Tema 5

Errores de especificación en la elección de los regresores

Contenido

5.1. Introducción

La primera especificación de un modelo de regresión implica tomar varias decisiones, a menudo previas a la confrontación de éste con los datos. Algunas de estas decisiones son:

- Elección de la variable dependiente.
- Elección de las variables explicativas.
- Medición de las variables.
- Forma funcional de la relación. Estabilidad.
- Especificación de las propiedades del término de error.

En los temas anteriores hemos especificado un modelo de regresión donde se satisfacen una serie de hipótesis básicas. Algunas de estas hipótesis pueden no mantenerse si las decisiones adoptadas son erróneas o porque simplemente, dadas las características de las variables del modelo y de los datos a utilizar, estas hipótesis pudieran no ser adecuadas. Esto puede influir negativamente en las propiedades del estimador utilizado y en la inferencia, siendo las decisiones posteriores sobre el modelo erróneas. En muchos casos la evaluación de un modelo puede estar influenciada por esta primera especificación. Por ello, es importante disponer de instrumentos o contrastes que nos permitan hacer un diagnóstico sobre si son aceptables ciertas decisiones o hipótesis adoptadas. Estos instrumentos pueden ser un análisis gráfico de los residuos o contrastes estadísticos donde se traten de detectar problemas de mala especificación.

En este tema nos vamos a centrar en ilustrar las implicaciones que pueden tener decisiones erróneas en términos de la elección de las variables explicativas o regresores. Para ello vamos a proponer que conocemos el modelo correcto y consideramos separadamente dos situaciones:

a) Omisión de variables explicativas relevantes. Analizaremos las implicaciones en el estimador MCO y en la validez de los contrastes de significatividad. Veremos la utilización del gráfico de residuos y algún contraste de mala especificación con algunos ejemplos empíricos.

b) Inclusión de variables irrelevantes. En este caso nos interesaremos por los efectos de haber incluido variables que sabemos no tendrían que estar en el modelo. La cuestión es cómo detectar y decidir en la práctica qué variables son o no relevantes. También discutiremos estas cuestiones utilizando un caso práctico.

Aunque teóricamente analizaremos cada uno de estos efectos por separado y asumiremos que conocemos la especificación correcta, en la práctica podemos tener combinados estos efectos.

5.2. Efectos de omisión de variables relevantes

Podemos seguir con nuestro ejemplo sobre el precio de la vivienda en el que queríamos explicar esta variable, medida en miles de dólares, en función de una serie de variables explicativas

como podían ser el tamaño de la vivienda $F2$, el número de habitaciones $BEDRMS$ y el número de baños $BATHS$. En principio, vamos a considerar que el modelo correcto para explicar el precio de la vivienda es

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \quad i = 1, \ldots, N \tag{5.1}$$

donde se satisfacen las hipótesis básicas pero se estima por MCO el siguiente,

$$P_i = \beta_1 + \beta_3\, BEDRMS_i + \beta_4\, BATHS_i + v_i \qquad i = 1, \ldots, N \tag{5.2}$$

En el modelo considerado a la hora de estimar se ha omitido la variable F2 o tamaño de la vivienda. Si esta variable es relevante entonces $\beta_2 \neq 0$ por lo que el error v_i recogerá la variable omitida, esto es $v_i = \beta_2 F2_i + u_i$, siendo $E(v_i) = \beta_2 F2_i \neq 0$. Luego en el modelo mal especificado no se satisface una de las hipótesis básicas. Esto a su vez implica que la covarianza entre las variables incluidas y el error del modelo (5.2) dependerá de la covarianza entre la variable omitida $F2_i$ y cada una de las incluidas $BEDRMS_i$ y $BATHS_i$. Si estas no son cero, esto introducirá un sesgo en los coeficientes estimados que será función de estas covarianzas. El signo del sesgo dependerá del signo del coeficiente β_2 y de los signos de estas covarianzas. Se puede demostrar que los sesgos de estimar por MCO β_3 y β_4 en el modelo (5.2) son

$$E(\hat{\beta}_3) - \beta_3 = \beta_2 \frac{S_{23}S_{44} - S_{24}S_{34}}{S_{33}S_{44} - S_{34}^2} \qquad E(\hat{\beta}_4) - \beta_4 = \beta_2 \frac{S_{24}S_{33} - S_{23}S_{34}}{S_{33}S_{44} - S_{34}^2} \tag{5.3}$$

