Econometría

Maestría en Economía - Maestría en Econometría

Lecture 1

Agenda

Introducción

- El objetivo principal de la inferencia causal es determinar el impacto que tiene alguna política pública o programa social.
- Pensemos en un ejemplo. El gobierno está pensando en un nuevo programa de capacitación dirigido a jóvenes sin educación formal. El programa capacita en algún oficio al trabajador con el objetivo de disminuir el tiempo que pasa hasta encontrar su primer empleo.
- En este caso la política es el nuevo programa de capacitación y su efectividad debiera verse en una reducción del tiempo de desempleo.
- El desafío principal de llevar a cabo una evaluación de impacto es la **identificación de una relación de causalidad entre el programa y los resultados de interés.**
- El modelo de resultados potenciales es un procedimiento que nos permite evaluar el impacto de las intervenciones sobre diferentes resultados de manera precisa.

Introducción

- Para identificar una relación de causalidad entre la implementación del programa y el resultado del mismo, el modelo de resultados potenciales se pregunta: **cuál hubiera sido el resultado de no haberse implementado el programa?**
- Supongamos un trabajador u.
- Denotemos con $Y_u(T)$ al tiempo que pasa (medido en meses) hasta que encuentra empleo después de haber recibido la capacitación del programa y con $Y_u(C)$ al tiempo que pasa desempleado hasta encontrar empleo de no haber recibido la capacitación del programa.
- Los valores $Y_u(T)$ y $Y_u(C)$ se conocen en la literatura como **resultados potenciales**.
- Ex-ante, el trabajador u potencialmente pasará $Y_u(T)$ ó $Y_u(C)$ meses en desempleo.
- Ex-post, nosotros solo observamos uno de esos resultados potenciales: $Y_u(T)$ si recibió la capacitación ó $Y_u(C)$ si no la recibió.
- Este problema se denomina en la literatura el **problema fundamental de la inferencia causal**.

Agenda

Modelo de inferencia causal

- Concepto fundamental: **exposición potencial**. Que cada unidad de la población pueda estar expuesta potencialmente a cualquiera de las causas.
- Para empezar a formalizar nuestro análisis vamos a considerar una **población** objetivo U de unidades denotadas por u.
- Las unidades u son los objetos básicos de estudio.
- Ejemplos de unidades son: individuos, empresas, familias, parcelas de tierra, equipos de laboratorio etc.
- Una **variable** es una función que se define sobre cada unidad u de U.
- El valor de una variable para una unidad u determinada es un número asignado por alguna medición aplicada sobre u.
- Por ejemplo, para cada unidad u en U hay asociado un valor $Y(u)$ de la variable Y. Si Y es la variable de interés, se la denomina variable de respuesta.

Modelo de inferencia causal

- Todas las distribuciones, probabilidades y valores esperados de las variables involucradas en el programa se calculan sobre U.
- Por simplicidad se asumirá que hay sólo dos niveles de tratamiento, expresados por T (el tratamiento) y C (el control) respectivamente.
- Sea S una variable que indica la causa a la cual fue expuesta cada $u \in U$, o sea $S = 1$ indica exposición al tratamiento y $S = 0$ indica exposición al control.
- En un estudio controlado, S está diseñado por el investigador. En un estudio no controlado S está determinado en alguna medida por factores que se encuentran más allá del control del investigador.
- En cualquiera de los dos casos un aspecto clave de la noción de causa en este modelo es que **el valor $S(u)$ para cada unidad 'pudo haber sido diferente'.**

Modelo de inferencia causal

- La interpretación de $Y_T(u)$ e $Y_C(u)$ para un elemento dado u es que $Y_T(u)$ es el valor de la variable respuesta que se hubiera observado si el individuo fue expuesto a la causa T mientras que $Y_C(u)$ es el valor de la variable respuesta que se hubiera observado **para el mismo individuo** si fue expuesto a la causa C.
- Entre los dos **resultados potenciales** correspondientes a los dos potenciales tratamientos solo **UN** resultado es observado. El otro, denominado **contrafáctico** no se observa.

Modelo de inferencia causal

- El efecto de la causa T sobre u medido por Y y en relación a la causa C es la diferencia entre $Y_T(u)$ e $Y_C(u)$. En el modelo estará representado por la siguiente diferencia:

$$Y_T(u) - Y_C(u) \tag{1}$$

- La diferencia anterior representa el efecto causal de T (con respecto a C) sobre u (medido a través de Y).
- La expresión (1) es la forma en que el modelo de inferencia causal expresa el concepto de causalidad más básico. Establece que T causa un efecto $[Y_T(u) - Y_C(u)]$ sobre u.
- Esta forma de definir el **efecto causal** usando dos resultados potenciales se denomina **causalidad contrafáctica**.

Modelo de inferencia causal

- Denotemos por $s_T = 1, 0,$ al tratamiento observado y por y_u a la respuesta observada:

$$
\begin{aligned}
y_u &= s_T \times Y_T(u) + (1 - s_T) \times Y_C(u), \;\; u = 1, \ldots, N \\
&= Y_C(u) + s_T \times [Y_T(u) - Y_C(u)], \;\; u = 1, \ldots, N \qquad (2)
\end{aligned}
$$

- Esta ecuación se conoce como "switching equation".

Problema Fundamental de la Inferencia Causal

Es imposible observar los valores de $Y_T(u)$ y $Y_C(u)$ sobre una misma unidad y por lo tanto es imposible observar el efecto de T sobre u.

Agenda

Modelo de inferencia causal

- La amenaza implícita en el PFIC es que la inferencia causal es imposible. Sin embargo, hay una "solución estadística"
- La solución estadística hace uso de la población U en un sentido típicamente estadístico.
- Se define el **efecto causal promedio** (*ATE - average treatment effect*) como la diferencia entre dos valores esperados, $E(Y_T)$ y $E(Y_C)$.

$$ATE = E[Y_T(u) - Y_C(u)] = E(Y_T) - E(Y_C) \tag{3}$$

Independencia

- Se considera que la población U de individuos es "grande" y los datos observados para cada individuo son los valores de S e Y.
- La información que obtenemos se refiere a:

$$E(Y|s_T = 1) = E(Y_T|s_T = 1), \qquad E(Y|s_T = 0) = E(Y_C|s_T = 0)$$

- Es esencial reconocer que $E(Y_T)$ y $E(Y_T|s_T = 1)$ no son equivalentes (idem para la causa C).
- Sin embargo, cuando los elementos de la población son asignados aleatoriamente a la causa C o T es posible verificar que la causa a la que está expuesta cada unidad u de la población será estadísticamente independiente de cualquier otra variable, incluyendo a Y_T e Y_C.
- Si el procedimiento de aleatorización se realiza correctamente entonces es posible que S sea independiente de Y_T e Y_C y otras variables en U.

Tabla de resultados potenciales

u	$S(u)$	$Y_T(u)$	$Y_C(u)$	Y
1	T	$Y_T(1)$	$Y_C(1)$	$Y_T(1)$
2	T	$Y_T(2)$	$Y_C(2)$	$Y_T(2)$
3	C	$Y_T(3)$	$Y_C(3)$	$Y_C(3)$
4	T	$Y_T(4)$	$Y_C(4)$	$Y_T(4)$
5	C	$Y_T(5)$	$Y_C(5)$	$Y_C(5)$
6	C	$Y_T(6)$	$Y_C(6)$	$Y_C(6)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	T	$Y_T(N)$	$Y_C(N)$	$Y_T(N)$
		$E(Y_T)$	$E(Y_C)$	

- Cuándo $E(Y_T) - E(Y_C) \stackrel{?}{=} E(Y_T|S=1) - E(Y_C|S=0)$

Tabla de resultados potenciales

u	$S(u)$	$Y_T(u)$	$Y_C(u)$	Y
1	1	$Y_T(1)$		$Y_T(1)$
2	1	$Y_T(2)$		$Y_T(2)$
3	0		$Y_C(3)$	$Y_C(3)$
4	1	$Y_T(4)$		$Y_T(4)$
5	0		$Y_C(5)$	$Y_C(5)$
6	0		$Y_C(6)$	$Y_C(6)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1	$Y_T(N)$		$Y_T(N)$
		$E(Y_T)$	$E(Y_C)$	

- Cuándo $E(Y_T) - E(Y_C) \stackrel{?}{=} E(Y_T|S=1) - E(Y_C|S=0)$

Tabla de resultados potenciales

u	$S(u)$	$Y_T(u)$	$Y_C(u)$	Y
1	1	$Y_T(1)$	$Y_C(1)$	$Y_T(1)$
2	1	$Y_T(2)$	$Y_C(2)$	$Y_T(2)$
3	0	$Y_T(3)$	$Y_C(3)$	$Y_C(3)$
4	1	$Y_T(4)$	$Y_C(4)$	$Y_T(4)$
5	0	$Y_T(5)$	$Y_C(5)$	$Y_C(5)$
6	0	$Y_T(6)$	$Y_C(6)$	$Y_C(6)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1	$Y_T(N)$	$Y_C(N)$	$Y_T(N)$
		$E(Y_T)$	$E(Y_C)$	

- Cuándo $E(Y_T) - E(Y_C) \stackrel{?}{=} E(Y_T|S=1) - E(Y_C|S=0)$

Tabla de resultados potenciales

u	$S(u)$	$Y_T(u)$	$Y_C(u)$	Y
1	1			$Y_T(1)$
2	1			$Y_T(2)$
3	0			$Y_C(3)$
4	1			$Y_T(4)$
5	0			$Y_C(5)$
6	0			$Y_C(6)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1			$Y_T(N)$
		$E(Y_T)$	$E(Y_C)$	

- Cuándo $E(Y_T) - E(Y_C) \stackrel{?}{=} E(Y_T|S=1) - E(Y_C|S=0)$

Tabla de resultados potenciales

u	$S(u)$	$Y_T(u)$	$Y_C(u)$	Y
1	1			$Y_T(1)$
2	1			$Y_T(2)$
3	0			$Y_C(3)$
4	1			$Y_T(4)$
5	0			$Y_C(5)$
6	0			$Y_C(6)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1			$Y_T(N)$
		$E(Y_T)$	$E(Y_C)$	$E(Y_T \mid S=1)$

- Cuándo $E(Y_T) - E(Y_C) \stackrel{?}{=} E(Y_T \mid S=1) - E(Y_C \mid S=0)$

Tabla de resultados potenciales

u	$S(u)$	$Y_T(u)$	$Y_C(u)$	Y
1	1			$Y_T(1)$
2	1			$Y_T(2)$
3	0			$Y_C(3)$
4	1			$Y_T(4)$
5	0			$Y_C(5)$
6	0			$Y_C(6)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1			$Y_T(N)$
		$E(Y_T)$	$E(Y_C)$	$E(Y_C \mid S=0)$

- Cuándo $E(Y_T) - E(Y_C) \stackrel{?}{=} E(Y_T \mid S=1) - E(Y_C \mid S=0)$

Independencia

- Si el supuesto de independencia es válido entonces

$$E(Y_T) = E(Y_T | s_T = 1)$$

y

$$E(Y_C) = E(Y_C | s_T = 0)$$

- Por lo tanto el efecto causal promedio se obtiene como

$$ATE = E(Y_s | s_T = 1) - E(Y_s | s_T = 0)$$

Si la aleatorización es posible (supuesto de independencia es válido), siempre se puede estimar el efecto causal promedio como una diferencia de medias.

