Tópicos de Economía Aplicada

Introducción II

La clase de hoy

- Una pregunta empírica
- Descripción de datos: una variable
- Descripción de datos: dos variables
- Regresión lineal
- Estimación causal

Además de introducir estos temas, comenzaremos a usar STATA y la encuesta EPH (ver "Materiales para clase" en el campus virtual, bajar clases-herramientas.zip y descomprimir en una carpeta; hoy usaremos parte del código 1EPH.do)

Una pregunta empírica

Análisis empírico

- Pregunta empírica:
- ¿Aumenta el ingreso con un título universitario? La pregunta empírica es también ¿en cuánto?
- Una variable relevante de resultado, y, es el salario o el ingreso. Una característica es el nivel educativo, x_1, pero también otra serie de características del individuo que pueden afectar al ingreso más allá de la educación, $x_2, x_3, \ldots$ que pueden ser observables, y una serie de inobservables que se resumen en una variable, u.
- Descripción: Una primera forma de explorar esta pregunta es hacer una descripción de estas variables y de sus relaciones.
- (Más adelante veremos un modelo de la decisión de estudiar y de los efectos de la educación)

Descripción de una variable

Análisis empírico - Descripción de los datos

Supongamos que x es una variable aleatoria. El valor x_j es una realización de x.

- La distribución de esta variable está caracterizada por la función de distribución acumulada

$$F(x_0) = \Pr(x \leq x_0)$$

que mide la probabilidad de que la realización de x sea menor o igual al valor x_0.

- Si la variable aleatoria es discreta, esta función de distribución se puede describir con la función de probabilidad

$$f_x(x_0) = \Pr(x = x_0).$$

- Si la variable aleatoria es continua entonces

$$F_x(b) - F_x(a) = \int_a^b f_x(x)\, dx$$

y en los puntos derivables, la derivada de la distribución acumulada es la densidad

$$f_x(x) = \frac{dF(x)}{dx}$$

Análisis empírico - Descripción de los datos: muestra

Apliquemos estos conceptos a una muestra. EPH, ingresos individuales (en logaritmo), nivel educativo. En STATA abrimos la base [link]:

```
. use Individual_t414.dta, clear
```

```
. brow CODUSU nro_hogar componente pondera ch03 ch06 nivel_ed p21 p47t
```

Editor de Datos (Navegación) - [Individual_t414]

Archivo Edición Ver Datos Herramientas

nivel_ed[11] 4

	CODUSU	nro_hogar	componente	pondera	ch03	ch06	nivel_ed	p21	p47t
1	380032	1	3	120	Hi...	20	S...	1,500	1,500
2	380032	1	1	120	Je...	48	S...	7,600	17,600
3	380032	1	2	120	CD...	45	S...	4,600	9,200
4	380034	1	6	435	Hi...	29	S...	0	0
5	380034	1	2	435	CD...	59	S...	4,500	8,000
6	380034	1	1	435	Je...	58	S...	500	500
7	380038	1	3	669	Hi...	25	S...	0	0
8	380038	1	2	669	CD...	51	S...	8,000	8,000
9	380038	1	4	669	Hi...	23	S...	0	0
10	380038	1	1	669	Je...	57	S...	6,000	6,000
11	380040	1	2	405	Hi...	23	S...	6,000	6,000
12	380040	1	1	405	Je...	52	P...	7,500	11,100
13	380040	1	3	405	Hi...	22	P...	0	0

Variables

Filtrar variables aquí

Nombre	Etiqueta
CODUSU	C◆digo para disti
nro_hogar	C◆digo para disti
componente	N◆mero de comp
pondera	Ponderaci◆n
ch03	Relaci◆n de pare
ch06	Edad en a◆os cur
nivel_ed	Nivel educativo
p21	Monto de ingreso

Variables Copia temporal

Propiedades

Variables

Variables: 9 de 173 Orden: Dataset Obs: 32.362 Filtro: Apagado Modo: Navegación CAP NUM

Análisis empírico - Descripción de los datos: muestra

- La variable nivel educativo es discreta. Podemos describir su distribución en una tabla. En STATA es tabular los datos (como es una encuesta con ponderadores los incluimos como fw, frequency weights)

```
. tab nivel_ed [fw=pondera]
```

- Un histograma es una representación gráfica de esta distribución

```
. hist nivel_ed [fw=pondera], discrete width(1)
start(1)
```

- La variable logaritmo del ingreso es continua, se representa graficando percentiles. Genera la variable y calcula percentiles y grafica:

```
. gen ling = log(p47t)
. xtile px_ling = ling, n(100)
. twoway scatter px_ling ling
```

Análisis empírico - Descripción de los datos

- La esperanza (o media) de una variable discreta es

$$\mathbb{E}(x) = \sum_j x_j f_X(x_j)$$

donde $f_X(x_j)$ es la función de probabilidad.

