dECON
Facultad de Ciencias Sociales
UNIVERSIDAD DE LA REPÚBLICA

Métodos Cuantitativos

El modelo de regresión lineal múltiple I

Prof. Luis Frones

Clases pasadas…

- Nos hemos centrado en el modelo de regresión lineal simple (dos variables, bivariado) :
  - cómo estimar sus parámetros usando OLS
  - bajo qué condiciones (supuestos) nuestros estimadores MCO son insesgados y eficientes
  - cómo estimar los errores estándar (una medida de precisión) de los estimadores OLS

- Pero los modelos de regresión lineal simple (dos variables) rara vez se utilizan en la práctica porque:
  - por lo general, hay más de un factor importante que afecta y
  - si todos los demás factores importantes se excluyen del modelo, ¿es razonable asumir **RLS4: media condicional cero** ($\mathrm{E}(u|x)=0$) ,es decir, que todos los factores excluidos que afectan y no están correlacionados con x?
  - si no podemos asumir esto, nuestros estimadores están sesgados

Próximas clases…

- Vamos a cambiar nuestro enfoque al modelo de regresión múltiple, el caballo de batalla de toda la econometría y, de hecho, de todas las ciencias sociales

- Nos permite obtener estimaciones de las relaciones entre y y muchos factores, en lugar de solo la relación entre y y x

- Una interpretación *causal ceteris paribus* de la estimación del efecto de x sobre y es más probable que sea válido cuando incluimos otras variables explicativas que pueden estar correlacionadas con x en el modelo porque el supuesto **RLS4: media condicional cero** ($\mathrm{E}(u|x)=0$) es más probable que sea válido

Ejemplo de un modelo con dos regresores

- Supongamos que estamos interesados en el efecto de la educación sobre los salarios:
  - Un modelo simple:

$$salario_i = \alpha_0 + \alpha_1 educ_i + u_i \qquad (1)$$

  - Un modelo múltiple:

$$salario_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + u_i \qquad (2)$$

- En el modelo (2) la experiencia fue sacada del término de error e incluida directamente en el modelo
- Esto nos permite:
  - investigar la relación entre experiencia y salario
  - investigar la relación entre educación y salarios, mientras se controla el efecto de la experiencia en los salarios.
  - Además, si trabajamos con el modelo (1), tendríamos que asumir que la experiencia no está correlacionado con la educación y, si esta suposición es inválida, nuestra los estimadores MCO estarían sesgados

Ejemplo de un modelo con dos regresores

- Decidimos estimar un modelo de regresión múltiple usando una base de datos del Wooldridge:

- Muestra (n): 526 adultos $salarios$ en $ la hora $educ\ y\ exper$ en años

- Aplicando MCO a este modelo y a la base de datos obtenemos:

$$\widehat{salario}_i = -3.39 + 0.64\, educ_i + 0.07 exper_i$$

Aplicando MCO al modelo simple y a la base de datos obtenemos:

$$\widehat{salario}_i = -0.9 + 0.54\, educ_i$$

Note la diferencia en las estimaciones de β_1, el efecto de educación sobre salario

Ejemplo de un modelo con dos regresores

- Decidimos estimar un modelo de regresión múltiple usando una base de datos del Wooldridge:

- Muestra (n): 526 adultos $salarios$ en $ la hora $educ\ y\ exper$ en años

- Aplicando MCO a este modelo y a la base de datos obtenemos:

$$\widehat{salario}_i = -3.39 + 0.64\, educ_i + 0.07 exper_i$$

- Si $educ = 0$ y $exper = 0$ el salario predicho es $-\$3.39$ por hora, es extraño pero nadie en la muestra tiene $educ = 0$ y $exper = 0$

➢ es una extrapolación fuera de la muestra y, por tanto, no relevante

- Manteniendo exper fija, $\Delta exper = 0$, un año de educación adicional predice que aumenta el salario en $\$0.64$ por hora, $\frac{\partial \widehat{salario}_i}{\partial educ} = 0.64$

- Manteniendo educ fija, $\Delta educ = 0$, un año de experiencia adicional predice que aumenta el salario en $\$0.07$ por hora, $\frac{\partial \widehat{salario}_i}{\partial exper} = 0.07$

Ejemplo de un modelo con dos regresores

- Decidimos estimar un modelo de regresión múltiple usando una base de datos del Wooldridge:
- Muestra (n): 526 adultos $salarios$ en $ la hora $educ\ y\ exper$ en años
- Aplicando MCO a este modelo y a la base de datos obtenemos:

$$\widehat{salario}_i = -3.39 + 0.64\, educ_i + 0.07 exper_i$$

- Pero gastar un año adicional de educación reduce los años potenciales de experiencia en el lugar de trabajo en uno. nadie
  - Manteniendo exper fija, $\Delta exper = 0$, un año de educación adicional predice que aumenta el salario en $\$0.64$ por hora, $\frac{\partial \widehat{salario}_i}{\partial educ} = 0.64$
  - Manteniendo educ fija, $\Delta educ = 0$, un año de experiencia adicional predice que aumenta el salario en $\$0.07$ por hora, $\frac{\partial \widehat{salario}_i}{\partial exper} = 0.07$