- La expresión anterior revela que se puede utilizar la información de distintos individuos para conocer ATE.

Agenda

El Director de Capacitación perfecto: ejemplo

- Los datos hipotéticos de la tabla muestran los resultados potenciales de dos programas de entrenamiento para el empleo:
  - Y_T: meses hasta conseguir el primer empleo después de recibir el nuevo programa de entrenamiento.
  - Y_C: meses hasta conseguir el primer empleo después de recibir el viejo programa de entrenamiento.

u	$Y_T(u)$	$Y_C(u)$
1	10	14
2	11	9
3	14	10
4	12	9
5	6	7
6	9	10
E(Y)	10.33	9.83

- **ATE** $= E(Y_T) - E(Y_C) = 10.33 - 9.83 = 0.5$, en promedio, el nuevo programa de capacitación no reduce el tiempo de desempleo.

El director de capacitación perfecto: ejemplo

- El **director de capacitación perfecto** elige para cada trabajador el mejor tratamiento (el que lo deja desempleado el menor tiempo posible).
- Qué observaríamos?

u	S(u)	$Y_T(u)$	$Y_C(u)$
1	1	10	?
2	0	?	9
3	0	?	10
4	0	?	9
5	1	6	?
6	1	9	?
$E(Y\|S)$		8.33	9.33

- **ATE** $= E(Y_T|S=1) - E(Y_C|S=0) = 8.33 - 9.33 = -1$, en promedio, el nuevo programa de capacitación reduce el tiempo de desempleo en un mes!.

El Director de Capacitación perfecto: ejemplo

- En el ejemplo del Director de Capacitación perfecto, el entrenamiento que cada unidad recibe depende del resultado potencial de esa unidad.
- El mecanismo de asignación, $S(u)$, no es independiente de los resultados potenciales.
- Esto provoca que el estimador del ATE sea sesgado.
- Sobre la base de los resultados observados, concluiríamos erróneamente que el nuevo programa de entrenamiento funciona.
- Como conseguir que el mecanismo de asignación sea independiente de los resultados potenciales? **Asigne aleatoriamente.**

El Director de Capacitación perfecto: ejemplo

- El verdadero ATE de la primera tabla es un parámetro (está calculado con los resultados potenciales).
- El ATE de la segunda tabla es una estimación (está calculado con los datos observados).
- Necesitamos un estimador que sea INSESGADO: si la asignación se repitiera una y otra vez, el promedio de los estimadores debiera ser igual al parámetro.

El Director de Capacitación perfecto: ejemplo

- En el ejemplo del Director de Capacitación perfecto hay 20 formas diferentes de asignar 3 unidades a la nueva capacitación y 3 a la vieja.

	$\#1 \Rightarrow$	111000	ATE estimado $= 3$
	$\#2 \Rightarrow$	110100	ATE estimado $= 2$
	$\dots$		
Director perfecto:	$\#7 \Rightarrow$	100011	ATE estimado $= -1$
	$\dots$		
	$\#20 \Rightarrow$	000111	ATE estimado $= -2$

- El promedio de los 20 ATE observados es 0.5 y coincide con el ATE verdadero.
- **El Director de Capacitación perfecto** siempre elige el #7.
- **El mecanismo de asignación independiente** selecciona el # aleatoriamente.

Agenda

Modelo de Regresión Simple

- Considere nuevamente la *switching equation (2)*

$$y_u = Y_C(u) + s_T \times [Y_T(u) - Y_C(u)]$$

- Escribamos los resultados potenciales como

$$\begin{aligned} Y_C(u) &= E[Y_C(u)] + \epsilon_C(u) && (4) \\ Y_T(u) &= E[Y_T(u)] + \epsilon_T(u) && (5) \end{aligned}$$

  piense que siempre podemos escribir a una variable aleatoria como su esperanza matemática más una variable de media cero.

- Reemplazando estos resultados potenciales en la ecuación (2) tenemos:

$$\begin{aligned} y_u &= E[Y_C(u)] + s_T(u) \times \{E[Y_T(u)] - E[Y_C(u)]\} + \epsilon_u \\ &= \overbrace{\beta_0} + s_T(u) \times \overbrace{\beta_1} + \epsilon_u && (6) \end{aligned}$$

  donde $\epsilon_u = \epsilon_C(u) + s_T(u) \times \{\epsilon_T(u) - \epsilon_C(u)\}$.

- Si se cumple el supuesto de independencia entre s_T y los resultados potenciales:

$$\begin{aligned}
\text{sesgo} &= E[\epsilon_u|s_T = 1] - E[\epsilon_u|s_T = 0] = E[\epsilon_T(u)|s_T = 1] - E[\epsilon_C(u)|s_T = 0] \\
&= E\{[Y_T(u) - E(Y_T(u))]|s_T = 1\} - E\{[Y_C(u) - E(Y_C(u))]|s_T = 0\} \\
&= \{E[Y_T(u)|s_T = 1] - E[Y_T(u)]\} - \{E[Y_C(u)|s_T = 0] - E[Y_C(u)]\} \\
&= 0
\end{aligned}$$

  y el sesgo es igual a cero.

- El parámetro de la pendiente en este modelo es igual a la diferencia de medias entre grupos y se puede expresar como:

$$\begin{aligned}
\beta_1 &= \frac{COV(y, s)}{VAR(s)} \\
&= E(y_1 - y_0)
\end{aligned}$$

- Es fundamental el supuesto de ignorabilidad del tratamiento (asignación aleatoria).

$$\begin{aligned}
\frac{COV(y,s)}{VAR(s)} &= \frac{E[(sy_1 + (1-s)y_0)s] - E[(sy_1 + (1-s)y_0)]E[s]}{p(1-p)} \\
&= \frac{E[s^2 y_1] + E[s(1-s)y_0] - E[sy_1]p + E[(1-s)y_0]p}{p(1-p)} \\
&= \frac{E[s^2 y_1](1-p) - E[(1-s)y_0]p}{p(1-p)} \\
&= \frac{E[E[sy_1|s]](1-p) - E[E[(1-s)y_0|s]]p}{p(1-p)} \\
&= \frac{E[y_1|s=1]P(s=1)(1-p) - E[y_0|s=0]P(s=0)p}{p(1-p)} \\
&= \frac{\{E[y_1|s=1] - E[y_0|s=0]\}p(1-p)}{p(1-p)} \\
&= \{E[y_1|s=1] - E[y_0|s=0]\} = E(y_1 - y_0)
\end{aligned}$$

bajo asignación aleatoria del tratamiento.

Modelo de Regresión Lineal

- En la práctica es muy difícil realizar una aleatorización de las unidades poblacionales.
- Piense en el ejemplo del programa de capacitación. Es muy poco probable que el gobierno designe aleatoriamente a aquellos a quienes les brindará la capacitación.
- En general siempre habrá cierta autoselección en el tratamiento.
- Si esto ocurre, entonces habrá alguna correlación entre s_T e y_u y el supuesto de independencia no se cumple.
- Aún en estos casos es posible recuperar el impacto del programa si se controla por las variables que inducen la correlación entre s_T e y_u.
- Para que el modelo de regresión tenga una interpretación causal necesitamos cambiar el supuesto de independencia.
- Supuesto de Independencia Condicional o Selección sobre Observables: llamemos w_u al vector de variables observables. Entonces

$$\{Y_T(u),\ Y_C(u)\} \perp\!\!\!\perp s_T(u)|w_u \tag{7}$$

- Lo que dice el supuesto de independencia condicional es que condicionando sobre las características w_u el sesgo de selección desaparece.
- En otras palabras, la asignación del tratamiento es tan buena como si se hubiera hecho un experimento aleatorizado entre unidades tratadas y no tratadas que resultan comparables en términos de sus características observables.
- Descomponiendo la parte aleatoria de los resultados potenciales, ϵ_u, en una parte lineal que depende de las características observables w_u y un término de error v_u tenemos:

$$\epsilon_u = w_u \gamma + v_u \tag{8}$$

donde γ es un vector de parámetros que satisfacen $E[\epsilon_u | w_u] = w_u \gamma$.

Modelo de Regresión Lineal

- Por lo tanto, en virtud del supuesto de independencia condicional

$$
\begin{aligned}
E[y_u|w_u, s_T] = E[y_u|w_u] &= \beta_0 + \beta_1 s_T(u) + E[\epsilon_u|w_u] \\
&= \beta_0 + \beta_1 s_T(u) + w_u\gamma \qquad (9)
\end{aligned}
$$

- De esta ecuación surge que se puede relajar el supuesto de independencia condicional al de **independencia condicional en media:**

$$
E[y_u|w_u, s_T] = E[y_u|w_u] \qquad (10)
$$

- El coeficiente β_1 sigue teniendo la interpretación de ser el efecto causal de interés.
- Esto es,

$$
E[y_u|w_u, s_T = 1] - E[y_u|w_u, s_T = 0] = (\beta_0 + \beta_1 + w_u\gamma) - (\beta_0 + w_u\gamma) = \beta_1 \qquad (11)
$$

Modelo de Regresión Lineal

- Un supuesto implícito en este análisis es que las características observables afecten a la variable de resultado de la misma manera en los grupos de tratamiento y de control.
- Supongamos que el efecto es diferente y denotemos por γ_1 y γ_0 a los vectores de parámetros asociados a w_u en el grupo de tratamiento y de control, respectivamente.
- Esto es,

$$\begin{aligned} E[y_u|w_u, s_T = 1] - E[y_u|w_u, s_T = 0] &= (\beta_0 + \beta_1 + w_u\gamma_1) - (\beta_0 + w_u\gamma_0) \\ &= \beta_1 + w_u(\gamma_1 - \gamma_0) \end{aligned} \tag{12}$$

- Definamos el **efecto tratamiento promedio condicional** como:

$$ATE(w) = \tau_w = E[y_u|w_u, s_T = 1] - E[y_u|w_u, s_T = 0] = E[y_1|w_u] - E[y_0|w_u] \tag{13}$$

- La primera igualdad es la que podemos calcular con los datos poblacionales y la segunda igualdad se cumple por el supuesto de independencia condicional.

