Tema 3

Modelo de Regresión Lineal Múltiple

Contenido

3.1. Introducción. Un ejemplo

En este tema consideramos introducir en el modelo de regresión, además del término constante, más de una variable explicativa por lo que pasamos del llamado modelo de regresión lineal simple al modelo de regresión lineal múltiple.

Comenzamos con el ejemplo que se ha seguido en el tema sobre el Modelo de Regresión Lineal Simple. El precio de una casa, en miles de dólares, (P) era la variable dependiente y las variables explicativas eran el término constante y el tamaño de la casa o el número de pies cuadrados del área habitable (F2). Ampliaremos el modelo incluyendo dos variables explicativas más, el número de habitaciones (BEDRMS) y el número de baños (BATHS) siendo el modelo de regresión lineal múltiple[1]

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \qquad i = 1, 2, \ldots, N \tag{3.1}$$

El modelo de regresión lineal general (MRLG), con K variables explicativas

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + u_i \qquad i = 1, 2, \ldots, N. \tag{3.2}$$

se puede escribir en notación matricial:

$$\underset{(N\times 1)}{Y} = \underset{(N\times K)}{X}\ \underset{(K\times 1)}{\beta} + \underset{(N\times 1)}{u}$$

donde cada uno de los elementos se definen:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1N} & X_{2N} & \cdots & X_{KN} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

Por el momento, seguimos suponiendo las mismas hipótesis básicas sobre el término de perturbación y sobre las variables explicativas o regresores, a saber:

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

ii) La perturbación sigue una distribución normal.

iii) Las variables X_2 a X_k no son estocásticas. Esto quiere decir que en muestras repetidas de N observaciones de $Y_i, X_{2i}, \ldots, X_{ki}$, las variables $X_{2i}, \ldots, X_{ki},\ i = 1, \ldots, N$ tomarían siempre los mismos valores. Este supuesto, junto a $E(u_i) = 0$, implica que los regresores y el término de perturbación están incorrelacionados.

iv) Los regresores son linealmente independientes, esto quiere decir que el rango de la matriz de datos de los regresores X es K tal que no tiene columnas repetidas ni unas son combinaciones lineales de otras.

v) Además se supone que se dispone de un número suficiente de observaciones para estimar los parámetros $\beta_j,\ \ j = 1, \ldots, K$, esto es $K < N$.

[1] Dado que seguimos con los mismos datos de sección cruzada utilizamos el subíndice $i = 1, \ldots, N$. La notación para datos de series temporales suele ser $t = 1, \ldots, T$.

Interpretación de cada uno de los coeficientes de regresión:

- Los parámetros β_j, $j = 2, \ldots, K$:

  Manteniendo constante el valor del resto de variables explicativas, si X_{ji} cambia en una unidad, Y_i se espera que cambie en media β_j unidades.

- El parámetro β_1 que acompaña al término constante recoge el valor esperado de la variable dependiente cuando el resto de variables explicativas o regresores incluidos toman el valor cero.

Siguiendo con el ejemplo, el modelo (3.1) se puede escribir en notación matricial:

$$\underset{(N\times 1)}{Y} = \underset{(N\times 4)}{X}\ \underset{(4\times 1)}{\beta} + \underset{(N\times 1)}{u}$$

donde cada uno de los elementos se definen:

$$Y = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_N \end{bmatrix} \quad X = \begin{bmatrix} 1 & F2_1 & BEDRMS_1 & BATHS_1 \\ 1 & F2_2 & BEDRMS_2 & BATHS_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & F2_N & BEDRMS_N & BATHS_N \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

Interpretación de los coeficientes:

- El coeficiente β_1 es el valor medio esperado de aquellas viviendas que no tienen ningún pie cuadrado de área habitable, ni habitaciones ni baños.

- El coeficiente β_2:

  Considerando dos casas con el mismo número de habitaciones y de baños, para aquella casa que tenga un pie cuadrado más de área habitable se espera que cambie en media su precio de venta en β_2 miles de dólares.

- El coeficiente β_3:

  Considerando dos casas con el mismo número de pies cuadrados de área habitable y número de baños, para aquella casa que tenga una habitación más se espera que cambie en media su precio de venta en β_3 miles de dólares.

- El coeficiente β_4:

  Considerando dos casas con el mismo número de pies cuadrados de área habitable y número de habitaciones, para aquella casa que tenga un baño más se espera que cambie en media su precio de venta en β_4 miles de dólares.

El análisis de regresión múltiple nos permite examinar el **efecto marginal** de una variable explicativa en particular, una vez hemos controlado por otras características recogidas en el resto de variables explicativas que mantenemos constantes. Por eso a veces al resto de regresores se les llama variables de control. Veremos más adelante cuándo es importante controlar por otras variables y qué problemas tendremos si las omitimos.

3.2. Estimación de Mínimos Cuadrados Ordinarios utilizando Gretl

Se dispone de una base de datos sobre el precio de venta de una vivienda y distintas características de 14 viviendas vendidas en la comunidad universitaria de San Diego en 1990. Son datos de sección cruzada y las variables que se consideran son:

P:	Precio de venta en miles de dólares (Rango 199.9 - 505)
F2:	Pies cuadrados de área habitable (Rango 1065 - 3000)
BEDRMS:	Número de habitaciones (Rango 3 - 4)
BATHS:	Número de baños (Rango 1,75 - 3)

Los datos para P y F2 son los mismos que los utilizados en el ejemplo del Tema 2 sobre el modelo de regresión lineal simple. Además tenemos información sobre dos nuevas variables que vamos a considerar incluir como explicativas en el modelo para el precio de la vivienda.

Comenzamos una sesión en Gretl para estimar este modelo con la muestra de 14 viviendas:

$$P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i \qquad i = 1, \ldots, 14$$

En la parte de arriba de la ventana principal de Gretl tenemos distintas opciones. Si posicionamos el cursor podemos ir eligiendo dentro de ellas.

1. Leemos los datos que están disponibles en Gretl como archivo de muestra:

   Archivo → *Abrir datos* → *Archivo de muestra*

   Elegir de Ramanathan el fichero *data4-1* proporcionados en el cuarto capítulo del libro de Ramanathan (2002). *Abrir*.

2. Podemos ver los datos de todas las variables. Las dos primeras columnas coinciden con los datos utilizados en el Tema 2.

P	F2	BEDRMS	BATHS
199.9	1065	3	1.75
228.0	1254	3	2.00
235.0	1300	3	2.00
285.0	1577	4	2.50
239.0	1600	3	2.00
293.0	1750	4	2.00
285.0	1800	4	2.75
365.0	1870	4	2.00
295.0	1935	4	2.50
290.0	1948	4	2.00
385.0	2254	4	3.00
505.0	2600	3	2.50
425.0	2800	4	3.00
415.0	3000	4	3.00

Tabla 3.1: Modelo (3.1). Datos de características de viviendas

3. Estimación por Mínimos Cuadrados Ordinarios (MCO).

   Modelo → Mínimos Cuadrados Ordinarios

   Se abre una nueva ventana. Utilizando el cursor, seleccionar de la lista de variables de la izquierda:

   - La variable dependiente (P) y pulsar elegir.
   - Las variables independientes o regresores de esta especificación y pulsar añadir cada vez. La variable Const es el término constante o variable que toma siempre valor uno. Por defecto ya está incluida pero si no se quisiera poner se podría excluir. Simplemente habría que seleccionarla con el cursor y dar a *Quitar*.

   Pinchar en *Aceptar*.

   Aparece una nueva ventana con los resultados de la estimación[2]. Iremos comentando los resultados mostrados. Situando el cursor en la parte de arriba de esta ventana podremos ver que hay distintos menús cuyas funciones estarán asociadas a esta regresión.

