Variables exógenas cualitativas

Manuel Artís Ortuño
Montserrat Guillén Estany

PID_00160618

Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

El modelo de regresión se fundamenta en una relación entre la variable endógena y las variables explicativas, donde estas variables pueden ser de tipo cuantitativo y/o cualitativo. En este módulo didáctico estudiaremos las características de los modelos en los cuales, entre las variables explicativas, hay variables cualitativas.

Para recordar los conceptos de modelo de regresión, variable endógena y variables explicativas, consultad los subapartados 2.1 y 2.2 del módulo "Modelo de regresión lineal múltiple: especificación.." de esta asignatura.

Las **variables cualitativas** miden características de los sujetos que no se pueden cuantificar con valores.

Ejemplos de variables cualitativas

Algunos ejemplos de variables cualitativas son los siguientes: el sector de actividad (en el caso de empresas), el tipo de contrato (para los asalariados) o el canal de distribución (en el caso de estudiar productos comerciales). Cuando modelizamos comportamientos individuales, atributos como el sexo, la nacionalidad o la clase social también son ejemplos típicos de variables explicativas cualitativas.

Las variables cualitativas se codifican (y se registran en soporte informático) con valores numéricos que no tienen sentido por sí mismos, pero que interpretamos como indicadores de las categorías posibles. Por ejemplo, en una base de datos de un colectivo de trabajadores, podemos identificar a los hombres con un 1 y a las mujeres con un 0. Debemos entender que esta codificación sólo es una equivalencia que definimos por comodidad, es decir, para no tener que escribir las palabras *hombre* y *mujer* en la casilla correspondiente.

Si queremos que algunas variables cualitativas formen parte del conjunto de variables explicativas en un modelo de regresión, debemos introducirlas adecuadamente en el modelo y transformarlas. El caso más sencillo es el de las **variables dicotómicas**, que pueden tomar únicamente dos valores. Para facilitar las interpretaciones de los resultados del modelo de regresión, las variables dicotómicas se codifican con 1 y 0, como en el ejemplo anterior. Las **variables politómicas**, que pueden tomar más de dos valores, requieren una transformación, definiendo unas variables ficticias dicotómicas, y no deben incluirse directamente en el proceso de estimación del modelo.

Consultad la definición de variable ficticia en el subapartado 1.2 de este módulo didáctico.

Objetivos

En este módulo didáctico explicaremos cómo se utilizan las variables exógenas cualitativas en un modelo de regresión y cómo se deben interpretar los resultados de la estimación de sus parámetros. Los objetivos que el estudiante debe alcanzar son los siguientes:

1. Saber cómo se introduce una variable cualitativa como explicativa en un modelo de regresión, tanto en el caso dicotómico como en el politómico.

2. Interpretar los parámetros asociados a variables ficticias en el caso aditivo.

3. Interpretar los parámetros asociados a variables ficticias en el caso multiplicativo.

4. Utilizar y entender las interacciones entre variables cualitativas.

5. Saber aplicar el uso de variables ficticias a la especificación de modelos de regresión, a la identificación de datos atípicos, a cambios estructurales o a la estacionalidad.

1. Modelos con variables exógenas cualitativas

El análisis econométrico mediante modelos de regresión estudia situaciones en las cuales una o más variables explicativas pueden tener un carácter cualitativo. A menudo las variables explicativas se denominan **atributos**. La **codificación de atributos** implica identificar cada categoría con un valor. Al realizar esta operación, debemos tener en cuenta que, habitualmente, las categorías no poseen una ordenación clara (países, marcas, sectores, etc.). Cuando asignamos un valor a cada grupo utilizando criterios arbitrarios, como, por ejemplo, el orden alfabético, automáticamente establecemos una prioridad y una ordenación de las categorías que no existen realmente.

Ejemplo de mala codificación de las categorías

Supongamos que queremos explicar el gasto en publicidad de un conjunto de empresas con un modelo de regresión en el que una variable exógena es la comunidad autónoma donde tienen la sede principal. No tiene sentido incluir en el modelo la variable con la codificación por orden alfabético de las diecisiete comunidades autónomas españolas. Si se hiciese así, la diferencia entre Aragón (codificada con un 2) y Andalucía (codificada con un 1) sería de una unidad. El País Vasco (que se codificaría con un 16) se situaría muy lejos de las dos comunidades anteriores. Es decir, implícitamente, supondríamos que el impacto de Aragón sería el doble que el de Andalucía, y que el del País Vasco sería dieciséis veces mayor. Este hecho provocaría graves problemas de interpretación de los resultados del modelo y supondría esta proporcionalidad entre los efectos que, seguramente, no es la que deseamos.

En un modelo de regresión nunca se puede utilizar una variable cualitativa politómica directamente como variable explicativa si su codificación numérica induce un orden (y una distancia) entre las diferentes categorías, que no existe en la realidad.

Consultad las variables politómicas en el subapartado 1.2 de este módulo didáctico.

En muchas situaciones queremos distinguir algunos momentos de tiempo respecto al resto (la estación de verano, la temporada alta, la época de rebajas, etc.). En estos casos hay que hallar una forma para identificar que los datos pertenecen a estos periodos temporales. Ocurre lo mismo cuando queremos identificar datos atípicos de cualquier tipo (temporales o no). Usaremos la misma metodología para resolver todas estas situaciones en la especificación de un modelo de regresión lineal.

1.1. Variables dicotómicas en el modelo de regresión

Lo primero que debe hacerse para poder utilizar **variables explicativas cualitativas** en un modelo de regresión es ver si son dicotómicas o politómicas. En el último caso habrá que definir las variables dicotómicas correspondientes.

Una **variable dicotómica** es aquella que sólo toma dos posibles valores. Habitualmente, las variables ficticias toman los valores 0 y 1.

Dummy

Las variables dicotómicas se llaman también *dummies,* vocablo que proviene del término inglés *dummy*, que podríamos traducir por 'bobo' o 'tonto'.

La razón de utilizar los valores 0 y 1 es la simplicidad en la interpretación de los resultados que se obtienen en la estimación de modelos de regresión si se aplica esta norma. Puesto que en el modelo de regresión lineal cada variable se multiplica por un coeficiente (o parámetro), cuando ésta toma el valor 0, el parámetro no tiene ningún efecto sobre el valor esperado de la variable dependiente. En otras palabras, la interpretación de los parámetros que acompañan a las variables dicotómicas es más sencilla con la codificación con 0 y 1. El parámetro indica cuál es el efecto diferencial sobre el valor esperado de la variable dependiente cuando el individuo tiene la característica identificada con el valor 1 de la variable ficticia, respecto al individuo que tiene la característica identificada con el valor 0.

Consultad la estimación del modelo de regresión lineal múltiple estándar en el apartado 2 del módulo "Modelo de regresión lineal múltiple: especificación..." de esta asignatura.

Evidentemente, las variables dicotómicas podrían codificarse con otros valores distintos de 0 y 1, pero la interpretación de la estimación por mínimos cuadrados ordinarios* de su parámetro asociado (que depende de estos valores) resulta más enrevesada.

*** Recordad que abreviamos *mínimos cuadrados ordinarios* con la sigla *MCO*.**

Ejemplo de introducción de variable dicotómica aditivamente

En un modelo de regresión, supongamos que queremos explicar los salarios de los individuos (Y_i) a partir de sus años de experiencia (X_{1i}) y de su sexo (D_i). El modelo es el siguiente:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 D_i + u_i \quad i = 1, ..., N. \tag{1.1}$$

Hemos definido la variable dicotómica que indica el sexo de la forma siguiente: D_i vale 1 si el individuo es un hombre y 0 si es una mujer.