- Usando la ley de expectativas iteradas podemos recuperar el ATE.

$$ATE = \tau = E\{E[y_1|w_u] - E[y_0|w_u]\} \tag{14}$$

- Para que la esperanza matemática en (14) se pueda calcular necesitamos que para cualquier combinación de valores de w_u en la población haya unidades tratadas y no tratadas.
- Esta última condición se denomina en la literatura de causalidad como **soporte común**.
- El soporte común descarta que el tratamiento pueda ser predecido por las características observables en forma determinística (lo que implicaría que todas las unidades están en el tratamiento ó están en el control).

Modelo de Regresión Lineal

- Matemáticamente el supuesto de soporte común (“superposición” u “overlapping”) es,

$$0 < \Pr\left(s_T(u) = 1 | w_u\right) < 1 \tag{15}$$

- La probabilidad condicional de recibir el tratamiento (también llamada “propensity score”) es mayor a cero y menor a uno de forma que s_T no es una función determinística de w_u y se cumple el supuesto de soporte común.

Agenda

Modelo de inferencia causal: Supuestos

- **Supuesto 0:** Para cada unidad u, tenemos: $(Y_T(u), Y_C(u), w_u)$, donde w_u son características observables de u. La muestra de datos $(Y_T(u), Y_C(u), w_u)_{u=1}^{n}$ es i.i.d. seleccionada desde una población P_o.

- **Supuesto 1 (Independencia Condicional en Media):**
$\mathbb{E}\left[Y_T(u) \mid w_u, s_T(u)\right] = \mathbb{E}\left[Y_T(u) \mid w_u\right]$ y $\mathbb{E}\left[Y_C(u) \mid w_u, s_T(u)\right] = \mathbb{E}\left[Y_C(u) \mid w_u\right]$.

- **Supuesto 2 (Superposición u "overlap"):**
$0 < \Pr\left(s_T(u) = 1 | w_u\right) < 1$.

Modelo de inferencia causal: identificación

- Dados los supuestos 1 y 2, se puede identificar el ATE.

$$\tau_{ATE}(w) = \mu_1(w_u) - \mu_0(w_u), \forall w_u$$

donde $\mu_s(w_u) = E\left[y_u \mid s_T = s, W = w_u\right]$

- Dado el Supuesto 1, las siguientes igualdades se cumplen ($s = 0, 1$):

$$\mu_s(w_u) = \mathbb{E}\left[y_u(s) \mid W = w_u\right] = \mathbb{E}\left[y_u(s) \mid s_T(u) = s, W = w_u\right] = \mathbb{E}\left[y_u \mid s_T(u) = s, W = w_u\right]$$

y $\mu_s(w_u)$ está identificado.

- Entonces, se puede calcular el efecto tratamiento promedio, τ, estimado primero el ATE para una subpoblación con covariables $W = w_u$,

$$\begin{aligned}
\tau(w) &\equiv \mathbb{E}\left[Y_T(u) - Y_C(u) \mid W = w_u\right] = \mathbb{E}\left[Y_T(u) \mid W = w_u\right] - \mathbb{E}\left[Y_C(u) \mid W = w_u\right] \\
&= \mathbb{E}\left[Y_T(u) \mid W = w_u, s_T(u) = 1\right] - \mathbb{E}\left[Y_C(u) \mid W = w_u, s_T(u) = 0\right] \\
&= \mathbb{E}\left[y_u \mid W = w_u, s_T(u) = 1\right] - \mathbb{E}\left[y_u \mid W = w_u, s_T(u) = 0\right]
\end{aligned}$$

Modelo de inferencia causal: identificación

- Note que para que esta estimación sea posible, necesitamos estimar $\mathbb{E}\left[y_u \mid W = w_u, s_T(u) = s\right]$ para todos los valores de w_u y s en el soporte de estas variables.

- Aquí es donde entra el Supuesto 2.

- Si el Supuesto 2 no se cumple para $W = w_u$, no es posible estimar las dos esperanzas matemáticas en esos puntos, $\mathbb{E}\left[y_u \mid W = w_u, s_T(u) = 1\right]$ y $\mathbb{E}\left[y_u \mid W = w_u, s_T(u) = 0\right]$, porque en esos valores de w_u habría solo observaciones del tratamiento o del control.

- En otras palabras, los supuestos 1 y 2 implican que dos unidades de diferentes grupos pero con las mismas variables observables deberían tener el mismo resultado potencial esperado: $E[Y(s) \mid s_T(u) = 1, W] = E[Y(s) \mid s_T(u) = 0, W]$, $s = 1, 0$.

Regresión Lineal Simple y Múltiple

- Estamos interesados en el coeficiente β_1 en las siguientes ecuaciones:

$$y_u = \beta_0 + \beta_1 s_T(u) + \epsilon_u \tag{16}$$

$$y_u = \beta_0 + \beta_1 s_T(u) + w_u\gamma + v_u \tag{17}$$

- En la literatura la ecuación (16) recibe el nombre de **modelo de regresión lineal simple (MRLS)** y la ecuación (17) se llama **modelo de regresión lineal múltiple (MRLM)**.
- y_u se conoce como variable dependiente o variable de resultado y $S_T(u)$ y w_u se denominan regresores o variables explicativas.
- La única diferencia entre los dos modelos está en el número de variables explicativas.

- La primera parte de la ecuación (16) [(17)], $\beta_0 + \beta_1 s_T(u)$ $[\beta_0 + \beta_1 s_T(u) + w_u\gamma]$, se denomina función de regresión poblacional.
- En un MRLS la función de regresión poblacional es una recta y en un MRLM un plano.
- La segunda parte de la ecuación ϵ_u $[v_u]$ se conoce como la parte aleatoria del modelo.
- La ordenada al origen β_0 y la pendiente de la recta β_1 [las pendientes de cada lado del plano β_1, γ] son los parámetros poblacionales.

Agenda

Modelo de Regresión Lineal Múltiple

- En economía un modelo vincula a una variable que se pretende explicar (y) con una o varias variables que la explican (x's).
- La variable explicada recibe el nombre de variable dependiente.
- Las variables que explican a la variable dependiente reciben el nombre de variables explicativas o regresores.
- Suponga que observamos n valores de estas variables. Sea y_i la i-ésima observación de la variable dependiente y sea $(x_{i1}, x_{i2}, \ldots, x_{ik})$ la i-ésima observación de los k regresores.
- El modelo se especifica a través de un conjunto de supuestos acerca del comportamiento de la variable dependiente y de las variables explicativas.

- **Supuesto 1 (linealidad)**

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad (i = 1, 2, \ldots, n). \tag{18}$$

donde los β's son parámetros desconocidos que deben ser estimados, y ϵ_i es un componente de error no observado con ciertas propiedades que se especifican más abajo.

- La parte de las variables explicativas en el modelo anterior, $(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$, se denomina función de regresión poblacional o regresión directamente.
- Los parámetros o coeficientes de la regresión representan los efectos marginales de los regresores. Por ejemplo, β_2 representa el cambio en la variable dependiente cuando el segundo regresor se incrementa en una unidad mientras que los otros regresores se mantienen constantes. En términos matemáticos,

$$\frac{\partial y_i}{\partial x_{i2}} = \beta_2.$$

- El supuesto de linealidad implica que el efecto marginal no depende del nivel de los regresores.
- El modelo en notación matricial. Definamos dos vectores columna k-dimensionales x_i' y β como:

$$x_i' = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}. \tag{19}$$

- Por definición: $x_i\beta = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$. Es decir que el modelo (18) se puede escribir como,

$$y_i = x_i\beta + \epsilon_i. \tag{20}$$

Modelo de Regresión Lineal Múltiple

- Ahora definamos dos vectores columna n-dimensionales,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{21}$$

- y la matriz $n \times k$,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}. \tag{22}$$

- Entonces, el supuesto de linealidad puede ser escrito como

$$y = x\beta + \epsilon, \tag{23}$$

- **Supuesto 2 (exogeneidad estricta)**

$$E(\epsilon_i|x) = 0, \quad i = 1, 2, \ldots, n. \tag{24}$$

- Note que la esperanza esta condicionada por los regresores para todas las observaciones.
- El supuesto de exogeneidad estricta tiene varias implicancias.

Modelo de Regresión Lineal Múltiple

- Usando la ley de expectativas iteradas, la exogeneidad estricta implica que la esperanza no condicionada del error es cero.

$$E(\epsilon_i) = 0, \forall i$$

- Bajo exogeneidad estricta, los regresores son ortogonales al error para todas las observaciones,

$$E(x_{jk}\epsilon_i) = 0, \ \ \forall\, i,\, j,\, \&\, k$$

- La característica de la exogeneidad estricta es el requerimiento de que los regresores sean ortogonales no solo al error para la misma observación $[E(x_{ik}\epsilon_i) = 0 \ \ \forall\, k]$, sino también a los errores de otras observaciones $[E(x_{jk}\epsilon_i) = 0 \ \forall\, k,\, \&\, i \neq j]$.

- **Supuesto 3 (no singularidad)**
  El rango de la matriz $k \times k$, $E(x'x)$ es k con probabilidad 1.
- Como $E(x'x)$ es una matriz simétrica de dimensión $k \times k$, el Supuesto 3 es equivalente a asumir que $E(x'x)$ es positiva definida. El supuesto puede establecerse directamente en términos del análogo muestral: $x'x/n$.
- **Supuesto 4 (varianza del error esférica)**
  Homocedasticidad

$$E(\epsilon_i^2|x) = \sigma^2 > 0 \quad i = 1, 2, \ldots, n \tag{25}$$

El supuesto de homocedasticidad dice que el segundo momento condicional, que en general es una función no lineal de x, es una constante.

- Este supuesto puede ser escrito en términos más familiares como,

$$\begin{aligned} Var(\epsilon_i|x) &= E(\epsilon_i^2|x) - E(\epsilon_i|x)^2 \text{ por definición de varianza} \\ &= E(\epsilon_i^2|x) = \sigma^2 > 0 \text{ por exogeneidad estricta.} \end{aligned} \quad (26)$$

- Usando la ley de expectativas iteradas,

$$E[E(\epsilon_i^2|x)] = E(\epsilon_i^2) = \sigma^2 > 0 \;\; i = 1, 2, \ldots, n \quad (27)$$

y como $Var(\epsilon_i) = E(\epsilon_i^2) - E(\epsilon_i)^2 = E(\epsilon_i^2) = \sigma^2 > 0$ el supuesto de homocedasticidad nos dice que la varianza condicional del error es igual a la varianza no condicionada que es constante.

- Ausencia de correlación serial

$$E(\epsilon_i \epsilon_j | x) = 0, \ \ \forall i \neq j \tag{28}$$

- Igual que el supuesto de homocedasticidad, el supuesto de ausenciac de correlación serial es equivalente a requerir que:

$$Cov(\epsilon_i, \epsilon_j | x) = 0, \ \ \forall i \neq j, \tag{29}$$

  la covarianza condicional de los errores del modelo es cero.

