TEMA 5
Variables ficticias

Cómo describir información cualitativa

- Muchas veces en el modelo de regresión aparecen factores cualitativos (sexo, raza, estado civil,….). En estos casos la información relevante se puede representar con la ayuda de **variables ficticias**.

- Las **variables ficticias** son variables binarias que toman valor 0,1.

- Al definir una variable ficticia debemos decidir a qué acontecimiento se le asigna el valor 1, y a cuál el 0.

- Ejemplo: la variable *sexo* es cualitativa. Para incluirla en un modelo de regresión hay que crear una variable ficticia que informe del sexo del individuo:

$$\text{mujer} = \begin{cases} 1 \text{ si es mujer} \\ 0 \text{ si es hombre} \end{cases}$$

- Utilizamos los valores 0 y 1 para describir información cualitativa porque ello conduce a modelos de regresión en los que los parámetros se prestan a interpretaciones muy naturales.

Variables ficticias aditivas y multiplicativas

Consideremos la siguiente ecuación de salarios:

$$\text{sal} = \beta_0 + \beta_1 \text{ne} + \varepsilon$$

Si queremos tener en cuenta el sexo para explicar el salario, tenemos que introducir **variables ficticias**.

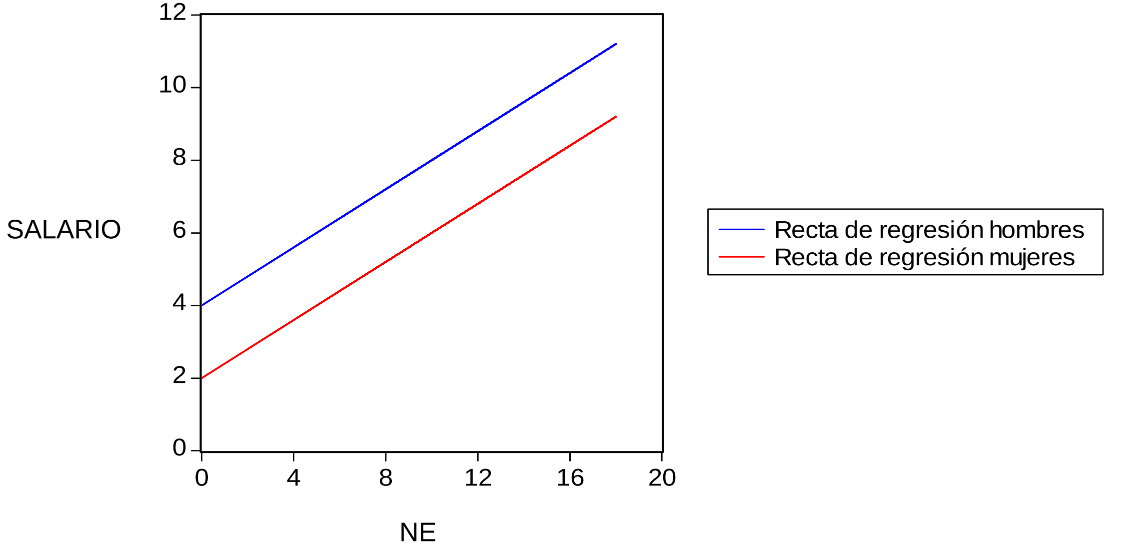
Ficticias aditivas:

Recogen un cambio en el término constante entre la ecuación de los hombres y la de las mujeres.

Ecuación hombres: $\text{sal} = \beta_0^H + \beta_1 \text{ne} + \varepsilon$

Ecuación mujeres: $\text{sal} = \beta_0^M + \beta_1 \text{ne} + \varepsilon$

Gráficamente:



La diferencia salarial entre hombres y mujeres no depende del nivel de estudios. El modelo refleja sólo que los hombres ganan un salario diferente, **en una cuantía fija**, al de las mujeres.

¿Cómo introducir ficticias aditivas?

