# 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**.

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

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

- Esta ecuación se conoce como "switching equation".

# Modelo de inferencia causal

**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:

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

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

## Modelo de Regresión Simple

- 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).

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

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

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

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

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

$$E[y_u|w_u, s_T] = E[y_u|w_u] \tag{10}$$

- El coeficiente $\beta_1$ sigue teniendo la interpretación de ser el efecto causal de interés.
- Esto es,

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

- 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 $\gamma_1$ y $\gamma_0$ 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):** $\mathbb{E}[Y_T(u) \mid w_u, s_T(u)] = \mathbb{E}[Y_T(u) \mid w_u]$ y $\mathbb{E}[Y_C(u) \mid w_u, s_T(u)] = \mathbb{E}[Y_C(u) \mid w_u]$.

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

# 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, $\tau$, 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 $\beta_1$ en las siguientes ecuaciones:

$$
\begin{aligned}
y_u &= \beta_0 + \beta_1 s_T(u) + \epsilon_u && (16) \\
y_u &= \beta_0 + \beta_1 s_T(u) + w_u \gamma + v_u && (17)
\end{aligned}
$$

- 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 $\epsilon_u$ [$v_u$] se conoce como la parte aleatoria del modelo.
- La ordenada al origen $\beta_0$ y la pendiente de la recta $\beta_1$ [las pendientes de cada lado del plano $\beta_1$, $\gamma$] 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.

# Modelo de Regresión Lineal Múltiple

- **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 $\beta$'s son parámetros desconocidos que deben ser estimados, y $\epsilon_i$ 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, $\beta_2$ 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,

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

# Modelo de Regresión Lineal Múltiple

- 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 $\beta$ 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 $[E(x_{jk}\epsilon_i) = 0 \;\; \forall\, k,\, \&\, i \neq j]$.

# Modelo de Regresión Lineal Múltiple

- **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.

# Modelo de Regresión Lineal Múltiple

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

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

- Usando la ley de expectativas iteradas,

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

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 $\beta$.
- 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:

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

# Modelo de Regresión Lineal Múltiple

- 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,

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

# Modelo de Regresión Lineal Múltiple

- 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

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

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

# 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 $\gamma_1$ en lugar de $\beta_1$ y el error a $u$ en lugar de $\epsilon$.
- Típicamente $\beta_1 \neq \gamma_1$ salvo casos especiales.

- Calculemos $\gamma_1$

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

  donde $\Gamma = [\mathbb{E}(x_1'x_1)]^{-1}\mathbb{E}(x_1'x_2)$ 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
$$\beta_1 = \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{1y\cdot 2} = \left[\mathbb{E}(u_1' u_1)\right]^{-1}\mathbb{E}\left(u_1' y\right) \quad (43)$$
- Se puede recuperar el coeficiente de $x_1$ en el MRLM de una regresión de $y$ sobre $u_1$.
- El coeficiente $\beta_1$ 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 $\beta_1$ 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:

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

La última desigualdad surge del supuesto 3.

- Definiciones

- 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 $\sigma^2$ (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 $\beta$ 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 $\sigma^2$, entonces

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

# Agenda

1. Una secuencia de variables aleatorias $\{x_N : N = 1, 2, \ldots\}$ 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 $N_\epsilon$, 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 $\{x_N : N = 1, 2, \dots\}$ 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 $\{Z_N : N = 1, 2, \dots\}$ 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 $Z_N' x_N = o_p(1)$.

# Teoría Asintótica: Convergencia en Probabilidad

- **Lema 4 (Teorema de Slutsky):** Sea $g : \mathbb{R}^K \rightarrow \mathbb{R}^J$ 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 $\{x_N : N = 1, 2, \ldots\}$ 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 $\{x_N : N = 1, 2, \ldots\}$ $K \times 1$ 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 $g : \mathbb{R}^K \to \mathbb{R}^J$ 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): $N^{-1/2} \sum_{j=1}^{N} w_j \xrightarrow{d} \text{Normal}(0, B)$, donde $B = Var(w_j) = E(w_j w_j')$.

# Mínimos Cuadrados Clásicos

- 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.

# Mínimos Cuadrados Clásicos

- 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,

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

- 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 $\beta_1$, $\hat{\beta}_1$, 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:

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

- y

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

- 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,

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

# Mínimos Cuadrados Clásicos

- El $R^2$ no centrado se define como,

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

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

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

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

- 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 $\alpha$ 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 $\alpha$ en la cola superior de la distribución.

# MCC: inferencia sobre combinación lineal de coeficientes

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.

# MCC: inferencia sobre combinación lineal de coeficientes

- 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 $\beta_2$ es de dimensión $g \times 1$ y $\beta_1$ 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 $\lambda$ es el vector de los $g$ multiplicadores de Lagrange.

# MCC: inferencia sobre combinación lineal de coeficientes

- 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

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

# MCC: inferencia sobre combinación lineal de coeficientes

- Esto significa que $\hat{\lambda}'\left(X_{22}M_1X_2\right)^{-1}\hat{\lambda}/\sigma^2$ se distribuye como una $\chi^2(g)$
- Reemplazando la varianza desconocida, $\sigma^2$ por una estimación consistente $\hat{\sigma}^2 = \frac{1}{n}e_R'e_R$ se define el estadístico del multiplicador de Lagrange:

$$LM = \hat{\lambda}'\left(X_2'M_1X_2\right)^{-1}\hat{\lambda}/\hat{\sigma}^2 \sim \chi^2(g) \tag{67}$$

- Note que el estadístico se puede escribir como

$$LM = n\frac{e_R'X_2\left(X_2'M_1X_2\right)^{-1}X_2'e_R}{e_R'e_R}. \tag{68}$$

- 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 $\beta_2$ 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

$$\begin{aligned} e_R'X\left(X'X\right)^{-1}X'e_R &= \begin{pmatrix} 0 & e_R'X_2 \end{pmatrix}\left(X'X\right)^{-1}\begin{pmatrix} 0 \\ X_2'e_R \end{pmatrix} \\ &= e_R'X_2\left(X_2'M_1X_2\right)^{-1}X_2'e_R \end{aligned}$$

- 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 $\chi^2$, con $g$ grados de libertad, que deje una probabilidad de $\alpha$ 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 $p = Prob(\chi^2 > LM)$

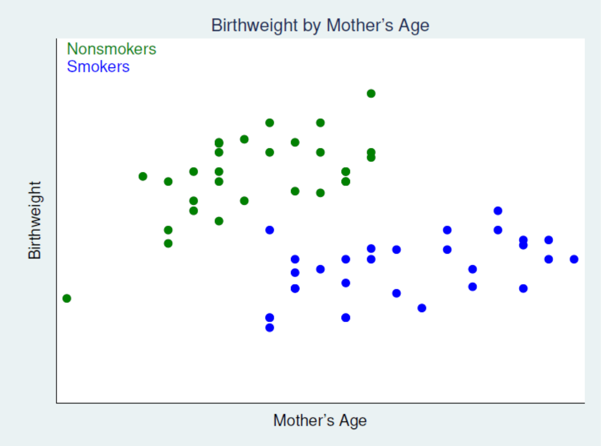
Paso 3 Regla de Decisión: Rechace $H_0$ si $p < \alpha$.

# 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.

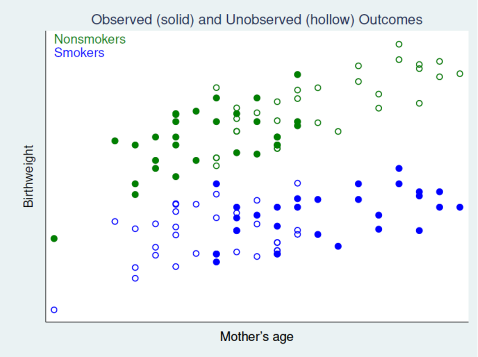
# Costos Económicos de Nacer con Peso Bajo

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



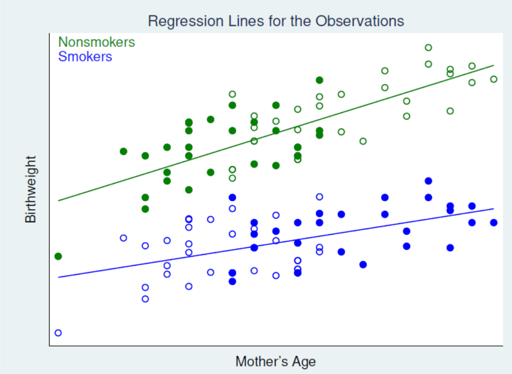
# Costos Económicos de Nacer con Peso Bajo

- 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

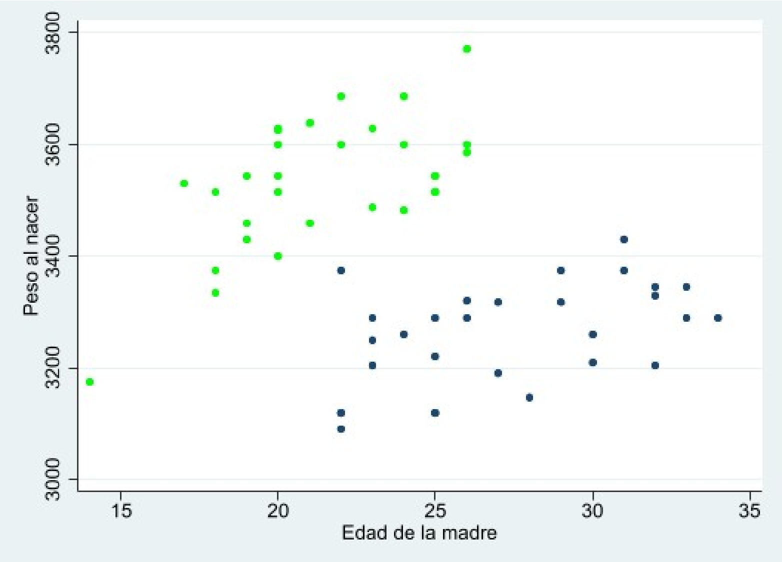
- Lo que hace el método de regresión es estimar una regresión (con los datos observados, círculos sólidos) del peso del bebé sobre la edad de la madre para el grupo de madres fumadoras y para el grupo de madres no fumadoras.
- Luego se usa la linea de regresión muestral de las madres fumadoras como resultado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



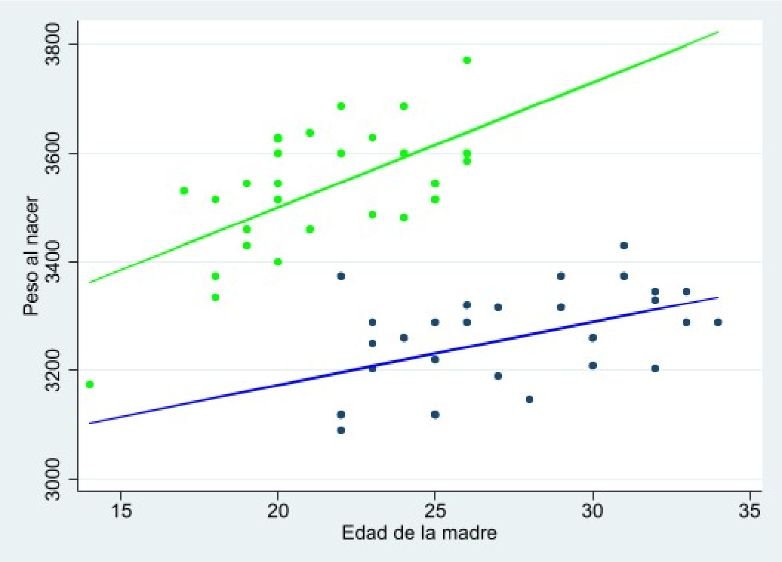
- La estimación del ATE condicional es directa a través del método de mínimos cuadrados clásicos.
- En la práctica estimamos $E[y \mid \mathbf{w}, s_T = 1]$ con el plano de regresión muestral, $\widehat{m}_1(\mathbf{w}, \hat{\delta}_1)$, con las observaciones del tratamiento y estimamos $E[y \mid \mathbf{w},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

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

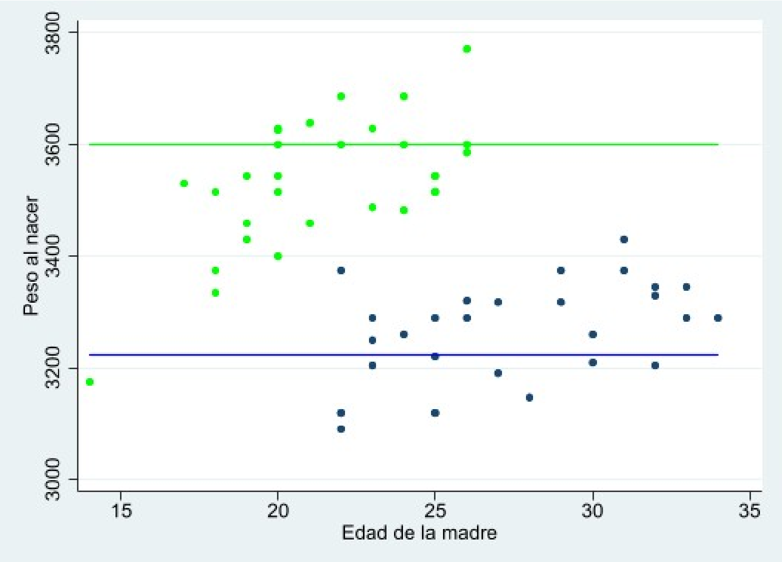
# Costos Económicos de Nacer con Peso Bajo



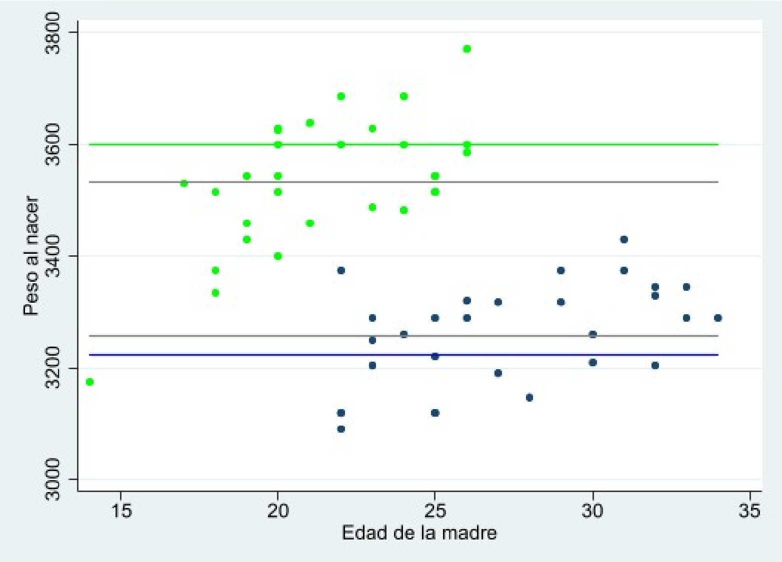
# Costos Económicos de Nacer con Peso Bajo



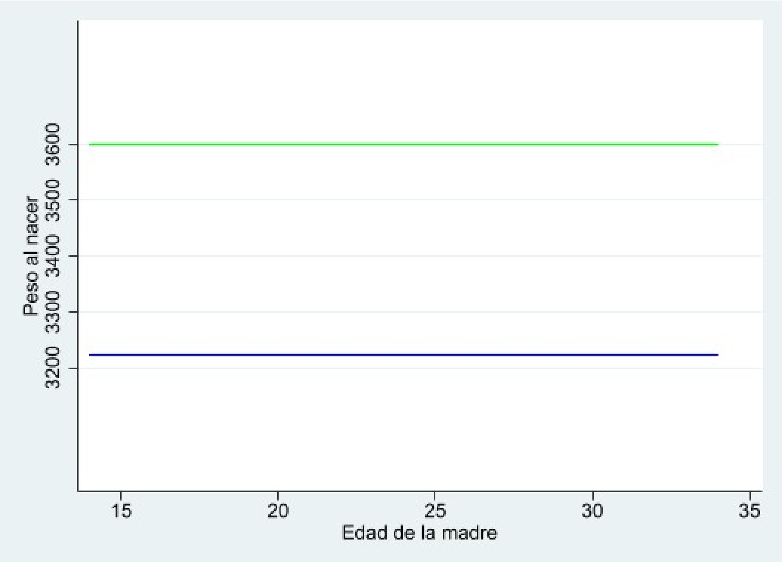
# Costos Económicos de Nacer con Peso Bajo



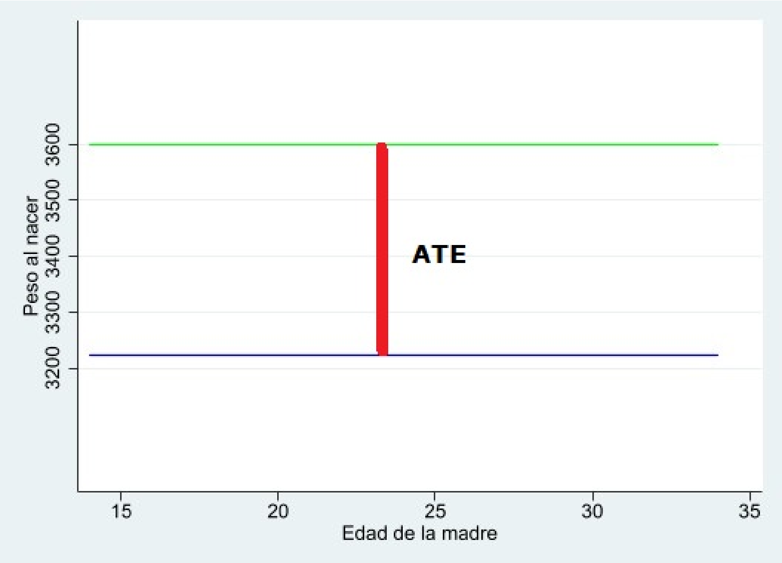
# Costos Económicos de Nacer con Peso Bajo



# Costos Económicos de Nacer con Peso Bajo



# Costos Económicos de Nacer con Peso Bajo



# Econometría

## Maestría en Economía - Maestría en Econometría

Lecture 2

# Agenda

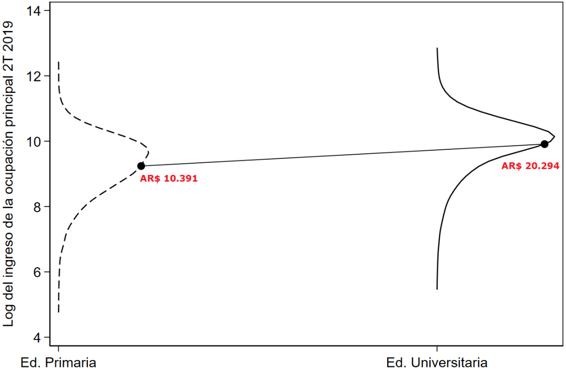
1. Variables Omitidas: Motivación
   - La Ecuación del Salario con Habilidad no Observada
   - Consecuencias de Ignorar la Presencia de Variables Omitidas
2. Soluciones al Problema de las Variables Omitidas
   - Variables “Proxy”
   - Variables Instrumentales
   - Mínimos Cuadrados en dos Etapas
3. El enfoque de la función de control para la endogeneidad
4. Errores no Esféricos
   - Efectos de Reducir el Tamaño de las Clases
   - Propiedades de MCC en Presencia de Errores no Esféricos
   - Inferencia en Presencia de Errores no Esféricos
   - Mínimos Cuadrados Generalizados

# Agenda

1. Variables Omitidas: Motivación
   - La Ecuación del Salario con Habilidad no Observada
   - Consecuencias de Ignorar la Presencia de Variables Omitidas
2. Soluciones al Problema de las Variables Omitidas
   - Variables “Proxy”
   - Variables Instrumentales
   - Mínimos Cuadrados en dos Etapas
3. El enfoque de la función de control para la endogeneidad
4. Errores no Esféricos
   - Efectos de Reducir el Tamaño de las Clases
   - Propiedades de MCC en Presencia de Errores no Esféricos
   - Inferencia en Presencia de Errores no Esféricos
   - Mínimos Cuadrados Generalizados

# Sesgo por variables omitidas

- ¿Cuál es el retorno monetario a la educación?
- En promedio, personas con mayor educación ganan más que personas menos educadas.
- Imagine, para empezar, que la educación es una decisión binaria: "obtengo educación primaria o universitaria"
- Si la educación se asigna aleatoriamente, entonces podríamos comparar el ingreso promedio de quienes reciben educación universitaria con aquellos que reciben educación primaria.



# Sesgo por variables omitidas

- En ausencia de aleatorización de la educación surgen dos potenciales sesgos al comparar el ingreso laboral promedio de ambos grupos de educación.
- Sesgos provocados por diferencias en características observables entre ambos grupos (se corrigen controlando por esas características)
- Sesgos provocados por diferencias en características no observables entre ambos grupos.
- Como ejemplo piense que un sesgo positivo surgiría si personas con mayor capacidad de ingresos (más productivos, más hábiles, etc.) obtuvieran más educación.
- En este caso la comparación del ingreso promedio de ambos grupos capturaría no solo el retorno a la educación sino también las diferencias en habilidad.

# Sesgo por variables omitidas

- La ecuación estructural típica para el salario es como sigue,

$$log(\text{wage}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 s + \gamma h + v$$

  donde $x$ representa años de experiencia en el mercado de trabajo, $s$ representa años de educación formal y $h$ es la habilidad natural del individuo. $v$ es el error estructural que satisface el supuesto de exogeneidad estricta: $E(v|x, s, h) = 0$.
- La teoría económica establece un perfil salarial creciente y cóncavo en experiencia, sugiriendo que $\beta_1 > 0$ y $\beta_2 < 0$ en la ecuación anterior. Además, la teoría del capital humano sugiere una relación directa entre salario y educación ($\beta_3 > 0$) y entre salario y habilidad ($\gamma > 0$).

# Sesgo por variables omitidas

- Empíricamente el problema para estimar una ecuación salarial como la anterior es que la habilidad de una persona no es observable. Por lo tanto, la ecuación estimable es,

$$log(\text{wage}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 s + u$$

donde $u = \gamma h + v$. En este modelo, en general $E(u|x, s) \neq 0$ debido a la probable correlación entre los años de educación y la habilidad de una persona.

- En econometría una variable explicativa $x_j$ se dice que es endógena si está correlacionada con el error de la ecuación.
- Empíricamente, la endogeneidad aparece frecuentemente cuando se presenta el problema descripto para la ecuación del salario. Es decir cuando tenemos el denominado problema de las variables omitidas.

# Agenda

# Mínimos Cuadrados Clásicos

- Considere el siguiente modelo estructural,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \gamma q + v \tag{1}$$

donde $E(v|x_1, x_2, \ldots, x_k, q) = 0$ y $q$ es la variable no observada.

- Nuestro interés es estimar correctamente los $\beta$'s que son los efectos parciales de las variables observadas manteniendo constantes el resto de las variables explicativas, incluyendo a $q$.
- El modelo que se podría estimar es,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \tag{2}$$

donde $u = \gamma q + v$.

- Sin pérdida de generalidad se puede asumir que $E(q) = 0$ de forma tal que $E(u) = 0$.

- En este modelo si $q$ está correlacionada con alguna de las variables explicativas, entonces $u$ estará correlacionado también y tenemos el problema de la endogeneidad.
- Sabemos que si no se satisface (al menos) el supuesto de exogeneidad contemporánea MCC no dará estimaciones consistentes de los parámetros.
- Escribamos la proyección lineal de $q$ en las $k$ variables explicativas observadas como,

$$q = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + r \tag{3}$$

  donde, por definición de proyección lineal, $E(r) = 0$ y $Cov(x_j, r) = 0$, $j = 1, 2, \ldots, k$.
- Sustituyendo la ecuación (3) en la (1) podemos ver que estimaría MCC aplicado sobre la ecuación (2).

