TEMA 5:
VARIABLES FICTICIAS

S. Álvarez, A. Beyaert, M. Camacho, M. González, A. Quesada
Departamento de Métodos Cuantitativos para la Economía y la Empresa
UNIVERSIDAD DE MURCIA

Lo que estudiaremos en este tema:

UNIVERSIDAD DE MURCIA

1. Cómo describir información cualitativa

2. Variables ficticias aditivas y multiplicativas

2.1 Introducción

2.2 Variables ficticias aditivas

2.3 Variables ficticias multiplicativas

2.4 Combinación de ficticias aditivas y multiplicativas

3. Cambio estructural y contraste de Chow

Bibliografía básica: Wooldridge, 2008, cap. 7

1. Cómo describir información cualitativa

- Las **variables ficticias** se utilizan para recoger factores cualitativos (sexo, raza, estado civil,…) en el modelo de regresión.

- Son variables binarias que toman valor 0,1.

- 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:

$$\text{mujer} = \begin{cases} 1 \text{ si es mujer} \\ 0 \text{ si no es mujer} \end{cases}$$

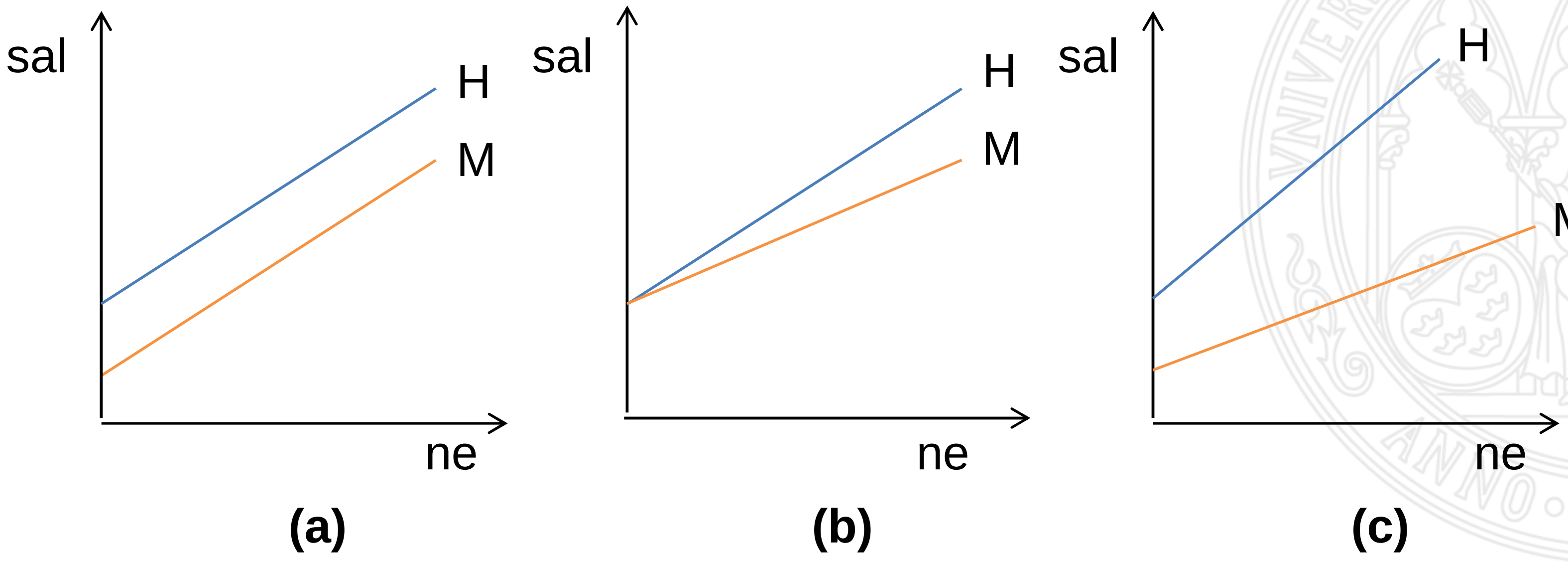
En este caso, *hombre* se conoce como categoría de referencia ó grupo base.

- Consideremos la siguiente ecuación de salarios:

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

- El sexo puede influir en la ecuación de salarios.