- Combinando los dos requerimientos, el supuesto 4 de varianza de los errores esférica establece que,

$$E(\epsilon \epsilon' | x) = \sigma^2 I_n. \tag{30}$$

- La función de regresión poblacional, $x\beta$, es el mejor predictor lineal (BLP) de la variable dependiente porque tiene el menor error medio cuadrático de predicción (MSPE) entre todos los predictores.
- Definamos el MSPE como,

$$S(\beta) = E[(y - x\beta)'(y - x\beta)] = E(y'y) - 2\beta' E(x'y) + \beta' E(x'x)\beta \tag{31}$$

- El BLP de y dado x, denotado por $\mathcal{P}(y|x)$, se encuentra minimizando $S(\beta)$ con respecto a β.
- Las condiciones de primer orden para la minimización son:

$$\frac{\partial S(\beta)}{\partial \beta'} = -2E(x'y) + 2E(x'x)\beta = 0. \tag{32}$$

- Por lo tanto:

$$\beta = \left[E(x'x)\right]^{-1} E(x'y) \tag{33}$$

- Entonces la proyección lineal de y dado x es

$$\mathcal{P}(y|x) = x\left[E(x'x)\right]^{-1} E(x'y) \tag{34}$$

- $\epsilon = y - x\beta$ es el error de predicción y satisface $E(x'\epsilon) = 0$.
- Es decir,

$$E[x'\epsilon] = E[x'(y - x\beta)] = E(x'y) - E(x'x)\left[E(x'x)\right]^{-1} E(x'y) = 0 \tag{35}$$

- Considere dividir x en una variable x_1 y el resto de las variables, x_2, esto es $x = [x_1, x_2]$.
- x_1 tiene dimensión $n \times 1$ y x_2 es de dimensión $n \times (k-1)$.
- El MRLM quedaría:

$$y = x_1\beta_1 + x_2\beta_2 + \epsilon, \text{ con } \mathbb{E}(x'\epsilon) = 0. \tag{36}$$

- Particione $\mathbb{E}(x'x)$ como

$$\mathbb{E}(x'x) = \begin{bmatrix} \mathbb{E}(x_1'x_1) & \mathbb{E}(x_1'x_2) \\ \mathbb{E}(x_2'x_1) & \mathbb{E}(x_2'x_2) \end{bmatrix} \tag{37}$$

- Similarmente

$$\mathbb{E}(x'y) = \begin{bmatrix} \mathbb{E}(x_1'y) \\ \mathbb{E}(x_2'y) \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_{1y} \\ \boldsymbol{Q}_{2y} \end{bmatrix} \tag{38}$$

- Usando la fórmula de la inversa de la matriz particionada tenemos

$$\begin{bmatrix} \mathbb{E}(x_1'x_1) & \mathbb{E}(x_1'x_2) \\ \mathbb{E}(x_2'x_1) & \mathbb{E}(x_2'x_2) \end{bmatrix}^{-1} \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{Q}^{11} & \boldsymbol{Q}^{12} \\ \boldsymbol{Q}^{21} & \boldsymbol{Q}^{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1} & -\boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot 1}^{-1} \end{bmatrix} \tag{39}$$

donde $\boldsymbol{Q}_{11\cdot 2} \overset{\text{def}}{=} \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$ y $\boldsymbol{Q}_{22\cdot 1} \overset{\text{def}}{=} \boldsymbol{Q}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$.

Modelo de Regresión Lineal Múltiple

- Por lo tanto

$$
\begin{aligned}
\boldsymbol{\beta} &= \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \\
&= \begin{bmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1} & -\boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot 1}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{Q}_{1y} \\ \boldsymbol{Q}_{2y} \end{bmatrix} \\
&= \begin{pmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1}\left(\boldsymbol{Q}_{1y} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{2y}\right) \\ \boldsymbol{Q}_{22\cdot 1}^{-1}\left(\boldsymbol{Q}_{2y} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{1y}\right) \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{1y\cdot 2} \\ \boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{2y\cdot 1} \end{pmatrix}
\end{aligned}
$$

Modelo de Regresión Lineal Múltiple

- Suponga la siguiente proyección lineal

$$y = x_1\gamma_1 + u, \quad \text{con} \quad E(x_1'u) = 0, \tag{40}$$

  donde cambiamos el coeficiente de x_1 a γ_1 en lugar de β_1 y el error a u en lugar de ϵ.
- Típicamente $\beta_1 \neq \gamma_1$ salvo casos especiales.

- Calculemos γ_1

$$\begin{aligned}
\gamma_1 &= \left[\mathbb{E}\left(x_1' x_1\right)\right]^{-1} \mathbb{E}\left(x_1' y\right) \\
&= \left[\mathbb{E}\left(x_1' x_1\right)\right]^{-1} \mathbb{E}\left(x_1' \left(x_1 \beta_1 + x_2 \beta_2 + e\right)\right) \\
&= \beta_1 + \left[\mathbb{E}\left(x_1' x_1\right)\right]^{-1} \mathbb{E}\left(x_1' x_2\right) \beta_2 \\
&= \beta_1 + \Gamma \beta_2
\end{aligned}$$

  donde $\Gamma = [\mathbb{E}\left(x_1' x_1\right)]^{-1} \mathbb{E}\left(x_1' x_2\right)$ es el vector de coeficientes de la proyección de x_2 dado x_1.

- En general $\gamma_1 = \beta_1 + \Gamma \beta_2 \neq \beta_1$ salvo que $\Gamma = 0$ lo que significa que la proyección de x_2 sobre x_1 tiene coeficientes iguales a cero o bien $\beta_2 = 0$ lo que implica que en (36) las variables en x_2 no explican y.
- $\Gamma \beta_2$ se conoce en la literatura como el **sesgo por variables omitidas**. Es la consecuencia de omitir una variable estadísticamente relevante.

- Considere ahora la proyección de x_1 dado x_2: $x_1 = x_2\gamma_2 + u_1, \quad$ con $E(x_2'u_1) = 0$

$$\begin{aligned}
\mathbb{E}(u_1'u_1) &= \mathbb{E}[(x_1 - x_2\gamma_2)'(x_1 - x_2\gamma_2)] = \mathbb{E}[x_1'(x_1 - x_2\gamma_2)] - \gamma_2'\mathbb{E}[x_2'(x_1 - x_2\gamma_2)] \\
&= \mathbb{E}[x_1'(x_1 - x_2\gamma_2)] - \gamma_2'\mathbb{E}[x_2'u_1] = \mathbb{E}[x_1'(x_1 - x_2\gamma_2)] \\
&= \mathbb{E}[x_1'x_1] - \mathbb{E}[x_1'x_2]\gamma_2 = \mathbb{E}[x_1'x_1] - \mathbb{E}[x_1'x_2]\left[\mathbb{E}\left(x_2'x_2\right)\right]^{-1}\mathbb{E}\left(x_2'x_1\right) \\
&= \boldsymbol{Q}_{11\cdot 2} = \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} \qquad (41)
\end{aligned}$$

- Además

$$\begin{aligned}
\mathbb{E}(u_1'y) &= \mathbb{E}[(x_1 - x_2\gamma_2)'y] = \mathbb{E}(x_1'y) - \gamma_2'\mathbb{E}(x_2'y) \\
&= \mathbb{E}(x_1'y) - \mathbb{E}(x_1'x_2)\left[\mathbb{E}\left(x_2'x_2\right)\right]^{-1}\mathbb{E}\left(x_2'y\right) \\
&= \boldsymbol{Q}_{1y} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{2y} = \boldsymbol{Q}_{1y\cdot 2} \qquad (42)
\end{aligned}$$

- Es decir
$$\beta_1 = \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{1y\cdot 2} = \left[\mathbb{E}(u_1' u_1)\right]^{-1}\mathbb{E}\left(u_1' y\right) \tag{43}$$
- Se puede recuperar el coeficiente de x_1 en el MRLM de una regresión de y sobre u_1.
- El coeficiente β_1 es el coeficiente de la proyección de y sobre u_1, donde u_1 es el error de la proyección de x_1 sobre el resto de los coeficientes x_2.
- u_1 puede interpretarse como el componente de x_1 que no está explicado por el resto de las variables.
- El coeficiente β_1 es el efecto lineal de x_1 (una vez que le sacamos el efecto del resto de las variables) sobre y.

Agenda

Mínimos Cuadrados Clásicos

- En la práctica no observamos a todas las unidades de la población sino que trabajamos con un subconjunto de las mismas: una muestra.
- En este contexto la esperanza matemática condicional de la variable dependiente debe estimarse.
- El método tradicional de estimación se denomina Mínimos Cuadrados Clásicos (MCC).

Agenda

Mínimos Cuadrados Clásicos

- El método de los mínimos cuadrados clásicos (MCC u OLS) consiste en estimar el valor de y con la regresión muestral $x\hat{\beta}$.
- La diferencia $y - x\hat{\beta} = e$ son los residuos del modelo.
- Técnicamente, el método de MCC consiste en minimizar la suma de residuos al cuadrado (RSS),

$$\begin{aligned} RSS(\hat{\beta}) &= (y - x\hat{\beta})'(y - x\hat{\beta}) \\ &= y'y - \hat{\beta}'x'y - y'x\hat{\beta} + \hat{\beta}'x'x\hat{\beta} \\ &= y'y - 2\hat{\beta}'x'y + \hat{\beta}'x'x\hat{\beta} \end{aligned} \tag{44}$$

- Las condiciones de primer orden para la minimización son:

$$\frac{\partial RSS(\hat{\beta})}{\partial \hat{\beta}'} = -2x'y + 2x'x\hat{\beta} = 0 \tag{45}$$

Mínimos Cuadrados Clásicos

- Las condiciones de primer orden dan origen a las denominadas ecuaciones normales,

$$x'x\hat{\beta} = x'y \tag{46}$$

- Usando el supuesto 3, las ecuaciones normales implican que,

$$\hat{\beta} = (x'x)^{-1}x'y. \tag{47}$$

  es el estimador de MCC.

- Reordenando (46) tenemos,

$$x'y - x'x\hat{\beta} = x'(y - x\hat{\beta}) = x'e = 0 \tag{48}$$

Mínimos Cuadrados Clásicos

- Las condiciones de segundo orden son:

$$\frac{\partial^2 RSS(\hat{\beta})}{\partial \hat{\beta}^2} = 2x'x > 0. \quad (49)$$

La última desigualdad surge del supuesto 3.