Ejemplo de un modelo con dos regresores

- Podemos usar el modelo estimado para predecir el efecto de cualquier combinación de cambios en las variables explicativas (regresores) sobre la variable dependiente

- ¿Entonces, cuál es el cambio de quedarse un año extra en educación ($\Delta educ = 1$) y, por tanto, renunciar a un año más de experiencia laboral ($\Delta exper = -1$) ?

$$\widehat{salario}_i = -3.39 + 0.64\, educ_i + 0.07 exper_i$$

- Respuesta:

$$\Delta\widehat{salario}_i = 0.64\, \Delta educ_i + 0.07 \Delta exper_i$$
$$\Delta\widehat{salario}_i = (0.64 \times 1) + (0.07 \times -1) = 0.57$$

- El efecto de permanecer en la educación un año más y, por lo tanto, perder un año de experiencia laboral es un aumento salarial por hora de $\$0.57$

Ejemplo de un modelo con dos regresores

- Decidimos estimar un modelo de regresión múltiple usando una base de datos del Wooldridge:

- Muestra (n): 526 adultos $salarios$ en $ la hora $educ\ y\ exper$ en años

- Aplicando MCO a este modelo y a la base de datos obtenemos:

$$\widehat{salario}_i = -3.39 + 0.64\, educ_i + 0.07 exper_i$$

- Si $educ = 0$ y $exper = 0$ el salario predicho es $-\$3.39$ por hora, es extraño pero nadie en la
  - es ur

- Para que estas estimaciones sean válidas, **el supuesto de media condicional zero** tiene que ser válido
- En este caso, **el supuesto de media condicional zero** es $\mathrm{E}(u|educ, exper) = 0$
- Esto es equivalente al supuesto RLS4, $\mathrm{E}(u|x) = 0$

- Man
  - aum

e que

- Manteniendo educ fija, $\Delta educ = 0$, un año de experiencia adicional predice que aumenta el salario en $\$0.07$ por hora, $\frac{\partial \widehat{salario}_i}{\partial exper} = 0.07$

Un modelo con dos regresores

- Cuando estimamos un modelo con dos regresores

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- Para que nuestros estimadores sean insesgados, el **supuesto de media condicional cero** $E(u|x_1, x_2) = 0$, debe ser válido. Es decir, para cualquier valor de x_1 y x_2 en la población, el valor promedio de u es zero

- Para el modelo

$$salario_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + u_i$$

el **supuesto de media condicional cero** es $\mathrm{E}(u_i|educ_i, exper_i) = 0$, es decir, otros factores omitidos del modelo no están correlacionados con educación y experiencia.

- ¿Pero que pasa con la habilidad innata no observable? Se omitió, y es probable que afecte los salarios y que al mismo tiempo esté correlacionada con educación y, por lo tanto, experiencia.

Un modelo con *k* regresores

- El modelo general con k regresores es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$$

  donde u es el término de error que representa todos los factores otros que $x_1, x_2, x_3, \ldots, x_k$ que afectan y

- No importa cuántos regresores tengamos en el modelo, siempre habrá factores no observados que no podemos controlar

- Estos están colectivamente representados por u

- El **supuesto de media condicional cero** sigue siendo necesario

- Para este modelo con k regresores el supuesto es: $E(u|x_1, x_2, \ldots x_k) = 0$, es decir, todos los factores no observados (u) no están correlacionados con todos y cada uno de las variables explicativas

- Si alguna de las variables explicativas se correlaciona con u, entonces no importa cuán grande sea k, los estimadores de MCO estarán sesgados.

La mecánica de los MCO

- Asumiendo que y de k variables explicativas de forma lineal, es decir:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$$

- Nuestro objetivo es estimar $\beta_0, \beta_1, \beta_2, \beta_3, \ldots, \beta_k$ en la **FRM**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \cdots + \hat{\beta}_k x_k$$

- MCO encuentra valores de $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ que minimizan la suma de los errores al cuadrado (*SSR* en inglés o *SRC* en español):

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_k x_{ki} \right) \right)^2$$

usando una muestra de n observaciones

- No vemos la solución matemática de esto, pero la intuición es similar al Modelo de Regresión Lineal Simple (encontrar la recta que minimice los errores)

La mecánica de los MCO

- En el modelo con $k = 2$: los datos se generan de acuerdo con

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- La FRM es: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

- Los estimadore MCO para el caso $k = 2$ son:

$$\hat{\beta}_1 = \frac{Cov(x_1 y)Var(x_2) - Cov(x_2 y)Cov(x_1 x_2)}{Var(x_1)Var(x_2) - Cov(x_1 x_2)^2}$$

$$\hat{\beta}_2 = \frac{Cov(x_2 y)Var(x_1) - Cov(x_1 y)Cov(x_1 x_2)}{Var(x_2)Var(x_1) - Cov(x_1 x_2)^2}$$

$$\hat{\beta}_0 = \bar{y} - (\hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2)$$