- $$y = (\beta_0 + \gamma\delta_0) + (\beta_1 + \gamma\delta_1)x_1 + (\beta_2 + \gamma\delta_2)x_2 + \cdots + (\beta_k + \gamma\delta_k)x_k + v + \gamma r$$

  donde el error $v + \gamma r$ cumple con el supuesto de exogeneidad estricta.
- De la ecuación anterior surge claramente que MCC aplicado en (2) dará estimadores consistentes, $\hat{\beta}_j$, de los parámetros $\beta_j + \gamma\delta_j$.
- Esta especificación es la más general que se puede tener. Muchas veces en la práctica la variable omitida solo está relacionada con alguna de las variables explicativas observadas. Esta especificación puede obtenerse haciendo ceros a los $\delta_j$ correspondientes en la ecuación (3) arriba.

# La Ecuación del Salario con Habilidad no Observada

- Volviendo al ejemplo del salario,

$$log(\text{wage}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 s + \gamma h + v$$

  la teoría económica sugiere correlación entre $s$ y $h$. Las personas con mayor habilidad alcanzan una educación más alta. Supongamos que la relación entre la habilidad y la educación es $h = \pi_0 + \pi_1 s$.

- Si omitimos $h$ en la estimación,

$$log(\text{wage}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 s + u$$

  MCC dará un estimador de $\beta_3$ sesgado y no consistente.

# La Ecuación del Salario con Habilidad no Observada

- Reemplazando $h$ en la ecuación del salario tenemos,

$$log(\text{wage}) = (\beta_0 + \pi_0\gamma) + \beta_1 x + \beta_2 x^2 + (\beta_3 + \pi_1\gamma)s + v$$

- En este ejemplo particular, MCC da estimaciones insesgadas y consistentes de $\beta_1$ y $\beta_2$ pero no asi de $\beta_3$. Si $\pi_1 > 0$ como sugiere la teoría económica, entonces MCC sobre-estimará el coeficiente asociado con los retornos a la educación.

# Agenda

# Solución usando variables proxy

- El problema de las variables omitidas puede ser solucionado si existe una variable proxy para la variable no observada.
- Los requerimientos formales para que una variable pueda ser considerada proxy de otra son dos.
  1. La variable proxy debe ser redundante en la ecuación estructural. Si $w$ es una variable proxy para $q$, el requerimiento de redundancia establece que $E(y|x, q, w) = E(y|x, q)$.
  2. La correlación entre la variable omitida $q$ y cada $x_j$ debe ser cero una vez que tomamos en consideración $w$. En términos de una proyección lineal este supuesto establece que $L(q|1, x_1, \ldots, x_k, w) = L(q|1, w)$.
- Es útil escribir el segundo punto en términos de una ecuación con error,

$$q = \theta_0 + \theta_1 w + r \tag{4}$$

donde por definición $E(r) = 0$ y $Cov(w, r) = 0$. Si $w$ es una variable proxy razonable de $q$, entonces $\theta_1 \neq 0$. La condición 2 de arriba establece además que $Cov(x_j, r) = 0 \;\; \forall j$.

# Solución usando variables proxy

- Para obtener una ecuación estimable podemos reemplazar (4) en (1),

$$y = (\beta_0 + \gamma\theta_0) + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \gamma\theta_1 w + (\gamma r + v)$$

  donde, bajo los supuestos realizados, el error de la ecuación, $\gamma r + v$, no está correlacionado con $x_j$, $\forall j$; redundancia establece que $w$ no está correlacionado con $v$ y por definición $w$ no está correlacionado con $r$.
- Por lo tanto el error satisface el supuesto de exogeneidad contemporánea y MCC aplicados en la ecuación anterior da estimaciones consistentes de $(\beta_0 + \gamma\theta_0)$, $\beta_1, \ldots, \beta_k$ y $\gamma\theta_1$.
- Entonces, bajo los supuestos de variables proxy, MCC estima en forma consistente el efecto parcial de las variables explicativas observadas $(x_j)$. En particular, en el ejemplo del salario empíricamente se utilizan los resultados de tests de inteligencia como proxy de habilidad.

# Agenda

# Solución usando variables instrumentales

- Las variables instrumentales son una forma de solucionar el problema de endogeneidad de las variables explicativas. En este sentido, es una corrección más general que la de las variables proxy porque no solo puede aplicarse en el caso de variables omitidas, sino en cualquier caso en el que exista endogeneidad de algún regresor (i.e. error de medición en variables explicativas, causalidad simultánea, etc.).
- Considere el siguiente modelo,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \tag{5}$$

donde $E(u) = 0$ y $Cov(u, x_j) = 0$, $j = 1, 2, \ldots, k-1$. En palabras, $x_k$ es potencialmente endógena en (5).

- Por lo tanto, MCC aplicados a (5) nos dará estimadores inconsistentes.

- Para usar el enfoque de IV con $x_k$ endógena, necesitamos una variable observable $z_1$, que no esté en (5) y que satisfaga dos condiciones:
  1. $z_1$ es una variable exógena en (5), es decir $Cov(z_1, u) = 0$.
  2. $\theta_1 \neq 0$ en la proyección lineal de la variable endógena, $x_k$, sobre todas las variables exógenas,

$$x_k = \delta_0 + \delta_1 x_1 + \cdots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + r_k \tag{6}$$

  donde, por definición, $E(r_k) = 0$ y $r_k$ no está correlacionado con $x_1, x_2, \ldots, x_{k-1}, z_1$.
- En palabras, $z_1$ está parcialmente correlacionada con $x_k$ una vez que el resto de las variables exógenas han sido tomadas en cuenta.
- Cuando $z_1$ satisface estas dos condiciones se dice una variable instrumental para $x_k$. La proyección lineal (6) se denomina ecuación de forma reducida para la variable endógena $x_k$.

# Solución usando variables instrumentales

- Reemplazando (6) en (5) tenemos,

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{k-1} x_{k-1} + \lambda_1 z_1 + v \tag{7}$$

  donde $v = u + \beta_k r_k$, $\alpha_j = \beta_j + \beta_k \delta_j$ y $\lambda_1 = \beta_k \theta_1$.
- Por nuestros supuestos, $v$ no está correlacionado con ninguna de las variables explicativas de (7) y por lo tanto MCC estima consistentemente los parámetros de la ecuación reducida de $y$.
- Algunas veces, estimar los parámetros de la ecuación reducida (7) tiene interés en si mismo pero en general se trata de estimar en forma consistente los parámetros de (5). Los supuestos hechos para IV también lo permiten.

# Solución usando variables instrumentales

- Para ver esto formalmente, escribamos (5) como,

$$y = x\beta + u$$

  donde $x = (1, x_1, \ldots, x_k)$ y $\beta' = (\beta_0, \beta_1, \ldots, \beta_k)$ son de dimensión $1 \times k + 1$.
- Definamos el vector de variables exógenas $z = (1, x_1, \ldots, x_{k-1}, z_1)$ y el vector de parámetros de la ecuación reducida $\delta' = (\delta_0, \delta_1, \ldots, \delta_{k-1}, \theta_1)$.
- Bajo los supuestos de (5) y el supuesto de que la variable instrumental $z_1$ es exógena se cumplen las siguientes condiciones de ortogonalidad:

$$E(z'u) = 0.$$

- Por lo tanto,

$$
\begin{aligned}
z'y &= z'x\beta + z'u \Rightarrow E(z'y) = E(z'x)\beta + E(z'u) \\
&\Rightarrow E(z'y) = E(z'x)\beta \qquad (8)
\end{aligned}
$$

  Donde $E(z'x)$ es de dimensión $k+1 \times k+1$, y $E(z'y)$ es de dimensión $k+1 \times 1$.

- La última expresión representa un sistema de $k+1$ ecuaciones lineales con $k+1$ incógnitas. El sistema tiene una solución única si y solo sí la matriz $E(z'x)$ tiene rango completo (i.e. rango $E(z'x) = k+1$).

# Solución usando variables instrumentales

- Note que $x = z\pi + r$ con

$$\pi = \begin{bmatrix} 1 & 0 & \cdots & 0 & \delta_0 \\ 0 & 1 & \cdots & 0 & \delta_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \delta_{k-1} \\ 0 & 0 & \cdots & 0 & \theta_1 \end{bmatrix}$$

  y $r = (0, \ldots, 0, r_k)$.
- Entonces $E(z'x) = E(z'(z\pi + r)) = E(z'z)\pi$ de forma tal que para que $E(z'x)$ tenga rango completo necesitamos que $E(z'z)$ tenga rango $k+1$, que es un supuesto estándar, y que $\pi$ tenga rango $k+1$ que está garantizado por el supuesto 2 de IV ($\theta_1 \neq 0$).
- En este caso, la solución del sistema de ecuaciones está dada por:

$$\beta = [E(z'x)]^{-1} E(z'y)$$

## Solución usando variables instrumentales

- Utilizando los análogos muestrales se obtiene el estimador de variables instrumentales,

$$\begin{aligned} \hat{\beta} &= \left(\frac{1}{n}\sum_{i=1}^{n} z_i' x_i\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} z_i' y_i \\ &= (z'x)^{-1}z'y = (z'x)^{-1}z'(x\beta + u) \\ &= \beta + (z'x)^{-1}z'u \end{aligned} \tag{9}$$

- Usando la WLLN,

$$\left(\frac{1}{n}\sum_{i=1}^{n} z_i' x_i\right) \xrightarrow{p} E(z'x)$$

que bajo los supuestos de IV tiene rango completo.

# Solución usando variables instrumentales

- Además,

$$\frac{1}{n}\sum_{i=1}^{n} z_i' u_i \xrightarrow{p} E(z'u) = 0$$

  y el estimador de IV es consistente.
- Volviendo al ejemplo de la ecuación de salarios, la omisión de la habilidad ($h$) provoca que la variable que mide educación ($s$) sea endógena en el modelo. Para obtener estimaciones consistentes en la ecuación salarial necesitamos un instrumento para $s$.
- Card(1995), por ejemplo, utiliza una variable binaria que indica si una persona creció en el vecindario de una universidad como variable instrumental de años de educación.

# Agenda

1. Variables Omitidas: Motivación
   - La Ecuación del Salario con Habilidad no Observada
   - Consecuencias de Ignorar la Presencia de Variables Omitidas
2. Soluciones al Problema de las Variables Omitidas
   - Variables "Proxy"
   - Variables Instrumentales
   - Mínimos Cuadrados en dos Etapas
3. El enfoque de la función de control para la endogeneidad
4. Errores no Esféricos
   - Efectos de Reducir el Tamaño de las Clases
   - Propiedades de MCC en Presencia de Errores no Esféricos
   - Inferencia en Presencia de Errores no Esféricos
   - Mínimos Cuadrados Generalizados

- Considere nuevamente la ecuación (5) con todos sus supuestos. Es decir, suponga que existe endogeneidad potencial de $x_k$.
- Supongamos que tenemos más de una variable instrumental para $x_k$. En particular, supongamos que $z_1, z_2, \ldots, z_M$ son variables tal que,

$$Cov(z_h, u) = 0, \quad h = 1, 2, \ldots, M. \tag{10}$$

  cada $z_h$ es exógena en la ecuación (5).
- Si cada una de estas variables tiene alguna correlación parcial con $x_k$, tenemos $M$ potenciales instrumentos.
- En realidad, hay muchos más que $M$ porque cualquier combinación lineal de $x_1, x_2, \ldots, x_{k-1}, z_1, z_2, \ldots, z_M$ no tiene correlación con $u$. Qué instrumento deberíamos utilizar?
- Bajo ciertos supuestos Mínimos cuadrados en dos etapas (2SLS) es el estimador de IV más eficiente.

# Mínimos Cuadrados en dos Etapas

- Definamos el vector de variables exógenas como antes, $z = (1, x_1, x_2, \ldots, x_{k-1}, z_1, z_2, \ldots, z_M)$ un vector de dimensión $1 \times L$ con $L = k + M$.
- Definamos la proyección lineal de la variable endógena sobre todas las variables exógenas,

$$x_k = \delta_0 + \delta_1 x_1 + \cdots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + \cdots + \theta_M z_M + r_k \tag{11}$$

donde, por definición, $E(r_k) = 0$ y $r_k$ no está correlacionado con ninguna de las variables en el lado derecho de la ecuación.

- Como ninguna combinación lineal de $z$ está correlacionada con $u$

$$x_k^* = \delta_0 + \delta_1 x_1 + \cdots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + \cdots + \theta_M z_M \tag{12}$$

tampoco lo estará.

- Si observáramos $x_k^*$ podríamos utilizarla como instrumento para $x_k$ en (5).

# Mínimos Cuadrados en dos Etapas

- Sin embargo, si no existen dependencias lineales exactas entre las variables exógenas se podrían estimar en forma consistente por MCC los parámetros de (11) y definir para cada observación $i$,

$$\hat{x}_{i,k} = \hat{\delta}_0 + \hat{\delta}_1 x_{i,1} + \cdots + \hat{\delta}_{k-1} x_{i,k-1} + \hat{\theta}_1 z_{i,1} + \cdots + \hat{\theta}_M z_{i,M} \tag{13}$$

- Ahora para cada observación $i$ definamos el vector $\hat{x}_i \equiv (1,\, x_{i,1},\, \ldots,\, x_{i,k-1}, \hat{x}_{i,k})$ y estimemos por IV,

$$\hat{\beta} = \left( \sum_{i=1}^{n} \hat{x}_i' x \right)^{-1} \sum_{i=1}^{n} \hat{x}_i' y = (\hat{x}' x)^{-1} \hat{x}' y. \tag{14}$$

- Este estimador IV es también un estimador de MCC.

$$\hat{x} = z\hat{\delta} = z(z'z)^{-1} z' x = P_z x$$

con $P_z$ una matriz idempotente y simétrica.

# Mínimos Cuadrados en dos Etapas

- Por lo tanto, $\hat{x}'x = x'P_z x = (P_z x)'P_z x = \hat{x}'\hat{x}$. Reemplazando esta última expresión en (14) se obtiene,

$$\hat{\beta} = (\hat{x}'\hat{x})^{-1}\hat{x}'y. \tag{15}$$

- El término, mínimos cuadrados en dos etapas viene de este procedimiento.
- Entonces $\hat{\beta}$ se puede obtener con los siguientes pasos
    1. Obtenga $\hat{x}_k$ de la regresión de $x_k$ sobre $x_1, \ldots, x_{k-1}, z_1, \ldots, z_M$. Este paso se denomina regresión de la primera etapa.
    2. Estime por MCC una regresión de $y$ sobre $x_1, \ldots, x_{k-1}, \hat{x}_k$. Esta es la regresión de la segunda etapa.
- El estimador de 2SLS y el de IV son idénticos si solo existe un instrumento para $x_k$.

- En términos generales podemos resumir los resultados de 2SLS como sigue. Considere el modelo,

$$y = x\beta + u$$

donde $x$ es de dimensión $1 \times k$ y varios elementos de $x$ pueden estar potencialmente correlacionados con $u$.

- Supuesto 1: Para algún vector $1 \times L$, $z$, $E(z'u) = 0$.
Note que el supuesto de exogeneidad estricta $E(u|z) = 0$ implica el supuesto 1.

- Supuesto 2: (a) rango $E(z'z) = L$; (b) rango $E(z'x) = k$.
Técnicamente, la parte (a) del supuesto 2 es necesaria pero no especialmente importante. La parte (b) del supuesto es la realmente importante porque es la condición de rango que permite la identificación de los parámetros del modelo.

# Mínimos Cuadrados en dos Etapas

- Definamos el vector de variables exógenas como antes, $z = (1, x_1, x_2, \ldots, x_{k-1}, z_1, z_2, \ldots, z_M)$ un vector de dimensión $1 \times L$ con $L = k + M$.
- Definamos la proyección lineal de $x$ sobre $z$ como $x = x^* + r$ con $x^* = z\pi = z[E(z'z)]^{-1}E(z'x)$ y $r = (0, \ldots, 0, r_k)$.
- Entonces, multiplicando el modelo por $x^{*'}$ y tomando esperanzas tenemos,

$$E(x^{*'}y) = E(x^{*'}x)\beta + E(x^{*'}u) = E(x^{*'}x)\beta$$

  y $\beta$ está identificado por $\beta = [E(x^{*'}x)]^{-1}E(x^{*'}y)$ si $E(x^{*'}x)$ no es singular.
- Ahora $E(x^{*'}x) = \pi' E(z'x) = E(z'x)'[E(z'z)]^{-1}E(z'x)$ y esta matriz no es singular si $E(z'x)$ tiene rango $k$, lo que está garantizado si se cumple el supuesto 2 (b).

- El estimador de 2SLS se puede escribir con la ecuación (14) o con,

$$
\begin{aligned}
\hat{\beta} &= \left[\left(\sum_{i=1}^{n} x_i' z_i\right)\left(\sum_{i=1}^{n} z_i' z_i\right)^{-1}\left(\sum_{i=1}^{n} z_i' x_i\right)\right]^{-1} \\
&\times \left(\sum_{i=1}^{n} x_i' z_i\right)\left(\sum_{i=1}^{n} z_i' z_i\right)^{-1}\left(\sum_{i=1}^{n} z_i' y_i\right) \\
&= \beta + \left[\left(n^{-1}\sum_{i=1}^{n} x_i' z_i\right)\left(n^{-1}\sum_{i=1}^{n} z_i' z_i\right)^{-1}\left(n^{-1}\sum_{i=1}^{n} z_i' x_i\right)\right]^{-1} \\
&\times \left(n^{-1}\sum_{i=1}^{n} x_i' z_i\right)\left(n^{-1}\sum_{i=1}^{n} z_i' z_i\right)^{-1}\left(n^{-1}\sum_{i=1}^{n} z_i' u_i\right) \qquad (16)
\end{aligned}
$$

# Mínimos Cuadrados en dos Etapas

- El estimador de 2SLS es consistente aplicando la WLLN a la ecuación anterior. Además apropiadamente re-escalado es asintóticamente normal.
- La normalidad asintótica de $\sqrt{n}(\hat{\beta} - \beta)$ se sigue de la normalidad asintótica de $n^{-1/2} \sum_{i=1}^{n} z_i' u_i$, que sigue del CLT bajo el supuesto 1,

$$n^{-1/2} \sum_{i=1}^{n} z_i' u_i \sim \text{Normal}(0, \sigma^2 E(z'z))$$

- Si agregamos el supuesto de varianza de los errores esférica podemos derivar la matriz de varianzas y covarianzas de los estimadores de 2SLS.
- Supuesto 3: $E(uu'|z) = \sigma^2$.
-

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \text{Normal}(0, \sigma^2 \{E(x'z)[E(z'z)]^{-1} E(z'x)\}^{-1}) \tag{17}$$

- La estimación de la matriz de varianzas y covarianzas asintótica de los estimadores de 2SLS se obtiene usando los análogos muestrales de las esperanzas y estimando consistentemente $\sigma^2$.
- Definiendo los residuos de la estimación de 2SLS como,

$$\hat{u}_i = y_i - x_i\hat{\beta}, \quad i = 1, 2, \ldots, n.$$

- La estimación consistente de $\sigma^2$ viene dada por,

$$\hat{\sigma}^2 = (n-k)^{-1}\sum_{i=1}^{n}\hat{u}_i^2$$

- Por lo tanto $\hat{\sigma}^2(\sum_{i=1}^{n}\hat{x}_i'\hat{x}_i)^{-1} = \hat{\sigma}^2(\hat{x}'\hat{x})^{-1}$ es un estimador válido de la matriz de varianzas y covarianzas de los estimadores de 2SLS.

# Mínimos Cuadrados en dos Etapas

- Bajo los supuestos 1, 2 y 3, los estimadores de 2SLS son eficientes dentro de la clase de todos los estimadores de IV que usan instrumentos lineales en $z$.
- Es posible detectar la presencia de variables explicativas endógenas?
- Test de Hausman (1978)
  Idea: bajo la hipótesis nula de no existencia de endogeneidad, el estimador de MCC, $\hat{\beta}_{MCC}$, y el estimador de variables instrumentales, $\hat{\beta}_{2SLS}$, son estimadores consistentes de $\beta$, y el estimador de MCC es el más eficiente.
- Si la hipótesis nula es falsa, el estimador de variables instrumentales, $\hat{\beta}_{2SLS}$, es el único consistente.
- Entonces, bajo la hipótesis nula ambos estimadores deberían diferir solo por error muestral. Es decir, aceptar la hipótesis nula del test es evidencia en favor de exogeneidad.

# Mínimos Cuadrados en dos Etapas

- Hausman sugiere utilizar un test de Wald. Supongamos que $V_{2SLS}$ es la matriz de varianzas y covarianzas asintótica del estimador de variables instrumentales y $V_{MCC}$ es la correspondiente al estimador de MCC. Entonces,

$$H = (\hat{\beta}_{MCC} - \hat{\beta}_{2SLS})'[V_{2SLS} - V_{MCC}]^{-1}(\hat{\beta}_{MCC} - \hat{\beta}_{2SLS}) \sim \chi^2(q)$$

  donde $q$ es la dimensión del vector $\hat{\beta}_{MCC}$.

- Si hay una sola variable potencialmente endógena, $x_k$, el estadístico de Hausman se reduce a,

$$t_H = \frac{(\hat{\beta}_{k,MCC} - \hat{\beta}_{k,2SLS})}{\sqrt{V_{k,2SLS} - V_{k,MCC}}} \sim \text{Normal}(0,\ 1)$$

# Función de Control

- En general el método de la función de control se utiliza para manejar la endogeneidad en modelos no lineales.
- El enfoque de la función de control utiliza regresores adicionales para romper la correlación entre las variables explicativas endógenas y los errores noobservables que afectan la variable dependiente.
- Como en el método de VI (o 2SLS) este enfoque también descansa en la existencia de variables exógenas que no aparecen en la ecuación estructural.
- Supongamos que $y_1$ es la variable dependiente, $y_2$ es una variable explicativa endógena y $\mathbf{z}$ es un vector de dimensión $1 \times L$ de variables exógenas (incluye un 1 en el primer elemento para la constante).
- Considere el siguiente modelo:

$$y_1 = \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \tag{18}$$

con $\mathbf{z}_1$ un subconjunto $1 \times L_1$ de $\mathbf{z}$.

# Función de Control

- Las variables en $\mathbf{z}$ son exógenas en el mismo sentido que con VI, es decir $E(\mathbf{z}'u_1) = 0$.
- Por lo tanto se puede estimar $(\boldsymbol{\delta}_1, \alpha_1)$ en forma consistente usando 2SLS (agregando la condición de rango estándar).
- En el método de la función de control la forma reducida de la variable endógena juega un rol fundamental:

$$y_2 = \mathbf{z}\pi_2 + v_2 \qquad (19)$$

$$\mathrm{E}\left(\mathbf{z}'v_2\right) = \mathbf{0} \qquad (20)$$

donde $\pi_2$ es $L \times 1$

- En este modelo la endogenidad de $y_2$ aparece si y solo si $u_1$ está correlacionado con $v_2$.
- Consideremos la proyección lineal de $u_1$ sobre $v_2$:

$$u_1 = \rho_1 v_2 + e_1 \qquad (21)$$

donde $\rho_1 = \mathrm{E}\left(v_2 u_1\right)/\mathrm{E}\left(v_2^2\right)$ es el coeficiente poblacional.

# Función de Control

- Por definición $\mathrm{E}(v_2 e_1) = 0$, y $\mathrm{E}(\mathbf{z}' e_1') = \mathbf{0}$ por que $u_1$ y $v_2$ no estan correlacionados con $\mathbf{z}$.
- Reemplazando (21) en (18) tenemos

$$y_1 = \mathbf{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1 \tag{22}$$

  donde $v_2$ aparece como variable explicativa.
- Note que $e_1$ no está correlacionado con $v_2$ ni con $\mathbf{z}$ y como $y_2$ es una función lineal de $\mathbf{z}$ y $v_2$, $e_1$ tampoco está correlacionado con $y_2$.
- La ecuación (22) sugiere una forma de estimar en forma consistente $(\boldsymbol{\delta}_1, \alpha_1)$: MCC
- Único problema: no observamos $v_2$
- El procedimiento entonces involucra dos pasos:
  1. Estimar por MCC la ecuación (19) y construir los residuos $\hat{v}_2$
  2. Estimar por MCC la ecuación (22) reemplazando $v_2$ por $\hat{v}_2$.