donde $S_{js} = \sum_i (X_{ji} - \bar{X}_j)(X_{is} - \bar{X}_s)$, siendo la covarianza muestral entre dos variables j, s si $j \neq s$, y la varianza muestral de la variable j si $j = s$. Como se puede apreciar, el sesgo en la estimación de ambos coeficientes depende de las covarianzas entre las variables relevante excluida $F2$ y cada una de las variables incluidas $BEDRMS$ y $BATHS$[1]. Además depende del coeficiente β_2 que en el modelo correcto (5.1) se esperaba fuera positivo, pero la dirección del signo de cada sesgo no es clara ya que depende del signo del cociente que acompaña a β_2. Para que no hubiera sesgo en la estimación de cualquiera de estos dos coeficientes **ambas variables incluidas**, $BEDRMS$ y $BATHS$ tendrían que estar **incorreladas con** el tamaño de la vivienda o **variable excluida**, cosa poco probable en este ejemplo.

[1] Si el modelo de partida correcto hubiera sido

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + u_i \quad i = 1, \ldots, N \tag{5.4}$$

pero hubiéramos considerado para estimar

$$P_i = \beta_1 + \beta_3 BEDRMS_i + v_i \quad i = 1, \ldots, N \tag{5.5}$$

entonces el sesgo en estimar β_3 en (5.5) sería simplemente

$$E(\hat{\beta}_3) - \beta_3 = \beta_2 \frac{S_{23}}{S_{33}} \tag{5.6}$$

El sesgo sigue dependiendo de la covarianza entre la variable omitida $F2$ y la incluida $BEDRMS$ dada por S_{23}. En este caso se puede esperar que el sesgo fuera positivo ya que tanto S_{23} como β_2 se esperan sean positivos. El efecto de omitir $F2$ o no controlar por el tamaño de la vivienda en el modelo (5.5) será sobreestimar el efecto marginal de tener una habitación más en la vivienda sobre el precio de ésta. Por tanto, el número de habitaciones estaría también de alguna forma representando el papel del tamaño de la vivienda, que no se ha incluido en el modelo. No se estimaría con sesgo si $S_{23} = 0$, cosa que no parece factible ya que el número de habitaciones estará correlacionado con el tamaño de la vivienda.

En cuanto al sesgo en la estimación del coeficiente que acompaña al término constante se puede demostrar que es[2]

$$E(\hat{\beta}_1) - \beta_1 = \beta_2 \left(\bar{X}_2 - \frac{S_{23}S_{44} - S_{24}S_{34}}{S_{33}S_{44} - S_{34}^2} \bar{X}_3 - \frac{S_{24}S_{33} - S_{23}S_{34}}{S_{33}S_{44} - S_{34}^2} \bar{X}_4 \right) \tag{5.7}$$

Vemos que en este caso aún siendo $S_{23} = S_{24} = 0$ el sesgo no se anularía, ya que todavía depende de la media de la variable omitida $\bar{X}_2$, que generalmente no va a ser cero. De este resultado se puede argumentar que el coeficiente que acompaña al término constante, generalmente va a recoger efectos de variables omitidas aún cuando esto no influya en la estimación del resto de parámetros o pendientes por estar estas variables incorreladas con las incluidas. Por ello, normalmente es conveniente no excluir el término constante, a no ser que se tengan fuertes razones teóricas para hacerlo.

Si se estiman con sesgo los coeficientes β_j, también serán incorrectos los contrastes de significatividad individual, conjunta y otro tipo de contrastes sobre los coeficientes del modelo utilizando estas estimaciones sesgadas. Ahora bien, ¿serán fiables los contrastes sobre las pendientes si se dan las condiciones para que los estimadores de estos parámetros no sean sesgados? La respuesta es que no, ya que aún dandose las condiciones de incorrelación entre regresores incluidos y variables relevantes excluidas, el estimador de la matriz de varianzas y covarianzas de esos coeficientes estimados seguirá siendo sesgada. Esto se debe a que el estimador del parámetro σ^2 utilizando la suma de cuadrados residual de la estimación del modelo mal especificado estará sesgado en cualquiera de los casos.

Luego vemos que en general las consecuencias de omitir variables relevantes en la especificación de un modelo son serias, especialmente en la inferencia.

