

Tema 6

Análisis con información cualitativa

(actualizadas el 31-08-2022)

Tema 6. Análisis con información cualitativa

6.1 Las variables ficticias

6.2 Interpretación del coeficiente de variables ficticias

6.3 Interacción entre una variable ficticia y otra continua

6.4 Múltiples categorías

6.5 Múltiples ficticias

Bibliografía

- Ezequiel Uriel (2013): Capítulo 5
- Wooldridge (2015): Capítulo 7
- Stock y Watson (2012): Capítulo 5 (epígrafe 5.3)

6.1 Las variables ficticias

Son variables que creamos para poder introducir en nuestros modelos información cualitativa

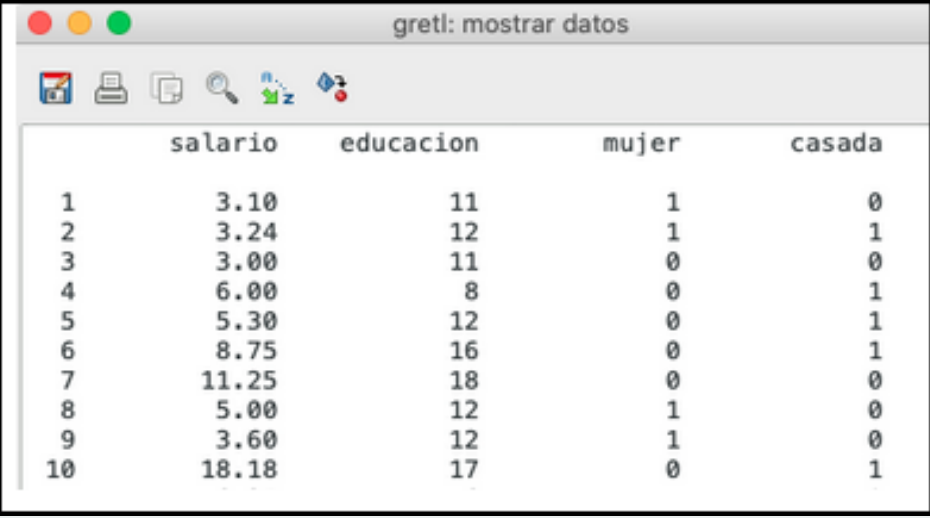
¿Qué son las variables ficticias?

- Hasta ahora las variables que hemos analizado han tenido un significado cuantitativo (salario, educación...).
- Pero en el trabajo empírico muchas veces necesita **incorporar factores cualitativos** en el modelo de regresión (por ejemplo: el género, el sector laboral, la ubicación 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 la observación posee una determinada característica**, y 0 si no la posee.

Algunos ejemplos

- Definimos la variable *Hombre* como una variable binaria que toma el valor 1 si el individuo es hombre y cero si es mujer.
- Definimos la variable *Urbano* como una variable binaria que toma el valor 1 si el individuo reside en una población de 150000 habitantes o más y cero en otro

Ejemplo: variables cuantitativas y cualitativas



The screenshot shows a window titled 'gretl: mostrar datos' with a toolbar containing icons for file operations, search, and visualization. Below the toolbar is a table with 10 rows of data. The columns are labeled 'salario', 'educacion', 'mujer', and 'casada'. The data is as follows:

	salario	educacion	mujer	casada
1	3.10	11	1	0
2	3.24	12	1	1
3	3.00	11	0	0
4	6.00	8	0	1
5	5.30	12	0	1
6	8.75	16	0	1
7	11.25	18	0	0
8	5.00	12	1	0
9	3.60	12	1	0
10	18.18	17	0	1

- Las variables *salario* y *educacion* son cuantitativas.
- *mujer* y *casada* son variables ficticias (o dummies), que permitirán incorporar a nuestro MLR información cualitativa (en este caso, el género o el estado civil del individuo).
- La variable *mujer* 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 (o atributos) en el modelo de regresión, sencillamente introduciremos las variables ficticias como si fuesen una variable más del modelo.
- Dado un atributo con q grupos o categorías, podemos definir q variables dummies.
- **PERO**, como nuestro MRL siempre incorpora término independiente (β_1), tendremos que incorporar al modelo solamente $(q - 1)$ variables dummies; si no incurriríamos en la *trampa de las ficticias*
- Lo veremos más adelante, pero ... las variables ficticias se pueden introducir en el modelo de forma **aditiva** o de forma **multiplicativa** (interactuando con otra variable, generalmente una variable cuantitativa)
- Cuántas ficticias incorporar y cómo incorporarlas, dependerá del fenómeno económico concreto que se quiera analizar.