Nota

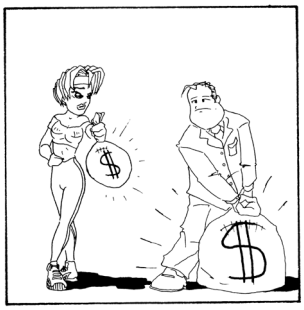
En el modelo suponemos que los individuos están ordenados según su sexo, de modo que los primeros N_1 individuos son hombres. En ese caso, el número de mujeres es $N - N_1$.

Para los hombres el modelo especificado es el siguiente:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 + u_i = \beta_0 + \beta_2 + \beta_1 X_{1i} + u_i \quad i = 1, ..., N_1. \tag{1.2}$$

En cambio, para las mujeres es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i \quad i = N_1 + 1, ..., N. \tag{1.3}$$



Observamos que el hecho de introducir la variable dicotómica relativa al sexo hace que el modelo proporcione una especificación que diferencia las dos categorías. La diferencia se encuentra en el término constante, que para los hombres es igual a $\beta_0 + \beta_2$, mientras que para las mujeres es igual a β_0. Por lo tanto, realizando esta especificación, el modelo establece que la influencia del sexo (que llamaremos *influencia aditiva*) sólo hace variar el nivel del término constante. En cambio, el efecto de los años de experiencia, la otra variable explicativa (X_1), es idéntico para ambos sexos. Observad el gráfico siguiente para entender cuándo conviene usar esta especificación.

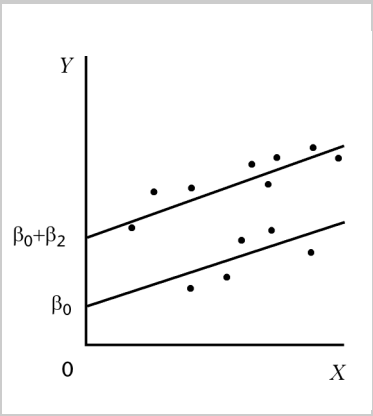
Variables dicotómicas

El modelo especificado con la variable dicotómica que corresponde al sexo establece que los salarios de las mujeres son inferiores a los de los hombres en condiciones de igualdad de experiencia y suponiendo que β_2 es positivo, como se ve a partir de las ecuaciones.

Todavía hay otro hecho importante. A menudo se cree que estimar un modelo de regresión con una variable dicotómica como D_i es equivalente a dividir

la muestra en las dos muestras (hombres y mujeres, en nuestro ejemplo) y estimar dos modelos de regresión separadamente. En la práctica los resultados no serán iguales, ya que las estimaciones MCO de los parámetros del resto de las variables explicativas no serán exactamente las mismas en los tres resultados: el modelo estimado con la variable dicotómica, el modelo estimado para los hombres y el modelo estimado para las mujeres. La razón es que el ajuste mínimo cuadrático de los dos submodelos proporciona estimaciones que permiten que todos los parámetros estimados sean diferentes para los hombres y para las mujeres.

Validez de la variable dicotómica



En el gráfico vemos representados unos datos sobre los salarios de un colectivo de trabajadores.

En el eje de las abscisas tenemos representados los años de experiencia del individuo, y en el eje de las ordenadas, su sueldo. En la representación de la nube de puntos se observa fácilmente que hay dos grupos bien diferenciados. Corresponden a los hombres y a las mujeres. Mientras que la influencia de los años de experiencia es la misma en ambos grupos (la pendiente de la recta de regresión es igual), el nivel salarial entre hombres y mujeres es distinto y se mantiene aproximadamente constante. Por lo tanto, es adecuado plantear el modelo introduciendo en él la variable dicotómica aditivamente.

Por lo tanto, no es cierto que el modelo 1.1 produzca las mismas estimaciones por el método de MCO que las que se obtienen estimando los otros dos modelos de regresión para los dos colectivos, hombres y mujeres, por separado.

Debemos tener en cuenta, también, que en el modelo que contiene todos los datos (hombres y mujeres) suponemos que la varianza del término de perturbación es la misma para los dos subgrupos. Cuando modelizamos ambos colectivos por separado, no utilizamos esta restricción.

El coeficiente que acompaña a una **variable dicotómica con efecto aditivo** en un modelo de regresión se interpreta como el aumento que se producirá en el valor esperado de la variable dependiente cuando el individuo pertenezca a la categoría identificada con el valor unitario de la variable endógena si el individuo perteneciese a la categoría complementaria (valor de la variable dicotómica igual a cero). Este aumento será el mismo, sean cuales sean los valores del resto de las variables explicativas.

Si queremos entender cuál será la interpretación de los contrastes estadísticos en el modelo 1.1, pensemos en lo que significan los parámetros de este modelo. Por ejemplo, el contraste sobre la significación del parámetro β_2 compara si un hombre tiene un nivel esperado de salario igual o diferente del de una mujer, suponiendo que tienen los mismos años de experiencia. El

Consultad la interpretación de los contrastes estadísticos en un modelo en el subapartado 2.7.2 del módulo "Modelo de regresión lineal múltiple: especificación.." de esta asignatura.

contraste del parámetro β_0 en el modelo 1.1 indica si el nivel esperado de salario para las mujeres es significativamente diferente de cero, cuando no se tiene experiencia (X_{1i}). Si se quisiese hacer lo mismo con los hombres (es decir, si se quisiese contrastar si el nivel de salario de los hombres es significativamente distinto de cero cuando no tienen experiencia), se debería contrastar la significación de $\beta_0 + \beta_2$, o sea, utilizar un **contraste de restricciones** con $\beta_0 + \beta_2 = 0$.

El parámetro β_2 en el modelo 1.1...

... indica cuál es el efecto diferencial, es decir, la diferencia de salario entre un hombre y una mujer si tienen los mismos años de experiencia.

Hasta este momento hemos visto cómo podemos introducir en un modelo de regresión una variable explicativa cualitativa, o atributo cualitativo, de dos categorías. Recordemos que de momento sólo tratamos del caso en que tenemos dos únicas alternativas posibles.

Ejemplo de variables dicotómicas

Dentro del ámbito empresarial, podríamos definir, por ejemplo, una variable dicotómica que diferenciase entre pertenecer al sector servicios o no, en cuyo caso emplearíamos la codificación con 1 y 0; o bien, si queremos distinguir las empresas multinacionales del resto, también usaremos el código igual a 1 para las multinacionales y 0 para el resto.

No existe ninguna norma que imponga cómo realizar la asignación de cada categoría de una variable cualitativa a un valor numérico. De todas maneras, existe la costumbre de utilizar ciertas pautas de codificación. En el caso de tener que codificar una respuesta afirmativa o negativa, se suele usar el 1 para las respuestas afirmativas y el 0 para las negativas. En el caso de tener diferenciaciones entre colectivos estableciendo un grupo frente al resto de las categorías, es usual emplear el 1 como el código del grupo y el 0 para el resto.

Ejemplo de diferenciación entre colectivos

Si queremos distinguir, por ejemplo, Cataluña del resto de las comunidades autónomas españolas, codificaremos con un 1 a los individuos de Cataluña y con un 0 a los del resto de las comunidades.