¿Cómo detectar que esto pueda estar ocurriendo? Una primera cuestión es tener en cuenta el modelo teórico de interés y pensar qué variables pueden faltar en el modelo empírico. Por otro lado, podemos ayudarnos de contrastes que puedan señalar la existencia de algún problema de mala-especificación[3].

Además, el análisis de los residuos nos puede ayudar a ver si hemos dejado fuera factores relevantes. Por ejemplo, podemos ver el gráfico de los residuos por observación y ver si estos presentan algún comportamiento sistemático que pueda apuntar en esa dirección.

Por ejemplo, consideremos los resultados de la estimación de los modelos (5.1) y (5.2) para explicar el precio de la vivienda[4]

[2] Ocurre lo mismo si consideramos que el modelo estimado es (5.5) y el verdadero modelo es (5.4).

[3] En este tema ilustraremos alguno de estos contrastes, aunque no todos. Incluso algunos contrastes diseñados para analizar si el término de error no está autocorrelacionado, puede capturar también otro tipo de cuestiones de mala especificación.

[4] Los valores entre paréntesis son los correspondientes estadísticos t de significatividad individual.

Variable	**Modelo (5.1)**	**Modelo (5.2)**
	Supuestamente Correcto	
CONSTANT	129,062	27,2633
	(1,462)	(0,182)
F2	0,1548	
	(4,847)	
BEDRMS	-21,588	-10,1374
	(-0,799)	(-0,216)
BATHS	-12,193	138,795
	(-0,282)	(2,652)
Suma de cuadrados de los residuos	16700,1	55926,4
Desviación típica de los residuos ($\hat{\sigma}$)	40,8657	71,3037
R^2	0,836	0,450706
$\bar{R}^2$	0,787	0,350834
F de significación conjunta	16,989	4,51285
Grados de libertad	10	11
Criterio de Akaike (AIC)	146,908	161,829
Criterio de Schwarz (BIC)	149,464	163,746

Tabla 5.1: Modelos (5.1) y (5.2) estimados para el precio de la vivienda

Como ya comentamos en el capítulo anterior, la omisión de la variable $F2$ empeora bastante el ajuste tanto en términos del R^2 como del $\bar{R}^2$, AIC y BIC. El coeficiente estimado que más ha cambiado es el que acompaña a la variable $BATHS$ pasando a tener signo positivo y ser significativamente distinto de cero. Parece que, dado que ambas variables representan también tamaño de la vivienda, el efecto indirecto de la omisión de esta variable puede estar siendo capturando más por el coeficiente de $BATHS$ que por el de $BEDRMS$.

Podemos mirar a las correlaciones entre la variable excluida $F2$ y las incluidas $BEDRMS$ y $BATHS$. En la ventana principal de Gretl donde tenemos estas variables, las seleccionamos con el botón izquierdo del ratón, mientras mantenemos la tecla de mayúsculas ⇑, y en *Ver* → *matriz de correlación* obtenemos

Coeficientes de correlación, usando las observaciones 1 - 14
valor crítico al 5 % (a dos colas) = 0,5324 para n = 14

F2	BEDRMS	BATHS	
1,0000	0,4647	0,7873	F2
	1,0000	0,5323	BEDRMS
		1,0000	BATHS

Vemos que, aunque tanto el número de habitaciones $BEDRMS$ como el número de baños $BATHS$ presenta una correlación positiva con la variable excluida, tamaño de la vivienda $F2$, es la variable $BATHS$ la que presenta una mayor correlación con esta última.

Seguidamente vamos a analizar diversos gráficos de los residuos del ajuste del modelo (5.2) donde hemos omitido $F2$ que parece ser relevante. De la estimación de este modelo en la ventana de estimación **gretl:modelo2** elegimos

Gráficos → *Grafico de residuos* → *Por número de observación*

que nos muestra el gráfico de residuos por observación según están las 14 observaciones ordenadas en la muestra. Lo podemos guardar posicionando el cursor dentro de la ventanta del gráfico y pinchando con el botón derecho del ratón, aparece un menú con distintas opciones y formatos para guardarlo.