4. Hay varios formatos para guardar los resultados, como por ejemplo un formato compatible con Microsoft Word mediante:

   Editar → Copiar → RTF(Ms Word)

   Abrir un documento con Microsoft Word. Elegir *Edición → Pegar*. Se pegarán todos los resultados de la ventana anterior. Guardar el documento y minimizar si se quiere volver a utilizar más tarde para pegar y guardar otros resultados.

3.3. Análisis de los resultados mostrados

En esta sección vamos a ir comentando los resultados que nos muestra el programa cuando utilizamos la opción de estimación por Mínimos Cuadrados Ordinarios. Algunos de estos resultados ya han sido comentados en el Tema 2 sobre el modelo de regresión lineal simple, pero nos servirá también de repaso. Una vez especificado el modelo, el programa Gretl muestra en la ventana **gretl:modelo1** la siguiente información sobre la estimación MCO del modelo con los datos del fichero elegido:

Modelo 1: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	129,062	88,3033	1,4616	0,1746
F2	0,154800	0,0319404	4,8465	0,0007
BEDRMS	−21,587	27,0293	−0,7987	0,4430
BATHS	−12,192	43,2500	−0,2819	0,7838

[2]Recordar que esta ventana puede ser minimizada para su posible utilización posterior o el modelo puede guardarse en la sesión como icono. Si la cerramos tendríamos que volver a hacer lo mismo para obtener de nuevo esta ventana y poder elegir dentro de las opciones asociadas a esta regresión.

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	16700,1
Desviación típica de los residuos ($\hat{\sigma}$)	40,8657
R^2	0,835976
$\bar{R}^2$ corregido	0,786769
$F(3, 10)$	16,9889
valor p para $F()$	0,000298587
Log-verosimilitud	−69,453
Criterio de información de Akaike	146,908
Criterio de información Bayesiano de Schwarz	149,464
Criterio de Hannan–Quinn	146,671

Algunos Gráficos.
En la ventana de resultados de estimación, Gretl nos ofrece la posibilidad de analizar el gráfico de residuos así como el gráfico de la variable observada y estimada tanto por observación como sobre las distintas variables que hay en la especificación del modelo. Por ejemplo elegimos

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

y obtenemos el gráfico de los residuos del modelo estimado para el precio de la vivienda a lo largo de las 14 observaciones de la muestra En el gráfico 3.1 se observa que los residuos se

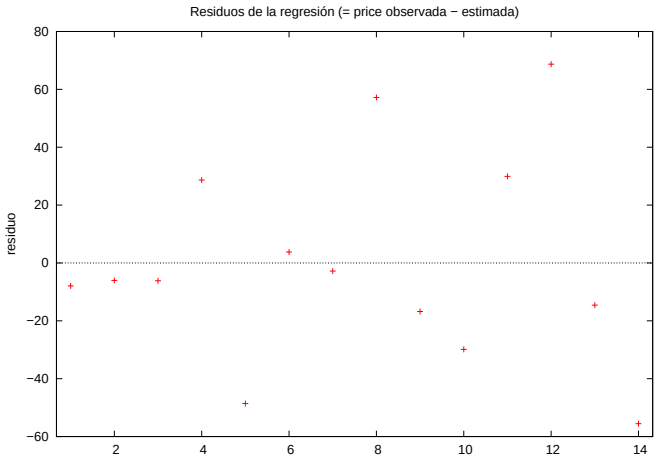

Gráfico 3.1: Gráfico de residuos por número de observación

disponen alrededor del valor cero ya que esta es su media muestral. La dispersión de estos residuos es mayor para las últimas viviendas en la muestra. Si elegimos

Gráficos → *Gráfico de residuos* → *Contra F2*

obtenemos el gráfico de los residuos sobre la variable *F2*. Este gráfico muestra que la dispersión de los residuos alrededor de su media muestral que es cero, aumenta a mayor valor de *F2*. Esto sugiere que la hipótesis básica sobre la varianza de la perturbación pueda no ser adecuada.

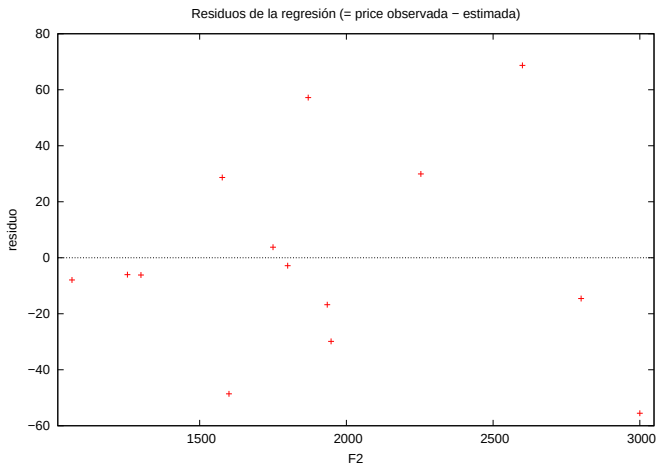

Gráfico 3.2: Gráfico de residuos contra la variable F2

Otro gráfico que ilustra la bondad del ajuste de nuestro modelo relativamente a los datos observados, es el gráfico de la variable estimada y observada por número de observación. Para obtener este gráfico elegimos

Gráficos → Gráfico de variable estimada y observada → por número de observación

De esta forma obtenemos el siguiente gráfico

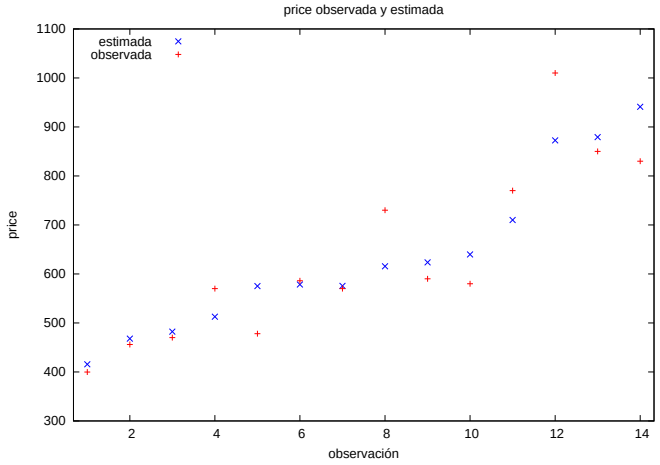

Gráfico 3.3: Gráfico de la variable estimada y observada por número de observación

En este gráfico se puede observar el valor estimado del precio de las viviendas en la muestra, dados los valores observados de las variables explicativas y el modelo estimado, en relación al precio observado. El ajuste parece empeorar para las últimas viviendas en la muestra. Si hacemos el gráfico de la variable estimada y observada contra la variable F2 que recoge el tamaño de las viviendas

Gráficos → Gráfico de variable estimada y observada → Contra F2

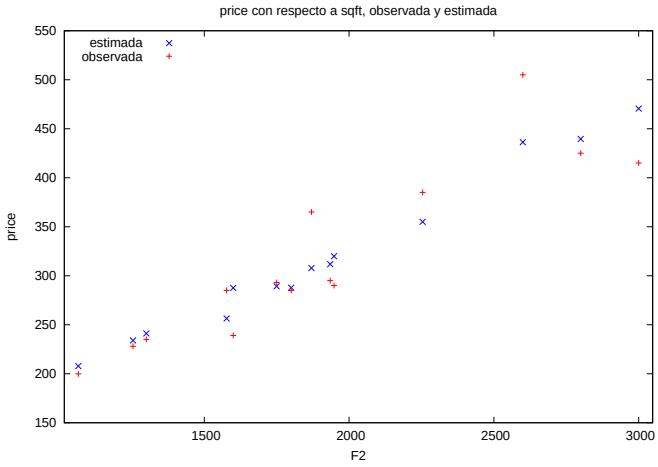

Gráfico 3.4: Gráfico de la variable estimada y observada contra F2

En el gráfico 3.4 se observa que el modelo se ajusta mejor a las observaciones asociadas a las viviendas de menor tamaño, ya que los valores estimados están más concentrados alrededor de los observados para esas viviendas. El ajuste es peor para viviendas de más de 2000 pies cuadrados.