En resumen, la introducción de una variable dicotómica aditivamente en un modelo es equivalente a suponer que los dos colectivos que diferencia la variable dicotómica siguen el mismo modelo de regresión de la población, salvo en lo que se refiere al término independiente.

A continuación, pasamos a ampliar el método presentado en dos sentidos diferentes: por un lado, podemos pensar en atributos que contengan más de dos categorías y, por otro lado, en formas de introducir las variables explicativas en esquemas no aditivos. Empezaremos por el caso de querer introducir una variable explicativa que tenga varias categorías, por ejemplo el estado civil (soltero, casado, viudo. etc.).

1.2. Variables cualitativas politómicas

Supongamos que la variable que pensamos utilizar como explicativa tiene más de dos categorías y que está codificada con valores que van desde 1 hasta el máximo de categorías posibles. Por ejemplo, en el caso del estado civil supongamos que tenemos el valor 1 para las personas solteras, el 2 para las

casadas y el 3 para el resto (viudas, separadas/divorciadas, etc.). En este caso no podemos utilizar directamente la variable, ya que impondríamos un orden y también una proporcionalidad entre los efectos que tal vez no deseamos. No es obligatoriamente cierto que la tercera categoría tenga el triple de efecto que la primera. Lo que hay que hacer es definir tantas variables dicotómicas como el número total de categorías excepto una.

A menudo las variables dicotómicas que definen las categorías de una variable politómica se denominan *variables ficticias*. Puesto que todas las variables ficticias son dicotómicas, es habitual hablar indistintamente de variables dicotómicas y de variables ficticias.

En el ejemplo del estado civil, como se muestra acto seguido en la tabla, hay que definir dos variables ficticias, D_2 y D_3, del modo siguiente:

$$D_2 = \begin{cases} 1 \text{ casados} \\ 0 \text{ otros} \end{cases}$$

$$D_3 = \begin{cases} 1 \text{ viudos, etc.} \\ 0 \text{ otros} \end{cases}$$

La primera variable identifica con un 1 a los individuos casados y con un 0 al resto*. La otra identifica con un 1 a los individuos separados, viudos, etc. y con un 0 a los solteros y los casados.

*** Tanto los solteros como los viudos, los separados, etc.**

Creación de dos variables ficticias

Estado civil	D_2	D_3
1	0	0
2	1	0
2	1	0
1	0	0
3	0	1
2	1	0
3	0	1
1	0	0
3	0	1

El primer individuo de la base de datos es un soltero; por lo tanto, la primera variable dicotómica tiene el valor 0, ya que el individuo no está casado, y la segunda también, porque tampoco es viudo ni está separado. El segundo individuo está casado; por lo tanto, tiene el valor 1 en la primera variable dicotómica y 0 en la otra. Observad que los individuos solteros tienen los valores 0 en las dos variables dicotómicas. Los individuos casados tienen el valor 1 en la primera y el 0 en la segunda. Los individuos del resto de los estados civiles tienen una 0 en la primera variable ficticia y un 1 en la segunda.

Los individuos que corresponden al valor 0...

... de todas las variables dicotómicas del modelo forman parte de lo que se denomina *categoría base* o *categoría de referencia*. En el ejemplo de variables dicotómicas, para representar el estado civil considerado en el texto, la categoría de referencia es la ley de los individuos solteros.

La información respecto al estado civil se puede dar de varias maneras equivalentes; proponemos las dos siguientes: con una única variable codificada con tres valores (la primera columna de la tabla anterior) o bien con dos variables dicotómicas como las anteriores D_2 y D_3. Puede aplicarse el mismo razonamiento para cualquier variable cualitativa. En general, se utiliza el cri-

terio de definir tantas variables ficticias como el número total de categorías menos una.

Existe una razón fundamental para no considerar a una de las categorías en la definición de las variables ficticias. Esta razón tiene relación con el tema de la multicolinealidad, que ya hemos estudiado. Sabemos que las observaciones de las variables explicativas de un modelo de regresión no pueden tener una relación lineal exacta entre sí. Si en el ejemplo anterior definiésemos una tercera variable ficticia (D_1) que fuese igual a 1 para las personas solteras e igual a 0 para el resto, la suma de las tres variables ficticias sería una variable igual a 1 en todas las observaciones. Por lo tanto, si introdujésemos las tres variables ficticias en un modelo de regresión con término constante, tendríamos claramente el problema de la multicolinealidad exacta, ya que la suma de las tres variables ficticias sería 1 (proporcional al término independiente) y no podría estimarse el modelo. Vemos que el mismo problema se repite sea cual sea el número de categorías de la variable cualitativa.

Consultad el concepto de multicolinealidad en los subapartados 2.1, 2.2 y 2.3 del módulo "Errores de especificación, multicolinealidad y observaciones atípicas" de esta asignatura.

El criterio para determinar el número de variables ficticias necesarias para tener en cuenta todas las categorías asociadas a la variable politómica es el siguiente: cuando se desea introducir una variable cualitativa de p categorías en un modelo de regresión, de deben definir $p - 1$ variables dicotómicas que identifiquen con un 1 a los individuos de cada categoría respectivamente y con un 0 al resto.

Obligatoriamente, siempre que se desee mantener el término constante en el modelo de regresión, se tendrá que excluir a una de las categorías en la construcción de las variables ficticias para evitar el problema de la multicolinealidad perfecta.

La categoría que no se utiliza para la definición de las variables ficticias se denomina **categoría de referencia** o **categoría base**.

Existen dos soluciones posibles al problema de introducir variables cualitativas:

1) La solución más usual es introducir las variables ficticias en el modelo de regresión con término independiente. Se habrán definido previamente tantas como el número de categorías totales menos una e identificarán a los individuos que pertenezcan a grupos diferentes, excepto a un grupo.

2) Alternativamente, aunque no es demasiado habitual, se podrán incluir tantas variables ficticias como categorías reúna el atributo, pero en ese caso se tendrá que eliminar el término independiente del modelo con el propósito de evitar el problema de la multicolinealidad perfecta.

En cualquier caso, la interpretación de los parámetros asociados a las variables ficticias no es la misma para las dos posibles especificaciones, como veremos ahora.

Interpretaciones diferentes para soluciones diferentes

Supongamos que, en el ejemplo anterior sobre el nivel esperado de salarios, queremos tener en cuenta el estado civil en lugar del sexo; en ese caso, podemos hacer dos cosas:

1) Podemos introducir en el modelo las variables D_2 y D_3:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \quad i = 1, ..., N. \tag{1.4}$$

En este modelo con término independiente, suponiendo que la variable ficticia D_2 indica con un 1 que el individuo está casado, entenderemos que el efecto medido por β_2 es la diferencia esperada que existiría entre una persona casada y una soltera si hubiese igualdad en el resto de las variables explicativas*. Como podemos ver según el modelo, esperaríamos que, con los mismos años de experiencia, las personas casadas ganasen β_2 más unidades que las solteras, mientras que las personas viudas o las separadas ganarían β_3 más unidades que las solteras. La diferencia entre un individuo casado y uno viudo, con los mismos años de experiencia, estaría determinada por $\beta_2 - \beta_3$.