- La esperanza de una variable continua es

$$\mathbb{E}(x) = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

o también puede escribirse

$$\mathbb{E}(x) = \int_{-\infty}^{\infty} x dF_X(x)$$

- La esperanza es la mejor predicción de x (minimiza el error cuadrático medio)
- En la muestra se computa un promedio

$$\bar{x} = \frac{1}{n} \sum_i x_i$$

- La EPH tiene ponderadores; es necesario ponderar cada observación, en este caso con pw, probability weights (después veremos por qué).

```
. mean p47 [pw=pondera]
```

Análisis empírico - Descripción de los datos

- La varianza de una variable discreta es

$$Var\left(x\right) = \sum_{j} \left(x_j - \mathbb{E}\left(x\right)\right)^2 f_X\left(x_j\right)$$

- La varianza de una variable continua es

$$Var\left(x\right) = \int_{-\infty}^{\infty} \left(x_j - \mathbb{E}\left(x\right)\right)^2 f_X\left(x_j\right) dx$$

- El desvío es

$$\sigma\left(x\right) = \sqrt{Var\left(x\right)}$$

- Skewness (asimetría):

$$S\left(x\right) = \frac{1}{\sigma\left(x\right)^3} \sum_{j} \left(x_j - \mathbb{E}\left(x\right)\right)^3 f_X\left(x_j\right)$$

- Curtosis (forma):

$$K\left(x\right) = \frac{1}{\sigma\left(x\right)^4} \sum_{j} \left(x_j - \mathbb{E}\left(x\right)\right)^4 f_X\left(x_j\right)$$

En STATA:

```
. sum p47t [fw=pondera], det
```

Descripción de dos variables

Análisis empírico - Descripción de los datos - Dos variables

- Si x e y son dos variables aleatorias no sólo se puede caracterizar cada una sino también su distribución conjunta.
- Un indicador relevante aquí es la covarianza:

$$\begin{aligned} Cov(x,y) &= \mathbb{E}\left((x-\mathbb{E}(x))(y-\mathbb{E}(y))\right) \\ &= \mathbb{E}(xy)-\mathbb{E}(x)\mathbb{E}(y) \end{aligned}$$

- En una muestra

$$c_{xy} = \frac{1}{n}\sum_i (x_i-\bar{x})(y_i-\bar{y})$$

- La correlación es

$$Corr(x,y) = \frac{Cov(x,y)}{\sigma(x)\sigma(y)}$$

- En STATA

```
. correlate ling aniosedu [fw=pondera]
```

- Una propiedad importante

$$Var(a+bx+cy) = b^2 Var(x) + c^2 Var(y) + 2bcCov(x,y)$$

Análisis empírico - Descripción de los datos - Esperanza condicional

- Cuando tenemos dos variables también podemos computar una media condicional (la mejor predicción de y si conocemos x)
- Supongamos x es discreta:

$$\mathbb{E}\left(y|x=x_j\right)=\int_{-\infty}^{\infty} y f_{Y|X}\left(y|x_j\right)dy$$

donde $f_{Y|X}\left(y|x_j\right)=\frac{f_{X,Y}\left(x_j,y\right)}{f_X\left(x_j\right)}$ es la distribución de y condicionada en $x=x_j$. Importante: $\int_{-\infty}^{\infty} f_{Y|X}\left(y|x_j\right)=1$.

