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.*

- 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:

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

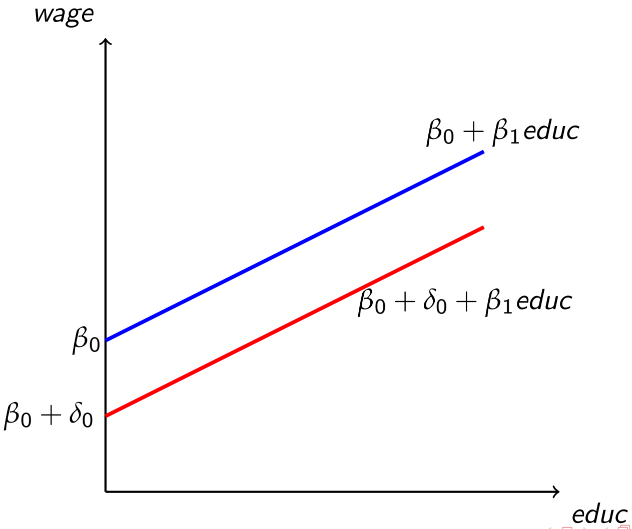
En este caso *female* no es una variable contínua, pero δ_0 tiene la misma interpretación que otros coeficientes. En particular, cuál es el cambio en *wage* cuando la variable *female* se incrementa *ceteris paribus* una unidad. En el caso particular de las dummies se obtiene:

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

Notar que este efecto es independiente del nivel de *educ*, es decir, se mantiene para todo nivel de educación.

Discriminación por género

$wage = \beta_0 + \delta_0 female + \beta_1 educ + u, \;\; \delta_0 < 0$



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 que reflejan dos categorías binarias diferentes (ejemplo género y casado). Definamos la interacción 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 $E[y|d_1 = 0, d_2 = 0] = \alpha$, $E[y|d_1 = 1, d_2 = 0] = \alpha + \gamma$, $E[y|d_1 = 0, d_2 = 1] = \alpha + \delta$, $E[y|d_1 = 1, d_2 = 1] = \alpha + \gamma + \delta + \phi$.
- Notar que ϕ se puede interpretar como estimador de diferencias en diferencias (d-en-d):
  $\phi = \{E[y|d_1 = 1, d_2 = 1] - E[y|d_1 = 0, d_2 = 1]\} - \{E[y|d_1 = 1, d_2 = 0] - E[y|d_1 = 0, d_2 = 0]\}$.

Estimador de diferencia en diferencias (d-en-d)

Esta metodología es muy común para establecer relaciones causales. El ejemplo típico es evaluar un cambio de política.

Tratamiento: Lo que queremos evaluar.
Grupo de control: Grupo de individuos que no tienen el tratamiento.
Grupo de tratamiento: Grupo de individuos que se ve afectado por el tratamiento.

$$y_{it} = \beta_0 + \beta_1 TREAT_i + \delta_0 AFTER_t + \delta_1 TREAT_i \times AFTER_t + u_{it}$$

- Consideremos una muestra de $i = 1, ..., N$ individuos donde cada individuo se lo ve en periodos $t = 1, ..., T$.
- $TREAT_i$ es una dummy que indica si el individuo i está en el grupo de tratamiento (o no). Tratamiento vs. control.
- $AFTER_t$ es una dummy que indica si el periodo t ya tiene el tratamiento aplicado (o no). Antes vs. después.

	Antes	Después	Después-Antes
Control	β_0	$\beta_0 + \delta_0$	δ_0
Tratamiento	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Tratamiento-Control	β_1	$\beta_1 + \delta_1$	δ_1

δ_1 es el estimador **d-en-d**.

Efecto de la ubicación de un incinerador de basura en los precios de las casas

- Ejemplo tomado de Kiel y McClain (1995) sobre el efecto de un incinerador de basura en los precios de las casas en North Andover, Massachusetts, para los años 1978 y 1981.
- La base de datos se puede acceder en:
  `use http://fmwww.bc.edu/ec-p/data/wooldridge/KIELMC, clear`
- Las variables son
  rprice: precio de las casas en términos reales.
  nearinc: variable dummy para cercanía al incinerador.
- Los rumores de que el incinerador se iba a construir empezaron **después** de 1978, su construcción empezó en 1981, y empezó a operar en 1985.
- Un estimador simple usaría sólo datos de 1981:

$$
\begin{aligned}
rprice &= \gamma_0 + \gamma_1\, nearinc + u \\
\widehat{rprice} &= \underset{(3093)}{101308} \; \underset{(5828)}{-30688}\, nearinc
\end{aligned}
$$

- ¿Es correcto afirmar que vivir cerca del incinerador causa que los precios disminuyan 30 mil dólares?

Efecto de la ubicación de un incinerador de basura en los precios de las casas

- El estimador previo no implica que el incinerador **cause** una disminución del precio de las casas. De hecho la misma regresión para 1978 da:

$$\begin{array}{rcc} \widehat{rprice} = & 82517 & -18824\ nearinc \\ & (2654) & (4744) \end{array}$$

- ¿Por qué? Porque el incinerador se construyó donde el precio de las casas ya era bajo.
- **Estimador d-en-d**:

$$\hat{\delta}_1 = (\widehat{rprice}_{81,nr} - \widehat{rprice}_{81,fr}) - (\widehat{rprice}_{78,nr} - \widehat{rprice}_{78,fr})$$

donde *nr* significa cerca del incinerador y *fr* sin (free) incinerador.