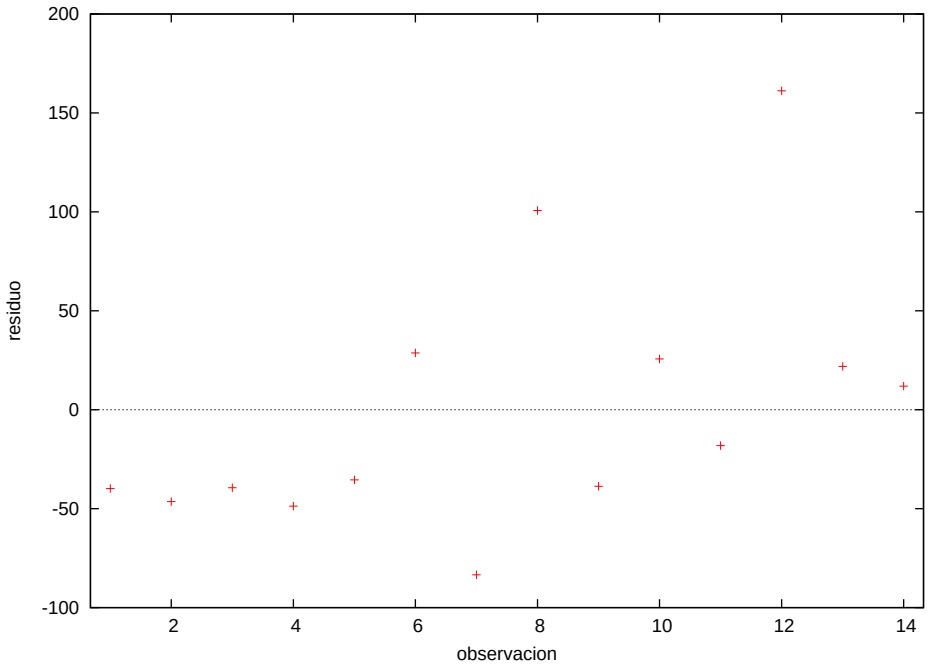

Gráfico 5.1: Gráfico de los residuos del Modelo (5.2) por observación

En el gráfico se puede apreciar que hay demasiados residuos negativos juntos al comienzo de la muestra y a medida que vamos hacia las últimas observaciones o viviendas, estos se concentran más en la parte positiva. Si observamos la disposición de las viviendas en la muestra, veremos que están ordenadas en función creciente del tamaño de la vivienda. Luego los residuos negativos estarían asociados en general con viviendas de menor tamaño y los positivos con viviendas de mayor tamaño. Esto sugiere un comportamiento sistemático en la disposición de los residuos alrededor de su media muestral que es cero.

El gráfico de los residuos sobre la variable $F2$ puede ayudar a ver si hay alguna relación. De hecho el gráfico nos mostrará la recta de regresión de los residuos sobre esta variable si es que existe una relación significativa. Para obtener el gráfico primero tenemos que guardar los residuos de la estimación del modelo (5.2). Para ello, en la ventana de estimación **gretl:modelo2** elegimos

Guardar → *Residuos*

y le damos un nombre a la serie de residuos. Esta serie aparecerá en la ventana principal **gretl** y la podremos utilizar posteriormente. En esta misma ventana elegimos

Ver → Gráficos → Grafico X-Y (scatter)

