ECONOMETRIA

Tema 4: ANÁLISIS DE REGRESIÓN CON INFORMACIÓN CUALITATIVA

César Alonso

Universidad Carlos III de Madrid

Introducción

- En el contexto del modelo de regresión, existen con frecuencia aspectos de interés que son de **naturaleza cualitativa** y que no pueden medirse numéricamente por medio de una variable cuantitativa.
- Las variables ficticias (o artificiales, o binarias o "dummy") se emplean para recoger información de carácter cualitativo:
  - ser hombre o mujer;
  - ser o no inmigrante;
  - estar o no estar casado;
  - residir en una determinada provincia o comunidad autónoma;
  - que una empresa pertenezca al sector manufacturero o al sector servicios
  - que una empresa tenga un determinado tamaño;
  - que una empresa cotice o no en bolsa;
  - etc.

Introducción

- Utilizando variables ficticias, podemos medir el efecto del factor cualitativo.
  - Además, podremos contrastar fácilmente si el efecto del factor cualitativo es relevante.
- Las variables ficticias se emplean en los modelos de regresión cuando queremos ver si el efecto de alguna/s de las X's sobre Y varía según alguna característica de la población (sexo, raza, tamaño de la empresa, etc).
  Típicamente, **las variables ficticias toman valor** 1 **en una categoría y valor** 0 **en el resto**. Por ejemplo:

$$Mujer = \begin{cases} 1 & \text{si el individuo es mujer} \\ 0 & \text{si el individuo es hombre} \end{cases}$$

Introducción

$$Hombre = \begin{cases} 1 & \text{si el individuo es hombre} \\ 0 & \text{si el individuo es mujer} \end{cases}$$

$$Pequeña = \begin{cases} 1 & \text{si la empresa es pequeña} \\ 0 & \text{en caso contrario} \end{cases}$$

$$Mediana = \begin{cases} 1 & \text{si la empresa es mediana} \\ 0 & \text{en caso contrario} \end{cases}$$

$$Grande = \begin{cases} 1 & \text{si la empresa es grande} \\ 0 & \text{en caso contrario} \end{cases}$$

- Podemos distinguir dos aspectos que pueden recogerse con ayuda de las variables artificiales:
  - Efecto aditivo (diferencias en el término constante)
  - Efecto interacción (diferencias en las pendientes)

Efecto aditivo

- Empleamos las variables ficticias para modelizar cambios en el término constante del modelo.
- Ya vimos un ejemplo cuando presentamos el modelo de regresión múltiple:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad i = 1, \ldots, n, \qquad \text{donde}$$

  - Y_i = salario *(o alguna transformación de éste)*,
  - X_{1i} = educación,
  - $X_{2i} = \text{mujer}_i = \begin{cases} 1 & \text{si el individuo es mujer} \\ 0 & \text{si el individuo es hombre} \end{cases}$

  Tenemos que:

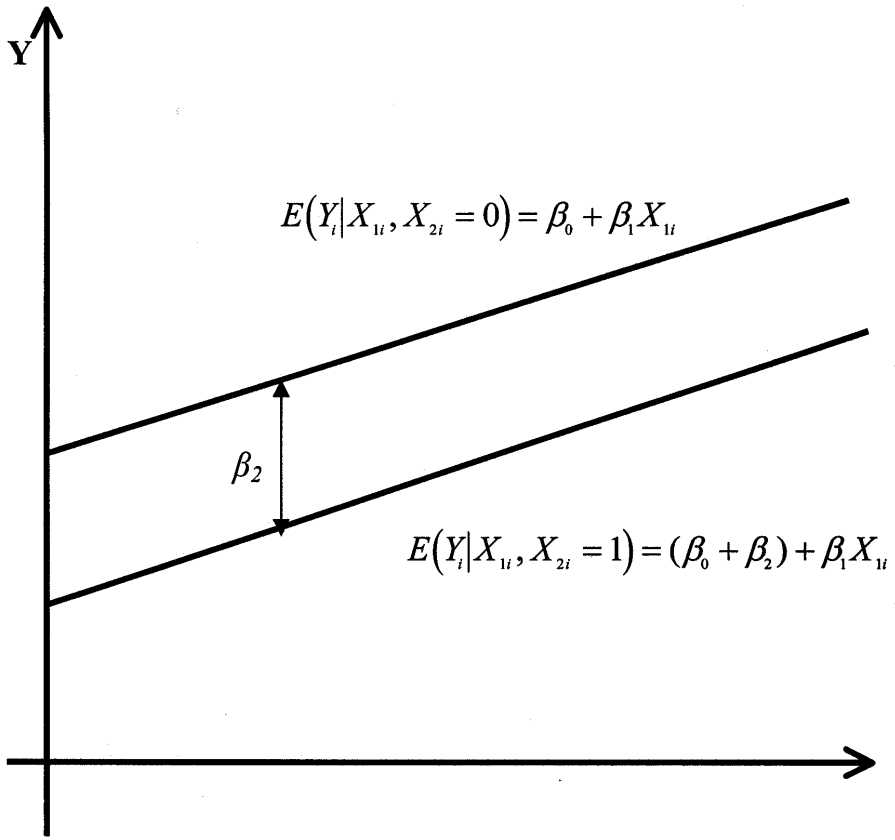
$$E(Y_i|X_{1i}, X_{2i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i},$$

