Unidad 4. Variables binarias o dummies

Erika R. Badillo
erika.badilloen@unaula.edu.co

Facultad de Economía
Universidad Autónoma Latinoamericana

En este tema

- Conceptualización general
- Modelación de factores y categorías

Lecturas

- Wooldridge, Jeffrey (2013). *Introducción a la econometría*. 5a edición, Cengage Learning. Cap. 6, 7
- Gujarati, D. y Porter, D. (2010). *Econometría*. 5a edición, Mc Graw Hill. Cap. 6 y 9

Variables binarias o dummies: conceptualización general

La inclusión de variables binarias (también llamadas *dummy* o falsas) en los modelos de regresión obedece a la necesidad de incorporar factores de naturaleza cualitativa que se traducen en cambios paramétricos. Algunos ejemplos:

- La ecuación de Mincer o de ingresos laborales puede ser diferente para hombres y mujeres (diferencias en el salario de reserva por discriminación) y el log del ingreso mínimo (o intercepto) puede ser diferente para cada **género**
- La demanda por carne puede variar según los **grupos religiosos**, las elasticidades precio e ingreso de cada grupo pueden ser diferentes
- Un **cambio estructural** en el tiempo puede ser el resultado de un factor cualitativo que induce el cambio paramétrico
  - Si se piensa en la función de consumo para Colombia de 1950 a 2000, es intuitivo afirmar que debido a migración campo-ciudad, transición demográfica o modernización del aparato financiero, la función de consumo de 1950 a 1970 no debe ser la misma que la correspondiente de 1971 a 2000
  - El consumo autónomo (intercepto) y la propensión marginal a consumir (la pendiente) de los dos períodos puede haber cambiado. Igual sucedería con los parámetros de la función de importaciones antes y después de la apertura económica en 1990
- **Raza**, **sector o industria** a la que pertenece una empresa, **región**, **etc.**

Variables binarias o dummies: conceptualización general

- La forma de incluir estos factores cualitativos es usando una variable que sólo tome el valor 0 y 1, y se denominan falsas, dicótomas, binarias o *dummies* $\Longrightarrow$ variables indicadores
- La escogencia de 0 y 1 no es arbitraria, proviene de la esencia del conteo. Cuando se esta contando algo, se suma 1 si ese algo esta y se suma 0 si ese algo no esta

$$\text{se puede asociar: } \begin{cases} 0 \text{ Ausencia} \\ 1 \text{ Presencia} \end{cases}$$

- Otro par de números (3 y 7 por ejemplo) no servirían para lo mismo, lo que puede ser arbitrario es la asignación del 0 y el 1
- Al definir una variable binaria hay que decidir a qué evento se le asigna el valor uno y a cuál el valor cero
- Cuando se usan variables binarias en los modelos se producen cambios en
  - el intercepto
  - la pendiente
  - intercepto y pendiente
  - funciones quebradas

i. Un factor dos categorías

Supóngase que se quiere incorporar al modelo de Mincer (ecuación de salarios) el factor cualitativo género. Existen tres posibilidades según el efecto que se quiere modelar

- cambio en el intercepto (en el log del salario mínimo)
- cambio en la pendiente (en la tasa de retorno de la educación)
- cambio de ambos, intercepto y pendiente

Lo que se intenta incorporar es una hipótesis de diferenciación por género en la ecuación de ingresos. Se define una variable binaria de la forma

$$bsexo_i = \begin{cases} 0 \text{ Mujer} \\ 1 \text{ Hombre} \end{cases}$$

i. Un factor dos categorías

A. Cambio en el intercepto

Sea $Y_i =$ log de los salarios
$X_{2i} =$ Años de educación aprobados

En el modelo $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$

β_1 : log tasa de salario mínima
β_2 : tasa de retorno de la educación
u_i : perturbación aleatoria con supuestos estándar

Al incorporar la variable binaria de género se tendría

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 bsexo_i + u_i$$

Es como si el modelo se convirtiese en dos submodelos

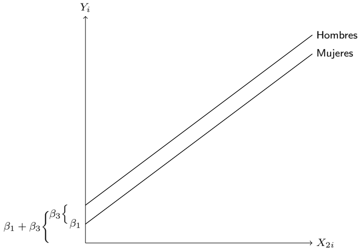
Hombres $(bsexo_i = 1) \Longrightarrow Y_i = (\beta_1 + \beta_3) + \beta_2 X_{2i} + u_i$
Mujeres $(bsexo_i = 0) \Longrightarrow Y_i = \beta_1 + \beta_2 X_{2i} + u_i$

