Tema 7

Variables Cualitativas

Contenido

7.1. Introducción. Un ejemplo

A lo largo del curso únicamente se han especificado modelos con variables de naturaleza cuantitativa, es decir, aquéllas que toman valores numéricos. Sin embargo, las variables también pueden ser cualitativas, es decir, pueden tomar valores no numéricos como categorías, clases o atributos. Por ejemplo, son variables cualitativas el género de las personas, el estado civil, la raza, el pertenecer a diferentes zonas geográficas, momentos históricos, estaciones del año, etc. De esta forma, el salario de los trabajadores puede depender del género de los mismos; la tasa de criminalidad puede venir determinada por la zona geográfica de residencia de los individuos; el PIB de los países puede estar influenciado por determinados acontecimientos históricos como las guerras; las ventas de un determinado producto pueden ser significativamente distintas en función de la época del año, etc.

En este tema, aunque seguimos manteniendo que la variable dependiente es cuantitativa, vamos a considerar que ésta puede venir explicada por variables cualitativas y/o cuantitativas.

Dado que las categorías de las variables no son directamente cuantificables, las vamos a cuantificar construyendo unas variables artificiales llamadas ficticias, binarias o dummies, que son numéricas. Estas variables toman arbitrariamente el valor 1 si la categoría está presente en el individuo y 0 en caso contrario[1].

$$D_i = \begin{cases} 1 & \text{si la categoría está presente} \\ 0 & \text{en caso contrario} \end{cases}$$

En este tema estudiamos la estimación, interpretación de los coeficientes y contrastes de hipótesis en modelos con presencia de variables cualitativas como regresores.

7.2. Modelo con una variable cualitativa

Consideremos el caso más sencillo, una variable cualitativa como único regresor del modelo. Vamos a suponer que queremos explicar el precio de la vivienda basándonos únicamente en si la vivienda tiene piscina o no[2]. Para ello, definimos la siguiente variable ficticia:

$$POOL_i = \begin{cases} 1 & \text{si la vivienda i-ésima tiene piscina} \\ 0 & \text{en caso contrario} \end{cases}$$

Abrimos el fichero de datos *data7-3* de Ramanathan (2002), que contiene datos para 14 viviendas sobre el precio de venta de la vivienda (PRICE), pies cuadrados habitables (SQFT), número de habitaciones (BEDRMS) y número de baños (BATHS), utilizados en capítulos anteriores y añade una variable ficticia que toma el valor 1 si la vivienda tiene piscina y 0 en caso contrario (POOL), una variable ficticia que toma el valor 1 si la vivienda tiene sala

[1] Las variables ficticias pueden tomar dos valores cualesquiera, sin embargo, la interpretación de los coeficientes es más sencilla si se consideran los valores 0 y 1.

[2] Por simplicidad vamos a ignorar el efecto del resto de variables que afectan al precio de la vivienda.

de estar y 0 en caso contrario (FAMROOM) y una variable ficticia que toma el valor 1 si la vivienda tiene chimenea y 0 en caso contrario (FIREPL). Seleccionamos las variables PRICE y POOL y observamos los valores de estas dos variables:

Obs	price	pool
1	199,9	1
2	228,0	0
3	235,0	1
4	285,0	0
5	239,0	0
6	293,0	0
7	285,0	0
8	365,0	1
9	295,0	0
10	290,0	0
11	385,0	1
12	505,0	1
13	425,0	0
14	415,0	0

Por ejemplo, la primera vivienda de la muestra tiene un precio de 199.900 dólares y tiene piscina (ya que la variable POOL toma el valor 1), mientras que la segunda no tiene piscina (la variable POOL toma el valor 0) y su precio de venta es de 228.000 dólares, etc.

Con los datos anteriores podemos obtener fácilmente que el precio medio de la vivienda es 317.493 dólares:

Estadísticos principales, usando las observaciones 1 - 14
para la variable price (14 observaciones válidas)

Media	Mediana	Mínimo	Máximo
317,49	291,50	199,90	505,00
Desv. Típ.	**C.V.**	**Asimetría**	**Exc. de curtosis**
88,498	0,27874	0,65346	−0,52983

Sin embargo, también es posible obtener el precio medio para las viviendas que tienen piscina, por un lado, y para las que no la tienen, por otro. Para ello, en primer, lugar se selecciona el precio para aquellas viviendas con piscina. Para ello, seleccionamos la variable PRICE, pinchamos en *Muestra → Definir a partir de v. ficticia...*, seleccionamos la variable POOL y aceptamos. De esta forma hemos seleccionado el precio para aquellas viviendas que tienen piscina[3]. A continuación, se obtienen los estadísticos principales:

[3]Para restablecer el tamaño muestral inicial pinchar en *Muestra → Recuperar el rango completo.*

Estadísticos principales, usando las observaciones 1 - 5
para la variable price (5 observaciones válidas)

Media	Mediana	Mínimo	Máximo
337,98	365,00	199,90	505,00
Desv. Típ.	**C.V.**	**Asimetría**	**Exc. de curtosis**
122,99	0,36390	0,15896	−1,2798

Para seleccionar el precio de las viviendas que no tienen piscina, pinchamos en *Muestra → Restringir a partir de criterio*, introducimos la condición $POOL = 0$ y aceptamos. Los estadísticos principales son los siguientes:

Estadísticos principales, usando las observaciones 1 - 9
para la variable price (9 observaciones válidas)

Media	Mediana	Mínimo	Máximo
306,11	290,00	228,00	425,00
Desv. Típ.	**C.V.**	**Asimetría**	**Exc. de curtosis**
68,959	0,225275	0,87575	−0,52255

Por tanto, el precio medio de las viviendas con piscina es de 337.980 dólares frente a los 306.110 de las viviendas sin piscina. Dado el modelo una vivienda con piscina es en promedio 31.869 dólares más cara que la que no tiene piscina. Notar que no se están teniendo en cuenta otros factores que pueden afectar al precio de la vivienda (número de pies cuadrados habitables, número de habitaciones, etc.).

