsilon$$

en donde hemos omitido las variables $\mathbf{X}_s$. Estas variables engrosarán la lista de factores omitidos que recoge el nuevo término de error, $\epsilon = \mathbf{X}_s\boldsymbol{\beta}_s + \mathbf{u}$, cuya esperanza claramente es distinta de un vector de ceros, $E(\epsilon) = \mathbf{X}_s\boldsymbol{\beta}_s$.

Proposición 51. *El estimador de mínimos cuadrados $\hat{\mathbf{b}}_r$ de $\boldsymbol{\beta}_r$ en el modelo incorrecto es, en general, un estimador sesgado, siendo el sesgo $B(\hat{\mathbf{b}}_r) = (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{X}_s\boldsymbol{\beta}_s$.*

Demostración. El estimador de mínimos cuadrados en el modelo incorrecto es

$$\hat{\mathbf{b}}_r = (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r\mathbf{y}$$

Para estudiar las propiedades estadísticas de este estimador, reemplazamos $\mathbf{y}$ por su expresión en el modelo correcto. Así,

$$\hat{\mathbf{b}}_r = (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r[\mathbf{X}_r\boldsymbol{\beta}_r + \mathbf{X}_s\boldsymbol{\beta}_s + \mathbf{u}] = \boldsymbol{\beta}_r + (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{X}_s\boldsymbol{\beta}_s + (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{u}$$

Tomando esperanza matemática

$$E(\hat{\mathbf{b}}_r) = \boldsymbol{\beta}_r + (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{X}_s\boldsymbol{\beta}_s$$

vemos que el estimador de $\boldsymbol{\beta}_r$ en el modelo incorrecto es sesgado, siendo el sesgo (*bias*, en inglés)

$$B(\hat{\mathbf{b}}_r) = (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{X}_s\boldsymbol{\beta}_s$$

□

DEFINICIÓN 43. *El sesgo causado por la omisión de variables relevantes se denomina sesgo de especificación, y recoge la influencia de* $\mathbf{X}_s$ *sobre* $\mathbf{y}$ *que estamos atribuyendo a* $\mathbf{X}_r$.

PROPOSICIÓN 52. *El estimador de* $\boldsymbol{\beta}_r$ *en el modelo incorrecto será insesgado cuando las variables omitidas y las variables incluidas sean ortogonales.*

DEMOSTRACIÓN. Las columnas de la matriz $(\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{X}_s$ de orden $r\times s$ contienen las pendientes estimadas en la regresión de cada variable explicativa omitida sobre las variables explicativas incluidas. Estas pendientes serán iguales a cero únicamente en el caso especial en que las matrices $\mathbf{X}_s$ y $\mathbf{X}_r$ sean ortogonales, $\mathbf{X}_r'\mathbf{X}_s = \mathbf{0}$. □

Observación 31. *El estimador de* $\boldsymbol{\beta}_r$ *en el modelo incorrecto también será insesgado cuando* $\boldsymbol{\beta}_s = \mathbf{0}_s$*, es decir, cuando no se comete ningún error de especificación.*

PROPOSICIÓN 53. *El estimador de* $\boldsymbol{\beta}_r$ *en el modelo incorrecto es más "eficiente"que el estimador de mínimos cuadrados de* $\boldsymbol{\beta}_r$ *en el modelo correcto.*

DEMOSTRACIÓN. Vamos a demostrar que la diferencia entre las matrices de varianzas y covarianzas de estos dos estimadores de $\boldsymbol{\beta}_r$ es una matriz semidefinida negativa. En el modelo incorrecto

$$\begin{aligned} V(\hat{\mathbf{b}}_r) =& E\left\{[\hat{\mathbf{b}}_r - E(\hat{\mathbf{b}}_r)][\hat{\mathbf{b}}_r - E(\hat{\mathbf{b}}_r)]'\right\} = E\left[(\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{u}\mathbf{u}'\mathbf{X}_r(\mathbf{X}_r'\mathbf{X}_r)^{-1}\right] \\ =& (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r' \underbrace{E\left[\mathbf{u}\mathbf{u}'\right]}_{\sigma_u^2\mathbf{I}_n} \mathbf{X}_r(\mathbf{X}_r'\mathbf{X}_r)^{-1} = \sigma_u^2(\mathbf{X}_r'\mathbf{X}_r)^{-1} \end{aligned}$$

mientras que en el modelo correcto

$$V(\hat{\boldsymbol{\beta}}_r) = \sigma_u^2(\mathbf{X}_r'\mathbf{M}_s\mathbf{X}_r)^{-1}$$