En esta situación
β_1 : log de la tasa salaria mínima de las mujeres
β_3 : cambio en log de la tasa salarial mínima de los hombres respecto a las mujeres
$\beta_1 + \beta_3$: log de la tasa salarial mínima de los hombres (es una combinación lineal paramétrica)

i. Un factor dos categorías

A. Cambio en el intercepto

Gráficamente tenemos



Lo que se esta modelando es un cambio en el intercepto manteniendo constante la pendiente

Lo que se hizo fue conservar el intercepto (β_1) y agregar una variable falsa ($bsexo_i$). Alternativamente se puede eliminar el intercepto e incluir dos variables binarias

$$bhombre_i = \begin{cases} 0 \text{ Mujer} \\ 1 \text{ Hombre} \end{cases} \qquad bmujer_i = \begin{cases} 0 \text{ Hombre} \\ 1 \text{ Mujer} \end{cases}$$

Observece que $bhombre_i + bmujer_i = 1$

i. Un factor dos categorías

A. Cambio en el intercepto

El modelo queda de la forma

$$Y_i = \gamma_2 X_{2i} + \gamma_3 bhombre_i + \gamma_4 bmujer_i + u_i$$

Nuevamente se tienen dos modelos

Mujeres $(bhombre_i = 0, bmujer_i = 1) \Longrightarrow Y_i = \gamma_4 + \gamma_2 X_{2i} + u_i$

Hombres $(bhombre_i = 1, bmujer_i = 0) \Longrightarrow Y_i = \gamma_3 + \gamma_2 X_{2i} + u_i$

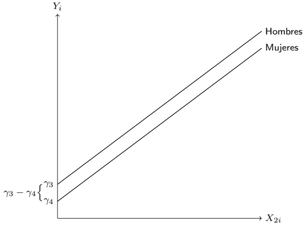
En esta situación

γ_2 : tasa de retorno de la educación, se supone igual para hombres y mujeres

γ_3 : log de la tasa salarial mínima para hombres

γ_4 : log de la tasa salarial mínima para mujeres

$\gamma_3 - \gamma_4$: diferencial del log de la tasa mínima de salario de hombres frente a mujeres



i. Un factor dos categorías

A. Cambio en el intercepto

Hay un cambio en el significado de los parámetros

- En la primera opción (intercepto + una binaria):
  - el intercepto corresponde al grupo con cero (β_1)
  - El coeficiente de la variable binaria es un diferencial (β_3) del grupo con 1 respecto al del 0
  - El intercepto del grupo con 1 es la suma de los dos anteriores ($\beta_1 + \beta_3$)
- En el segundo caso (no intercepto y dos binarias)
  - el coeficiente de cada variable es el respectivo intercepto (γ_3 para hombres y γ_4 para mujeres)
  - si se quiere indagar sobre el diferencial se construye $\gamma_3 - \gamma_4$

En el primer caso verificar si el log de la tasa salarial mínima de los hombres es diferente de cero implica el siguiente contraste

$$H_o : \beta_3 = 0$$
$$H_A : \beta_3 \neq 0$$

En el segundo caso se indaga si los hombres tienen un log de la tasa mínima de salario mayor que el de las mujeres, lo que implica contrastar

$$H_o : \gamma_3 - \gamma_4 = 0$$
$$H_A : \gamma_3 - \gamma_4 \neq 0$$

i. Un factor dos categorías

A. Cambio en el intercepto

Qué sucede si se utilizan las dos opciones anteriores al mismo tiempo: se conserva el intercepto y se incluyen las dos variables binarias

$$Y_i = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 bhombre_i + \gamma_4 bmujer_i + u_i$$

La matriz **X** del modelo tendría la siguiente estructura (suponemos primero mujeres (M) y después hombres $(N - M)$)

$$\mathbf{X}_{N\text{x}4} = \begin{bmatrix} 1 & X_{21} & 0 & 1 \\ 1 & \vdots & 0 & 1 \\ 1 & X_{2M} & 0 & 1 \\ \dots & \dots & \dots & \dots \\ 1 & X_{2M+1} & 1 & 0 \\ 1 & \vdots & 1 & 0 \\ 1 & X_{2N} & 1 & 0 \end{bmatrix}$$