# Función de Control

- Note que en el procedimiento anterior tenemos

$$
\begin{aligned}
y_1 &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1 \\
&= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1(y_2 - \mathbf{z}\pi_2) + e_1 \\
&= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1(y_2 - \mathbf{z}\pi_2 \pm \mathbf{z}\hat{\pi}_2) + e_1 \\
&= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1(y_2 - \mathbf{z}\hat{\pi}_2) + e_1 + \rho_1(\mathbf{z}\hat{\pi}_2 - \mathbf{z}\pi_2) \\
&= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1\hat{v}_2 + \text{error} \qquad (23)
\end{aligned}
$$

  donde error $= e_1 + \rho_1(\mathbf{z}\hat{\pi}_2 - \mathbf{z}\pi_2)$ depende del error muestral de $\hat{\pi}_2$, salvo que $\rho_1 = 0$.
- Esto implica que la matriz de varianzas y covarianzas de los coeficientes estimados en el segundo paso del procedimiento deberá tomar en cuenta este error muestral.
- La estimación por MCC de (23) es una ejemplo de **estimador de la función de control**.
- La inclusión de los residuos $\hat{v}_2$ "controla" por la endogeneidad de $y_2$ en la ecuación original (aunque lo hace con error muestral porque $\pi_2 \neq \hat{\pi}_2$).

# Función de Control

- Como comparan los enfoques de 2SLS y función de control?
- Se puede mostrar algebraicamente que los estimadores de $(\boldsymbol{\delta}_1, \alpha_1)$ son exactamente los mismos.
- La ecuación (23) provee otra forma de contrastar por endogeneidad: $H_0 : \rho_1 = 0$ versus $H_1 : \rho_1 \neq 0$ con un estadístico $t$ de significatividad individual.
- Se puede mostrar que este contrate es igual al test de Hausman que describimos para variables instrumentales.

# Agenda

# La Política

- Recordemos el ejemplo de una política educativa que consiste en reducir el tamaño de las clases en la educación primaria y su efecto sobre el aprendizaje de los alumnos.
- La política es tener clases con menos alumnos por profesor. La variable que mide el aprendizaje son las notas en pruebas estandarizadas de fin de año.
- En la práctica, la implementación más común de esta medida se realiza en dos etapas:
  1. 1ra Etapa: se eligen aleatoriamente algunas escuelas de la población de escuelas.
  2. 2da Etapa: Se asignan aleatoriamente las escuelas elegidas en la primera etapa a los grupos de tratamiento y de control.
- En este contexto, los alumnos de una misma escuela o clase tienden a tener puntajes en las pruebas que están correlacionados ya que están sujetos a algunas de las mismas influencias ambientales y de origen familiar.

- En términos matemáticos el modelo es

$$y_i = \alpha + \beta s_i + u_i \tag{24}$$

donde $y_i$ es la nota del alumno $i$ y $s_i$ es la asignación aleatoria de la política.

- Hasta ahora, para hacer inferencia estadística, supusimos que $u_i$ tenía varianza constante y covarianzas iguales a cero.
- Lo que sucede con la implementación de la política es que ahora la $cov(u_i, u_j) \neq 0$ si los alumnos $i$ y $j$ pertenecen a la misma escuela.
- ¿Qué sucede con la inferencia en este modelo si pasa esto?

# Agenda

# Propiedades de MCC

- Considere el siguiente modelo,

$$y = x\beta + u, \;\; E(u|x) = 0, \;\; E(uu'|x) = \sigma^2\Omega. \tag{25}$$

- El estimador de MCC es,

$$\begin{aligned}\hat{\beta} &= (x'x)^{-1}x'y \\ &= \beta + (x'x)^{-1}x'u\end{aligned} \tag{26}$$

- Los estimadores de MCC son insesgados y consistentes,

$$\begin{aligned}E(\hat{\beta}|x) &= \beta + (x'x)^{-1}x'E(u|x) \\ &= \beta\end{aligned} \tag{27}$$

y aplicando la ley de expectativas iteradas, $E(\hat{\beta}) = \beta$.

- $$\begin{aligned}\hat{\beta} &= \beta + \left(n^{-1}\sum_{i=1}^{n} x_i' x_i\right)^{-1} n^{-1}\sum_{i=1}^{n} x_i' u_i \\ &\xrightarrow{p} \beta + [E(x'x)]^{-1}E(x'u) \\ &\xrightarrow{p} \beta \end{aligned} \tag{28}$$

  donde la primera convergencia se obtiene aplicando la WLLN y la segunda sigue del supuesto de exogeneidad estricta.
- Apropiadamente re-escalado, los estimadores de MCC son asintóticamente normales,

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta) &= + \left(n^{-1}\sum_{i=1}^{n} x_i' x_i\right)^{-1} n^{-1/2}\sum_{i=1}^{n} x_i' u_i \\ &\xrightarrow{d} \text{Normal}(0, [E(x'x)]^{-1}E(x'uu'x)[E(x'x)]^{-1})\end{aligned}$$

# Propiedades de MCC

- La matriz de varianzas y covarianzas asintótica de $\hat{\beta}$ puede escribirse como,

$$\begin{aligned} Var[\sqrt{n}(\hat{\beta}-\beta)] &= [E(x'x)]^{-1}E[E(x'uu'x|x)][E(x'x)]^{-1} \\ &= [E(x'x)]^{-1}E[x'E(uu'|x)x][E(x'x)]^{-1} \\ &= \sigma^2[E(x'x)]^{-1}E[x'\Omega x][E(x'x)]^{-1} \end{aligned} \quad (29)$$

- Note que si la varianza de los errores fuera esférica, la ecuación anterior se reduce a la matriz de varianzas y covarianzas asintótica que obtuvimos para MCC: $\sigma^2[E(x'x)]^{-1}$.
- Entonces, una consecuencia de la heterocedasticidad y/o de la correlación serial es que matriz de varianzas y covarianzas asintótica convencional de MCC es incorrecta.

# Agenda

- La inferencia estadística convencional de MCC no es válida en presencia de errores no esféricos.
- Para poder hacer inferencia estadística en este modelo necesitamos tests estadísticos robustos ante la presencia de heterocedasticidad y/o correlación serial.
- Para esto necesitamos estimar la matriz de varianzas y covarianzas asintótica de MCC correcta,

$$\widehat{Var(\hat{\beta})} = \left(\sum_{i=1}^{n} x_i' x_i\right)^{-1} \left(\sum_{i=1}^{n} x_i' \hat{u}_i \hat{u}_i' x_i\right) \left(\sum_{i=1}^{n} x_i' x_i\right)^{-1} \tag{30}$$

donde $\hat{u}$ son los residuos de la estimación por MCC. Esta es la matriz de varianzas y covarianzas robusta ante la presencia de heterocedasticidad y correlación serial de White (1980).

- La raiz cuadrada de los elementos de la diagonal principal de (30) son los errores estándar robustos. Los estadísticos $t$ se calculan de la forma usual con estos errores estándar robustos.
- Para realizar contrastes sobre combinación lineal de coeficientes el estadístico de Wald se construye con la fórmula habitual,

$$W = (R\hat{\beta} - r)'[R\widehat{Var(\hat{\beta})}R']^{-1}(R\hat{\beta} - r)/\#r \tag{31}$$

donde $\widehat{Var(\hat{\beta})}$ está definida por la ecuación (30)

# Agenda

# Mínimos Cuadrados Generalizados

- Una alternativa a la estimación por MCC y la inferencia estadística robusta es utilizar el método de mínimos cuadrados generalizados (MCG).
- Considere la estimación del mismo modelo que antes,

$$y = x\beta + u, \;\; E(u|x) = 0, \;\; E(uu'|x) = \sigma^2\Omega.$$

  y asumamos por un momento que los elementos de $\Omega$ son conocidos.
- Como $\Omega$ es definida positiva, su inversa también lo es. Por lo tanto, es posible encontrar una matriz no singular $P$ tal que:

$$\Omega^{-1} = P'P. \tag{32}$$

- Premultiplicando (25) por $P$ se obtiene,

$$y_* = x_*\beta + u_*, \tag{33}$$

  donde $y_* = Py$, $x_* = Px$ y $u_* = Pu$.

# Mínimos Cuadrados Generalizados

- El modelo transformado cumple con el supuesto de exogeneidad estricta,

$$\begin{aligned} E(u_*|x_*) &= E(u_*|x) = E(Pu|x) \\ &= PE(u|x) = 0 \end{aligned}$$

- Además, de acuerdo con (32), $\Omega = P^{-1}(P')^{-1}$. Por lo tanto,

$$\begin{aligned} Var(u_*) &= E(Puu'P') \\ &= \sigma^2 P\Omega P' \\ &= \sigma^2 PP^{-1}(P')^{-1}P' \\ &= \sigma^2 I_n \end{aligned} \tag{34}$$

y el modelo transformado cumple con los supuestos del modelo de regresión lineal múltiple y puede ser estimado por MCC.

- El método de MCG minimiza,

$$\begin{aligned} RSS(\hat{\beta}) &= (y_* - x_*\hat{\beta})'(y_* - x_*\hat{\beta}) \\ &= (y_*' - \hat{\beta}'x_*')(y_* - x_*\hat{\beta}) \\ &= y_*'y_* - \hat{\beta}'x_*'y_* - y_*'x_*\hat{\beta} + \hat{\beta}'x_*'x_*\hat{\beta} \\ &= y_*'y_* - 2\hat{\beta}'x_*'y_* + \hat{\beta}'x_*'x_*\hat{\beta} \end{aligned} \tag{35}$$

- El estimador de MCG es,

$$\begin{aligned} \hat{\beta} &= (x_*'x_*)^{-1}x_*'y_* \\ &= (x'\Omega^{-1}x)^{-1}x'\Omega^{-1}y \end{aligned} \tag{36}$$

- y usando la teoría desarrollada para MCC,

$$Var(\hat{\beta}) = \sigma^2(x_*'x_*)^{-1} = \sigma^2(x'\Omega^{-1}x)^{-1},$$

# Mínimos Cuadrados Generalizados

- Una estimación consistente de la varianza del estimador de MCG es,

$$\widehat{Var(\hat{\beta})} = s^2(x_*'x_*)^{-1} = s^2(x'\Omega^{-1}x)^{-1}, \tag{37}$$

con,

$$\begin{aligned} s^2 &= (y_* - x_*\hat{\beta})'(y_* - x_*\hat{\beta})/(n-K) \\ &= [P(y - x\hat{\beta})]'[P(y - x\hat{\beta})]/(n-K) \\ &= [(y - x\hat{\beta})]'\Omega^{-1}[(y - x\hat{\beta})]/(n-K). \end{aligned}$$

- La raiz cuadrada de los elementos de la diagonal principal de (37) son los errores estándar de los estimadores de MCG y pueden utilizarse para construir los estadísticos $t$.
- Restricciones lineales del tipo $H_0 : R\beta = r$ pueden contrastarse utilizando el test de Wald,

$$W = (R\hat{\beta} - r)'[Rs^2(x'\Omega^{-1}x)^{-1}R']^{-1}(R\hat{\beta} - r)/\#r \tag{38}$$

# Mínimos Cuadrados Generalizados

- En términos generales, los resultados de MCG pueden resumirse de la siguiente manera.
- Para mostrar consistencia necesitamos reforzar el supuesto 3 de no singularidad.
  Supuesto 3': $\Omega$ es positiva definida y $E(x'\Omega^{-1}x)$ no es singular.
- Usando la WLLN,

$$n^{-1}\sum_{i=1}^{n} x_i'\Omega x_i \xrightarrow{p} E(x'\Omega^{-1}x) \equiv A$$

y por el supuesto 3',

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

# Mínimos Cuadrados Generalizados

- Usando la WLLN y el supuesto de exogeneidad estricta,

$$n^{-1}\sum_{i=1}^{n} x_i'\Omega u_i \xrightarrow{p} E(x'\Omega^{-1}u) = 0$$

- Prueba:

$$\begin{aligned} vec[E(x'\Omega^{-1}u)] &= [E(u' \otimes x')]vec(\Omega^{-1}) \\ &= [E(u \otimes x)']vec(\Omega^{-1}) = 0 \end{aligned}$$

- Por lo tanto,

$$\begin{aligned} \hat{\beta} &= \beta + (n^{-1}x'\Omega^{-1}x)^{-1}n^{-1}x'\Omega^{-1}u \\ &\xrightarrow{p} \beta + A^{-1}E(x'\Omega^{-1}u) = \beta \end{aligned} \tag{39}$$

y el estimador de MCG es consistente.

# Mínimos Cuadrados Generalizados

- La distribución asintótica se obtiene desde,

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta) &= (n^{-1}x'\Omega^{-1}x)^{-1}n^{-1/2}x'\Omega^{-1}u \\ &\xrightarrow{d} A^{-1}\text{Normal}(0,\, \sigma^2 A) \\ &\xrightarrow{d} \text{Normal}(0,\, \sigma^2 A^{-1}) \end{aligned} \tag{40}$$

donde se utilizó el CLT para,

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

con $A \equiv E(x'\Omega^{-1}x)$.

- Por lo tanto, la matriz de varianzas y covarianzas asintótica del estimador de MCG es,

$$Var(\hat{\beta}) = \sigma^2 A^{-1}/n$$

# Mínimos Cuadrados Generalizados Estimados

- Remark: Note que para poder obtener la consistencia del estimador de MCG es necesario asumir que las variables explicativas son estríctamente exógenas. Este supuesto es más fuerte que el necesario para obtener consistencia de MCC que es, como vimos, exogeneidad contemporánea.
- Hasta ahora se asumió que $\Omega$ era conocida. En general, en la práctica, esto no es así y se necesita una estimación consistente de la misma.
- El método que utiliza una estimación consistente de $\Omega$ es conocido como mínimos cuadrados generalizados estimados (MCGE) ó feasible generalized least squares (FGLS).
- Vamos a mostrar, en lo que sigue, que MCGE es asintóticamente equivalente a MCG.

- El estimador de MCGE es,

$$
\begin{aligned}
\tilde{\beta} &= (x'\hat{\Omega}^{-1}x)^{-1}x'\hat{\Omega}^{-1}y \\
&= \beta + (x'\hat{\Omega}^{-1}x)^{-1}x'\hat{\Omega}^{-1}u \\
&= \beta + (n^{-1}\sum_{i=1}^{n} x_i'\hat{\Omega}^{-1}x_i)^{-1}n^{-1}\sum_{i=1}^{n} x_i'\hat{\Omega}^{-1}u_i, \Rightarrow \\
\sqrt{n}(\hat{\beta}-\beta) &= (n^{-1}\sum_{i=1}^{n} x_i'\hat{\Omega}^{-1}x_i)^{-1}n^{-1/2}\sum_{i=1}^{n} x_i'\hat{\Omega}^{-1}u_i \qquad (41)
\end{aligned}
$$

- Comparando el segundo término de (41) con el correspondiente a MCG tenemos,

$$
n^{-1/2}\sum_{i=1}^{n} x_i'\hat{\Omega}^{-1}u_i - n^{-1/2}\sum_{i=1}^{n} x_i'\Omega^{-1}u_i = n^{-1/2}\sum_{i=1}^{n} x_i'(\hat{\Omega}^{-1}-\Omega^{-1})u_i
$$

# Mínimos Cuadrados Generalizados Estimados

- Note que la ecuación anterior puede escribirse como,

$$\left[n^{-1/2}\sum_{i=1}^{n}(u_i \otimes x_i)'\right] vec(\hat{\Omega}^{-1} - \Omega^{-1}) \quad = \quad o_p(1) \Rightarrow$$

$$n^{-1/2}\sum_{i=1}^{n} x_i'\hat{\Omega}^{-1}u_i = n^{-1/2}\sum_{i=1}^{n} x_i'\Omega^{-1}u_i + o_p(1) \tag{42}$$

- Usando el mismo argumento,

$$n^{-1}\sum_{i=1}^{n} x_i'\hat{\Omega}^{-1}x_i = n^{-1}\sum_{i=1}^{n} x_i'\Omega^{-1}x_i + o_p(1) \tag{43}$$

- Usando los resultados anteriores tenemos,

$$\begin{aligned}\sqrt{n}(\tilde{\beta}-\beta) &= \left(n^{-1}\sum_{i=1}^{n} x_i'\Omega^{-1}x_i\right)\left(n^{-1/2}\sum_{i=1}^{n} x_i'\Omega^{-1}u_i\right) + o_p(1) \\ &= \sqrt{n}(\hat{\beta}-\beta) + o_p(1) \\ &\Rightarrow \sqrt{n}(\tilde{\beta}-\hat{\beta}) = o_p(1) \qquad (44)\end{aligned}$$

  que los estimadores de MCG y de MCGE son asintóticamente equivalentes ($\sqrt{n}$ asintóticamente equivalentes).

- Empíricamente, la equivalencia asintótica de los estimadores de MCG y MCGE implica que para realizar inferencia estadística sobre $\beta$ usando MCGE, no hay que preocuparse de que $\hat{\Omega}$ sea un estimador de $\Omega$.
- En otras palabras,

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

# Mínimos Cuadrados Generalizados Estimados

- Bajo MCGE la estimación de $A$ viene dada por,

$$\tilde{A} \equiv n^{-1} \sum_{i=1}^{n} x_i' \hat{\Omega} x_i$$

- En general la estimación de $\Omega$ se hace imponiendo alguna estructura en la matriz (e.g. heterocedasticidad o correlación serial).
- Por ejemplo, si hay heterocedasticidad, una estimación consistente de $A$ es,

$$\tilde{A} = \hat{\Gamma}_0 + \sum_{j=1}^{m} \omega(j, m)(\hat{\Gamma}_j + \hat{\Gamma}_j'),$$

donde,

$$\hat{\Gamma}_j = \frac{1}{n} \sum_{i=j+1}^{n} x_i' \hat{u}_i \hat{u}_{i-j}' x_{i-j}, \;\; j = 0, 1, 2, \ldots, m$$

# Mínimos Cuadrados Generalizados Estimados

- $\omega(j, m)$ es una ventana a definir y $m$ es un parámetro de truncamiento.
- Por ejemplo,

$$\omega(j,m) = \begin{cases} 1 & j = 1,\, 2,\, \ldots,\, m. \qquad \text{uniform window} \\ 1 - \frac{j}{m+1}, & j = 1, 2, \ldots, m. \text{ Bartlett (Newey-West) window} \end{cases}$$

- $\hat{u}_i$ $(i = 1, 2, \ldots, n)$ son los residuos de la estimación por MCC.
- Cómo detectar la presencia de heterocedasticidad y/ó correlación serial en el modelo?
- Vamos a desarrollar el test de White para detectar heterocedasticidad y el test de Breusch-Godfrey para detectar correlación serial.

- Test de White

$$
\begin{aligned}
H_0 &: \text{No existe heterocedasticidad} \\
H_1 &: \text{Existe heterocedasticidad}
\end{aligned}
$$

Este contraste asume que la forma funcional de la heterocedasticidad es lineal en todas las variables explicativas del modelo, sus cuadrados y sus productos cruzados.

- Por ejemplo, suponga el siguiente modelo,

$$y_i = \alpha_0 + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + u_i$$

- La forma funcional de la heterocedasticidad de White es,

$$\sigma_i^2 = \gamma_0 + \gamma_1 x_{i,1} + \gamma_2 x_{i,2} + \gamma_3 x_{i,1}^2 + \gamma_4 x_{i,2}^2 + \gamma_5 x_{i,1} x_{i,2}$$

# Mínimos Cuadrados Generalizados Estimados

- En este modelo la hipótesis nula se puede escribir como: $H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_5 = 0$, y la alternativa como, $H_1$ : al menos un $\gamma_i \neq 0$, $i = 1, 2, \ldots, 5$.
- El procedimiento de White es como sigue:
  1. Estimar el modelo original por MCC y obtener la serie de residuos y de sus cuadrados.
  2. Estimar la ecuación de la forma funcional de la varianza reemplazando la variable dependiente por la serie de residuos al cuadrado obtenida en el paso anterior.
  3. Construir el estadístico $LM = n \times R^2 \sim \chi^2_q$. El $R^2$ es el de la regresión del Paso 2 y $q$ es el número de parámetros iguales a cero en la hipótesis nula.

- Suponga el siguiente modelo estructural,

$$y_i = \alpha_0 + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + u_i \tag{45}$$

- Entonces, la forma funcional de la heterocedasticidad de White es,

$$\sigma_i^2 = \gamma_0 + \gamma_1 x_{i,1} + \gamma_2 x_{i,2} + \gamma_3 x_{i,1}^2 + \gamma_4 x_{i,2}^2 + \gamma_5 x_{i,1} x_{i,2} \tag{46}$$

- FGLS White
  1. Estimar (45) por OLS y obtener las estimaciones de los parámetros del modelo.
  2. Calcular los residuos del modelo y elevarlos al cuadrado, $\hat{u}_i^2$.
  3. Estimar (46) por OLS usando $\hat{u}_i^2$ como proxy de $\sigma_i^2$.
  4. Usar las estimaciones de la regresión auxiliar y obtener las variancias ajustadas como:

$$\widehat{\hat{u}_i^2} \equiv \hat{\sigma}_i^2 = \hat{\gamma}_0 + \hat{\gamma}_1 x_{i,1} + \hat{\gamma}_2 x_{i,2} + \hat{\gamma}_3 x_{i,1}^2 + \hat{\gamma}_4 x_{i,2}^2 + \hat{\gamma}_5 x_{i,1} x_{i,2}$$

- FGLS White (Cont.)
  5. Transformar las variables de (46) dividiendolas por $\hat{\sigma}_i^2$ y estimar por OLS,

$$\frac{\hat{u}_i^2}{\hat{\sigma}_i^2} = \gamma_0 \frac{1}{\hat{\sigma}_i^2} + \gamma_1 \frac{x_{i,1}}{\hat{\sigma}_i^2} + \gamma_2 \frac{x_{i,2}}{\hat{\sigma}_i^2} + \gamma_3 \frac{x_{i,1}^2}{\hat{\sigma}_i^2} + \gamma_4 \frac{x_{i,2}^2}{\hat{\sigma}_i^2} + \gamma_5 \frac{x_{i,1}x_{i,2}}{\hat{\sigma}_i^2} + \nu_i$$

  6. Con los $\tilde{\gamma}$'s de la estimación anterior calcular

$$\widehat{\frac{\hat{u}_i^2}{\hat{\sigma}_i^2}} \equiv \tilde{\sigma}_i^2 = \tilde{\gamma}_0 + \tilde{\gamma}_1 x_{i,1} + \tilde{\gamma}_2 x_{i,2} + \tilde{\gamma}_3 x_{i,1}^2 + \tilde{\gamma}_4 x_{i,2}^2 + \tilde{\gamma}_5 x_{i,1} x_{i,2}$$

  7. Usar $\frac{1}{\tilde{\sigma}_i}$ como ponderadores para estimar (45).

- Correlación serial de orden uno

  Supongamos el mismo modelo de antes,

$$y_i = \alpha_0 + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + u_i$$

En este modelo, la correlación serial de orden uno se especifica como:

$$u_i = \rho u_{i-1} + \epsilon_i, \; |\rho| < 1.$$

- Test de Breusch-Godfrey

$$H_0 : \text{No existe correlación serial de orden uno } (\rho = 0)$$
$$H_1 : \text{Existe correlación serial de orden uno } (\rho \neq 0)$$

- El procedimiento de Breusch-Godfrey es como sigue:
  1. Estimar el modelo original por MCC y obtener la serie de residuos y la serie de residuos rezagada una observación.
  2. Estime una ecuación auxiliar que tenga como variable dependiente a la serie de residuos del paso anterior y como variables independientes a todas las variables explicativas del modelo original más la serie de residuos rezagada una observación.
  3. Construir el estadístico $LM = (n-1) \times R^2 \sim \chi^2_{q=1}$. El $R^2$ es el de la regresión del Paso 2 y $q = 1$ es el número de parámetros iguales a cero en la hipótesis nula.
- Si se rechazara la hipótesis nula una estimación consistente viene dada por el procedimiento de Cochrane-Orcutt.