Visión general con un ejemplo (una característica con dos categorías)

- Supongamos que se quiere contrastar si hay **discriminación por género** en la determinación de los salarios. Supongamos, además, que el género es una característica con **sólo 2 grupos(!!!)**
- Si hay 2 grupos, **podemos definir 2 dummies** (H y M). En la práctica **sólo necesitaremos una** de las 2 dummies
- ¿Qué dummy introducimos? La que queramos, PERO, el grupo que no tenga dummy será la **categoría de referencia**
- Podemos introducir la dummy de forma **aditiva o multiplicativa**
- Modelo con **dummy aditiva**:

$$\text{salario}_i = \beta_1 + \beta_2 \text{educacion}_i + \delta_1 \text{mujer}_i + u_i$$

- Modelo con **dummy multiplicativa**, interactuando con la variable *educacion*

$$\text{salario}_i = \beta_1 + \beta_2 \text{educacion}_i + \gamma_1 (\text{educacion}_i \times \text{mujer}_i) + u_i$$

- ¿Cuál es la **interpretación de los coeficientes** que acompañan a las dummies?
Lo veremos **poco a poco con ejemplos**

6.2 Interpretación del coeficiente de variables ficticias

La interpretación es (un poco) diferente al de las variables cuantitativas

Ejemplo 1: una característica con dos categorías y dummy aditiva

- Hemos planteado: $\text{salario}_i = \beta_1 + \beta_2 \text{educacion}_i + \delta_1 \text{mujer}_i + u_i$
- Como suponemos que $E(u_i) = 0$ entonces:

$$E(\text{salario}_i | \text{mujer}_i = 1) = \beta_1 + \beta_2 \text{educacion}_i + \delta_1$$

$$E(\text{salario}_i | \text{mujer}_i = 0) = \beta_1 + \beta_2 \text{educacion}_i$$

- Por tanto

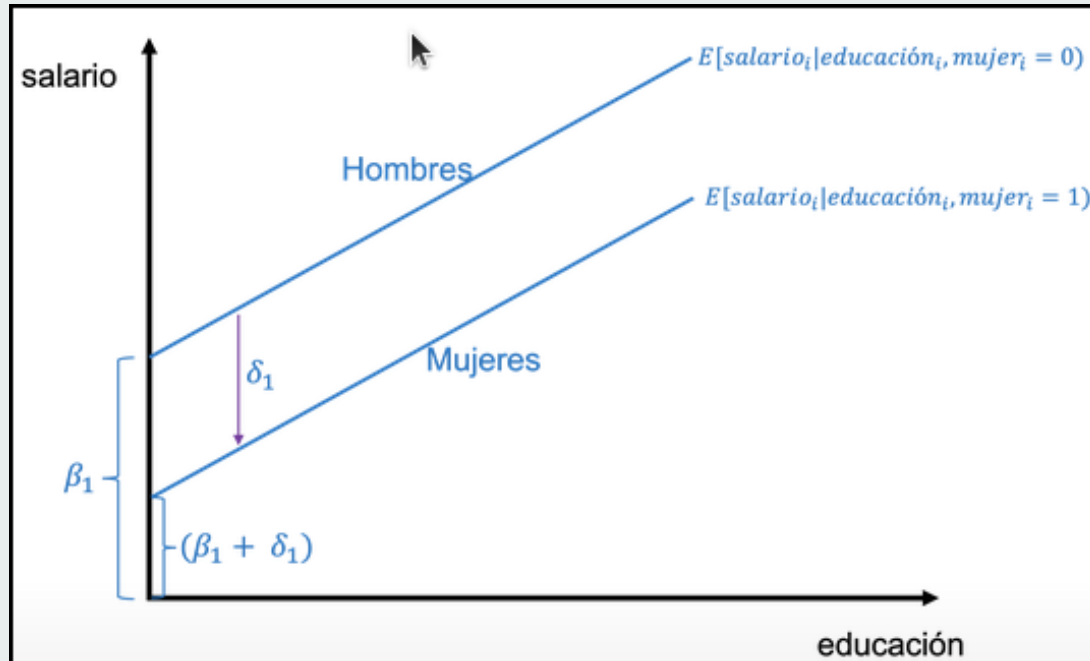
$$\delta_1 = E(\text{salario}_i | \text{mujer}_i = 1) - E(\text{salario}_i | \text{mujer}_i = 0)$$