*** En este caso la única variable explicativa es, aparte de las dicotómicas relativas al estado civil, X_1.**

2) Podemos eliminar el término constante en el modelo y entonces no se tiene categoría de referencia. En este caso hay que emplear una tercera variable D_4 como variable ficticia que identifica a las personas solteras, y el modelo resultante queda especificado del modo siguiente:

$$Y_i = \delta_1 X_{1i} + \delta_2 D_{2i} + \delta_3 D_{3i} + \delta_4 D_{4i} + u_i \quad i = 1, ..., N. \tag{1.5}$$

En el modelo 1.5 hay que interpretar δ_2 como el valor esperado de la variable dependiente (el salario) para las personas casadas cuando X_1 vale cero (sin experiencia). Los parámetros δ_3 y δ_4 se interpretarían análogamente para la categoría de las personas viudas/separadas y de las solteras, respectivamente. No obstante, observad que, si en el modelo hay más de una variable explicativa cualitativa, esta última forma de especificar el modelo es ciertamente incómoda.

Equivalencia de las dos soluciones

Tal como hemos especificado los modelos 1.4 y 1.5, las equivalencias entre los parámetros se muestran en la tabla siguiente:

Valor esperado del sueldo cuando $X_j = 0$		
Modelo	**1.4**	**1.5**
Solteros	β_0	δ_4
Casados	$\beta_0 + \beta_2$	δ_2
Otros (viudos, etc.)	$\beta_0 + \beta_3$	δ_3

La recomendación general es utilizar una categoría de referencia y mantener el término independiente en el modelo, ya que simplifica mucho las interpretaciones de los resultados de las estimaciones de los parámetros. De todos modos, estas interpretaciones siempre dependerán de cómo se hayan definido las variables ficticias.

Recordad que las interpretaciones...

... de los parámetros de las variables ficticias, las deberemos referir siempre respecto a la categoría base, si el modelo tiene un término constante.

1.2.1. Variables ficticias espaciales

Una aplicación de la metodología explicada hasta ahora se emplea cuando deseamos diferenciar unidades muestrales que se encuentran situadas en zonas o espacios diferentes. El procedimiento es el mismo que acabamos de explicar, y lo reproducimos a continuación:

- Se codifica la zona, ya que se trata de una variable cualitativa posiblemente con más de dos categorías.

- A continuación, se definen las variables ficticias correspondientes, tantas como el total de las zonas excepto una, la categoría de referencia.

- Si sólo se quieren considerar efectos de cambio del nivel esperado de la variable endógena uniformemente para todos los valores de las variables exógenas, las variables ficticias se introducen aditivamente en el modelo.

Todas las interpretaciones de los resultados medirán el diferencial respecto a la zona de referencia que no se ha identificado con una variable ficticia. La decisión de cuál es la zona de referencia depende de la situación concreta que se estudie, pero no es recomendable establecer que la categoría de referencia sea la más heterogénea o la que queda menos delimitada.

Variables ficticias espaciales en un modelo referido a empresas

Si en un modelo de regresión referido a empresas deseamos introducir variables explicativas que indiquen cuál es su sector de actividad (agrícola, industrial y de servicios), definimos dos variables ficticias; por ejemplo:

$$C_1 = \begin{cases} 1 & \text{agrícola} \\ 0 & \text{otros} \end{cases}$$

$$C_2 = \begin{cases} 1 & \text{industrial} \\ 0 & \text{otros} \end{cases}$$

Entonces, en este caso, la categoría base es la correspondiente al sector de servicios.

Si, análogamente, en un modelo de regresión referido a empresas queremos introducir variables explicativas que indiquen en qué lugar tienen su principal mercado (podríamos definir cuatro zonas: Europa, América, África y Asia-Oceanía), construiremos tres variables ficticias, y dejaremos el término independiente en el modelo de regresión; por ejemplo:

$$C_1 = \begin{cases} 1 & \text{América} \\ 0 & \text{otros} \end{cases}$$

$$C_2 = \begin{cases} 1 & \text{África} \\ 0 & \text{otros} \end{cases}$$

$$C_3 = \begin{cases} 1 & \text{Àsia-Oceanía} \\ 0 & \text{otros} \end{cases}$$

En este último caso, la categoría base es la correspondiente a Europa.

1.2.2. Variables ficticias temporales

Si se tiene una base de datos cuyas variables hacen referencia a instantes del tiempo diferentes, las variables ficticias temporales se pueden utilizar para identificar intervalos o periodos del tiempo concretos. Igualmente, estos periodos se codificarán con un 1 y el resto, con un 0. Cuando los coeficientes que acompañan a estas variables se introduzcan en un modelo de regresión aditivamente, medirán el impacto sobre la variable endógena de la situación temporal indicada por la variable ficticia respecto al resto de los periodos.

Esta metodología es útil, como veremos con los ejemplos posteriores, para considerar cambios estructurales o impactos instantáneos en los modelos.

Variables ficticias temporales en un modelo referido a empresas

Si deseamos introducir en un modelo de regresión referido a la evolución de una empresa a lo largo del tiempo una variable ficticia (E_1) que señale, por ejemplo, que en el mes de agosto no hay actividad, definimos:

$$E_1 = \begin{cases} 1 & \text{agosto} \\ 0 & \text{otros meses} \end{cases}$$

Finalmente, incluiremos esta variable directamente en el modelo. Dado que distinguimos agosto del resto de los meses, utilizamos esta variable dicotómica.

1.3. Interpretación de los coeficientes de variables ficticias

En los subapartados anteriores, hemos insistido en la introducción en el modelo de regresión de variables dicotómicas, bien como variables explicativas cualitativas con dos categorías únicas, bien como variables ficticias que se introducen simultáneamente para poder tener en cuenta las múltiples categorías posibles de una variable politómica.

A continuación, plantearemos algún ejemplo de cómo se interpretan los resultados de las estimaciones de modelos que incluyan este tipo de variables. Asimismo, ampliaremos la metodología dada en el párrafo anterior, para poder considerar posibles efectos multiplicativos con otras variables e interacciones entre variables cualitativas diferentes.

1.3.1. Interpretación del esquema aditivo y multiplicativo

Supongamos que continuamos con el ejemplo empleado a lo largo de este apartado sobre el salario de un colectivo. Ahora, introduciremos un nuevo elemento que complique el modelo: esperamos que el impacto que tienen los años de experiencia sobre el nivel esperado de salario no sea el mismo para los hombres que para las mujeres. Entonces, todavía deberemos distinguir entre los dos casos siguientes: suponer que es el mismo, es decir, que cuando no hay experiencia el salario esperado es el mismo para los hombres que para

Consultad el ejemplo de introducción de una variable dicotómica aditivamente en un modelo en el subapartado 1.1 de este módulo didáctico.

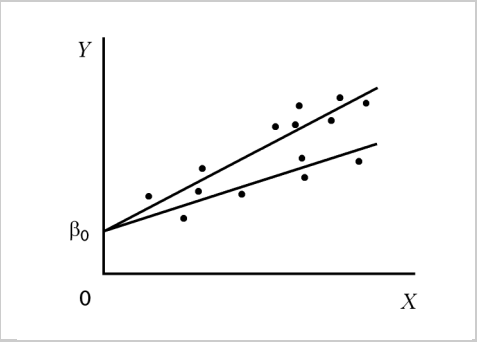
las mujeres, o bien que es diferente. La apariencia de los datos es la que se muestra en los gráficos siguientes, donde vemos representados unos datos sobre los salarios de un colectivo de trabajadores.

1. Variación del diferencial del nivel salarial con la experiencia

En este gráfico vemos que la influencia de los años de experiencia no es la misma en ambos grupos (la pendiente de la recta de regresión no es igual).

Se aprecia que, si existe igualdad en cuanto a los años de experiencia, el diferencial de nivel salarial entre hombres y mujeres crece, se hace mayor.