- Procedimiento de Cochrane-Orcutt
  Considere el mismo modelo de los ejemplos anteriores,

$$y_i = \alpha_0 + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + u_i \tag{47}$$

$$u_i = \rho u_{i-1} + \epsilon_i, \ |\rho| < 1. \tag{48}$$

1. Estimar el modelo original (47) por MCC y obtener la serie de residuos, $\hat{u}_i = y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_{i,1} - \hat{\alpha}_2 x_{i,2}$, y la serie de residuos rezagada una observación, $\hat{u}_{i-1}$.
2. Estimar la ecuación (48) por MCC, reemplazando a $u_i$ y $u_{i-1}$ por $\hat{u}_i$ y $\hat{u}_{i-1}$, respectivamente. Obtener $\hat{\rho}$.
3. Transforme las variables del siguiente modo: $y_i^* = y_i - \hat{\rho} y_{i-1}$, $x_{i,1}^* = x_{i,1} - \hat{\rho} x_{i-1,1}$, $x_{i,2}^* = x_{i,2} - \hat{\rho} x_{i-1,2}$ y $c^* = 1 - \hat{\rho}$.
4. Estimar por MCC el modelo transformado,

$$y_i^* = \alpha_0 c^* + \alpha_1 x_{i,1}^* + \alpha_2 x_{i,2}^* + u_i^*$$

y obtenga nuevos estimadores $\hat{\hat{\alpha}}_0$, $\hat{\hat{\alpha}}_1$, $\hat{\hat{\alpha}}_2$.

# Mínimos Cuadrados Generalizados Estimados

- Procedimiento de Cochrane-Orcutt (continuación)
  5. Construir nuevos residuos $\hat{\hat{u}}_i = y_i - \hat{\hat{\alpha}}_0 - \hat{\hat{\alpha}}_1 x_{i,1} - \hat{\hat{\alpha}}_2 x_{i,2}$ y sus rezagos $\hat{\hat{u}}_{i-1}$.
  6. Vuelva al Paso 2 y repita el procedimiento.
  7. El procedimiento termina cuando en dos iteraciones sucesivas el valor estimado para $\rho$ es el mismo.
- Los estimadores obtenidos con este procedimiento son insesgados, consistentes y asintóticamente eficientes.
- Remark: Tanto el test de Breusch-Godfrey como la corrección de Cochrane-Orcutt pueden generalizarse a correlación serial de orden mayor a uno.

# Econometría

## Maestría en Economía - Maestría en Econometría

Lecture 3

# Agenda

# Agenda

1. Estimación por Máxima Verosimilitud
   - Introducción
   - Estimación
   - Propiedades
   - Máxima Verosimilitud: MCC
   - Máxima Verosimilitud: Inferencia Estadística
   - La Motivación Económica
2. Costos Económicos de Nacer con Peso Bajo
3. Modelo de Variable Dependiente Binaria e Inferencia Causal

# Definiciones

- Suponga que $x = (x_1, x_2, \ldots, x_n)'$ es una muestra aleatoria de

$$f(x_t; \theta), \quad t = 1, 2, \ldots, n$$

- Definamos:
  - Función de Verosimilitud (FV): $L(x; \theta) = f(x_1, x_2, \ldots, x_n; \theta)$
  - Logaritmo de la FV: $ln[L(x; \theta)]$
  - Función Score: $\frac{\partial ln[L(x;\theta)]}{\partial \theta}$
- El estimador de Máxima Verosimilitud (MLE) se obtiene de:

$$\hat{\theta}_{MLE} : argmax_{\theta \in \Theta} ln[L(x; \theta)]$$

# Agenda

1. Estimación por Máxima Verosimilitud
   - Introducción
   - Estimación
   - Propiedades
   - Máxima Verosimilitud: MCC
   - Máxima Verosimilitud: Inferencia Estadística
   - La Motivación Económica
2. Costos Económicos de Nacer con Peso Bajo
3. Modelo de Variable Dependiente Binaria e Inferencia Causal

# Máxima Verosimilitud: Estimación

- En el caso de que la función de verosimilitud sea diferenciable, entonces el estimador MLE se obtiene de la solución a las siguientes ecuaciones:
  a) $\frac{\partial ln[L(x;\theta)]}{\partial\theta} = 0$
  b) $H(\theta) = \frac{\partial^2 ln[L(x;\theta)]}{\partial\theta^2}|_{\theta=\hat{\theta}} < 0$
- Ejemplo: suponga que $(x_1, x_2, \ldots, x_n)$ es una muestra aleatoria de una $N(\mu, \sigma^2)$. Definamos a $\theta = (\mu, \sigma^2)'$. Entonces,

$$\begin{aligned}
L(x;\theta) &= f(x_1, x_2, \ldots, x_n; \theta) = \prod_{t=1}^{n} f(x_t; \theta) \\
&= \prod_{t=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(x_t-\mu)^2] \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp[-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(x_t-\mu)^2]
\end{aligned}$$

# Máxima Verosimilitud: Estimación

- Por lo tanto,

$$ln[L(x;\theta)] = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(x_t - \mu)^2$$

- Las condiciones de primer orden vienen dadas por,

$$\begin{aligned}
\frac{\partial ln[L(x;\theta)]}{\partial \mu} &= -\frac{1}{\sigma^2}\sum_{t=1}^{n}(x_t - \mu) = 0 \Longrightarrow \hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} x_t \\
\frac{\partial ln[L(x;\theta)]}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{n}(x_t - \mu)^2 = 0
\end{aligned}$$

- De la segunda ecuación tenemos,

$$\Longrightarrow \sigma^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t-\mu)^2$$

$$\Longrightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t-\hat{\mu})^2$$

- Las condiciones de segundo orden son,

$$\frac{\partial^2 ln[L(x;\theta)]}{\partial\mu^2} = -\frac{n}{\sigma^2}; \quad \frac{\partial^2 ln[L(x;\theta)]}{\partial\sigma^4} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{n}(x_t-\mu)^2$$

y

$$\frac{\partial^2 ln[L(x;\theta)]}{\partial\mu\partial\sigma^2} = -\frac{1}{\sigma^4}\sum_{t=1}^{n}(x_t-\mu)$$

# Máxima Verosimilitud: Estimación

- Evaluando las derivadas segundas en $\theta = \hat{\theta}$ tenemos:

$$\sum_{t=1}^{n}(x_t - \hat{\mu}) = 0 \Longrightarrow \frac{\partial^2 ln[L(x;\theta)]}{\partial\mu\partial\sigma^2}|_{\theta=\hat{\theta}} = 0$$

$$\sum_{t=1}^{n}(x_t - \hat{\mu})^2 = n\hat{\sigma}^2 \Longrightarrow \frac{\partial^2 ln[L(x;\theta)]}{\partial\sigma^4}|_{\theta=\hat{\theta}} = -\frac{n}{2\hat{\sigma}^4}$$

- La matriz hessiana queda entonces,

$$H(\hat{\theta}) = \begin{bmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{bmatrix}$$

$-\frac{n}{\hat{\sigma}^2} < 0$ y determinante de $H(\hat{\theta}) = \frac{n^2}{2\hat{\sigma}^6} > 0$ y $H(\hat{\theta})$ es negativa definida.

# Máxima Verosimilitud: Propiedades

- Invariancia: Sea $g : \Theta \to \mathbb{R}^k$ y $\hat{\theta}$ el estimador MLE de $\theta$ entonces $g(\hat{\theta})$ es el MLE de $g(\theta)$.
- En general los estimadores MLE no son insesgados. Ejemplo,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2$$

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-1)$$

- Por lo tanto,

$$E(\frac{n\hat{\sigma}^2}{\sigma^2}) = n-1 \Longrightarrow \frac{n}{\sigma^2}E(\hat{\sigma}^2) = n-1 \Longrightarrow E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

y

$$E(\hat{\mu}) = E(\frac{1}{n}\sum_{t=1}^{n}x_t) = \mu$$

# Máxima Verosimilitud: Propiedades

- Eficiencia. Teorema de Cramer-Rao: Sea $\hat{\theta}$ un estimador insesgado de $\theta = (\theta_1, \theta_2, \ldots, \theta_k)'$. Entonces bajo ciertas condiciones de regularidad,

$$Var(\hat{\theta}) - I_n(\theta)^{-1} \geq 0$$

  donde

$$I_n(\theta) = E\left[\left(\frac{\partial ln[L(x;\theta)]}{\partial\theta}\right)\left(\frac{\partial ln[L(x;\theta)]}{\partial\theta}\right)'\right] = E\left[-\frac{\partial^2 ln[L(x;\theta)]}{\partial\theta\partial\theta'}\right]$$

  es la matriz de información de Fisher.

-

$$I_n(\theta) = -E\begin{bmatrix} \frac{\partial^2 ln[L(x;\theta)]}{\partial\theta_1^2} & \frac{\partial^2 ln[L(x;\theta)]}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 ln[L(x;\theta)]}{\partial\theta_1\partial\theta_k} \\ \frac{\partial^2 ln[L(x;\theta)]}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 ln[L(x;\theta)]}{\partial\theta_2^2} & \cdots & \frac{\partial^2 ln[L(x;\theta)]}{\partial\theta_2\partial\theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 ln[L(x;\theta)]}{\partial\theta_k\partial\theta_1} & \frac{\partial^2 ln[L(x;\theta)]}{\partial\theta_k\partial\theta_2} & \cdots & \frac{\partial^2 ln[L(x;\theta)]}{\partial\theta_k^2} \end{bmatrix}$$

# Máxima Verosimilitud: Propiedades

- Por lo tanto $Var(\hat{\theta}_i) \geq I_n(\theta)^{-1}_{ii}$.
- Un estimador insesgado se dice completamente eficiente si su varianza "alcanza la cota de Cramer-Rao" $Var(\hat{\theta}_i) = I_n(\theta)^{-1}_{ii}$.
- Si el estimador es sesgado, entonces su eficiencia se calcula con el MSE

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + Bias(\hat{\theta})^2$$

- en este caso, $MSE(\hat{\theta}) \geq CRLB(\theta)$,

$$CRLB(\theta) = \left(1 + \frac{\partial Bias(\theta)}{\partial \theta}\right)^2 / E\left[-\frac{\partial^2 ln[L(x;\theta)]}{\partial \theta^2}\right]$$

- Ejemplo (continuación):

$$H(\theta) = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{t=1}^{n}(x_t-\mu) \\ -\frac{1}{\sigma^4}\sum_{t=1}^{n}(x_t-\mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{n}(x_t-\mu)^2 \end{bmatrix}$$

- Por definición, $I_n(\theta) = -E[H(\theta)]$. Entonces,

$$E(-\frac{n}{\sigma^2}) = -\frac{n}{\sigma^2}, \quad E[-\frac{1}{\sigma^4}\sum_{t=1}^{n}(x_t-\mu)] = -\frac{1}{\sigma^4}\sum_{t=1}^{n}E(x_t-\mu) = 0$$

y

$$\begin{aligned} E[\frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{n}(x_t-\mu)^2] &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{n}E(x_t-\mu)^2 \\ &= \frac{n}{2\sigma^4} - \frac{n\sigma^2}{\sigma^6} = -\frac{n}{2\sigma^4} \end{aligned}$$

# Máxima Verosimilitud: Propiedades

- $$I_n(\theta) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}, \quad I_n(\theta)^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$
- Como $E(\hat{\mu}) = E(\frac{1}{n}\sum_{t=1}^{n} x_t) = \mu$ podemos chequear que,

  $$Var(\hat{\mu}) = \frac{1}{n^2}\sum_{t=1}^{n} Var(x_t) = \frac{\sigma^2}{n} = I_n(\theta)^{-1}_{11}$$

  y el estimador de $\mu$ es completamente eficiente.
- Para el estimador de la varianza tenemos $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{1}{n}\sigma^2$.
- Por lo tanto,

  $$\frac{\partial Bias}{\partial \sigma^2} = -\frac{1}{n}$$

- Entonces,

$$CRLB = (1 - 1/n)^2 / \frac{n}{2\sigma^4} = \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n}$$

- Además, $Var(\hat{\sigma}^2) = \frac{2(n-1)\sigma^4}{n^2}$, por lo tanto

$$MSE = \frac{2(n-1)\sigma^4}{n^2} + (-\frac{1}{n}\sigma^2)^2 \neq CRLB$$

y el estimador de la varianza no es completamente eficiente.

- Alcanza $s^2$ (el estimador de la varianza de MCC) la cota de Cramer-Rao?
- Sabemos que $s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(x_t - \hat{\mu})^2$ es un estimador insesgado de $\sigma^2$.

# Máxima Verosimilitud: Propiedades

- Para saber si $s^2$ es completamente eficiente calculamos,

$$Var(\frac{(n-1)s^2}{\sigma^2}) = 2(n-1), \implies \frac{(n-1)^2}{\sigma^4} Var(s^2) = 2(n-1)$$

$$\implies Var(s^2) = \frac{2\sigma^4}{(n-1)} \neq I_n(\theta)^{-1}_{22}$$

y el estimador de la varianza de MCC tampoco es completamente eficiente.

- **Propiedades asintóticas:** Sea $\hat{\theta}_n$ el estimador MLE de $\theta$ entonces:
  a) $\hat{\theta}_n \xrightarrow{p} \theta$.
  b) $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N[0, V(\theta)]$.
  c) $V(\theta) = I(\theta)^{-1}$.

donde $I(\theta) = \lim_{n\to\infty}[\frac{1}{n} I_n(\theta)]$ es la matriz de información de Fisher asintótica.

- Estimadores de $I(\theta)$,

  (i)

$$-\frac{1}{n}\frac{\partial^2 ln[L(x,\hat{\theta})]}{\partial\theta\partial\theta'} \xrightarrow{p} I(\theta)$$

  (ii)

$$\frac{1}{n}\sum_{t=1}^{n}\left[\left(\frac{\partial ln[L(x,\hat{\theta})]}{\partial\theta}\right)\left(\frac{\partial ln[L(x,\hat{\theta})]}{\partial\theta}\right)'\right] \xrightarrow{p} I(\theta)$$

- Estimadores de $Var(\hat{\theta}_n)$

$$\left(\frac{\partial^2 ln[L(x,\hat{\theta})]}{\partial\theta\partial\theta'}\right)^{-1}$$

ó,

$$\left[\sum_{t=1}^{n}\left(\frac{\partial ln[L(x,\hat{\theta})]}{\partial\theta}\right)\left(\frac{\partial ln[L(x,\hat{\theta})]}{\partial\theta}\right)'\right]^{-1}$$

# Máxima Verosimilitud: Propiedades

- Ejemplo (continuación)

$$\hat{\theta}_n = \left[ \begin{array}{c} \hat{\mu} \\ \hat{\sigma}^2 \end{array} \right]$$

-
$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N[0, I(\theta)^{-1}]$$

-
$$I(\theta) = \lim_{n\to\infty}[\frac{1}{n}I_n(\theta)] = \lim_{n\to\infty}\frac{1}{n}\left[ \begin{array}{cc} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{array} \right] = \left[ \begin{array}{cc} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{array} \right]$$

-
$$\left[ \begin{array}{c} \sqrt{n}(\hat{\mu} - \mu) \\ \sqrt{n}(\hat{\sigma}^2 - \sigma^2) \end{array} \right] \xrightarrow{d} N\left[ \left( \begin{array}{c} 0 \\ 0 \end{array} \right), \left( \begin{array}{cc} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{array} \right) \right]$$

# Máxima Verosimilitud: Propiedades

- ó
$$\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} \sim AN\left[\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}\right]$$
- Note que el estimador de la varianza es asintóticamente eficiente.

# Agenda

# Máxima Verosimilitud: MCC

- **Ejemplo 2: MCC.**

$$y = x\beta + u, \quad u \sim N(0, \sigma^2 I_n) \quad \theta = (\beta', \sigma^2)'$$

- $$f(y) = f(u) = (2\pi\sigma^2)^{n/2} \exp[-\frac{1}{2\sigma^2}(y - x\beta)'(y - x\beta)] \equiv L(x; \theta)$$

- $$ln[L(x; \theta)] = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}(y'y - \beta'x'x\beta - 2\beta'x'y)$$

# Máxima Verosimilitud: MCC

- Condiciones de primer orden:

$$\frac{\partial ln[L(x;\theta)]}{\partial \beta} = -\frac{1}{2\sigma^2}(2x'x\beta - 2x'y) = \frac{1}{\sigma^2}(x'y - x'x\beta) = 0 \tag{1}$$

$$\frac{\partial ln[L(x;\theta)]}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - x\beta)'(y - x\beta) = 0 \tag{2}$$

- $$(1) \Longrightarrow x'y - x'x\beta = 0 \Longrightarrow \hat{\beta} = (x'x)^{-1}x'y \tag{3}$$

$$(2)y(3) \Longrightarrow \hat{\sigma}^2 = \frac{1}{n}(y - x\hat{\beta})'(y - x\hat{\beta}) = \frac{\hat{u}'\hat{u}}{n} \tag{4}$$

- Condiciones de segundo orden,

$$\frac{\partial^2 ln[L(x;\theta)]}{\partial\beta\partial\beta'} = -\frac{1}{\sigma^2}x'x; \qquad \frac{\partial^2 ln[L(x;\theta)]}{\partial\sigma^4} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}u'u \tag{5}$$

$$\frac{\partial^2 ln[L(x;\theta)]}{\partial\beta\partial\sigma^2} = -\frac{1}{2\sigma^4}(x'y - x'x\beta) = -\frac{1}{2\sigma^4}x'(y - x\beta) = -\frac{1}{2\sigma^4}x'u \tag{6}$$

- Esperanzas,

$$E\left(\frac{\partial^2 ln[L(x;\theta)]}{\partial\beta\partial\beta'}\right) = E(-\frac{1}{\sigma^2}x'x) = -\frac{1}{\sigma^2}x'x$$

$$E\left(\frac{\partial^2 ln[L(x;\theta)]}{\partial\sigma^4}\right) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}E(u'u) = \frac{n}{2\sigma^4} - \frac{n\sigma^2}{\sigma^6} = -\frac{n}{2\sigma^4}$$

- $$E\left(\frac{\partial^2 ln[L(x;\theta)]}{\partial\beta\partial\sigma^2}\right) = -\frac{1}{2\sigma^4}E(x'u) = 0 \tag{7}$$

- Por lo tanto, la matriz de información de Fisher muestral es,

$$I_n(\theta) = -E\left(\frac{\partial^2 ln[L(x;\theta)]}{\partial\theta\partial\theta'}\right) = \begin{bmatrix} \frac{1}{\sigma^2}x'x & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

- La inversa de la matriz de información de Fisher es,

$$I_n(\theta)^{-1} = \begin{bmatrix} \sigma^2(x'x)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

- Propiedades de los Estimadores
- $$E(\hat{\beta}) = \beta, \quad Var(\hat{\beta}) = \sigma^2 (x'x)^{-1} = I_n(\theta)^{-1}_{11}$$

  y el estimador de MV de $\beta$ es completamente eficiente.

  $$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k) \Longrightarrow E(\frac{n\hat{\sigma}^2}{\sigma^2}) = (n-k) \Longrightarrow E(\hat{\sigma}^2) = \frac{n-k}{n}\sigma^2 \neq \sigma^2$$

  $$E(\hat{\sigma}^2) = \frac{n-k}{n}\sigma^2 = \sigma^2 - \frac{k}{n}\sigma^2 \Longrightarrow Bias(\hat{\sigma}^2) = -\frac{k}{n}\sigma^2$$

- Entonces,

  $$Var(\frac{n\hat{\sigma}^2}{\sigma^2}) = 2(n-k) \Longrightarrow Var(\hat{\sigma}^2) = \frac{2(n-k)}{n^2}\sigma^4$$

# Máxima Verosimilitud: MCC

- y el MSE es,

$$MSE(\hat{\sigma}^2) = \frac{2(n-k)}{n^2}\sigma^4 + (-\frac{k}{n}\sigma^2)^2 = \frac{2(n-k)+nk^2}{n^2}\sigma^4$$

- La cota de Cramer-Rao es,

$$\begin{aligned} CRLB &= \left(1 + \frac{\partial Bias(\theta)}{\partial \theta}\right)^2 / E\left[-\frac{\partial^2 ln[L(x;\theta)]}{\partial \theta^2}\right] \\ &= (1-\frac{k}{n})^2/(n/2\sigma^4) = \frac{2(n-k)^2\sigma^4}{n^3} \end{aligned}$$

El estimador de la varianza es sesgado y su MSE no alcanza la cota de Cramer-Rao.

# Máxima Verosimilitud: MCC

- Es el estimador de la varianza de MCC es completamente eficiente?
- $$\frac{(n-k)s^2}{\sigma^2} \sim \chi^2(n-k); \quad s^2 = \frac{\hat{u}'\hat{u}}{(n-k)}$$
- Esperanza,
$$E(\frac{(n-k)s^2}{\sigma^2}) = n-k \Longrightarrow E(s^2) = \sigma^2$$
- Varianza,
$$Var(\frac{(n-k)s^2}{\sigma^2}) = 2(n-k) \Longrightarrow Var(s^2) = \frac{2\sigma^4}{n-k} \neq I_n(\theta)^{-1}_{22} = \frac{2\sigma^4}{n}$$
El estimador es insesgado pero tampoco es completamente eficiente.

- Propiedades Asintóticas:

$$\hat{\beta} \xrightarrow{p} \beta, \quad s^2 \xrightarrow{p} \sigma^2$$

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N[0,\, I(\theta)^{-1}],\ \ \hat{\theta} = (\hat{\beta}',\, s^2)'$$

$$I(\theta) = \lim_{n\to\infty}[\frac{1}{n}I_n(\theta)] = \left[\begin{array}{cc} \frac{1}{\sigma^2}Q & 0 \\ 0 & \frac{1}{2\sigma^4} \end{array}\right], \quad Q = \lim_{n\to\infty}\frac{1}{n}x'x$$

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N[0,\, \sigma^2 Q^{-1}],\ \ \sqrt{n}(s^2 - \sigma^2) \xrightarrow{d} N[0,\, 2\sigma^4]$$

Ambos estimadores son asintóticamente insesgados y eficientes.

# Máxima Verosimilitud: MCC

- MCC con variables explicativas aleatorias.

$$y = x\beta + u, \quad u \sim N(0, \sigma^2 I_n) \quad \theta = (\beta', \sigma^2)'$$

- Supuestos:
  1. $x$ aleatorias con densidad $f(x)$.
  2. $f(x)$ no depende de $\theta$.
  3. $x$ y $u$ se distribuyen en forma independiente.
-

$$L(y, x; \theta) = f(y|x)f(x) = f(u|x)f(x) = f(u)f(x)$$

- El logaritmo de la función de verosimilitud queda,

$$\begin{aligned} Ln[L(y, x; \theta)] &= Ln[f(u)] + Ln[f(x)] \\ &= -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}(y - x\beta)'(y - x\beta) \\ &+ Ln[f(x)] \end{aligned}$$

- Por lo tanto,

$$\hat{\beta} = (x'x)^{-1}x'y, \quad \hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n}, \quad I_n(\theta) = \begin{bmatrix} \frac{1}{\sigma^2}E(x'x) & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

$$I(\theta) = \lim_{n\to\infty}[\frac{1}{n}I_n(\theta)] = \begin{bmatrix} \frac{1}{\sigma^2}Q & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}, \quad Q = \lim_{n\to\infty}\frac{1}{n}E(x'x)$$

- Propiedades asintóticas,

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^2$$

# Máxima Verosimilitud: MCC

- $$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N[0,\, I(\theta)^{-1}], \;\; \hat{\theta} = (\hat{\beta}',\, \hat{\sigma}^2)'$$
  $$I(\theta) = \lim_{n\to\infty}[\frac{1}{n} I_n(\theta)] = \begin{bmatrix} \frac{1}{\sigma^2}Q & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}, \quad Q = \lim_{n\to\infty} \frac{1}{n} E(x'x)$$
  $$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N[0,\, \sigma^2 Q^{-1}], \;\; \sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N[0,\, 2\sigma^4]$$
- Todos estos resultados asumieron normalidad del término de error. Sin este supuesto no hay garantías de que los estimadores de MV sean los de MCC o que el estimador de MCC alcance la cota de Cramer-Rao.
- Si el término de error no es normal, todo lo anterior implica que los estimadores de MV maximizan una función de verosimilitud que está mal especificada.

