

# Capítulo 7

## Variables instrumentales

# Variables instrumentales

- En ausencia del problema de autoselección:

$$Y_i = B_0 + B_1 D_i + B_2 X_{1i} + \dots + B_{K+1} X_{Ki} + U_i$$

- $D_i$ : Indicador de participación en el programa.
- $Y_i$ : Variable de resultado.
- $X_{1i}, \dots, X_{ki}$ : Vector de variables observadas del individuo.
- $U_i$ : Error, elementos no observados o no medidos.
  - Determinan la variables de resultado.
  - No están contenidos en el vector  $X_i$ .

# Variables instrumentales

- En ausencia del problema de autoselección:

$$Y_i = B_0 + B_1 D_i + B_2 X_{1i} + \dots + B_{K+1} X_{Ki} + U_i$$

- $B_1$ : Efecto del tratamiento.
  - Se obtiene al comparar la media de la variable de resultado del grupo de tratamiento y la del grupo de control.

# Variables instrumentales

- Problema de autoselección:
  - Un elemento del error está correlacionado con la participación en el programa.
  - Los estimadores del programa por mínimos cuadrados ordinarios (MCO) son inconsistentes y sesgados.
  - Ejemplo: La habilidad innata del individuo está correlacionada con la participación en el programa, pero la habilidad innata no está registrada en las bases de datos.

# Variables instrumentales

El método de variables instrumentales consiste en encontrar una parte de la variación en  $D_i$  que no esté correlacionada con el error  $U_i$ , y estimar el efecto del programa utilizando sólo esa parte de la variación que sí se puede considerar exógena. Por lo tanto, necesitamos encontrar una parte de la variación en  $D_i$  que se deba a una variable exógena, conocida como variable instrumental,  $Z_i$ .

# Variables instrumentales

- Una variable instrumental debe cumplir dos condiciones:
  - **Relevancia:**  $Cov(D_i, Z_i) \neq 0$ 

El instrumento tiene un buen poder de predicción de la probabilidad de participación en el programa.
  - **Validez:**  $Cov(D_i, U_i) = 0$ 

El término del error no está correlacionado con la variable instrumental  $Z_i$ , porque ésta es exógena.

# Ejemplo: *Canasta 1*

- El programa *Canasta* entrega un mercado a los hogares para mejorar el estado nutricional de los niños.
- La participación en el programa es voluntaria.
- La aplicación al programa es costosa en términos de tiempo.
  - **Existe sesgo de selección:** Las madres participantes probablemente son más proactivas.

# Ejemplo: *Canasta 1*

- El indicador de participación está correlacionado con variables no observables de la madre:
  - Motivación.
  - Disposición para invertir en la nutrición de sus hijos.
- Estas variables también afectan la salud y la nutrición de los niños.  
  
→ El estimador del programa a través de MCO estaría sesgado.



# Ejemplo: *Canasta 1*

- Condiciones para la variable instrumental:
  - Debe explicar la probabilidad de participar en el programa *Canasta*.
  - No debe estar relacionada con la motivación de las madres a invertir tiempo para mejorar la salud de sus hijos (proactivas).
  - No debe afectar el estado nutricional de los niños directamente.
- Variable instrumental:
  - Distancia del hogar a la oficina administradora del programa.

# Variables instrumentales

- Fuente de variación exógena:
  - Afecta la participación en el programa.
  - Es independiente de las variables observadas y no observadas de las familias.
- Ejemplo: *Canasta 1*
  - Algunos individuos participan en el programa porque viven cerca de la oficina administradora, independientemente de si están más motivados o no.
  - La distancia del hogar familiar a la oficina no afecta el estado nutricional de los niños.

# Variables instrumentales

El efecto del programa se identifica a partir de un *subconjunto* de individuos. El subconjunto de individuos que cambia su decisión de participar en el programa, debido a la cercanía de su hogar a la oficina administradora.

→ El efecto estimado es un efecto LOCAL y no promedio.

