

Tema 6. Análisis con información cualitativa.

6.1. Las variables ficticias.

6.2. Interpretación del coeficiente de variables ficticias.

6.3. Múltiples categorías .

6.4. Interacción de ficticias

Bibliografía:

Ezequiel Uriel (2013): Capítulo 5

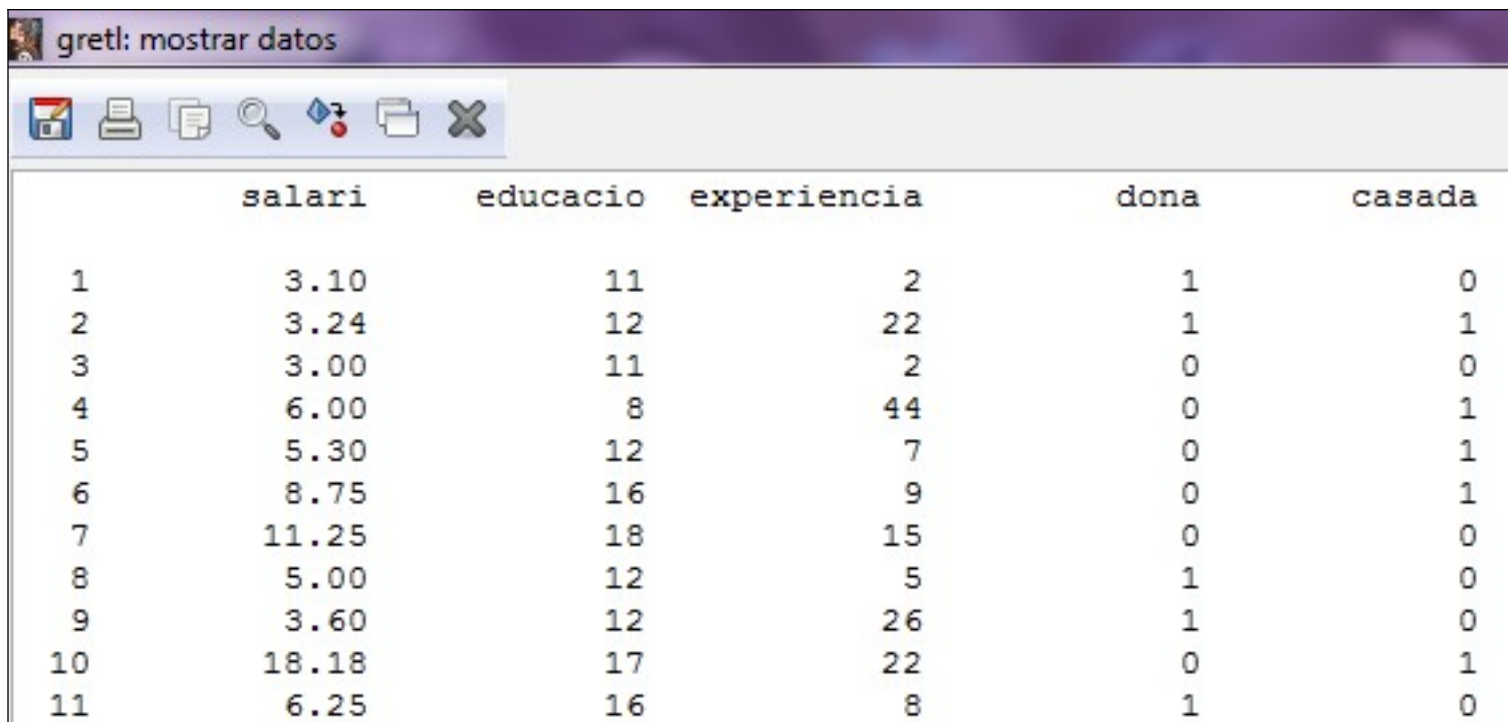
Wooldridge (2015): Capítulo 7

Stock y Watson (2012): Capítulo 5 (epígrafe 5.3)

Variables ficticias

- Hasta ahora las variables que hemos analizado han tenido un significado cuantitativo (salario, educación etc...).
- Pero en el trabajo empírico muchas veces necesita incorporar factores cualitativos en el modelo de regresión (por ejemplo: el género, el tamaño, la situación geográfica, la estación del año).
- ¿Cómo? Mediante la creación e introducción en el modelo de una serie de variables, conocidas como variables ficticias, también llamadas variables artificiales o variables dummy.
- Estas variables ficticias tomarán el valor 1 si el individuo posee una determinada característica, y 0 si no la posee.
- Por ejemplo: podemos definir la variable Hombre como una variable binaria que tome el valor 1 si el individuo es hombre y cero si es mujer. También podemos definir la variable mujer o la variable casado, o la variable “grande”, o la variable “en la costa”....