Hay que definir una variable binaria (0,1) que informe sobre el "sexo" de los individuos. Si escogemos a los hombres como categoría de referencia, definimos la variable ficticia:

$$\text{mujer} = \begin{cases} 1 \text{ si es mujer} \\ 0 \text{ si es hombre} \end{cases}$$

El modelo con la ficticia aditiva es:

$$\boxed{\text{sal} = \beta_0 + \delta_0 \text{mujer} + \beta_1 \text{ne} + \varepsilon}$$

- Para *hombres* el modelo es: $\text{sal} = \beta_0 + \beta_1 \text{ne} + \varepsilon$
- Para *mujeres* el modelo es: $\text{sal} = (\beta_0 + \delta_0) + \beta_1 \text{ne} + \varepsilon$

δ_0: diferencial lineal entre el salario de una mujer y un hombre, independiente del nivel de educación. Si hay discriminación salarial a favor del hombre $\delta_0 < 0$.

En lugar de introducir la variable ficticia *mujer* se puede introducir la variable ficticia *hombre*:

$$\text{hombre} = \begin{cases} 1 \text{ si es hombre} \\ 0 \text{ si es mujer} \end{cases}$$

El modelo con la ficticia aditiva es:

$$\boxed{\text{sal} = \beta_0 + \delta_0 \,\text{hombre} + \beta_1 \text{ne} + \varepsilon}$$

En este caso, la categoría de referencia son las mujeres. Si hay discriminación salarial a favor del hombre $\delta_0 > 0$.

No importa si se escoge *hombre* o *mujer* como categoría de referencia, lo importante es saber cuál es el grupo de referencia para interpretar bien los parámetros:

β_0 : ordenada en el origen para el grupo de referencia.

δ_0 : diferencial lineal con respecto al grupo de referencia.

¿Se pueden incluir ambas variables a la vez en la ecuación?

$$sal = \beta_0 + \beta_0^H hombre + \beta_0^M mujer + \beta_1 ne + \varepsilon$$

NO, porque la ecuación presentaría multicolinealidad perfecta (**Trampa de las variables ficticias**).

SÍ se pueden incluir las dos ficticias si se elimina el término constante:

$$\boxed{sal = \beta_0^H hombre + \beta_0^M mujer + \beta_1 ne + \varepsilon}$$

β_0^H : ordenada en el origen para los hombres.

β_0^M : ordenada en el origen para las mujeres.

¿Cómo contrastar si existe discriminación salarial?

Depende del modelo:

(a) $\boxed{sal = \beta_0 + \delta_0 mujer + \beta_1 ne + \varepsilon}$

El contraste es: $\begin{cases} H_0 : \delta_0 = 0 \\ H_A : \delta_0 \neq 0 \end{cases}$

(b) $\boxed{sal = \beta_0^H hombre + \beta_0^M mujer + \beta_1 ne + \varepsilon}$

El contraste es: $\begin{cases} H_0 : \beta_0^H = \beta_0^M \\ H_A : \beta_0^H \neq \beta_0^M \end{cases}$

- ¿Cómo se interpretan los coeficientes de las ficticias si la **variable dependiente está en logaritmos**?

$$\log(\text{sal}) = \beta_0 + \delta_0 \text{mujer} + \beta_1 \text{ne} + \varepsilon$$

$100\delta_0$: diferencial salarial porcentual entre hombres y mujeres con el mismo nivel de educación.