- En una muestra es el promedio pero restringiendo la muestra a todas las observaciones con $x=x_j$
- Por ejemplo, el log del ingreso promedio de los graduados universitarios

```
. sum ling if nivel_ed==6 [fw=pondera]
```

- La distribución condicionada es la distribución del ingreso por educación (para los graduados universitarios)
- Cuando la variable x es continua el concepto es idéntico pero en la implementación en una muestra se usan otros métodos (que veremos más adelante)

Análisis empírico - Descripción de los datos - Esperanza condicional

- La esperanza condicionada puede verse como una función: Conditional Expectation Function (CEF) o función de esperanza condicionada

$$\mu(x) = \mathbb{E}(y|x)$$

- Una de sus propiedades útiles es la ley de esperanzas iteradas:

$$\mathbb{E}(y) = \mathbb{E}(\mathbb{E}(y|x))$$

donde

$$\mathbb{E}(\mathbb{E}(y|x)) = \int \mathbb{E}(y|x) f_x(x) dx$$

- En palabras para una muestra: el promedio total es igual que un promedio (ponderado) del promedio para cada grupo (ejemplo ingreso por grupo)

Análisis empírico - Descripción de los datos - Esperanza condicional

- Tabla del promedio condicionado por nivel educativo.
- El promedio ponderado de los valores de la tabla es el promedio total: ($\bar{y} = \frac{\sum_j \bar{y}_j w_j}{\sum_j w_j}$, donde j es el nivel educativo, $\bar{y}$ es un promedio y w_j es la cantidad de observaciones de cada grupo, columna (2))

nivel_ed	(1) p47t	(2) obs
-	2,234	43778
Primaria Completa	3,773	2558680
Secundaria Incompleta	3,914	2484380
Secundaria Completa	5,060	3660164
Superior Universitaria Incompleta	4,647	2253661
Superior Universitaria Completa	9,220	2578232

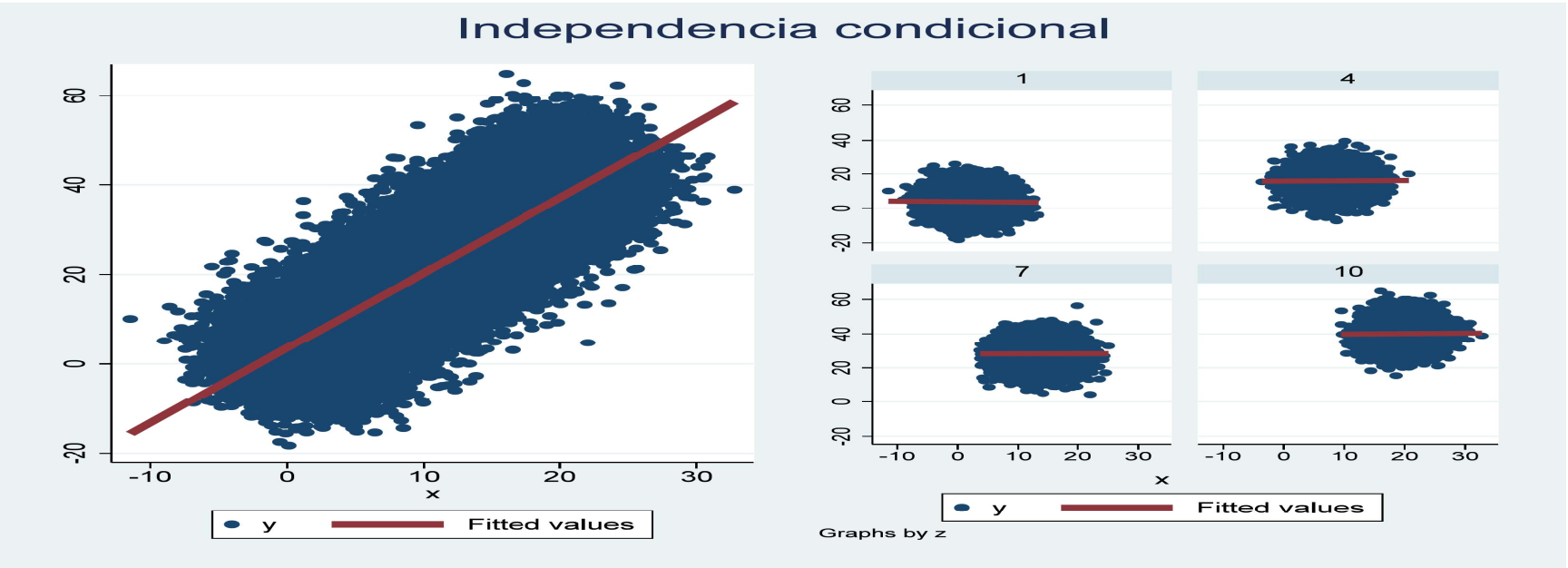
```
. sum p47t [fw=obs]
 Var. | Obs       Mean     Std. Dev Min     Max
 p47t | 14171408 5234.901 1957.489 2233.73 9220.45
```