Datos: variables cuantitativas y cualitativas



	salari	educacio	experiencia	dona	casada
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
6	8.75	16	9	0	1
7	11.25	18	15	0	0
8	5.00	12	5	1	0
9	3.60	12	26	1	0
10	18.18	17	22	0	1
11	6.25	16	8	1	0

- Dona y casada son 2 variables ficticias (o dummies), permitirán incorporar a nuestro MLR información cualitativa (en este caso, el género o el estado civil).
- La variable dona se ha definido de la siguiente manera: toma el valor 1 si el individuo es mujer y toma el valor cero si el individuo no es mujer (en este caso, si es hombre).
- ¿Cómo se define la variable casada?

¿Cómo incorporar información cualitativa al modelo de regresión?

- Para incorporar información cualitativa en el modelo regresión, sencillamente introduciremos las variables ficticias como si fuese una variable más del modelo.
- Cuantas ficticias incorporar y como incorporarlas, dependerá del fenómeno económico concreto que se quiera analizar.

Ejemplo: una característica con dos categorías

- **Ejemplo:** se quiere contrastar si hay discriminación por género en la determinación de los salarios. Una posibilidad es plantear el siguiente modelo:

$$\text{Salari}_i = \beta_1 + \delta_1 \text{dona}_i + \beta_2 \text{educ}_i + u_i$$

- ¿Cuál es la interpretación de δ_1 ? δ_1 es la diferencia en el salario (en promedio) entre hombres y mujeres, dado el mismo nivel de educación (y el mismo u). Es decir, si δ_1 es distinto de cero habrá discriminación salarial. Veámoslo con detalle:

Ejemplo: una característica con dos categorías (discriminación laboral)

- Hemos planteado: $\text{Salari}_i = \beta_1 + \delta_1 \text{dona}_i + \beta_2 \text{educ}_i + u_i$
- En términos de esperanzas: si $E(u \mid \text{educ}, \text{mujer}) = 0$ entonces:

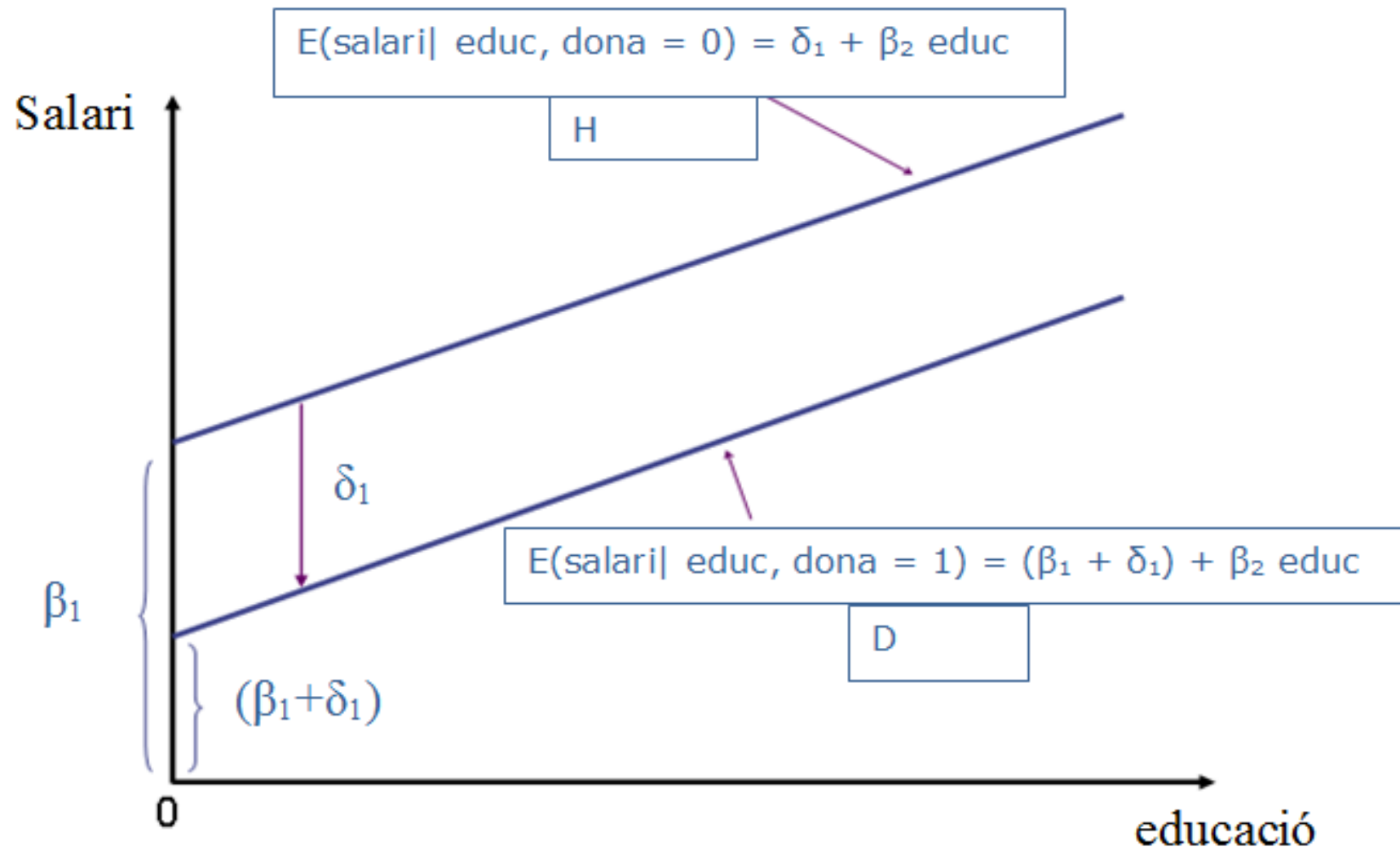
$$E(\text{salario} \mid \text{educ}, \text{mujer}=1) = \beta_1 + \delta_1 + \beta_2 \text{educ}_i$$

$$E(\text{salario} \mid \text{educ}, \text{mujer}=0) = \beta_1 + \quad + \beta_2 \text{educ}_i$$

$$\delta_1 = E(\text{salario} \mid \text{educ}, \text{mujer}=1) - E(\text{salario} \mid \text{educ}, \text{mujer}=0)$$

- Gráficamente la ordenada en el origen será distinta para hombres y mujeres

Habrà discriminación en contra de las mujeres si $\delta_1 < 0$. Las mujeres (para el mismo nivel de los demás factores) ganarán menos en promedio.

Ejemplo con $\delta_1 < 0$ 

Contrastes sobre ficticias ¿Hay realmente discriminación?

- El introducir ficticias no cambia nada en la mecánica de estimación por MCO ni en la forma de efectuar los contrastes. La única diferencia respecto a los regresores cuantitativos es la interpretación del coeficiente.

modelo 2

Modelo 2: MCO, usando las observaciones 1-526

Variable dependiente: salari

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	-1.73448	0.753620	-2.302	0.0218	**
educacio	0.602580	0.0511174	11.79	1.33e-028	***
experiencia	0.0642417	0.0104003	6.177	1.32e-09	***
dona	-2.15552	0.270305	-7.974	9.74e-015	***
Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086		
Suma de cuad. residuos	4945.672	D.T. de la regresión	3.078062		
R-cuadrado	0.309304	R-cuadrado corregido	0.305334		
F(3, 522)	77.91966	Valor p (de F)	1.15e-41		
Log-verosimilitud	-1335.736	Criterio de Akaike	2679.472		
Criterio de Schwarz	2696.533	Crit. de Hannan-Quinn	2686.152		

¿Cuántas ficticias?: Trampa de las ficticias

- En el ejemplo de la discriminación salarial hemos introducido la variable ficticia “mujer”. ¿Porque no hemos introducido las dos ficticias “hombre” y “mujer” a la vez?
- Intuitivamente porque las dos variables proporcionan la misma información y más técnicamente porque si introduyésemos una ficticia para cada categoría (hombre/mujer) se crearía un problema de multicolinealidad perfecta en el modelo de regresión, ya que $\text{hombre} + \text{mujer} = 1$.
- Por lo tanto, si el modelo tiene constante, sólo se pueden introducir en el modelo tantas ficticias como categorías menos una.
- Si se incorporan al modelo tantas ficticias como categorías, se genera multicolinealidad perfecta. A esta situación se le conoce como la trampa de las variables ficticias.