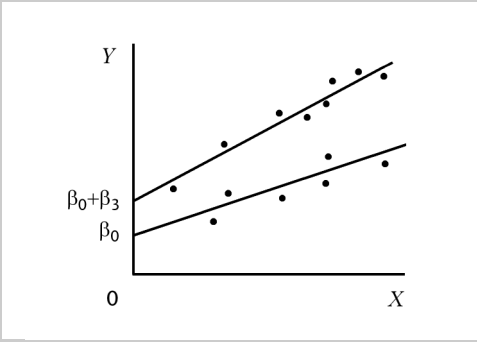
De todos modos, la diferencia para los que no tienen nada de experiencia o tienen muy poca es prácticamente nula.



2. Variación del diferencial del nivel salarial con la experiencia

En este gráfico vemos que la situación respecto al salario ya es diferente cuando no hay experiencia y, dado que la pendiente de la nube de puntos (y, por tanto, de la recta de regresión) no es igual para los dos subgrupos, supondremos que el efecto de la experiencia no es el mismo para los hombres que para las mujeres.

Se aprecia que, si hay igualdad en los años de experiencia, el diferencial de nivel salarial entre hombres y mujeres aún crece más.



En el eje de las abscisas, tenemos los años de experiencia del individuo, y en el eje de las ordenadas, su sueldo. Como en el ejemplo citado, la nube de puntos muestra dos grupos diferentes, que corresponden a los hombres y a las mujeres.

En las situaciones descritas en los gráficos anteriores es adecuado plantear un **modelo con efectos multiplicativos**, es decir, se debe introducir la variable dicotómica multiplicativa en el modelo.

Los modelos que hay que utilizar en estas situaciones son los que se escriben a continuación. Recordemos que en la ecuación 1.1 hemos definido una variable dicotómica (D) que identifica con un 1 a los hombres y con un 0 a las mujeres. En primer lugar, detallaremos los modelos y a continuación indicaremos cuáles son sus ventajas y su interpretación. Los modelos posibles son los siguientes:

Nota

Para comprender bien los efectos asociados a cada modelo, cambiad D_i por 1 cuando queráis ver los efectos para los hombres. Análogamente, poned D_i igual a 0 para las mujeres.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 D_i + u_i \quad i = 1, ..., N, \tag{1.6}$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i} D_i + u_i \quad i = 1, ..., N, \tag{1.7}$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i} D_i + \beta_3 D_i + u_i \quad i = 1, ..., N. \tag{1.8}$$

La interpretación de cada modelo es la siguiente:

1) El modelo 1.6 es el **modelo aditivo**, donde el efecto del sexo (β_2) es constante para todos los años de experiencia y sólo hace cambiar el término independiente. Este modelo corresponde al comportamiento del gráfico del ejemplo de introducción de una variable dicotómica aditivamente en un modelo, del cual hemos partido.

2) En el **modelo multiplicativo** 1.7, el sexo afecta al impacto de los años de experiencia sobre el sueldo esperado; es decir, para los hombres este impacto será igual a $\beta_1 + \beta_2$, mientras que para las mujeres será igual a β_1. Esta situación corresponde al gráfico 1 de este subapartado.

3) En el **modelo aditivo y multiplicativo** 1.8, no tendremos el mismo efecto de la variable explicativa de los años de experiencia, y ahora el sueldo esperado para un hombre sin experiencia será igual a $\beta_0 + \beta_3$, mientras que para una mujer será igual a β_0 (gráfico 2).

Conclusiones

Veamos las conclusiones de la interpretación de los modelos:

- En el modelo aditivo: cuando no se posee experiencia, el sueldo esperado es β_0 para las mujeres y $\beta_0+\beta_2$ para los hombres.
- En el modelo aditivo y multiplicativo: para los hombres tener un año más de experiencia hace aumentar el sueldo esperado en $\beta_1+\beta_2$ unidades. En cambio, para las mujeres el aumento esperado de sueldo por cada año de experiencia es de β_1 unidades.

En conclusión, podemos decir que cada modelización con variables dicotómicas necesita una especificación adecuada al comportamiento que deseemos establecer.

Debemos darnos cuenta de que, en los casos anteriores, el modelo 1.8 es el más general de todos y de que, por medio de contrastes individuales o múltiples de parámetros, podemos contrastar si los otros dos modelos son adecuados. Por ejemplo, si estimamos el modelo 1.8 e inferimos $\beta_3 = 0$, estamos contrastando, en realidad, el modelo 1.7 en la hipótesis nula.

En la práctica, hay muchas posibilidades, y el principal inconveniente de usar variables cualitativas explicativas es que, si el número de categorías es alto*, y si además se quieren considerar efectos multiplicativos, el modelo especificado debe contener un número muy elevado de parámetros. Como ya hemos visto, no se recomienda que el número de variables explicativas del modelo sea excesivo.

*** Por lo tanto, también lo es el número de variables ficticias introducidas en el modelo.**

En cuanto al número de variables explicativas del modelo, consultad el subapartado 2.3.2 del módulo "Modelo de regresión lineal múltiple: especificación.." de esta asignatura.

1.3.2. Interpretación de las interacciones

Cuando queremos utilizar distintas variables explicativas cualitativas en un modelo de regresión, es usual pensar que puede haber efectos multiplicativos entre ellas. En este caso, los efectos se denominan **efectos cruzados** o **interacciones**.

Ahora combinaremos los ejemplos utilizados en los apartados anteriores. Si la variable dicotómica *S* indica el sexo de un individuo (le hemos dado este nombre para evitar confusiones) y las variables dicotómicas que llamaremos D_2 y D_3 son las dos variables ficticias del estado civil, con la misma definición que hemos visto antes, podríamos tener el modelo siguiente:

Consultad el ejemplo de introducción de una variable dicotómica aditivamente en un modelo y la definición de D_2 y D_3 en los subapartados 1.1 y 1.2 de este módulo didáctico, respectivamente.

$$Y_i = \beta_0 + \beta_1 D_{2i} + \beta_2 D_{3i} + \beta_3 S_i + u_i \quad i = 1, ..., N. \tag{1.9}$$

No hemos introducido la variable de los años de experiencia con el propósito de simplificar el modelo. En este modelo sólo hay efectos aditivos, es decir:

- β_1 representa la diferencia del valor esperado de la variable endógena entre los individuos casados y los solteros.
- β_2 representa la diferencia entre los individuos separados/viudos y los solteros.
- β_0 es el nivel esperado de la variable endógena para las mujeres solteras, es decir, la categoría base.
- β_3 es la diferencia del sueldo esperado entre hombres y mujeres que tienen el mismo estado civil.

Es importante que entendamos estas interpretaciones correctamente. Resulta muy práctico escribir los valores de las variables ficticias para cada uno de los grupos y ver que, si éstas son cero, el efecto del parámetro desaparece. Una forma de hacerlo es construir una tabla como la siguiente:

Modelo sin interacción		
	E(*Y*)	
Estado civil	**Hombres**	**Mujeres**
Solteros	$\beta_0 + \beta_3$	β_0
Casados	$\beta_0 + \beta_1 + \beta_3$	$\beta_0 + \beta_1$
Viudos/separados	$\beta_0 + \beta_2 + \beta_3$	$\beta_0 + \beta_2$

Como vemos en la tabla anterior, siempre esperamos que los hombres ganen β_3 unidades más que las mujeres que tienen el mismo estado civil.