## Ejemplo: *Canasta 2*

- **Población elegible:** La participación en el programa se determinó con base en un sorteo entre los elegibles; pero:
  - Algunos ganadores del programa deciden no reclamar el mercado.
  - Algunos perdedores insisten hasta que les dan el mercado.
- La decisión de algunos ganadores de no reclamar el mercado y de los perdedores de insistir para obtenerlo está relacionada con variables observadas y no observadas que también afectan el estado nutricional de los niños.

## Ejemplo: *Canasta 2*

- Variable instrumental:

$$Z_i = \begin{cases} 1 & \text{Si ganó el sorteo} \\ 0 & \text{De lo contrario} \end{cases}$$

- Condiciones para la variable instrumental:

- Está altamente correlacionada con la probabilidad de participación efectiva en el programa *Canasta*.
- No está relacionada con las variables no observadas (no medidas) de las madres que afectan directamente el estado nutricional de sus hijos. Por ej., la motivación de las madres o su interés en el desarrollo de sus hijos.

# Variables instrumentales

- **Supuesto de monotonicidad:**
  - El estatus del tratamiento  $D_i$  es una función monótona creciente del nivel de  $Z_i$ .
  - Esto implica, por ejemplo, que si un individuo no participa en el programa cuando ganó la rifa, entonces tampoco lo haría en el caso en que no hubiera resultado favorecido en el sorteo.
  - Necesario para garantizar que se identifica el efecto del programa.

# Supuesto de monotonicidad

- **Variable instrumental:**  $Z_i = \begin{cases} 1 & \text{Si ganó el sorteo} \\ 0 & \text{De lo contrario} \end{cases}$
- $D_i$ : Indicador de participación en el programa.
- **Tipos de individuos:**
  - Los que siempre participan:  $D_i = 1 \mid Z_i = 1 \text{ \& } D_i = 1 \mid Z_i = 0$ .
  - Los que nunca participan:  $D_i = 0 \mid Z_i = 1 \text{ \& } D_i = 0 \mid Z_i = 0$ .
  - Los cooperativos:  $D_i = 1 \mid Z_i = 1 \text{ \& } D_i = 0 \mid Z_i = 0$ .
  - Los desafiantes:  $D_i = 0 \mid Z_i = 1 \text{ \& } D_i = 1 \mid Z_i = 0$ .

# Supuesto de monotonicidad

- El supuesto de monotonicidad requiere que no exista el grupo de los desafiantes.
- El supuesto no se cumple cuando:
  - Los individuos no participaron cuando ganaron la rifa.
  - Los individuos participaron cuando perdieron la rifa.



# Variables instrumentales

## Estimador

- Suponga un modelo sencillo de la variable de resultado:

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

- Entonces:

$$\text{cov}(Y_i, Z_i) = \text{cov}(\beta_0 + \beta_1 D_i + u_i, Z_i)$$

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)}$$

Se conoce como el  
estimador de Wald

# Análogo muestral

- El análogo muestral es el estimador de variables instrumentales:

$$\hat{\beta}_1 = \frac{\sum_i (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_i (Z_i - \bar{Z})(D_i - \bar{D})}$$

- $D_i$ : Indicador de participación en el programa.
- $Y_i$ : Variable de resultado.
- $Z_i$ : Variable instrumental.

# Variables instrumentales mínimos cuadrados en dos etapas (MC2E)

Este estimador se puede obtener con base en un procedimiento de dos etapas:

- **Primera etapa:**

$$\hat{D}_i = \hat{\alpha}_0 + \hat{\alpha}_1 Z_i + \hat{\alpha}_2 X_{1i} + \dots + \hat{\alpha}_{K+1} X_{Ki}$$

- Se predice la decisión de participación atribuible a cambios en el instrumento.
- $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_{k+1}$  son los estimadores de MCO de la regresión del indicador de tratamiento sobre las variables exógenas y la variable instrumental.