El sencillo análisis anterior podemos realizarlo mediante un análisis de regresión. Podemos especificar un modelo econométrico utilizando la variable ficticia POOL como regresor, estimarlo, hacer inferencia e ir incorporando otras características que pueden afectar a los precios de las viviendas. Para comenzar, consideramos el siguiente modelo de regresión lineal simple:

$$PRICE_i = \alpha_1 + \alpha_2 POOL_i + u_i \qquad i = 1, \ldots, 14 \tag{7.1}$$

Interpretación y estimación de los coeficientes

En nuestro ejemplo, la función de regresión poblacional varía en función de si la vivienda tiene piscina o no:

- $E(PRICE_i|i$ es una vivienda con piscina$) = \alpha_1 + \alpha_2$, puesto que la variable POOL toma el valor 1 y $E(u_i) = 0$.
- $E(PRICE_i|i$ es una vivienda sin piscina$) = \alpha_1$, puesto que la variable POOL toma el valor 0 y $E(u_i) = 0$.

Por tanto, los coeficientes se interpretan como sigue:

- α_1: precio medio de una vivienda sin piscina.
- $\alpha_1 + \alpha_2$: precio medio de una vivienda con piscina.
- α_2: diferencia en el precio medio de una vivienda con piscina con respecto a una que no la tiene.

Utilizando las ecuaciones normales que derivamos en el Tema 2 para estimar el modelo de regresión simple y teniendo en cuenta que al ser POOL una variable ficticia que toma valores 0 y 1 coincide con su cuadrado, obtenemos que los estimadores de los coeficientes del modelo (7.1) se pueden calcular a partir de simples medias muestrales[4]:

- $\hat{\alpha}_1 = \overline{PRICE}_{nopool} = 306{,}111 \quad \Rightarrow$ precio estimado medio de las viviendas sin piscina.
- $\hat{\alpha}_2 = \overline{PRICE}_{pool} - \overline{PRICE}_{nopool} = 337{,}980 - 306{,}111 = 31{,}869 \quad \Rightarrow$ diferencia estimada en el precio medio de las viviendas con piscina con respecto a las que no la tienen.

En efecto, si estimamos el modelo por Mínimos Cuadrados Ordinarios utilizando Gretl obtenemos que las estimaciones de los coeficientes son las siguientes:

Modelo (7.1): estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: price

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	306,111	30,2077	10,1335	0,0000
pool	31,8689	50,5471	0,6305	0,5402

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	98550,5
Desviación típica de los residuos ($\hat{\sigma}$)	90,6231
R^2	0,0320632
$\bar{R}^2$ corregido	−0,0485982
Grados de libertad	12
Log-verosimilitud	−81,880
Criterio de información de Akaike	167,760
Criterio de información Bayesiano de Schwarz	169,038

Que coinciden con las calculadas utilizando los valores obtenidos en ambas submuestras mediante los Estadísticos Principales:

$$\underset{(estad.\,t)}{\widehat{PRICE}_i} = \underset{(10{,}13)}{306{,}111} + \underset{(0{,}63)}{31{,}869} POOL_i \qquad i = 1, \ldots, 14$$

[4] $\overline{PRICE}_{pool}$ es la media muestral del precio de las viviendas con piscina, de igual forma $\overline{PRICE}_{nopool}$ es la media muestral del precio de las viviendas sin piscina.

El modelo (7.1) no es la única especificación correcta posible para explicar las variaciones del precio de la vivienda en función de si tiene piscina o no. Al igual que hemos definido la variable ficticia POOL, podemos crear la variable NOPOOL, tomando el valor 1 si la vivienda no tiene piscina y 0 en caso contrario. Con esta nueva variable podemos especificar los dos modelos siguientes:

$$PRICE_i = \gamma_1 + \gamma_2 NOPOOL_i + u_i \qquad i = 1, \ldots, 14 \tag{7.2}$$

$$PRICE_i = \beta_1 POOL_i + \beta_2 NOPOOL_i + u_i \qquad i = 1, \ldots, 14 \tag{7.3}$$

La interpretación de los coeficientes se haría de forma análoga a como hemos visto para el modelo (7.1). Notar que la equivalencia entre los coeficientes de los distintos modelos (7.1), (7.2) y (7.3) es la siguiente:

- $E(PRICE_i|i$ es una vivienda con piscina$) = \alpha_1 + \alpha_2 = \gamma_1 = \beta_1$
- $E(PRICE_i|i$ es una vivienda sin piscina$) = \alpha_1 = \gamma_1 + \gamma_2 = \beta_2$

Una especificación que no sería adecuada es la siguiente:

$$PRICE_i = \alpha + \beta_1 POOL_i + \beta_2 NOPOOL_i + u_i \qquad i = 1, \ldots, 14$$

ya que si analizamos la matriz de datos X para este modelo observamos que la suma de la segunda y tercera columnas es igual a la primera y tendríamos un problema de multicolinealidad exacta, por lo que la matriz $X'X$ no sería invertible. En estas circunstancias no se podría obtener una única solución para $\hat{\alpha}, \hat{\beta}_1$ y $\hat{\beta}_2$ del sistema de ecuaciones normales.