- <u>Definiciones</u>

- El valor estimado para la observación i se define como, $\hat{y}_i \equiv x_i \hat{\beta}$. El vector de valores estimados $\hat{y}$ es igual a $x\hat{\beta}$. Por lo tanto el vector de residuos es $e = y - \hat{y}$.
- La estimación mínimo cuadrática de σ^2 (la varianza del término de error), se denota s^2 y es la suma de los residuos al cuadrado dividida por $n - k$:

$$s^2 = \frac{RSS(\hat{\beta})}{n-k} = \frac{e'e}{n-k}$$

- La raiz cuadrada de s^2, s, se denomina error estándar de la regresión.
- El error de estimación se define como $\hat{\beta} - \beta$ y esta dado por la siguiente expresión,

$$
\begin{aligned}
\hat{\beta} - \beta &= (x'x)^{-1}x'y - \beta \\
&= (x'x)^{-1}x'(x\beta + \epsilon) - \beta \\
&= \beta + (x'x)^{-1}x'\epsilon - \beta \\
&= (x'x)^{-1}x'\epsilon
\end{aligned}
\tag{50}
$$

Agenda

Mínimos Cuadrados Clásicos

- Ausencia de Sesgo: Bajo los supuestos 1 a 3, $E(\hat{\beta}|x) = \beta$.
- Bajo los supuestos 1 a 4, $Var(\hat{\beta}|x) = \sigma^2(x'x)^{-1}$.
- Teorema de Gauss-Markov Bajo los supuestos 1 a 4, el estimador de MCC de β es eficiente dentro de la clase de estimadores lineales insesgados. Esto es, para cualquier estimador insesgado lineal en y,

$$Var(\hat{\beta}|x) \leq Var(\tilde{\beta}|x)$$

- Bajo los supuestos 1 a 4, $Cov(\hat{\beta}, e|x) = 0$.
- Si s^2 es una estimación de σ^2, entonces

$$\widehat{Var(\hat{\beta}|x)} = s^2(x'x)^{-1}$$

Agenda

Teoría Asintótica: Convergencia en Probabilidad

1. Una secuencia de variables aleatorias $\{x_N : N = 1, 2, \dots\}$ converge en probabilidad a la constante a si para todo $\epsilon > 0$,

$$P[|x_N - a| > \epsilon] \longrightarrow 0 \text{ cuando } N \longrightarrow \infty$$

- En general, escribimos $x_N \xrightarrow{p} a$ y decimos que a es el plímite de x_N.

2. En el caso especial en que $a = 0$, también decimos que $\{x_N\}$ es $o_p(1)$ (o pequeña p 1). En este caso escribimos $x_N = o_p(1)$ ó $x_N \xrightarrow{p} 0$.
3. Una secuencia de variables aleatorias $\{x_N\}$ está limitada en probabilidad (bounded in probability) sí y solo sí para cada $\epsilon > 0$, existe un $b_\epsilon < \infty$ y un entero N_ϵ, tal que $P[|x_N| \geq b_\epsilon] < \epsilon$ para todo $N \geq N_\epsilon$.

- En este caso escribimos $x_N = O_p(1)$ ($\{x_N\}$ es o grande p 1).

- Lema 1: si $x_N \xrightarrow{p} a$, entonces $x_N = O_p(1)$.

4. Una secuencia aleatoria $\{x_N : N = 1, 2, \ldots\}$ es $o_p(N^\delta)$ para $\delta \in \mathbb{R}$, si $N^{-\delta} x_N = o_p(1)$.

- Lema 2: si $w_N = o_p(1)$, $x_N = o_p(1)$, $y_N = O_p(1)$, y $z_N = O_p(1)$, entonces
  (i) $w_N + x_N = o_p(1)$;
  (ii) $y_N + z_N = O_p(1)$;
  (iii) $y_N \times z_N = O_p(1)$;
  (iv) $x_N \times z_N = o_p(1)$.
- Todas las definiciones anteriores se aplican elemento por elemento a secuencias de vectores y matrices.
- Lema 3: Sea $\{Z_N : N = 1, 2, \ldots\}$ una secuencia de matrices $J \times K$ tal que $Z_N = o_p(1)$, y sea $\{x_N\}$ una secuencia de vectores aleatorios $J \times 1$ tal que $x_N = O_p(1)$. Entonces $Z_N' x_N = o_p(1)$.

- Lema 4 (Teorema de Slutsky): Sea $g : \mathbb{R}^K \to \mathbb{R}^J$ una función contínua en algun punto $c \in \mathbb{R}^K$. Sea $\{x_N : N = 1, 2, \ldots\}$ una secuencia de vectores aleatorios $K \times 1$ tal que $x_N \xrightarrow{p} c$. Entonces $g(x_N) \xrightarrow{p} g(c)$ cuando $N \longrightarrow \infty$. En otras palabras: $plim\, g(x_N) = g(plim\, x_N)$ si $g(\cdot)$ es contínua en $plim\, x_N$.
- Definición 1: Sea $(\Omega, \mathfrak{F}, P)$ un espacio de probabilidad. Una secuencia de eventos $\{\Omega_N : N = 1, 2, \ldots\} \subset \mathfrak{F}$ se dice que ocurre con probabilidad aproximándose a uno (w.p.a 1) sí y solo sí $P(\Omega_N) \longrightarrow 1$ cuando $N \longrightarrow \infty$.
- Corolario 1: Sea $\{Z_N : N = 1, 2, \ldots\}$ una secuencia de matrices aleatorias $K \times K$, y sea A una matriz invertible no aleatoria $K \times K$. Si $Z_N \xrightarrow{p} A$ entonces:
  (1) Z_N^{-1} existe w.p.a 1
  (2) $Z_N^{-1} \xrightarrow{p} A^{-1}$

Teoría Asintótica: Convergencia en Distribución

- Definición 2: Una secuencia de variables aleatorias $\{x_N : N = 1, 2, \dots\}$ converge en distribución a la variable aleatoria contínua x sí y solo sí $F_N(\xi) \longrightarrow F(\xi)$ cuando $N \longrightarrow \infty$ para todo $\xi \in \mathbb{R}$.
- Donde F_N es la función de distribución acumulada de x_N y F es la función de distribución acumulada de x. En este caso escribimos: $x_N \xrightarrow{d} x$.
- Definición 3: Una secuencia de vectores aleatorios $\{x_N : N = 1, 2, \dots\}$ $K \times 1$ converge en distribución al vector aleatorio contínuo x sí y solo sí para cualquier vector no aleatorio $K \times 1$, c tal que $c'c = 1$, $c'x_N \xrightarrow{d} c'x$ y escribimos $x_N \xrightarrow{d} x$.
- Lema 5: Si $x_N \xrightarrow{d} x$, donde x es cualquier vector aleatorio $K \times 1$, entonces $x_N = O_p(1)$.

- Lema 6 (Continuous mapping theorem): Sea $\{x_N\}$ una secuencia de vectores aleatorios de dimensión $K \times 1$, tal que $x_N \xrightarrow{d} x$. Si $g : \mathbb{R}^K \rightarrow \mathbb{R}^J$ es una función contínua, entonces $g(x_N) \xrightarrow{d} g(x)$.
- Corolario 2: Si $\{z_N\}$ es una secuencia de vectores aleatorios de dimensión $K \times 1$, tal que $z_N \xrightarrow{d} \text{Normal}(0, V)$
  Entonces:
  (1) Para cualquier matriz no aleatoria A, $(K \times M)$ $A' z_N \xrightarrow{d} \text{Normal}(0, A'VA)$.
  (2) $z_N' V^{-1} z_N \xrightarrow{d} \chi^2_K$.
- Lema 7: Sean $\{x_N\}$ y $\{z_N\}$ secuencias de vectores aleatorios de dimensión $K \times 1$. Si $z_N \xrightarrow{d} z$ y $x_N - z_N \xrightarrow{p} 0$, entonces $x_N \xrightarrow{d} z$.

Teoría Asintótica: Convergencia en Distribución

- Teorema 1: Sea $\{w_j : j = 1, 2, \ldots\}$ una secuencia de vectores aleatorios $G \times 1$, independientes, idénticamente distribuidos tal que $E(|w_{jg}|) < \infty$, $g = 1, 2, \ldots, G$. Entonces la secuencia satisface la ley débil de los grandes números (WLLN): $N^{-1}\sum_{j=1}^{N} w_j \xrightarrow{p} \mu_w$, donde $\mu_w = E(w_j)$.
- Teorema 2 (Lindeberg-Levy): $\{w_j : j = 1, 2, \ldots\}$ una secuencia de vectores aleatorios $G \times 1$, independientes, idénticamente distribuidos tal que $E(w_{jg}^2) < \infty$, $g = 1, 2, \ldots, G$ y $E(w_j) = 0$. Entonces la secuencia satisface el teorema central del límite (CLT): $N^{-1/2}\sum_{j=1}^{N} w_j \xrightarrow{d}$ Normal$(0, B)$, donde $B = Var(w_j) = E(w_j w_j')$.

- De (47) tenemos,

$$\begin{aligned}\hat{\beta} &= (x'x)^{-1}x'y = \left(n^{-1}\sum_{i=1}^{n} x_i'x_i\right)^{-1}\left(n^{-1}\sum_{i=1}^{n} x_i'y_i\right)\\ &= \beta + \left(n^{-1}\sum_{i=1}^{n} x_i'x_i\right)^{-1}\left(n^{-1}\sum_{i=1}^{n} x_i'\epsilon_i\right)\end{aligned}$$

- Bajo el Supuesto 3, $x'x$ no es singular con probabilidad uno y por el Corolario 1,

$$\text{plim}\left(n^{-1}\sum_{i=1}^{n} x_i'x_i\right)^{-1} = A^{-1}, \quad \text{con } A \equiv E(x'x).$$

Mínimos Cuadrados Clásicos

- Usando la WLLN y el supuesto de exogeneidad estricta,

$$\text{plim}\left(n^{-1}\sum_{i=1}^{n} x_i'\epsilon_i\right) = E(x'\epsilon) = 0.$$

- Usando el teorema de Slutzky,

$$\text{plim}\hat{\beta} = \beta + A^{-1} \times E(x'\epsilon) = \beta.$$

- Note que en realidad para demostrar consistencia del estimador de MCC no es necesario asumir exogeneidad estricta. Uno podría relajar este supuesto a la condición de ortogonalidad poblacional: $E(x'\epsilon) = 0$.
- Sin embargo, bajo este nuevo supuesto el estimador de MCC no es necesariamente insesgado.

Mínimos Cuadrados Clásicos

- La distribución asintótica del estimador de MCC viene dada desde:

$$
\begin{aligned}
\hat{\beta} &= \beta + \left(n^{-1} \sum_{i=1}^{n} x_i' x_i \right)^{-1} \left(n^{-1} \sum_{i=1}^{n} x_i' \epsilon_i \right) \\
\sqrt{n}(\hat{\beta} - \beta) &= \left(n^{-1} \sum_{i=1}^{n} x_i' x_i \right)^{-1} \left(n^{-1/2} \sum_{i=1}^{n} x_i' \epsilon_i \right)
\end{aligned}
$$

- Bajo el supuesto de exogeneidad estricta $\{(x_i'\epsilon_i) : i = 1, 2, \ldots\}$ es una secuencia de variables i.i.d. con media igual a cero.
- Bajo el supuesto de varianza de los errores esférica la secuencia de variables anteriores tiene varianza finita.