# Variables instrumentales

## MC2E

- Segunda etapa:

$$Y_i = \beta_0 + \beta_1 \hat{D}_i + \beta_2 X_{1i} + \dots + \beta_{K+1} X_{Ki} + U_i$$

Se utiliza la variación en  $D_i$  que se debe al instrumento para identificar el efecto del programa. Esta regresión se estima por MCO.

$\hat{D}_i$  es la predicción de  $D_i$  que resulta de la primera etapa.

# Variables instrumentales

## MC2E

- $Z_i$  se conoce como restricción de exclusión, pues:
  - Explica la decisión de participar en el programa.
  - No afecta directamente la variable de resultado  $Y_i$ .
- Cuando la variable explicativa de interés (endógena) es continua, la primera etapa se corre por MCO.

# Variables instrumentales

## MC2E

- Cuando la variable explicativa de interés es dicotómica ( $D_i$ ), entonces:
  - El modelo de probabilidad lineal en la primera etapa:
    - Es heterocedástico por construcción.
    - Produce predicciones de la probabilidad negativas o mayores a uno.
  - En cambio, el modelo de elección discreta no lineal:
    - Es logit o probit.
    - Produce predicciones razonables.
    - Se utiliza la probabilidad predicha de participación en el programa como instrumento.

# Elección de los instrumentos

- Instrumentos más usados:
  - Disponibilidad o facilidad de acceso al programa.
  - Diferencias en los precios de acceso al programa:
    - Copago.
    - Costos de desplazamiento.
    - Tiempos de espera.
  - Políticas de asignación aleatorias.

# Evaluación de la variable instrumental

- El investigador debe:
  - Proveer evidencia de que la variable instrumental es relevante.
    - Mostrar que el instrumento es estadísticamente significativo en la primera etapa.
  - Proveer evidencia de que la variable instrumental es exógena.
    - Demostrar que  $Z_i$  y  $U_i$  no están correlacionadas.
    - Problema:  $U_i$  no se observa.



# Pruebas de relevancia

- **Regresión simple:**
  - Se evalúa la significancia estadística de los coeficientes asociados a instrumentos con base en pruebas  $t$ .
  - Se evalúa la significancia estadística del conjunto de instrumentos (si tiene más de uno) con base en la prueba  $F$ .

# Pruebas de exogeneidad

## Modelos sobreidentificados

El modelo está sobreidentificado si se dispone de más de un instrumento para la participación en el programa.

### Prueba de sobreidentificación de Sargan:

1. Estimar el efecto del programa por MC2E utilizando solo un instrumento.
2. Calcular el error predicho con base en los estimadores de MC2E.

# Prueba de sobreidentificación (Sargan)

3. Verificar que el instrumento excluido no está correlacionado con el término del error:

Regresión del término del error contra todas las variables exógenas.

4. Construir prueba chi cuadrado:  $\chi^2 = N.R^2 \sim \chi_q^2$

- $R^2$ : Ajuste del modelo del término del error contra todas las variables exógenas.
- $N$ : Número de observaciones.
- $q$ : Número de instrumentos adicionales.
- $H_o$ : Al menos uno de los instrumentos no es exógeno:
  - Si se rechaza por lo menos un instrumento no es exógeno.
  - En ese caso, el estimador del efecto del programa es sesgado.


# Prueba de sobreidentificación (Sargan)

- Si el instrumento está correlacionado con el término del error no es un instrumento válido para  $D_i$ .
- La prueba de sobreidentificación no dice nada sobre el otro instrumento:
  - Si están escogidos bajo la misma lógica el resultado se puede aplicar a ambos instrumentos.
  - Ejemplo: Educación de la madre y del padre.

# Variables instrumentales: Problemas potenciales

1. Es difícil encontrar variables instrumentales válidas y relevantes.
2. Si el instrumento es débil (es decir, no predice bien la participación), el estimador de variables instrumentales puede amplificar el sesgo.

$$\hat{\beta}_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)}$$

$$\hat{\beta}_1 = \beta_1 + \frac{\text{cov}(u_i, Z_i)}{\text{cov}(D_i, Z_i)}$$


## Variables instrumentales: Problemas potenciales

3. Las variables instrumentales deben cumplir el supuesto de monotonicidad, que es inverificable.
4. El estimador de variables instrumentales identifica un efecto local y no un efecto promedio.