- Existencia de **otras variables explicativas** en el modelo:

Indica la interpretación de δ_0 en los dos modelos siguientes y compara su significado:

$$\log(\text{sal}) = \beta_0 + \delta_0 \text{mujer} + \varepsilon$$

$$\log(\text{sal}) = \beta_0 + \delta_0 \text{mujer} + \beta_1 \text{ne} + \beta_2 \text{exper} + \varepsilon$$

Podemos incluir **varias variables ficticias** en la misma ecuación. Por ejemplo, en la ecuación de salarios podemos incluir también el hecho de si el individuo trabaja en el sur o no:

$$\log(\text{sal}) = \beta_0 + \delta_0 \text{mujer} + \delta_1 \text{sur} + \beta_1 \text{ne} + \varepsilon$$

$$\text{sur} = \begin{cases} 1 & \text{si trabaja en el sur} \\ 0 & \text{si no trabaja en el sur} \end{cases}$$

$100\delta_0$: diferencia salarial porcentual entre mujeres y hombres, manteniendo fijos el lugar de trabajo y la educación.

$100\delta_1$: diferencia salarial porcentual entre los individuos que trabajan en el sur y los que no, manteniendo fijos el sexo y la educación.

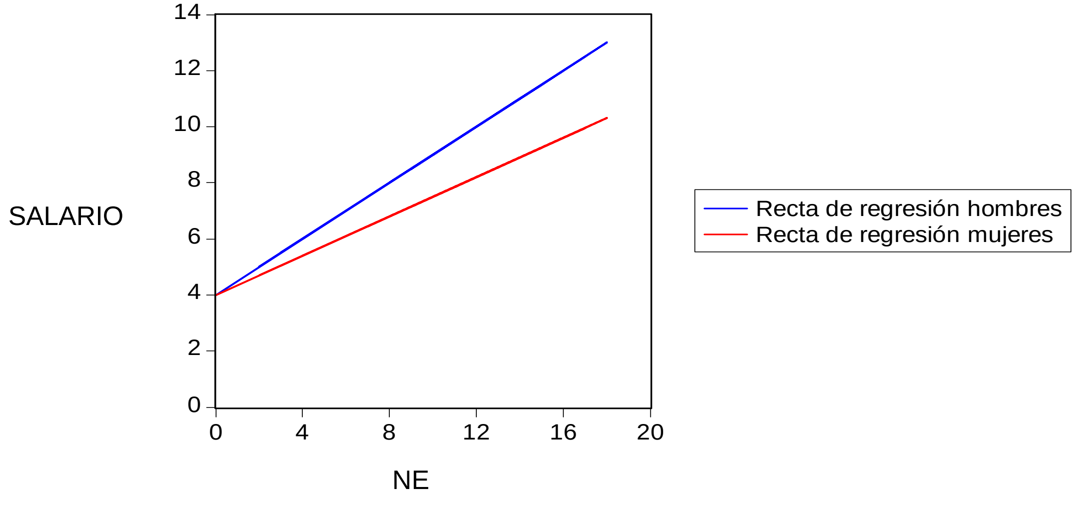
Ficticias multiplicativas:

Recogen un cambio en la pendiente entre la ecuación de los hombres y la de las mujeres.

Ecuación hombres: $sal = \beta_0 + \beta_1^H ne + \varepsilon$

Ecuación mujeres: $sal = \beta_0 + \beta_1^M ne + \varepsilon$

Gráficamente:



La diferencia salarial entre hombres y mujeres depende del nivel de estudios. Los rendimientos de la educación de los hombres son diferentes a los de las mujeres.

¿Cómo introducir ficticias multiplicativas?

Hay que multiplicar la variable ficticia correspondiente al sexo por la variable nivel de estudios. Considerando a los hombres como grupo de referencia:

$$\text{mujer} \cdot \text{ne} = \begin{cases} \text{ne si es mujer} \\ 0 \text{ si es hombre} \end{cases}$$

El modelo con la ficticia multiplicativa es:

$$\boxed{\text{sal} = \beta_0 + \beta_1 \text{ne} + \delta_1 \text{mujer} \cdot \text{ne} + \varepsilon}$$

- Para *hombres* el modelo es: $\text{sal} = \beta_0 + \beta_1 \text{ne} + \varepsilon$
- Para *mujeres* el modelo es: $\text{sal} = \beta_0 + (\beta_1 + \delta_1)\text{ne} + \varepsilon$

δ_1: diferencial proporcional al nivel de estudios entre el salario de una mujer y un hombre. Diferencia en la rentabilidad de la educación entre mujeres y hombres. Si la rentabilidad de la educación es menor para las mujeres $\delta_1 < 0$.