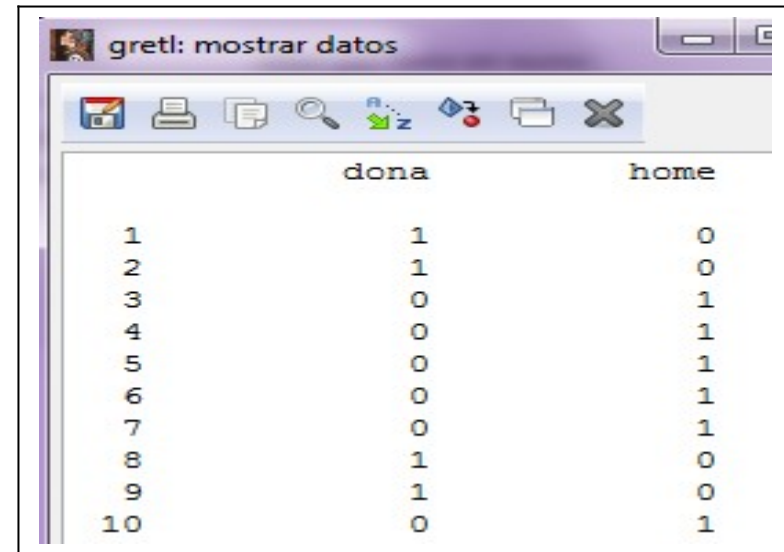
¿Qué ficticia hay que introducir en el modelo?: categoría de referencia

- Ya sabemos que si no queremos caer en la trampa de las ficticias, hay que introducir una *dummy* menos que categorías, pero qué ficticia introduzco en el modelo ¿hombre o mujer?
- La categoría que no tendrá *dummy* es elección del investigador, no afecta a los resultados, aunque sí a la interpretación de los coeficientes de las variables ficticias.
- **La categoría que no tiene *dummy* se llama grupo o categoría de referencia.**
- El coeficiente que acompaña a una *dummy* indica la diferencia en el (valor esperado del) *regresando* entre esa categoría y la categoría de referencia
- En nuestro ejemplo, la variable introducida es “mujer” lo que hace que la categoría de referencia sean los hombres. Por lo tanto, el coeficiente que acompaña a “mujer” indica la diferencia en la constante entre las mujeres y la categoría de referencia (hombres).

Ejemplo: cambiando la categoría de referencia (“mujer”)

- **Ejemplo:** supongamos, para contrastar la existencia de discriminación salarial, se especifica el siguiente modelo:

$$\text{Salari}_i = \beta_1 + \gamma_1 \text{home}_i + \beta_2 \text{educ}_i + u_i$$