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

Contraste de hipótesis

Los contrastes de hipótesis se realizan con la metodología estudiada en los capítulos previos. Por ejemplo, si quisiéramos contrastar en el modelo (7.1) si hay diferencias significativas en

el precio medio de la vivienda entre aquéllas que tienen piscina y las que no, la hipótesis de contraste es $H_0 : \alpha_2 = 0$.[5] Este contraste se puede realizar utilizando el estadístico t habitual cuyo *valor-p* es 0,5402, por lo que no se rechaza la hipótesis nula para un nivel de significación del 5 %, es decir, el precio medio de la vivienda no es significativamente diferente por el hecho de tener piscina. Alternativamente, se puede realizar el contraste utilizando el estadístico F basado en las sumas de cuadrados de los residuos siendo en este caso el modelo (7.1) el modelo no restringido mientras que el modelo restringido es $PRICE_i = \alpha_1 + u_i \quad i = 1, \ldots, 14$.

7.2.1. Incorporación de variables cuantitativas

En el modelo (7.1) el único regresor para explicar el precio de la vivienda es una característica cualitativa, el hecho de tener o no piscina sin embargo, en un modelo pueden convivir variables cualitativas y cuantitativas. Vamos a comenzar añadiendo un regresor cuantitativo, la variable SQFT (número de pies cuadrados habitables de la vivienda) y manteniendo la variable ficticia POOL afectando a la ordenada.

Cambio en la ordenada

Suponer que el precio de la vivienda únicamente depende de si tiene piscina o no es poco realista, por lo que añadimos como regresor a la variable cuantitativa SQFT (número de pies cuadrados habitables de la vivienda) de la siguiente manera:

$$PRICE_i = \alpha_1 + \alpha_2 POOL_i + \beta\, SQFT_i + u_i \quad i = 1, \ldots, 14 \tag{7.4}$$

Estimación e interpretación de los coeficientes:

La función de regresión poblacional se puede expresar como:

- $E(PRICE_i|i$ es una vivienda con piscina$) = \alpha_1 + \alpha_2 + \beta\, SQFT_i$
- $E(PRICE_i|i$ es una vivienda sin piscina$) = \alpha_1 + \beta\, SQFT_i$

Por tanto podemos interpretar α_1 como el precio esperado de una vivienda sin piscina y cero pies cuadrados, α_2 como el diferencial en el precio esperado en una vivienda por el hecho de tener piscina, manteniendo el número de pies cuadrados habitables constante. A igual número de pies cuadrados habitables el hecho de tener piscina se puede considerar una mejora en la vivienda por lo que sería preferida, así tener piscina es una característica que sube el precio de la vivienda y esperaríamos que α_2 tuviese signo positivo. Finalmente interpretamos β como la variación en el precio esperado de una vivienda por incrementar su superficie en un pie cuadrado. Esperaríamos signo positivo, a mayor superficie mayor precio esperado para la vivienda. Gráficamente, obtenemos dos rectas con igual pendiente, β, y distinta ordenada como podemos observar en el Gráfico 7.1:

[5] Equivalentemente, $H_0 : \gamma_2 = 0$ ó $H_0 : \beta_1 = \beta_2$ para los modelos (7.2) y (7.3), respectivamente.

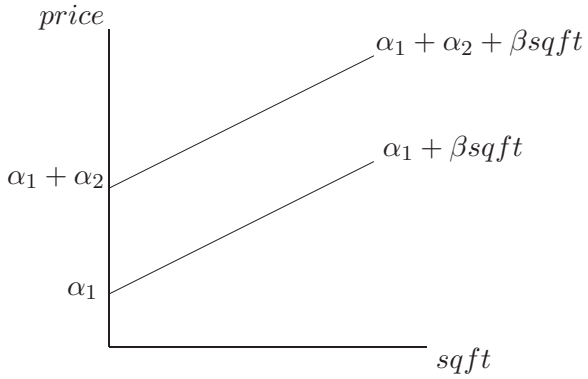

Gráfico 7.1: Cambio en ordenada

El resultado de la estimación del modelo (7.4) por Mínimos Cuadrados Ordinarios es:

Modelo (7.4): estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: price

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	22,6728	29,5058	0,7684	0,4584
pool	52,7898	16,4817	3,2029	0,0084
sqft	0,144415	0,0141849	10,1809	0,0000

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	9455,36
Desviación típica de los residuos ($\hat{\sigma}$)	29,3186
R^2	0,907132
$\bar{R}^2$ corregido	0,890247
$F(2, 11)$	53,7238
Log-verosimilitud	−65,472
Criterio de información de Akaike	136,944
Criterio de información Bayesiano de Schwarz	138,861

El modelo estimado es:

$$\underset{(estad.\,t)}{\widehat{PRICE}_i} = \underset{(0,768)}{22,673} + \underset{(3,203)}{52,790} POOL_i + \underset{(10,181)}{0,144}\, SQFT_i$$

donde se puede observar que ambos regresores son significativos para explicar el precio medio de la vivienda y tienen los signos adecuados[6]. Por tanto, existen diferencias significativas en el precio medio de la vivienda que tiene piscina con respecto a la que no la tiene.

Los coeficientes estimados se interpretan como sigue:

[6]El valor de los estadísticos t para los coeficientes de ambos regresores es superior al valor crítico de una distribución t de Student de $N - K = 14 - 3 = 11$ grados de libertad para un nivel de significación del 5 %, que es 2,201.