Ahora introduciremos efectos cruzados (las interacciones) entre las variables, con el fin de posibilitar un modelo más general:

$$Y_i = \beta_0 + \beta_1 D_{2i} + \beta_2 D_{3i} + \beta_3 S_i + \beta_4 D_{2i} S_i + \beta_5 D_{3i} S_i + u_i \quad i = 1, ..., N. \tag{1.10}$$

En este modelo, los efectos de las interacciones hacen que el modelo sea más amplio. Lo mismo sucedía en el caso del esquema multiplicativo que hemos tratado antes. Los parámetros de las interacciones tienen interpretaciones

Consultad el esquema multiplicativo en el subapartado 1.3.1 de este módulo didáctico.

más complejas. En este caso, hay que establecer del modo siguiente cuál es la categoría de referencia para las dos variables cualitativas: para el sexo lo son las mujeres y para el estado civil, las personas solteras; por lo tanto, todas las interacciones se interpretarán con respecto a la categoría conjunta siguiente: mujeres solteras.

Para poder entender mejor cómo se interpretarían los parámetros estimados, veamos la tabla siguiente:

Modelo con interacción		
	E(*Y*)	
Estado civil	**Hombres**	**Mujeres**
Solteros	$\beta_0 + \beta_3$	β_0
Casados	$\beta_0 + \beta_1 + \beta_3 + \beta_4$	$\beta_0 + \beta_1$
Viudos/separados	$\beta_0 + \beta_2 + \beta_3 + \beta_5$	$\beta_0 + \beta_2$

Para los individuos solteros, el sueldo esperado de los hombres es β_3 unidades más elevado que el de las mujeres. Para los individuos casados, el sueldo esperado de los hombres es $\beta_3 + \beta_4$ unidades más elevado que el de las mujeres. El parámetro β_4 muestra el efecto diferencial hombre/mujer de las personas casadas respecto al efecto diferencial hombre/mujer de las solteras. El parámetro β_4 se activará (es decir, no será nulo) cuando tengamos un hombre casado, ya que entonces tanto D_{2i} como S_i serán igual a 1. Por lo tanto, este parámetro indica la diferencia que hay que añadir a β_3 para obtener el valor esperado de la variable endógena cuando se desea calcular el valor esperado para un hombre casado.

Por otra parte, para los individuos separados/viudos esperaremos que los hombres ganen $\beta_3 + \beta_5$ unidades más que las mujeres. Un hombre viudo gana $\beta_2 + \beta_5$ unidades más que un hombre soltero. Una mujer viuda gana β_2 unidades más que una mujer soltera. Por lo tanto, β_5 hace referencia a la diferencia entre el incremento esperado de salario de un hombre viudo/separado y el de un hombre soltero, respecto a la diferencia que existiría si fuesen mujeres.

Recordemos que la diferencia esperada de salario entre un hombre soltero y una mujer soltera es β_3. Si queremos hacer las comparaciones respecto a un hombre casado y una mujer soltera, debemos añadir a β_0 (el valor de una mujer soltera) el incremento β_3 por el hecho de ser hombre, el incremento β_1 por el hecho de estar casado y el incremento β_4 por el hecho de ser un hombre casado.

Cuando queremos introducir interacciones entre dos variables cualitativas, hay que usar tantas variables ficticias como categorías observadas menos una. A continuación se deben incluir estas variables ficticias y todos los posibles productos cruzados entre las variables ficticias de atributos diferentes.

La interpretación concreta de los parámetros es como la que ejemplificamos aquí, siempre y cuando se hayan utilizado las codificaciones iguales a 0 y 1. En otro caso, la comprensión de los resultados puede complicarse más. Observamos también que, si llegase el caso y especificásemos un modelo sólo con los efectos cruzados (sin las variables ficticias simples), la interpretación sería más difícil. Esta especificación se denomina **especificación no jerárquica**.

La dificultad en el procedimiento de estimación no es más elevada cuando hay variables explicativas cualitativas. Una vez que hemos definido las variables ficticias y hemos especificado el modelo procurando que no se produzca multicolinealidad, el método de estimación MCO no tiene ninguna otra complicación añadida. En este contexto interpretamos los contrastes estadísticos sobre la significación individual de los parámetros como contrastes sobre la existencia de diferencias en el valor esperado de la variable endógena para las diferentes categorías.

1.4. Utilización de variables ficticias

La utilización de variables ficticias no se limita sólo a la necesidad de distinguir grupos diferentes dentro de la muestra (sexo, estado civil, país de origen, etc.). Como ya se ha mencionado antes, en ciertas situaciones las podemos emplear también para distinguir periodos temporales o localizaciones diferentes. A continuación mostramos algunos otros usos posibles de las variables ficticias en el contexto del modelo de regresión.

Consultad el uso de variables ficticias para distinguir periodos temporales o localizaciones diferentes en los subapartados 1.2.1 y 1.2.2 de este módulo didáctico.

1.4.1. Datos atípicos

Como ya hemos visto, los datos atípicos constituyen un problema del modelo de regresión, porque pueden distorsionar los resultados de la estimación.

Consultad la problemática de los datos atípicos en el subapartado 2.4 del módulo "Errores de especificación, multicolinealidad y observaciones atípicas" de esta asignatura.

Existen distintas posibilidades de tratamiento una vez que ya se han detectado los datos atípicos. Si se desea conservarlos en la base de datos, pero eliminar su influencia sobre las estimaciones, se pueden emplear variables ficticias.

Supongamos que en una base de datos disponemos de N observaciones. Supongamos también que la observación i_0-ésima es un dato atípico. En este caso se define la variable ficticia que es igual a 1 para la observación i_0-ésima e igual a 0 para el resto, como presentamos a continuación:

$$D_i = \begin{cases} 1 & \text{si} \quad i = i_0 \\ 0 & \text{si} \quad i \neq i_0 \end{cases}$$

Un dato puede ser atípico...

... porque ha habido un error de medida, porque se trata de un individuo especial o porque está en un momento temporal en que ha ocurrido un acontecimiento extraordinario.

Por lo tanto, se trata de una variable ficticia que tiene todos sus valores iguales a cero y un único valor igual a la unidad. Si se introduce esta variable ficticia en el conjunto de las variables explicativas del modelo, se elimina la influencia de esta observación.

La ventaja de este procedimiento es que el residuo de la estimación será cero para esta observación. Esto se debe al hecho de que el parámetro que acompaña a la variable ficticia se ajusta mediante el criterio MCO; por lo tanto, su valor estimado es el que hace menor el residuo para esta observación. Puesto que el parámetro sólo tiene efecto para esta observación, ya que la variable ficticia es cero para el resto de las observaciones, es posible convertir en cero su residuo. En consecuencia, el modelo ya ajustará exactamente esta observación y dejará de influir en la estimación del resto de los parámetros y sus estadísticos correspondientes. Igualmente, se habrá aislado su efecto en este único parámetro y se podrá estimar el impacto de esta observación o acontecimiento concreto.

Los inconvenientes son que los resultados obtenidos para el resto de los parámetros serían los mismos que si se hubiese eliminado la observación y que el modelo tiene una variable explicativa adicional.