# Cuasi (Pseudo) Máxima Verosimilitud: MCC

- En este caso, el método se denomina de Cuasi-Máxima Verosimilitud y tiene las siguientes propiedades.
- Todos los resultados de muestra finita que vimos para MCC estimados por MV pueden re-interpretarse como resultados de muestra finita del cuasi-MLE cuando el error es incorrectamente especificado como teniendo distribución normal.

# Agenda

- Sea $L(x, \theta)$ la función de densidad conjunta del vector de variables aleatorias $x = (x_1, \ldots, x_n)'$, caracterizadas por el vector de parámetros $\theta$ de dimensión $(k \times 1)$.
- Queremos contrastar las siguientes hipótesis

$$H_0 : \phi(\theta) = 0 \quad vs. \quad H_1 : \phi(\theta) \neq 0$$

  donde $\phi(\cdot)$ es un vector de dimensión $q \times 1$ de funciones diferenciables con $q < k$.
- Definamos a $\hat{\theta}$ como el estimador MLE sin restricciones, i.e. resuelve $argmax_\theta \ Ln[L(x, \theta)]$.
- Definamos a $\tilde{\theta}$ como el estimador MLE restringido, i.e. resuelve $argmax_\theta \ Ln[L(x, \theta)], \ \ s.t. \, \phi(\theta) = 0$.

# Test de Wald

- Haciendo una expansión de Taylor de primer orden de $\phi(\hat{\theta})$ alrededor del verdadero vector de parámetros $\theta$, tenemos

$$\begin{aligned} \phi(\hat{\theta}) &= \phi(\theta) + F(\theta)'(\hat{\theta} - \theta) + o_p(1) \Longrightarrow \\ \sqrt{n}\left[\phi(\hat{\theta}) - \phi(\theta)\right] &= F(\theta)'\sqrt{n}(\hat{\theta} - \theta) + o_p(1) \end{aligned} \tag{8}$$

donde $F(\theta) = \frac{\partial\phi(\theta)'}{\partial\theta}$ con *rango* $[F(\theta)] = q$.

- De las propiedades asintóticas de los estimadores de máxima verosimilitud sabemos que:

$$\sqrt{n}(\hat{\theta} - \theta) \sim N(0, \, I(\theta)^{-1}) \tag{9}$$

- De (8) y (9) tenemos:

$$\sqrt{n}\left[\phi(\hat{\theta}) - \phi(\theta)\right] \sim N[0, \, F(\theta)'I(\theta)^{-1}F(\theta)]$$

# Test de Wald

- Bajo la hipótesis nula $\phi(\theta) = 0$ por lo tanto,

$$\sqrt{n}\phi(\hat{\theta}) \sim N[0,\ F(\theta)'I(\theta)^{-1}F(\theta)] \tag{10}$$

- Usando la forma cuadrática de variables aleatorias normales tenemos

$$n\phi(\hat{\theta})'[F(\theta)'I(\theta)^{-1}F(\theta)]^{-1}\phi(\hat{\theta}) \sim \chi^2(q) \tag{11}$$

- Podemos aproximar consistentemente los términos del corchete del estadístico anterior evaluando en el estimador de máxima verosimilitud $\hat{\theta}$,

$$W = n\phi(\hat{\theta})'[F(\hat{\theta})'I(\hat{\theta})^{-1}F(\hat{\theta})]^{-1}\phi(\hat{\theta}) \sim \chi^2(q) \tag{12}$$

# Test del Multiplicador de Lagrange (LM)

- Se resuelve el siguiente problema:

$$max_\theta \, Ln[L(x,\, \theta)] \;\; s.t. \;\phi(\theta) = 0$$

$$\mathbb{L}(\theta,\, \lambda) = Ln[L(x,\, \theta)] + \lambda' \phi(\theta)$$

- La solución satisface el siguiente conjunto de ecuaciones:

$$\begin{aligned} \frac{\partial Ln[L(x,\, \theta)]}{\partial \theta}|_{\tilde{\theta}} + F(\tilde{\theta})\tilde{\lambda} &= 0 \\ \phi(\tilde{\theta}) &= 0 \end{aligned} \tag{13}$$

donde $\tilde{\theta}$ es la solución del problema de maximización condicionada.

- El tests LM está basado en la idea de que $\tilde{\lambda}$ apropiadamente ponderado tiene distribución asintótica normal.

# Test del Multiplicador de Lagrange (LM)

- Tomemos una expansión de Taylor de primer orden de $\phi(\hat{\theta})$ y $\phi(\tilde{\theta})$ alrededor del verdadero vector de parámetros $\theta$. Ignorando los términos $o_p(1)$ tenemos,

$$\sqrt{n}\phi(\hat{\theta}) = \sqrt{n}\phi(\theta) + F(\theta)'\sqrt{n}(\hat{\theta} - \theta) \quad (14)$$

$$\sqrt{n}\phi(\tilde{\theta}) = \sqrt{n}\phi(\theta) + F(\theta)'\sqrt{n}(\tilde{\theta} - \theta) \quad (15)$$

- De las condiciones de primer orden sabemos que $\phi(\tilde{\theta}) = 0$ de forma tal que restando (15) de (14) obtenemos

$$\sqrt{n}\phi(\hat{\theta}) = F(\theta)'\sqrt{n}(\hat{\theta} - \tilde{\theta}) \quad (16)$$

- Por otro lado, tomando una expansión de Taylor de primer orden de $\frac{\partial Ln[L(x,\theta)]}{\partial\theta}|_{\theta=\hat{\theta}}$ y $\frac{\partial Ln[L(x,\theta)]}{\partial\theta}|_{\theta=\tilde{\theta}}$ alrededor del verdadero vector de parámetros $\theta$ tenemos (ignorando los términos $o_p(1)$)

$$\frac{\partial Ln[L(x,\theta)]}{\partial \theta}\Big|_{\theta=\hat{\theta}} = \frac{\partial Ln[L(x,\theta)]}{\partial \theta} + \frac{\partial^2 Ln[L(x,\theta)]}{\partial \theta \partial \theta'}(\hat{\theta}-\theta) \tag{17}$$

$$\frac{\partial Ln[L(x,\theta)]}{\partial \theta}\Big|_{\theta=\tilde{\theta}} = \frac{\partial Ln[L(x,\theta)]}{\partial \theta} + \frac{\partial^2 Ln[L(x,\theta)]}{\partial \theta \partial \theta'}(\tilde{\theta}-\theta) \tag{18}$$

- Note que,

$$\begin{aligned}
\frac{\partial Ln[L(x,\theta)]}{\partial \theta}\Big|_{\theta=\hat{\theta}} &= \frac{\partial Ln[L(x,\theta)]}{\partial \theta} + \frac{\partial^2 Ln[L(x,\theta)]}{\partial \theta \partial \theta'}(\hat{\theta}-\theta) \Longrightarrow \\
\frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\theta)]}{\partial \theta}\Big|_{\theta=\hat{\theta}} &= \frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\theta)]}{\partial \theta} + \frac{1}{n}\frac{\partial^2 Ln[L(x,\theta)]}{\partial \theta \partial \theta'}\sqrt{n}(\hat{\theta}-\theta) \Longrightarrow \\
\frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\theta)]}{\partial \theta}\Big|_{\theta=\hat{\theta}} &= \frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\theta)]}{\partial \theta} - I(\theta)\sqrt{n}(\hat{\theta}-\theta)
\end{aligned} \tag{19}$$

- De la misma manera

$$\frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\theta)]}{\partial\theta}|_{\theta=\tilde{\theta}} = \frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\theta)]}{\partial\theta} - I(\theta)\sqrt{n}(\tilde{\theta}-\theta) \tag{20}$$

- De las condiciones de primer orden de la maximización no restringida sabemos que $\frac{\partial Ln[L(x,\theta)]}{\partial\theta}|_{\theta=\hat{\theta}} = 0$. Restando (19) de (20) tenemos,

$$\frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\theta)]}{\partial\theta}|_{\theta=\tilde{\theta}} = -I(\theta)\sqrt{n}(\tilde{\theta}-\hat{\theta}) = I(\theta)\sqrt{n}(\hat{\theta}-\tilde{\theta}) \tag{21}$$

- Por lo tanto

$$\sqrt{n}(\hat{\theta}-\tilde{\theta}) = I(\theta)^{-1}\frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\theta)]}{\partial\theta}|_{\theta=\tilde{\theta}} \tag{22}$$

# Test del Multiplicador de Lagrange (LM)

- De (16) y (22) tenemos

$$\sqrt{n}\phi(\hat{\theta}) = F(\theta)' I(\theta)^{-1} \frac{1}{\sqrt{n}} \frac{\partial Ln[L(x,\theta)]}{\partial\theta}\Big|_{\theta=\tilde{\theta}} \tag{23}$$

- Usando (13) obtenemos

$$\begin{aligned} \sqrt{n}\phi(\hat{\theta}) &= -F(\theta)' I(\theta)^{-1} F(\tilde{\theta}) \frac{\tilde{\lambda}}{\sqrt{n}} \\ &\Longrightarrow -F(\theta)' I(\theta)^{-1} F(\theta) \frac{\tilde{\lambda}}{\sqrt{n}} \end{aligned} \tag{24}$$

- Entonces,

$$\frac{\tilde{\lambda}}{\sqrt{n}} = -\left[F(\theta)' I(\theta)^{-1} F(\theta)\right]^{-1} \sqrt{n}\phi(\hat{\theta}) \tag{25}$$

# Test del Multiplicador de Lagrange (LM)

- De (10), bajo la hipótesis nula $\sqrt{n}\phi(\hat{\theta}) \sim N[0,\ F(\theta)'I(\theta)^{-1}F(\theta)]$, por lo tanto

$$\frac{\tilde{\lambda}}{\sqrt{n}} \sim N\left[0,\ \left(F(\theta)'I(\theta)^{-1}F(\theta)\right)^{-1}\right] \tag{26}$$

- Nuevamente, usando la forma cuadrática de variables normales se obtiene

$$\frac{1}{n}\tilde{\lambda}'[F(\theta)'I(\theta)^{-1}F(\theta)]\tilde{\lambda} \xrightarrow{d}_{H_0} \chi^2(q) \tag{27}$$

- Usando (13) otra forma del estadístico LM es

$$\frac{1}{n}\frac{\partial Ln[L(x,\ \theta)]}{\partial\theta'}\Big|_{\theta=\tilde{\theta}} I(\theta)^{-1}\frac{\partial Ln[L(x,\ \theta)]}{\partial\theta}\Big|_{\theta=\tilde{\theta}} \xrightarrow{d}_{H_0} \chi^2(q) \tag{28}$$

- Para que el test LM sea operativo en la práctica debemos evaluar la matriz de información de Fisher en la estimación restringida consistente, $\tilde{\theta}$.

# Test del Multiplicador de Lagrange (LM)

- Podemos aproximar $I(\theta)$ con:

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 Ln[L(x_i,\,\theta)]}{\partial\theta\partial\theta'}|_{\theta=\tilde{\theta}} \tag{29}$$

- O con

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial Ln[L(x_i,\,\theta)]}{\partial\theta}|_{\theta=\tilde{\theta}}\frac{\partial Ln[L(x_i,\,\theta)]}{\partial\theta'}|_{\theta=\tilde{\theta}} \tag{30}$$

- Si elegimos la segunda aproximación el test LM queda,

$$LM = \frac{1}{n}\frac{\partial Ln[L(x,\,\tilde{\theta})]}{\partial\theta'}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial Ln[L(x_i,\,\tilde{\theta})]}{\partial\theta}\frac{\partial Ln[L(x_i,\,\tilde{\theta})]}{\partial\theta'}\right]^{-1}\frac{\partial Ln[L(x,\,\tilde{\theta})]}{\partial\theta} \tag{31}$$

## Test del Multiplicador de Lagrange (LM)

- Note que,

$$\frac{\partial Ln[L(x,\tilde{\theta})]}{\partial\theta'} = \sum_{i=1}^{n}\frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial\theta'}$$

por lo que el test LM queda,

$$LM =$$

$$\sum_{i=1}^{n}\frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial\theta'}\left[\sum_{i=1}^{n}\frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial\theta}\frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial\theta'}\right]^{-1}\sum_{i=1}^{n}\frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial\theta}$$

- Recuerde la definición del $R^2$ no centrado para el modelo $y = X\beta + u$:

$$R^2 = \frac{\hat{y}'\hat{y}}{y'y} = \frac{(X\hat{\beta})'X\hat{\beta}}{y'y} = \frac{y'X(X'X)^{-1}X'X(X'X)^{-1}X'y}{y'y} = \frac{y'X(X'X)^{-1}X'y}{y'y} \tag{32}$$

# Test del Multiplicador de Lagrange (LM)

- Defina $y = \mathbf{1} = (1\,1\,\ldots\,1)'$ al vector de unos de dimensión $n \times 1$ y $X = \left[\frac{\partial Ln[L(x_1,\tilde{\theta})]}{\partial \theta'}\ \frac{\partial Ln[L(x_2,\tilde{\theta})]}{\partial \theta'}\ \ldots\ \frac{\partial Ln[L(x_n,\tilde{\theta})]}{\partial \theta'}\right]'$ a la matriz $n \times k$ de valores de las variables explicativas.
- En la ecuación del $R^2$ no centrado tenemos:

$$y'X = \mathbf{1}'X = \sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\,\tilde{\theta})]}{\partial \theta'}$$

y

$$X'X = \sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\,\tilde{\theta})]}{\partial \theta} \frac{\partial Ln[L(x_i,\,\tilde{\theta})]}{\partial \theta'}$$

# Test del Multiplicador de Lagrange (LM)

- Entonces,

$$R^2 = \frac{\mathbf{1}'X(X'X)^{-1}X'\mathbf{1}}{\mathbf{1}'\mathbf{1}} =$$

$$\frac{\sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta'} \left[\sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta'}\right]^{-1} \sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta}}{n}$$

$$\Longrightarrow LM = nR^2 \sim \chi^2(q)$$

- Este resultado sugiere que el test LM no es más que el $R^2$ no centrado de una regresión auxiliar de un vector de unos sobre los *scores*, evaluados en la estimación de máxima verosimilitud restringida, multiplicado por el número de observaciones.

- Estadístico del cociente de verosimilitud (LR test)

$$\mu = max_{\phi(\theta)=0} L(x,\,\theta)/max_{\theta} L(x,\,\theta) = \frac{L(x,\,\tilde{\theta})}{L(x,\,\hat{\theta})}$$

- Entonces el estadístico de contraste es,

$$LR = -2Ln(\mu) = 2\{Ln[L(x,\,\hat{\theta})] - Ln[L(x,\,\tilde{\theta})]\} \xrightarrow{d}_{H_0} \chi^2(q)$$

- Note que, haciendo una expansión de Taylor de segundo orden

$$\begin{aligned} Ln[L(x,\,\tilde{\theta})] &\simeq Ln[L(x,\,\hat{\theta})] + (\hat{\theta} - \tilde{\theta})\frac{\partial Ln[L(x,\,\hat{\theta})]}{\partial \hat{\theta}} \\ &+ \frac{1}{2}(\hat{\theta} - \tilde{\theta})'\frac{\partial^2 Ln[L(x,\,\hat{\theta})]}{\partial \hat{\theta} \partial \hat{\theta}'}(\hat{\theta} - \tilde{\theta}) \end{aligned}$$

- La ecuación anterior implica que,

$$\begin{aligned} -2\{Ln[L(x,\hat{\theta})] - Ln[L(x,\tilde{\theta})]\} &= (\hat{\theta}-\tilde{\theta})'\frac{\partial^2 Ln[L(x,\hat{\theta})]}{\partial\hat{\theta}\partial\hat{\theta}'}(\hat{\theta}-\tilde{\theta}) \Longrightarrow \\ 2\{Ln[L(x,\hat{\theta})] - Ln[L(x,\tilde{\theta})]\} &= n(\hat{\theta}-\tilde{\theta})' I(\theta)(\hat{\theta}-\tilde{\theta}) \end{aligned}$$

- Donde hicimos uso de,

$$\frac{\partial Ln[L(x,\hat{\theta})]}{\partial\hat{\theta}} = 0$$

$$I(\theta) = -\frac{1}{n}\frac{\partial^2 Ln[L(x,\hat{\theta})]}{\partial\hat{\theta}\partial\hat{\theta}'}$$

- Por lo tanto, como $\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N[0, I(\theta)^{-1}]$, se desprende que,

$$n(\hat{\theta}-\tilde{\theta})' I(\theta)(\hat{\theta}-\tilde{\theta}) \xrightarrow{d}_{H_0} \chi^2(q)$$

# Agenda

# Sesgo de Selección por Truncamiento Incidental

- En economía un caso emblemático de aplicación de máxima verosimilitud es el modelo de oferta de trabajo de las mujeres (Gronau, 1974; Heckman, 1976). Este modelo consiste de dos ecuaciones, una ecuación de salarios que representa la diferencia entre el salario de mercado de una persona y su salario de reserva en función de características tales como la edad, educación, número de hijos etc.
- La segunda ecuación es una ecuación de horas deseadas de trabajo que depende del salario, de la presencia de hijos pequeños en el hogar, del estado civil, etc.
- El problema del truncamiento es que en la segunda ecuación observamos las horas reales solo si la persona está trabajando. Esto es, solo si el salario de mercado excede al salario de reserva. En este caso se dice que la variable horas en la segunda ecuación está incidentalmente truncada.

# Sesgo de Selección por Truncamiento Incidental

- Definiciones: Suponga que $y$ y $z$ tienen una distribución bivariada con correlación $\rho$. Nosotros estamos interesados en la distribución de $y$ dado que $z$ excede un determinado valor. Esto es, la función de densidad conjunta de $y$ y $z$ es:

$$f(y, z|z > a) = \frac{f(y, z)}{Pr(z > a)}$$

- Teorema 20.4 (Greene, 1997, Cap. 20, página 975): Si $y$ y $z$ tienen una distribución normal bivariada con medias $\mu_y$ y $\mu_z$, desvíos estándar $\sigma_y$ y $\sigma_z$ y correlación $\rho$, entonces:

$$E(y|z > a) = \mu_y + \rho\sigma_y\lambda(\alpha_z)$$

$$Var(y|z > a) = \sigma_y^2[1 - \rho^2\delta(\alpha_z)],$$

donde, $\alpha_z = (a - \mu_z)/\sigma_z$, $\lambda(\alpha_z) = \phi(\alpha_z)/[1 - \Phi(\alpha_z)]$ y $\delta(\alpha_z) = \lambda(\alpha_z)[\lambda(\alpha_z) - \alpha_z]$.

# Sesgo de Selección por Truncamiento Incidental

- Para poner el ejemplo de la oferta de trabajo de las mujeres en un marco general de análisis, digamos que la ecuación que determina la selección muestral es:

$$z_i^* = \gamma' w_i + u_i,$$

donde $z_i^*$ es la diferencia entre el salario de mercado y el salario de reserva de la persona $i$.

- La ecuación de interés es,

$$y_i = \beta' x_i + \epsilon_i,$$

donde $y_i$ es la oferta de trabajo (en horas) de la persona $i$.

- La regla es que $y_i$ es observada solo cuando $z_i^*$ es mayor a cero.

- Asumamos que $u_i$ y $\epsilon_i$ tienen distribución normal bivariada con medias cero y correlación $\rho$. Aplicando el teorema 20.4 tenemos,

$$\begin{aligned}
E(y_i|y_i \text{es observada}) &= E(y_i|z_i^* > 0) \\
&= E(y_i|u_i > -\gamma' w_i) \\
&= \beta' x_i + E(\epsilon_i|u_i > -\gamma' w_i) \\
&= \beta' x_i + \rho\sigma_\epsilon \lambda_i(\alpha_u) \\
&= \beta' x_i + \beta_\lambda \lambda_i(\alpha_u)
\end{aligned}$$

donde $\alpha_u = -\gamma' w_i/\sigma_u$ y $\lambda_i(\alpha_u) = \phi(\gamma' w_i/\sigma_u)/\Phi(\gamma' w_i/\sigma_u)$.

- Entonces, la ecuación de interés puede escribirse como,

$$y_i|z_i^* > 0 = \beta' x_i + \beta_\lambda \lambda_i(\alpha_u) + v_i$$

donde $v_i$ es un término de error con media cero.

# Sesgo de Selección por Truncamiento Incidental

- Como queda claro de este desarrollo, estimar por MCC la ecuación de interés usando solo los datos observados, produce estimadores inconsistentes de $\beta$ por el argumento estándar de variables omitidas (i.e. estamos omitiendo $\lambda_i(\cdot)$).
- Cómo podemos obtener estimaciones consistentes de la ecuación de horas de trabajo utilizando solo los datos observados.
- En principio tenemos un problema similar al de la variable habilidad omitida en la ecuación del salario. En este caso, la variable $\lambda_i(\cdot)$ no es observada.
- Note que aún cuando observáramos $\lambda_i(\cdot)$, MCC no nos daría estimadores eficientes porque los errores de la ecuación de interés, $v_i$, son heterocedásticos (i.e. $Var(v_i) = \sigma_\epsilon^2(1 - \rho^2\delta_i)$ de acuerdo al teorema 20.4 de Greene).

# Sesgo de Selección por Truncamiento Incidental

- Una posible solución es estimar la ecuación de selección para obtener los $\hat{\gamma}$ y construir la variable omitida como $\hat{\lambda}_i = \phi(\hat{\gamma}' w_i)/\Phi(\hat{\gamma}' w_i)$. Luego en un segundo paso estimar la ecuación de interés por MCC en una regresión de $y$ sobre $x$ y $\hat{\lambda}$.
- El único problema de esta solución es que la variable dependiente de la ecuación de selección, $z_i^*$, no es observada. Lo que podemos observar es $z_i = 1$ si $z_i^* > 0$, es decir si la persona está trabajando; o $z_i = 0$ si $z_i^* < 0$ si la persona no está trabajando.
- Es decir que el modelo que podemos estimar es:

$$z_i = \gamma' w_i + u_i, \tag{33}$$

donde $z_i$ es una variable binaria.

# Modelo de Probabilidad Lineal

- Una posibilidad es estimar la ecuación de selección por MCC. Para ver porque esta no es la mejor solución considere la esperanza condicional de la variable dependiente.

$$E(z_i|w_i) = \gamma' w_i$$

  Además desde la definición de esperanza matemática $E(z_i|w_i) = 0 * (1 - P_i) + 1 * P_i = P_i$ donde $P_i$ es la probabilidad de que $z_i = 1$. Entonces,

$$P_i = Pr[z_i = 1|w_i] = E(z_i|w_i) = \gamma' w_i$$

  de aquí el nombre de Modelo de Probabilidad Lineal.

- El problema es que la estimación por MCC de $z$, $\hat{z}_i = \hat{\gamma}' w_i$, es la estimación de la probabilidad de que la persona $i$ se encuentre trabajando. En la práctica, como MCC estima una recta, los valores de $\hat{z}_i$ pueden caer fuera del intervalo (0, 1).