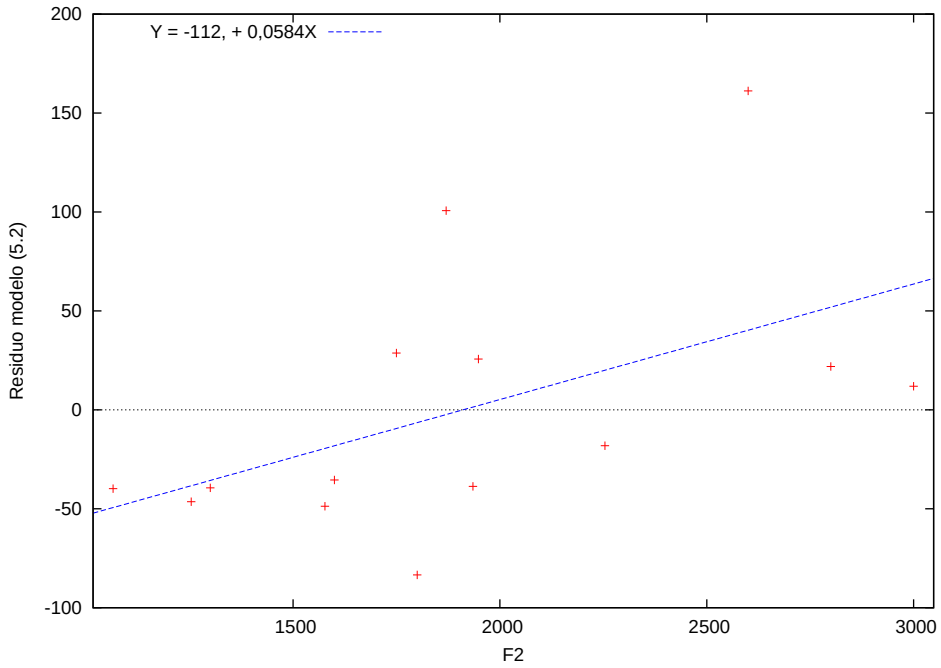

Gráfico 5.2: Gráfico de los residuos del Modelo (5.2) sobre F2

En la ventana que aparecerá posteriormente, especificamos que variable se representa en el eje de ordenadas *eje X*, en este caso $F2$, y en el eje de abcisas o *eje Y*, en este caso los residuos de la estimación del Modelo (5.2). En este gráfico podemos apreciar que hay una relación positiva significativa entre los residuos de la estimación del modelo (5.2) y la variable $F2$ omitida en ese modelo. De hecho, la línea que aparece en el gráfico representa la recta de regresión de los residuos sobre esa variable. Esto indica que cierto componente residual puede ser explicado por la variable que no hemos incluido.

Lo detectado en estos gráficos puede ser contrastado utilizando el siguiente contraste que se debe a Engle (1982). Este contraste utiliza el R^2 de la regresión auxiliar de los residuos del modelo que se está analizando sobre la variable o variables que sospechamos puedan ser candidatas a ser incluidas en él por ser relevantes. En nuestro caso sería realizar la regresión

$$\hat{u}_i = \delta_1 + \delta_2 F2_i + \xi_i \qquad i = 1, \ldots, N \tag{5.8}$$

El estadístico de contraste es NR^2 donde el R^2 es el coeficiente de determinación de esta regresión auxiliar. La distribución exacta del estadístico, bajo la hipótesis nula de que la variable $F2$ no es una variable relevante a incluir en el modelo, no es conocida pero se puede aproximar por la distribución χ^2 con un grado de libertad[5]. Esta aproximación será mejor cuanto mayor sea el tamaño muestral.

[5]En general, los grados de libertad serán el número de regresores de la regresión auxiliar sin contar el término constante.

En el ejemplo que nos ocupa esta regresión auxiliar la podemos obtener con Gretl eligiendo

Modelo → Minimos Cuadrados Ordinarios

y en la ventana que emerge elegir como variable dependiente la serie de residuos de la estimación del modelo (5.2) que teníamos guardada y como regresores a $F2$ además de la constante. Los resultados de esta regresión auxiliar (5.8) para el ejemplo que nos ocupa son

$$\widehat{\mathrm{u}}_i = \underset{(-1,995)}{-111,588} + \underset{(2,078)}{0,0583946}\,\mathrm{F2}_i$$
$$N = 14 \quad R^2 = 0,264584$$

Si queremos guardar el valor muestral NR^2 podemos hacerlo en esa misma ventana eligiendo

*Guardar → T*R-cuadrado*

El valor muestral del estadístico $NR^2 = 3,70417$ se muestra en la ventana principal con el resto de variables. Este valor habrá que compararlo en este caso con el valor crítico $\chi^2_{(1)\alpha}$ utilizando en el contraste un nivel de significación α concreto.

Para buscar el valor crítico en las tablas de la Chi-cuadrado con 1 grado de libertad podemos elegir en la ventana principal de Gretl, *Herramientas → Tablas Estadísticas* y en la ventana que aparece seleccionar la chi-cuadrado especificando 1 grado de libertad. Aparece una ventana con los valores críticos de la distribución Chi-cuadrado para distintos niveles de significación.

También podemos obtener el *valor-p* dado el valor muestral del estadístico. En la ventana principal de nuevo en *Herramientas → Buscador de valores-p*, y en la ventana que aparece seleccionar la chi-cuadrado especificando en la primera casilla 1 grado de libertad y el valor muestral del estadístico en la segunda casilla. Aparece una ventana con la siguiente información: Chi-cuadrado(1): área a la derecha de 3,70417 = 0,0542767 (a la izquierda: 0,945723).