- $\hat{\alpha}_1 = 22,673 \Rightarrow$ el precio medio estimado de las viviendas sin piscina y con cero pies cuadrados habitables es 22.673 dólares.
- $\hat{\alpha}_2 = 52,790 \Rightarrow$ se estima que entre dos viviendas con el mismo número de pies cuadrados habitables el precio medio de una con piscina es 52.790 dólares más caro que el de una sin piscina.
- $\hat{\beta} = 0,144 \Rightarrow$ el precio medio estimado de una vivienda se incrementa en 144 dólares al aumentar en un pie cuadrado habitable la vivienda.

Cambio en la ordenada y en la pendiente

También es posible pensar que la variación en el precio de las viviendas ante el incremento en un pie cuadrado habitable sea diferente para aquéllas que tienen piscina. En este caso se especifica el siguiente modelo, donde la variable ficticia POOL afecta tanto a la ordenada como a la pendiente de la recta:

$$PRICE_i = \alpha_1 + \alpha_2 POOL_i + \beta_1 SQFT_i + \beta_2 POOL \cdot SQFT_i + u_i \qquad i = 1, \ldots, 14 \qquad (7.5)$$

La interacción $POOL \cdot SQFT$ mide el número de pies cuadrados habitables para las viviendas que tienen piscina, mientras que toma el valor 0 para las que no la tienen.

Estimación e interpretación de los coeficientes:

Una vez definida la interacción $POOL \cdot SQFT$ en Gretl, estimamos el modelo (7.5):

Modelo (7.5): estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: price

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	77,1332	25,6379	3,0086	0,0131
pool	−82,648	39,7759	−2,0779	0,0644
sqft	0,116667	0,0125934	9,2641	0,0000
pool· sqft	0,0722955	0,0203274	3,5566	0,0052

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	4174,72
Desviación típica de los residuos ($\hat{\sigma}$)	20,4321
R^2	0,958997
$\bar{R}^2$ corregido	0,946696
$F(3, 10)$	77,9615
Log-verosimilitud	−59,749
Criterio de información de Akaike	127,499
Criterio de información Bayesiano de Schwarz	130,055

La función de regresión poblacional se puede expresar como:

- $E(PRICE_i | i$ es una vivienda con piscina$) = \alpha_1 + \alpha_2 + (\beta_1 + \beta_2) SQFT_i$

- $E(PRICE_i|i \text{ es una vivienda sin piscina}) = \alpha_1 + \beta_1 SQFT_i$

El parámetro poblacional α_1 se interpreta como el precio esperado de una vivienda sin piscina y con cero pies cuadrados habitables. α_2 mide el diferencial en el precio esperado de una vivienda con cero pies cuadrados habitables por el hecho de tener piscina. Esperaríamos que ambos coeficientes tuviesen signo positivo por las razones argumentadas anteriormente.

β_1 se interpreta como la variación en el precio esperado de una vivienda sin piscina por incrementar su superficie en un pie cuadrado habitable mientras que β_2 mide el diferencial en la variación en el precio esperado de una vivienda ante un incremento de su superficie en un pie cuadrado por el hecho de tener piscina. Esperaríamos que ambos coeficientes tuviesen signo positivo, a mayor superficie de la vivienda mayor precio esperado. Si además la vivienda tiene piscina el cambio en el precio esperado por pie cuadrado más de superficie será mayor ya que la posesión de piscina es una mejora.

La representación gráfica corresponde a dos rectas que varían tanto en el punto de corte con el eje de ordenadas como en la pendiente:

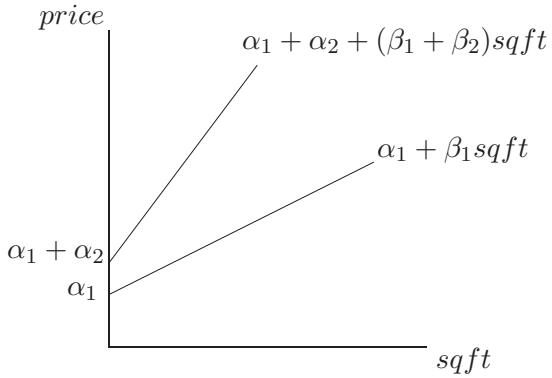

Gráfico 7.2: Cambio en ordenada y en pendiente

Interpretación de los coeficientes estimados:

- $\hat{\alpha}_1 = 77,133 \Rightarrow$ el precio medio estimado de las viviendas que no tienen piscina y con cero pies cuadrados habitables es 77.133 dólares.
- $\hat{\alpha}_2 = -82,648 \Rightarrow$ entre dos viviendas con 0 pies cuadrados habitables el precio medio estimado de una con piscina es 82.648 dólares más barato que el de una sin piscina.
- $\hat{\beta}_1 = 0,117 \Rightarrow$ al incrementar en un pie cuadrado la superficie habitable, el precio medio estimado de una vivienda sin piscina aumenta en 117 dólares.
- $\hat{\beta}_2 = 0,072 \Rightarrow$ al incrementar en un pie cuadrado la superficie habitable, el precio medio estimado de una vivienda con piscina aumenta en 72 dólares.

Contraste de hipótesis

La hipótesis nula para contrastar si tener piscina influye significativamente en el precio medio de las viviendas es $H_0 : \alpha_2 = \beta_2 = 0$. El resultado del contraste es:

Contraste de omisión de variables –
  Hipótesis nula: los parámetros son cero para las variables
    pool
    poolsqft
  Estadístico de contraste: $F(2, 10) = 16{,}886$
  con valor p = $P(F(2, 10) > 16{,}886) = 0{,}000622329$

por lo que se rechaza la hipótesis nula para un nivel de significación del 5 % y por lo tanto tener piscina es una variable significativa para explicar el precio de las viviendas.

