ECONOMETRÍA I

Tema 5: Análisis de regresión múltiple con información cualitativa

Patricia Moreno
Juan Manuel Rodriguez Poo
Alexandra Soberon
Departamento de Economía

Modelos con información cualitativa

- En los modelos de regresión podemos estar interesados en ver si el efecto de alguna de las X sobre Y varía según alguna característica de la población (sexo, raza, tamaño de la empresa, etc).
- Utilizando variables ficticias (binarias o "dummy") podemos ser capaces de medir el efecto del factor cualitativo, así como constrastar si el efecto del factor cualitativo es relevante.
- Las variables ficticias toman valor 1 en una categoría y valor 0 en el resto. Ejemplo:

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

donde *hombre* $= 0$ es el grupo base.

Modelos con información cualitativa

- Consideramos un modelo de regresión múltiple en el cual queremos determinar el efecto del sexo y de la educación sobre los salarios:

$$wage = \beta_0 + \beta_1 educ + \delta_0 female + u$$

donde δ_0 es el parámetro asociado a la variable ficticia.

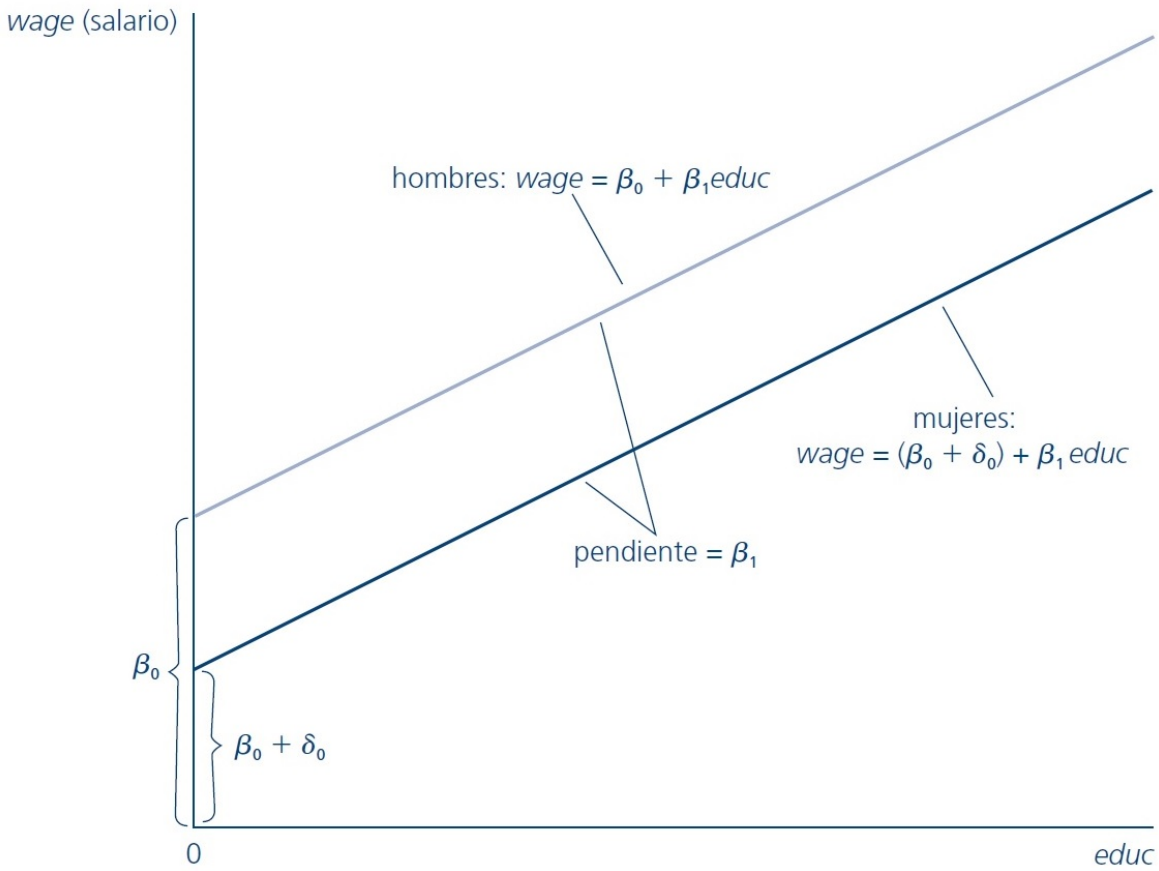
- δ_0 refleja la diferencia existente en el salario/hora entre una mujer y un hombre, para un nivel dado de educación. Suponiendo que $E(u|female, educ) = 0$,

$$\delta_0 = \underbrace{E(wage|female = 1, educ)}_{\beta_0+\delta_0 female+\beta_1 educ} - \underbrace{E(wage|female = 0, educ)}_{\beta_0+\delta_0 female}.$$

- $\delta_0 < 0$ describe un cambio en el término constante entre hombre y mujeres (aunque ambos tienen la misma pendiente, β_1).