- Como $z_i$ solo adopta dos valores, $z_i = 1$ ó $z_i = 0$, los errores del modelo pueden ser $u_i = 1 - \gamma' w_i$ ó $u_i = -\gamma' w_i$ con probabilidades $P_i$ y $1 - P_i$ respectivamente. Entonces la distribución de los errores es,

| $u_i$ | $f(u_i)$ |
|---|---|
| $1 - \gamma' w_i$ | $\gamma' w_i$ |
| $-\gamma' w_i$ | $1 - \gamma' w_i$ |

- Una consecuencia de la distribución anterior es que

$$\begin{aligned} Var(u_i) &= \gamma' w_i (1 - \gamma' w_i)^2 + (1 - \gamma' w_i)(\gamma' w_i)^2 \\ &= \gamma' w_i (1 - \gamma' w_i) \\ &= E(z_i | w_i)[1 - E(z_i | w_i)]. \end{aligned}$$

y los errores son heterocedásticos.

# Modelo de Probabilidad Lineal

- Golberger (1964) sugirió el siguiente procedimiento de estimación: Primero, estime (33) por MCC. Segundo, calcule $\hat{z}_i(1-\hat{z}_i)$ y use MCGE.
- Los problemas del MPL son los siguientes
  1. En la práctica $\hat{z}_i(1-\hat{z}_i)$ puede ser cero o negativo y el método de MCGE no puede aplicarse.
  2. Como la esperanza condicional de la variable dependiente $E(z_i|w_i)$ se interpreta como una probabilidad, en la práctica su estimación por MCGE puede arrojar valores fuera del intervalo $(0, 1)$.
- Una forma alternativa para formular el modelo que resuelve los problemas anteriores es considerar que $z_i = 1$ cuando $z_i^* > 0$ de forma tal que

$$
\begin{aligned}
P_i = Pr[z_i = 1] &= Pr[z_i^* > 0] = Pr[\gamma' w_i + u_i > 0] \\
&= Pr[u_i > -\gamma' w_i] = Pr[u_i < \gamma' w_i] \\
&= \Phi(\gamma' w_i)
\end{aligned}
$$

# Modelo Probit

- A diferencia de lo que ocurría con el MPL en este caso existe una relación no lineal con los parámetros del modelo y por lo tanto no puede usarse el método de mínimos cuadrados para la estimación.
- Para realizar la estimación del modelo debemos recurrir al método de máxima verosimilitud.
- Como sabemos la función de probabilidad de los errores y tenemos una muestra aleatoria (es decir, compuesta por variables aleatorias independientes) la función de verosimilitud es simplemente la multiplicación de las funciones de probabilidad para todas las observaciones que hay en la muestra.
- $$L(\gamma; w) = \prod_{i=1}^{n} \Phi(\gamma' w_i)^{z_i} [1 - \Phi(\gamma' w_i)]^{(1-z_i)}$$

# Modelo Probit: Estimación

- El logaritmo natural de la función de verosimilitud es,

$$l(\hat{\gamma}; w_i) = \sum_{i=1}^{n} \left[ z_i ln\{\Phi(\hat{\gamma}' w_i)\} + (1 - z_i) ln\{1 - \Phi(\hat{\gamma}' w_i)\} \right]$$

- Las condiciones de primer orden para la maximización de esta función son,

$$S(\hat{\gamma}) = \frac{\partial l(\cdot)}{\partial \hat{\gamma}} \quad = \quad \sum_{i=1}^{n} \left[ \frac{z_i - \Phi(\hat{\gamma}' w_i)}{\Phi(\hat{\gamma}' w_i)[1 - \Phi(\hat{\gamma}' w_i)]} \phi(\hat{\gamma}' w_i) \right] w_i = 0$$

- Como se puede observar en las condiciones de primer orden las incógnitas, $\hat{\gamma}$, entran en forma no lineal y por lo tanto no pueden resolverse usando métodos lineales.

# Sesgo de Selección por Truncamiento Incidental

- Para resolver las condiciones de primer orden hay que recurrir a algoritmos numéricos.
- Amemiya (1985) demostró que la función de verosimilitud del modelo Probit es globalmente cóncava por lo que las condiciones de segundo orden para un máximo se cumplen y no es necesario chequearlas.
- Sin embargo, necesitamos las condiciones de segundo orden para calcular la matriz de información,

$$
\begin{aligned}
I(\hat{\gamma}) &= E\left(-\frac{\partial^2 l(\cdot)}{\partial\hat{\gamma}\partial\hat{\gamma}'}\right) \\
&= \sum_{i=1}^{n} \frac{[\phi(\hat{\gamma}' w_i)]^2}{\Phi(\hat{\gamma}' w_i)[1-\Phi(\hat{\gamma}' w_i)]} w_i w_i'
\end{aligned}
$$

- Entonces la matriz de varianzas y covarianzas asintótica viene dada por $[I(\hat{\gamma})]^{-1}$.

# Sesgo de Selección por Truncamiento Incidental

- Adicionalmente, usando el algoritmo de Newton-Raphson, comenzando con un valor inicial $\hat{\gamma}_0$ el nuevo valor de $\hat{\gamma}_1$ se obtiene,

$$\hat{\gamma}_1 = \hat{\gamma}_0 + [I(\hat{\gamma}_0)]^{-1} S(\hat{\gamma}_0)$$

  este procedimiento iterativo se repite hasta converger.

- Volviendo al ejemplo de la estimación de la oferta de trabajo de las mujeres, Heckman (1979) sugirió utilizar el siguiente procedimiento en dos etapas.
  1. Estimar la ecuación de selección usando un Probit para obtener estimaciones de $\gamma$. Luego para cada observación de la muestra se construye,

$$\hat{\lambda}_i = \frac{\phi(\hat{\gamma}' w_i)}{\Phi(\hat{\gamma}' w_i)}$$

     También vamos a necesitar $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i + \hat{\gamma}' w_i)$.
  2. Estime $\beta$ y $\beta_\lambda$ por MCC en una regresión de $y$ sobre $x$ y $\hat{\lambda}$.

- Para poder hacer inferencia estadística en la regresión del segundo paso hay que tener en cuenta dos problemas: heterocedasticidad en $v_i$ y el hecho de que una de las variables explicativas de la regresión está construída a partir de una estimación anterior.
- Heckman (1979) deriva la verdadera matriz de varianzas y covarianzas de los estimadores del segundo paso.
- Recuerde que $\widehat{Var(v_i)} = \hat{\sigma}_\epsilon^2(1-\hat{\rho}^2\hat{\delta}_i)$ usando el teorema (20.4) de Greene. Donde $\hat{\sigma}_\epsilon^2 = \frac{e'e}{n} + \bar{\delta}\hat{\beta}_\lambda^2$, y $\bar{\delta} = \frac{1}{n}\sum_i \hat{\delta}_i$.
- Entonces la estimación correcta de la matriz de varianzas y covarianzas del segundo paso es,

$$Var[\hat{\beta},\, \hat{\beta}_\lambda] = \hat{\sigma}_\epsilon^2 (\sum_i x_i^{*'} x_i^*)^{-1}\left[\sum_i (1-\hat{\rho}^2\hat{\delta}_i) x_i^* x_i^{*'} + Q\right](\sum_i x_i^{*'} x_i^*)^{-1}$$

- Donde,

$$Q = \hat{\rho}^2 (X^{*'} \Delta W)[Avar(\hat{\gamma})](W' \Delta X^*)$$

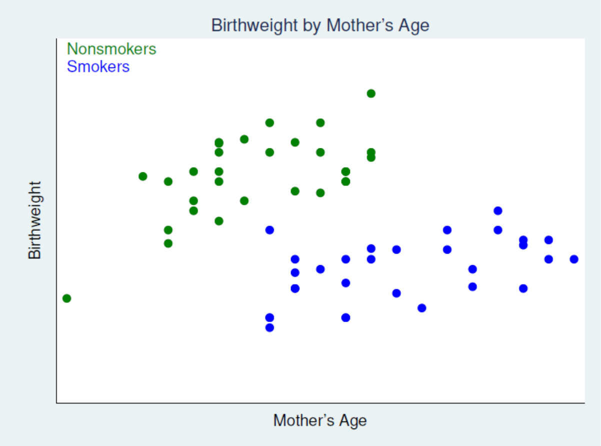
y $\Delta$ es una matriz diagonal con $\hat{\delta}_i$ en la diagonal principal.

# 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.

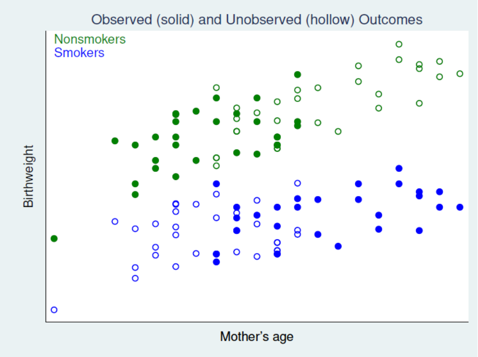
# Costos Económicos de Nacer con Peso Bajo

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



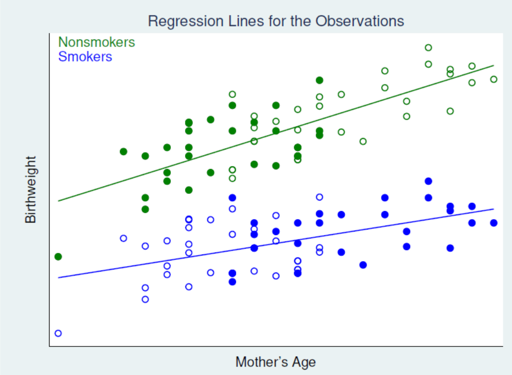
# Costos Económicos de Nacer con Peso Bajo

- 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.



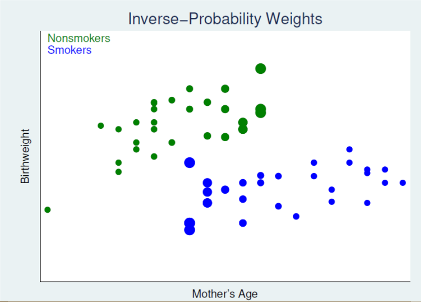
# Estimación

- 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.

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

# Inversa del propensity score como ponderador

- Una alternativa a esto es ponderar las esperanzas matemáticas por la inversa de la probabilidad condicional de recibir el tratamiento (**propensity score**).
- El método del propensity score ve a los resultados potenciales no observados (círculos sin relleno) como observaciones faltantes y usa ponderadores para corregir las estimaciones de las esperanzas para las unidades tratadas y no tratadas.
- El método aplica mayor ponderación a los círculos verdes sólidos correspondientes a madres de mayor edad y menor ponderación a los correspondientes a madres más jóvenes.



# Inversa del propensity score como ponderador

- En términos formales, una forma de estimar el ATE es utilizando la inversa del propensity score ($\pi(x)$) como ponderador.
- Recordando que $sy = sy_1$ tenemos

$$
\begin{aligned}
E\left[\frac{sy}{\pi(x)}\bigg| x\right] &= E\left[\frac{sy_1}{\pi(x)}\bigg| x\right] = E\left\{E\left[\frac{sy_1}{\pi(x)}\bigg| x, s\right]\bigg| x\right\} \\
&= E\left\{\frac{sE(y_1|x,\ s)}{\pi(x)}\bigg| x\right\} = E\left\{\frac{sE(y_1|x)}{\pi(x)}\bigg| x\right\} \\
&= E\left\{\frac{s}{\pi(x)}\bigg| x\right\} E(y_1|x) = E(y_1|x)
\end{aligned}
$$

- Haciendo uso de $E(s|x) = \pi(x)$.
- Usando un argumento similar se puede mostrar que:

$$
E\left[\frac{(1-s)y}{1-\pi(x)}\bigg| x\right] = E(y_0|x)
$$

- Usando álgebra se puede mostrar que,

$$E(y_1 - y_0|x) = E\left[\left.\frac{[s - \pi(x)]y}{\pi(x)[1 - \pi(x)]}\right| x\right]$$

- Y usando expectativas iteradas,

$$\tau_{ATE} = E(y_1 - y_0) = E\left[\frac{[s - \pi(x)]y}{\pi(x)[1 - \pi(x)]}\right] \tag{34}$$

# Inversa del propensity score como ponderador

- Note que $\pi(x) = Pr[s = 1|x]$ es la probabilidad de recibir el tratamiento y ex-ante no es observable.
- Lo único que observamos en la práctica es si la persona recibe el tratamiento ($s = 1$) o no lo recibe ($s = 0$).
- Supongamos que una persona se autoselecciona en el tratamiento dependiendo de la utilidad que le brinda.
- Denotemos por

$$U_i^T = x_i \gamma_T + u_i^T \tag{35}$$

a la utilidad que le brinda al individuo $i$ autoseleccionarse en el tratamiento y de la misma manera

$$U_i^C = x_i \gamma_C + u_i^C \tag{36}$$

a la utilidad de no recibir el tratamiento.

# Inversa del propensity score como ponderador

- Entonces

$$\begin{aligned}\pi(x) &= Pr[s=1|x] = Pr[U_i^T > U_i^C] = Pr[u_i^C - u_i^T \leq x_i(\gamma_T - \gamma_C)] \\ &= Pr[u_i \leq x_i\gamma] = \Phi(x_i\gamma)\end{aligned}$$

- Donde asumimos que el error $u_i$ tiene distribución normal estándar (modelo Probit).
- En términos de nuestro ejemplo,

$$\Phi(x_i\gamma) = \int_{-\infty}^{x_i\gamma} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

existe una relación nolineal entre la probabilidad de recibir el tratamiento y los parámetros del modelo.

# Modelo Probit: estimación

- Para realizar la estimación del modelo debemos recurrir al método de máxima verosimilitud.
- La función de verosimilitud es la multiplicación de la probabilidad de recibir el tratamiento para todas las observaciones que hay en la muestra.
- $$L(\gamma; x) = \prod_{i=1}^{n} \Phi(x_i\gamma)^{s_i}[1 - \Phi(x_i\gamma)]^{(1-s_i)}$$

# Modelo Probit: Estimación

- El logaritmo natural de la función de verosimilitud es,

$$l(\hat{\gamma}; x_i) = \sum_{i=1}^{n} [s_i ln\{\Phi(x_i\hat{\gamma})\} + (1 - s_i) ln\{1 - \Phi(x_i\hat{\gamma})\}]$$

- Las condiciones de primer orden para la maximización de esta función son,

$$S(\hat{\gamma}) = \frac{\partial l(\cdot)}{\partial \hat{\gamma}} \quad = \quad \sum_{i=1}^{n} \left[ \frac{s_i - \Phi(x_i\hat{\gamma})}{\Phi(x_i\hat{\gamma})[1 - \Phi(x_i\hat{\gamma})]} \phi(x_i\hat{\gamma}) \right] x_i = 0$$

- Como se puede observar en las condiciones de primer orden las incógnitas, $\hat{\gamma}$, entran en forma no lineal y por lo tanto no pueden resolverse usando métodos lineales.

# Modelo Probit: Estimación

- Para resolver las condiciones de primer orden hay que recurrir a algoritmos numéricos.
- Amemiya (1985) demostró que la función de verosimilitud del modelo Probit es globalmente cóncava por lo que las condiciones de segundo orden para un máximo se cumplen y no es necesario chequearlas.
- Sin embargo, necesitamos las condiciones de segundo orden para calcular la matriz de información,

$$
\begin{aligned}
I(\hat{\gamma}) &= E\left(-\frac{\partial^2 l(\cdot)}{\partial\hat{\gamma}\partial\hat{\gamma}'}\right) \\
&= \sum_{i=1}^{n} \frac{[\phi(x_i\hat{\gamma})]^2}{\Phi(x_i\hat{\gamma})[1-\Phi(x_i\hat{\gamma})]} x_i x_i'
\end{aligned}
$$

- Entonces la matriz de varianzas y covarianzas asintótica viene dada por $[I(\hat{\gamma})]^{-1}$.

# Sesgo de Selección por Truncamiento Incidental

- Adicionalmente, usando el algoritmo de Newton-Raphson, comenzando con un valor inicial $\hat{\gamma}_0$ el nuevo valor de $\hat{\gamma}_1$ se obtiene,

$$\hat{\gamma}_1 = \hat{\gamma}_0 + [I(\hat{\gamma}_0)]^{-1} S(\hat{\gamma}_0)$$

  este procedimiento iterativo se repite hasta converger.

- Una vez estimado el propensity score usando un Probit estimamos el ATE,

$$\hat{\tau}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} \left[ \frac{[s - \hat{\pi}(x)]y}{\hat{\pi}(x)[1 - \hat{\pi}(x)]} \right] \tag{37}$$

# Variable Dependiente Binaria y Efectos Marginales

- En los modelos lineales los coeficientes que acompañan a las variables explicativas tienen la interpretación de ser directamente el efecto marginal de un cambio en la variable explicativa sobre la variable dependiente.
- En el caso de modelos no lineales como el Probit los coeficientes que acompañan a las variables explicativas no tienen esta interpretación.
- Para ver esto, considere el siguiente modelo:

$$\Phi(\alpha + \beta x_i + z\gamma) = \int_{-\infty}^{\alpha+\beta x_i+z\gamma} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

  donde $z$ es un vector de variables explicativas adicionales a $x$ y $\gamma$ un vector de parámetros poblacionales.
- En este caso el efecto marginal viene dado por la derivada parcial de la probabilidad ex-ante con respecto a alguna variable explicativa.

$$\frac{\partial Pr[y_i = 1|\cdot]}{\partial x_i} = \phi(\alpha + \beta x_i + z\gamma) \times \beta$$

# Variable Dependiente Binaria y Efectos Marginales

- La expresión anterior corresponde a una variable explicativa $x_i$ continua.
- Si $x_i$ es una variable discreta (binaria) el efecto marginal viene dado por:

$$Pr[y_i = 1|x_i = 1] - Pr[y_i = 1|x_i = 0] = \{\Phi(\alpha + \beta + z\gamma) - \Phi(\alpha + z\gamma)\} \tag{38}$$

- Note que a diferencia de los modelos lineales, el efecto marginal es una función y no una constante.
- Para obtener un resultado numérico puntual, es estándar en la práctica calcular el denominado efecto marginal promedio.
- En términos matemáticos,

$$E\left\{Pr[y_i = 1|x_i = 1] - Pr[y_i = 1|x_i = 0]\right\} = E\left\{\Phi(\alpha + \beta + z\gamma) - \Phi(\alpha + z\gamma)\right\} \tag{39}$$

- Supongamos ahora que queremos analizar la eficacia de una vacuna contra el COVID-19 en un experimento aleatorizado.
- La "política" es la vacunación.
- La variable de resultado es si el individuo se contagia o no de COVID.
- Aquí el resultado potencial es una variable binaria y en un modelo lineal el ATE viene dado por:

$$\begin{aligned} ATE &= E(Y_s(u)|s_T = 1) - E(Y_s(u)|s_T = 0) = E(Y_T(u)) - E(Y_C(u)) \\ &= Pr(Y_s = 1|s_T = 1) - Pr(Y_s = 1|s_T = 0) \qquad (40) \end{aligned}$$

  donde $u$ denota al individuo, $Y_s$ es una variable binaria que adopta el valor unitario si el individuo se contagia de COVID-19 y vale cero si no se contagia y $s_T = 1$ si el individuo está vacunado.
- La segunda igualdad viene dada por el hecho de que la esperanza matemática condicional de un modelo de variable dependiente binaria es igual a la probabilidad de ocurrencia del evento analizado.

# Variable Dependiente Binaria y Efectos Marginales

- Supongamos que estimamos las probabilidades utilizando un modelo Probit

$$E\left[\mathrm{Y}_i \mid \mathrm{X}_i, \mathrm{S}_i\right] = \Phi\left[\mathrm{X}_i'\beta_0 + \beta_1 \mathrm{S}_i\right] \tag{41}$$

- Ahora la "switching equation" es

$$E\left[\mathrm{Y}_i \mid \mathrm{X}_i, \mathrm{S}_i\right] = \Phi\left[\mathrm{X}_i'\beta_0\right] + \left\{\Phi\left[\mathrm{X}_i'\beta_0 + \beta_1\right] - \Phi\left[\mathrm{X}_i'\beta_0\right]\right\} \times \mathrm{S}_i \tag{42}$$

- y el ATE es

$$\begin{aligned} ATE &= E\left\{Pr(Y_i = 1|\mathrm{X}_i, \mathrm{S}_i = 1) - Pr(Y_i = 1|\mathrm{X}_i, \mathrm{S}_i = 0)\right\} \\ &= E\left\{\Phi\left[\mathrm{X}_i'\beta_0 + \beta_1\right] - \Phi\left[\mathrm{X}_i'\beta_0\right]\right\} \end{aligned} \tag{43}$$

# Variable Dependiente Binaria y Efectos Marginales

- El principal resultado de esta discusión es que estimar el efecto tratamiento promedio usando un modelo Probit involucra calcular el efecto marginal promedio.
- En la práctica no hay grandes diferencias entre el efecto marginal promedio y el coeficiente sobre la variable de asignación del tratamiento en un modelo de regresión lineal.
- Esta regularidad empírica más la complejidad adicional para hacer inferencia estadística (i.e. se necesitan errores estándar para los efectos marginales promedio) hace que muchos investigadores cuando trabajan con evaluaciones de impacto prefieran aproximar la estimación de las esperanzas matemáticas condicionales usando un modelo de probabilidad lineal.

# Econometría

## Maestría en Economía - Maestría en Econometría

Lecture 4

# Agenda

1. Ecuaciones Aparentemente no Relacionadas (SURE)
   - Motivación
   - Casos Especiales
   - Estimación por FGLS
2. Ecuaciones Simultáneas
   - Motivación
   - Estimación
   - Propiedades
   - Inferencia
   - Identificación

# Agenda

1. Ecuaciones Aparentemente no Relacionadas (SURE)
   - Motivación
   - Casos Especiales
   - Estimación por FGLS
2. Ecuaciones Simultáneas
   - Motivación
   - Estimación
   - Propiedades
   - Inferencia
   - Identificación

# Testing the CAPM

- El Capital Asset Pricing Model (CAPM) o modelo de determinación del precio de los activos desarrollado por Sharpe (1964) y Litner (1965) es ampliamente utilizado en finanzas para estimar el costo de capital para las firmas y evaluar el desempeño de portafolios de acciones en el mercado.
- El CAPM se construye sobre el modelo de elección de portafolios desarrollado por Harry Markowitz (1959). En el modelo de Markowitz, un inversor selecciona un portafolio en el período $t-1$ que produce un retorno estocástico en el período $t$.
- El modelo asume que los inversores son adversos al riesgo y cuando eligen su portafolio están interesados solamente en la media y la varianza del rendimiento de su inversión de un período.
- Como resultado, los inversores eligen portafolios eficientes en media y varianza en el sentido que (i) minimizan la varianza del rendimiento del portafolio, dado el retorno esperado y (ii) maximizan el rendimiento esperado, dada la varianza.

# Testing the CAPM

- El modelo CAPM establece que el retorno esperado de cualquier activo es lineal en su covarianza con el retorno esperado del portafolio que representa el mercado. Esto es,

$$E[R_i] = r_f + \beta_i(E[R_M] - r_f), \quad i = 1, \ldots, n$$

donde $E[R_i]$ y $E[R_M]$ son los retornos esperados del activo $i$ y del portafolio que representa al mercado, respectivamente. $r_f$ es la tasa libre de riesgo y $\beta_i$ es la covarianza entre el retorno del activo $i$ y el retorno del mercado, dividido por la varianza del retorno del mercado. En finanzas este último término se denomina el beta del activo.