3.3.1. Coeficientes estimados

Las estimaciones obtenidas de los coeficientes que se muestran en la segunda columna están asociados a cada una de las variables explicativas que figuran al lado en la primera columna. Dadas las realizaciones muestrales de la variable dependiente $Y_i \equiv P_i$, y explicativas, $X_{2i} \equiv F2_i$, $X_{3i} \equiv BEDRMS_i$, $X_{4i} \equiv BATHS_i$, las estimaciones se obtienen de minimizar la suma de cuadrados de los residuos con respecto a los coeficientes desconocidos $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$. Estos coeficientes estimados se han obtenido de utilizar el siguiente criterio de estimación por el método de Mínimos Cuadrados Ordinarios

$$\min_{\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\beta}_4} \sum_{i=1}^{N} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \hat{\beta}_4 X_{4i})^2$$

Las condiciones de primer orden de este problema resultan en cuatro ecuaciones con cuatro incógnitas.

$$\begin{aligned}
\sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} + \hat{\beta}_4 \sum X_{4i} \\
\sum Y_i X_{2i} &= \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{3i} X_{2i} + \hat{\beta}_4 \sum X_{4i} X_{2i} \\
\sum Y_i X_{3i} &= \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2 + \hat{\beta}_4 \sum X_{4i} X_{3i} \\
\sum Y_i X_{4i} &= \hat{\beta}_1 \sum X_{4i} + \hat{\beta}_2 \sum X_{2i} X_{4i} + \hat{\beta}_3 \sum X_{3i} X_{4i} + \hat{\beta}_4 \sum X_{4i}^2
\end{aligned}$$

Estas ecuaciones se conocen con el nombre de **Ecuaciones Normales**. Al igual que en el modelo de regresión lineal simple, la primera ecuación o primera condición asociada al término constante implica que la suma de los residuos debe de ser cero. El resto de ecuaciones

implican que los residuos tienen que ser ortogonales a cada una de las variables explicativas. En conjunto, estas condiciones implican que los residuos de la estimación MCO están incorrelacionados con los regresores. En términos matriciales se pueden escribir como:

$$X'Y = (X'X)\hat{\beta} \quad \Leftrightarrow \quad X'(Y - X\hat{\beta}) = 0 \quad \Leftrightarrow \quad X'\hat{u} = 0$$

Si las cuatro ecuaciones son linealmente independientes, el rango de $(X'X)$ es igual a $K = 4$, y por lo tanto existe una única solución a este sistema de ecuaciones. La solución será el estimador MCO del vector de parámetros β.

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$$

Sustituyendo los valores muestrales del fichero *data4-1* para Y y X darían lugar a las estimaciones obtenidas de los coeficientes.

Para el modelo especificado en la ecuación (3.1), la relación estimada es

$$\widehat{\mathrm{P}}_i = 129,062 + 0,1548\,\mathrm{SQFT}_i - 21,588\,\mathrm{BEDRMS}_i - 12,193\,\mathrm{BATHS}_i \tag{3.3}$$

Aunque hemos utilizado los mismos datos para P y F2 que en el Tema 2, el incluir las dos nuevas variables explicativas en el modelo ha hecho que las estimaciones de los coeficientes asociados al término constante y a F2 hayan cambiado[3].

Esto ocurre porque las nuevas variables BEDRMS y BATHS están correlacionadas con la ya incluida F2 y su media es distinta de cero[4].

Si esto no ocurriera y $\sum X_{3i} = \sum X_{4i} = \sum X_{2i}X_{3i} = \sum X_{2i}X_{4i} = 0$, las ecuaciones normales quedarían de la siguiente forma

$$\begin{aligned}
\textstyle\sum Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \textstyle\sum X_{2i} && \Leftrightarrow \textstyle\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i}) = 0 \\
\textstyle\sum Y_i X_{2i} &= \hat{\beta}_1 \textstyle\sum X_{2i} + \hat{\beta}_2 \textstyle\sum X_{2i}^2 && \Leftrightarrow \textstyle\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i})X_{2i} = 0 \\
\textstyle\sum Y_i X_{3i} &= \hat{\beta}_3 \textstyle\sum X_{3i}^2 + \hat{\beta}_4 \textstyle\sum X_{4i}X_{3i} \\
\textstyle\sum Y_i X_{4i} &= \hat{\beta}_3 \textstyle\sum X_{3i}X_{4i} + \hat{\beta}_4 \textstyle\sum X_{4i}^2
\end{aligned}$$

[3] En el caso de considerar un MRLS solamente con F2 además de la constante se obtenía

$$\underset{(37,285)}{\widehat{\mathrm{P}} = 52,3509} + \underset{(0,018733)}{0,138750\,\mathrm{F2}}$$

$$T = 14 \quad \bar{R}^2 = 0,8056 \quad F(1,12) = 54,861 \quad \hat{\sigma} = 39,023$$

(Desviaciones típicas entre paréntesis)

[4] Usando las observaciones 1 - 14, la matriz de correlaciones entre BEDRMS, BATHS y $F2$ es

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

y las medias muestrales de BEDRMS y BATHS son:

Variable	Media
BEDRMS	3,64286
BATHS	2,35714

Dadas esas condiciones, las dos últimas ecuaciones no dependen de $\hat{\beta}_1$ ni de $\hat{\beta}_2$ y las dos primeras ecuaciones normales coinciden con las que se obtenían en el Tema 2 para el modelo de regresión lineal simple. Por lo tanto, en ese caso se obtendría la misma solución para $\hat{\beta}_1$ y $\hat{\beta}_2$ que en el MRLS inlcuyendo solamente el término constante y $F2 \equiv X_2$ y entonces las mismas estimaciones de esos coeficientes. Por lo tanto, en general no da lo mismo incluir o no otras variables en el modelo a la hora de estimar el efecto de una variable sobre la variable dependiente.

Interpretación de los coeficientes estimados.
El coeficiente estimado que acompaña a la variable F2, variable que recoge el tamaño total de la vivienda, es positivo y parece ser el signo adecuado. Si consideramos dos viviendas con el mismo número de baños y habitaciones, parece razonable pensar que aquella con mayor área habitable tenga un precio mayor. Esto indica que las habitaciones serán más grandes.

Los signos de los coeficientes asociados a BEDRMS y BATHS son negativos. Podemos pensar que si aumenta el número de habitaciones o el número de baños, esto indicaría una vivenda más lujosa y por lo tanto debería de aumentar el valor de la vivienda. Pero hay que tener en cuenta que a la hora de interpretar un coeficiente de regresión asociado a uno de los regresores estamos manteniendo constante el resto de variables explicativas.

Si la misma superficie habitable se tiene que dividir para poder incluir una nueva habitación, el resultado será que cada habitación será más pequeña. El signo del coeficiente estimado indica que un comprador medio valora negativamente tener más habitaciones a costa de un menor tamaño de éstas. Lo mismo se puede interpretar en el caso del coeficiente que acompaña a BATHS.

Interpretación de los coeficientes estimados:

- El coeficiente estimado $\hat{\beta}_1 = 129,062$ indica el precio medio estimado en miles de euros, de aquellas viviendas que no tienen ningún pie cuadrado de área habitable, ni habitaciones ni baños.