- Es decir, δ_1 es **la diferencia** en promedio, o en términos esperados, entre el salario de una mujer y un hombre, asumiendo que tienen la misma educación.
- Es decir, en promedio (y para el mismo nivel educativo):
 - Si $\delta_1 < 0$ habría brecha de género en contra de la mujer.
 - Si $\delta_1 > 0$ habría brecha de género a favor de la mujer.
 - Si $\delta_1 = 0$ no hay brecha de género.

Ejemplo 1 (con brecha de genero en contra de las mujeres)

- Gráficamente la ordenada en el origen será distinta para hombres y mujeres.
- Habrá brecha de género en contra de las mujeres si $\delta_1 < 0$. Las mujeres (para el mismo nivel de los demás factores) obtendrán un menor salario en promedio.

\



Ejemplo 1: contrastes sobre ficticias ¿Hay realmente brecha de género?

- 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 1: MCO, usando las observaciones 1-526

Variable dependiente: salario

	coeficiente	Desv. típica	Estadístico t	valor p
const	0,622817	0,672533	0,9261	0,3548
educacion	0,506452	0,0503906	10,05	7,56e-22 ***
mujer	-2,27336	0,279044	-8,147	2,76e-15 ***
Media de la vble. dep.	5,896103	D.T. de la vble. dep.	3,693086	
Suma de cuad. residuos	5307,161	D.T. de la regresión	3,185520	
R-cuadrado	0,258819	R-cuadrado corregido	0,255985	
F(2, 523)	91,31542	Valor p (de F)	9,66e-35	
Log-verosimilitud	-1354,289	Criterio de Akaike	2714,578	
Criterio de Schwarz	2727,374	Crit. de Hannan-Quinn	2719,588	

¿Cuántas ficticias hay que introducir? Trampa de las ficticias (Ejemplo 1)

- En el ejemplo de la brecha 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.
- 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 (Ejemplo 1)

- Ya sabemos que si no queremos caer en la trampa de las ficticias, hay que introducir una variable ficticia menos que categorías, pero qué ficticia introduzco en el modelo ¿hombre o mujer?
- La categoría que no tendrá variable ficticia 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 variable ficticia se llama grupo o categoría de referencia.**
- El coeficiente que acompaña a una variable ficticia indica la diferencia en el valor (esperado) del *regresando* entre la categoría de la variable ficticia 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 de salario entre las mujeres y la categoría de referencia (hombres).

Ejemplo 1: cambiando la categoría de referencia a mujer

- Si en el modelo introducimos la variable *hombre*, que toma el valor 1 si el individuo es hombre y toma el valor cero si el individuo no es hombre (en este caso, si es mujer), simplemente cambia el signo del coeficiente de la variable ficticia.

$$\text{salario}_i = \beta_1 + \beta_2 \text{educacion}_i + \gamma_1 \text{hombre}_i + u_i$$

Modelo 2: MCO, usando las observaciones 1-526

Variable dependiente: salario

	coeficiente	Desv. típica	Estadístico t	valor p	
const	-1,65055	0,652317	-2,530	0,0117	**
educacion	0,506452	0,0503906	10,05	7,56e-22	***
hombre	2,27336	0,279044	8,147	2,76e-15	***
Media de la vble. dep.	5,896103	D.T. de la vble. dep.	3,693086		
Suma de cuad. residuos	5307,161	D.T. de la regresión	3,185520		
R-cuadrado	0,258819	R-cuadrado corregido	0,255985		
F(2, 523)	91,31542	Valor p (de F)	9,66e-35		
Log-verosimilitud	-1354,289	Criterio de Akaike	2714,578		
Criterio de Schwarz	2727,374	Crit. de Hannan-Quinn	2719,588		

6.3 Interacción entre una variable ficticia y otra continua

Con el ejemplo 1 hemos trabajado con dummies aditivas, ahora introduciremos las dummies de forma multiplicativa

¿Las pendientes tienen que ser iguales entre categorías?