# Testing the CAPM

- El CAPM tiene tres implicancias directas:
  - Los retornos esperados de todos los activos están linealmente relacionados con sus betas y no hay ninguna otra variable que tenga poder explicativo adicional.
  - La prima de riesgo de mercado es positiva.
  - En términos de un modelo de regresión de la prima de riesgo de cada activo sobre la prima de riesgo del mercado, la constante del modelo debe ser cero.
- Los tests de "primera generación" consistían en regresiones de corte transversal de los retornos promedio sobre los beta de cada activo (que eran estimados previamente con regresiones de series temporales) y otras variables de control para chequear la primera implicancia mencionada arriba.
- Estos tests tenian un problema principal, los errores en las variables estimadas en el primer paso (los betas).

# Testing the CAPM

- Los tests de "segunda generación" consisten en regresiones multivariantes. Por ejemplo, para $m$ activos y $n$ períodos temporales tenemos,

$$r_{i,t} = \alpha_i + \beta_i r_{M,t} + u_{i,t}, \quad i = 1, \ldots, m, \quad t = 1, \ldots, n$$

  donde $r_{i,t}$ es la prima de riesgo real del activo $i$ en el período $t$ y $r_{M,t}$ es la prima de riesgo real del mercado en el período $t$.
- En términos vectoriales,

$$r_i = \alpha_i + \beta_i r_M + u_i, \quad i = 1, \ldots, m,$$

  Estas $m$ ecuaciones aparentemente no están relacionadas. En realidad la relación entre ellas viene dada por la relación entre los errores de las mismas.

# Testing the CAPM

- Si el CAPM se cumple entonces todos los $\alpha_i = 0$ esto significa que no existen otros factores, aparte de la prima de riesgo del mercado, que expliquen las primas de riesgo de los activos.
- El estadístico de contraste para esta hipótesis es un estadístico de Wald estándar,

$$W = \hat{\alpha}'[\widehat{Var(\hat{\alpha})}]^{-1}\hat{\alpha} \sim \chi^2(m)$$

Si este estadístico se desvía significativamente de cero, entonces concluímos que el CAPM no explica completamente las variaciones en los retornos de los activos.

# Ecuaciones aparentemente no relacionadas

- El modelo de ecuaciones aparentemente no relacionadas (SURE) o modelo de Zellner (1962) general, consiste de $m$ ecuaciones de regresión que individualmente satisfacen todos los supuestos del modelo de regresión lineal múltiple.
- $$\underset{(n\times 1)}{y_i} = \underset{(n\times k_i)}{X_i}\underset{(k_i\times 1)}{\beta_i} + \underset{(n\times 1)}{u_i}, \quad i=1,2,\ldots,m \tag{1}$$
- Si $u_{it}$ es el t-ésimo elemento de $u_i$, asumimos que $(u_{1t}, u_{2t}, \ldots, u_{mt})$ es $i.i.d$ con
  - $E(u_{it}) = 0$
  - $E(u_{it}u_{js}) = \begin{cases} \sigma_{ij} \text{ if } t=s \\ 0 \text{ if } t\neq s \end{cases}$

# Ecuaciones aparentemente no relacionadas

- *Stacking* las $m$ ecuaciones tenemos:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}_{(nm \times 1)} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_m \end{bmatrix}_{(nm \times \sum_{i=1}^{m} k_i)} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}_{(\sum_{i=1}^{m} k_i \times 1)}$$

$$+ \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}_{(nm \times 1)}$$

- Esto es,

# Ecuaciones aparentemente no relacionadas

- $$y = X\beta + u \tag{2}$$
  - $E(u) = 0$
  - $E(uu') = V = \Sigma \otimes I_n = \begin{bmatrix} \sigma_{11} I_n & \sigma_{12} I_n & \cdots & \sigma_{1m} I_n \\ \sigma_{21} I_n & \sigma_{22} I_n & \cdots & \sigma_{2m} I_n \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{m1} I_n & \sigma_{m2} I_n & \cdots & \sigma_{mm} I_n \end{bmatrix}$
- El estimador de MCG (GLS) de $\beta$ es,

$$\begin{aligned} \hat{\beta}_{GLS} &= (X'V^{-1}X)^{-1}X'V^{-1}y \\ &= [X'(\Sigma^{-1} \otimes I_n)X]^{-1}X'(\Sigma^{-1} \otimes I_n)y \end{aligned}$$

- Y su matriz de varianzas y covarianzas

$$Var(\hat{\beta}_{GLS}) = (X'V^{-1}X)^{-1} = [X'(\Sigma^{-1} \otimes I_n)X]^{-1} \tag{3}$$

# Agenda

1. Ecuaciones Aparentemente no Relacionadas (SURE)
   - Motivación
   - Casos Especiales
   - Estimación por FGLS
2. Ecuaciones Simultáneas
   - Motivación
   - Estimación
   - Propiedades
   - Inferencia
   - Identificación

# GLS $\equiv$ MCC

- Si los errores de las ecuaciones no están correlacionados ($\sigma_{ij} = 0$ para $i \neq j$)

$$\hat{\beta}_{GLS} = [X'(\Sigma^{-1} \otimes I_n)X]^{-1} X'(\Sigma^{-1} \otimes I_n)y =$$

$$= \begin{bmatrix} \sigma_{11}X_1'X_1 & 0 & \cdots & 0 \\ 0 & \sigma_{22}X_2'X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{mm}X_m'X_m \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{11}X_1'y_1 \\ \sigma_{22}X_2'y_2 \\ \vdots \\ \sigma_{mm}X_m'y_m \end{bmatrix} =$$

$$\begin{bmatrix} \sigma_{11}^{-1}(X_1'X_1)^{-1} & 0 & \cdots & 0 \\ 0 & \sigma_{22}^{-1}(X_2'X_2)^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{mm}^{-1}(X_m'X_m)^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{11}X_1'y_1 \\ \sigma_{22}X_2'y_2 \\ \vdots \\ \sigma_{mm}X_m'y_m \end{bmatrix}$$

$$\hat{\beta}_{GLS} = \begin{bmatrix} (X_1'X_1)^{-1}X_1'y_1 \\ (X_2'X_2)^{-1}X_2'y_2 \\ \vdots \\ (X_m'X_m)^{-1}X_m'y_m \end{bmatrix}$$

# GLS $\equiv$ MCC

- Si $X_1 = X_2 = \cdots = X_m$.
- Definamos $X_1 = X_2 = \cdots = X_m = X_0$.
- $$\begin{aligned}
\hat{\beta}_{GLS} &= [X'(\Sigma^{-1} \otimes I_n)X]^{-1} X'(\Sigma^{-1} \otimes I_n)y \\
&= [(I_n \otimes X_0)'(\Sigma^{-1} \otimes I_n)(I_n \otimes X_0)]^{-1}(I_n \otimes X_0)'(\Sigma^{-1} \otimes I_n)y \\
&\qquad \text{por } X = X_0 \\
&= [(\Sigma^{-1} \otimes X_0')(I_n \otimes X_0)]^{-1}(\Sigma^{-1} \otimes X_0')y \\
&\qquad \text{por } (A \otimes B)(C \otimes D) = (AC \otimes BD) \\
&= (\Sigma^{-1} \otimes X_0'X_0)^{-1}(\Sigma^{-1} \otimes X_0')y \\
&\qquad \text{por } (A \otimes B)(C \otimes D) = (AC \otimes BD) \\
&= [\Sigma \otimes (X_0'X_0)^{-1}](\Sigma^{-1} \otimes X_0')y \\
&\qquad \text{por } (A \otimes B)^{-1} = (A^{-1} \otimes B^{-1}) \\
&= [I_n \otimes (X_0'X_0)^{-1}X_0']y = (X'X)^{-1}X'y = \hat{\beta}
\end{aligned}$$

# Agenda

1. Ecuaciones Aparentemente no Relacionadas (SURE)
   - Motivación
   - Casos Especiales
   - Estimación por FGLS
2. Ecuaciones Simultáneas
   - Motivación
   - Estimación
   - Propiedades
   - Inferencia
   - Identificación

# Ecuaciones aparentemente no relacionadas

- MCGE (FGLS) de dos Etapas
- Primer paso: construya los residuos de MCC (OLS) de la ecuación $i$,

$$\hat{u}_i = y_i - X_i\hat{\beta}$$

- Segundo Paso: estime consistentemente los elementos de $\Sigma$ con,

$$\hat{\sigma}_{ij} = \frac{1}{n}\hat{u}_i'\hat{u}_j$$

- Tercer Paso: el estimador de MCGE es,

$$\hat{\beta}_{FGLS} = [X'(\hat{\Sigma}^{-1} \otimes I_n)X]^{-1}X'(\hat{\Sigma}^{-1} \otimes I_n)y$$

# Ecuaciones aparentemente no relacionadas

- Como hemos visto bajo condiciones generales MCG y MCGE tienen la misma distribución asintótica.

$$\sqrt{n}(\hat{\beta}_{FGLS} - \beta) \xrightarrow{d} N[0, \lim_{n\to\infty}(\frac{1}{n}X'V^{-1}X)^{-1}], \text{ cuando } n \to \infty$$

- La inferencia estadística se hace como siempre. $H_0 : R\beta = r$,

$$W = (R\hat{\beta}_{FGLS} - r)'[R\widehat{Var(\hat{\beta}_{FGLS})}R']^{-1}(R\hat{\beta}_{FGLS} - r) \xrightarrow{d} \chi^2(q)$$

# Agenda

1. Ecuaciones Aparentemente no Relacionadas (SURE)
   - Motivación
   - Casos Especiales
   - Estimación por FGLS
2. Ecuaciones Simultáneas
   - Motivación
   - Estimación
   - Propiedades
   - Inferencia
   - Identificación

# Modelo de Oferta de Trabajo

- Considere el siguiente modelo de oferta de trabajo:

$$h = \gamma_1 w + z_1\delta_1 + u_1 \tag{4}$$

$$w = \gamma_2 h + z_2\delta_2 + u_2 \tag{5}$$

donde $(h, w)$ denotan los valores de equilibrio de horas trabajadas y salario ofrecido en el mercado.

- Es poco probable que podamos asumir que a un individuo le hacen una oferta salarial exógena y luego él decide cuanto trabajar usando (4).
- En general tanto las horas trabajadas como el salario se determinan conjuntamente.
- Esto es, en la mayoría de los casos $u_1$ estará correlacionado con $w$ y $u_2$ con $h$. En otras palabras, $w$ probablemente sea una variable endógena en (4) y $h$ sea endógena en (5).

# Modelo de Oferta de Trabajo

- Note que este ejemplo puede escribirse de la siguiente manera:

$$y_1 = X_1\beta_1 + u_1 \quad (6)$$
$$y_2 = X_2\beta_2 + u_2 \quad (7)$$

  que luce como un sistema SURE de dos ecuaciones.
- La única diferencia es que tanto $X_1$ como $X_2$ contienen variables endógenas y exógenas.
- Como no se cumple el supuesto de exogeneidad (estricta o contemporanea) la estimación de estas ecuaciones por MCC o MCGE producirá estimadores inconsistentes.

# Agenda

- Pasemos a discutir el modelo general

$$y_i = X_i\beta + u_i$$

donde $y_i$ es un vector de dimensión $G \times 1$, $X_i$ es una matriz $G \times K$ y $u_i$ es un vector $G \times 1$ de errores.

- Las siguientes condiciones de ortogonalidad, más el supuesto de rango estándar, son la base para estimar consistentemente los parámetros del modelo.
- Supuesto SIV.1: $E(Z_i'u_i) = 0$, donde $Z_i$ es una matriz $G \times L$ de variables instrumentales conocidas.
Supuesto SIV.2: rango$E(Z_i'X_i) = K$.
Para el supuesto de rango es necesaria la condición de orden $L \geq K$.

# Ecuaciones Simultáneas: Estimación

- Reescribamos el modelo en forma matricial,

$$y_i \equiv \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iG} \end{bmatrix} X_i \equiv \begin{bmatrix} x_{i1} & 0 & \cdots & 0 \\ 0 & x_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_{iG} \end{bmatrix} u_i \equiv \begin{bmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iG} \end{bmatrix} \beta \equiv \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_G \end{bmatrix}$$

- Asumamos que para cada ecuación tenemos un conjunto de variables instrumentales $z_j$ $(1 \times L_j)$ que son exógenas en el sentido que,

$$E(z_j' u_j) = 0, \;\; j = 1, \ldots, G$$

- Entonces la matriz de instrumentos $(Z_i)$ $(G \times L)$ tendrá la forma,

$$Z_i = \begin{bmatrix} z_{i1} & 0 & \cdots & 0 \\ 0 & z_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_{iG} \end{bmatrix}$$

- Para cada $i$, $Z_i' u_i = (z_{i1} u_{i1}, z_{i2} u_{i2}, \ldots, z_{iG} u_{iG})$ y por lo tanto $E(Z_i' u_i) = 0$. También tenemos,

$$E(Z_i' X_i) = \begin{bmatrix} E(z_{i1}' x_{i1}) & 0 & \cdots & 0 \\ 0 & E(z_{i2}' x_{i2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & E(z_{iG}' x_{iG}) \end{bmatrix}$$

  donde $E(z_{ij}' x_{ij})$ es de dimensión $L_j \times K_j$
- El supuesto SIV.2 requiere que esta matriz tenga rango completo. Esto es, el supuesto se satisface si *Rango* $E(z_{ij}' x_{ij}) = K_j$, $j = 1, 2, \ldots, G$.
- Las condiciones de ortogonalidad dadas por SIV.1 sugieren que una estrategia de estimación puede ser resolver el siguiente conjunto de condiciones poblacionales,

$$E[Z_i'(y_i - X_i \beta)] = 0$$

# Ecuaciones Simultáneas: Estimación

- Que $\beta$ es una solución para la ecuación anterior surge de SIV.1 y que es la única solución sigue de SIV.2.
- Por lo tanto aplicando el principio de analogía el estimador de $\beta$ es el que soluciona,

$$\frac{1}{N}\sum_{i=1}^{N}[Z_i'(y_i - X_i\hat{\beta})] = 0 \tag{8}$$

- La expresión (8) es un conjunto de $L$ ecuaciones lineales en las $K$ incógnitas $\hat{\beta}$.
- Caso 1: $L = K$ (tenemos exactamente el mismo número de instrumentos que de variables explicativas).
- Entonces si $\sum_{i=1}^{N} Z_i' X_i$ no es singular, El estimador de $\beta$ es,

$$\hat{\beta} = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i' X_i\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} Z_i' y_i = (Z'X)^{-1} Z' Y$$

# Ecuaciones Simultáneas: Estimación

- En la ecuación anterior $Z$ es la matriz $NG \times L$ obtenida "stacking" $Z_i$ $(i = 1, 2, \ldots, N)$, $X$ es la matriz $NG \times K$ obtenida "stacking" $X_i$ $(i = 1, 2, \ldots, N)$, e $Y$ es el vector $NG \times 1$ obtenido "stacking" $y_i$ $(i = 1, 2, \ldots, N)$.
- Este estimador se conoce como estimador de variables instrumentales en sistemas (SIV).
- Aplicando la WLLN se puede demostrar que si se cumplen los supuestos SIV.1 y SIV.2 el estimador es consistente.
- Caso 2: L > K. Esto significa que hay más instrumentos que variables explicativas y solo en casos especiales la ecuación (SIV) puede resolverse. En su lugar lo que se hace en la práctica es elegir de forma de hacer el vector en (SIV) lo "más chico" posible.

# Ecuaciones Simultáneas: Estimación

- Una primera idea es minimizar el largo Euclídeo cuadrado eligiendo $\hat{\beta}$ de forma de minimizar,

$$\left[\sum_{i=1}^{N} Z_i'(y_i - X_i\hat{\beta})\right]' \left[\sum_{i=1}^{N} Z_i'(y_i - X_i\hat{\beta})\right]$$

- Este método produce estimadores consistentes bajo los supuestos SIV.1 y SIV.2 pero no los más eficientes.
- Una clase más general de estimadores se obtiene usando una matriz ponderadora. Definamos la matriz $\hat{W}$ simétrica y de dimensión $L \times L$.
- Un estimador del método generalizado de momentos (GMM) de $\beta$ es un vector que resuelve,

$$\min_{\hat{\beta}} \left[\sum_{i=1}^{N} Z_i'(y_i - X_i\hat{\beta})\right]' \hat{W} \left[\sum_{i=1}^{N} Z_i'(y_i - X_i\hat{\beta})\right] \tag{9}$$

# Ecuaciones Simultáneas: Estimación

- Como la expresión (9) es una función cuadrática de $\hat{\beta}$ tiene una solución cerrada. Por lo tanto, el estimador de GMM es,

$$\hat{\beta} = (X'Z\hat{W}Z'X)^{-1}X'Z\hat{W}Z'Y \tag{10}$$

- Asumiendo que $X'Z\hat{W}Z'X$ no es singular para mostrar la consistencia de este estimador necesitamos el siguiente supuesto,
  Supuesto SIV.3: $\hat{W} \xrightarrow{p} W$, donde $W$ es una matriz de dimensión $L \times L$, no aleatoria, simétrica y positiva definida.
- En las aplicaciones este último supuesto se va a seguir de la aplicación de la WLLN.
- Bajo los supuestos SIV.1-SIV.3 el estimador de GMM es consistente.

# Agenda

# Ecuaciones Simultáneas: Propiedades

- Escribamos (10) como,

$$\hat{\beta} = \left[ (\frac{1}{N}\sum_{i=1}^{N} X_i' Z_i)\hat{W}(\frac{1}{N}\sum_{i=1}^{N} Z_i' X_i) \right]^{-1} (\frac{1}{N}\sum_{i=1}^{N} X_i' Z_i)\hat{W}(\frac{1}{N}\sum_{i=1}^{N} Z_i' y_i)$$

- y reemplazando $y_i$ por el modelo se obtiene,

$$\hat{\beta} - \beta = \left[ (\frac{1}{N}\sum_{i=1}^{N} X_i' Z_i)\hat{W}(\frac{1}{N}\sum_{i=1}^{N} Z_i' X_i) \right]^{-1} (\frac{1}{N}\sum_{i=1}^{N} X_i' Z_i)\hat{W}(\frac{1}{N}\sum_{i=1}^{N} Z_i' u_i)$$

- Usando SIV.2 $C \equiv E(Z_i' X_i)$ tiene rango $K$ y por el supuesto SIV.3 $C'WC$ tiene rango $K$ y por lo tanto no es singular.
- Aplicando la WLLN,

$$\begin{aligned} plim\hat{\beta} &= \beta + (C'WC)^{-1} C'W\, plim(\frac{1}{N}\sum_{i=1}^{N} Z_i' u_i) \\ &= \beta + (C'WC)^{-1} C'W0 = \beta \end{aligned}$$

# Ecuaciones Simultáneas: Propiedades

- Bajo estos mismos supuestos, el estimador de GMM es asintóticamente normal.
- De (10) tenemos,

$$\sqrt{N}(\hat{\beta}-\beta)=\left[(\frac{1}{N}\sum_{i=1}^{N}X_i'Z_i)\hat{W}(\frac{1}{N}\sum_{i=1}^{N}Z_i'X_i)\right]^{-1}(\frac{1}{N}\sum_{i=1}^{N}X_i'Z_i)\hat{W}(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}Z_i'u_i)$$

- Usando $N^{-1/2}\sum_{i=1}^{N}Z_i'u_i \xrightarrow{d} N(0,\Lambda)$, la distribución asintótica del estimador GMM es normal con matriz de varianzas y covarianzas asintótica,

$$AVar[\sqrt{N}(\hat{\beta}-\beta)]=(C'WC)^{-1}C'W\Lambda WC(C'WC)^{-1} \tag{11}$$

donde $\Lambda \equiv E(Z_i'u_iu_i'Z_i)=Var(Z_i'u_i)$.

# Ecuaciones Simultáneas: Propiedades

- Si $\hat{\Lambda}$ es un estimador consistente de $\Lambda$ entonces la varianza asintótica de $\sqrt{N}(\hat{\beta} - \beta)$ se puede estimar consistentemente con,

$$\left(\frac{X'Z}{N}\hat{W}\frac{Z'X}{N}\right)^{-1}\frac{X'Z}{N}\hat{W}\hat{\Lambda}\hat{W}\frac{Z'X}{N}\left(\frac{X'Z}{N}\hat{W}\frac{Z'X}{N}\right)^{-1}$$

- La fórmula anterior puede simplificarse eligiendo $\hat{W}$ apropiadamente.
- Una de las alternativas es elegir $\hat{W}$ como,

$$\hat{W} = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i' Z_i\right)^{-1} = \left(\frac{Z'Z}{N}\right)^{-1}$$

que es un estimador consistente de $[E(Z_i'Z_i)]^{-1}$.

- El supuesto SIV.3 solo requiere que $E(Z_i'Z_i)$ exista y no sea singular.

# Ecuaciones Simultáneas: Propiedades

- Reemplazando la expresión elegida para $\hat{W}$ en (10) tenemos,

$$\hat{\hat{\beta}} = [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'Y \tag{12}$$

- Que es el estimador de mínimos cuadrados en dos etapas para sistemas de ecuaciones (S2SLS).
- El estimador de S2SLS no es necesariamente el estimador asintóticamente más eficiente.
- Para ver esto intuitivamente considere (11) que es la ecuación que queremos que sea mínima.
- Supuesto SIV.4: $\hat{W} = \Lambda^{-1}$.

# Ecuaciones Simultáneas: Propiedades

- Bajo los supuestos SIV.1-SIV.4, el estimador de GMM es el más eficiente entre todos los estimadores de GMM y su varianza asintótica se reduce a,

$$AVar[\sqrt{N}(\hat{\beta} - \beta)] = (C'\Lambda^{-1}C)^{-1}$$

- Cualquier estimador consistente de $\Lambda$ nos da un estimador GMM eficiente. Un procedimiento común es el siguiente.

- 1. Sea $\hat{\hat{\beta}}$ el estimador de S2SLS.
  2. Obtenga el vector de residuos, $(G \times 1)$, $\hat{\hat{u}}_i = y_i - X_i\hat{\hat{\beta}}$, $i = 1, 2, \ldots, N$.
  3. Un estimador consistente de $\Lambda$ es $\hat{\Lambda} = \frac{1}{N}\sum_{i=1}^{N} Z_i'\hat{\hat{u}}_i\hat{\hat{u}}_i'Z_i$.
  4. Elija $\hat{W} = \hat{\Lambda}^{-1} = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\hat{\hat{u}}_i\hat{\hat{u}}_i'Z_i\right)^{-1}$ y use esta matriz para obtener el estimador GMM.

# Ecuaciones Simultáneas: Propiedades

- La varianza asintótica del estimador de GMM se estima consistentemente con,

$$\left[(X'Z)\left(\sum_{i=1}^{N} Z_i'\hat{\hat{u}}_i\hat{\hat{u}}_i'Z_i\right)(Z'X)\right]^{-1}$$

donde $\hat{\hat{u}}_i = y_i - X_i\hat{\hat{\beta}}$

- Note que este estimador no pone ninguna restricción sobre la matriz de varianzas y covarianzas de los errores. Usualmente el estimador recibe el nombre de Estimador Chi-Cuadrado Mínimo.
- En el enfoque tradicional de ecuaciones simultáneas, se asume que $Var(u_i|Z_i)$ es constante.
- Este supuesto nos da un estimador que estará en el medio entre el de S2SLS y el estimador más eficiente.
- El estimador se conoce como mínimos cuadrados en tres etapas y es un estimador GMM que usa una matriz de ponderación particular.