  Tres posibilidades:



2. Variables ficticias aditivas y multiplicativas
2.2 Variables ficticias aditivas

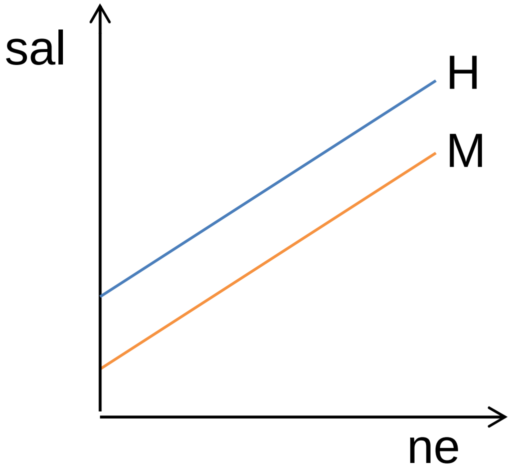
UNIVERSIDAD DE **MURCIA**

- Las **variables 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:



- En este caso, la diferencia salarial por razón de sexo no depende del nivel de estudios.

- Sea 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:

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

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

β_0 : ordenada en el origen del modelo de hombres.

$(\beta_0 + \delta_0)$: ordenada en el origen del modelo de mujeres.

δ_0 : diferencia entre el salario de una mujer respecto de un hombre con el mismo 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*:

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

- El modelo con la ficticia aditiva es:

$$sal = \beta_0 + \delta_0\, hombre + \beta_1 ne + \varepsilon$$

- En este caso, la categoría de referencia son las mujeres.

β_0: ordenada en el origen del modelo de mujeres.

$(\beta_0 + \delta_0)$: ordenada en el origen del modelo de hombres.

δ_0: diferencia entre el salario de un hombre respecto de una mujer con el mismo nivel de educación. Si hay discriminación salarial a favor del hombre, $\delta_0 > 0$.

- ¿Se puede estimar el modelo que incluye ambas variables a la vez?

CASO 1: $\text{sal} = \beta_0 + \beta_0^H \text{hombre} + \beta_0^M \text{mujer} + \beta_1 \text{ne} + \varepsilon$

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

CASO 2: $\text{sal} = \beta_0^H \text{hombre} + \beta_0^M \text{mujer} + \beta_1 \text{ne} + \varepsilon$

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

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

No hay categoría de referencia.

UNIVERSIDAD DE MURCIA

- ¿Cómo contrastar si existe discriminación salarial?

Depende del modelo:

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

El contraste es: $\begin{cases} H_0 : \delta_0 = 0 \\ H_1 : \delta_0 \neq 0 \end{cases}$

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

El contraste es: $\begin{cases} H_0 : \beta_0^H = \beta_0^M \\ H_1 : \beta_0^H \neq \beta_0^M \end{cases}$

¿Cómo contrastaría si existe discriminación salarial a favor del hombre?

UNIVERSIDAD DE MURCIA

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

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

$100\delta_0$: diferencia porcentual del salario de una mujer con respecto de un hombre con el mismo nivel de educación.

- Podemos incluir **varias variables ficticias** en la misma ecuación:

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

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

$100\delta_0$: diferencia porcentual del salario de una mujer respecto de un hombre con la misma educación y que trabajan en la misma zona.

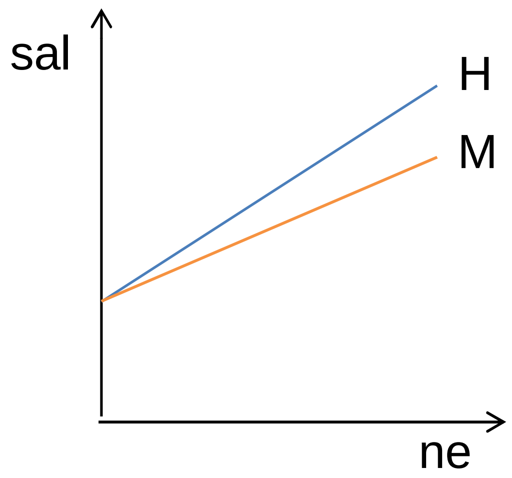
$100\delta_1$: diferencia porcentual del salario entre un individuo que trabaja en el sur con respecto a otro que no, con el mismo sexo y la misma educación.

- Las **variables 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:



- En este caso, la diferencia salarial por razón de sexo depende del nivel de estudios.

- La ficticia multiplicativa se define como:

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

- El modelo con la ficticia multiplicativa es:

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

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

β_1 : pendiente del modelo de hombres.

$(\beta_1 + \delta_1)$: pendiente del modelo de mujeres.