La mecánica de los MCO

- En el modelo con $k = 2$: los datos se generan de acuerdo con

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- La FRM es:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- Los estimadore MCO para el caso $k = 2$ son:

$$\hat{\beta}_1 = \frac{Cov(x_1 y)Var(x_2) - Cov(x_2 y)Cov(x_1 x_2)}{Var(x_1)Var(x_2) - Cov(x_1 x_2)^2}$$

$$\hat{\beta}_2 = \frac{Cov(x_2 y)Var(x_1) - Cov(x_1 y)Cov(x_1 x_2)}{Var(x_2)Var(x_1) - Cov(x_1 x_2)^2}$$

Notar la simetría

$$\hat{\beta}_0 = \bar{y} - (\hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2)$$

La mecánica de los MCO

- En el modelo con $k = 2$: los datos se generan de acuerdo con

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- La FRM es: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

- Los estimadore MCO para el caso $k = 2$ son:

$$\hat{\beta}_1 = \frac{Cov(x_1 y)Var(x_2) - Cov(x_2 y)Cov(x_1 x_2)}{Var(x_1)Var(x_2) - Cov(x_1 x_2)^2}$$

$$\hat{\beta}_2 = \frac{Cov(x_2 y)Var(x_1) - Cov(x_1 y)Cov(x_1 x_2)}{Var(x_2)Var(x_1) - Cov(x_1 x_2)^2}$$

$$\hat{\beta}_0 = \bar{y} - (\hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2)$$

Centrándonos en $\hat{\beta}_1$, note que las diferencias tienen que ver con considerar la correlación entre x_1 y x_2

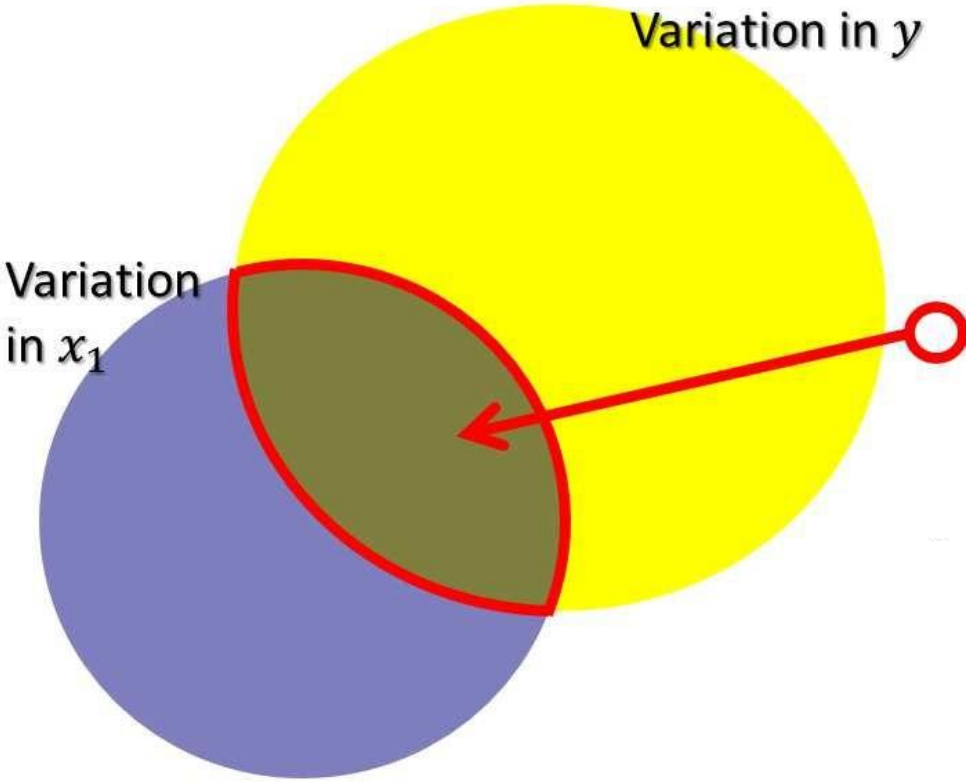
Estimadores MCO del modelo lineal simple

$$\hat{\beta}_1 = \frac{Cov(x_1, y)}{Var(x_1)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

La mecánica de los MCO: una interpretación visual

En el modelo simple: $y = \beta_0 + \beta_1 x_1 + u$, FRM: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ $\hat{\beta}_1 = \frac{Cov(x_1, y)}{Var(x_1)}$