Análisis empírico - Descripción de los datos - Esperanza condicional

- Independencia: Dos variables y, x son independientes en media si
$$\mathbb{E}(y) = \mathbb{E}(y|x)$$
- Dos variables independientes en media serán incorrelacionadas $cov(y, x) = 0$.
- Llamamos ortogonales, $y \perp x$, a dos variables independientes en media
- Independencia condicional, $y \perp x|z$. Ejemplo: salarios y conciencia ambiental estan correlacionados. Sin embargo, la relación es porque los más educados tienen mayor conciencia ambiental y más salarios. Controlando por educación, estas dos variables no correlacionan (salarios y conciencia ambiental son independientes, condicional en educación).

Independencia condicional:

Un ejemplo con datos simulados: En el panel izquierdo se muestra un gráfico de dispersión entre dos variables y, x con correlación positiva. En el panel derecho se muestran gráficos de dispersión entre las mimas dos variables y, x pero condicional en x (para cada valor de x, que es una variable discreta). En este ejemplo, x e y no son ortogonales ($y \not\perp x$), pero sí son ortogonales o independientes, condicional en z ($y \perp x|z$).



Regresión

Análisis empírico - Descripción de los datos - Dos variables

- La función de esperanza condicionada puede ser lineal.
- Supongamos que es el caso y que podemos escribir

$$y = a + bx + u$$

donde a y b son parámetros y u es una variable aleatoria. Entonces,

$$\begin{aligned} \mathbb{E}(y|x) &= \mathbb{E}(a + bx + u|x) \\ &= a + bx + \mathbb{E}(u|x) \end{aligned}$$

- Cuando $\mathbb{E}(u|x) = 0$ entonces el CEF es lineal.

Análisis empírico - Descripción de los datos - Dos variables

- Definamos la regresión en la población como la solución de la minimización de cuadrados

$$\beta = \arg\min_{b} \mathbb{E}\left[(y - X'b)^2\right]$$

donde β es $k \times 1$, y X es un vector $k \times 1$.

- La solución considerando muchas variables (vector X), es

$$\beta = \mathbb{E}\left[(XX')^{-1}\right]\mathbb{E}\left[(Xy)^{-1}\right]$$

- Para una única variable (y una constante) la pendiente será

$$\beta = \frac{cov(y, x)}{var(x)}$$

- Dos características de la regresión vinculadas directamente con la función de esperanzas condicionadas:
  - Si la función de esperanza condicionada es lineal, una regresión recupera la función
  - Si la función de esperanza condicionada no es lineal, una regresión nos brinda la mejor aproximación lineal a esa función

Regresión lineal. Cómo?

Regresión y predicción en STATA:

```
. regress p47t aniosedu [pw=pondera]
. predict p47hat, xb
```

Incluyendo controles

```
. regress p47t aniosedu edad edad2 sexo
```

Relación causal

La estimación de relaciones causales

Una segunda pregunta, mucho más avanzada y difícil, es la de la relación causal entre las variables de interés (salarios y educación, por ejemplo).
¿Qué relaciones u objetos son de interés? En general el objeto de interés es

$$\frac{\partial \mathbb{E}\left(y|X\right)}{\partial x_1}$$

(si x_1 es una variable continua) es decir, cuánto aumenta el resultado cuando cambia un input, sólo uno, dejando constante todas las demás variables, incluso los inobservables (en el ejemplo, cuánto aumenta el salario cuando aumenta la educación).
Este efecto puede ser medido según las medias condicionadas o según una regresión.

Problema de variables omitidas

Sin embargo, el resultado de una regresión lineal no necesariamente será un efecto causal.
Consideremos $\mathbb{E}(y|x_1)$.
Una de las formas de aproximarnos al objeto de interés es considerar la media de los salarios según años de educación o también por nivel educativo (universitario incompleto, universitario completo).
Por ejemplo $\hat{\beta}_1 = \mathbb{E}(y|x_1 = UC) - \mathbb{E}(y|x_1 = UI)$. Sin embargo, si y depende de muchas otras variables, por ejemplo x_2 (educación del padre), y x_2 cambia con la educación (la educación de un individuo se correlaciona con la educación de la generación anterior) entonces no vamos a estar aproximándonos correctamente al objeto de interés.
Lo importante es que la derivada de interés debe responder a la pregunta de: ¿cuál será el salario si un individuo pasa de no estudiar la universidad a terminar una carrera? Si todo el mayor salario de un individuo universitario viene porque su padre es universitario y esto le genera más contactos y oportunidades laborales entonces en realidad conviene que sólo los individuos con padres universitarios estudien!