Modelos con información cualitativa

Gráfica de $wage = \beta_0 + \delta_0 female + \beta_1 educ$ en la que $\delta_0 < 0$.



Fuente: Wooldridge (2005)

Modelos con información cualitativa

- **Trampa de las ficticias**: aparece cuando se incluyen dos variables ficticias para describir el mismo grupo (multicolinealidad perfecta).

$$female + male = 1.$$

- **Interpretación**: $\widehat{wage} = 7,10 - 2,51 female$
  - El término constante (7,10) es el salario medio para los hombres (grupo de referencia).
  - El coeficiente $2,51$ es la diferencia entre salarios medios de hombres y mujeres de la muestra. Es decir, las mujeres ganan, en media, $2,51$\$ menos por hora.

Dummies para categorías múltiples

- Las variables ficticias pueden ser usadas para controlar por alguna característica con categorías múltiples. Ejemplo: sexo y estado civil.
- Ejemplo: Si queremos distinguir las diferencias salarias entre 4 grupos de individuos (hombres casados, mujeres casadas, hombres solteros y mujeres solteras) el modelo a estimar sería:

$$\begin{aligned} log(wage) &= \beta_0 + \delta_1 marrmale + \delta_2 marrfem + \delta_3 sinfem + \beta_1 educ \\ &+ \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u \end{aligned}$$

- Para evitar la trampa de las dummies incluimos 3 de los 4 grupos de individuos. Grupo de referencia: hombres solteros.
- $\delta_2 - \delta_1$: diferencia salarial entre mujeres solteras y casadas.

Dummies para categorías múltiples

- Problema: no podemos contrastar directamente si esta diferencia es significativa.
- Para determinar la significatividad de esta diferencia salarial tendríamos que reestimar la ecuación tomando uno de estos grupos como grupo base (ej: mujeres casadas):

$$\begin{aligned} low(wage) &= \beta_0 + \delta_0 marrmale + \delta_3 singmale + \delta_2 singfem + \beta_1 educ \\ &+ \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u. \end{aligned}$$

- **Hipótesis nula**: $H_0 : \delta_2 = 0$.

Incorporar información ordinal con dummies

- **Variables ordinales**: aquellas variables que distinguen las categorías de acuerdo a un determinado criterio. Ejemplo: CR es la clasificación crediticia de los bancos asignada por las agencias financieras (Moody y SP)
- ¿Cómo podemos incorporar (CR) en un modelo para explicar el tipo de interés de los bonos municipales (MBR)?

$$MBR = \beta - 0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + u$$

- δ_j es la diferencia en MBR entre un municipio de crédito j y otro de crédito $j - 1$.
- $CR = 0$ es el grupo de referencia y sólo incluimos 4 ficticias.

Interacciones con variables ficticias

- Interactuar con variables ficticias es similar subdividir un grupo.
- Tenemos ficticias para male (hombre), hsgrad (título bachillerato) y colgrad (graduado escolar). Podemos añadir en el modelo $male * hsgrad$ y $male * colgrad$.
- Siendo mujeres con título de bachillerato el grupo de control, el modelo a estimar es

$$\begin{aligned} Y &= \beta_0 + \delta_1 male + \delta_2 hsgrad + \delta_3 colgrad + \delta_4 male * hsgrad \\ &+ \delta_5 male * colgrad + \beta_1 X + u \end{aligned}$$