Se observa que Columna(1)=Columan(3)+Columna(4), lo cual implica que rango de la matriz **X** no es de 4 sino de 3, con lo cual hay un problema de multicolinealidad perfecta:

$$(X'X)_{4\text{x}4} \text{ es singular}$$
$$(X'X)^{-1}_{4\text{x}4} \text{ no existe}$$

Este caso se conoce como la trampa de las variables *dummies*

i. Un factor dos categorías

B. Cambio en la pendiente

La manera de incorporar cambios en la pendiente es agregar el producto de la variable binaria por la correspondiente variable explicativa. Por ejemplo, modelando diferentes tasas de retornos a la educación por género, el modelo queda de la forma:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{2i} bsexo_i + u_i$$

Nuevamente es un modelo que contiene dos "submodelos"
Mujeres $(bsexo_i = 0) \Longrightarrow Y_i = \beta_1 + \beta_2 X_{2i} + u_i$
Hombres $(bsexo_i = 1) \Longrightarrow Y_i = \beta_1 + (\beta_2 + \beta_3) X_{2i} + u_i$

En esta situación
β_1 : log de la tasa salarial mínima, se supone igual para hombres y mujeres
β_2 : tasa de retorno de la educación de las mujeres
β_3 : cambio en la tasa de retorno de la educación de hombres respecto a mujeres
$\beta_2 + \beta_3$: tasa de retorno de la educación de los hombres

i. Un factor dos categorías

C. Cambio en el intercepto y la pendiente

La intuición indica que se debe reunir los dos casos anteriores: agregar una variable binaria (o eliminar el intercepto y agregar dos binarias) y la binaria multiplicada por la variable independiente. El modelo queda de la forma:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 bsexo_i + \beta_4 X_{2i} bsexo_i + u_i$$

Mujeres $(bsexo_i = 0) \Longrightarrow Y_i = \beta_1 + \beta_2 X_{2i} + u_i$

Hombres $(bsexo_i = 1) \Longrightarrow Y_i = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) X_{2i} + u_i$

En esta situación
β_1 : log de la tasa salarial mínima de las mujeres
β_2 : tasa de retorno de la educación de las mujeres
β_3 : cambio en log de la tasa salarial mínima de hombres respecto a mujeres
β_4 : cambio en la tasa de retorno de la educación de hombres respecto a mujeres
$\beta_1 + \beta_3$: log de la tasa salarial mínima de los hombres
$\beta_2 + \beta_4$: Tasa de retorno de la educación de los hombres

El modelo conjunto es equivalente a estimar dos regresiones por separado

ii. Un factor varias categorías

Si en lugar de tener dos categorías se tuvieran varias, al modelar cambios en el intercepto se puede proceder de manera similar al caso de dos categorias:

- si hay P categorías, definir $P-1$ variables binarias y conservar el intercepto
- incluir P variables binarias y eliminar el intercepto

En cualquier caso hay que evitar la trampa de las variables *dummies*

Ejemplo: supongamos que se quiere modelar los ingresos laborales para individuos con 3 diferentes niveles educativos. El factor sería la educación y las categorías: primaria, secundaria y superior

La pregunta entonces es: existen diferencias en los ingresos laborales entre individuos con diferentes niveles educativos? En otras palabras los parámetros de la función de salarios cambian de grupo a grupo de individuos con diferentes niveles educativos?

ii. Un factor varias categorías

Se define
Y_i : log del salario
X_{21} : experiencia en años del individuo

Se definen las siguientes variables binarias

$$bpri_i = \begin{cases} 1 \text{ primaria} \\ 0 \text{ otro caso} \end{cases} \quad bsec_i = \begin{cases} 1 \text{ secundaria} \\ 0 \text{ otro caso} \end{cases} \quad bsup_i = \begin{cases} 1 \text{ superior} \\ 0 \text{ otro caso} \end{cases}$$

En el modelo de RLM se conserva el intercepto y al haber 3 categorías se incluyen 2 variables binarias. La categoría a la cual no se le incluye la variable binaria se vuelve el patrón de referencia del modelo. El modelo queda de la forma:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 bsec_i + \beta_4 bsup_i + u_i$$

El modelo incluye 3 submodelos:
Secundaria $(bsec_i = 1, bsup_i = 0) \Longrightarrow Y_i = (\beta_1 + \beta_3) + \beta_2 X_{2i} + u_i$
Superior $(bsec_i = 0, bsup_i = 1) \Longrightarrow Y_i = (\beta_1 + \beta_4) + \beta_2 X_{2i} + u_i$
Primaria $(bsec_i = 0, bsup_i = 0) \Longrightarrow Y_i = \beta_1 + \beta_2 X_{2i} + u_i$