- Usando el modelo general:

$$\begin{array}{lllllll} \text{rprice} & = \beta_0 & + \beta_1 & \text{nearinc} + \delta_0 & \text{y81} & + \delta_1 & \text{nearinc*y81} + \text{u} \\ \widehat{\text{rprice}} & = 82517 & - 18824 & \text{nearinc} + 18790 & \text{y81} & - 11864 & \\ & (2727) & (4875) & (4050) & & (7466) & \end{array}$$

Card y Krueger (1994)

Estudio de cambios en el salario mínimo sobre empleo

David Card y Alan B. Krueger (1994), "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey," American Economic Review 84(4), 772-793.
Ver una discusión avanzada en
http://econ.lse.ac.uk/staff/spischke/ec524/evaluation3.pdf

Card y Krueger (1994)

- Datos de EEUU. Hay salario mínimo a nivel federal y a nivel de cada estado. Aplica el mayor.
- En noviembre de 1989 se aprobó una ley para incrementar el salario mínimo a nivel federal de $3.35 por hora a $3.80 (en abril 1990), $4.25 por hora (en abril 1991), y al principio de 1990 se decidió incrementarlo a $5.05 por hora (en abril de 1992).
- Sin embargo había diferencias a nivel estadual en cuanto al salario mínimo vigente (Pennsylvania y New Jersey)
- ¿Qué tipo de firmas iban a ser más afectadas? Aquellas que usaban trabajo no calificado (*unskilled* labor).
- Entonces estudiaron: Burger King, KFC, Wendy's, Roy Rogers
    1. son los empleadores líderes para trabajadores de bajo salario
    2. cumplen con la regulación de salarios mínimos
    3. los trabajos son relativamente homogéneos
    4. es relativamente fácil construir una muestra de franquicias de esas cadenas

Card y Krueger (1994)

- **Tratamiento**: incremento del salario mínimo efectivo
- **Grupo de control**: restaurantes en Pennsylvania
- **Grupo de tratamiento**: restaurantes en New Jersey
- Mapa of New Jersey
- **Antes**: Wave 1, February 15 - March 4, 1992
- **Después**: Wave 2, November 5 - December 31, 1992
- Efecto sobre el empleo: Los autores no encuentran ningún efecto significativo sobre el empleo.
- Impacto sobre otros beneficios no salariales: Podría ser que no hay efecto sobre el empleo porque los restaurantes reducen otros beneficios, o reducen training y las promociones. No hay evidencia.
- Impacto sobre precios: En un modelo competitivo incrementar el salario implica incrementar los precios. Tampoco hay evidencia de esto.

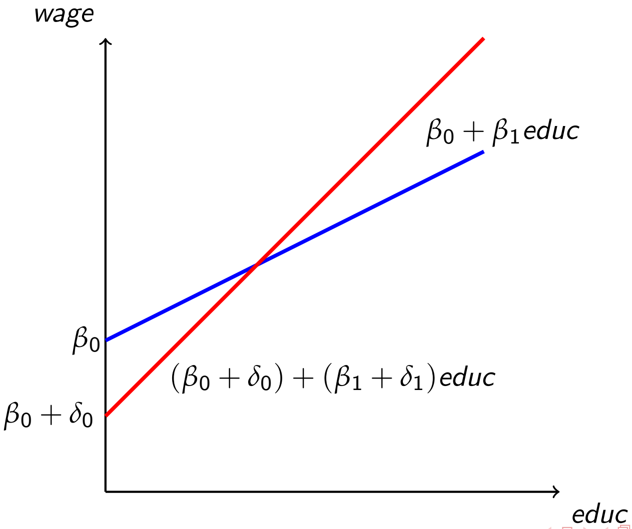
Efectos individuales y compuestos: interacciones

- 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.
- Este modelo permite dos pendientes de acuerdo a la clasificación de d, β y $\beta + \delta$.
- Notar que $E[y|d=0] = \alpha + \beta E[x|d=0]$ y que $E[y|d=1] = \alpha + \gamma + (\beta + \delta)E[x|d=1]$. También, $E[y|d=0,x] = \alpha + \beta x$ y $E[y|d=1,x] = \alpha + \gamma + (\beta + \delta)x$. ¿Cuál es la diferencia entre estos términos?

Discriminación por género $wage = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \times educ + u, \ \delta_0 < 0, \delta_1 > 0$



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:

$$\log \textit{wage} = \beta_0 + \beta_1 \log \textit{educ} + u$$

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?
- $\beta_1 = \frac{\Delta wage}{\Delta educ} = \frac{\Delta wage}{\Delta educ * \frac{12}{12}}$.
- $\beta_1 / 12 = \frac{\Delta wage}{\Delta educ * 12} = \frac{\Delta wage}{\Delta educm} = \gamma_1$.
- ¿Qué pasa con las constantes?

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 $\beta_1 = \gamma_1/100$