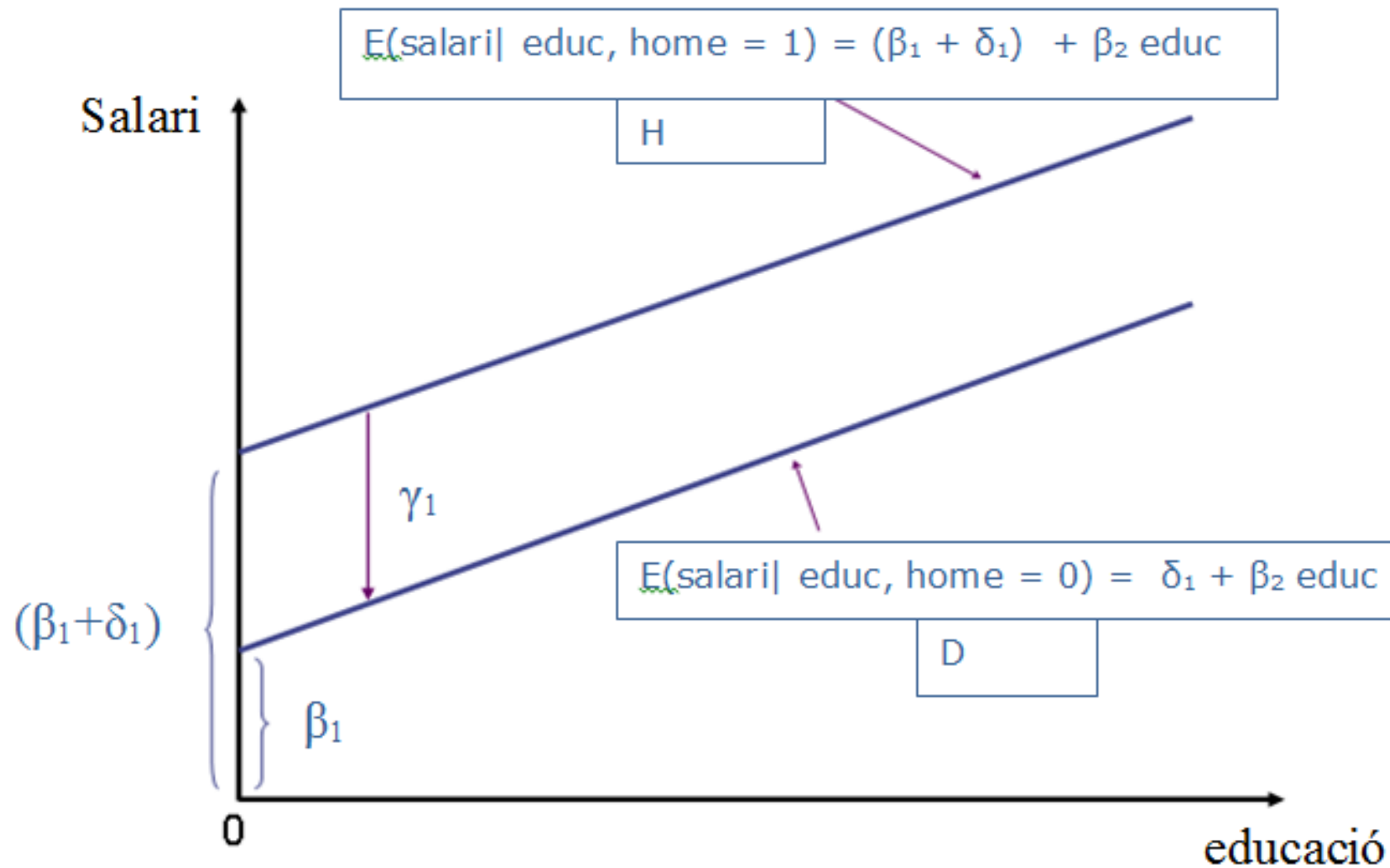


The screenshot shows a window titled 'gretl: mostrar datos' displaying a dataset with two columns, 'dona' and 'home', and 10 rows of data. The 'dona' column represents women (1) and men (0), while the 'home' column represents men (1) and women (0). This setup effectively makes 'home' the reference category for men.

	dona	home
1	1	0
2	1	0
3	0	1
4	0	1
5	0	1
6	0	1
7	0	1
8	1	0
9	1	0
10	0	1

- Ahora la categoría de referencia es “mujer”. En este caso β_1 es la ordenada en el origen para las mujeres (la categoría de referencia); mientras que, el parámetro que acompaña a la *dummy* (γ_1) es la diferencia en el salario entre los hombres y la categoría de referencia (mujeres).
- Ahora, para que haya discriminación (en contra de las mujeres) γ_1 tendrá que ser positivo.

Ejemplo: existe discriminación ($\gamma_1 > 0$) siendo “mujer” la categoría de referencia



Cambio en la categoría de referencia

Modelo 1: MCO, usando las observaciones 1-526

Variable dependiente: salari

	Coeficiente	Desv. Típica	Estadístico t	Valor p	

const	-3.89000	0.727144	-5.350	1.32e-07	***
educacio	0.602580	0.0511174	11.79	1.33e-028	***
experiencia	0.0642417	0.0104003	6.177	1.32e-09	***
home	2.15552	0.270305	7.974	9.74e-015	***
Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086		
Suma de cuad. residuos	4945.672	D.T. de la regresión	3.078062		
R-cuadrado	0.309304	R-cuadrado corregido	0.305334		
F(3, 522)	77.91966	Valor p (de F)	1.15e-41		
Log-verosimilitud	-1335.736	Criterio de Akaike	2679.472		
Criterio de Schwarz	2696.533	Crit. de Hannan-Quinn	2686.152		

¿Las pendientes tienen que ser iguales entre categorías?