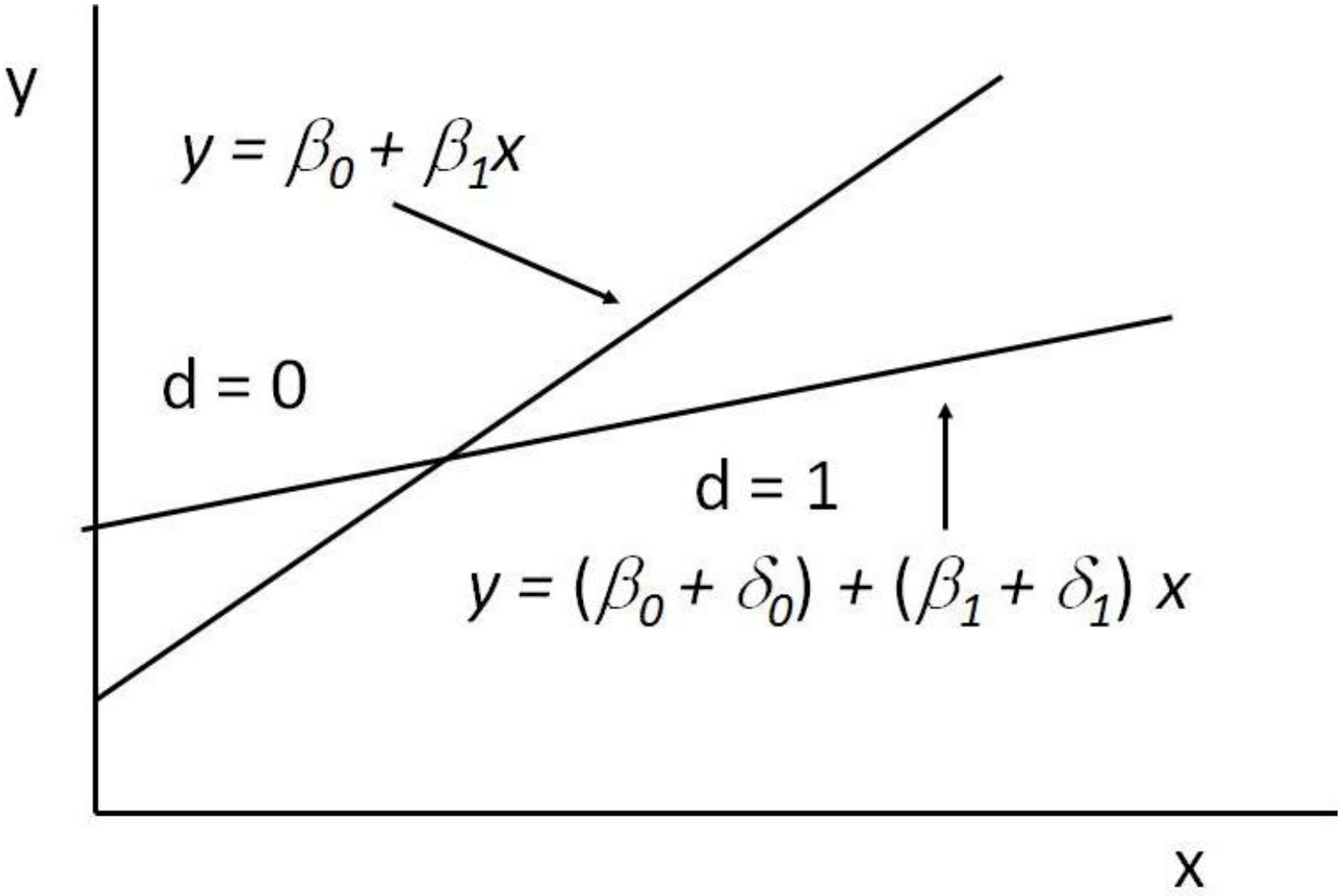
- Esta formulación es interesante porque permite contrastar directamente la significatividad de la interacción entre el sexo y el nivel educativo.

Variables ficticias con pendientes distintas

- Las variables ficticias también pueden interactuar con variables explicativas que no son binarias.
- Esto permite trabajar con ecuaciones con distintas pendientes para cada uno de los grupos.
- Si queremos contrastar si la rentabilidad de la educación es la misma para hombres y para mujeres (sin abandonar la posibilidad de que exista un diferencial salarial por género):

$$log(wage) = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female)educ + u.$$

Variables ficticias con pendientes distintas



Fuente: Wooldridge (2005)

Variables ficticias con pendientes distintas

- Ecuación para hombres:

$$log(wage) = \beta_0 + \beta_1 educ + u.$$

- Ecuacción para mujeres:

$$log(wage) = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) educ + u.$$

- Modelo a estimar:

$$log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female * educ + u.$$

donde

  - δ_0: diferencia entre los términos constantes de hombres y mujeres.
  - δ_1: diferencia entre la pendiente de *educ* para hombres y mujeres.