- El coeficiente estimado $\hat{\beta}_2 = 0,154800$:

  Considerando dos casas con el mismo número de habitaciones y de baños, para aquella casa que tenga un pie cuadrado más de área habitable se estima que en media su precio de venta se incremente en 154.800 dólares.

- El coeficiente estimado $\hat{\beta}_3 = -21,5875$:

  Si aumenta el número de habitaciones, manteniendo constante el tamaño de la vivienda y el número de baños, el precio medio se estima disminuirá en 21.588 dólares.

- El coeficiente $\hat{\beta}_4 = -12,1928$:

  Manteniendo el tamaño de la vivienda y el número de habitaciones constante, añadir un baño completo más significa tener habitaciones más pequeñas, por lo que el precio medio se estima disminuirá en 12.193 dólares.

¿Se mantendría el signo del coeficiente que acompaña a BEDRMS si no incluimos la variable F2 ni BATHS?
Pues seguramente no, porque en ese caso no estamos controlando por esa variable en la regresión, y como hemos visto F2 y BEDRMS están correlacionados. Por lo tanto más habitaciones implicaría mayor superficie de piso, y por lo tanto más precio en media. Lo mismo ocurriría si solamente incluimos BATHS. Ahora bien, ¿qué ocurriría si excluimos solamente F2 y dejamos las otras dos variables explicativas? Veremos las implicaciones que tiene omitir o no controlar por variables relevantes en un tema posterior.

Estimación del incremento medio en el precio de la vivienda ante cambios en las variables explicativas.
Utilizando los resultados (3.3) de la estimación del modelo (3.1), si manteniendo el número de baños tenemos dos habitaciones más y aumenta el área habitable en 500 pies cuadrados, el cambio en el precio medio estimado de una vivienda será de 34.224 dólares, esto es

$$\widehat{\triangle \mathrm{P}}_i = 0,1548 \triangle \mathrm{F2}_i - 21,588 \triangle \mathrm{BEDRMS}_i = (0,1548 \times 500) - (21,588 \times 2) = 34,224$$

3.3.2. Desviaciones típicas e intervalos de confianza

Por el momento nos hemos centrado en la interpretación de las estimaciones puntuales. Pero también tenemos que tener en cuenta que estas estimaciones son realizaciones muestrales de un estimador, que es una variable aleatoria. Por lo tanto, pueden estar sujetas a variación muestral ya que distintas muestras puedan dar lugar a distintas realizaciones muestrales. Estas estimaciones de un mismo vector de parámetros β estarán distribuidas con mayor o menor variación alrededor de su valor poblacional siguiendo cierta distribución de probabilidad.

Bajo las hipótesis básicas que hemos enumerado al principio de este tema, el valor poblacional del vector de parámetros β es la media de la distribución ya que $\hat{\beta}_{MCO}$ es un estimador insesgado. Su distribución es una Normal y la matriz de varianzas y covarianzas viene dada por la expresión $V(\hat{\beta}_{MCO}) = \sigma^2(X'X)^{-1}$. Esto se suele denotar como

$$\hat{\beta}_{MCO} \sim N(\beta, \sigma^2(X'X)^{-1}) \tag{3.4}$$

La varianza de las perturbaciones, σ^2, es un parámetro desconocido. Un estimador insesgado de la misma bajo las hipótesis básicas es

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N-K}$$

donde $\hat{u} = Y - X\hat{\beta}_{MCO}$ es el vector de residuos. El programa, en la ventana gretl:modelo1 muestra las realizaciones muestrales de la suma de cuadrados de los residuos (SCR), $\hat{u}'\hat{u} = 16700,1$ y de la desviación típica de los residuos $\sqrt{\hat{\sigma}^2} = 40,8657$.

Un estimador insesgado, bajo las hipótesis básicas, de la matriz de varianzas y covarianzas de $\hat{\beta}_{MCO}$ es

$$\hat{V}(\hat{\beta}_{MCO}) = \hat{\sigma}^2(X'X)^{-1}$$

En la ventana de resultados de la estimación del modelo por MCO, **gretl:modelo1**, podemos obtener la realización muestral de este estimador $\hat{V}(\hat{\beta}_{MCO}) = \hat{\sigma}^2(X'X)^{-1}$ eligiendo:

Análisis → Matriz de covarianzas de los coeficientes

Se abre una nueva ventana, **gretl:covarianzas de los coeficientes**, donde se muestra la estimación de las varianzas (elementos de la diagonal principal) y covarianzas (elementos fuera de la diagonal principal) de los coeficientes de regresión $\hat{\beta}$, como se muestra en la Tabla 3.2. Dado que es una matriz simétrica, solamente aparecen los valores por encima de la diagonal principal. La raíz cuadrada de los elementos de la diagonal principal son los mismos

Matriz de covarianzas de los coeficientes

const	F2	BEDRMS	BATHS	
7797,47	0,670891	−1677,1	−1209,3	const
	0,00102019	−0,0754606	−0,995066	F2
		730,585	−356,40	BEDRMS
			1870,56	BATHS

Tabla 3.2: Modelo (3.1). Estimación de la matriz de covarianzas de $\hat{\beta}$

valores que los mostrados en la tercera columna de la ventana **gretl:modelo1**. Por ejemplo, la varianza estimada del coeficiente $\hat{\beta}_2$ asociado a F2 es $\widehat{var}(\hat{\beta}_2) = 0,00102019$ y su raíz cuadrada es su desviación típica estimada $\widehat{des}(\hat{\beta}_2) = 0,0319404$.

También podemos obtener estimaciones de las covarianzas entre los coeficientes estimados. Por ejemplo, la covarianza estimada entre los coeficientes $\hat{\beta}_2$ asociado a F2 y $\hat{\beta}_4$ asociado a $BATHS$ es igual a $\hat{cov}(\hat{\beta}_2, \hat{\beta}_4) = -0,995066$.

Intervalos de confianza:

Seguidamente vamos a ver cómo podemos obtener intervalos de confianza para cada coeficiente individual. ¿Qué nos indican estos intervalos? ¿Cuál es su utilidad?

Bajo las hipótesis básicas, se puede demostrar que la variable aleatoria

$$\frac{\hat{\beta}_j - \beta_j}{\widehat{des}(\hat{\beta}_j)} \sim t(N-K) \tag{3.5}$$

donde $\widehat{des}(\hat{\beta}_j)$ es la desviación típica estimada del estimador $\hat{\beta}_j$ y $t(N-K)$ denota la distribución t de Student de $(N-K)$ grados de libertad. Esto es válido para cualquiera de los coeficientes β_j , $j = 1, \ldots, K$.

Denotamos por $c = t_{(N-K)\alpha/2}$ la ordenada de la distribución t de Student con $N-K$ grados de libertad, tal que deja a la derecha una probabilidad de $\alpha/2$, esto es $P(t > c) = \alpha/2$. Esto implica que:

$$Pr\left(-c \leq \frac{\hat{\beta}_j - \beta_j}{\widehat{des}(\hat{\beta}_j)} \leq c\right) = Prob\left(\hat{\beta}_j - c\,\widehat{des}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c\,\widehat{des}(\hat{\beta}_j)\right) = 1-\alpha \tag{3.6}$$

Por lo tanto, un intervalo de confianza del $(1-\alpha)$ por ciento para un coeficiente cualquiera β_j viene dado por

$$IC(\beta_j)_{1-\alpha} = \left[\hat{\beta}_j \pm c\,\widehat{des}(\hat{\beta}_j)\right]$$

El cálculo de los intervalos de confianza para los coeficientes de regresión del modelo se conoce con el nombre de **estimación por intervalo**. Un intervalo de confianza nos dice que, con

probabilidad $(1-\alpha)$ se estima que el parámetro β_j estará dentro de ese rango de valores. Este intervalo puede ser demasiado amplio, y esto dependerá de la precisión con la que estimemos los parámetros recogido en $\widehat{des}(\hat{\beta}_j)$. Es importante tener en cuenta que la validez de estos intervalos de confianza depende de que se satisfagan las hipótesis básicas.