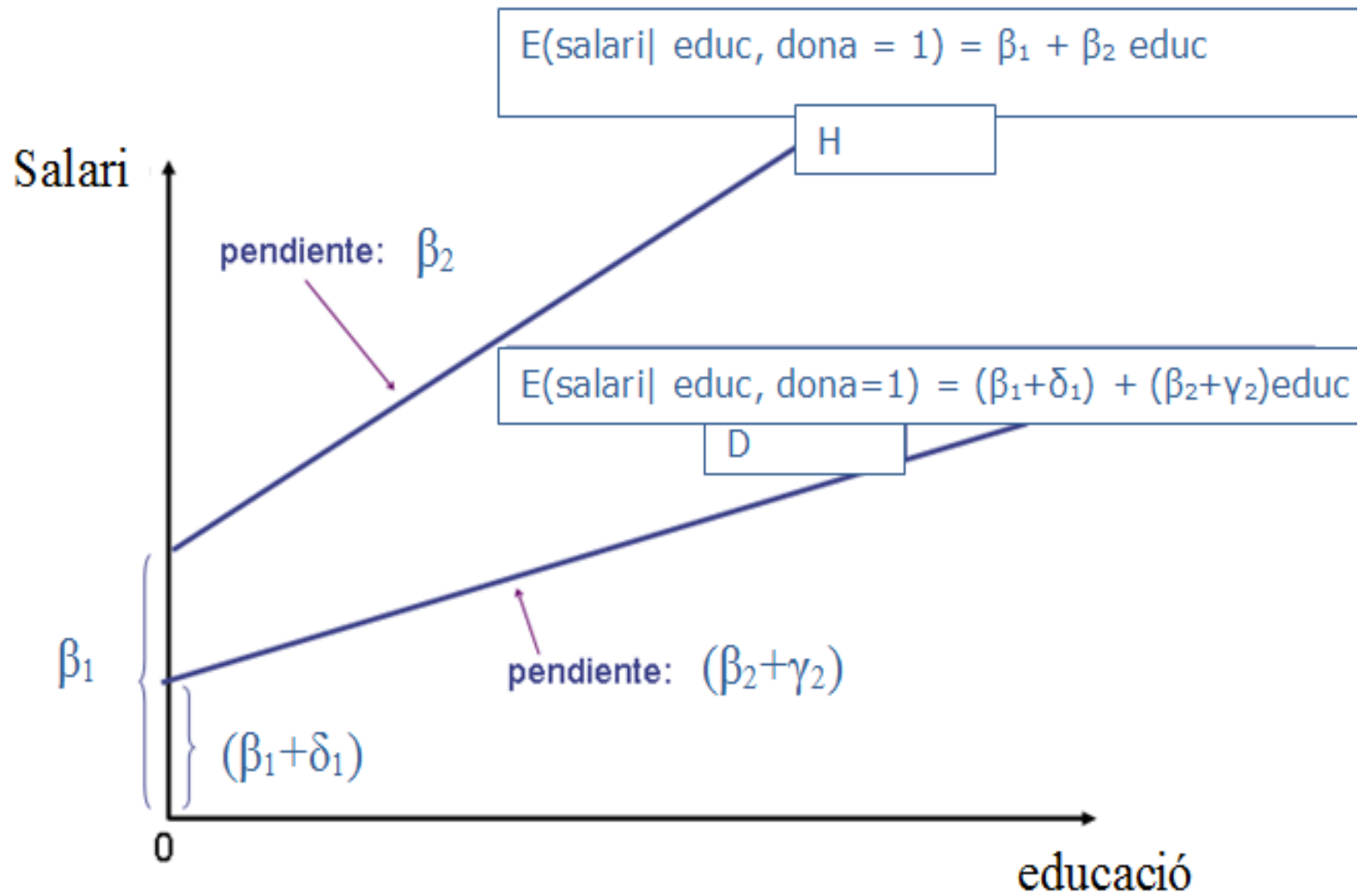
- En el ejemplo que hemos usado hemos planteado un modelo que permitía distintos interceptos (ordenadas) por categorías, pero nada impide que también pueda haber diferencias en la pendiente.
- Para introducir diferencias en el intercepto hemos introducidos las *dummys* en forma **aditiva** (ellas solas, sólo acompañadas de su parámetro).
- Para introducir diferentes pendientes, las variables ficticias han de interactuar con los otros regresores; es decir, se han de introducir en el modelo multiplicando a algún regresor (**dummy multiplicativa**). Por ejemplo:

$$\text{salario} = \beta_1 + \beta_2 \text{educ} + \delta_2(\text{educ} \times \text{mujer}) + u$$