# Variables instrumentales: Problemas potenciales

En particular:

El efecto corresponde al grupo de los individuos cooperativos que generalmente no se puede identificar.

- El grupo sobre el cual se hace inferencia acerca del impacto del programa es desconocido.

# Variables instrumentales

## Ejemplo: *Canasta*

### Programa *Canasta*:

- Tratamiento: Entrega un mercado a la familia.
- Variable de resultado: Puntaje  $-Z$  de la talla según la edad.
- La participación en el programa es voluntaria:
  - Autoselección.
  - El estimador de MCO podría estar sesgado.



# Variables instrumentales

## Ejemplo: *Canasta*

- **Variables instrumentales:**
  - Número de oficinas operadoras del programa en el municipio:
    - Disminuye la congestión.
    - Hace que la solicitud sea rápida y fácil.
  - Distancia desde el hogar de la familia hasta la oficina administradora más cercana:
    - Los costos de transporte son menores.
    - Hace que la solicitud sea más fácil para las madres.

**Y no están relacionadas con la motivación-preocupación de las madres.**

# Conclusiones

Ventajas	Desventajas
<ul style="list-style-type: none"><li>• Permite estimaciones del efecto causal de programas con base en datos no experimentales.</li><li>• Se puede estimar aun en presencia de variables no observadas que determinan la participación en el programa y la variable de resultado.</li></ul>	<ul style="list-style-type: none"><li>• Es difícil encontrar instrumentos válidos.</li><li>• Para el investigador es difícil demostrar la validez de sus instrumentos.</li><li>• Es un estimador local y, por lo tanto, tiene limitaciones en términos del alcance de la interpretación del estimador.</li></ul>

# Ejemplo: Servicio militar y salario como civil (Angrist, 1989)

- **Pregunta:** Efecto del servicio militar en salarios posteriores. Pero participar en el servicio militar es endógeno, dado que el individuo parcialmente decide.
- **Este estudio** utiliza un sorteo que se llevó a cabo para reclutar a los reservistas. Ciertos números del sorteo implicaban que el reservista podía ser convocado, y otros números que no.
- Sin embargo, algunos individuos que “ganaban” el sorteo y podían ser llamados no cumplían el servicio militar por exenciones (como problemas de salud).
- El servicio militar estaba sólo parcialmente determinado por la lotería.

# Ejemplo: Servicio militar y salario como civil

- **Instrumento: Lotería**
  - Algunos números de la lotería eran elegibles para ser llamados por el ejército.
  - Algunos números de la lotería no.
- **Resultado:**
  - El servicio militar reduce los salarios civiles en el largo plazo.
  - No tiene efecto en los veteranos de raza negra o de ascendencia latina.

# Ejemplo: Hogares comunitarios (Attanasio y Vera-Hernández, 2004)

- **Variables de resultado:**
  - Talla según la edad.
  - Peso según la talla.
  - Peso según la edad.
- **Variables exógenas:**
  - Distancia al colegio o escuela.
  - Distancia al hospital más cercano.
  - Distancia a centros de capacitación.
  - Distancia a otros establecimientos donde se prestan servicios del Estado.

# Ejemplo: Hogares comunitarios (Attanasio y Vera-Hernández, 2004)

- **Instrumento:**
  - Distancia del hogar de residencia al hogar comunitario más cercano.
- **Resultado:**
  - La distancia afecta claramente la participación en el programa de los hogares rurales estudiados.

# Ejemplo: Hogares comunitarios (Attanasio y Vera-Hernández, 2004)

Efecto estimado del programa *Hogares comunitarios*

	<b>MCO</b>	<b>Variables instrumentales (VI)</b>
Talla según la edad	-0.059	0.486
	(0.050)	(0.156)
Peso según la edad	0.006	0.274
	(0.046)	(0.170)

Errores estándar entre paréntesis.