Otra cuestión añadida que aparece en la identificación de observaciones atípicas es su utilidad práctica. Si para una observación cualquiera i_0 fijamos el valor de la variable dependiente en cero y definimos una variable ficticia que es igual a cero para el resto de las observaciones e igual a 1 para ésta, la estimación MCO del parámetro de la variable ficticia es una estimación del error de predicción que cometería el modelo para esta observación si se pretendiese realizar una predicción de la variable endógena. El error estándar de la estimación del parámetro proporcionaría la estimación del error de predicción puntual. De esta manera, los errores de predicción y sus intervalos de confianza se pueden calcular fácilmente con cualquier programa informático que facilite las estimaciones MCO de un modelo de regresión.

Si hubiese más de una observación atípica...

... se podrían poner más variables dicotómicas (una para cada una) o definir una variable dicotómica única que las identificase a todas. En este último caso los residuos no necesariamente tendrían que valer cero.

Ejemplo de tratamiento de datos atípicos con variables dicotómicas

A partir de los datos utilizados en el módulo "Errores de especificación, multicolinealidad y observaciones atípicas", hemos considerado el último ejemplo que contiene un dato atípico. Hemos especificado el mismo modelo que en aquel apartado, pero incluyendo en él una variable dicotómica que toma el valor 1 para aquella observación.

Consultad la estimación del modelo 2.3 a partir de los datos de la tabla del subapartado 2.4 del módulo "Errores de especificación, multicolinealidad y observaciones atípicas" de esta asignatura.

Observad los resultados y veréis cómo la estimación del modelo produce las mismas estimaciones de los parámetros que cuando esta observación no existía y que, además, su residuo vale ahora cero.

El modelo estimado es el siguiente:

$$Y_{5i} = \beta_1 + \beta_2 X_{5i} + \beta_3 D_i + u_i \quad \forall i = 1, ..., N.$$

La variable D_i vale 1 para la observación 26 y 0 para el resto.

Estimación del modelo con el dato atípico

Variable dependiente: Y_5

Número de observaciones: 26

VARIABLES	COEFICIENTES	ERROR STD.	ESTADÍSTICO T	SIGNIFICACIÓN
C	**27.030013**	**0.239810**	**112.71416**	0.0000
X_5	**1.8741038**	**0.392824**	**4.7708407**	0.0001
D	-0.448093	0.239810	-0.498834	0.6231

R-squared	**0.626888**	Mean of dependent var	28.10958
Adjusted R-squared	0.594435	**S.D. of dependent var**	**0.998368**
S.E.of regression	**0.629501**	**Sum of squared resid**	**9.114254**
		F-statistic	**19.32120**
		Prob(F-statistic)	**0.000096**

1.4.2. Cambio estructural

En el caso del cambio estructural, el procedimiento empleado es análogo al anterior. La única diferencia es la definición de la variable ficticia.

Consultad el concepto de permanencia estructural en el subapartado 2.2.4 del módulo "Modelo de regresión lineal múltiple: especificación.." de esta asignatura.

La identificación de un cambio estructural se lleva a cabo de la manera siguiente: para los periodos anteriores al cambio, la variable ficticia se iguala a cero y, para los periodos en que el cambio ya se ha producido, se codifica como 1. El parámetro asociado a esta variable ficticia mide el impacto del cambio estructural en la esperanza matemática de la variable endógena. Su contraste individual evaluará la significación estadística del cambio estructural y, en definitiva, su existencia.

En el caso de querer establecer efectos multiplicativos del cambio estructural, se procede igual que como se ha explicado antes para las variables ficticias.

Consultad el método empleado para introducir efectos multiplicativos en un modelo en el subapartado 1.3.1 de este módulo.

Variables ficticias y contraste de Chow

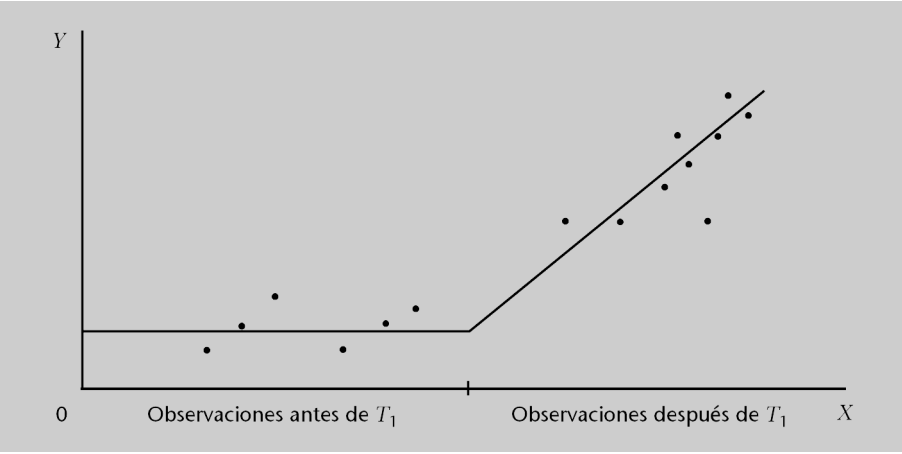
Consultad el contraste de Chow de permanencia estructural y el modelo 3.34 del módulo "Modelo de regresión lineal múltiple: especificación.." de esta asignatura.

Hay mucha relación entre el contraste de Chow (de permanencia estructural) y la definición de variables ficticias para detectar un cambio estructural.

Supongamos el mismo modelo que ya habíamos utilizado para hacer el contraste de Chow, concretamente el modelo siguiente:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad t = 1, ..., T.$$

Supongamos que deseamos contrastar la existencia de un cambio estructural, es decir, que los parámetros de la población se mantienen iguales a lo largo de los dos subperiodos: el primero desde 1 hasta T_1 y el segundo desde $T_1 + 1$ hasta T. Esta hipótesis puede verse en el gráfico que se muestra a continuación:



En esta situación podríamos definir una variable dicotómica D_t que fuese igual a 0 para el primer subperiodo e igual a 1 para el segundo, es decir:

$$D_t = \begin{cases} 1 & \text{si} \quad t = 1, ..., T_1 \\ 0 & \text{si} \quad t = T_1 + 1, ..., T \end{cases}$$

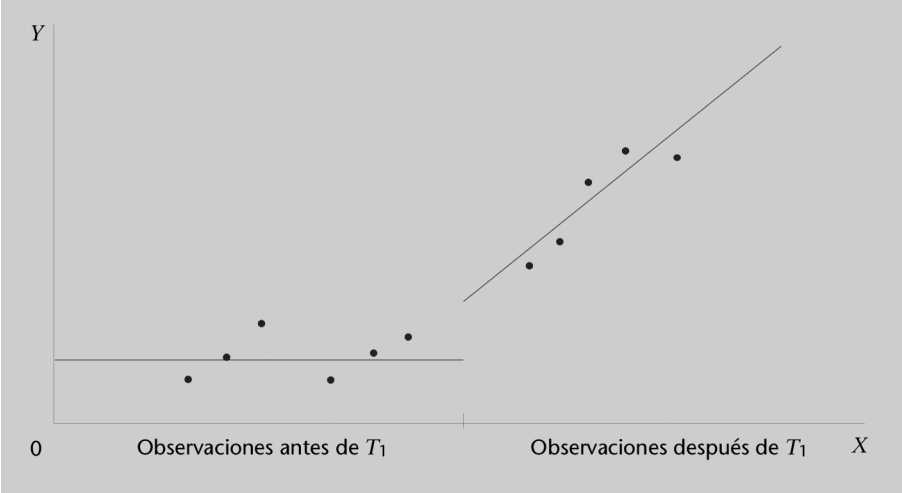
A continuación, se especificaría el modelo incluyendo en él el efecto multiplicativo de esta variable ficticia, con el fin de captar el cambio de pendiente en la segunda submuestra.