- Por lo tanto, aplicando el CLT tenemos,

$$\left(n^{-1/2}\sum_{i=1}^{n} x_i'\epsilon_i\right) \xrightarrow{d} \text{Normal}(0,\ \sigma^2 A)$$

- De la demostración de consistencia sabemos que,

$$\left(n^{-1}\sum_{i=1}^{n} x_i' x_i\right)^{-1} - A^{-1} = o_p(1)$$

- Por lo tanto,

$$\sqrt{n}(\hat{\beta} - \beta) = A^{-1}\left(n^{-1/2}\sum_{i=1}^{n} x_i'\epsilon_i\right) + o_p(1)$$

Mínimos Cuadrados Clásicos

- De estas ecuaciones surge que

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \text{Normal}(0, \, \sigma^2 A^{-1})$$

Note que, usando el análogo muestral, la varianza asintótica de $\hat{\beta}$ coincide con la varianza condicional que encontramos antes.

- Considere ahora el modelo particionado

$$y = x\beta + \epsilon = [x_1 \ x_2] \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \epsilon \tag{51}$$

- **Teorema de Frish-Waugh-Lovell:** El estimador de MCC de β_1, $\hat{\beta}_1$, se puede obtener desde la ecuación anterior o mediante los siguientes pasos:
  1. Regrese y sobre x_2 y obtenga los residuos $\hat{u}_1$
  2. Regrese x_1 sobre x_2 y obtenga los residuos $\tilde{x}_1$
  3. Regrese $\hat{u}_1$ sobre $\tilde{x}_1$ y obtenga $\hat{\beta}_1$

- Reemplace en las ecuaciones (41) y (42) las esperanzas matemáticas por sus análogos muestrales para obtener:

$$\begin{aligned} \hat{u}_1'\hat{u}_1 &= (x_1'x_1) - (x_1'x_2)\left[(x_2'x_2)\right]^{-1}(x_2'x_1) \\ &= x_1'\left[I - x_2\left[(x_2'x_2)\right]^{-1}x_2'\right]x_1 = x_1'M_2x_1 \end{aligned} \tag{52}$$

- y

$$\begin{aligned} \hat{u}_1'y &= (x_1'y) - (x_1'x_2)\left[(x_2'x_2)\right]^{-1}(x_2'y) \\ &= x_1'\left[I - x_2\left[(x_2'x_2)\right]^{-1}x_2'\right]y = x_1'M_2y \end{aligned} \tag{53}$$

- Entonces

$$\begin{aligned} \hat{\beta}_1 &= \left[x_1'M_2x_1\right]^{-1}y'M_2x_1 = \left[x_1'M_2M_2x_1\right]^{-1}x_1'M_2M_2y \\ &= \left[\tilde{x}_1'\tilde{x}_1\right]^{-1}\tilde{x}_1'\hat{u}_1 \end{aligned} \tag{54}$$

- Medidas de Bondad del Ajuste
  Una medida de la variabilidad de la variable dependiente es la denominada suma de cuadrados totales ($TSS = y'y$). Esta suma de cuadrados puede descomponerse de la siguiente forma,

$$\begin{aligned} y'y &= (\hat{y}+e)'(\hat{y}+e) \\ &= \hat{y}'\hat{y}+e'\hat{y}+\hat{y}'e+e'e \\ &= \hat{y}'\hat{y}+2\hat{\beta}'x'e+e'e \\ &= \hat{y}'\hat{y}+e'e \qquad (55) \end{aligned}$$

- El R^2 no centrado se define como,

$$R_{nc}^2 = 1 - \frac{e'e}{y'y} = \frac{\hat{y}'\hat{y}}{y'y} \tag{56}$$

- Tanto $\hat{y}'\hat{y}$ como $e'e$ son expresiones no negativas $0 \leq R_{nc}^2 \leq 1$. $\hat{y}'\hat{y}$ recibe el nombre de suma de cuadrados explicados por la regresión (ESS).
- El R_{nc}^2 tiene la interpretación de ser la fracción de la variabilidad de la variable dependiente que es atribuible a la variación en las variables explicativas.

Mínimos Cuadrados Clásicos

- Coeficiente de determinación o R^2 centrado

  Si la regresión incluye un término constante, entonces la suma de cuadrados totales se define como $TSS = (y - \bar{y})'(y - \bar{y})$ y la suma de cuadrados explicados por la regresión se define como $ESS = (\hat{y} - \bar{y})'(\hat{y} - \bar{y})$.

- El coeficiente de determinación o R^2 centrado se define como,

$$R_c^2 = 1 - \frac{e'e}{(y - \bar{y})'(y - \bar{y})} = \frac{(\hat{y} - \bar{y})'(\hat{y} - \bar{y})}{(y - \bar{y})'(y - \bar{y})} \tag{57}$$

Agenda

Mínimos Cuadrados Clásicos: inferencia

- Supuesto 5 (normalidad del término de error)
  La distribución de ϵ condicionada en x es normal.
- Los supuestos 2, 4 y 5 tomados en conjunto implican que la distribución de ϵ condicionada en x es
$$\epsilon | x \sim \text{Normal}(0, \sigma^2 I_n) \tag{58}$$
- Como la distribución de ϵ condicionada en x no depende de x, se sigue que ϵ y x son independientes.
- Una consecuencia del punto anterior es que la distribución marginal de ϵ es $N(0, \sigma^2 I_n)$.
- Recuerde que de (50) sabemos que el error de estimación, $\hat{\beta} - \beta$, es lineal en ϵ. Como ϵ es normal, entonces el error de estimación también tiene distribución normal.

MCC: inferencia sobre coeficientes individuales

- Entonces,
$$\hat{\beta} - \beta \sim \text{Normal}(0, \sigma^2(x'x)^{-1}) \tag{59}$$
- Contrastes de Hipótesis sobre Coeficientes Individuales
La hipótesis nula a considerar es del tipo $H_0 : \beta_k = \bar{\beta}_k$, donde $\bar{\beta}_k$ es algún valor conocido.
- La hipótesis alternativa es del tipo, $H_1 : \beta_k \neq \bar{\beta}_k$.
- Usando (59) e imponiendo la hipótesis nula tenemos:
$$\hat{\beta}_k - \beta_k \sim \text{Normal}(0, \sigma^2[(x'x)^{-1}]_{kk})$$
donde $[(x'x)^{-1}]_{kk}$ es el elemento de la diagonal principal de $(x'x)^{-1}$ ubicado en la fila k y la columna k.

MCC: inferencia sobre coeficientes individuales

- Bajo la hipótesis nula, el estadístico de contraste,

$$z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2[(x'x)^{-1}]_{kk}}} \tag{60}$$

  tiene distribución Normal Estándar.
- Utilizando este estadístico de contraste z_k, podemos determinar si el error muestral $\hat{\beta}_k - \beta_k$ es demasiado grande como para ocurrir por casualidad.
- El concepto de "muy grande" se establece probabilísticamente si el estadístico de constraste toma un valor que no es representativo para una realización proveniente de la distribución de z_k.
- Si σ^2 es desconocido, una idea natural es reemplazarlo por su estimador mínimo cuadrático s^2.

MCC: inferencia sobre coeficientes individuales

- El estadístico de contraste, después de reemplazar σ^2 por s^2, se denomina estadístico-t.
- El denominador del nuevo estadístico se llama error estándar del estimador mínimo cuadrático de $\hat{\beta}_k$,

$$
\begin{aligned}
se(\hat{\beta}_k) &= \sqrt{s^2[(x'x)^{-1}]_{kk}} \\
&= \sqrt{\text{elemento } (k,\ k) \text{ de } \widehat{Var(\hat{\beta}_k)}}
\end{aligned}
$$

- Como s^2 es una variable aleatoria, su inclusión en el estadístico de contraste cambia la distribución muestral del mismo.

MCC: inferencia sobre coeficientes individuales

- <u>Distribución del Estadístico-t</u>. Supongamos que los supuestos 1 a 5 se cumplen. Entonces, bajo la hipótesis nula el estadístico-t definido como

$$t_k \equiv \frac{\hat{\beta}_k - \beta_k}{\sqrt{s^2[(x'x)^{-1}]_{kk}}}, \tag{61}$$

  se distribuye como una $t(n-k)$ (distribución t con $n-k$ grados de libertad).

- El contraste de hipótesis basado en el estadístico-t se denomina test-t y se resuelve como sigue

Paso 1 Dado el valor hipotetizado, $\bar{\beta}_k$, de β_k, construya el estadístico-t como en (61). Un desvío muy grande de t_k de cero es un signo de la falla de la hipóótesis nula. El siguiente paso especifica cuán grande es muy grande.

Paso 2 Encuentre el valor crítico, $t_{\alpha/2}(n-k)$, en la distribución $t(n-k)$ como el intervalo alrededor de cero que deja $\alpha/2\,\%$ de cada lado de la distribución. Esto es,

$$Prob(-t_{\alpha/2}(n-k) < t < t_{\alpha/2}(n-k)) = 1-\alpha$$

Paso 3 Regla de Decisión: Rechace H_0 si $t_k < -t_{\alpha/2}(n-k)$ ó $t_k > t_{\alpha/2}(n-k)$.

- p-Valor. La regla de decisión del test-t también puede ser formulada en términos del p-valor (p).

Paso 2 Calcule $p = Prob(t > |t_k|)$

Paso 3 Regla de Decisión: Rechace H_0 si $p < \alpha$.

MCC: inferencia sobre combinación lineal de coeficientes

- Hipótesis Lineales: Supongamos que queremos contrastar hipótesis que son combinaciones lineales de coeficientes. Estas hipótesis pueden escribirse como,

$$H_0 : R\beta = r \tag{62}$$

donde los valores de R y de r son conocidos. Note que (62) puede ser un sistema de ecuaciones.

- Denominamos al número de ecuaciones, que es la dimensión de r, por $\#r$. Por lo tanto R es de dimensión $\#r \times k$.
- Para contrastar estas hipótesis utilizamos un test-F ó **test de Wald**. El estadístico de contraste se define como,

$$\begin{aligned} W &= \frac{(R\hat{\beta} - r)'[R(x'x)^{-1}R']^{-1}(R\hat{\beta} - r)/\#r}{s^2} \\ &= (R\hat{\beta} - r)'[R\widehat{Var}(\hat{\beta})R']^{-1}(R\hat{\beta} - r)/\#r \end{aligned} \tag{63}$$

MCC: inferencia sobre combinación lineal de coeficientes

- Supongamos que se cumplen los supuestos 1 a 5. Bajo la hipótesis nula (62) el estadístico de contraste, W, definido en la ecuación anterior se distribuye como una $F(\#r, n-k)$ (distribución F con $\#r$ y $n-k$ grados de libertad).
- Si la hipótesis nula es verdadera, deberíamos esperar que $R\hat{\beta} - r$ fuera muy pequeño, por lo que valores muy altos de W deberían ser tomados como evidencia en contra de la hipótesis nula.
- El contraste de hipótesis basado en W se resuelve como sigue,

Paso 1 Calcule el estadístico-W con la fórmula (63).