En lugar de introducir la variable ficticia *mujer·ne* se puede introducir la variable ficticia *hombre·ne* :

$$\mathrm{hombre} \cdot \mathrm{ne} = \begin{cases} \mathrm{ne\ si\ es\ hombre} \\ 0 \mathrm{\ si\ es\ mujer} \end{cases}$$

El modelo con la ficticia multiplicativa es:

$$\boxed{\mathrm{sal} = \beta_0 + \beta_1 \mathrm{ne} + \delta_1 \mathrm{hombre} \cdot \mathrm{ne} + \varepsilon}$$

En este caso, la categoría de referencia son las mujeres. Si la rentabilidad de la educación es menor para las mujeres $\delta_1 > 0$

No importa si se escoge *hombre* o *mujer* como categoría de referencia, lo importante es saber cuál es el grupo de referencia para interpretar bien los parámetros:

β_1 : pendiente (rendimientos de la educación) para el grupo de referencia.

δ_1 : diferencia en la pendiente con respecto al grupo de referencia.

¿Se pueden incluir ambas variables a la vez en la ecuación?

$$sal = \beta_0 + \beta_1 ne + \beta_1^H hombre \cdot ne + \beta_1^M mujer \cdot ne + \varepsilon$$

NO, porque la ecuación presentaría multicolinealidad perfecta (**Trampa de las variables ficticias**).

SÍ se pueden incluir las dos ficticias si se elimina la variable *ne*:

$$\boxed{sal = \beta_0 + \beta_1^H hombre \cdot ne + \beta_1^M mujer \cdot ne + \varepsilon}$$

β_1^H: pendiente para los hombres.

β_1^M: pendiente para las mujeres.

¿Cómo contrastaría la existencia de discriminación salarial en los dos modelos siguientes?

$$sal = \beta_0 + \beta_1 ne + \delta_1 hombre \cdot ne + \varepsilon$$

$$sal = \beta_0 + \beta_1^H hombre \cdot ne + \beta_1^M mujer \cdot ne + \varepsilon$$

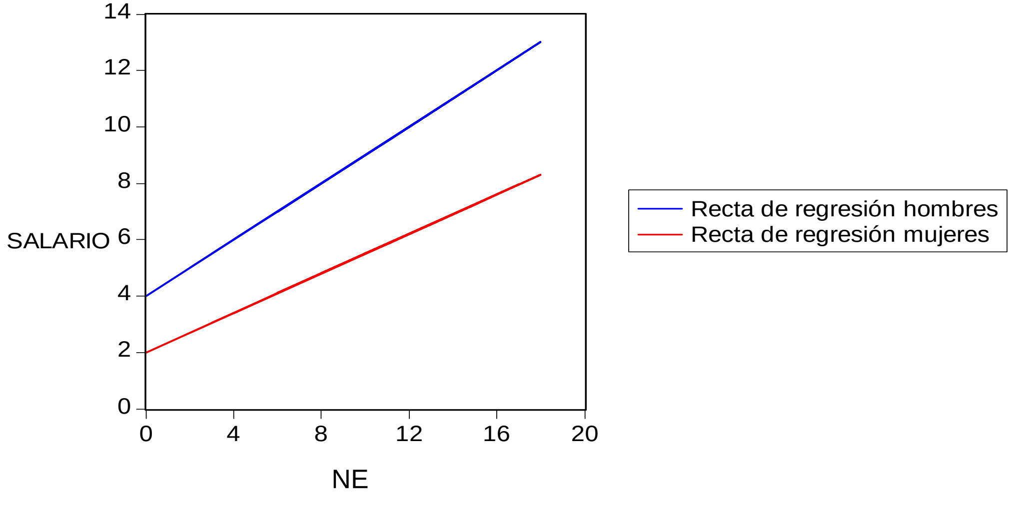
Ficticias aditivas y multiplicativas:

Recogen un cambio en la constante y en la pendiente entre la ecuación de los hombres y la de las mujeres.