Siguiendo con el ejemplo del modelo (3.1) para el precio de la vivienda, Gretl nos permite obtener directamente los intervalos de confianza del 95 por ciento para los coeficientes. El resultado mostrado en la Tabla 3.3 se obtiene eligiendo en la ventana **gretl:modelo1**

Análisis → Intervalos de confianza para los coeficientes

Variable	Coeficiente	Intervalo de confianza 95 %	
		bajo	alto
const	129,062	−67,690	325,814
F2	0,154800	0,0836321	0,225968
BEDRMS	−21,587	−81,812	38,6376
BATHS	−12,192	−108,56	84,1742

Tabla 3.3: Modelo (3.1): Estimación por intervalo de los coeficientes.

A su vez, utilizando los resultados mostrados en la ventana **gretl:modelo1**

Modelo 1: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	129,062	88,3033	1,4616	0,1746
F2	0,154800	0,0319404	4,8465	0,0007***
BEDRMS	−21,587	27,0293	−0,7987	0,4430
BATHS	−12,192	43,2500	−0,2819	0,7838

podemos obtener intervalos de confianza para cada uno de los coeficientes, dado un nivel de confianza $(1-\alpha)$, por ejemplo del 95 por ciento[5]. Los intervalos de confianza obtenidos son:

$$\begin{aligned} \beta_1: &\quad 129,0620 \pm (2,228 \times 88,3033) \\ \beta_2: &\quad 0,1548 \pm (2,228 \times 0,0319404) \\ \beta_3: &\quad -21,5875 \pm (2,228 \times 27,0293) \\ \beta_4: &\quad -12,1928 \pm (2,228 \times 43,2500) \end{aligned}$$

El intervalo de confianza además se puede utilizar para contrastar la hipótesis de que el parámetro β_j tome determinado valor. Si el valor del parámetro bajo la hipótesis nula

[5] Al 95 por ciento de confianza, $(\alpha/2 = 0,025)$, el valor en las tablas de la distribución t de Student con 10 grados de libertad es $c = t_{(10)0,025} = 2,228$. Recordar que Gretl permite acceder a algunos valores tabulados de distintas distribuciones, Normal, t-Student, Chi-cuadrado, F de Snedecor. En la ventana principal **gretl** en *Herramientas → Tablas estadísticas*. En el caso de la t de Student hay que introducir los grados de libertad (gl). Los valores mostrados corresponden a los valores de $\alpha/2$ de 0,10-0,05-0,025-0,01-0,001.

está dentro del intervalo de confianza, no podemos rechazar esa hipótesis al nivel de significación α. Dada la muestra y nuestra especificación del modelo, no podemos rechazar con una confianza del 95 por ciento, excepto para el parámetro asociado a F2, que el coeficiente asociado a cada una de estas variables sea igual a cero ya que este valor está dentro del intervalo de confianza. ¿Quiere decir entonces que el valor poblacional de cada uno de esos parámetros es cero? La respuesta es NO, ya que por esa misma regla de tres el parámetro β_j debería de tomar cada uno de los valores en el intervalo.

3.3.3. Significatividad individual y conjunta

Contrastes de significatividad individual

Uno de los principales objetivos de un primer análisis de regresión es la de contrastar si son o no estadísticamente relevantes los factores que hemos considerado como explicativos de la variable dependiente en cuestión, dada la especificación de nuestro modelo. Podemos considerar individualmente cada regresor y contrastar:

$$H_0\colon\ \beta_j = 0$$
$$H_a\colon\ \beta_j \neq 0$$

donde la hipótesis nula implica que, dada la especificación del modelo una vez se ha controlado por el resto de factores incluidos como variables explicativas, el efecto marginal de la variable X_j sobre el valor medio de la variable dependiente es cero.

Dado que en la hipótesis alternativa se contempla la posibilidad de que el coeficiente, de ser distinto de cero, pueda ser indistintamente negativo o positivo, el contraste es a dos colas. Normalmente en estos contrastes, conocidos con el nombre de contrastes de significatividad individual, se considera esta alternativa.

El estadístico de contraste y su distribución bajo la hipótesis nula es:

$$t_j = \frac{\hat{\beta}_j}{\widehat{des}(\hat{\beta}_j)} \overset{H_0}{\sim} t(N-K) \tag{3.7}$$

Una vez obtenido el valor muestral del estadístico, t_j^m, ¿cómo decidimos si rechazar o no la hipótesis nula?

- Se elige un nivel de significación α que indicaría nuestra elección de la probabilidad de error de tipo I (rechazar la hipótesis nula cuando esta fuera cierta) o tamaño del contraste. Obtenemos el valor crítico o umbral $c = t_{(N-K)\alpha/2}$ tal que $Pr(t_j > c) = \alpha/2$.
- Rechazamos la hipótesis nula a un nivel de significación α, si en valor absoluto la realización muestral del estadístico es mayor que el valor crítico $|t_j^m| > c$. No rechazamos la hipótesis nula en caso contrario.

Si no se rechaza la hipótesis nula, en el lenguaje econométrico se dice que la variable que acompaña al coeficiente en cuestión no es significativa o que el coeficiente no es significativamente distinto de cero al α por ciento de significación. Si por el contrario se rechaza la hipótesis nula, se dice que la variable es significativa o que el coeficiente es significativamente distinto de cero.

Otra forma de llevar a cabo el contraste es utilizar el *valor-p*. Este valor es una probabilidad e indica cuál sería el menor nivel de significación que se tendría que elegir para rechazar la hipótesis nula, dada la realización muestral del estadístico. Si el contraste es a dos colas, el *valor-p* es dos veces el área a la derecha de la realización muestral del estadístico en valor absoluto, en la distribución de éste bajo la hipótesis nula, esto es

$$\textit{valor-p} = 2\ \Pr(\mathrm{t_j} > \mathrm{t_j^m}|\mathrm{H_0})$$

Si el contraste es a una cola, el *valor-p* sería el área a la derecha de la realización muestral del estadístico en valor absoluto, en la distribución de éste bajo la hipótesis nula, esto es $Pr(t_j > t_j^m|H_0)$. A mayor *valor-p*, mayor sería la probabilidad de error de tipo I si elegimos rechazar la hipótesis nula. Luego a mayor *valor-p* menor evidencia contra la hipótesis nula y por el contrario a menor *valor-p* mayor evidencia contra la hipótesis nula.

¿Cuál será la regla de decisión del contraste mirando al *valor-p*?

Rechazar la hipótesis nula si el *valor-p* es menor que el nivel de significación elegido y no rechazarla en caso contrario.

Esta es exactamente la misma regla de decisión que antes. Elegido un nivel de significación, si el valor muestral es mayor en valor absoluto que el valor crítico c, querrá decir que dos veces la probabilidad que deja a la derecha el valor muestral es más pequeño que ese nivel de significación.

Siguiendo con nuestro ejemplo, vamos a comentar qué nos indican la cuarta y quinta columna que aparecían en la ventana de resultados de la estimación por MCO del modelo (3.1) **gretl:modelo1**.