Paso 2 Encuentre el valor crítico en una distribución F, con $\#r$ grados de libertad en el numerador y $n-k$ grados de libertad en el denominador, que deje una probabilidad de α en la cola superior de la distribución.

Paso 3 Rechace H_0 si el estadístico-W calculado en el Paso 1 es mayor que el valor crítico $F_{\alpha}(\#r, n-k)$.

MCC: inferencia sobre combinación lineal de coeficientes

- p-Valor. La regla de decisión del test de Wald también puede ser formulada en términos del p-valor (p).

Paso 2 Calcule $p = Prob(F > W)$

Paso 3 Regla de Decisión: Rechace H_0 si $p < \alpha$.

- El test de Wald puede reformularse en los siguientes términos:

Paso 1 Establezca dos modelos, el modelo original (denominado no restringido) y un modelo alternativo que cumpla con la hipótesis nula (denominado restringido). Estime los dos modelos y obtenga la suma de residuos al cuadrado (RSS_{nr} y RSS_r).

Paso 2 Construya el estadístico-W con la siguiente fórmula,

$$W = \frac{(RSS_r - RSS_{nr})/\#r}{RSS_{nr}/(n-k)} \tag{64}$$

Paso 3 Encuentre el valor crítico en una distribución F, con $\#r$ grados de libertad en el numerador y $n-k$ grados de libertad en el denominador, que deje una probabilidad de α en la cola superior de la distribución.

Paso 4 Rechace H_0 si W calculado en el Paso 2 es mayor que el valor crítico $F_\alpha(\#r, n-k)$.

- Intuición: Si la hipótesis nula es verdadera, deberíamos esperar que $RSS_r - RSS_{nr}$ fuera muy pequeño, por lo que valores muy altos de W deberían ser tomados como evidencia en contra de la hipótesis nula.
- Que sucede si no se cumple el supuesto de normalidad de los errores?
- Como hemos mostrado, sin el supuesto de normalidad, la distribución de los estimadores de MCC es normal. Por lo tanto, las distribuciones de los estadísticos de contraste que derivamos en forma exacta se verifican asintóticamente.

- Test del Multiplicador de Lagrange (test LM
- Considere el siguiente modelo:

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon, \quad H_0 : \beta_2 = 0 \tag{65}$$

donde β_2 es de dimensión $g \times 1$ y β_1 de dimensión $k - g \times 1$. Asuma que X_1 incluye una columna de unos (es decir, el modelo restringido tiene constante).

- La minimización restringida es entonces

$$\Lambda(\beta_1, \beta_2, \lambda) = (y - X_1\beta_1 - X_2\beta_2)'(y - X_1\beta_1 - X_2\beta_2) + 2\lambda'\beta_2, \tag{66}$$

donde λ es el vector de los g multiplicadores de Lagrange.

- Condiciones de primer orden,

$$\begin{aligned}
\frac{\partial \Lambda}{\partial \beta_1} &= -2X_1'\left(y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2\right) = 0, \\
\frac{\partial \Lambda}{\partial \beta_2} &= -2X_2'\left(y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2\right) + 2\hat{\lambda} = 0 \\
\frac{\partial \Lambda}{\partial \lambda} &= 2\hat{\beta}_2 = 0.
\end{aligned}$$

- Sustituyendo $\hat{\beta}_2 = 0$ de la última ecuación en la primera: $X_1'\left(y - X_1\hat{\beta}_1\right) = 0$
- Esto es $\hat{\beta}_1 = \hat{\beta}_R = (X_1'X_1)^{-1} X_1'y$ es el estimador restringido que surge de regresar y sobre X_1.

MCC: inferencia sobre combinación lineal de coeficientes

- Denotemos por $e_R = y - X_1 \hat{\beta}_R$ a los residuos del modelo restringido.
- Entonces las tres condiciones de primero orden se pueden escribir como:

$$X_1' e_R = 0, \quad \hat{\lambda} = X_2' e_R, \quad \hat{\beta}_2 = 0$$

- Bajo la hipótesis nula de $\beta_2 = 0$ se sigue que

$$e_R = y - X_1 \hat{\beta}_R = M_1 y = M_1 \left(X_1 \beta_1 + \varepsilon \right) = M_1 \varepsilon$$

donde $M_1 - X_1 \left(X'_1 X_1 \right)^{-1} X'_1$.

- Si se cumple el supuesto 5 de normalidad, $\varepsilon \sim N\left(0, \sigma^2 I\right)$ y $e_R \sim N\left(0, \sigma^2 M_1\right)$, entonces

$$\hat{\lambda} = X_2' e_R \sim \mathrm{N}\left(0, \sigma^2 X_2' M_1 X_2\right).$$

- Esto significa que $\hat{\lambda}' \left(X_{22} M_1 X_2\right)^{-1} \hat{\lambda}/\sigma^2$ se distribuye como una $\chi^2(g)$
- Reemplazando la varianza desconocida, σ^2 por una estimación consistente $\hat{\sigma}^2 = \frac{1}{n} e_R' e_R$ se define el estadístico del multiplicador de Lagrange:

$$LM = \hat{\lambda}' \left(X_2' M_1 X_2\right)^{-1} \hat{\lambda}/\hat{\sigma}^2 \sim \chi^2(g) \tag{67}$$

- Note que el estadístico se puede escribir como

$$LM = n \frac{e_R' X_2 \left(X_2' M_1 X_2\right)^{-1} X_2' e_R}{e_R' e_R}. \tag{68}$$

- El estadístico LM se puede recuperar con el siguiente procedimiento
  1. Estime el modelo restringido y genere los residuos
  2. Estime una regresión con los residuos de (1) como variable dependiente y todas las variables explicativas del modelo no restringido
  3. Construya el LM como la multiplicación del número de observaciones por el R^2 de la regresión en (2)

MCC: inferencia sobre combinación lineal de coeficientes

- El modelo restringido tiene constante y como por condición de primer orden $X_1' e_R = 0$ la media de los residuos restringidos e_R es cero.
- La suma de cuadrados totales (TSS) de la regresión del paso (2) es igual a $TSS = \sum_i (e_{Ri} - \bar{e}_R)^2 = \sum_i e_{Ri}^2 = e_R' e_R$
- La regresión de e_R sobre X en el modelo $e_R = X\gamma + \omega$ da $\hat{\gamma} = (X'X)^{-1} X' e_R$ con plano de regresión muestral $\hat{e}_R = X\hat{\gamma} = X(X'X)^{-1} X' e_R$ y como X tiene constante la media de $\hat{e}_R$ es cero.
- Por lo tanto la suma de cuadrados explicados por la regresión es

$$ESS = \hat{e}_R' \hat{e}_R = e_R' X \left(X'X\right)^{-1} X' e_R$$

- Lo único que resta mostrar es que ESS se puede escribir como $e_R' X_2 \left(X_2' M_1 X_2\right)^{-1} X_2' e_R$.

MCC: inferencia sobre combinación lineal de coeficientes

- Note que $X = (X_1 \; X_2)$ y por las condiciones de primer orden

$$X' e_R = \begin{pmatrix} X_1' e_R \\ X_2' e_R \end{pmatrix} = \begin{pmatrix} 0 \\ X_2' e_R \end{pmatrix}$$

- Usando Frish-Waugh-Lovell el estimador de β_2 en el modelo original es $\hat{\beta}_2 = [x_2' M_1 x_2]^{-1} \, x_2' M_1 y$ y su matriz de varianzas y covarianzas es $Var(\hat{\beta}_2) = \sigma^2 \, (X_2' M_1 X_2)^{-1}$.
- Como la matriz de varianzas y covarianzas de $(\hat{\beta}_1 \; \hat{\beta}_2)$ en el modelo original es $\sigma^2 \, (X'X)^{-1}$, $(X_2' M_1 X_2)^{-1}$ es el bloque diagonal, $g \times g$, de abajo de $(X'X)^{-1}$.

MCC: inferencia sobre combinación lineal de coeficientes

- Combinando estos resultados tenemos

$$\begin{aligned} e_R' X \left(X'X\right)^{-1} X' e_R &= \left(0 \;\; e_R' X_2\right) \left(X'X\right)^{-1} \begin{pmatrix} 0 \\ X_2' e_R \end{pmatrix} \\ &= e_R' X_2 \left(X_2' M_1 X_2\right)^{-1} X_2' e_R \end{aligned}$$

- Por lo tanto

$$LM = n \frac{e_R' X \left(X'X\right)^{-1} X' e_R}{e_R' e_R} = n \frac{ESS}{TSS} = nR^2, \tag{69}$$

donde R^2 es el coeficiente de determinación de la regresión auxiliar

$$e_R = X_1 \gamma_1 + X_2 \gamma_2 + \omega \tag{70}$$

- El contraste de hipótesis basado en el LM se resuelve como sigue,

Paso 1 Calcule el estadístico-LM con la fórmula (69).

Paso 2 Encuentre el valor crítico en una distribución χ^2, con g grados de libertad, que deje una probabilidad de α en la cola superior de la distribución.

Paso 3 Rechace H_0 si el estadístico-LM calculado en el Paso 1 es mayor que el valor crítico $\chi^2_\alpha(g)$.

- p-Valor. La regla de decisión del test del Multiplicador de Lagrange también puede ser formulada en términos del p-valor (p).

Paso 2 Calcule $p = Prob(\chi^2 > LM)$

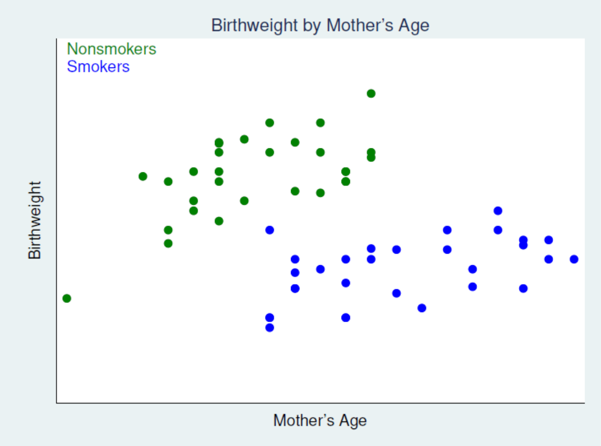
Paso 3 Regla de Decisión: Rechace H_0 si $p < \alpha$.