  con lo cual:

$$E(Y_i|X_{1i},\ \text{mujer}) = E(Y_i|X_{1i}, X_{2i} = 1) = (\beta_0 + \beta_2) + \beta_1 X_{1i},$$
$$E(Y_i|X_{1i},\ \text{hombre}) = E(Y_i|X_{1i}, X_{2i} = 0) = \beta_0 + \beta_1 X_{1i}$$

Efecto aditivo

- $\beta_2 = E(Y_i|X_{1i}, \text{mujer}) - E(Y_i|X_{1i}, \text{hombre})$
  es la diferencia, en media, entre el salario de una mujer y el de un hombre, para un mismo nivel educativo.
- Suponiendo $\beta_2 < 0$, tendríamos el siguiente gráfico:



Efecto aditivo

- Otras dos formulaciones alternativas de este mismo modelo serían:

1. $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{3i} + \varepsilon_i, \quad i = 1, \ldots, n$
   donde:
   $$X_{3i} = \text{hombre}_i = \begin{cases} 1 & \text{si el individuo es hombre} \\ 0 & \text{si el individuo es mujer} \end{cases}.$$

Efecto aditivo

Ahora tenemos que:

$$E(Y_i|X_{1i}, X_{2i}) = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{3i},$$

con lo cual:

$$\begin{aligned} E(Y_i|X_{1i}, \text{mujer}) &= E(Y_i|X_{1i}, X_{3i} = 0) = \alpha_0 + \alpha_1 X_{1i}, \\ E(Y_i|X_{1i}, \text{hombre}) &= E(Y_i|X_{1i}, X_{3i} = 1) = (\alpha_0 + \alpha_2) + \alpha_1 X_{1i} \end{aligned}$$

$\alpha_2 = E(Y_i|X_{1i}, \text{hombre}) - E(Y_i|X_{1i}, \text{mujer})$ es la diferencia, en media, entre el salario de un hombre y el de una mujer, para un mismo nivel educativo.

Obviamente:

$$\begin{aligned} \alpha_1 &= \beta_1 \\ \alpha_0 &= \beta_0 + \beta_2 \\ \alpha_0 + \alpha_2 &= \beta_0 \end{aligned}$$

Efecto aditivo

2. $Y_i = \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{3i} + \varepsilon_{i,} \quad i = 1, \ldots, n$

Tenemos que:

$$E(Y_i|X_{1i}, X_{2i}, X_{3i}) = \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{3i},$$

con lo cual:

$$E(Y_i|X_{1i}, \text{mujer}) = E(Y_i|X_{1i}, X_{2i} = 1, X_{3i} = 0) = \delta_2 + \delta_1 X_{1i},$$
$$E(Y_i|X_{1i}, \text{hombre}) = E(Y_i|X_{1i}, X_{2i} = 0, X_{3i} = 1) = \delta_3 + \delta_1 X_{1i}$$

$(\delta_3 - \delta_2) = E(Y_i|X_{1i},\text{hombre}) - E(Y_i|X_{1i},\text{mujer})$ es la diferencia, en media, entre el salario de un hombre y el de una mujer, para un mismo nivel educativo.