En esta situación
β_1 : log de la tasa salarial mínima de los individuos con primaria
β_2 : tasa de retorno de la experiencia, asumida igual independiente del nivel educativo
β_3 : diferencia en log de la tasa salarial mínima de individuos con secundaria respecto a los de primaria
β_4 : diferencia en log de la tasa salarial mínima de individuos con superior respecto a los de primaria
$\beta_1 + \beta_3$: log de la tasa salarial mínima de los individuos con secundaria
$\beta_1 + \beta_4$: log de la tasa salarial mínima de los individuos con superior

ii. Un factor varias categorías

La segunda opción es eliminar el intercepto, el modelo entonces queda de la forma

$$Y_i = \beta_2 X_{2i} + \beta_3 bpri_i + \beta_4 bsec_i + \beta_5 bsup_i + u_i$$

El modelo de nuevo incluye 3 submodelos:
Primaria $(bpri_i = 1, bsec_i = 0, bsup_i = 0) \Longrightarrow Y_i = \beta_3 + \beta_2 X_{2i} + u_i$
Secundaria $(bpri_i = 0, bsec_i = 1, bsup_i = 0) \Longrightarrow Y_i = \beta_4 + \beta_2 X_{2i} + u_i$
Superior $(bpri_i = 0, bsec_i = 0, bsup_i = 1) \Longrightarrow Y_i = \beta_5 + \beta_2 X_{2i} + u_i$

En esta situación
β_2 : tasa de retorno de la experiencia, asumida igual independiente del nivel educativo
β_3 : log de la tasa salarial mínima de individuos con primaria
β_4 : log de la tasa salarial mínima de individuos con secundaria
β_5 : log de la tasa salarial mínima de individuos con superior

iii. Varios factor y varias categorías

- Es una generalización de los dos casos anteriores
- En la medida que se tendrán bastantes variables binarias es muy importante la trampa de las variables binarias
- Lo más sencillo será conservar el intercepto y en cada factor excluir una variable *dummy*. De esta forma se garantiza que habrá siempre un patrón de referencia en el cual todas las variables falsas son cero
- Otra opción, menos clara, es eliminar el intercepto, para un factor incluir todas las variables binarias y para el resto de factores excluir una variable binaria
- Una tercera opción, menos referenciada en la literatura, es cruzar todos los factores, creando uno solo con todas las posibles combinaciones entre categorías

iii. Varios factor y varias categorías

Supongamos que se quiere modelar los salarios de los trabajadores atendiendo a diferencias en género y posición dentro del hogar, esto es hombre, mujer, jefe de hogar y no jefe

$$bhombre_i = \begin{cases} 1 \text{ Hombre} \\ 0 \text{ Mujer} \end{cases} \quad bjefehog_i = \begin{cases} 1 \text{ Jefe} \\ 0 \text{ No jefe} \end{cases}$$

A. Conservando el intercepto

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 bhombre_i + \beta_4 bjefehog_i + u_i$$

B. Cambio en pendiente

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 (bhombre_i * X_{2i}) + \beta_4 (bjefehog_i * X_{2i}) + u_i$$

C. Cambio en intercepto y pendiente

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 bhombre_i + \beta_4 bjefehog_i + \beta_5 (bhombre_i * X_{2i}) + \beta_6 (bjefehog_i * X_{2i}) + u_i$$

Ejemplo - Stata

Se tiene una base de datos de corte transversal de 526 trabajadores correspondientes a 1976 para los Estados unidos. $wage$ son los salarios en dólares por hora y $educ$ los años de educación

i. Un factor dos categorías

A. Cambio en el intercepto (intercepto + una binaria)

$$lwage = \beta_1 + \beta_2 educ + \beta_3 male + u$$

$$male = \begin{cases} 1 \text{ Male} \\ 0 \text{ Female} \end{cases}$$

```
cd "C:\..."
use "WAGE1.dta", clear

reg lwage educ male
```

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =  112.19
       Model |  44.5315181     2  22.2657591           Prob > F      =  0.0000
    Residual |  103.798233   523  .198466985           R-squared     =  0.3002
-------------+------------------------------           Adj R-squared =  0.2975
       Total |  148.329751   525   .28253286           Root MSE      =   .4455