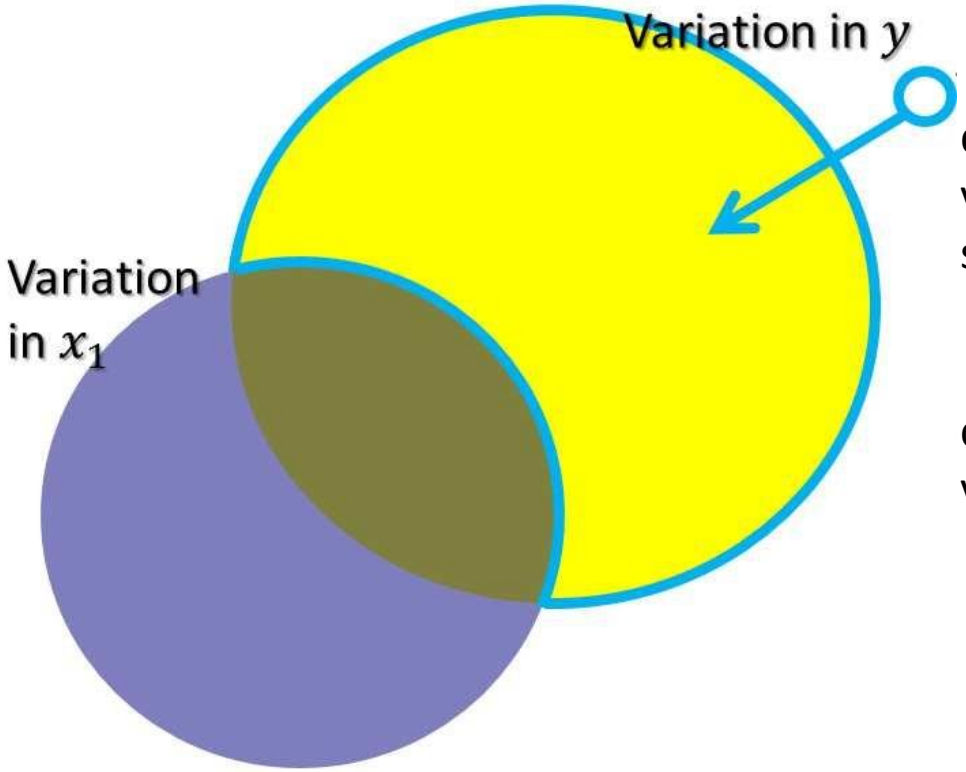


El solapamiento representa la variación en y que puede ser explicada por la variación en x_1, es decir, $Cov(y, x_1)$

El tamaño de este solapamiento refleja la cantidad de información usada por MCO para generar $\hat{\beta}_0$ y $\hat{\beta}_1$

La mecánica de los MCO: una interpretación visual

En el modelo simple: $y = \beta_0 + \beta_1 x_1 + u$, FRM: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ $\hat{\beta}_1 = \frac{Cov(x_1, y)}{Var(x_1)}$