- Si se quiere especificar un modelo que permita diferencias entre grupos tanto en la ordenada en el origen como en la pendiente, se deberá introducir la ficticia tanto en forma aditiva como multiplicativa. Por ejemplo:

$$\text{salario} = \beta_1 + \delta_1 \text{mujer} + \beta_2 \text{educ} + \delta_2(\text{educ} \times \text{mujer}) + u$$

Ejemplo con $\delta_0 < 0$ y $\delta_1 < 0$ (“hombre” grupo de referencia)



Ejemplo de ficticias multiplicativas

- ¿Hay diferencias en la rentabilidad o efecto de la educación entre mujeres y hombres?

Modelo 3: MCO, usando las observaciones 1-526

Variable dependiente: salari

Desviaciones típicas robustas ante heterocedasticidad, variante

	Coeficiente	Desv. Típica	Estadístico t	Val
const	0.200496	0.871723	0.2300	0.81
dona	-1.19852	1.46081	-0.8205	0.41
educacio	0.539476	0.0733959	7.350	7.69
edu_dona	-0.0859990	0.123818	-0.6946	0.48
Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.6	
Suma de cuad. residuos	5300.170	D.T. de la regresión	3.1	
R-cuadrado	0.259796	R-cuadrado corregido	0.2	
F(3, 522)	49.07709	Valor p (de F)	5.9	
Log-verosimilitud	-1353.942	Criterio de Akaike	271	
Criterio de Schwarz	2732.946	Crit. de Hannan-Quinn	272	

gretl: mostrar datos

	educacio	dona	edu_dona
1	11	1	11
2	12	1	12
3	11	0	0
4	8	0	0
5	12	0	0
6	16	0	0
7	18	0	0
8	12	1	12
9	12	1	12
10	17	0	0
11	16	1	16
12	13	1	13

Una ficticia con múltiples categorías

- **Ejemplo** (discriminación salarial por raza): podemos dividir a la población en tres (o más) grupos, raza blanca, negra e hispana. Para incorporar esta información cualitativa en un modelo de regresión, hay que definir las correspondientes variables ficticias.
- Recordar que sólo hay que introducir en el modelo una ficticia menos que categorías.
- La categoría que no tenga su *dummy* en el modelo será el grupo de referencia.
- Las ficticias se pueden introducir de forma aditiva y/o multiplicativa.
- Veamos un ejemplo...

Ejemplo: una ficticia con múltiples categorías (solo aditivas)

Modelo 4: MCO, usando las observaciones 1-526

Variable dependiente: salari

Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	-0.175746	0.733805	-0.2395	0.8108
negre	-0.622091	0.453779	-1.371	0.1710
hisp	-1.47274	0.293803	-5.013	7.36e-07 ***
educacio	0.528523	0.0603190	8.762	2.69e-017 ***
Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086	
Suma de cuad. residuos	5739.640	D.T. de la regresión	3.315943	
R-cuadrado	0.198421	R-cuadrado corregido	0.193814	
F(3, 522)	34.08791	Valor p (de F)	3.92e-20	
Log-verosimilitud	-1374.892	Criterio de Akaike	2757.785	
Criterio de Schwarz	2774.846	Crit. de Hannan-Quinn	2764.465	

Ejemplo: una ficticia con múltiples categorías (solo multiplicativas)

Modelo 6: MCO, usando las observaciones 1-526

Variable dependiente: salari

Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	Desv. Típica	Estadístico t	Valor p

const	-0.746556	0.713165	-1.047	0.2957
educacio	0.458631	0.0629836	7.282	1.22e-012 ***
edu_x_negre	0.0545374	0.0386720	1.410	0.1591
edu_x_blanc	0.115018	0.0255442	4.503	8.28e-06 ***
Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086	
Suma de cuad. residuos	5735.617	D.T. de la regresión	3.314781	
R-cuadrado	0.198983	R-cuadrado corregido	0.194379	
F(3, 522)	33.44478	Valor p (de F)	8.73e-20	
Log-verosimilitud	-1374.708	Criterio de Akaike	2757.416	
Criterio de Schwarz	2774.477	Crit. de Hannan-Quinn	2764.096	