- En el ejemplo 1 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 **ficticias en forma aditiva** (ellas solas 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 alguna variable cuantitativa (**ficticias multiplicativas**).

Ejemplo 2: sólo dummies multiplicativas

$$\text{salario}_i = \beta_1 + \beta_2 \text{educacion}_i + \delta_2 (\text{educacion}_i \times \text{mujer}_i) + u_i$$

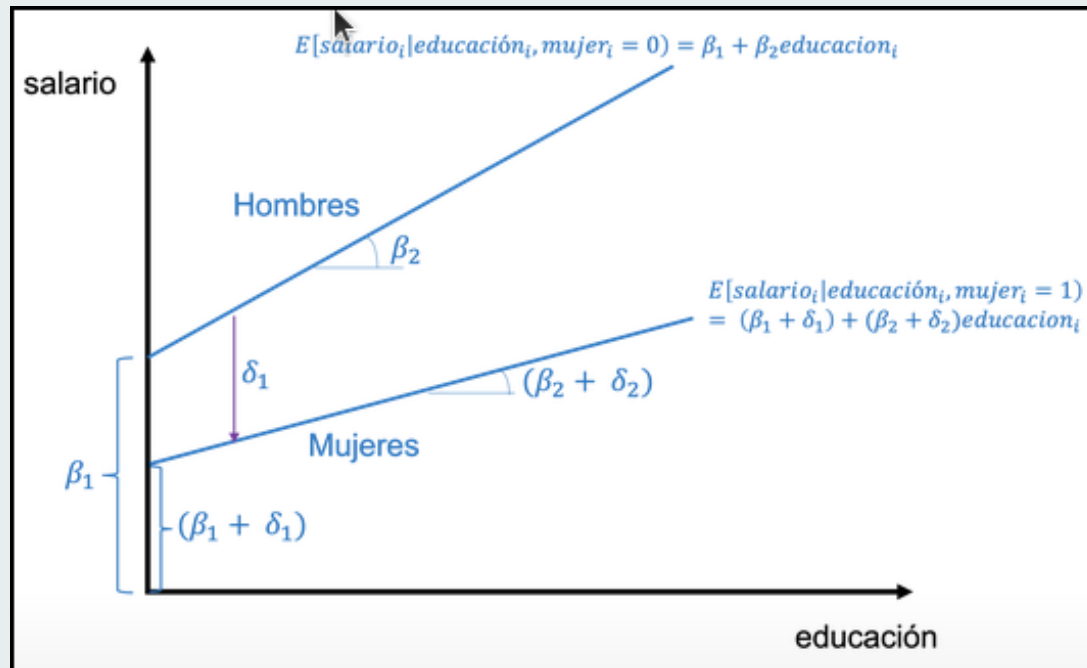
- Ahora, en el modelo 2, el efecto marginal de un año más de educación es: $\beta_2 + \delta_2 \text{mujer}_i$. De forma que:
 - Para una mujer ($\text{mujer}_i = 1$), un año más de educación aumenta su salario en $\beta_2 + \delta_2$.
 - Para un hombre ($\text{mujer}_i = 0$), un año más de educación aumenta su salario en β_2 .
- Dependiendo del signo de δ_2 , el efecto de un año más de educación será mayor entre los hombres o las mujeres.

Ejemplo 3: dummy aditiva y multiplicativa

- 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:

$$\text{salario}_i = \beta_1 + \beta_2 \text{educacion}_i + \delta_1 \text{mujer}_i + \delta_2 (\text{educacion}_i \times \text{mujer}_i) + u_i$$

- Ejemplo gráfico** si asumimos que $\delta_1 < 0$ y $\delta_2 < 0$



Ejemplo 3: estimación con Gretl

Modelo 3: MCO, usando las observaciones 1-526

Variable dependiente: salario

	coeficiente	Desv. típica	Estadístico t	valor p
const	0,200496	0,843562	0,2377	0,8122
educacion	0,539476	0,0642229	8,400	4,24e-16 ***
mujer	-1,19852	1,32504	-0,9045	0,3661
educ_mujer	-0,0859990	0,103639	-0,8298	0,4070
Media de la vble. dep.	5,896103	D.T. de la vble. dep.	3,693086	
Suma de cuad. residuos	5300,170	D.T. de la regresión	3,186469	
R-cuadrado	0,259796	R-cuadrado corregido	0,255542	
F(3, 522)	61,07022	Valor p (de F)	7,44e-34	
Log-verosimilitud	-1353,942	Criterio de Akaike	2715,885	
Criterio de Schwarz	2732,946	Crit. de Hannan-Quinn	2722,565	

\

- ¿Cuál es el efecto marginal de la educación en las mujeres? ¿Y en los hombres?
- ¿Los efectos marginales de la educación en hombres y mujeres difieren?