En lugar de comparar estas dos matices, comparamos las inversas

$$V(\hat{\mathbf{b}}_r)^{-1} - V(\hat{\boldsymbol{\beta}}_r)^{-1} = \sigma_u^{-2}\mathbf{X}_r'\mathbf{X}_r - \sigma_u^{-2}\mathbf{X}_r'\mathbf{M}_s\mathbf{X}_r = \sigma_u^{-2}\mathbf{X}_r'[\mathbf{I}_n - \mathbf{M}_s]\mathbf{X}_r = \sigma_u^{-2}\mathbf{X}_r'[\mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s']\mathbf{X}_r$$

Definiendo $\mathbf{C} = \mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{X}_r$, obtenemos la matriz semidefinida positiva

$$V(\hat{\mathbf{b}}_r)^{-1} - V(\hat{\boldsymbol{\beta}}_r)^{-1} = \sigma_u^{-2}\mathbf{C}'\mathbf{C}$$

que será una matriz nula cuando $\mathbf{X}_s'\mathbf{X}_r = \mathbf{0}$. □

PROPOSICIÓN 54. *El estimador de* σ_u^2 *en el modelo incorrecto*

$$\hat{\sigma}_\epsilon = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n-r} = \frac{\mathbf{y}'\mathbf{M}_r\mathbf{y}}{n-r}$$

es sesgado, siendo el sesgo no negativo.

DEMOSTRACIÓN. La suma de cuadrados de los residuos en el modelo incorrecto, $\mathbf{y}'\mathbf{M}_r\mathbf{y}$, puede escribirse como

$$\mathbf{y}'\mathbf{M}_r\mathbf{y} = [\mathbf{X}_r\boldsymbol{\beta}_r + \mathbf{X}_s\boldsymbol{\beta}_s + \mathbf{u}]'\mathbf{M}_r[\mathbf{X}_r\boldsymbol{\beta}_r + \mathbf{X}_s\boldsymbol{\beta}_s + \mathbf{u}]$$

Material publicado bajo licencia Creative Commons

en donde se ha reemplazando $\mathbf{y}$ por su expresión en el modelo correcto. Como $\mathbf{M}_r\mathbf{X}_r = 0$, tenemos

$$\mathbf{y}'\mathbf{M}_r\mathbf{y} = [\mathbf{X}_s\boldsymbol{\beta}_s + \mathbf{u}]'\mathbf{M}_r[\mathbf{X}_s\boldsymbol{\beta}_s + \mathbf{u}] = \boldsymbol{\beta}_s'\mathbf{X}_s'\mathbf{M}_r\mathbf{X}_s\boldsymbol{\beta}_s + \mathbf{u}'\mathbf{M}_r\mathbf{u}_s + 2\boldsymbol{\beta}_s'\mathbf{X}_s'\mathbf{M}_r\mathbf{u}$$

en donde hemos usado el resultado $\mathbf{u}'\mathbf{M}_r\mathbf{X}_s\boldsymbol{\beta}_s = \boldsymbol{\beta}_s'\mathbf{X}_s'\mathbf{M}_r\mathbf{u}$.