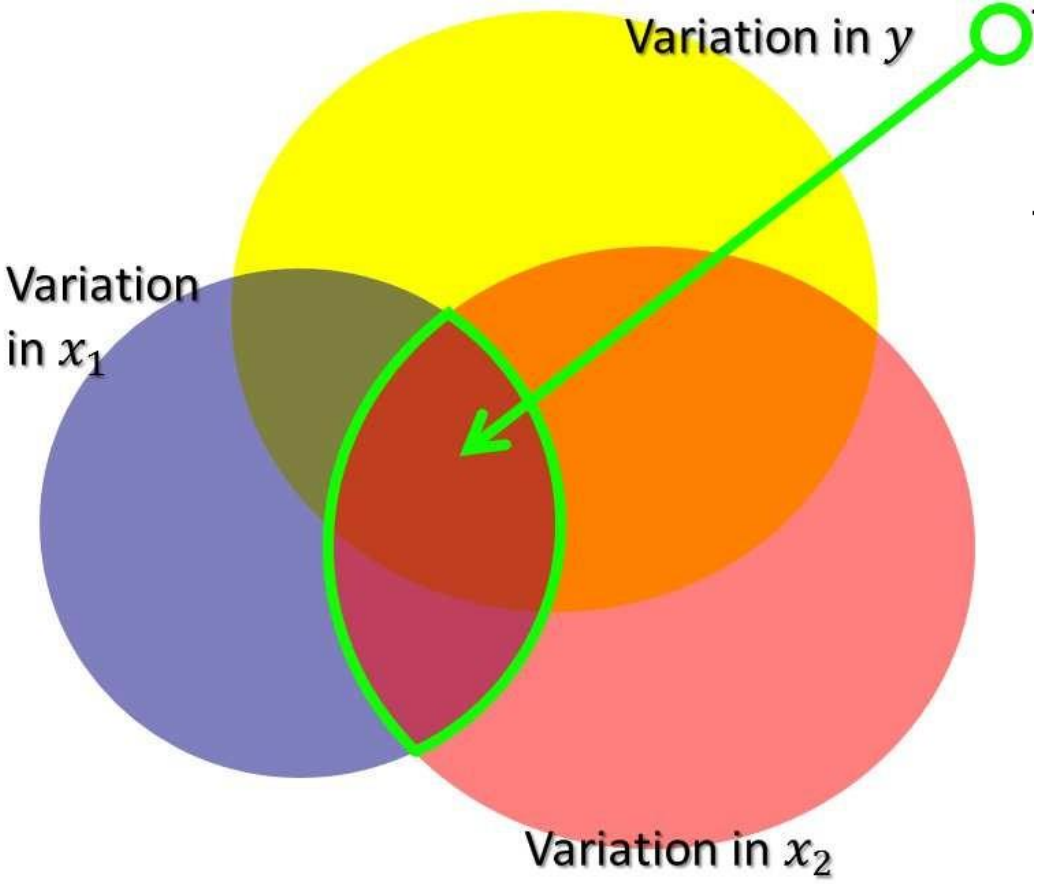


Esta región representa la variación en y que no puede ser explicada por la variación en x_1. Es la variación de y que se debe a u, o a como cambia u, (σ^2).

Cuánto más grande esta región, es decir, cuanto menor el solapamiento, mayor la varianza de nuestro estimador

La mecánica de los MCO: una interpretación visual

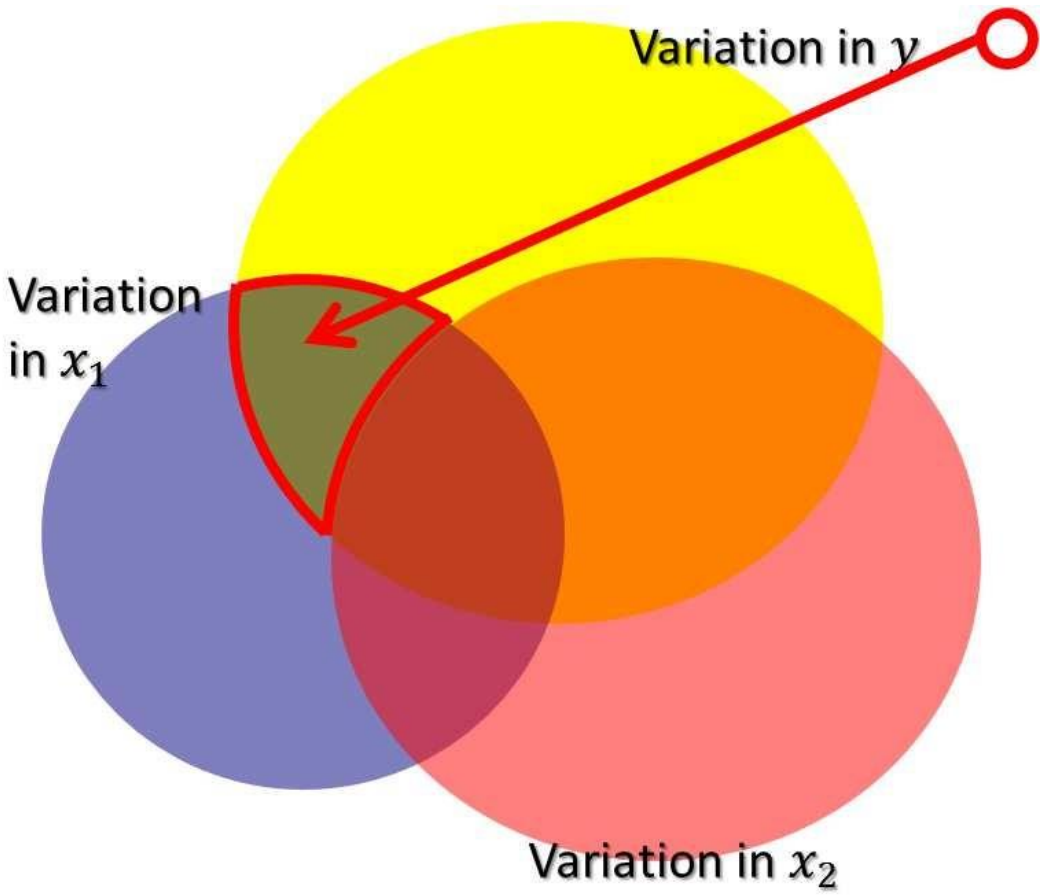
En el modelo de regresión múltiple: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$



Esta región representa la correlación entre x_1 y x_2, parte de la razón de por qué debemos incluir x_2 en el modelo (recordar el supuesto de media condicional cero)