δ_1: diferencia en el rendimiento de la educación de las mujeres respecto de los hombres.

Plantee un modelo donde la variable ficticia multiplicativa sea *hombre·ne* e interprete sus coeficientes.

- ¿Se puede estimar el modelo que incluye ambas variables a la vez?

CASO 1: $$\text{sal} = \beta_0 + \beta_1 \text{ne} + \beta_1^H \text{hombre} \cdot \text{ne} + \beta_1^M \text{mujer} \cdot \text{ne} + \varepsilon$$

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

CASO 2: $$\text{sal} = \beta_0 + \beta_1^H \text{hombre} \cdot \text{ne} + \beta_1^M \text{mujer} \cdot \text{ne} + \varepsilon$$

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

β_1^H: pendiente del modelo de hombres.
β_1^M: pendiente del modelo de mujeres.

No hay categoría de referencia.

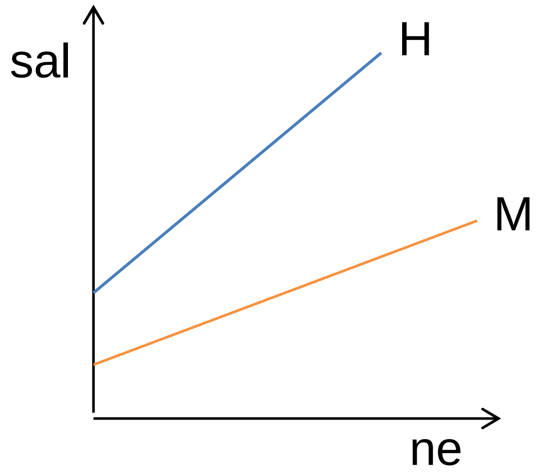
UNIVERSIDAD DE MURCIA

- Si el sexo cambia la pendiente y la constante de la ecuación, tendríamos:

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:



- Podemos modelizarlo añadiendo ficticias aditivas y multiplicativas.

- El modelo con las ficticias aditivas y multiplicativas es:

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

- 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: parte de la diferencia salarial de una mujer respecto de un hombre que es independiente del nivel de educación.

δ_1: parte de la diferencia salarial de una mujer respecto de un hombre que depende del nivel de educación. Es la diferencia en el rendimiento de la educación de una mujer respecto de un hombre.

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

- Existe ***cambio estructural*** cuando los coeficientes del modelo no son iguales en submuestras diferentes.

$$y_i = \beta_0^{(1)} + \beta_1^{(1)} x_{1i} + \ldots + \beta_k^{(1)} x_{ki} + \varepsilon_i, \quad i \in \text{submuestra } 1$$

$$y_i = \beta_0^{(2)} + \beta_1^{(2)} x_{1i} + \ldots + \beta_k^{(2)} x_{ki} + \varepsilon_i, \quad i \in \text{submuestra } 2$$

- Ejemplo: ecuación de salarios con coeficientes distintos para hombres y mujeres.

$$sal_i = \beta_0^{H} + \beta_1^{H} ne_i + \varepsilon_i, \quad \text{si el individuo es hombre}$$

$$sal_i = \beta_0^{M} + \beta_1^{M} ne_i + \varepsilon_i, \quad \text{si el individuo es mujer}$$

- El ***contraste de Chow*** permite contrastar si hay cambio en la estructura del modelo.

$$H_0: \text{no hay cambio estructural}$$
$$H_1: \text{hay cambio estructural}$$

- Bajo la H_0 no hay cambio, y el ***modelo restringido*** (MR) es:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, \ldots, N$$

- Bajo la H_1 hay cambio estructural, y el ***modelo no restringido*** (MNR) es:

$$y_i = \beta_0^{(1)} + \beta_1^{(1)} x_{1i} + \ldots + \beta_k^{(1)} x_{ki} + \varepsilon_i, \quad i \in \text{submuestra } 1$$
$$y_i = \beta_0^{(2)} + \beta_1^{(2)} x_{1i} + \ldots + \beta_k^{(2)} x_{ki} + \varepsilon_i, \quad i \in \text{submuestra } 2$$

- Podemos escribir los dos modelos contenidos en el **MNR** como uno solo utilizando **variables ficticias**.

