Capítulo 5

Especificación inadecuada del modelo

5.1. Introducción.

En lo que antecede hemos dado por supuesto que el modelo lineal que se estima es el "correcto", es decir, que la variable aleatoria Y efectivamente se genera de la siguiente manera:

$$Y = \beta_0 X_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \epsilon. \tag{5.1}$$

En la práctica, sin embargo, no tenemos un conocimiento preciso del mecanismo que genera las Y's. Tenemos, todo lo más, una lista de variables susceptibles de formar parte de la ecuación (6.1) en condición de regresores.

De ordinario, por ello, incurriremos en errores en la especificación, que pueden ser de dos naturalezas:

1. Incluir en (6.1) regresores irrelevantes.

2. Omitir en (6.1) regresores que hubieran debido ser incluidos.

Estudiamos en lo que sigue el efecto de estos dos tipos de mala especificación.

5.2. Inclusión de regresores irrelevantes.

Supongamos que

$$\vec{Y} \quad = \quad X\vec{\beta} + \vec{\epsilon} \tag{5.2}$$

pese a lo cual decidimos estimar el modelo

$$\vec{Y} \quad = \quad X\vec{\beta} + Z\vec{\gamma} + \vec{\epsilon} \tag{5.3}$$

¿Qué ocurre con los estimadores de los parámetros $\vec{\beta}$?

Al estimar el modelo sobreparametrizado (6.3) obtendríamos:

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} X' \\ Z' \end{pmatrix} \vec{Y} \tag{5.4}$$

En el caso particular de columnas Z ortogonales a las columnas en X, los estimadores de $\vec{\beta}$ proporcionados por (6.3) son idénticos a los que se obtendrían de (6.2). En efecto, si existe tal ortogonalidad, la matriz inversa en (6.4) es una matriz diagonal por bloques y $\hat{\beta} = (X'X)^{-1}X'\vec{Y}$.

Fuera de este caso particular, los estimadores de $\vec{\beta}$ procedentes de (6.4) son diferentes a los que se obtendría de estimar (6.2).

Sin embargo, (6.4) *proporciona estimadores insesgados*, sean cuales fueren los regresores irrelevantes añadidos[1]. En efecto, sustituyendo (6.2) en (6.4) tenemos:

$$\begin{aligned} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} &= \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} X' \\ Z' \end{pmatrix} \left[\begin{pmatrix} X & Z \end{pmatrix} \begin{pmatrix} \vec{\beta} \\ \vec{0} \end{pmatrix} + \vec{\epsilon} \right] && (5.5) \\ &= \begin{pmatrix} \vec{\beta} \\ \vec{0} \end{pmatrix} + \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} X'\vec{\epsilon} \\ Z'\vec{\epsilon} \end{pmatrix}. && (5.6) \end{aligned}$$

Al tomar valor medio en la ecuación anterior obtenemos:

$$\begin{aligned} E[\hat{\beta}] &= \vec{\beta}, && (5.7) \\ E[\hat{\gamma}] &= \vec{0}. && (5.8) \end{aligned}$$

De la misma ecuación (6.6) obtenemos que la matriz de covarianzas del vector $(\hat{\beta}' \ \hat{\gamma}')'$ es:

$$\Sigma = \sigma^2 \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1}. \tag{5.9}$$

[1]De los que lo único que supondremos es que no introducen combinaciones lineales exactas que hagan inestimables los parámetros.

El bloque superior izquierdo de (6.9) es la matriz de covarianzas de los $\hat{\beta}$ obtenidos en el modelo sobreparametrizado. Debemos comparar dicho bloque con $\sigma^2(X'X)^{-1}$, matriz de covarianzas de los $\hat{\beta}$ obtenidos al estimar el modelo (6.2).