Tomando ahora esperanza matemática

$$E(\mathbf{y}'\mathbf{M}_r\mathbf{y}) = \boldsymbol{\beta}_s'\mathbf{X}_s'\mathbf{M}_r\mathbf{X}_s\boldsymbol{\beta}_s + E(\mathbf{u}'\mathbf{M}_r\mathbf{u}) + 2E(\boldsymbol{\beta}_s'\mathbf{X}_s'\mathbf{M}_r\mathbf{u}) = \boldsymbol{\beta}_s'\mathbf{X}_s'\mathbf{M}_r\mathbf{X}_s\boldsymbol{\beta}_s + (n-r)\sigma_u^2$$

De aquí,

$$E(\hat{\sigma}_\epsilon^2) = \frac{E(\mathbf{y}'\mathbf{M}_r\mathbf{y})}{n-r} = \sigma_u^2 + \frac{1}{n-r}\boldsymbol{\beta}_s'\mathbf{X}_s'\mathbf{M}_r\mathbf{X}_s\boldsymbol{\beta}_s$$

El sesgo $\boldsymbol{\beta}_s'\mathbf{X}_s'\mathbf{M}_r\mathbf{X}_s\boldsymbol{\beta}_s/(n-r) \geq 0$ es una forma cuadrática semidefinica porque la matriz $\mathbf{X}_s'\mathbf{M}_r\mathbf{X}_s$ es del tipo $\mathbf{C}'\mathbf{C}$ con $\mathbf{C} = \mathbf{M}_r\mathbf{X}_s$. □

Observación 32. *Cuando las variables en $\mathbf{X}_r$ y $\mathbf{X}_s$ son ortogonales, $\mathbf{X}_r'\mathbf{X}_s = \mathbf{0}$, el sesgo de $\hat{\sigma}_\epsilon^2$ es $\boldsymbol{\beta}_s'\mathbf{X}_s'\mathbf{X}_s\boldsymbol{\beta}_s/(n-r)$ siendo $\mathbf{X}_s'\mathbf{X}_s$ una matriz semidefinida positiva, si hay multicolinealidad exacta entre las variables explicativas omitidas, o definida positiva, si no la hay.*

La solución al problema de la omisión de variable relevante parece simple: incluirlas en el modelo. Sin embargo, cuando especificamos un modelo econométrico seguimos las indicaciones de la teoría económica y de nuestro sentido común, y no somos conscientes de que nos estamos olvidando de ciertas variables. Para empeorar las cosas, la omisión de variables relevantes puede confundirse con otras violaciones de los supuestos básicos, que requieren tratamientos muy diferentes. Si, por ejemplo, las variables omitidas están correlacionadas con las incluidas, entonces $\mathbf{X}_r$ y $E(\epsilon) = \mathbf{X}_s\boldsymbol{\beta}_s$ también están correlacionados y no podemos mantener el supuesto de regresores fijo. Por otra parte, con datos de series temporales, la perturbación $E(\epsilon)$ presentará autocorrelación. Estudiaremos estos errores de especificación en capítulos posteriores.

5.2. Inclusión de variables irrelevantes

Suponemos, ahora, que el modelo correcto es

$$\mathbf{y} = \mathbf{X}_r\boldsymbol{\beta}_r + \mathbf{u}$$

pero incluimos erróneamente un conjunto de variables explicativas irrelevantes

$$\mathbf{y} = \mathbf{X}_r\boldsymbol{\beta}_r + \mathbf{X}_s\boldsymbol{\beta}_s + \boldsymbol{\epsilon}$$

Vamos a estudiar las propiedades estadísticas de los estimadores de $\boldsymbol{\beta}_r$ y $\boldsymbol{\beta}_s$ en el modelo incorrecto o sobreespecificado.

Proposición 55. *Los estimadores de los parámetros asociados a variables relevantes son insesgados, mientras que los estimadores de los parámetros asociados a variables irrelevantes tienen un valor esperado nulo.*

Demostración. Los estimadores de mínimos cuadrados de $\boldsymbol{\beta}_r$ y $\boldsymbol{\beta}_s$ son

$$\hat{\mathbf{b}}_r = (\mathbf{X}_r'\mathbf{M}_s\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{M}_s\mathbf{y} \quad \text{y} \quad \hat{\mathbf{b}}_s = (\mathbf{X}_s'\mathbf{M}_r\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{M}_r\mathbf{y}$$