# Ecuaciones Simultáneas: Propiedades

- Definamos $\hat{\hat{u}}_i = y_i - X_i \hat{\hat{\beta}}$ como los residuos de la estimación por S2SLS y la matriz de dimensión $G \times G$,

$$\hat{\Omega} \equiv \frac{1}{N} \sum_{i=1}^{N} \hat{\hat{u}}_i' \hat{\hat{u}}_i$$

- Note que este último estimador es un estimador consistente de $\Omega$ usando argumentos similares a los que utilizamos para MCG.
- La matriz de ponderación utilizada por 3SLS es,

$$\hat{W} = \left( \frac{1}{N} \sum_{i=1}^{N} Z_i' \hat{\Omega} Z_i \right)^{-1} = \left( \frac{Z'(I_N \otimes \hat{\Omega})Z}{N} \right)^{-1} \tag{13}$$

- El estimador de 3SLS es,

$$\hat{\hat{\hat{\beta}}} = [X'Z(Z'(I_N \otimes \hat{\Omega})Z)^{-1} Z'X]^{-1} X'Z(Z'(I_N \otimes \hat{\Omega})Z)^{-1} Z'Y \tag{14}$$

# Ecuaciones Simultáneas: Propiedades

- Si se cumple el Supuesto SIV.5:

$$E(Z_i' u_i u_i' Z_i) = E(Z_i' \Omega Z_i), \quad \text{con } \Omega = E(u_i u_i')$$

  el estimador de 3SLS (14) es eficiente.
- Una condición suficiente para que SIV.5 se satisfaga es que $E(u_i u_i' | Z_i) = E(u_i u_i)$.
- El supuesto SIV.5 implica que el supuesto SIV.4 se cumple porque (13) estima consistentemente $\Lambda^{-1}$.
- Bajo los supuestos SIV.1-SIV.3 y SIV.5 el estimador 3SLS es un estimador de GMM óptimo.
- El estimador de la matriz de varianzas y covarianzas asintóticas del estimador de 3SLS es,

$$\left[ (X'Z) \left( \sum_{i=1}^{N} Z_i' \hat{\Omega} Z_i \right)^{-1} (Z'X) \right]^{-1} = [X'Z\{Z'(I_N \otimes \hat{\Omega})Z\}^{-1} Z'X]$$

# Ecuaciones Simultáneas: Propiedades

- Remark 1: sin el supuesto SIV.5 el estimador 3SLS es generalmente menos eficiente asintóticamente que el estimador chi-cuadrado mínimo y el estimador de la varianza asintótica para 3SLS no es apropiado.
- Remark 2: Aún con el supuesto SIV.5, el estimador de 3SLS NO es más eficiente asintóticamente que el estimador chi-cuadrado mínimo. Ambos estimadores son asintóticamente equivalentes.
- Remark 3: Si el estimador chi-cuadrado mínimo nunca es peor asintóticamente que el estimador de 3SLS, por qué se usa 3SLS?
    - Históricamente uno de los primeros estimadores. GMM a partir de la década de los 80.
    - 3SLS puede tener mejores propiedades en muestra chica.

# Agenda

# Ecuaciones Simultáneas: Inferencia

- Los estadísticos estándar se construyen utilizando el estimador de GMM y la estimación de su matriz de varianzas y covarianzas apropiada.
- Una forma alternativa de contrastar restricciones lineales se basa en la diferencia en la función objetivo de GMM con y sin las restricciones impuestas.
- Para aplicar este estadístico se debe utilizar el estimador de GMM óptimo, de forma tal que estime consistentemente $[Var(Z_i' u_i)]^{-1}$.
- Usando resultados de teoría asintótica anteriores, como $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} Z_i' u_i \xrightarrow{d} N(0, \Lambda)$,

$$\left( \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Z_i' u_i \right)' \hat{W} \left( \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Z_i' u_i \right) \xrightarrow{d} \chi^2_L$$

# Ecuaciones Simultáneas: Inferencia

- Definamos a $\hat{\beta}$ como el estimador GMM usando la matriz de ponderación óptima $\hat{W}$ obtenida sin imponer las restricciones y a $\tilde{\beta}$ como el estimador GMM usando la misma matriz ponderadora $\hat{W}$ pero obtenida con las restricciones impuestas.
- Bajo la hipótesis nula el estadístico de distancia GMM tiene la siguiente distribución,

$$\left[ (\sum_{i=1}^{N} Z_i' \tilde{u}_i)' \hat{W} (\sum_{i=1}^{N} Z_i' \tilde{u}_i) - (\sum_{i=1}^{N} Z_i' \hat{u}_i)' \hat{W} (\sum_{i=1}^{N} Z_i' \hat{u}_i) \right] / N \xrightarrow{d} \chi^2_Q$$

  Donde $\hat{u}_i = y_i - X_i \hat{\beta}$ son los residuos del modelo no restringido y $\tilde{u}_i = y_i - X_i \tilde{\beta}$ son los residuos del modelo restringido.
- Si hay más instrumentos que variables explicativas endógenas, entonces se pueden contrastar las restriciones de sobreidentificación.

- En nuestro modelo si $X_i$ es $G \times K$ y $Z_i$ es $G \times L$, entonces hay $L > K$ restricciones de sobreidentificación.
- Para contrastar la hipótesis nula $H_0 : E(Z_i' u_i) = 0$ se utiliza el siguiente estadístico,

$$\left( \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Z_i' \hat{u}_i \right)' \hat{W} \left( \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Z_i' \hat{u}_i \right) \xrightarrow{d} \chi^2_{L-K}$$

# Agenda

# Ecuaciones Simultáneas: Identificación

- Para motivar el análisis consideremos un modelo de oferta y demanda de trabajo,

$$h^s(w) = \gamma_1 \log(w) + z_{(1)}\delta_{(1)} + u_1 \quad (15)$$

$$h^d(w) = \gamma_2 \log(w) + z_{(2)}\delta_{(2)} + u_2 \quad (16)$$

- Asumiendo que las horas observadas, $h$, y los salarios observados, $w$ igualan la oferta y la demanda: $h = h^s(w) = h^d(w)$.
- Definiendo $y_1 = h$ e $y_2 = log(w)$ tenemos:,

$$y_1 = \gamma_1 y_2 + z_{(1)}\delta_{(1)} + u_1 \quad (17)$$

$$y_1 = \gamma_2 y_2 + z_{(2)}\delta_{(2)} + u_2 \quad (18)$$

- Que necesitamos para identificar los parámetros de la curva de oferta?

# Ecuaciones Simultáneas: Identificación

- Intuitivamente no podemos identificar la función de oferta de la de demanda si $z_{(1)}$ y $z_{(2)}$ contienen exactamente los mismos elementos.
- Formalmente, supongamos que las curvas tienen pendientes diferentes. Restando la ecuación (18) de la (17) y dividiendo ambos miembros por $(\gamma_2 - \gamma_1)$ tenemos,

$$y_2 = z_{(1)}\pi_{21} + z_{(2)}\pi_{22} + v_2 \tag{19}$$

  donde $\pi_{21} \equiv \delta_{(1)}/(\gamma_2 - \gamma_1)$, $\pi_{22} \equiv -\delta_{(2)}/(\gamma_2 - \gamma_1)$ y $v_2 \equiv (u_1 - u_2)/(\gamma_2 - \gamma_1)$.
- La ecuación (19) es la forma reducida de $y_2$ porque lo expresa como función de variables exógenas y un error, $v_2$, que es ortogonal a todas esas variables.

# Ecuaciones Simultáneas: Identificación

- Dada la ecuación (17) y la forma reducida (19) podemos usar la condición de identificación presentada para el caso de una sola ecuación.
- Esa condición decía que para poder identificar los parámetros de la curva de oferta (17) necesitamos que la ecuación de la forma reducida contenga al menos una variable exógena que no esté en (17).
- El último punto significa que al menos un elemento de $z_{(2)}$ que no está en $z_{(1)}$ debe tener un coeficiente diferente de cero en (19).
- Podemos establecer la misma condición de identificación pero utilizando las ecuaciones estructurales. Como $\pi_{22}$ es proporcional a $\delta_{(2)}$, entonces en la ecuación (18) al menos un elemento de $z_{(2)}$ que no está en $z_{(1)}$ debe tener un coeficiente diferente de cero.

# Ecuaciones Simultáneas: Identificación

- En palabras, para identificar la curva de oferta de trabajo se necesita que en la función de demanda haya una variable exógena que no aparezca en la de oferta.
- La identificación de la curva de demanda es similar, esto es, se necesita que la curva de oferta contenga una variable exógena que no esté presente en la de demanda.
- Ahora extenderemos el análisis a un sistema general.

# Ecuaciones Simultáneas: Identificación

- Considere el siguiente sistema de ecuaciones,

$$
\begin{aligned}
y_1 &= y_{(1)}\gamma_{(1)} + z_{(1)}\delta_{(1)} + u_1 \\
&\vdots \\
y_G &= y_{(G)}\gamma_{(G)} + z_{(G)}\delta_{(G)} + u_G
\end{aligned}
$$

  Donde $y_{(h)}$ es $1 \times G_h$, $\gamma_{(h)}$ es $G_h \times 1$, $z_{(h)}$ es $1 \times M_h$ y $\delta_{(h)}$ es $M_h \times 1$, $h = 1, 2, \ldots, G$.
- El vector $y_{(h)}$ denota las variables endógenas que aparecen en el lado derecho de la ecuación $h$. Por convención $y_{(h)}$ puede contener cualquier variable endógena excepto $y_h$.
- Las variables en $z_{(h)}$ son las variables exógenas que aparecen en la ecuación $h$.
- El vector de dimensión $1 \times M$ de todas las variables exógenas $z$ se asume que satisface $E(z'u_g) = 0$, $g = 1, 2, \ldots, G$.
- Además, sin pérdida de generalidad asumimos que $E(z'z)$ no es singular.

# Ecuaciones Simultáneas: Identificación

- Ahora extendemos la identificación al sistema presentado. Comencemos identificando la primera ecuación del sistema,

$$y_1 = y_{(1)}\gamma_{(1)} + z_{(1)}\delta_{(1)} + u_1 = x_{(1)}\beta_{(1)} + u_1$$

- Asumiendo que existe, escribamos la forma reducida para $y_{(1)}$ como,

$$y_{(1)} = z\Pi_{(1)} + v_{(1)}$$

donde $E(z'v_{(1)}) = 0$.
- Definamos la matriz $S_{(1)}$ de dimensión $M \times M_1$ que contiene solo unos y ceros de forma tal que $z_{(1)} = zS_{(1)}$.
- La condición de rango que vimos puede establecerse como: Rango $E(z'x_{(1)}) = K_1$, donde $K_1 = G_1 + M_1$.
- $E(z'x_{(1)}) = E[z'(z\Pi_{(1)}, zS_{(1)})] = E(zz)[\Pi_{(1)}|S_{(1)}]$
Como $E(z'z)$ tiene rango completo $M$, la condición de rango puede reescribirse como: Rango $[\Pi_{(1)}|S_{(1)}] = G_1 + M_1$.

# Ecuaciones Simultáneas: Identificación

- En palabras, la matriz $[\Pi_{(1)}|S_{(1)}]$ tiene que tener rango completo.
- Note que $[\Pi_{(1)}|S_{(1)}]$ es de dimensión $M \times (G_1 + M_1)$, por lo tanto, una condición necesaria para tener rango completo es que $M \geq G_1 + M_1$ ó $M - M_1 \geq G_1$.
- Esta última es la condición que ya habíamos encontrado antes, el número de variables exógenas que no aparecen en la primera ecuación, $M - M_1$, debe ser al menos tan grande como el número de variables endógenas que aparecen en el lado derecho de la primera ecuación, $G_1$.
- La condición de orden es una condición necesaria para la identificación pero no es suficiente. Es decir, si no se cumple la ecuación no estará identificada y si se cumple, podría estar identificada.

# Ecuaciones Simultáneas: Identificación

- Como hicimos para el caso de las curvas de oferta y demanda de trabajo, podemos expresar la condición de rango en lugar de utilizar los coeficientes de la forma reducida, utilizando los coeficientes de las ecuaciones estructurales.
- Escribamos el sistema sin ninguna restricción,

$$
\begin{aligned}
y\gamma_1 + z\delta_1 + u_1 &= 0 \\
&\vdots \\
y\gamma_G + z\delta_G + u_G &= 0
\end{aligned}
$$

donde $y = (y_1, y_2, \ldots, y_G)$ es el vector $1 \times G$ de todas las variables endógenas y $z = (z_1, z_2, \ldots, z_M)$ es el vector $1 \times M$ de todas las variables exógenas; $\gamma_g$ es $G \times 1$ y $\delta_g$ es $M \times 1$ para todo $g = 1, 2, \ldots, G$.

# Ecuaciones Simultáneas: Identificación

- Reescribiendo el sistema en forma matricial tenemos,

$$y\Gamma + z\Delta + u = 0$$

  donde $u = (u_1, \ldots, u_G)$, $\Gamma$ es $G \times G$ con columna $g$ dada por $\gamma_g$ y $\Delta$ es $M \times G$ con columna $g$ dada por $\delta_g$.
- Asumamos que $\Gamma$ no es singular de forma tal de que exista forma reducida y sea $\Sigma = E(u'u)$ la matriz, $G \times G$, de varianzas y covarianzas de $u$.
- El sistema en forma reducida queda,

$$y = z(-\Delta\Gamma^{-1}) + u(-\Gamma^{-1}) \equiv z\Pi + v$$

- Definamos $\Lambda = E(v'v) = \Gamma^{-1'}\Sigma\Gamma^{-1}$ como la matriz de varianzas y covarianzas de la forma reducida.

# Ecuaciones Simultáneas: Identificación

- La pregunta relevante es, bajo que condiciones podemos recuperar los parámetros estructurales desde la forma reducida?.
- La respuesta es que sin algunas restricciones sobre los parámetros no podremos identificar la forma estructural.
- Para ver esto más claramente definamos la matriz $F$ no singular de dimensión $G \times G$ y post-multipliquemos el sistema de ecuaciones estructurales,

$$y\Gamma F + z\Delta F + uF = 0, \text{ ó } y\Gamma^* + z\Delta^* + u^* = 0$$

- Es fácil comprobar que esta última ecuación tiene exactamente la misma forma reducida que el sistema original.
- Este resultado significa que sin restricciones en los parámetros estructurales existen muchas estructuras equivalentes, en el sentido de que llevan a la misma forma reducida. De hecho existen tantas estructuras equivalentes como matrices no singulares $F$.

# Ecuaciones Simultáneas: Identificación

- Si $B \equiv \begin{pmatrix} \Gamma \\ \Delta \end{pmatrix}$ es la matriz $(G \times M) \times G$ de parámetros estructurales, definimos a $F$ como una transformación lineal admisible si se cumple que,
  1. $BF$ satisface todas las restricciones de $B$
  2. $F'\Sigma F$ satisface todas las restricciones de $\Sigma$.
- Para identificar el sistema necesitamos suficiente información sobre los parámetros estructurales de forma tal que $F = I_G$ sea la única transformación lineal admisible.
- Como antes, consideremos la identificación de la primera ecuación del sistema,

$$y\gamma_1 + z\delta_1 + u_1 = 0 \quad \text{ó}$$

$$\gamma_{11}y_1 + \gamma_{12}y_2 + \cdots + \gamma_{1G}y_G + \delta_{11}z_1 + \delta_{12}z_2 + \cdots + \delta_{1M}y_M + u_1 = 0$$

# Ecuaciones Simultáneas: Identificación

- La primera restricción que vamos a imponer es la restricción de normalización haciendo uno de los coeficientes de $\gamma_1$ igual a $-1$.
- Sea $\beta_1 \equiv (\gamma_1', \delta_1')'$ el vector, $(G+M) \times 1$, de parámetros estructurales de la primera ecuación con la restricción de normalización impuesta. Esto es, hay $(G+M)-1$ parámetros desconocidos en $\beta_1$.
- Asumamos que las restricciones sobre $\beta_1$ pueden expresarse como: $R_1\beta_1 = 0$. Donde $R_1$ es una matriz $J_1 \times (G+M)$ de constantes conocidas.
- Es decir que $J_1$ es el número de restricciones sobre $\beta_1$ (además de la normalización). Asumiendo que el *Rango* $R_1 = J_1$ no tenemos restricciones redundantes.
- Lo que queremos averiguar es cuando son estas restricciones suficientes como para identificar los parámetros de $\beta_1$.
- Nuevamente definamos $F$ como una matriz de dimensión $G \times G$ escrita en términos de sus columnas: $F = (f_1, f_2, \ldots, f_G)$.

# Ecuaciones Simultáneas: Identificación

- Considere la siguiente transformación lineal de $B$, $B^* = BF$, tal que la primera columna de $B^*$ es $\beta_1^* = Bf_1$.
- Necesitamos determinar una condición tal que la ecuación $R_1\beta_1 = 0$ nos permita distinguir a $\beta_1$ de cualquier otro $\beta_1^*$.
- El vector $\beta_1^*$ satisface las restricciones lineales de $R_1$ si y solo si, $R_1\beta_1^* = R_1(Bf_1) = (R_1B)f_1 = 0$ que se satisface naturalmente cuando $f_1 = e_1 = (1, 0, 0, \ldots, 0)'$ que implica $\beta_1^* = \beta_1$.
- Decir que la condición anterior se satisface para $f_1 = e_1$ significa que el espacio nulo de $R_1B$ tiene dimensión uno.
- En términos del rango de la matriz esto es lo mismo que decir que:
  $Rango\ R_1B = G - 1$
  Esta es la condición de rango para la identificación de $\beta_1$ en la primera ecuación estructural.

# Ecuaciones Simultáneas: Identificación

- Note que la matriz $R_1 B$ puede escribirse como,
  $R_1 B = [R_1 \beta_1, R_1 \beta_2, \ldots, R_1 \beta_G]$
  donde $\beta_g$ es el vector $(G + M) \times 1$ de parámetros estructurales en la ecuación $g$.
- Ahora, por nuestros supuestos sabemos que $R_1 \beta_1 = 0$, por lo que la primera columna de $R_1 B$ es de ceros y la matriz $R_1 B$ no puede tener rango mayor a $G - 1$.
- También sabemos que $\Gamma$ tiene rango completo (i.e. no es singular) de forma tal que $B$ necesariamente tiene rango $G$. Por lo tanto, para que la condición de rango se cumpla necesitamos que *Rango* $R_1 \geq G - 1$.
- Nosotros asumimos que *Rango* $R_1 = J_1$ por lo que una condición necesaria para que la primera ecuación esté identificada es: $J_1 \geq G - 1$.
- $J_1 \geq G - 1$ es la condición de orden general en un sistema.

# Ecuaciones Simultáneas: Identificación

- Note que cuando las restricciones sobre $\beta_1$ son la de normalización y restricciones de exclusión la matriz $R_1$ consiste de ceros y unos de forma tal que el número de filas en $R_1$ iguala el número de variables endógenas excluidas, $G - G_1 - 1$, más el número de variables exógenas excluidas, $M - M_1$.
- En otras palabras: $J_1 = (G - G_1 - 1) + (M - M_1)$, entonces la condición de orden es,
  $(G - G_1 - 1) + (M - M_1) \geq G - 1 \longrightarrow$
  $(M - M_1) \geq G - 1 - (G - G_1 - 1) \longrightarrow$
  $(M - M_1) \geq G_1$.
  Es decir, el número de variables exógenas excluidas debe ser igual o mayor al número de variables endógenas incluidas en la ecuación.

# Ecuaciones Simultáneas: Identificación

- Resumen de los pasos necesarios para chequear la identificación de la primera ecuación del sistema:
  1. Normalice uno de los elementos de $\gamma_1$ igual a $-1$.
  2. Defina la matriz $R_1$ de dimensión $J_1 \times (G + M)$ tal que $R_1\beta_1 = 0$.
  3. Si $J_1 < G - 1$ la ecuación no está identificada.
  4. Si $J_1 \geq G - 1$ la ecuación podría estar identificada.
  5. Sea $B$ la matriz de todos los parámetros estructurales con la condición de normalización impuesta. Si *Rango* $R_1 B = G - 1$, entonces la ecuación está identificada.
- Cuando la condición de rango se satisface, es útil refinar el significado de identificación. En este sentido cuando $J_1 = G - 1$ decimos que la ecuación está exactamente identificada.
- Si $J_1 > G - 1$ decimos que la ecuación está sobreidentificada. En este caso tenemos $J_1 - (G - 1)$ restricciones de sobreidentificación.

# Ecuaciones Simultáneas: Identificación

- Lo visto hasta ahora es el caso más importante de identificación en ecuaciones simultáneas: identificar la ecuación usando solo restricciones de exclusión.
- No obstante, hay ejemplos en la literatura donde la endogenidad está causada por variables omitidas o errores de medición y donde la teoría económica impone restricciones entre ecuaciones.
- Un ejemplo es la estimación de un sistema de oferta y demanda utilizando una encuesta de ingresos y gastos.
- Tipicamente la característica de estas encuestas es que no se pregunta el precio del artículo comprado sino que se reporta el valor unitario del mismo.
- Usar el valor unitario en lugar del precio produce un error de medición en la variable.
- En estos casos las restricciones entre ecuaciones se pueden utilizar para identificar las ecuaciones.

# Ecuaciones Simultáneas: Identificación

- Considere el siguiente sistema de ecuaciones,

$$
\begin{aligned}
y_1 &= \gamma_{12} y_2 + \delta_{11} z_1 + \delta_{12} z_2 + \delta_{13} z_3 + u_1 && (20) \\
y_2 &= \gamma_{21} y_1 + \delta_{21} z_1 + \delta_{22} z_2 + u_2 && (21)
\end{aligned}
$$

  donde cada $z_j$ no está correlacionada con $u_1$ y $u_2$.
- Sin información adicional, la ecuación (20) no está identificada y la ecuación (21) está identificada si $\delta_{13} \neq 0$.
- Ahora supongamos que $\delta_{12} = \delta_{22}$. Como $\delta_{22}$ está identificado en (21) podemos tratarlo como un parámetro conocido en la identificación de (20). Esto es, se puede escribir (20) como

$$y_1 - \delta_{12} z_2 = \gamma_{12} y_2 + \delta_{11} z_1 + \delta_{13} z_3 + u_1 \qquad (22)$$

  donde ahora $y_1 - \delta_{12} z_2$ se conoce.

# Ecuaciones Simultáneas: Identificación

- Ahora, el lado derecho de (22) tiene una variable endógena, $y_2$, y dos exógenas $z_1$ y $z_3$.
- Como $z_2$ ahora está excluida del lado derecho, se puede utilizar como instrumento para $y_2$, si aparece en la forma reducida de $y_2$.
- Este es el caso si $\delta_{12} = \delta_{22} \neq 0$
- Este procedimiento sugiere una forma de estimar el sistema:
- Primero: estime la ecuación (21) por 2SLS usando $(z_1, z_2, z_3)$ como instrumentos y denote por $\hat{\delta}_{22}$ al estimador de $\delta_{22}$.
- Segundo: estime

$$y_1 - \hat{\delta}_{12} z_2 = \gamma_{12} y_2 + \delta_{11} z_1 + \delta_{13} z_3 + \text{error}$$

  por 2SLS usando $(z_1, z_2, z_3)$ como instrumentos.
- Como $\hat{\delta}_{22} \longrightarrow \delta_{22}$, cuando $\delta_{12} = \delta_{22} \neq 0$, este último procedimiento produce estimadores consistentes de $\gamma_{12}$, $\delta_{11}$ y $\delta_{13}$.

# Ecuaciones Simultáneas: Identificación

- También es posible estimar el sistema imponiendo la restricción:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_2 & z_1 & z_2 & z_3 & 0 & 0 \\ 0 & 0 & z_2 & 0 & y_1 & z_1 \end{pmatrix} \begin{pmatrix} \gamma_{12} \\ \delta_{11} \\ \delta_{12} \\ \delta_{13} \\ \gamma_{21} \\ \delta_{21} \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

  y usando GMM o 3SLS.
- El parámetro $\delta_{22}$ no aparece porque se impuso la restricción $\delta_{12} = \delta_{22}$.
- La matriz de instrumentos es $I_2 \otimes z$ y contiene solo las variables exógenas como instrumentos en cada ecuación.
- Como $I_2 \otimes z$ tiene 6 columnas (hay seis elementos en el vector de parámetros) la condición de orden se satisface.