- Dos posibilidades:

MNR (a): $y_i = \beta_0^{(1)} D_{1i} + \beta_0^{(2)} D_{2i} + \beta_1^{(1)} D_{1i} x_{1i} + \beta_1^{(2)} D_{2i} x_{1i} + \ldots + \beta_k^{(1)} D_{1i} x_{ki} + \beta_k^{(2)} D_{2i} x_{ki} + \varepsilon_i$

MNR (b): $y_i = \beta_0^{(2)} + \delta_0 D_{1i} + \beta_1^{(2)} x_{1i} + \delta_1 D_{1i} x_{1i} + \ldots + \beta_k^{(2)} x_{ki} + \delta_k D_{1i} x_{ki} + \varepsilon_i$

donde

$$D_{1i} = \begin{cases} 1 \text{ si } i \in \text{ submuestra } 1 \\ 0 \text{ si } i \notin \text{ submuestra } 1 \end{cases}, \quad D_{2i} = \begin{cases} 1 \text{ si } i \in \text{ submuestra } 2 \\ 0 \text{ si } i \notin \text{ submuestra } 2 \end{cases}$$

Nota: Reparamentrizando con $D_{2i} = 1 - D_{1i}$ el MNR (a) se llega al MNR (b).

- **Hipótesis del contraste de Chow**. Dos posibilidades:

MNR (a): $y_i = \beta_0^{(1)} D_{1i} + \beta_0^{(2)} D_{2i} + \beta_1^{(1)} D_{1i} x_{1i} + \beta_1^{(2)} D_{2i} x_{1i} + \ldots + \beta_k^{(1)} D_{1i} x_{ki} + \beta_k^{(2)} D_{2i} x_{ki} + \varepsilon_i$

$$H_0 : \beta_0^{(1)} = \beta_0^{(2)}, \ldots, \beta_k^{(1)} = \beta_k^{(2)}$$
$$H_1 : \text{No } H_0$$

MNR (b): $y_i = \beta_0^{(2)} + \delta_0 D_{1i} + \beta_1^{(2)} x_{1i} + \delta_1 D_{1i} x_{1i} + \ldots + \beta_k^{(2)} x_{ki} + \delta_k D_{1i} x_{ki} + \varepsilon_i$

$$H_0 : \delta_0 = 0, \ldots, \delta_k = 0$$
$$H_1 : \text{No } H_0$$

- En ambos casos, **el estadístico de contraste** es el mismo:

$$F = \frac{\left(SCE_R - SCE_{NR}\right)/K}{SCE_{NR}/\left(N-2K\right)} \underset{H_0}{\sim} F_{q,N-K}$$

- SCE_R: suma de cuadrados de los residuos del modelo sin ficticias (MR).

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1,\ldots,N$$

- SCE_{NR}: suma de cuadrados de los residuos del modelo con ficticias. Dos posibilidades: **MNR (a)** o **MNR (b)**.

- K: número de restricciones bajo la hipótesis nula (q=K).

- N-$2K$: grados de libertad del MNR.

- Ejemplo: contraste de Chow en la ecuación de salarios

MR: $\text{sal}_i = \beta_0 + \beta_1 \text{ne}_i + \varepsilon_i, \quad i=1,\ldots,N \quad \Rightarrow \text{SCE}_R$

MNR: $\text{sal}_i = \beta_0 + \delta_0 \text{mujer}_i + \beta_1 \text{ne}_i + \delta_1 \text{mujer}_i \cdot \text{ne}_i + \varepsilon_i, \quad i=1,\ldots,N \Rightarrow \text{SCE}_{NR}$

- Las hipótesis nula y alternativa son: $H_0 : \delta_0 = \delta_1 = 0$

  $H_1 : \text{No } H_0$

- El estadístico de contraste es: $F = \dfrac{(\text{SCE}_R - \text{SCE}_{NR})/2}{\text{SCE}_{NR}/(N-4)} \underset{H_0}{\sim} F_{q,N-K}$

- Un RH_0 implica que ecuación salarial de los hombres no es la misma que la de las mujeres (al menos un coeficiente es diferente).

Lo que hemos aprendido:

UNIVERSIDAD DE MURCIA

- Representar información cualitativa mediante el uso de variables ficticias

- Modelizar cambios tanto en la constante (ficticias aditivas) como en las pendientes (ficticias multiplicativas) de un modelo de regresión en función de las categorías representadas

- Interpretación de los parámetros de las variables ficticias:
  - Cuando se incluyen tantas ficticias como categorías menos una, los parámetros miden diferenciales respecto a la referencia
  - Cuando se incluyen tantas ficticias como categorías, los parámetros miden los niveles de cada categoría

- Realizar un contraste de cambio estructural mediante el uso de variables ficticias