Obviamente:

$$\delta_1 = \alpha_1 = \beta_1$$
$$\delta_2 = \alpha_0 = \beta_0 + \beta_2$$
$$\delta_3 = \alpha_0 + \alpha_2 = \beta_0$$

Efecto aditivo

- Sin embargo, nótese que un modelo como

$$Y_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \varepsilon_i, \quad i = 1, \ldots, n$$

NO sería válido, ya que habría multicolinealidad exacta:

$$X_{2i} + X_{3i} = 1 \quad \forall i = 1, \ldots, n$$

- ¿Cómo contrastaríamos si existen diferencias en media entre el salario-hora de un hombre y de una mujer, para un mismo nivel educativo? Para cada una de las tres representaciones posibles del mismo modelos, tendríamos:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad H_0 : \beta_2 = 0$$

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{3i} + \varepsilon_i \qquad H_0 : \alpha_2 = 0$$

$$Y_i = \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{3i} + \varepsilon_i \qquad H_0 : \delta_2 = \delta_3$$

Efecto interacción

- Empleamos las variables ficticias para modelizar cambios en el efecto de las X's sobre Y (en las pendientes del modelo).
- Veamos un ejemplo con efectos aditivos e interacción:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{4i} + \varepsilon_i, \quad i = 1, \ldots, n$$

donde:

$$X_{2i} = \text{mujer}_i = \begin{cases} 1 & \text{si el individuo es mujer} \\ 0 & \text{si el individuo es hombre} \end{cases}$$

$$X_{4i} = X_{1i} \times X_{2i} = \begin{cases} X_{1i} & \text{si el individuo es mujer} \\ 0 & \text{si el individuo es hombre} \end{cases}$$

- Tenemos que:

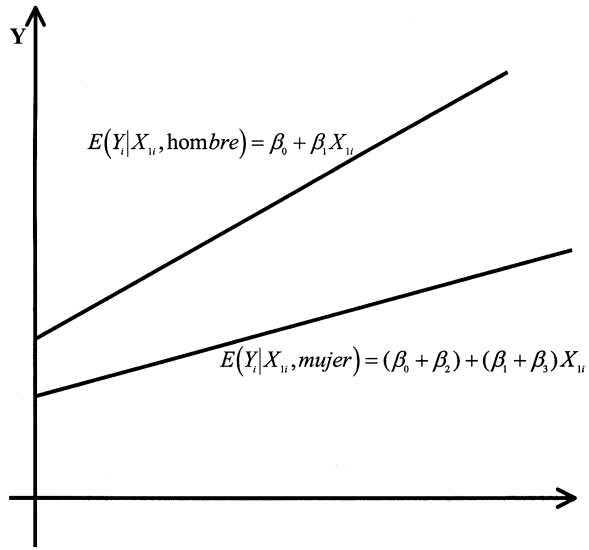
$$E(Y_i|X_{1i}, X_{2i}, X_{4i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{4i},$$

$\Rightarrow$

$$E(Y_i|X_{1i}, \text{mujer}) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{1i},$$
$$E(Y_i|X_{1i}, \text{hombre}) = \beta_0 + \beta_1 X_{1i},$$

Efecto interacción

- β_2 mide la diferencia en el término constante entre hombres y mujeres.
- β_3 mide la diferencia en la pendiente entre hombres y mujeres:
  Si la educación (X_1) aumenta 1 unidad, el salario-hora varía en media en:
  - $\beta_1 + \beta_3$ unidades en las mujeres
  - β_1 unidades en los hombres
- Suponiendo $\beta_2 < 0$, $\beta_3 < 0$:



Efecto interacción

- Este gráfico ilustraría una situación de discriminación salarial en contra de las mujeres, donde la brecha salarial aumenta con el nivel de educación X_1.
- Si Y fuera una función del salario, la diferencia vertical entre ambas rectas mediría
  - La diferencia salarial media (en euros) entre hombres y mujeres con igual nivel de educación, si $Y =$ Salario (en euros).
  - La diferencia salarial media (en tanto por uno) entre hombres y mujeres con igual nivel de educación, si $Y = \ln(\text{Salario})$.

