Análisis de Regresión Múltiple con Información Cualitativa: Variables Binarias o Ficticias

Carlos Velasco[1]

[1] Departamento de Economía
Universidad Carlos III de Madrid

Econometría I
Máster en Economía Industrial
Universidad Carlos III de Madrid
Curso 2007/08

Resumen

1. Cómo describir información cualitativa
2. Una variable ficticia independiente única
3. Cómo usar variables ficticias para categorías múltiples
4. Interacciones con variables ficticias

1. Cómo describir información cualitativa

- Los factores cualitativos aparecen a menudo bajo la forma de información binaria
  - un individuo es un hombre o mujer
  - un individuo posee o no un ordenador personal
  - una empresa ofrece o no un determinado plan de pensiones a los empleados
- La información se presenta por medio de una variable **binaria** o variable cero-uno.
- En Econometría a estas variables se las llama variables **ficticias**.

Definición de Variables ficticias

- Hay que decidir a qué acontecimiento se le asigna el valor uno y a cuál le corresponde el cero.
- El nombre de la variable indica a qué acontecimiento le corresponde el valor uno:
  - Si definimos *female* entonces tomaría el valor 1 para las mujeres y 0 para los hombres.
  - Si definimos *male* entonces tomaría el valor 1 para los hombres y 0 para las mujeres.
- Ambas variables contienen la misma información, y es mejor que definir la variable *gender*, porque es ambiguo su significado.

Ejemplo

- Si *married* indica si el individuo está casado, tenemos un ejemplo:
- Datos WAGE1:

person	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
⋮	⋮	⋮	⋮	⋮	⋮
525	11.56	16	5	0	1
526	3.50	14	5	1	0

2. Una variable ficticia independiente única

¿Cómo incorporar información cualitativa en los modelos de regresión?

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

- Usamos la notación δ_0 para indicar que es el parámetro de la variable ficticia *female*.
- Solo hay dos factores que afectan a los salarios: sexo y educación.
- δ_0 es la diferencia entre el salario por hora de una mujer y de un hombre, a un nivel dado de educación (y con el mismo término de error).
- δ_0 determina si hay **discriminación** en contra de las mujeres: si $\delta_0 < 0$, para el mismo nivel de los demás factores, las mujeres ganan menos en promedio que los hombres.

Interpretación coeficiente de la variable ficticia con Esperanzas condicionales

- Suponiendo que $E(u|female, educ) = 0$,

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

- Una formulación equivalente sería

$$\delta_0 = E(wage|mujer, educ) - E(wage|hombre, educ).$$

Interpretación gráfica del coeficiente de la variable ficticia

- δ_0 describe un cambio en el término constante entre hombres y mujeres:
  - Mujeres: $\beta_0 + \delta_0$.
  - Hombres: β_0.
- (Pero ambos tienen la misma pendiente, β_1).

¿Cuántas variables ficticias?

- Se necesita una única variable ficticia para obtener diferentes términos constantes.
- Si se incluyen dos variables ficticias tendríamos multicolinealidad perfecta porque

$$female + male = 1,$$

  es decir *male* es una función lineal perfecta de *female*.
- Esta es la denominada **trampa de las ficticias**, que aparece cuando se usan demasiadas variables para describir un número dado de grupos.

¿Qué variables ficticias?

- En el ejemplo los hombres son el **grupo base** o **grupo de referencia**, respecto al cuál se hacen las comparaciones.
- Este es el grupo cuyo término constante no depende de la variables ficticias.
- Podríamos haber tomado a las mujeres:

$$wage = \alpha_0 + \gamma_0 male + \beta_1 educ + u.$$

- O haber eliminado el término constante e incluir dos ficticias:

$$wage = \beta_0 male + \alpha_0 female + \beta_1 educ + u$$

que no es interesante porque es más complicado contrastar si las ordenadas en el origen son iguales.

Otras variables

Si hay más variables explicativas,

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

la interpretación de δ_0 no cambia.

Contraste de discriminación en contra de las mujeres: test de la t sobre

$$H_0 : \delta_0 = 1 \quad \text{en contra de} \quad H_1 : \delta_0 < 0.$$

EJEMPLO: WAGE1 (Ecuación de salario por hora).

Modelos sólo con variables ficticias

EJEMPLO: WAGE1

Si se estima el modelo sólo con la variable *female* :

$$\widehat{wage} = \underset{(0,21)}{7,10} - \underset{(0,30)}{2,51}\, female$$

- El término constante 7.10 es el salario medio para los hombres (grupo de referencia).
- El coeficiente de la variable ficticia es la diferencia entre salarios medios de hombres y mujeres en la muestra: las mujeres ganan 2.51$ menos por hora.
- **Contraste de comparación de medias** entre los dos grupos: estadístico t del coeficiente de la v. ficticia:

$$t = \frac{-2,51}{0,30} = -8,37$$

  que es estadísticamente muy significativo.
- Necesita el supuesto de homoscedasticidad (igual variabilidad de los salarios de hombres y mujeres).