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 X_t D_t + u_t \quad t = 1, ..., T.$$

Entonces, el contraste individual del efecto multiplicativo sería totalmente equivalente al contraste de Chow.

Análogamente, si existiesen más variables explicativas, habría que incluir todos los efectos multiplicativos correspondientes entre cada variable explicativa y la variable ficticia, y realizar el contraste de cambio estructural observando el contraste sobre el conjunto de parámetros asociados a los efectos multiplicativos o aditivos.

Supongamos que deseamos especificar un modelo similar al anterior, pero adaptándolo a la estructura del gráfico siguiente:



Entonces, el modelo sería el siguiente:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 X_t D_t + \beta_4 D_t + u_t \quad t = 1, ..., T.$$

Con la utilización de la variable dicotómica puede verse si la causa del cambio estructural es un cambio de pendiente o bien un cambio de nivel. Esto se consigue haciendo un contraste estadístico de significación individual del parámetro β_3 y del parámetro β_4, respectivamente.

1.4.3. Estacionalidad

Para detectar o poder tratar problemas de estacionalidad, se pueden utilizar variables ficticias que tomen el valor 0 fuera del periodo estacional e igual a 1 en el periodo estacional. Una vez que se ha introducido la variable en el modelo, su parámetro se interpreta como el impacto de la época estacional designada en el valor esperado de la variable dependiente.

Consultad el ejemplo de variables ficticias temporales en un modelo referido a empresas en el apartado 1.2.2 de este módulo didáctico como ejemplo concreto donde se usan variables ficticias para remediar la estacionalidad.

Ejemplo de tratamiento de problemas de estacionalidad

Uno de los ejemplos clásicos de la utilización de variables ficticias para tratar problemas de estacionalidad es un caso en el que se quiere explicar el volumen de ventas de helados de una empresa del sector, en función del precio del producto y también del trimestre del año.

Habría que incorporar en el modelo tres variables ficticias. Sería recomendable que la primera identificase la primavera; la segunda, el otoño, y la tercera, el invierno. De este modo, todas las comparaciones se harían con respecto al verano (el periodo en el que los helados tienen más éxito). Si introdujésemos las tres variables ficticias anteriores en el modelo de regresión, sus parámetros asociados identificarían la caída esperada de ventas en cada trimestre respecto al trimestre de verano. Fijémonos en que, si sólo incorporásemos una variable dicotómica que identificase la estación del verano, estaríamos suponiendo que el resto de las estaciones tienen el mismo comportamiento.

El modelo podría hacerse más general si también se tuviesen en consideración efectos multiplicativos.

1.4.4. El modelo de efectos fijos

Cuando tenemos datos sobre N individuos a lo largo de T periodos, decimos que tenemos un **panel de datos**. Por ejemplo, podríamos hacer un seguimiento de precios de venta y unidades vendidas de un determinado producto para quince empresas productoras a lo largo de un periodo de veinticuatro meses.

Una situación usual se produce cuando deseamos especificar un modelo de regresión con el total de $N \cdot T$ observaciones y queremos eliminar los efectos temporales y de discrepancia entre las unidades. En el ejemplo, esto significaría que en el estudio de la relación precio/demanda no queremos tener en cuenta los factores debidos al mes en que nos encontramos ni los efectos particulares de ciertas empresas. Para hacerlo, podemos definir $N - 1$ y $T - 1$ variables ficticias que identifiquen todas las unidades (excepto la de referencia) y todos los periodos (excepto el de referencia), y podemos incluirlas aditivamente en el modelo de regresión. Este modelo se denomina **modelo de efectos fijos**.

En el ejemplo, se definen, en primer lugar, catorce variables dicotómicas, donde cada una identifica a una de las empresas, y se deja a una empresa en la categoría de referencia. Es recomendable dejar a una de las empresas más importantes en la categoría base, ya que todas las interpretaciones de los parámetros se harán con respecto a ella. Por otra parte, tomando el primer mes como el de referencia, se definen veintitrés variables ficticias que identifiquen a cada uno de los veintitrés meses restantes.

Observad que, en este ejemplo, el número total de observaciones es $15 \cdot 24 = 360$ y el número de parámetros que hay que estimar es igual a $2 + 23 + 14 = 39$. Hemos añadido dos parámetros, ya que consideramos un término independiente y otro parámetro de efecto de la demanda sobre el precio. El resto de los parámetros asociados a las variables ficticias se interpretará como los efectos singulares debidos al mes (o a la empresa) que identifican, respecto al periodo (o empresa) de referencia.

Los coeficientes de las variables dicotómicas que identifican unidades y periodos en un modelo de efectos fijos se interpretan como desplazamientos de la recta de regresión debidos a efectos (fijos) de variables no observables.

Los efectos que se han definido en el ejemplo...

... del modelo de efectos fijos se podrían llegar a considerar aleatorios, pero en este módulo no desarrollaremos la metodología para poder estimarlos.

El principal inconveniente del modelo de efectos fijos es que la inclusión de las variables ficticias hace aumentar el número de parámetros que hay que estimar en el modelo y, en consecuencia, el número de grados de libertad disminuye.

Glosario

codificación
Asignación de valores numéricos (a menudo enteros) a categorías, atributos o cualidades observadas. Habitualmente las variables dicotómicas se codifican poniendo 1 en una categoría y 0 en la otra.

efecto aditivo
Parámetro asociado a una variable ficticia incluida como variable explicativa en un modelo sin combinarla con ninguna otra variable. Indica un cambio de nivel de la variable endógena.

efecto multiplicativo
Parámetro asociado a una variable ficticia que multiplica otra variable para formar parte del conjunto de las variables explicativas de un modelo. Indica un cambio de pendiente de la variable endógena.

interacción
Variable que se sostiene multiplicando dos variables dicotómicas asociadas a diferentes atributos. Con la codificación habitual, indica los individuos observados que cumplen las dos condiciones identificadas por 1 en las dos variables dicotómicas.

modelo de efectos fijos
Modelo de regresión lineal múltiple utilizado para datos de panel. Incluye variables ficticias que indican el grupo al que pertenece cada observación y también otras que indican el periodo en el que se ha observado. Los parámetros de estas variables ficticias se interpretan como los efectos debidos a variables no observables.

panel de datos
Conjunto de observaciones sobre N individuos o unidades a lo largo de T periodos de tiempo.

variable dicotómica
Variable que sólo toma dos valores posibles. Para la codificación se usan a menudo los valores con 0 y 1. En general se llama también *ficticia*, especialmente cuando se utiliza como herramienta para especificar un modelo con variables explicativas cualitativas.

variable exógena cualitativa
Variable explicativa de un MRLM que indica un atributo no numérico del individuo o un periodo de tiempo.

variable ficticia
Variable que sólo toma dos valores posibles. Para la codificación se usan a menudo los valores con 0 y 1. Se utiliza para especificar modelos en los cuales hay que introducir una variable cualitativa con múltiples categorías o bien para identificar casos o grupos de casos. También se la llama *variable dicotómica* o *dummy*.

Bibliografía

Gujarati, D. (1990). *Econometría* (2.ª ed., cap. 12). Bogotá: McGraw-Hill.

Johnston, J. (1987). *Métodos de econometría* (trad. J. Sánchez Fernández). Barcelona: Vicens-Vives.

Novales, A. (1993). *Econometría* (2.ª ed., cap. 4.11). Madrid: McGraw-Hill.