Efecto interacción

- ¿Cómo se contrastaría si las variaciones unitarias en la educación generan el mismo efecto medio sobre el salario-hora en hombres y en mujeres?

$$H_0 : \beta_3 = 0$$

- ¿Cómo se contrastaría si el término constante es el mismo para hombres y para mujeres?

$$H_0 : \beta_2 = 0$$

- ¿Cómo se contrastaría si el modelo de determinación salarial es el mismo en hombres y en mujeres?

$$H_0 : \beta_2 = \beta_3 = 0$$

Efecto interacción

- **Comentarios**:
  - Igual que hemos visto con el efecto aditivo, existen otras formulaciones alternativas de este mismo modelo.
    - Por ejemplo:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{3i} + \alpha_3 X_{5i} + \varepsilon_i, \quad i = 1, \ldots, n$$

donde:

$$X_{3i} = \text{hombre}_i = \begin{cases} 1 & \text{si el individuo es hombre} \\ 0 & \text{si el individuo es mujer} \end{cases}$$

$$X_{5i} = X_{1i} \times X_{3i} = \begin{cases} X_{1i} & \text{si el individuo es hombre} \\ 0 & \text{si el individuo es mujer} \end{cases}$$

    - O, alternativamente:

$$Y_i = \delta_1 X_{2i} + \delta_2 X_{3i} + \delta_3 X_{4i} + \delta_4 X_{5i} + \varepsilon_{i,} \quad i = 1, \ldots, n$$

    - Sin embargo, NO sería válido un modelo como:

$$Y_i = \gamma_1 X_{2i} + \gamma_2 X_{3i} + \gamma_3 X_{4i} + \gamma_4 X_{5i} + \gamma_5 X_{1i} + \varepsilon_i, \quad i = 1, \ldots, n,$$

ya que habría multicolinealidad exacta:

$$X_{4i} + X_{5i} = X_{1i} \quad \forall i = 1, \ldots, n$$

Efecto interacción

- **Podríamos tener más de dos categorías**. Por ejemplo, supongamos que las empresas se distribuyen en tres sectores distintos:

$$V_i = \beta_0 + \beta_1 S1_i + \beta_2 S2_i + \beta_3 P_i + \beta_4 (P_i \times S1_i) + \beta_5 (P_i \times S2_i) + \varepsilon_i,$$

donde:

V_i = ventas de la empresa

P_i = gastos en publicidad de la empresa

$$S1_i = \begin{cases} 1 & \text{si la empresa pertenece al sector 1} \\ 0 & \text{si la empresa pertenece al sector 2 ó 3} \end{cases}$$

$$S2_i = \begin{cases} 1 & \text{si la empresa pertenece al sector 2} \\ 0 & \text{si la empresa pertenece al sector 1 ó 3} \end{cases}$$

Entonces:

$$E(V_i|P_i, \text{sector 1}) = (\beta_0 + \beta_1) + (\beta_3 + \beta_4) P_i$$
$$E(V_i|P_i, \text{sector 2}) = (\beta_0 + \beta_2) + (\beta_3 + \beta_5) P_i$$
$$E(V_i|P_i, \text{sector 3}) = \beta_0 + \beta_3 P_i$$

- En esta representación del modelo, al incluir tanto el término constante como P_i, sólo incluimos efectos aditivos y efectos interacción para dos de los sectores:
  - β_0 es el término constante del sector cuya variable ficticia ignoramos (Sector 3).
  - β_3 es la pendiente (el efecto de la publicidad) del sector cuya variable ficticia ignoramos (Sector 3).
  - Las ordenadas en el origen para los otros sectores 1 y 2 son $\beta_0 + \beta_1$ y $\beta_0 + \beta_2$, respectivamente.
  - Las pendientes (el efecto de la publicidad) para los otros sectores 1 y 2 son $\beta_3 + \beta_4$ y $\beta_3 + \beta_5$, respectivamente.
- Una representación alternativa y equivalente (entre otras):

$$\begin{aligned} V_i &= \delta_1 S1_i + \delta_2 S2_i + \delta_3 S3_i + \delta_4 (P \times S1_i) + \\ &\quad + \delta_5 (P_i \times S2_i) + \delta_6 (P_i \times S3_i) + \varepsilon_i \end{aligned}$$