Modelo 1: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	129,062	88,3033	1,4616	0,1746
F2	0,154800	0,0319404	4,8465	0,0007***
BEDRMS	−21,587	27,0293	−0,7987	0,4430
BATHS	−12,192	43,2500	−0,2819	0,7838

Los valores obtenidos en la cuarta columna se obtienen de dividir los correspondientes valores de la segunda y tercera columnas esto es, la estimación del coeficiente dividida por su desviación típica estimada. Esta sería la realización muestral del estadístico t_j bajo la hipótesis nula de que el valor poblacional del parámetro β_j asociado a esa variable es igual a cero.

La quinta columna es el *valor-p* asociado a cada coeficiente, siendo el contraste de significatividad individual a dos colas. Habitualmente se eligen como niveles de significación el 1 %, 5 % y 10 % siendo el 5 % el más utilizado. Gretl indica con uno, dos o tres asteriscos cuando se rechaza la hipótesis nula al 10 %, al 5 %, o al 1 % respectivamente.

En este caso solamente es significativa la variable F2 al 1 % y se indica con tres asteriscos. El *valor-p* asociado a esta variable es más pequeño que 0,01 y por lo tanto que 0,05 y que 0,1.

Para el resto de coeficientes no se rechazaría la hipótesis nula. Los coeficientes asociados al término constante, BEDRMS y BATHS no serían significativamente distintos de cero ni

siquiera al 10 %. El *valor-p* asociado es mayor que 0,1. Estos valores oscilan entre 0,175 y 0,784 por lo que, si rechazásemos la hipótesis nula de que cada uno de estos coeficientes es cero, habría desde un 17,5 a un 78,4 por ciento de probabilidad de cometer el error de rechazar esa hipótesis siendo cierta.

Si miramos a los valores críticos en cada uno de estos niveles de significación tenemos que:

$$\begin{array}{ll} \alpha = 0,01 & t_{(10)0,005} = 3,169 \\ \alpha = 0,05 & t_{(10)0,025} = 2,228 \\ \alpha = 0,1 & t_{(10)0,05} = 1,812 \end{array}$$

Excepto en el caso de la variable F2, el valor muestral de los estadísticos t_j en valor absoluto es más pequeño que cualquiera de estos valores críticos. Por lo tanto solamente se rechaza la hipótesis nula de que el coeficiente asociado a la variable SQFT sea igual a cero. Esto parece indicar que dado que el número de habitaciones y de baños está ya recogido en el tamaño de la vivienda, una vez incluimos esta variable el tener más o menos habitaciones o baños no tiene un efecto marginal significativo en el precio medio de ésta. Lo normal es tener una vivienda con un número de habitaciones y baños proporcional a su tamaño.

Esto mismo concluimos mirando a los intervalos de confianza, aunque en ese caso el nivel de significación elegido sólo fue del 5 por ciento.

Contraste de significación conjunta

Otro estadístico que se muestra en la ventana de resultados de la estimación es el valor del estadístico F (3, 10) = 16,9889 con *valor-p* = 0, 000299. ¿Cómo se calcula este estadístico? ¿Qué hipótesis nula se está contrastando?

La hipótesis nula que se está contrastando es que conjuntamente todos lo coeficientes, excepto el asociado al término constante, sean cero. En nuestro ejemplo en concreto

$$\begin{array}{l} H_0\colon\ \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a\colon\ \text{alguna de las igualdades no se cumple} \end{array}$$

Este estadístico se puede considerar como un contraste general de bondad de ajuste del modelo. Si la hipótesis nula no se rechaza podemos concluir que ninguna de las variables en conjunto puede explicar la **variación** en el precio de la vivienda. Esto significa que es un modelo muy pobre y que debiera de ser reformulado.

Estamos excluyendo de la hipótesis nula el parámetro que acompaña al término constante. El modelo bajo la hipótesis nula, al que llamaremos Modelo Restringido es:

$$\texttt{Modelo Restringido} \qquad P_i = \beta_1 + u_i \qquad\qquad i = 1, 2, \ldots, N \tag{3.8}$$

Este modelo incluye solamente un término constante como regresor y le compararemos con el Modelo No Restringido (3.1). El estimador MCO del parámetro β_1 en el modelo restringido es aquél que

$$\min_{\hat{\beta}_1} \sum_{i=1}^{N} (Y_i - \hat{\beta}_1)^2$$

En este caso tenemos solamente un parámetro a estimar por lo que sólo hay una ecuación normal,

$$\sum_i Y_i = N\hat{\beta}_1 \tag{3.9}$$

cuya solución es

$$\hat{\beta}_{1,R} = \frac{1}{N}\sum_i Y_i = \bar{Y}$$

El coeficiente estimado que acompaña al término constante nos recoge simplemente la media muestral de la variable dependiente. El residuo correspondiente al modelo restringido es $\hat{u}_{i,R} = Y_i - \hat{\beta}_{1,R} = Y_i - \bar{Y}$, por lo que la suma de cuadrados residual coincide con la suma de cuadrados total o variación total de la variable dependiente. Esto implica que la suma de cuadrados explicada o variación explicada con la estimación de este modelo (3.8) es nula

$$SCR_R = \sum_i \hat{u}_{i,R}^2 = \sum_i (Y_i - \overline{Y})^2 = SCT \qquad \Rightarrow \qquad SCE_R = 0$$

Por último, y teniendo en cuenta como se define el coeficiente de determinación R^2

$$R^2 = 1 - \frac{\sum_i \hat{u}_i^2}{\sum_i (Y_i - \bar{Y})^2}$$

para este modelo el coeficiente de determinación es igual a cero[6]. Dado que en el modelo solamente incluimos un regresor que no varía, éste no puede explicar **variación** o varianza de la variable dependiente. Si estimamos con Gretl el modelo (3.8) obtenemos los siguientes resultados:

Modelo 2: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	317,493	23,6521	13,4234	0,0000

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	101815,
Desviación típica de los residuos ($\hat{\sigma}$)	88,4982
R^2	0,000000
$\bar{R}^2$ corregido	0,000000
Grados de libertad	13
Log-verosimilitud	−82,108
Criterio de información de Akaike	166,216
Criterio de información Bayesiano de Schwarz	166,855
Criterio de Hannan–Quinn	166,157

[6]Esto es así dado que $\sum_i \hat{u}_{i,R}^2 = \sum_i (Y_i - \bar{Y})^2 \Rightarrow R_R^2 = 1 - \frac{\sum_i \hat{u}_{i,R}^2}{\sum_i (Y_i - \bar{Y})^2} = 1 - 1 = 0$.

Podemos comprobar que la estimación del coeficiente que acompaña al término constante coincide con la media muestral de la variable dependiente ($\bar{P} = 317,493$). La desviación típica de los residuos coincide con la desviación típica de la variable dependiente, ya que la suma de cuadrados residual coincide con la suma de cuadrados total, $SCR_R = \sum_i \hat{u}_{i,R}^2 = \sum_i (Y_i - \overline{Y})^2 = 101815$, y también los grados de libertad de ambas, $T - K = T - 1 = 13$. Por lo tanto,

$$\sqrt{\frac{\sum_i \hat{u}_{i,R}^2}{13}} = \sqrt{\frac{\sum_i (Y_i - \overline{Y})^2}{13}} = 88,4982$$

Por último, el coeficiente de determinación R^2 es igual a cero.