Haciendo uso del Teorema A.3, pág. 230, vemos que el bloque que nos interesa de (6.9) es σ^2 multiplicado por

$$(X'X)^{-1} + (X'X)^{-1}X'Z[Z'Z - Z'X(X'X)^{-1}X'Z]^{-1}Z'X(X'X)^{-1}.$$

Por simple inspección vemos que el segundo sumando es una matriz definida no negativa[2], y por tanto la expresión anterior tendrá en su diagonal principal elementos no menores que los de la diagonal principal de $(X'X)^{-1}$. En consecuencia, la inclusión de regresores irrelevantes no disminuye, y en general incrementa, las varianzas de los estimadores de los parámetros relevantes. No afecta sin embargo a su insesgadez.

De cuanto antecede se deduce que

$$\left(\vec{Y} - \begin{pmatrix} X & Z \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix}\right) \tag{5.10}$$

es un vector aleatorio de media cero. Denominando,

$$\begin{aligned} L &= \begin{pmatrix} X & Z \end{pmatrix}, \\ \hat{\delta} &= \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix}, \end{aligned}$$

un desarrollo enteramente similar al realizado en el Teorema 7.1, pág. 72, muestra que en el modelo sobreparametrizado

$$SSE = \vec{Y}\,'(I - L(L'L)^{-1}L')\vec{Y} = \vec{\epsilon}\,'(I - L(L'L)^{-1}L')\vec{\epsilon} \tag{5.11}$$

es, bajo los supuestos habituales más normalidad, una forma cuadrática con distribución $\sigma^2\chi^2_{N-(p+q)}$, en que p y q son respectivamente los rangos de X y Z. En consecuencia,

$$\hat{\sigma}^2 = \frac{SSE}{N-(p+q)} \tag{5.12}$$

[2]Llamemos G a dicho segundo sumando. Para mostrar que es definida no negativa, basta ver que para cualquier $\vec{a}$ se verifica $\vec{a}'G\vec{a} \geq 0$. Pero $\vec{a}'G\vec{a} = \vec{b}'(Z'Z - Z'X(X'X)^{-1}XZ)^{-1}\vec{b}$ con $\vec{b} = Z'X(X'X)^{-1}\vec{a}$; ya sólo tenemos que comprobar que $(Z'Z - Z'X(X'X)^{-1}XZ)^{-1}$ es definida no negativa, o equivalentemente que $(Z'Z - Z'X(X'X)^{-1}XZ)$ lo es. Esto último es inmediato: $(Z'Z - Z'X(X'X)^{-1}XZ) = Z'(I - X(X'X)^{-1}X)Z$, y $\vec{d}\,'Z'(I - X(X'X)^{-1}X)Z\vec{d}$ puede escribirse como $\vec{e}\,'(I - X(X'X)^{-1}X)\vec{e}$ con $\vec{e} = Z\vec{d}$. La matriz de la forma cuadrática en $\vec{e}$ es la conocida matriz de coproyección, definida no negativa por ser idempotente (con valores propios cero o uno).

es un estimador insesgado de σ^2. El único efecto adverso de la inclusión de los q regresores irrelevantes ha sido la pérdida de otros tantos grados de libertad.

5.3. Omisión de regresores relevantes.

Sea $X = (X_1 \vdots X_2)$ una matriz de diseño particionada en sendos bloques de p y r columnas. Sea $\vec{\beta}\,' = (\vec{\beta}\,'_1 \vdots \vec{\beta}\,'_2)$ el correspondiente vector de $p+r$ parámetros. Consideremos el caso en que el modelo "correcto" es

$$\vec{Y} \;=\; X\vec{\beta} + \vec{\epsilon} = X_1\vec{\beta}_1 + X_2\vec{\beta}_2 + \vec{\epsilon}, \tag{5.13}$$

pese a lo cual estimamos el modelo "escaso"

$$\vec{Y} \;=\; X_1\vec{\beta}_1 + \vec{\epsilon}. \tag{5.14}$$

Estimar (6.14) es lo mismo que estimar (6.13) junto con las restricciones $h : \vec{\beta}_2 = \vec{0}$, expresables así:

$$\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \vec{\beta}_1 \\ \vec{\beta}_2 \end{pmatrix} \;=\; \begin{pmatrix} \vec{0} \\ \vec{0} \end{pmatrix} \tag{5.15}$$