Por lo tanto, como el *valor-p* obtenido es $0,0542767$ que, aunque poco, es algo mayor que $0,05$, no se rechazaría la hipótesis nula de que $F2$ sea una variable importante a añadir al modelo al 5 %, pero sí al 10 % al ser el *valor-p* en ese caso menor que ese nivel de significación. Vemos que la hipótesis nula se rechazaría al 10 % de significación ya que el valor muestral en ese caso $NR^2 = 3,70417$ sería mayor que el valor crítico $\chi^2_{(1)0,1} = 2,706$, aunque no se rechazaría al 5 %. Luego existe cierta evidencia de que $F2$ sea una variable relevante a añadir en el modelo.

¿Cómo cambiarían los gráficos (5.1) y (5.2) si consideramos los residuos del modelo (5.1) que incluye a la variable F2? Estos corresponden a los gráficos de la Figura (5.3). En este caso la disposición de los residuos positivos y negativos es más aleatoria alrededor de su media muestral. Por otro lado, el gráfico de los residuos del modelo (5.1) sobre la variable $F2$ ya no muestra esa relación positiva entre ambas variables.

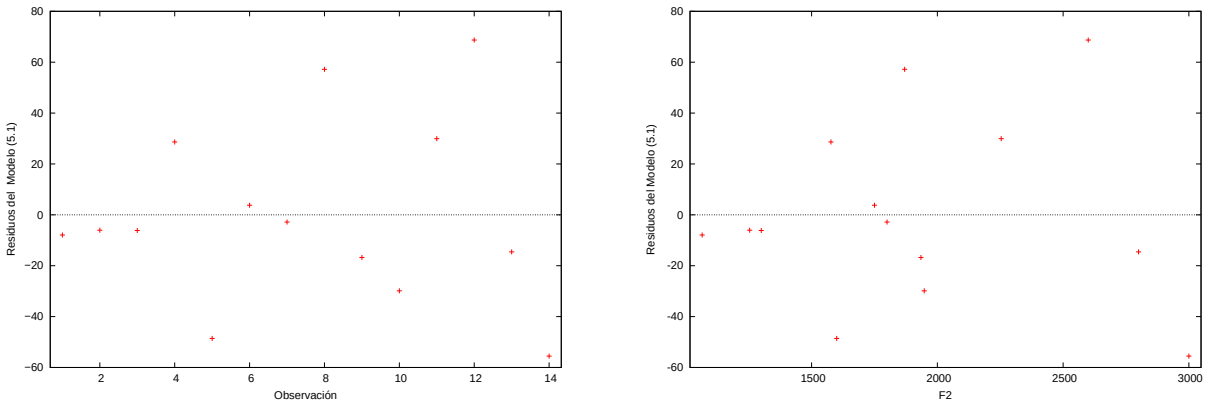

Gráfico 5.3: Gráficos de los residuos del Modelo (5.1) sobre observación y sobre F2

5.3. Efectos de inclusión de variables irrelevantes

Supongamos ahora que el modelo correcto para el precio de la vivienda es

$$P_i = \beta_1 + \beta_2 F2_i + u_i \quad i = 1, \ldots, N \tag{5.9}$$

donde se satisfacen las hipótesis básicas, pero incluimos en la regresión una variable más que no es relevante, $BEDRMS$. El modelo que ajustamos es

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + u_i \quad i = 1, \ldots, N \tag{5.10}$$

En este modelo se siguen satisfaciendo las hipótesis básicas, ya que el valor poblacional del coeficiente que acompaña a la variable $BEDRMS$ es cero al no ser una variable relevante, por lo que el término de error no cambia. Pero en la regresión se estimarán todos los coeficientes, también los de las variables irrelevantes y la estimación puntual de β_3 no será en general cero. ¿Qué consecuencias tendrá este error de especificación?

- En este caso, los estimadores de todos los coeficientes son insesgados, por lo que $E(\hat{\beta}_j) = \beta_j \ \forall j$. En particular, $E(\hat{\beta}_3) = 0$.

- La matriz de varianzas y covarianzas se estimará correctamente con el estimador habitual. Por lo que tanto los intervalos de confianza como los procedimientos habituales de contraste sobre los coeficientes β_j siguen siendo válidos.

- El coste de este error de especificación es la pérdida de eficiencia en la estimación. Si se comparan las varianzas de los coeficientes estimados en el modelo incorrecto relativamente al correctamente especificado, estas serán mayores en el primero. Por ejemplo, se puede demostrar que esta pérdida de eficiencia depende de la correlación entre $F2$ y $BEDRMS$ siendo mayor cuanto mayor sea esta correlación.