Un estadístico general de contraste de restricciones lineales es aquél que compara las sumas de cuadrados de residuos de la estimación del modelo restringido y del modelo no restringido, teniendo en cuenta los grados de libertad en la estimación de cada modelo,(gl_R) y (gl_{NR}) respectivamente[7]

$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-K)} \overset{H_0}{\sim} \mathcal{F}(q,\, N-K) \tag{3.10}$$

donde $q = (gl_R - gl_{NR})$ es el número de restricciones bajo la hipótesis nula y $N - K = gl_{NR}$. Si dividimos numerador y denominador por la suma de cuadrados total SCT y utilizamos los siguientes resultados:

a) $1 - R^2 = SCR_{NR} \,/\, SCT$ y en este caso $1 - R_R^2 = 1 - 0 = 1$.

b) $gl_R - gl_{NR} = (N-1) - (N-K) = K - 1$ que es el número de restricciones bajo la hipótesis nula.

el estadístico general (3.10) nos queda para este contraste en concreto igual a

$$F = \frac{R^2/(K-1)}{(1-R^2)/(N-K)} = \frac{R^2}{(1-R^2)}\frac{(N-K)}{(K-1)} \overset{H_0}{\sim} \mathcal{F}(K-1,\, N-K) \tag{3.11}$$

En nuestro ejemplo sobre el precio de la vivienda, $K - 1 = 3$ que es el número de restricciones bajo la hipótesis nula y $N - K = 14 - 4 = 10$. Dado el resultado mostrado $F(3, 10) = 16,9889$ (valor p $= 0,000299$), si consideramos el valor-p se rechazaría la hipótesis nula a cualquier nivel de significación razonable, en particular al $\alpha = 0,05$ ya que este valor es mayor que el *valor-p* obtenido. Si utilizamos el valor crítico $\mathcal{F}_{(3,10)0,05} = 3,71$ obtenemos el mismo resultado ya que el valor muestral del estadístico es mayor que el valor crítico. Esto indica que al menos uno de los coeficientes, aparte del asociado al término constante, es distinto de cero.

Aunque hemos utilizado en esta sección el coeficiente de determinación en relación al estadístico de significación conjunta, en la siguiente sección vamos a hablar de su utilización junto con el coeficiente de determinación corregido y otros estadísticos para la selección entre distintos modelos.

[7] En temas posteriores veremos la utilización de este estadístico para contrastar otro tipo de restricciones lineales.

3.4. Bondad de ajuste y selección de modelos

En los temas anteriores se ha presentado el coeficiente de determinación como una medida de bondad de ajuste que es invariante a unidades de medida[8]. Este coeficiente se define como la proporción de variación explicada por la regresión del total de variación a explicar en la muestra de la variable dependiente. Si hay término constante en el modelo,

$$R^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} = 1 - \frac{\sum_i \hat{u}_i^2}{\sum_i (Y_i - \bar{Y})^2} \qquad 0 \leq R^2 \leq 1$$

Este indicador tiene que ser considerado como uno más a tener en cuenta a la hora de valorar si un modelo es adecuado, pero no debemos darle más importancia de la que tiene. Obtener un valor del R^2 cercano a 1 no indica que nuestros resultados puedan ser fiables. Por ejemplo, podemos tener problemas de no satisfacerse alguna hipótesis básica y nuestra inferencia no ser válida.

Por otro lado, obtener un valor más o menos alto del coeficiente de determinación puede estar influido por el tipo de datos que estemos analizando. Normalmente con datos de series temporales, donde las variables pueden presentar tendencias similares en el tiempo, es fácil obtener R^2 altos, mientras que con datos de sección cruzada eso no suele ocurrir ya que normalmente las variables presentan mayor dispersión.

Por otro lado, si queremos utilizar el R^2 para comparar distintos modelos, estos deben de tener la misma variable dependiente ya que así tendrán igual suma de cuadrados total. Aún así, esta medida adolece del problema de aumentar su valor al añadir una nueva variable explicativa, sea cual sea su aportación al modelo. Además no tiene en cuenta que hay que estimar un nuevo parámetro con el mismo número de observaciones.

Para tener en cuenta este problema se suele utilizar el R^2 corregido por grados de libertad. Esta medida tiene en cuenta los grados de libertad tanto de la suma de cuadrados residual, $(N - K)$, como de la suma de cuadrados total, $(N - 1)$. Se define como

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(N-K)}{\sum (Y_i - \bar{Y})^2/(N-1)} = 1 - \frac{N-1}{N-K}(1 - R^2) \qquad -\infty < \bar{R}^2 \leq R^2$$

El $\bar{R}^2$ puede disminuir si el incluir una nueva variable no compensa la pérdida de grados de libertad al tener que estimar un nuevo parámetro[9]. El coeficiente de determinación corregido $\bar{R}^2$ no tomará valores mayores que el R^2 pero sí puede tomar valores negativos. Esto último indicaría que el modelo no describe adecuadamente el proceso que ha generado los datos.

Hasta el momento hemos ido comentado los resultados que normalmente se muestran en la estimación de un modelo. Una forma de presentarlos es la siguiente:

$$\underset{(estad.\,t)}{\widehat{\mathrm{P}}} = \underset{(1{,}462)}{129{,}062} + \underset{(4{,}847)}{0{,}154800}\,\mathrm{F2} - \underset{(-0{,}799)}{21{,}5875}\,\mathrm{BEDRMS} - \underset{(-0{,}282)}{12{,}1928}\,\mathrm{BATHS}$$

$$N = 14 \quad R^2 = 0{,}8359 \quad \bar{R}^2 = 0{,}7868 \quad F(3, 10) = 16{,}989$$

[8] Esto no ocurre con otras medidas como puede ser la desviación típica de los residuos, $\hat{\sigma} = \sqrt{SCR/N-K)}$ ya que la suma de cuadrados de los residuos no es invariante a un cambio de escala en las variables.

[9] Se puede demostrar que si el valor absoluto del estadístico t de significatividad individual asociado a una variable es menor que la unidad, eliminar esta variable del modelo aumentará el $\bar{R}^2$ mientras que si es mayor que la unidad lo reducirá.

Una alternativa a presentar los estadísticos t de significatividad individual, aunque suele ser lo más habitual, es mostrar las desviaciones típicas estimadas de los coeficientes o los valores p correspondientes.

Otros criterios de selección de modelos que muestra Gretl son los criterios de información de Akaike (AIC), Bayesiano de Schwarz (BIC) y de Hannan-Quinn (HQC). Estos criterios se calculan en función de la suma de cuadrados residual y de algún factor que penalice por la pérdida de grados de libertad. Un modelo más complejo, con más variables explicativas, reducirá la suma de cuadrados residual pero aumentará el factor de penalización. Utilizando estos criterios se escogería aquel modelo con un menor valor de AIC, BIC o HQC. Normalmente no suelen dar la misma elección, siendo el criterio AIC el que elige un modelo con mayor número de parámetros.

Selección de un modelo para el precio de la vivienda.

Vamos a continuar con nuestro ejemplo sobre el precio de la vivienda y comparar distintas especificaciones, para seleccionar una especificación entre varias propuestas. Para ello, utilizamos distintos indicadores que hemos visto hasta ahora, significatividad individual, conjunta, coeficientes de determinación y criterios de información. Podemos considerar que estos indicadores nos ayudan a valorar la especificación en términos de la contribución de las variables explicativas incluidas en el modelo[10].

