Información cualitativa y modelos no lineales

Gabriel V. Montes-Rojas

Variable binaria o dummy

Un factor cualitativo (vs. uno cuantitativo) es un factor cuya información tiene que ser codificada en forma numérica.

Definición: Una variable que toma valores **0** *y* **1** *se define como VARIABLE DUMMY. La categoría que tiene valor 0 se llama CATEGORIA BASE (por ejemplo masculino para sexo).*

- Ej. *Sexo. female* es una variable binaria que tiene 1 si sexo femenino, 0 si sexo masculino. No importa cual es 1 o 0, lo importante es que distinga.
- Ej. *Estado civil.* Para categorizar estado civil se puede necesitar más de dos valores. 0 soltera/o, 1 casada/o, 2 divorciada/o, 3 viuda/o.
- Ej. *Nacionalidad.* Para categorizar la nacionalidad se necesita una variable que tome más de dos valores. 0 Argentina, 1 Uruguay, 2 Brasil, 3 Paraguay, 4 Chile, 5 otros.

En los dos últimos casos más de una dummy. Como regla, si hay Q categorías necesitamos $Q-1$ dummies. (ver más abajo)

Variable binaria o dummy

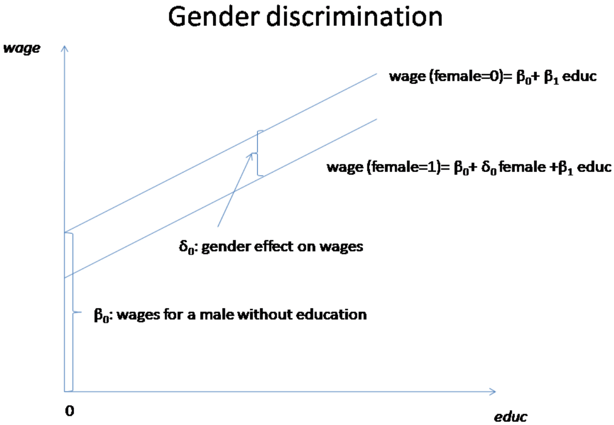
Consideremos el modelo:

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

En este caso *female* no es una variable contínua, pero δ_0 tiene la misma interpretación que otros coeficientes. En particular, cuál es el cambio en *wage* cuando la variable *female* se incrementa *ceteris paribus* una unidad. En el caso particular de las dummies se obtiene:

$$\delta_0 = E(wage|female = 1, educ) - E(wage|female = 0, educ)$$

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



Variable binaria o dummy

Ejercicio: Probar que el modelo

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

y

$$wage = \beta_0' + \alpha_0 male + \beta_1' educ + e$$

donde $male = 1 - female$, cumplen las relaciones $\beta_0' + \alpha_0 = \beta_0, \beta_0 + \delta_0 = \beta_0', \beta_1 = \beta_1'$.

Esto significa que la selección de la categoría base no tiene ningún efecto sobre los resultados. Sólo para el intercepto.

Pregunta: ¿Cuál es el problema con este modelo?

$$wage = \beta_0 + \delta_0 female + \alpha_0 male + \beta_1 educ + u$$

Efectos individuales y compuestos: interacciones

Las variables dummy pueden ser combinadas para efectos compuestos.

Supongamos que D_1 and D_2 son dos variables dummy. Definamos la interación como $D_1 \times D_2$. Consideremos el modelo

$$y = \alpha + \gamma D_1 + \delta D_2 + \phi(D_1 \times D_2) + u$$

¿Cómo se interpretan $\alpha, \gamma, \delta, \phi$? Notar que $E[y|D_1 = 0, D_2 = 0] = \alpha$,
$E[y|D_1 = 1, D_2 = 0] = \alpha + \gamma$, $E[y|D_1 = 0, D_2 = 1] = \alpha + \delta$,
$E[y|D_1 = 1, D_2 = 1] = \alpha + \gamma + \delta + \phi$.

Supongamos que queremos contrastar si una variable continua, X, tiene distintas pendientes en distintos grupos, dados por la variable dummy D.

$$y = \alpha + \gamma D + \beta X + \delta(D \times X) + u$$

$(D \times X)$ es la interacción.