Evaluación de Programas

- Se quiere conocer el efecto de programas sociales y económicos sobre los individuos, empresas, etc.
- En el caso más simple hay dos grupos:
  - El **grupo de control**, que no participa en el programa.
  - El **grupo experimental** o grupo de tratamiento, que sí que toma parte en el programa.
- Generalmente la selección de los grupos no se hace al azar como en ciencias experimentales.
- En ese caso se incluye regresores adicionales para controlar el efecto de otros factores.
- EJEMPLO: Efectos de la subvenciones de formación sobre las horas de formación JTRAIN:

$$hrsemp = \beta_0 + \beta_1 grant + \beta_2 \log(sales) + \beta_3 \log(employ) + u$$

donde $grant = 1$ si la empresa recibió subvención en 1988. ¿Es el efecto medido causal?

Interpretación de los coeficientes de las variables ficticias explicativas si la var. dependiente es log(y)

- EJEMPLO: HPRICE1

$$\begin{aligned}\widehat{\log(price)} &= \underset{(0,65)}{5,56} + \underset{(0,038)}{0,168}\log(lotsize) + \underset{(0,093)}{0,707}\log(sqrft) \\ &\quad + \underset{(0,029)}{0,027}bdrms + \underset{(0,045)}{0,054}colonial \\ n &= 88, \quad R^2 = 0,649\end{aligned}$$

donde *colonial* es una variable ficticia que vale 1 si la casa es de estilo colonial.

- Para niveles de *lotsize*, *sqrft* y *bdrms*, la diferencia en log (*price*) entre una casa de estilo colonial, y otra que no lo es de 0.054: una casa de estilo colonial será un 5.4 % más cara.
- Si la diferencia porcentual es grande, el **diferencial exacto** es:

$$100\left\{\exp\left(\hat{\beta}_j\right) - 1\right\}\ \%$$

por lo que $100\{\exp(0,054) - 1\} \approx 5,6\,\%$.

3. Cómo usar variables ficticias para categorías múltiples

En la ecuación

$$\log(\textit{wage}) = \beta_0 + \delta_0 \textit{female} + \beta_1 \textit{educ} + \beta_2 \textit{exper} + \beta_3 \textit{tenure} + u$$

podemos incluir la variable ficticia *married*.

El coeficiente de *married* proporciona el diferencial salarial porcentual aproximado entre los que están casados y los que no lo están, manteniendo fijos el sexo, *educ*, *exper* y *tenure*.

EJEMPLO: ecuación del logaritmo del salario

Queremos distinguir diferencias salariales entre cuatro grupos: hombres casados, mujeres casadas, hombres solteros y mujeres solteras

■ Seleccionamos un grupo de referencia: hombres solteros.

■ Debemos definir variables ficticias para cada uno de los grupos restantes (y eliminar *female*):

- *marrmale* : hombres casados.
- *marrfe* : mujeres casadas.
- *singfem*: mujeres solteras.

$$\begin{aligned}\log(wage) = {} & \beta_0 + \delta_0 marrmale + \delta_1 marrfem + \delta_2 singfem \\ & + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 \\ & + \beta_4 tenure + \beta_5 tenure^2 + u\end{aligned}$$

Cálculo de la diferencia entre grupos

- Diferencia mujeres solteras y casadas: $\delta_2 - \delta_1$.
- Pero no podemos contrastar directamente si esta diferencia es significativa.
- Habría que reestimar la ecuación tomando uno de estos grupos como grupo de referencia (por ejemplo mujeres casadas):

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

  donde ahora $H_0 : \delta_2 = 0$.
- **Regla general**: para distinguir entre g grupos hay que incluir $g - 1$ variables ficticias.

Incorporar información ordinal con variables ficticias

Variables ordinales: distinguen diferentes categorías de acuerdo a un determinado criterio.
EJEMPLO: clasificación de la calidad del endeudamiento por parte de las agencias financieras (Moody y SP).
Si *CR* es la variable que recoge la clasificación (entre 0 y 4) , ¿cómo incorporar la información de *CR* en un modelo para explicar el tipo de interés de los bonos municipales, *MBR*?

Primera posibilidad:

$$MBR = \beta_0 + \beta_1 CR + otros \;\; factores.$$

β_1 : es el cambio en puntos porcentuales de *MBR* cuando *CR* se incrementa en una unidad, otros factores constantes.
■ Pero *no* está claro el significado de un incremento en una unidad de *CR* : ¿es igual la diferencia entre 4 y 3 que la hay entre 2 y 3?

Otro enfoque

Sustituir *CR* or un número reducido de variables cualitativas:

- $CR_1 = 1$ si $CR = 1$; $CR_1 = 0$ si no.
- $CR_2 = 1$ si $CR = 2$; $CR_2 = 0$ si no.
- $CR_3 = 1$ si $CR = 3$; $CR_3 = 0$ si no.
- $CR_4 = 1$ si $CR = 4$; $CR_4 = 0$ si no.