También se puede contrastar mediante un contraste de significatividad individual si el incremento en un pie cuadrado de superficie afecta al precio de manera diferente según la vivienda tenga o no piscina, para ello podemos contrastar $H_0 : \beta_2 = 0$. Como vemos en los resultados de la estimación del modelo este coeficiente es significativo, como esperábamos la influencia de la superficie habitable de una vivienda en su precio varía si la vivienda tiene piscina o no. Por otro lado, $\hat{\alpha}_2$ no tiene el signo esperado y a su vez no es significativo a nivel individual, aparentemente el hecho de incluir la variable ficticia en la pendiente ha restado significatividad a la discriminación en la ordenada.

7.3. Modelo con dos o más variables cualitativas

Al igual que ocurría con los regresores cuantitativos sobre una variable endógena pueden influir más de una variable cualitativa. Por ejemplo en el precio de una vivienda podría influir no sólo el hecho de tener o no piscina, su superficie habitable, el número de habitaciones, el número de baños, si no también si tiene o no chimenea, si tiene o no ascensor o la zona de la ciudad donde esté situada.

7.3.1. Varias categorías

Supongamos que creemos que la zona de la ciudad donde esté situada la vivienda es un determinante de su precio. Pensemos por ejemplo en precios de viviendas situadas en una gran ciudad en la que podemos distinguir como zonas a la zona centro, zona norte, zona sur, zona este y zona oeste. En general el centro de las ciudades es una zona valorada por ser el centro neurálgico económico-comercial y el resto de zonas se valorará en función del tipo de viviendas que recoja y sus comunicaciones, por ejemplo en una ciudad como Madrid esperaríamos mayor precio en el centro, norte y oeste que en el sur o en el este que agrupan a barrios, en general, con menor nivel económico y peor comunicados. Para el ejemplo necesitamos definir cinco variables ficticias una para cada zona ya que la situación geográfica de la vivienda la hemos

dividido en cinco categorías[7].

Definimos las siguiente variables:

$$D_{1i} = \begin{cases} 1 & \text{si la vivienda i-ésima está situada en la zona centro} \\ 0 & \text{en caso contrario} \end{cases}$$
$$D_{2i} = \begin{cases} 1 & \text{si la vivienda i-ésima está situada en la zona norte} \\ 0 & \text{en caso contrario} \end{cases}$$
$$D_{3i} = \begin{cases} 1 & \text{si la vivienda i-ésima está situada en la zona sur} \\ 0 & \text{en caso contrario} \end{cases}$$
$$D_{4i} = \begin{cases} 1 & \text{si la vivienda i-ésima está situada en la zona este} \\ 0 & \text{en caso contrario} \end{cases}$$
$$D_{5i} = \begin{cases} 1 & \text{si la vivienda i-ésima está situada en la zona oeste} \\ 0 & \text{en caso contrario} \end{cases}$$

Si además de la situación geográfica de la vivienda creemos que la superficie habitable influye en su precio podemos definir, por ejemplo, el siguiente modelo:

$$PRICE_i = \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \alpha_5 D_{5i} + \beta\, SQFT_i + u_i \quad i = 1, \ldots, N \quad (7.6)$$

Donde β se interpreta de la forma habitual y α_1 se interpreta como el precio esperado de una vivienda con cero pies cuadrados situada en la zona centro, así $\alpha_i \quad i = 1, \ldots, 5$ se interpretan como el precio esperado de una vivienda con cero pies cuadrados situadas en la zona correspondiente, centro, norte, sur, este u oeste.

En la especificación (7.6) se ha optado por no incluir término independiente en el modelo e incluir las cinco variables ficticias para no incurrir en un problema de multicolinealidad exacta como se expuso en el punto anterior pero, podríamos especificar un modelo con término independiente siempre y cuando dejemos fuera una de las variables ficticias o categorías para no tener dicho problema. Por ejemplo una especificación alternativa sería:

$$PRICE_i = \alpha + \alpha_2^\star D_{2i} + \alpha_3^\star D_{3i} + \alpha_4^\star D_{4i} + \alpha_5^\star D_{5i} + \beta\, SQFT_i + u_i \qquad i = 1, \ldots, N \quad (7.7)$$

En el modelo anterior la interpretación del parámetro poblacional β no varía, α se interpreta como el precio esperado de una vivienda con cero pies cuadrados situada en la zona centro, $\alpha_i^\star \quad i = 2, \ldots, 5$ se interpretan como el diferencial en el precio esperado de una vivienda, a igual superficie habitable, por estar situada en la zona norte, (sur, este y oeste respectivamente) con respecto a una vivienda situada en la zona centro. Qué variable ficticia (o categoría) dejemos fuera no es relevante siempre y cuando interpretemos adecuadamente los parámetros. Naturalmente podemos afectar las variables ficticias a la variable cuantitativa como en el caso anterior siempre y cuando no incurramos en multicolinealidad exacta.

[7] En el ejemplo anterior la vivienda tenía o no piscina, solo había dos casos posibles y por tanto sólo había dos categorías.