La mecánica de los MCO: una interpretación visual

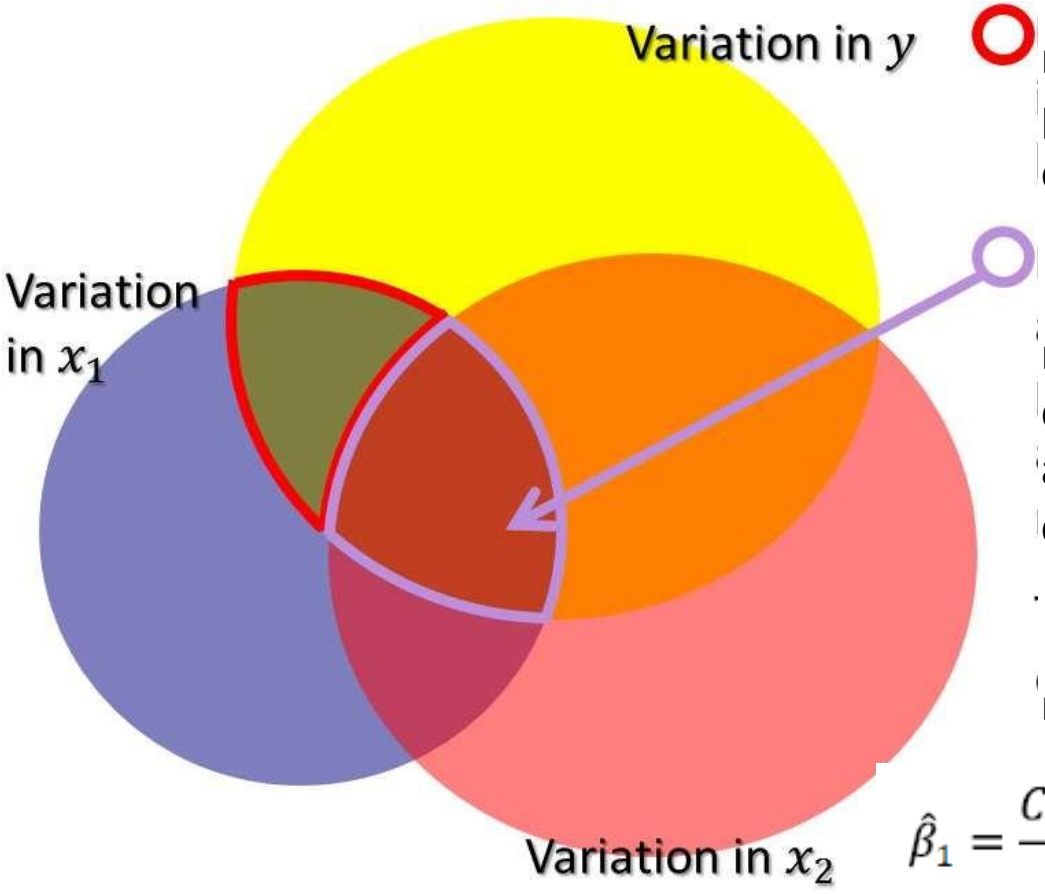
En el modelo de regresión múltiple: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$



Pero ahora el solapamiento que representaba la variación en y que puede ser explicada por la variación en x_1 se ha achicado

La mecánica de los MCO: una interpretación visual

En el modelo de regresión múltiple: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$



Pero ahora el solapamiento que representaba la variación en y que puede ser explicada por la variación en x_1 se ha achicado

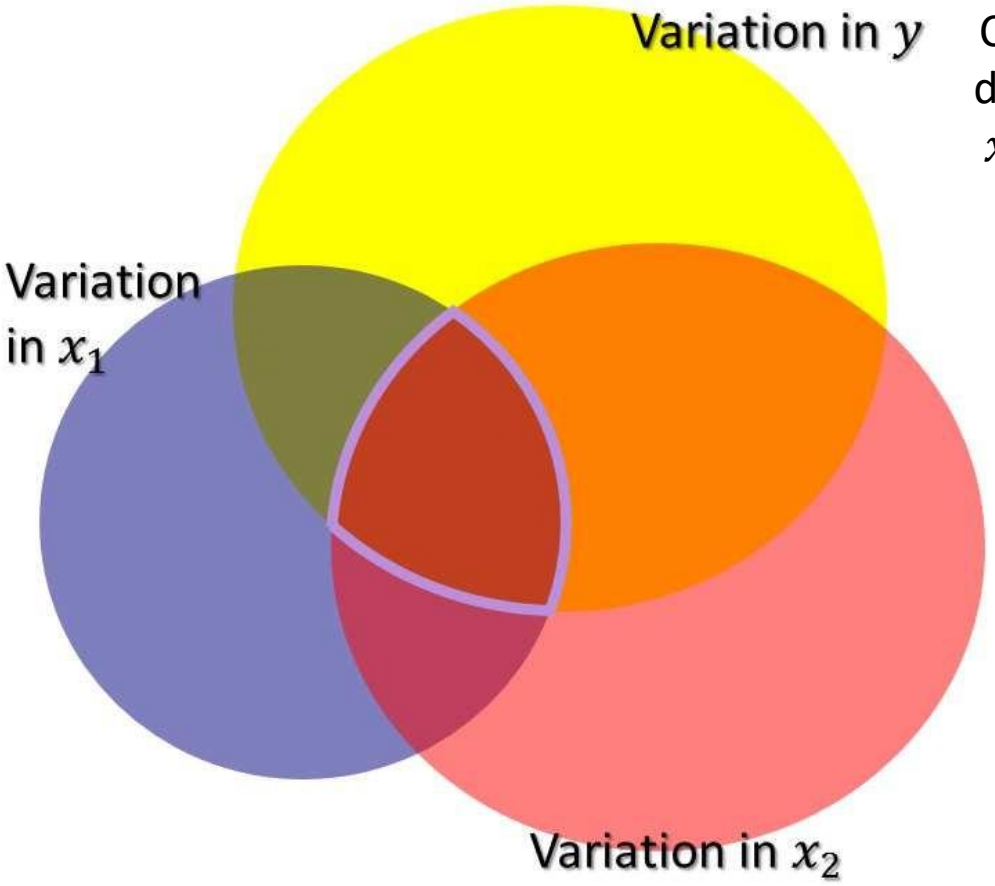
Ahora, cuando calculamos $\hat{\beta}_1$, necesitamos tomar en cuenta la porción de variación en y que puede ser atribuible a x_1 o x_2, o a una combinación