Contraste de diferencias entre grupos

- **Objetivo**: Contrastar si una función de regresión es distinta para cada grupo (ej: atletas y no atletas). En otras palabras, contrastar la significatividad conjunta de dummies y de sus interaciones con todas las variables X.
- Ejemplo: queremos determinar la nota media de la universidad (GPA) entre los atletas universitarios hombres y mujeres estimamos

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_4 tothrs + u,$$

donde *sat* es la nota del SAT, *hsperc* es el percentil de la clasificación a la que pertenece el instituto y *tothrs* es el número de horas de clase de las asignaturas universitarias.

Contraste de diferencias entre grupos

- **Modelo no restringido**: Modelo en el que el término constante y todas las pendientes difieren entre grupos poblacionales:

$$\begin{aligned} cumgpa &= \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female * sat \\ &+ \beta_2 hsperc + \delta_2 female * hsperc + \beta_3 tothrs \\ &+ \delta_3 female * tothrs + u. \end{aligned}$$

- **Hipótesis nula**:

$$H_0 : \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0.$$

Si uno de los δ_j es diferente de cero, el modelo es diferente para hombres y mujeres.

- **Modelo restringido**:

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u.$$

Contraste de diferencias entre grupos

- **Método**: contraste conjunto de la F comparando el modelo restringido y el no restringido basándonos en la Suma de Cuadrados de los Residuos (SCR).
- La SCR del modelo no restringido puede obtenerse de dos regresiones separadas ($SCR = SCR_1 + SCR_2$):
  - SCR_1: la resultante de estimar el modelo no restringido para el Grupo 1 (female).
  - SCR_2: la resultante de estimar el modelo no restringido para el Grupo 2 (male).

Contraste de diferencias entre grupos

- **Estadístico de Chow**:

$$F = \frac{[SCR_R - (SCR_1 + SCR_2)]}{SCR_1 + SCR_2} \times \frac{n - 2(k+1)}{k+1},$$

donde SCR_R es la SCR del modelo restringido.

- El test de Chow es simplemente un contraste de F para restricciones de exclusión. La diferencia es que ahora $SCR_{NR} = SCR_1 + SCR_2$.
- Este contraste tiene $k + 1$ restricciones (una por cada uno de los coeficientes de la pendiente y otra por el intercepto).
- El modelo no restringido estimaría 2 interceptos y 2 coeficientes de la pendiente distintos. De este modo, los grados de libertad son $n - 2k - 2$.

Modelo de probabilidad lineal

- En el modelo de probabilidad lineal, el valor predicho de Y es interpretado como la probabilidad predicha de que $Y = 1$

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, \cdots, N$$

donde β_1 es el cambio en la probabilidad predicha cuando X cambia en una unidad.

$$E(Y|X) = 1 \times Pr(Y = 1|X) + 0 \times Pr(Y = 0|X) = Pr(Y = 1|X).$$

- Bajo el supuesto $E(u_i|X_i)$,

$$\begin{aligned} E(Y_i|X_i) &= E(\beta_0 + \beta_1 X_i + u_i|X_i) = \beta_0 + \beta_1 X_i \\ Pr(Y_i = 1|X_i) &= \beta_0 + \beta_1 X_i. \end{aligned}$$

Modelo de probabilidad lineal

- Tratando de determinar la probabilidad de que un individuo sea arrestado obtenemos

$$\begin{aligned}\widehat{arr86} &= 0,441 - 0,162 pcnv + 0,0061 avgsen - 0,0023 tottime \\ &- 0,022 ptime86 - 0,043 qemp86\end{aligned}$$

donde *pcnv* es la proporción de condenas previas, *avgsen* la sentencia media en detenciones anteriores (meses), *tottime* son los meses pasados en prisión hasta 1986, *ptime*86 son los meses pasados en prisión en 1986 y *qemp*86 es el número de trimestres que tuvo un trabajo legal.

Modelo de probabilidad lineal

- Interpretación: ceteris paribus, 1 mes más en prisión reduce la probabilidad de detención en $0,022$ meses.
- Dado el caracter binario de Y, este modelo vulnera el supuesto RLM.5 de Gauss-Markov (homocedasticidad). En este caso,

$$Var(Y|X) = \rho(x)[1 - \rho(x)],$$

donde $\rho(x)$ es la probabilidad de éxito.