Contraste de hipótesis

Para contrastar en el modelo (7.6) que por ejemplo no existen diferencias significativas en el precio medio de la vivienda por su situación la hipótesis de contraste es $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5$. Hipótesis que podemos contrastar utilizando el estadístico F basado en las sumas de cuadrados de los residuos siendo en este caso el modelo (7.6) el modelo no restringido mientras que el modelo restringido sería $PRICE_i = \alpha_1 + \beta\, SQFT_i + u_i \quad i = 1, \ldots, N$. El mismo contraste puede llevarse a cabo en el modelo (7.7) con la hipótesis $H_0 : \alpha_2^\star = \alpha_3^\star = \alpha_4^\star = \alpha_5^\star = 0$ siendo el modelo no restringido el modelo (7.7) y el restringido $PRICE_i = \alpha + \beta\, SQFT_i + u_i \quad i = 1, \ldots, N$.

7.3.2. Varios conjuntos de variables ficticias

Supongamos que ampliamos el modelo (7.4) incorporando regresores que podrían explicar el precio de la vivienda como por ejemplo el número de habitaciones, el número de baños, que la vivienda tenga sala de estar o no y que tenga chimenea o no. Las dos primeras son variables ficticias que pueden definirse así:

$$FIREPL_i = \begin{cases} 1 & \text{si la vivienda i-ésima tiene chimenea} \\ 0 & \text{en caso contrario} \end{cases}$$
$$FAMROOM_i = \begin{cases} 1 & \text{si la vivienda i-ésima tiene sala de estar} \\ 0 & \text{en caso contrario} \end{cases}$$

Mientras que el número de baños y el número de habitaciones se definen como en los temas anteriores:

$BEDRMS$ número de habitaciones de la vivienda i-ésima
$BATHS$ número de cuartos de baño de la vivienda i-ésima

Con todas ellas podemos definir el siguiente modelo para explicar el precio de la vivienda:

$$\begin{aligned} PRICE_i = {} & \gamma_1 + \gamma_2 POOL_i + \gamma_3 FAMROOM_i + \gamma_4 FIREPL_i \\ & + \beta_1 SQFT_i + \beta_2 BEDRMS_i + \beta_3 BATHS_i + u_i \quad i = 1, \ldots, 14 \end{aligned} \tag{7.8}$$

Donde lo primero a notar es que en el modelo (7.8), afectando a la ordenada, conviven tres conjuntos de variables ficticias con dos categorías cada una, el hecho de tener o no piscina, el hecho de tener o no chimenea y el hecho de tener o no sala de estar, de las cuales sólo se incluye una de cada conjunto y se mantiene el término independiente.

Esta forma de definir el modelo es muy cómoda ya que sigue manteniendo los resultados de los modelos con término independiente y permite una fácil interpretación de los coeficientes que acompañan a las variables ficticias. Así, γ_i $i = 2, 3, 4$ recogen el diferencial en el valor esperado de una vivienda por el hecho de poseer la característica correspondiente manteniéndose constante el resto de variables.

El resultado de la estimación es:

Modelo (7.8): estimaciones MCO utilizando las 14 observaciones 1–14
Variable dependiente: price

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	39,0571	89,5397	0,4362	0,6758
pool	53,1958	22,0635	2,4110	0,0467
famroom	−21,344	42,8734	−0,4979	0,6338
firepl	26,1880	53,8454	0,4864	0,6416
sqft	0,146551	0,0301014	4,8686	0,0018
bedrms	−7,0455	28,7363	−0,2452	0,8134
baths	−0,263691	41,4547	−0,0064	0,9951

Media de la var. dependiente	317,493
D.T. de la variable dependiente	88,4982
Suma de cuadrados de los residuos	9010,24
Desviación típica de los residuos ($\hat{\sigma}$)	35,8773
R^2	0,911504
$\bar{R}^2$ corregido	0,835650
$F(6, 7)$	12,0166
valor p para $F()$	0,00221290
Log-verosimilitud	−65,134
Criterio de información de Akaike	144,269
Criterio de información Bayesiano de Schwarz	148,743

La interpretación de los coeficientes estimados es la siguiente:

- $\hat{\gamma}_1 = 39,057$: el precio medio estimado de las viviendas sin piscina, baños, habitaciones, sala de estar ni chimenea y con 0 pies cuadrados habitables es de 39.057 dólares.

- $\hat{\gamma}_2 = 53,1958$: la diferencia estimada en el precio medio de las viviendas con piscina con respecto a las que no la tienen, siendo iguales en el resto de características (pies cuadrados habitables, número de habitaciones, número de baños, existencia de sala de estar y/o chimenea) es de 53.196 dólares.

- $\hat{\gamma}_3 = -21,34$: el precio medio estimado de una vivienda con sala de estar es 21.340 dólares inferior al de una sin sala de estar, siendo idénticas en el resto de características. Esto se debe a que, al mantener constante el número de pies cuadrados de la vivienda y el número de habitaciones y baños, incluir una sala de estar hará que el resto de habitaciones o baños sean de menor tamaño.

- $\hat{\gamma}_4 = 26,188$: el precio medio estimado de una vivienda con chimenea es 26.188 dólares más caro que el de una sin chimenea, siendo idénticas en el resto de características.

- $\hat{\beta}_1 = 0,147$: el precio medio estimado de una vivienda se incrementa en 147.000 dólares al aumentar en 1 pie cuadrado habitable su superficie, permaneciendo constantes el número de baños y habitaciones y el resto de características de la vivienda.

- $\hat{\beta}_2 = -7,046$: el precio medio estimado de una vivienda disminuye en 7.046 dólares al aumentar en 1 el número de habitaciones, permaneciendo constantes el número de baños y los pies cuadrados habitables y el resto de características de la vivienda. Esto se debe a que las habitaciones serán de menor tamaño .