6.4 Múltiples categorías

**Hemos visto ejemplos con variables cualitativas con sólo dos grupos:
amplíemos a múltiples grupos**

Una característica, pero con múltiples categorías

- **Ejemplo:** vamos a determinar si el tiempo dedicado al ocio depende de los estudios del sujeto. La variable estudios indica el máximo nivel de estudios alcanzados y esta recogida como una variable cualitativa con **tres grupos:** primarios (PRI), secundarios (SEC) y universitarios (UNI).
- Para incorporar esta información cualitativa en un modelo de regresión, hay que definir las correspondientes variables ficticias.
- Recordar que hay que introducir en el modelo **una variable ficticia menos que categorías** tiene la variable cualitativa.
- La categoría que no tenga su ficticia en el modelo será el **grupo de referencia**
- Las ficticias se pueden introducir de forma aditiva y/o multiplicativa.
- Veamos un ejemplo...

Ejemplo: una característica con múltiples categorías (dummies aditivas)

```
Modelo 1: MCO, usando las observaciones 1-1200
Variable dependiente: OCI
```

	coeficiente	Desv. típica	Estadístico t	valor p	
const	-367,649	117,669	-3,124	0,0018	***
RENTA	0,0727363	0,00553279	13,15	5,71e-37	***
SEC	375,885	68,6544	5,475	5,33e-08	***
UNI	921,918	105,315	8,754	6,90e-18	***
Media de la vble. dep.	1721,191	D.T. de la vble. dep.	1249,409		
Suma de cuad. residuos	1,03e+09	D.T. de la regresión	926,9725		
R-cuadrado	0,450919	R-cuadrado corregido	0,449541		
F(3, 1196)	327,3945	Valor p (de F)	3,7e-155		
Log-verosimilitud	-9899,032	Criterio de Akaike	19806,06		
Criterio de Schwarz	19826,42	Crit. de Hannan-Quinn	19813,73		

- ¿Cómo se ha definido la variable SEC? ¿y UNI?
- ¿Cuál es la categoría de referencia?
- Interpreta el coeficiente de SEC (el tiempo de ocio está medido en minutos semanales).
- Interpreta el coeficiente de UNI.

Ejemplo: una característica con múltiples categorías (dummies multiplicativas)

Modelo 2: MCO, usando las observaciones 1-1200				
Variable dependiente: OCI				
	coeficiente	Desv. típica	Estadístico t	valor p

const	-22,5268	140,497	-0,1603	0,8726
RENTA	0,0556471	0,00696825	7,986	3,26e-15 ***
RENTA_SEC	0,0184970	0,00300693	6,151	1,05e-09 ***
RENTA_UNI	0,0335877	0,00371855	9,032	6,56e-19 ***
Media de la vble. dep.	1721,191	D.T. de la vble. dep.	1249,409	
Suma de cuad. residuos	1,02e+09	D.T. de la regresión	925,6776	
R-cuadrado	0,452452	R-cuadrado corregido	0,451078	
F(3, 1196)	329,4273	Valor p (de F)	7,0e-156	
Log-verosimilitud	-9897,354	Criterio de Akaike	19802,71	
Criterio de Schwarz	19823,07	Crit. de Hannan-Quinn	19810,38	

- ¿Cuál es la ecuación del modelo estimado? Representelo gráficamente
- ¿Cuál es el efecto marginal de la renta sobre el ocio? (la renta está expresada en miles de euros)
- ¿Cuál es el efecto marginal **estimado** de la renta sobre el ocio entre los sujetos con estudios primarios?
- ¿Y entre los sujetos con estudios secundarios? ¿Y universitarios?