Esto es lo que estos elementos representan en la fórmula

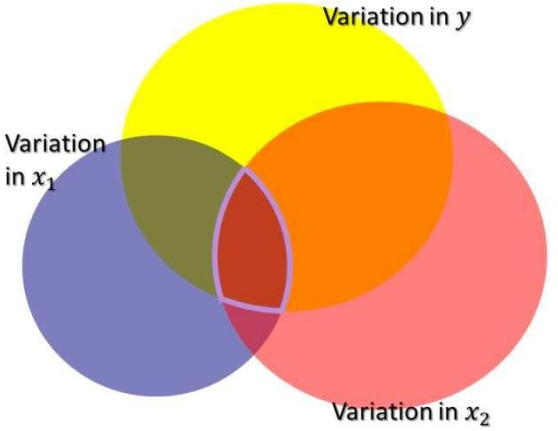
$$\hat{\beta}_1 = \frac{Cov(x_1 y)Var(x_2) - Cov(x_2 y)Cov(x_1 x_2)}{Var(x_1)Var(x_2) - Cov(x_1 x_2)^2}$$

La mecánica de los MCO: una interpretación visual

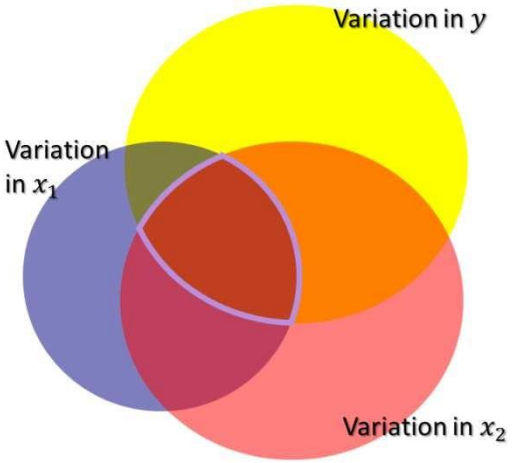
- La correlación entre x_1 y x_2 no sesga las estimaciones
- Sin embargo, cuanto mayor la correlación, menos información hay para los MCO para usar en la estimación -> mayor la varianza de los estimadores



Correlación débil entre x_1 y x_2



Correlación fuerte entre x_1 y x_2



Tres importantes implicaciones de pasar desde el modelo simple al modelo múltiple

1. Los estimadores cambian: en general, los estimadores de las regresiones en el modelo simple y múltiple no son iguales

2. Una interpretación *ceteris paribus* (causal) de los estimadores (y las estimaciones que generan) es más probable que sea correcta

3. Los estimadores (y las estimaciones que generan) tienen una interpretación “eliminando parcialmente”

1. Los estimadores cambian: en general, los estimadores de regresiones simples y múltiples no son iguales

- Modelo Simple (MRS): $\hat{\beta}_1 = \frac{Cov(x_1, y)}{Var(x_1)}$
- Modelo Múltiple (MRM): $\hat{\beta}_1 = \frac{Cov(x_1 y)Var(x_2) - Cov(x_2 y)Cov(x_1 x_2)}{Var(x_1)Var(x_2) - Cov(x_1 x_2)^2}$

Los estimadores son diferentes

Aplicando MCO al modelo múltiple en la base de datos:

$$\widehat{salario}_i = -3.39 + 0.64\, educ_i + 0.07 exper_i$$

Aplicando MCO al modelo simple y a la base de datos:

$$\widehat{salario}_i = -0.9 + 0.54\, educ_i$$

1. Los estimadores cambian: en general, los estimadores de regresiones simples y múltiples no son iguales

- Modelo Simple (MRS): $\hat{\beta}_1 = \frac{Cov(x_1, y)}{Var(x_1)}$
- Modelo Múltiple (MRM): $\hat{\beta}_1 = \frac{Cov(x_1 y)Var(x_2) - Cov(x_2 y)Cov(x_1 x_2)}{Var(x_1)Var(x_2) - Cov(x_1 x_2)^2}$