En este caso $CR = 0$ es el grupo de referencia y sólo incluimos 4 v. ficticias para 5 grupos:

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + otros\ factores.$$

δ_j : es la diferencia en *MBR* (otros factores fijos) entre una municipalidad de crédito j y otra de crédito $j - 1$.

Contraste de un efecto parcial constante

Restricciones a contrastar:

$$\delta_2 = 2\delta_1, \quad \delta_3 = 3\delta_1, \quad \delta_4 = 4\delta_1$$

Modelo restringido:

$$\begin{aligned} MBR &= \beta_0 + \delta_1 (CR_1 + CR_2 + CR_3 + CR_4) + otros\ factores \\ &= \beta_0 + \delta_1 CR + otros\ factores. \end{aligned}$$

que es el modelo con la variable original de crédito.

Solución: contraste de la F con $q = 3$.

4. Interacciones en las que intervienen variables ficticias

Las variables ficticias puede interactuar igual que lo hacen las variables con significado cuantitativo.
EJEMPLO: Ecuación para el logaritmo del salario, WAGE1.
El modelo se puede reformular interactuando las v. ficticias *female* and *married* :

$$\begin{aligned} \log(\mathit{wage}) &= \beta_0 + \gamma_0 \mathit{female} + \gamma_1 \mathit{married} \\ &\quad + \gamma_2 \mathit{female} * \mathit{married} + \cdots \end{aligned}$$

indicando que la prima por estar casado depende del sexo (o al revés).

- Esta formulación es interesante porque permite contrastar directamente la significatividad de la interacción entre sexo y estado civil.
- Hombres solteros sigue siendo el grupo de referencia, por lo que la ecuación es adecuada para contrastar diferenciales entre ese grupo y cualquier otro.

Cómo permitir pendientes distintas

- Las variables ficticias también se usan para permitir ecuaciones con diferentes pendientes para cada uno de los grupos.
- EJEMPLO: Ecuación de salarios: se quiere 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 entre hombres y entre mujeres):

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

- Ecuación para hombres:

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

- Ecuación para mujeres:

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

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

¿Cómo estimar el modelo con diferentes pendientes?

Podemos escribir

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

donde los regresores son el término constante, *female*, *educ* y la

interacción $female * educ$ (que es igual a cero para todos los hombres e igual a *educ* para todas las mujeres).
Hipótesis: la rentabilidad de la educación es la misma para las mujeres y los hombres:

$$H_0 : \delta_1 = 0,$$

que indica que la pendiente de log(*wage*) respecto a *educ* es igual para hombres que para mujeres.
Esta hipótesis no impone restricciones sobre los términos constantes de ambos grupos (δ_0) : puede existir diferencia salarial entre ambos grupos, pero ha de ser independiente del nivel de educación.

Ejemplo: ecuación del logaritmo del salario por hora, WAGE1

$$\begin{aligned}\log(\mathit{wage}) &= \beta_0 + \delta_0 \mathit{female} + \beta_1 \mathit{educ} + \delta_1 \mathit{female} * \mathit{educ} \\ &\quad + \beta_2 \mathit{exper} + \beta_3 \mathit{exper}^2 \\ &\quad + \beta_4 \mathit{tenure} + \beta_5 \mathit{tenure}^2 + u\end{aligned}$$

- Para contrastar si hay **diferencias salariales** no hay que contrastar individualmente la significatividad de δ_0 y δ_1, si no hacer un contraste conjunto, ya que *female* y *female* $*$ *educ* estarán muy correlacionadas y, por ejemplo, efecto de *female* se estimará con mucha menos precisión que antes.
- Además δ_0 mide el efecto diferencial de la educación entre hombres y mujeres cuando *educ* $= 0$ que es un caso poco interesante (y no existente en la muestra): sería mejor sustituir otro valor para *educ*, como la media

Contraste de diferencias entre grupos en funciones de regresión

Hipótesis: dos poblaciones o dos grupos siguen la misma función de regresión, contra la alternativa de que una o más pendientes difieren entre grupos.
EJEMPLO: Modelo para describir la nota media entre la universidad (*GPA*) de los atletas universitarios masculinos y femeninos:

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

donde *stat* es el resultado en el SAT, *hsperc* es el percentil de la clasificación a la que pertenece el instituto y *tothrs* es el número total de horas de clase de las asignaturas universitarias.
■ Modelo en el que el término constante y todas las pendientes puedan diferir de un grupo a otro:

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

Contraste de diferencias entre grupos en funciones de regresión

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.

Método: contraste conjunto de la F, comparando el modelo restringido bajo H_0 (sin la variable *female*) y el no restringido.

- Precaución para interpretar los coeficientes del modelo no restringido: δ_0 es la diferencia entre hombres y mujeres cuando $sat = hsperc = tothrs = 0$.
- El estadístico F también se puede calcular haciendo regresiones separadas (Contraste de Chow.)