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0772033   .0070472    10.96   0.000     .0633591    .0910475
        male |   .3608654   .0390245     9.25   0.000     .2842015    .4375294
       _cons |   .4654039   .0912268     5.10   0.000     .2861879    .6446199
```

i. Un factor dos categorías

A. Cambio en el intercepto (no intercepto y dos binarias)

$$lwage = \beta_2 educ + \beta_3 male + \beta_4 female + u$$

$$male = \begin{cases} 1 \text{ Hombre} \\ 0 \text{ Mujer} \end{cases} \qquad female = \begin{cases} 0 \text{ Hombre} \\ 1 \text{ Mujer} \end{cases}$$

```
recode female (0=1 "Male") (1=0 "Female"), gen(male)

reg lwage educ male female, noconst
```

```
      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(  3,    523)  =   2402.65
       Model |  1430.54175         3   476.84725   Prob > F        =    0.0000
    Residual |  103.798233       523  .198466985   R-squared       =    0.9323
-------------+----------------------------------   Adj R-squared   =    0.9320
       Total |  1534.33998       526  2.91699617   Root MSE        =     .4455

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0772033   .0070472    10.96   0.000     .0633591    .0910475
        male |   .8262694   .0940541     8.79   0.000     .6414991     1.01104
      female |   .4654039   .0912268     5.10   0.000     .2861879    .6446199
```

Ejemplo - Stata

i. Un factor dos categorías

B. Cambio en la pendiente

$$lwage = \beta_1 + \beta_2 educ + \beta_3 educ * male + u$$

```
gen educXmale = educ*male

reg lwage educXmale
```

Source	SS	df	MS
Model	30.6869654	1	30.6869654
Residual	117.642786	524	.224509134
Total	148.329751	525	.28253286

Number of obs =	526
F(1, 524) =	136.68
Prob > F =	0.0000
R-squared =	0.2069
Adj R-squared =	0.2054
Root MSE =	.47382

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educXmale	.0358103	.003063	11.69	0.000	.029793	.0418276
_cons	1.384715	.0290373	47.69	0.000	1.327671	1.441759

i. Un factor dos categorías

C. Cambio en el intercepto y la pendiente

$$lwage = \beta_1 + \beta_2 educ + \beta_3 male + \beta_4 educ * male + u$$

```
reg lwage educ male educXmale
```

Source	SS	df	MS
Model	44.531522	3	14.8438407
Residual	103.798229	522	.198847183
Total	148.329751	525	.28253286

Number of obs	=	526
F(3, 522)	=	74.65
Prob > F	=	0.0000
R-squared	=	0.3002
Adj R-squared	=	0.2962
Root MSE	=	.44592

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0771639	.0113831	6.78	0.000	.0548015	.0995262
male	.3600645	.1854296	1.94	0.053	-.0042154	.7243444
educXmale	.0000641	.0145035	0.00	0.996	-.0284283	.0285565
_cons	.4658902	.1429974	3.26	0.001	.184969	.7468113

Es equivalente a estimar dos regresiones por separado, una cuando $male = 1$ y otra cuando $male = 0$

Ejemplo - Stata

ii. Un factor varias categorías

$$lwage = \beta_1 + \beta_2 exper + \beta_3 secundaria + \beta_4 superior + u$$

$$primaria = \begin{cases} 1 \text{ primaria} \\ 0 \text{ otro caso} \end{cases} \quad secundaria = \begin{cases} 1 \text{ secundaria} \\ 0 \text{ otro caso} \end{cases} \quad superior = \begin{cases} 1 \text{ superior} \\ 0 \text{ otro caso} \end{cases}$$

```
gen primaria = .
replace primaria = 1 if educ>=0 & educ<=5
replace primaria = 0 if primaria==.

gen secundaria = .
replace secundaria = 1 if educ>=6 & educ<=13
replace secundaria = 0 if secundaria==.

gen superior = .
replace superior = 1 if educ>=14 & educ<=18
replace superior = 0 if superior==.
```

Source	SS	df	MS		
				Number of obs =	526
				F(3, 522) =	41.51
Model	28.5691716	3	9.52305719	Prob > F =	0.0000
Residual	119.76058	522	.229426398	R-squared =	0.1926
				Adj R-squared =	0.1880
Total	148.329751	525	.28253286	Root MSE =	.47898

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exper	.0080989	.0015939	5.08	0.000	.0049677	.0112302
secundaria	.4796754	.1744276	2.75	0.006	.1370091	.8223417
superior	.944581	.177853	5.31	0.000	.5951854	1.293977
_cons	.8601617	.1803265	4.77	0.000	.505907	1.214416