- $\hat{\beta}_3 = -0,264$: el precio medio estimado de una vivienda disminuye en 264 dólares al aumentar en 1 el número de baños, permaneciendo constantes el número de habitaciones y los pies cuadrados habitables el resto de características de la vivienda. De nuevo, las habitaciones serán de menor tamaño.

Contraste de hipótesis

Para contrastar, por ejemplo, que no existen diferencias significativas en el precio medio de la vivienda por el hecho de tener chimenea, se realiza un contraste de significatividad individual de la variable FIREPL. En este caso, observando el *valor-p* correspondiente, 0,6416, se puede concluir que a un nivel de significación del 5 %, no existen diferencias significativas en el precio medio de una vivienda por el hecho de tener chimenea.

Si comparamos los modelos (7.4) y (7.8), ninguna de las variables añadidas en el último modelo es significativa individualmente[8]. Además, el $\bar{R}^2$ es inferior. El contraste de significatividad conjunta para las variables añadidas se puede realizar con el estadístico F basado en las sumas de cuadrados residuales de los modelos restringido (modelo (7.4)) y no restringido (modelo (7.8)). En este caso, el resultado es:

Contraste de omisión de variables –
  Hipótesis nula: los parámetros son cero para las variables
    bedrms
    baths
    famroom
    firepl
  Estadístico de contraste: $F(4, 7) = 0,0864517$
  con valor p = $P(F(4, 7) > 0,0864517) = 0,983881$

por lo que no se rechaza la hipótesis nula de que las variables añadidas al modelo (7.4) son conjuntamente no significativas. Al omitir dichas variables el modelo mejora en cuanto a la significación de sus coeficientes y el $\bar{R}^2$. Por tanto, manteniendo las variables POOL y SQFT, la inclusión del resto (FIREPL, FAMROOM, BATHS, BEDRMS) no añade capacidad explicativa al modelo.

[8] Un problema añadido es que tenemos un bajo tamaño muestral, T=14, y hemos aumentado significativamente el número de parámetros a estimar, K=7, por lo que tenemos muy pocos grados de libertad.

7.4. Contraste de cambio estructural

En ocasiones puede ocurrir que la relación entre la variable dependiente y los regresores cambie a lo largo del periodo muestral, es decir, puede que exista un cambio estructural. Por ejemplo, si estamos analizando el consumo de tabaco y durante el período muestral se ha producido una campaña de salud pública informando sobre los peligros que conlleva el consumo de tabaco, podemos pensar que tras dicha campaña el comportamiento de la demanda de tabaco haya cambiado, reduciéndose significativamente. Si esto ocurre no podemos especificar una única función de demanda para todo el período muestral si no que deberíamos especificar dos funciones, una hasta la campaña antitabaco y otra para el período siguiente. Por tanto, ante sospechas de que exista un cambio estructural, debemos de contrastar la estabilidad de los parámetros de nuestra relación.

El contraste de cambio estructural, llamado habitualmente contraste de Chow, puede realizarse de manera sencilla mediante el estadístico de sumas de cuadrados de los residuos sin más que especificar adecuadamente el modelo restringido y el no restringido. También podemos llevarlo a cabo utilizando variables ficticias. Veamos un ejemplo.

El fichero *data7-19* contiene datos para 1960-1988 sobre la demanda de tabaco y sus determinantes en Turquía. Las variables de interés para el ejemplo son las siguientes:

Q: consumo de tabaco por adulto (en kg).
Y: PNB real per cápita en liras turcas de 1968.
P: precio real del kilogramo de tabaco, en liras turcas.
$D82$: variable ficticia que toma valor 1 a partir de 1982.

A mediados de 1981 el gobierno turco lanza una campaña de salud pública advirtiendo de los peligros de salud que conlleva el consumo de tabaco. Nuestro objetivo es determinar si existen cambios en la demanda de tabaco tras la campaña institucional en cuyo caso la especificación:

$$LnQ_t = \alpha + \beta LnY_t + \gamma LnP_t + u_t \qquad t = 1960, \ldots, 1988 \tag{7.9}$$

no es correcta para todo el período muestral y deberíamos especificar dos ecuaciones:

$$LnQ_t = \alpha_1 + \beta_1 LnY_t + \gamma_1 LnP_t + u_{1t} \qquad t = 1960, \ldots, 1981 \tag{7.10}$$
$$LnQ_t = \alpha_2 + \beta_2 LnY_t + \gamma_2 LnP_t + u_{2t} \qquad t = 1982, \ldots, 1988 \tag{7.11}$$

Si existe cambio estructural rechazaríamos $H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ y $\gamma_1 = \gamma_2$

Este contraste podemos llevarlo a cabo utilizando el estadístico F basado en las sumas de cuadrados de los residuos siendo en este caso el modelo restringido el recogido en la ecuación (7.9) mientras que el modelo no restringido está constituido por las ecuaciones (7.10) y (7.11). Utilizando Gretl una vez abierto el fichero de datos y tomado las correspondientes transformaciones estimaríamos el modelo (7.9) por MCO y en la ventana de resultados de la estimación elegimos:

Contrastes $\longrightarrow$ *Contraste de Chow*

A la pregunta *Observación en la cual dividir la muestra* contestaríamos *1982* y la correspondiente devolución es:

Modelo (7.9): estimaciones MCO utilizando las 29 observaciones 1960-1988
Variable dependiente: lnQ