Ejemplo: una ficticia con múltiples categorías (aditivas y multiplicativas)

Modelo 7: MCO, usando las observaciones 1-526

Variable dependiente: salari

	Coeficiente	Desv. Típica	Estadístico t	Valor p

const	1.09065	1.69014	0.6453	0.5190
educacio	0.582259	0.0722663	8.057	5.38e-015 ***
blanc	-1.95190	1.93497	-1.009	0.3136
hisp	-2.54481	2.05753	-1.237	0.2167
edu_x_negre	-0.212829	0.155081	-1.372	0.1705
edu_x_hisp	-0.0693380	0.117062	-0.5923	0.5539
Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086	
Suma de cuad. residuos	5718.460	D.T. de la regresión	3.316178	
R-cuadrado	0.201379	R-cuadrado corregido	0.193700	
F(5, 520)	26.22441	Valor p (de F)	1.19e-23	
Log-verosimilitud	-1373.920	Criterio de Akaike	2759.840	
Criterio de Schwarz	2785.432	Crit. de Hannan-Quinn	2769.860	

Múltiples ficticias

- Nada impide que nuestro modelo incorpore varios tipos de información cualitativa o características
- El mecanismo es el mismo que con una característica: definir las correspondientes ficticias e introducir para cada característica tantas *dummys* como categoría menos una. Para cada atributo tendremos una categoría de referencia
- **Ejemplo:** discriminación salarial por sexo (hombre/mujer) y estado civil (soltero/casado).
- Nada cambia en cuanto a la mecánica solo que ¿cuidado con la multicolinealidad?
- Cuando hay múltiples ficticias surge la posibilidad de que las dos características interactúen (**efecto interacción**)

Ejemplo: múltiples ficticias (sexo y estado civil)

Modelo 8: MCO, usando las observaciones 1-526

Variable dependiente: salari

Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	Desv. Típica	Estadístico t	Valor p	

const	-2.12781	0.715292	-2.975	0.0031	***
educacio	0.494954	0.0595780	8.308	8.46e-016	***
home	2.08699	0.257623	8.101	3.88e-015	***
casada	1.18153	0.257484	4.589	5.59e-06	***
Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086		
Suma de cuad. residuos	5137.567	D.T. de la regresión	3.137209		
R-cuadrado	0.282504	R-cuadrado corregido	0.278381		
F(3, 522)	49.97140	Valor p (de F)	2.10e-28		
Log-verosimilitud	-1345.748	Criterio de Akaike	2699.495		
Criterio de Schwarz	2716.556	Crit. de Hannan-Quinn	2706.175		

Ejemplo: múltiples ficticias y efecto interacción (¿mujer casada?)

Modelo 9: MCO, usando las observaciones 1-526

Variable dependiente: salari

Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	Desv. Típica	Estadístico t	Valor p
-----	-----	-----	-----	-----
const	-1.02442	0.777706	-1.317	0.1883
educacio	0.493559	0.0583005	8.466	2.60e-016 ***
dona	-0.368964	0.371298	-0.9937	0.3208
casada	2.64107	0.401705	6.575	1.19e-010 ***
dona_casada	-2.82883	0.497760	-5.683	2.20e-08 ***
Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086	
Suma de cuad. residuos	4894.020	D.T. de la regresión	3.064884	
R-cuadrado	0.316517	R-cuadrado corregido	0.311270	
F(4, 521)	39.35195	Valor p (de F)	8.34e-29	
Log-verosimilitud	-1332.975	Criterio de Akaike	2675.950	
Criterio de Schwarz	2697.276	Crit. de Hannan-Quinn	2684.300	

¿Hay diferencias entre los distintos grupos?

- A veces el objetivo del análisis consiste en contrastar si las dos poblaciones o grupos siguen la misma función de regresión
- Esto se traduce en contrastar la significatividad conjunta de todas las *dummys* tanto aditivas como multiplicativas.
- Evidentemente esto se puede hacer estimando el modelo con *dummys* (modelo general) y el modelo sin ninguna *dummy* (modelo restringido) y construir el correspondiente estadístico F.