Ecuación hombres: $\text{sal} = \beta_0^H + \beta_1^H \text{ne} + \varepsilon$

Ecuación mujeres: $\text{sal} = \beta_0^M + \beta_1^M \text{ne} + \varepsilon$

Gráficamente:



El modelo con la ficticias aditivas y multiplicativas es:

$$\boxed{sal = \beta_0 + \delta_0 mujer + \beta_1 ne + \delta_1 mujer \cdot ne + \varepsilon}$$

- Para *hombres* el modelo es: $sal = \beta_0 + \beta_1 ne + \varepsilon$
- Para *mujeres* el modelo es: $sal = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) ne + \varepsilon$

δ_0: diferencial lineal entre el salario de una mujer y un hombre, independiente del nivel de educación.

δ_1: diferencial en la rentabilidad de la educación entre mujeres y hombres.

Plantee una especificación alternativa al modelo anterior e interprete los coeficientes. Explique cómo contrastar la existencia de discriminación salarial en la especificación propuesta.

Variables ficticias para categorías múltiples

Si la **variable cualitativa tiene g categorías** hay que **incluir g-1 variables ficticias** en el modelo.

Ejemplo: si queremos incluir en la ecuación de salarios el estado civil, tenemos que definir 3 variables ficticias en el modelo con término constante (categoría de referencia = casados):

$$\text{soltero} = \begin{cases} 1 \text{ si es soltero} \\ 0 \text{ si no es soltero} \end{cases} \quad \text{divorciado} = \begin{cases} 1 \text{ si es divorciado} \\ 0 \text{ si no es divorciado} \end{cases} \quad \text{viudo} = \begin{cases} 1 \text{ si es viudo} \\ 0 \text{ si no es viudo} \end{cases}$$

$$\boxed{\log(\text{sal}) = \beta_0 + \delta_0 \text{mujer} + \delta_1 \text{soltero} + \delta_2 \text{divorciado} + \delta_3 \text{viudo} + \beta_1 \text{ne} + \varepsilon}$$

El término constante del modelo es la constante para el grupo de referencia (hombres casados).

El coeficiente de la variable ficticia para un grupo particular representa la diferencia estimada entre el término constante de ese grupo y el grupo de referencia.

Incorporar información ordinal:

Las variables ficticias permiten introducir en el modelo **variables ordinales** que toman un **reducido número de valores**.

Ejemplo: Supongamos que no conocemos los años exactos de educación sino solamente el grado que el estudiante ha alcanzado:

$$ne = \begin{cases} 0 & \text{sin estudios} \\ 1 & \text{primaria} \\ 2 & \text{secundaria} \\ 3 & \text{diplomatura} \\ 4 & \text{licenciatura} \end{cases}$$

Definiendo variables ficticias para cada nivel de estudios (categoría de referencia = *sin estudios*), tenemos:

$$\text{pri} = \begin{cases} 1 \text{ si educ} = 1 \\ 0 \text{ resto} \end{cases} \qquad \text{dip} = \begin{cases} 1 \text{ si educ} = 3 \\ 0 \text{ resto} \end{cases}$$

$$\text{sec} = \begin{cases} 1 \text{ si educ} = 2 \\ 0 \text{ resto} \end{cases} \qquad \text{lic} = \begin{cases} 1 \text{ si educ} = 4 \\ 0 \text{ resto} \end{cases}$$

$$\boxed{\log(\text{sal}) = \beta_0 + \delta_0 \text{mujer} + \delta_1 \text{prim} + \delta_2 \sec + \delta_3 \text{dip} + \delta_4 \text{lic} + \varepsilon} \qquad (1)$$

Este modelo permite que el salto de un ciclo de estudios a otro pueda tener un efecto diferente, por lo que es mucho más flexible que:

$$\boxed{\log(\text{sal}) = \beta_0 + \delta_0 \text{mujer} + \beta_1 \text{ne} + \varepsilon} \qquad (2)$$

Demuestre que cuando el salto de un ciclo de estudios a otro tiene un efecto constante en el salario, el modelo (1) se puede escribir como el modelo (2).