Problema de variables omitidas

Una analogía matemática... supongamos:

$$y = g\left(x_1, x_2\right)$$

donde x_2 es una función de x_1, de forma tal que $x_2 = x_2(x_1)$.
Entonces,

$$\frac{dy}{dx_1} = \frac{\partial g\left(x_1, x_2\right)}{\partial x_1} + \frac{\partial g\left(x_1, x_2\right)}{\partial x_2}\frac{\partial x_2}{\partial x_1}$$

En este caso, el efecto total de x_1 es $\frac{dy}{dx_1}$, e incluye el efecto de x_1 a través de x_2.
Pero nos interesa medir sólo el primer término $\frac{\partial g(x_1, x_2)}{\partial x_1}$, dejando fijo el valor de x_2.

Problema de variables omitidas - Ejemplo

Supongamos que $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, con u ortogonal a las x, pero también $x_2 = \delta_0 + \delta_1 x_1 + \varepsilon$ con ε ortogonal a x_1 y a u.
Supongamos que observamos y, x_1 pero no x_2, entonces la media condicionada es

$$\frac{\partial \mathbb{E}\left(y|x_1\right)}{\partial x_1} = \beta_1 + \beta_2 \delta_1$$

porque condicionamos sólo a lo observable.
Con $\beta_2 \neq 0$ y $\delta_1 \neq 0$, hay una diferencia sistemática entre el objeto de interés (β_1) y lo que estimaríamos en una regresión si sólo observamos x_1 (en promedio estimaríamos $\beta_1 + \beta_2 \delta_1$); este es un que sobre-estimará el efecto de x_1 sobre y si $\beta_2 > 0$ y $\delta_1 > 0$.
Notar que si observamos tanto x_1 como x_2, condionamos sobre ambas variables y

$$\frac{\partial \mathbb{E}\left(y|x_1, x_2\right)}{\partial x_1} = \beta_1$$

porque $\mathbb{E}\left(x_2|x_1, x_2\right) = x_2$.

Supuestos

¿Qué tiene que ocurrir para que el coeficiente de la regresión sea el efecto marginal?

- Los inobservables deben ser independientes de la variable x_1, $\mathbb{E}(u|x) = 0$, o también $u \perp x$ (u y x son ortogonales).
- Los inobservables son independientes e idénticamente distribuidos, $u \sim iid$.

Aclaración: Sesgo

El sesgo de un estimador es la diferencia entre su valor esperado y el valor verdadero del parámetro de interés.
Un ejemplo básico:

- supongamos una población determinada (los encuestados en nuestra EPH con ingresos positivos) y un parámetro de interés (la media de los ingresos individuales): $\mu = \mathbb{E}(\ln(y))$.
- consideremos un estimador que es función de los datos de una muestra de tamaño n del universo: $\hat{\hat{\mu}} = \ln\left(\frac{1}{n}\sum_i x_i\right)$

El sesgo será $\mathbb{E}(\hat{\hat{\mu}}) - \mu$. En este caso, el sesgo es positivo: se tenderá a sobre-estimar el parámetro.

La desigualdad de Jensen establece que si $g(.)$ es una función cóncava e y una variable aleatoria, entonces $g(\mathbb{E}(y)) \geq \mathbb{E}(g(y))$. Otro ejemplo: supongamos que S^2 es un estimador insesgado del parámetro σ^2 (varianza, $\mathbb{E}(S^2) = \sigma^2$), entonces la medida $\sqrt{S^2}$ será un estimador sesgado del desvío σ:
$\mathbb{E}(S) = \mathbb{E}(\sqrt{S^2}) \leq \sqrt{\mathbb{E}(S^2)} = \sigma$.

Referencias

Leer:

Angrist & Pischke, "Mostly Harmless Econometrics: An empiricit's Companion", Cap 1