Ejemplo: múltiples categorías (aditivas y multiplicativas)

Modelo 3: MCO, usando las observaciones 1-1200				
Variable dependiente: OCIO				
	coeficiente	Desv. típica	Estadístico t	valor p
-----	-----	-----	-----	-----
const	-42,8484	199,264	-0,2150	0,8298
UNI	269,614	381,921	0,7059	0,4804
RENTA	0,0566196	0,00971030	5,831	7,09e-09 ***
SEC	-94,4295	319,670	-0,2954	0,7677
RENTA_SEC	0,0219018	0,0137495	1,593	0,1114
RENTA_UNI	0,0256156	0,0134440	1,905	0,0570 *
Media de la vble. dep.	1721,191	D.T. de la vble. dep.	1249,409	
Suma de cuad. residuos	1,02e+09	D.T. de la regresión	926,1398	
R-cuadrado	0,452821	R-cuadrado corregido	0,450530	
F(5, 1194)	197,6204	Valor p (de F)	1,5e-153	
Log-verosimilitud	-9896,949	Criterio de Akaike	19805,90	
Criterio de Schwarz	19836,44	Crit. de Hannan-Quinn	19817,40	

- ¿Cuál es la ecuación del modelo?
- Escribe la ecuación estimada para el valor esperado de ocio de un sujeto con estudios primarios
- Ahora para un sujeto con estudios universitarios.

6.5 Múltiples ficticias

Hemos visto ejemplos con varias dummies, pero todas las dummies estaban relacionadas con una sola variable cualitativa. Ahora vamos a introducir en el modelo varias variables cualitativas (o características).

Varias variables cualitativas en el modelo

- Nada impide que nuestro modelo incorpore varios tipos de información cualitativa o características.
- El mecanismo es el mismo que con una variable cualitativa: definir las correspondientes ficticias e introducir para cada variable cualitativa tantas ficticias como categorías menos una. Para cada atributo tendremos una categoría de referencia
- Ejemplo: brecha 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)

$$\text{salario}_i = \beta_1 + \beta_2 \text{educacion}_i + \delta_1 \text{mujer} + \gamma_1 \text{casada}_i + u_i$$

Modelo 1: MCO, usando las observaciones 1-526				
Variable dependiente: salario				
	coeficiente	Desv. típica	Estadístico t	valor p
-----	-----	-----	-----	-----
const	-0,0408223	0,681355	-0,05991	0,9522
educacion	0,494954	0,0497037	9,958	1,67e-21 ***
mujer	-2,08699	0,278456	-7,495	2,86e-13 ***
casada	1,18153	0,284630	4,151	3,86e-05 ***
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	5,137209	
R-cuadrado	0,282504	R-cuadrado corregido	0,278381	
F(3, 522)	68,51016	Valor p (de F)	2,28e-37	
Log-verosimilitud	-1345,748	Criterio de Akaike	2699,495	
Criterio de Schwarz	2716,556	Crit. de Hannan-Quinn	2706,175	

- ¿Cuál es la diferencia salarial entre un hombre casado y otro soltero?
- ¿Cuál es la diferencia salarial entre una mujer casada y otra soltera?
- ¿Cuál es la diferencia salarial entre una mujer soltera y un hombre casado?

Ejemplo: múltiples ficticias y efecto interacción (mujer casada)

$$\text{salario}_i = \beta_1 + \beta_2 \text{educac}_i + \delta_1 \text{mujer} + \gamma_1 \text{casada}_i + \alpha_1 (\text{mujer}_i \times \text{casada}_i) + u_i$$

Modelo 2: MCO, usando las observaciones 1-526

Variable dependiente: salario

	coeficiente	Desv. típica	Estadístico t	valor p
const	-1,02442	0,693110	-1,478	0,1400
educacion	0,493559	0,0485586	10,16	2,91e-22 ***
mujer	-0,368964	0,433412	-0,8513	0,3950
casada	2,64107	0,399356	6,613	9,33e-11 ***
mujer_casada	-2,82883	0,555558	-5,092	4,96e-07 ***
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)	60,31807	Valor p (de F)	7,38e-42	
Log-verosimilitud	-1332,975	Criterio de Akaike	2675,950	
Criterio de Schwarz	2697,276	Crit. de Hannan-Quinn	2684,300	

- ¿Cuál es la diferencia salarial entre un hombre casado y otro soltero?
- ¿Cuál es la diferencia salarial entre una mujer casada y otra soltera?
- ¿Cuál es la diferencia salarial entre una mujer casada y un hombre soltero?