Material publicado bajo licencia Creative Commons

y pueden escribirse, reemplazando $\mathbf{y}$ por su expresión en el modelo correcto, como

$$\hat{\mathbf{b}}_r = (\mathbf{X}_r'\mathbf{M}_s\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{M}_s[\mathbf{X}_r\boldsymbol{\beta}_r + \mathbf{u}] = \boldsymbol{\beta}_r + (\mathbf{X}_r'\mathbf{M}_s\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{M}_s\mathbf{u}$$
$$\hat{\mathbf{b}}_s = (\mathbf{X}_s'\mathbf{M}_r\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{M}_r[\mathbf{X}_r\boldsymbol{\beta}_r + \mathbf{u}] = (\mathbf{X}_s'\mathbf{M}_r\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{M}_r\mathbf{u}$$

Tomando esperanza matemática

$$E(\hat{\mathbf{b}}_r) = \boldsymbol{\beta}_r + (\mathbf{X}_r'\mathbf{M}_s\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{M}_sE(\mathbf{u}) = \boldsymbol{\beta}_r$$
$$E(\hat{\mathbf{b}}_s) = (\mathbf{X}_s'\mathbf{M}_r\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{M}_rE(\mathbf{u}) = \mathbf{0}$$

□

PROPOSICIÓN 56. *El estimador de $\boldsymbol{\beta}_r$ en el modelo sobreespecificado es menos eficiente que el estimador de $\boldsymbol{\beta}_r$ en el modelo correcto.*

DEMOSTRACIÓN. Ya hemos demostrado que al añadir nuevas variables a un modelo de regresión la eficiencia de los estimadores puede disminuir, pero nunca aumentar. □

PROPOSICIÓN 57. *El estimador de mínimos cuadrados de la varianza del error σ_u^2 en el modelo sobreespecificado es insesgado.*

DEMOSTRACIÓN. En el modelo incorrecto

$$\hat{\sigma}_\epsilon = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n-k} = \frac{\mathbf{y}'\mathbf{M}\mathbf{y}}{n-k}$$

La suma de cuadrados de los residuos $\mathbf{y}'\mathbf{M}\mathbf{y}$ puede escribirse como

$$\mathbf{y}'\mathbf{M}\mathbf{y} = [\mathbf{X}_r\boldsymbol{\beta}_r + \mathbf{u}]'\mathbf{M}[\mathbf{X}_r\boldsymbol{\beta}_r + \mathbf{u}]$$

en donde se ha reemplazando $\mathbf{y}$ por su expresión en el modelo correcto. Como $\mathbf{M}\mathbf{X} = \mathbf{M}[\mathbf{X}_r|\mathbf{X}_s] = \mathbf{0}$, tenemos

$$\mathbf{y}'\mathbf{M}\mathbf{y} = \mathbf{u}'\mathbf{M}\mathbf{u}$$

Tomando esperanza matemática

$$E(\mathbf{u}'\mathbf{M}\mathbf{u}) = (n-k)\sigma_u^2$$

De aquí,

$$E(\hat{\sigma}_\epsilon^2) = \frac{E(\mathbf{y}'\mathbf{M}\mathbf{y})}{n-k} = \sigma_u^2$$

□

Comparando las dos tipos de errores de especificación vemos que la consecuencia final es la misma: los estadísticos t y F están distorsionados. En el modelo sobreespecificado la distorsión proviene de la sobreestimación de las varianzas. Si el estimador es consistente, esta varianza tiende a cero cuando el tamaño muestral tiende a infinito. Parece, por tanto, razonable pensar que la distorsión en varianza será mayor en muestras pequeñas y menor en muestras muy grandes. En el caso del modelo infraespecificado la distorsión además proviene del sesgo del estimador, que depende de la relación entre variables incluidas y omitidas. Esta relación no tiene porqué disminuir con el tamaño muestral. A modo de conclusión, si el tamaño muestral es suficientemente grande, parece menos grave cometer errores de especificación por exceso que por defecto.