Efectos de interacción

Si hay interacción entre dos o más variables ficticias, el efecto de una ellas depende del valor que tomen las otras y viceversa.

Modelo sin efectos de interacción:

Modelo de salarios con educación (ordinal) y sexo (categoría de referencia = *hombre sin estudios*):

$$\log(\text{sal}) = \beta_0 + \delta_0 \text{mujer} + \delta_1 \text{prim} + \delta_2 \text{sec} + \delta_3 \text{dip} + \delta_4 \text{lic} + \varepsilon$$

Como no hay efectos de interacción, el efecto del sexo no depende del nivel de estudios y el efecto del nivel de estudios es el mismo para hombres y mujeres.

Veamos a continuación una tabla que clarifica la interpretación de los coeficientes, donde se representa el término constante del modelo en los diferentes grupos.

	Sin estudios	Prim.	Sec.	Dip.	Lic.
Hombre	β_0	$\beta_0 + \delta_1$	$\beta_0 + \delta_2$	$\beta_0 + \delta_3$	$\beta_0 + \delta_4$
Mujer	$\beta_0 + \delta_0$	$\beta_0 + \delta_0 + \delta_1$	$\beta_0 + \delta_0 + \delta_2$	$\beta_0 + \delta_0 + \delta_3$	$\beta_0 + \delta_0 + \delta_4$

Es un **modelo muy restringido** porque la diferencia salarial entre hombres y mujeres del mismo nivel educativo es siempre δ_0 .

¿Cuál es la diferencia salarial entre un hombre diplomado y un hombre licenciado?

¿Cuál es la diferencia salarial entre una mujer diplomada y una mujer licenciada?

Modelo con efectos de interacción:

Permite que la diferencia salarial entre hombres y mujeres dependa del nivel de estudios:

$$\log(\text{sal}) = \beta_0 + \delta_0 \text{mujer} + \delta_1 \text{prim} + \delta_2 \text{sec} + \delta_3 \text{dip} + \delta_4 \text{lic} + \\ + \theta_1 \text{mujer} \cdot \text{prim} + \theta_2 \text{mujer} \cdot \text{sec} + \theta_3 \text{mujer} \cdot \text{dip} + \theta_4 \text{mujer} \cdot \text{lic} + \varepsilon$$

	Sin estudios	Prim.	Sec.	Dip.	Lic.
Hombre	β_0	$\beta_0 + \delta_1$	$\beta_0 + \delta_2$	$\beta_0 + \delta_3$	$\beta_0 + \delta_4$
Mujer	$\beta_0 + \delta_0$	$\beta_0 + \delta_0 + \delta_1 + \theta_1$	$\beta_0 + \delta_0 + \delta_2 + \theta_2$	$\beta_0 + \delta_0 + \delta_3 + \theta_3$	$\beta_0 + \delta_0 + \delta_4 + \theta_4$

La diferencia salarial entre hombres y mujeres es:

Sin estudios: δ_0

Primaria: $\delta_0 + \theta_1$

Secundaria: $\delta_0 + \theta_2$

Diplomatura: $\delta_0 + \theta_3$

Licenciatura: $\delta_0 + \theta_4$

¿Cuál es la diferencia salarial entre un hombre diplomado y un hombre licenciado?

¿Cuál es la diferencia salarial entre una mujer diplomada y una mujer licenciada?

¿Cómo contrastaría si la discriminación salarial en función del sexo depende del nivel de estudios?