  En particular, para β_2 el ratio de la varianza del estimador de este coeficiente en el modelo incorrecto (5.10) sobre la varianza del estimador en el modelo correcto (5.9) es

$$\frac{var(\hat{\beta}_2)_{(10)}}{var(\hat{\beta}_2)_{(9)}} = \frac{1}{\left(1 - \rho_{23}^2\right)} \geq 1 \tag{5.11}$$

siendo $0 \leq \rho_{23}^2 \leq 1$ el coeficiente de correlación al cuadrado entre $F2$ y $BEDRMS$. En el caso de los datos que estamos utilizando *data4-1* sobre 14 viviendas este ratio es $\left(1/\left(1-(0,5323)^2\right)\right) = 1,4$, luego hay cierta pérdida de eficiencia en la estimación de β_2 en el modelo (5.10) relativamente a (5.9). La inclusión de la variable supuestamente irrelevante $BEDRMS$ hace que estimemos con menor precisión el coeficiente β_2. Lo mismo ocurre con el coeficiente β_1.

¿Cómo podemos detectar la presencia de variables innecesarias?

Una posibilidad es comenzar por un modelo relativamente general y utilizar los contrastes de significatividad individual, así como las medidas de bondad de ajuste $\bar{R}^2$ o los criterios de información AIC o BIC por ejemplo. Estos indicadores nos pueden ayudar en la toma de esta decisión. Los resultados obtenidos de la estimación de los modelos (5.9) y (5.10) se muestran en la tabla (5.2)[6]. Considerando que nuestro modelo de partida es el modelo más general, **Modelo (5.10)**, y utilizando el contraste de significatividad individual para el coeficiente que acompaña a $BEDRMS$, podríamos considerar que esta variable no es relevante en explicar la variación en el precio de la vivienda una vez hemos incluido el tamaño de ésta. Eliminar esta variable del modelo también mejora el resto de indicadores de ajuste, mayor $\bar{R}^2$, menores AIC y BIC. Se puede observar también que las desviaciones típicas estimadas se reducen bastante. Por otro lado, tanto en el modelo (5.10) como en el (5.9), la variable $F2$ es significativa indicando su relevancia en explicar la variación en el precio de la vivienda.

Variable	**Modelo (5.9)**	**Modelo (5.10)**
	supuestamente correcto	
CONSTANT	52,351	121,179
	(1,404)	(1,511)
	[37,28]	[80,1778]
F2	0,13875	0,14831
	(7,407)	(6,993)
	[0,0187]	[0,0212]
BEDRMS		-23,911
		(-0,970)
		[24,642]
Suma de cuadrados de los residuos	18273,6	16832,8
Desviación típica de los residuos ($\hat{\sigma}$)	39,023	39,1185
R^2	0,821	0,835
$\bar{R}^2$	0,806	0,805
F de significación conjunta	54,861	27,767
Grados de libertad	12	11
Criterio de Akaike (AIC)	144,168	145,019
Criterio de Schwarz (BIC)	145,447	146,936

Tabla 5.2: Modelos estimados para el precio de la vivienda.

[6]Entre paréntesis estadísticos t y entre corchetes las desviaciones típicas estimadas.

La aproximación de ir de un modelo más general a uno más restringido suele ser más conveniente que la aproximación contraria. En el caso de comenzar por un modelo más reducido e ir añadiendo variables secuencialmente, decidiendo mantenerlas o no en función de si son o no significativas, se corre el peligro de lo que se conoce con el nombre inglés de *data mining* o *torturar a los datos*.

El problema en la aproximación contraria es que, si el modelo de partida es demasiado general y los regresores están muy correlacionados, la precisión con la que estimemos los parámetros puede ser poca. Por esa falta de precisión en la estimación podemos tener coeficientes no significativamente distintos de cero, no siendo capaces de identificar el efecto de esas variables ya que la potencia de los contrastes de significación puede ser muy poca[7]. No rechazar en ese caso la hipótesis nula no es evidencia de que esas variables no sean relevantes sino de que el contraste tiene poca potencia.

[7]Este problema será tratado más en detalle en el tema de Multicolinealidad.

Bibliografía

Engle, R. F. (1982), "A general approach to Lagrangian Multiplier Modelo Diagnostics", *Journal of Econometrics*, vol. 20, pp. 83-104.