En consecuencia, podemos deducir cuanto necesitamos saber haciendo uso de los resultados en la Sección 5.3. Las siguientes conclusiones son así inmediatas:

- El estimador $\hat{\beta}_1^{(h)}$ obtenido en el modelo "escaso" (6.14) es, en general, sesgado. El sesgo puede obtenerse haciendo uso de (5.11). Tenemos así que

$$\begin{pmatrix} \hat{\beta}_1^{(h)} \\ \vec{0} \end{pmatrix} \;=\; \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - \vec{0}),$$

  y en consecuencia

$$E[\hat{\beta}_1^{(h)} - \vec{\beta}_1] \;=\; -\left[(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}\begin{pmatrix} \vec{0} \\ \vec{\beta}_2 \end{pmatrix}\right]_{(p\times 1)} \tag{5.16}$$

  en que $[M]_{(p\times q)}$ designa el bloque superior izquierdo con p filas y q columnas de la matriz M. La ecuación (6.16) muestra que el sesgo introducido depende de la magnitud de los parámetros asociados a los regresores omitidos.

- La ecuación (6.16) muestra también que hay un caso particular en que $\hat{\beta}_1^{(h)}$ es insesgado para $\vec{\beta}_1$; cuando las columnas de X_1 y las de X_2 son ortogonales, $X_1'X_2 = 0$, la matrix $(X'X)^{-1}$ es diagonal por bloques, y

$$(X'X)^{-1}A' = \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} \tag{5.17}$$

  tiene sus primeras p filas de ceros. Ello hace que el bloque considerado en (6.16) esté formado por ceros.

- El estimador de la varianza de la perturbación

$$\hat{\sigma}^2 = \frac{SSE}{N-p} = \frac{(\vec{Y} - X_1\hat{\beta}_1^{(h)})'(\vec{Y} - X_1\hat{\beta}_1^{(h)})}{N-p} \tag{5.18}$$

  no es insesgado. En efecto, puede verse que no es de aplicación a (6.18) el Teorema 3.3, pág. 23.

5.4. Consecuencias de orden práctico

Los resultados de las dos Secciones anteriores pueden ayudarnos a tomar decisiones a la hora de especificar un modelo. Hemos visto que sobreparametrizar no introduce sesgos: tan sólo incrementa la varianza de los estimadores y resta grados de libertad. Errar "por exceso" tendrá por ello en general consecuencias menos graves, y tanto menos importantes cuanto mayor sea el tamaño muestral. La pérdida de un grado de libertad adicional originada por la inclusión de un parámetro es menos importante cuando los grados de libertad restantes $(N - p)$ siguen siendo muchos.

La sóla circunstancia en que la inclusión de un regresor innecesario puede perjudicar gravemente la estimación se presenta cuando la muestra es muy pequeña o el parámetro adicional es aproximadamente combinación lineal de los ya presentes. A esta última cuestión volveremos en el Capítulo 10.

Omitir regresores relevantes tiene consecuencias en general más graves y que no se atenúan al crecer el tamaño muestral: el sesgo de $\hat{\beta}_1^{(h)}$ en el modelo "escaso" (6.14) no decrece hacia cero al crecer N.

En este capítulo hemos rastreado las consecuencias de dos posibles errores de especificación "puros": falta o sobra de regresores. En la práctica los dos tipos de errores se pueden presentar conjuntamente y sus efectos se combinan.

Conocidos los problemas de una mala especificación se plantea el problema de cómo lograr una buena. Esta cuestión se trata en el Capítulo 13. Algunas técnicas de análisis gráfico de residuos que pueden ser de ayuda en la especificación de modelos se consideran en la Sección 14.2.1.