STATA: dummies

Una variable dummy se implementa como cualquier otra variable independiente. Supongamos que queremos ver el efecto de la variable z, que tiene categorías múltiples. $Z \in 0, 1, 2, ..., J$

- Para ver la distribución de z en la muestra:
  `tab z`
- Para ver los valores de y para distintos z en la muestra:
  `tab z, summ(y)`
- Para ver un histograma de z:
  `hist z`
- En forma general, si tenemos más de dos categorías, ej. Q, necesitamos $Q - 1$. Esto se implementa automáticamente en STATA
  `xi: reg y i.z x1 x2 x3`
  Nota: Por default, STATA omite el valor de z del primer grupo. Pero esto se puede cambiar (por ej. z=2)
  `char z[omit] 2`
- Más detalles:
  `http://www.stata.com/help.cgi?xi`

Modelos cuadráticos

Consideremos el siguiente modelo:

$$wage = \beta_0 + \beta_1 exper + \beta_2 exper^2 + u$$

En este caso,

$$\frac{\partial E(wage|exper)}{\partial exper} = \beta_1 + 2\beta_2 exper$$

En palabras, el efecto de *exper* sobre *wage* no es lineal, y el efecto lineal (pendiente) depende de los valores de *exper*.

Logaritmos

Consideremos el siguiente modelo log-lineal:

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

Resultado: $\frac{d \log wage}{d\ educ} = \frac{\frac{d\ wage}{wage}}{d\ educ}$

En general funciona la siguiente aproximación: $\frac{d\ wage}{wage} \approx \frac{\Delta\ wage}{wage} \approx$ % cambio en *wage*

β_1: Es el cambio porcentual en *wage* ante un cambio de una unidad en *educ*.

$$\begin{array}{rcc} lwage = & .584^{***}+ & .083^{***} educ \\ & (.097) & (.0076) \\ & < 0.000 > & < 0.000 > \\ & [6.0] & [10.9] \end{array}$$

(error estándar); $< p - valor >$; $[t - valor]$; * significancia 10%; ** significancia 5%; *** significancia 1%

Logaritmos

Sin embargo, la aproximación sólo funciona para pequeños cambios en la variable independiente. El cálculo exacto es

$$\%\hat{\Delta} y = 100[exp(\hat{\beta}_1 \Delta x) - 1]$$

$$exp(.083) - 1 = .087 \neq .083$$

Logaritmos

Ahora consideremos el modelo log-log:

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

Pregunta: ¿Qué significa β_1 en este modelo?

Ejemplos

http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge7.html

STATA: modelos de variables no lineales

- Para implementar modelos cuadráticos se debe crear el cuadrado de la variable. Por ejemplo,
  ```
  gen exper2=exper*exper
  reg wage educ exper exper2
  ```
- Para implementar logaritmos se debe transformar la variable en log. Por ejemplo,

  ```
  gen lwage=ln(wage)
  reg lwage educ
  gen leduc=ln(educ)
  reg wage leduc
  reg lwage leduc
  ```

Cambio en las unidades de medida

Es muy importante saber cuales son la unidades de medida en X e Y para interpretar correctamente los parámetros estimados.

Ej.: Supongamos que `educ` se mide en meses, en vez de años como teníamos antes. Definamos `educm=12*educ`

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

$$wage = \gamma_0 + \gamma_1 educm + u$$

- ¿Cómo se comparan β y γ?
- $\beta_1 = \frac{\Delta wage}{\Delta educ} = \frac{\Delta wage}{\Delta educ * \frac{12}{12}}$.
- $\beta_1 / 12 = \frac{\Delta wage}{\Delta educ * 12} = \frac{\Delta wage}{\Delta educm} = \delta_1$.

Cambio en las unidades de medida

Ej.: supongamos que `wage` se mide en centavos. Definamos `wagec=wage*100`

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

$$wagec = \gamma_0 + \gamma_1 educ + u$$

- ¿Cómo se comparan β y γ?
- $wage * 100 = \beta_0 * 100 + \beta_1 educ * 100 = wagec = \gamma_0 + \gamma_1 * educ$ entonces $\beta_1 = \gamma_1 / 100$