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	−4,58987	0,724913	−6,332	0,00001***
lnY	0,688498	0,0947276	7,268	0,00001***
lnP	0,485683	0,101394	−4,790	0,00006***

Media de la var. dependiente = 0,784827
Desviación típica de la var. dependiente. = 0,108499
Suma de cuadrados de los residuos = 0,0949108
Desviación típica de los residuos = 0,0604187
R-cuadrado = 0,712058
R-cuadrado corregido = 0,689908
Estadístico F (2, 26) = 32,148 (valor p < 0,00001)
Estadístico de Durbin-Watson = 1,00057
Coef. de autocorr. de primer orden. = 0,489867
Log-verosimilitud = 41,8214
Criterio de información de Akaike (AIC) = -77,6429
Criterio de información Bayesiano de Schwarz (BIC) = -73,541
Criterio de Hannan-Quinn (HQC) = -76,3582

Contraste de Chow de cambio estructural en la observación 1982 -
Hipótesis nula: no hay cambio estructural
Estadístico de contraste: F(3, 23) = 20,1355
con valor p = P(F(3, 23) > 20,1355) = 1,25619e-006

El estadístico calculado es $F_c = 20,135 > F_{0,05(3,23)}$ por lo que rechazamos H_0 para un nivel de significatividad del 5 %, es decir existe cambio estructural, la campaña institucional ha tenido efecto y la demanda de tabaco en Turquía de 1960 a 1988 queda especificada por las ecuaciones (7.10) y (7.11). Los resultados de la estimación mínimo cuadrática de estas ecuaciones son los siguientes:

$$\underset{(estad.\,t)}{\widehat{LnQ}_t} = \underset{(-10,614)}{-5,024} + \underset{(11,587)}{0,735}\,LnY_t - \underset{(-4,227)}{0,381}\,LnP_t \qquad t = 1960,\ldots,1981 \quad SCR_1 = 0,01654$$

$$\underset{(estad.\,t)}{\widehat{LnQ}_t} = \underset{(2,170)}{8,837} - \underset{(-1,941)}{0,953}\,LnY_t + \underset{(0,654)}{0,108}LnP_t \qquad t = 1982,\ldots,1988 \quad SCR_2 = 0,00965$$

7.4.1. Cambio estructural utilizando variables ficticias

Alternativamente, el contraste anterior podríamos haberlo realizado mediante la variable ficticia $D82$ especificando el siguiente modelo donde $t = 60,\ldots,88$:

$$LnQ_t = \beta_1 + \beta_2 LnY_t + \beta_3 LnP_t + \beta_1^\star D82_t + \beta_2^\star D82_t \cdot LnY_t + \beta_3^\star D82_t \cdot LnP_t + u_t \quad (7.12)$$

En el cual, si existe cambio estructural rechazaríamos $H_0 : \beta_1^\star = \beta_2^\star = \beta_3^\star = 0$. De nuevo el contraste puede realizarse con el estadístico F habitual de sumas residuales donde el modelo no restringido es el (7.12) y el modelo restringido es

$$LnQ_t = \beta_1 + \beta_2 LnY_t + \beta_3 LnP_t + u_t \tag{7.13}$$

Utilizando Gretl, el proceso después de abierto el fichero de datos, tomado logaritmos y construido las interacciones $D82 \cdot LnY$ y $D82 \cdot LnP$, sería: estimaríamos el modelo (7.12) por MCO y en la ventana de resultados de la estimación haríamos

Contrastes $\longrightarrow$ *Omitir variables*

elegiríamos $D82$, $D82 \cdot LnY$ y $D82 \cdot LnP$ y obtendríamos el siguiente resultado:

Modelo 1: estimaciones MCO utilizando las 29 observaciones 1960-1988
Variable dependiente: lnQ

Variable	Coeficiente	Desv. típica	Estadístico t	valor p
const	−4,58987	0,724913	−6,332	0,00001***
lnY	0,688498	0,0947276	7,268	0,00001***
lnP	0,485683	0,101394	−4,790	0,00006***

Media de la var. dependiente = 0,784827
Desviación típica de la var. dependiente. = 0,108499
Suma de cuadrados de los residuos = 0,0949108
Desviación típica de los residuos = 0,0604187
R-cuadrado = 0,712058
R-cuadrado corregido = 0,689908
Estadístico F (2, 26) = 32,148 (valor p < 0,00001)
Estadístico de Durbin-Watson = 1,00057
Coef. de autocorr. de primer orden. = 0,489867
Log-verosimilitud = 41,8214
Criterio de información de Akaike (AIC) = -77,6429
Criterio de información Bayesiano de Schwarz (BIC) = -73,541
Criterio de Hannan-Quinn (HQC) = -76,3582

Comparación entre el modelo (7.12) y el modelo (7.13):
Hipótesis nula: los parámetros de regresión son cero para las variables
D82
D82Y
D82P
Estadístico de contraste: F(3, 23) = 20,1355, con valor p = 1,25619e-006
De los 3 estadísticos de selección de modelos, 0 han mejorado.

Dado el *valor-p* rechazamos la hipótesis nula para un nivel de significatividad del 5 % y existe cambio estructural. La demanda de tabaco en Turquía de 1960 a 1988 queda mejor especificada por el modelo (7.12). O lo que es lo mismo las ecuaciones (7.10) y (7.11) si no utilizamos

la variable ficticia $D82$ en la especificación del modelo. Notar que ambas especificaciones son idénticas, son dos formas alternativas y por lo tanto equivalentes de especificar la demanda de tabaco en Turquía para ese periodo temporal.

Bibliografía

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