Material publicado bajo licencia Creative Commons

5.3. Forma funcional incorrecta

Imaginemos que la relación estadística entre dos variables Y_t y X_t viene dada por la regresión polinomial

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \cdots + \beta_k X_t^k + u_t$$

pero por error especificamos la regresión lineal simple

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$$

En este modelo, la omisión de las $k-1$ variables relevantes puede interpretarse como un error en la especificación de la forma función: estamos especificando una relación lineal entre Y_t y X_t, cuando la relación entre ambas es no lineal

Ramsey propuso un contraste muy simple para detectar este error de especificación, el test RESET (del inglés, REgression Epecification Error Test). Los pasos son los siguientes:

1. Estimar la regresión simple de Y_t sobre X_t, calcular el coeficiente de determinación, digamos R_0^2, y generar los valores ajustados $\hat{Y}_t$.
2. Estimar la regresión de Y_t sobre 1, X_t, $\hat{Y}_t^2$, ..., Y_t^r y calcular el coeficiente de determinación, R_1^2.
3. Calcular el estadístico F
$$F = \frac{(R_1^2 - R_0^2)/(k-1)}{(1-R_1^2)/(n-k-1)}$$
4. El modelo está erróneamente especificado si $F > c$, en donde c es el valor crítco tal que $Prob(F_{k-1,n-k-1} > c) = \alpha$.

Resumen

1. Las consecuencias de la omisión de variables relevantes son:
   a) los estimadores de los parámetros asociados a las variables incluidas son sesgados,
   b) la matriz de varianzas y covarianzas de estos estimadores es "menor"que la que se obtendría en el modelo correcto,
   c) la varianza de las perturbaciones está sobreestimada,
   d) los contrastes de hipótesis basados en los estadísticos t y F no son válidos.

   La inclusión de variables irrelevantes tiene las siguientes consecuencias:
   a) los estimadores de los parámetros asociados a las variables relevantes son insesgados, mientras que los de las variables irrelevantes tienen un valor esperado nulo,
   b) la matriz de varianzas y covarianzas de los estimadores parámetros relevantes en el modelo sobreespecificado es "mayor"que la correspondiente en el modelo verdadero,
   c) la varianza del error es insesgada,
   d) los contrastes de hipótesis basados en los estadísticos t y F están sesgados hacia la no significación.

Apuntes de Econometría. LADE y LE. Curso 2008-2009.
Material publicado bajo licencia Creative Commons

Palabras clave

Matriz de diseño
Error de especificación
Omisión de variable relevante
Inclusión de variable irrelevante
Infraespecificación
Sobreespecificación
Sesgo de especificación
Forma funcional incorrecta
Test RESET

Ejercicios

1. Discuta la siguiente proposición: cuando el tamaño muestral es muy grande, es menos grave la inclusión de variables irrelevantes que la omisión de variables relevantes.
2. Obtenga el sesgo del estimador de la varianza residual en un modelo de regresión con omisión de variables relevantes.
3. Sea la ecuación de demanda

$$q_i = \alpha + \beta p_i + \gamma r_i + u_i, \quad i = 1, \ldots, N$$

donde las variables explicativas precio p_i y renta r_i se consideran no estocásticas y el término de error u_i cumple las propiedades siguientes

$$E(u_i) = 0, \quad E(u_i^2) = \sigma^2, \quad E(u_i u_j) = 0 \, \forall i \neq j.$$

a) Demuestre que si esta ecuación se estima sin la variable relevante renta, entonces la esperanza matemática del estimador MCO del parámetro β es

$$E(\hat{\beta}) = \beta + \gamma \frac{\sum_{i=1}^{N} \tilde{p}_i \tilde{r}_i}{\sum_{i=1}^{N} \tilde{p}_i^2}$$

donde $\tilde{p}_i = p_i - \bar{p}$ y $\tilde{r}_i = r_i - \bar{r}$.

b) Demuestre que si esta ecuación se estima con la variable relevante renta, entonces

$$V(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{N} \tilde{p}_i^2 (1 - \rho_{pr}^2)}$$

donde ρ_{pr} es el coeficiente de correlación simple entre las variables precio y renta.

Apuntes de Econometría. LADE y LE. Curso 2008-2009.
Material publicado bajo licencia Creative Commons