Costos Económicos de Nacer con Peso Bajo

- Considere un caso hipotético como el de Almond et al. (2005).
- La pregunta que se quiere responder es si fumar durante el embarazo afecta el peso de un recién nacido.
- Las unidades en este caso son mujeres embarazadas, algunas de las cuales fumaron durante el embarazo.
- La variable de resultado es el peso del bebé al nacer.

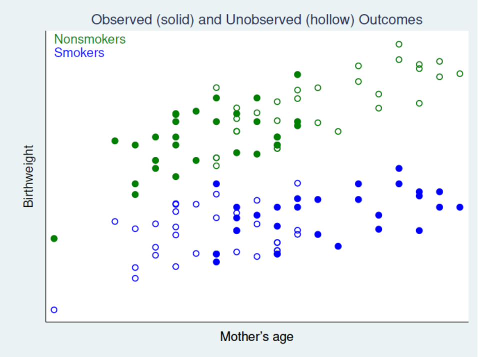
Costos Económicos de Nacer con Peso Bajo

- La figura muestra el peso del bebé al nacer para madres fumadoras y no fumadoras como función de la edad de la madre.



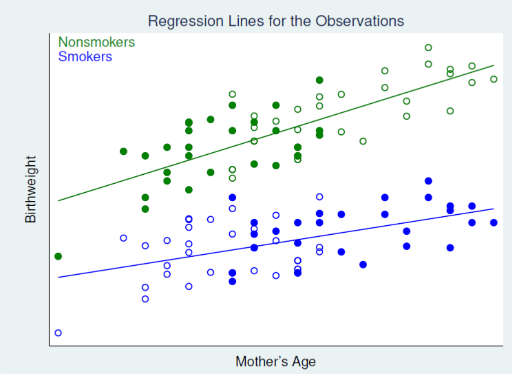
Costos Económicos de Nacer con Peso Bajo

- La figura sugiere que las mujeres fumadoras tienden a tener mayor edad que las no fumadoras.
- Para las mujeres fumadoras de mayor edad y las no fumadoras más jóvenes no parece haber un soporte común.
- Supongamos que observamos los resultados potenciales de cada mujer embarazada:



Costos Económicos de Nacer con Peso Bajo

- Lo que hace el método de regresión es estimar una regresión (con los datos observados, círculos sólidos) del peso del bebé sobre la edad de la madre para el grupo de madres fumadoras y para el grupo de madres no fumadoras.
- Luego se usa la linea de regresión muestral de las madres fumadoras como resultado contrafáctico de las madres no fumadoras y viceversa.

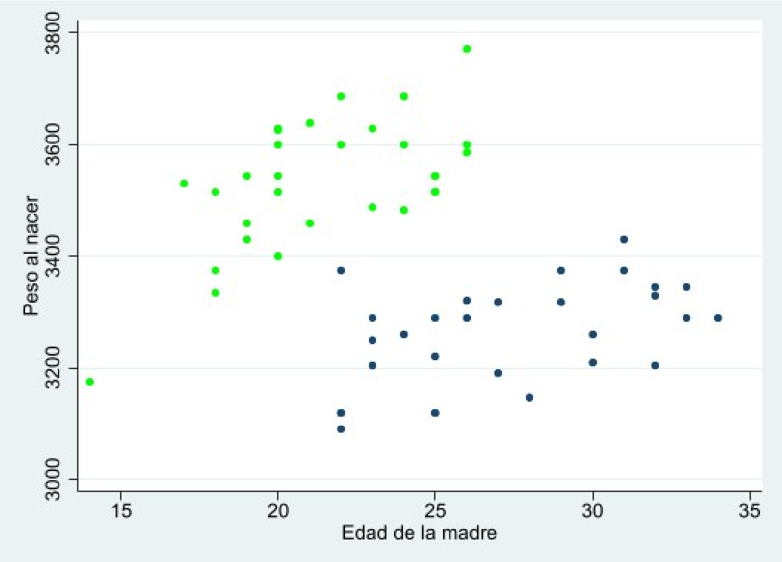


Estimación

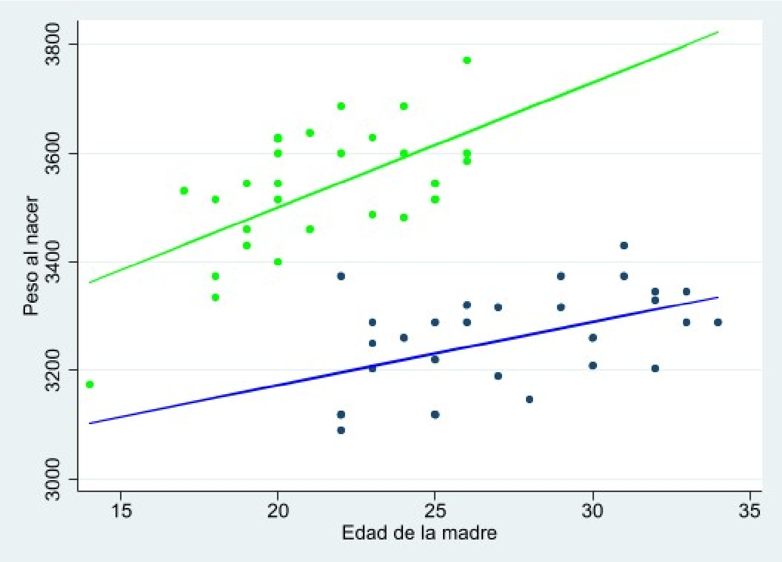
- La estimación del ATE condicional es directa a través del método de mínimos cuadrados clásicos.
- En la práctica estimamos $E[y \mid \mathbf{w}, s_T = 1]$ con el plano de regresión muestral, $\widehat{m}_1(\mathbf{w}, \hat{\delta}_1)$, con las observaciones del tratamiento y estimamos $E[y \mid \mathbf{w}, s_T = 0]$ con el plano de regresión muestral, $\widehat{m}_0(\mathbf{w}, \hat{\delta}_0)$, con las observaciones del control.

$$\widehat{\tau}_w = \frac{1}{N} \sum_{i=1}^{N} \left[\widehat{m}_1(\mathbf{w_i}, \hat{\delta}_1) - \widehat{m}_0(\mathbf{w_i}, \hat{\delta}_0) \right]$$

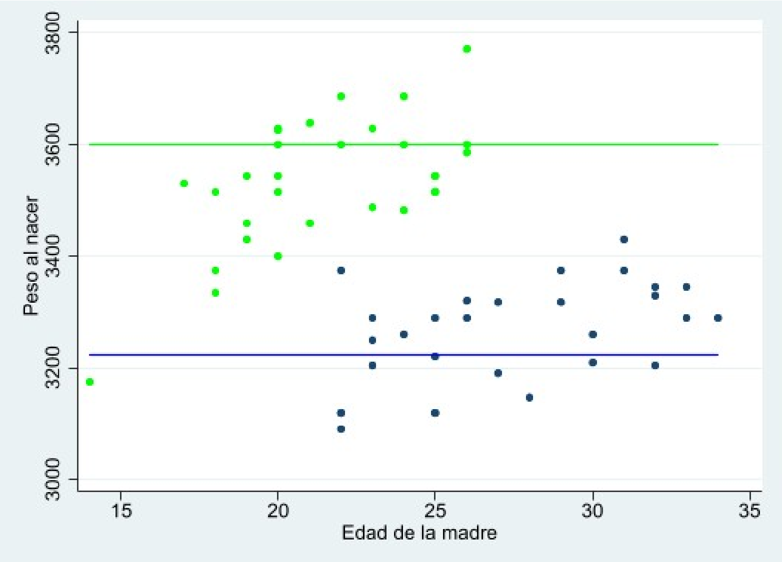
Costos Económicos de Nacer con Peso Bajo



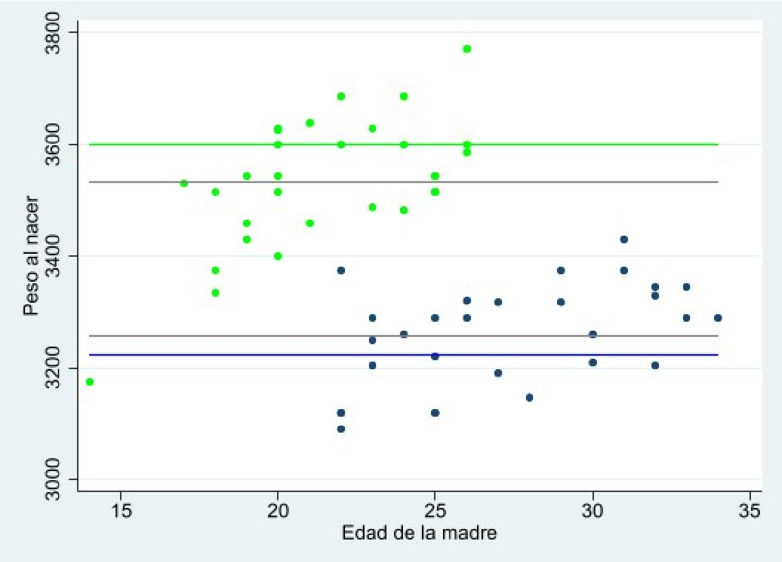
Costos Económicos de Nacer con Peso Bajo



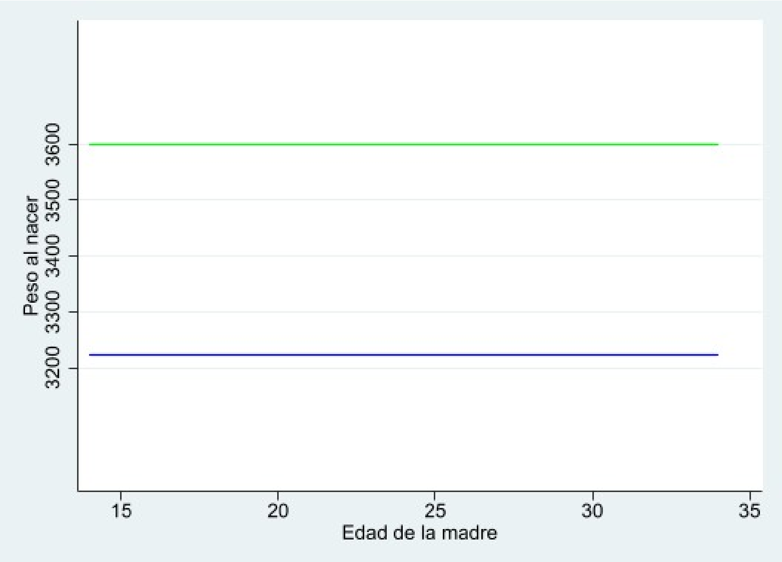
Costos Económicos de Nacer con Peso Bajo



Costos Económicos de Nacer con Peso Bajo



Costos Económicos de Nacer con Peso Bajo



Costos Económicos de Nacer con Peso Bajo