Los estimadores son diferentes al menos que:

- x_1 y x_2 estén incorrelacionados, es decir, $Cov(x_1, x_2) = 0$, entonces los estimadores del MRM y MRS son iguales
- el efecto *ceteris paribus* del cambio de x_2 en $\hat{y}$ es cero, es decir, $\hat{\beta}_2 = 0$

(Esto también aplica a los estimadores de β_2)

2. Una interpretación *ceteris paribus* (causal) de los estimadores del MRM es más probable que sea correcta

- Modelo Simple (MRS): $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$
- Modelo Múltiple (MRM): $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

En el MRM, hemos **controlado** por el efecto de x_2 en y. En el MRS no.

Es por esto que una interpretación *ceteris paribus* del estimador β_1 de la estimación MRM sea probablemente más correcto.

Aplicando MCO al modelo múltiple en la base de datos:

$$\widehat{salario}_i = -3.39 + 0.64\, educ_i + 0.07 exper_i$$

Aplicando MCO al modelo simple y a la base de datos:

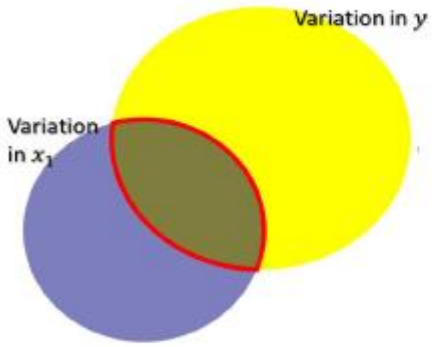
$$\widehat{salario}_i = -0.9 + 0.54\, educ_i$$

3. Los estimadores (y las estimaciones que generan) tienen una interpretación “eliminando parcialmente"

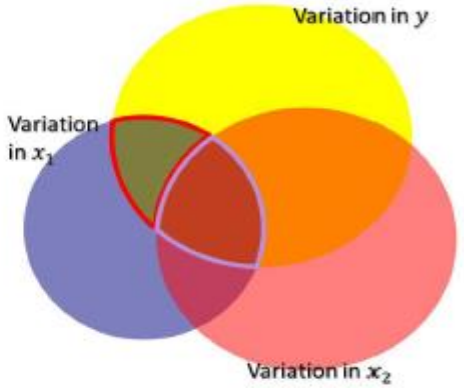
- Modelo Simple (MRS): $\hat{\beta}_1 = \frac{Cov(x_1, y)}{Var(x_1)}$
- Modelo Múltiple (MRM): $\hat{\beta}_1 = \frac{Cov(x_1 y)Var(x_2) - Cov(x_2 y)Cov(x_1 x_2)}{Var(x_1)Var(x_2) - Cov(x_1 x_2)^2}$

En el MRM, "eliminamos parcialmente" esa parte de la variación que podría ser explicada por x_1 o x_2, o una mezcla de ambos

MRS



MRM



3. Los estimadores (y las estimaciones que generan) tienen una interpretación “eliminando parcialmente"

- Modelo Simple (MRS): $\hat{\beta}_1 = \frac{Cov(x_1, y)}{Var(x_1)}$
- Modelo Múltiple (MRM): $\hat{\beta}_1 = \frac{Cov(x_1 y)Var(x_2) - Cov(x_2 y)Cov(x_1 x_2)}{Var(x_1)Var(x_2) - Cov(x_1 x_2)^2}$

En el MRM, "eliminamos parcialmente" esa parte de la variación que podría ser explicada por x_1 o x_2, o una mezcla de ambos
Esto hace que la estimación sea más exigente, pero es un paso importante. hacia una interpretación *ceteris paribus*

Aplicando MCO al modelo múltiple en la base de datos:

$$\widehat{salario}_i = -3.39 + 0.64\, educ_i + 0.07 exper_i$$

Aplicando MCO al modelo simple y a la base de datos:

$$\widehat{salario}_i = -0.9 + 0.54\, educ_i$$

En resumen….

- El poder del análisis de regresión múltiple MCO debería quedar claro

- Cuando incluimos un factor adicional en una regresión, estamos controlando por su efecto sobre la variable dependiente y esto nos mueve hacia una interpretación *ceteris paribus* creíble de los estimadores de MCO y las estimaciones que generan

- En los experimentos **controlamos** por los factores que no nos interesan manteniéndolos constantes en la etapa de generación de datos

- En el análisis de regresión múltiple MCO podemos controlar los factores que no nos interesan en la etapa de análisis

➢ La regresión múltiple MCO permite a los cientistas sociales hacer de forma no experimental lo que hacen los experimentalistas hacen en un experimento de laboratorio controlado.

Lecturas

- Wooldridge capítulo 3