Vamos a estimar las siguientes especificaciones o modelos alternativos para explicar el precio de la vivienda:

Modelo 1 $P_i = \beta_1 + \beta_2\, F2_i + u_i$

Modelo 2 $P_i = \beta_1 + \beta_2\, F2_i + \beta_3\, BEDRMS_i + u_i$

Modelo 3 $P_i = \beta_1 + \beta_2 F2_i + \beta_3 BEDRMS_i + \beta_4 BATHS_i + u_i$

Modelo 4 $P_i = \beta_1 + \beta_3\, BEDRMS_i + \beta_4\, BATHS_i + u_i$

Estos cuatro modelos difieren en las variables explicativas incluidas. El **Modelo 3** es el más general e incluye al resto de modelos. Esto quiere decir que cada uno de los restantes se obtiene imponiendo una o más restricciones sobre los coeficientes de este modelo. En este caso son restricciones de exclusión, es decir que algún coeficiente o coeficientes son iguales a cero. A este tipo de modelos se les llama modelos anidados. Los resultados de la estimación del **Modelo 3** con Gretl son los siguientes:

Modelo 3: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	129,062	88,3033	1,4616	0,1746
F2	0,154800	0,0319404	4,8465	0,0007
BEDRMS	−21,587	27,0293	−0,7987	0,4430
BATH	−12,192	43,2500	−0,2819	0,7838

[10]Estos no son los únicos indicadores. Por ejemplo, analizar el gráfico de residuos o utilizar diversos contrastes de algunas de las hipótesis básicas son elementos importantes a la hora de evaluar los resultados de la especificación y estimación de un modelo.

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	16700,1
Desviación típica de los residuos ($\hat{\sigma}$)	40,8657
R^2	0,835976
$\bar{R}^2$ corregido	0,786769
$F(3, 10)$	16,9889
valor p para $F()$	0,000298587
Log-verosimilitud	−69,453
Criterio de información de Akaike	146,908
Criterio de información Bayesiano de Schwarz	149,464
Criterio de Hannan–Quinn	146,671

El **Modelo 1** es el más reducido y también está incluido en los modelos 2 y 3, no así en el 4. Estos son los resultados de su estimación:

Modelo 1: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	52,3509	37,2855	1,4041	0,1857
F2	0,138750	0,0187329	7,4068	0,0000

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	18273,6
Desviación típica de los residuos ($\hat{\sigma}$)	39,0230
R^2	0,820522
$\bar{R}^2$ corregido	0,805565
Grados de libertad	12
Log-verosimilitud	−70,084
Criterio de información de Akaike	144,168
Criterio de información Bayesiano de Schwarz	145,447
Criterio de Hannan–Quinn	144,050

El **Modelo 2** está anidado en el 3. Los resultados de la estimación de este modelo se muestran a continuación:

Modelo 2: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	121,179	80,1778	1,5114	0,1589
F2	0,148314	0,0212080	6,9933	0,0000
BEDRMS	−23,910	24,6419	−0,9703	0,3527

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	16832,8
Desviación típica de los residuos ($\hat{\sigma}$)	39,1185
R^2	0,834673
$\bar{R}^2$ corregido	0,804613
$F(2, 11)$	27,7674
valor p para $F()$	5,02220e-05
Log-verosimilitud	−69,509
Criterio de información de Akaike	145,019
Criterio de información Bayesiano de Schwarz	146,936
Criterio de Hannan–Quinn	144,841

Finalmente el **Modelo 4** solamente está anidado en el modelo 3. Los resultados de la estimación por MCO son:

Modelo 4: estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: P

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	27,2633	149,652	0,1822	0,8588
BEDRMS	−10,137	46,9811	−0,2158	0,8331
BATHS	138,795	52,3450	2,6515	0,0225

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	55926,4
Desviación típica de los residuos ($\hat{\sigma}$)	71,3037
R^2	0,450706
$\bar{R}^2$ corregido	0,350834
$F(2, 11)$	4,51285
valor p para $F()$	0,0370619
Log-verosimilitud	−77,914
Criterio de información de Akaike	161,829
Criterio de información Bayesiano de Schwarz	163,746
Criterio de Hannan–Quinn	161,651

Comparación de los resultados para los modelos 1,2 y 3.

- Se observa que a medida que se introducen más variables explicativas, la suma de cuadrados residual va disminuyendo y el coeficiente de determinación R^2 aumenta.

- En términos del coeficiente de determinación R^2, en el **Modelo 1** el tamaño de la vivienda (F2) explica el $82,1\,\%$ de la variación en los precios de la vivienda, pasando a ser de un $83,6\,\%$ al incluir el número de habitaciones (BEDRMS) y número de baños (BATHS).

- A medida que se incluyen más variables explicativas, primero BEDRMS y luego BATHS, el coeficiente de determinación corregido $\bar{R}^2$ disminuye y la desviación típica de los residuos aumenta[11]. Esto indica que la ganancia en un mayor valor del R^2 o menor suma de cuadrados residual no se compensa en ningún caso por la pérdida de grados de libertad.

- En cuanto a la significatividad individual, en los tres modelos la única variable significativa a los niveles de significación habituales es F2[12]. Así, una vez hemos controlado por el tamaño de la vivienda, las variables BEDRMS y BATHS no afectan significativamente el precio de la vivienda.

- El estadístico F de significación conjunta señala en los tres casos no aceptar la hipótesis nula de que todos los coeficientes excepto el asociado al término constante son igual a cero. Al menos hay un coeficiente que es significativamente distinto de cero. Por lo obtenido en los contrastes de significatividad individual, sabemos que éste es el coeficiente que acompaña a F2.

  Si nos fijamos, a medida que vamos del **Modelo 1** al **3**, el valor muestral del estadístico F disminuye. Esto es lógico, ya que este estadístico es función del R^2 pero también de los grados de libertad. Otra vez estaría recogiendo que, a medida que aumenta el número de parámetros a estimar K, las diferencias en R^2 son demasiado pequeñas para compensar la disminución en el ratio $(N-K)/(K-1)$. Ahora bien, en general, las diferencias en el estadístico F no son relevantes. Lo que es de interés es el resultado del contraste.

- Si consideramos los criterios de información AIC, BIC y HQC, de los tres modelos el elegido es el **Modelo 1**, reafirmando lo que indica el $\bar{R}^2$. La ganancia en un mejor ajuste, o una menor suma de cuadrados residual, no es suficiente para compensar el factor que penaliza en función de grados de libertad.

  Dado que el tamaño de la vivienda depende del número de habitaciones y de baños, este resultado parece indicar que una vez se controla por F2 indirectamente esta variable incluye casi todo lo que pueden aportar BEDRMS y BATHS.

¿Qué ocurre con el Modelo 4?

En este modelo no hemos incluido la variable F2, que en el análisis anterior era la variable que más explica el precio de la vivienda y hemos dejado las variables que no eran significativas una vez que incluíamos esta variable. Podríamos argumentar que de esta forma se podría analizar el efecto de BEDRMS y BATHS, ya que F2 parecía recoger la información relevante de estas dos variables.

Si lo comparamos con el **Modelo 3**, que es en el que está anidado el **Modelo 4**, se obtiene menor valor de R^2 y $\bar{R}^2$, mayor valor de AIC, BIC y HQC, mayor suma de cuadrados residual y mayor desviación típica de los residuos. Todos ellos señalan en la misma dirección siendo, en términos de estos criterios, peor modelo el 4. Vemos que el omitir F2 empeora mucho

[11] Notar que los estadísticos t asociados a cada coeficiente son menores que uno en valor absoluto.

[12] Por ejemplo, con nivel de significación del 5 por ciento los valores críticos serían para el modelo **1** $t_{(12)0,025} = 2,179$, para el **Modelo 2** $t_{(11)0,025} = 2,201$ y para el **Modelo 3** $t_{(10)0,025} = 2,228$.

el ajuste sin compensar por la ganancia en grados de libertad. Además cambia sustancialmente la estimación y la significatividad del coeficiente que acompaña a BATHS, pasando la estimación de signo positivo a negativo y ser significativamente distinto de cero al 5 % de significación. ¿Qué puede estar ocurriendo? ¿Serán esta estimación y este contraste fiables si hemos omitido una variable que parece ser relevante? ¿Se verán afectadas las propiedades del estimador MCO por esta omisión? Todo esto lo veremos en el tema de error de especificación.

Bibliografía

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5ª edn., South-Western.