PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Taller 4: Material Complementario Variables Instrumentales

Datos para la evaluación de impacto en políticas públicas
Diplomado en Ciencia de Datos para Políticas Públicas

07 de Octubre 2022

Para estimar efectos causales, hay tres tipos de sesgos:

1. **Sesgo de variable omitida**: Sesgo producido por una variable no observada que está correlacionada con el tratamiento.
2. **Sesgo de causalidad simultánea (endogeneidad)**: X causa Y e Y causa X.
3. **Sesgo de error en las variables**: X está medida con error.

Las variables instrumentales se pueden hacer cargo de estos sesgos.

¿Qué es una variable instrumental?

Si tenemos selección en inobservables, entonces no existe una estrategia de condicionamiento que nos permita estimar los efectos causales. En otras palabras, si consideramos el siguiente modelo:

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

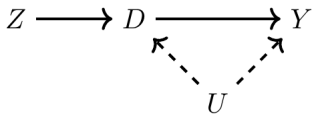
Entonces, si la variable D_i está correlacionada con μ_i, la variable instrumental intentará quebrar X_i en dos partes:

- La parte **que quizás** está correlacionada con μ_i
- La parte **no** está correlacionada con μ_i

Si tenemos:

- Una variable instrumental Z que afecta a D
- Además, Z afecta a Y solo a través de D
- A su vez, Z es independiente de las variables que determinan Y excepto D (**restricción de exclusión**)

En términos gráficos:



Si podemos aislar la parte que no está correlacionada con μ, podemos estimar β_1 (el efecto del tratamiento)de modo consistente. En otras palabras las variables instrumentales solo identifican un efecto causal para cualquier grupo de unidades cuyos comportamientos se modifican como resultado del instrumento (cumplidores o *compliers*)

Lo central para considerar un buen instrumento es la **restricción de exclusión**. Sin embargo, este supuesto no es comprobable de forma empírica por lo que se debe defender a través de la teoría y la lógica.

Los instrumentos deberían sentirse raros

Es probable que nadie se confunda cuando se le diga que el tamaño de la familia reducirá la oferta de mano de obra femenina. En otras palabras, las mujeres que tienen más hijos probablemente trabajan fuera del hogar con menos frecuencia que las que tienen menos hijos. Sin embargo, **¿qué pensarían si dijeramos que las madres cuyos dos primeros hijos eran del mismo sexo estaban menos empleadas que aquellas cuyos dos hijos tenían una proporción de sexos equilibrada?**

Probablemente estarían confundidos porque, después de todo, **¿qué tiene que ver la composición de género de los dos primeros hijos con que una mujer trabaje fuera del hogar?**

No obstante, la composición de género de los dos primeros hijos es importante para una familia tiene preferencias sobre la diversidad de género. Las familias en las que los dos primeros hijos eran varones tienen más probabilidades de volver a intentarlo con la esperanza de tener una niña. Y lo mismo para dos chicas. En la medida en que a los padres les gustaría tener al menos un niño y una niña, tener dos niños podría hacer que lo volvieran a intentar.

Un buen instrumento (dos niños) solo cambia el resultado al cambiar primero alguna variable de tratamiento endógeno (tamaño de la familia), lo que nos permite identificar el efecto causal del tamaño de la familia en algún resultado (oferta de trabajo). Entonces, sin el conocimiento de la variable endógena, las relaciones entre el instrumento y el resultado no tienen mucho sentido. ¿Por qué? Porque el instrumento es irrelevante para los determinantes del resultado excepto por su efecto sobre la variable de tratamiento endógeno.

Entonces:

- **Relevancia del instrumento**:
$$corr(Z_i X_i) \neq 0$$
- **Exogeneidad del instrumento**:
$$corr(Z_i \mu_i) = 0$$

Estimación: Two stage least squares (2SLS)

Nuestra ecuación de interés:

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

Paso 1: aislar la parte de X_i que no está correlacionada con μ_i: regresión entre Z_i y X_i con OLS:

$$X_i = \pi_0 + \pi_1 Z_i + \epsilon_i$$

Donde Z_i no está correlacionado con μ_i y $\pi_0 + \pi_1 Z_i$ no está correlacionado con μ_i. A partir de esta regresión estimar las predicciones $\hat{X}_i$.

Paso 2: Usar las predicciones de $\hat{X}_i$ en vez de X_i en la eucación de interés con OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \mu_i$$

Esto nos permite estimar estimar β_1 como el efecto causal del tratamiento de modo consistente. Algunas consideraciones:

- 2SLS ayuda a comprender mejor la intuición de las variables instrumentales.
- 2SLS utilizó solo los valores ajustados de los regresores endógenos para la estimación.
- Los valores ajustados se basaron en todas las variables utilizadas en el modelo, incluido el instrumento excluible.
- Como los instrumentos son exógenos en el modelo estructural, entonces que los propios valores ajustados también se han vuelto exógenos. Esto nos permite estimar el efecto causal del tratamiento.

Dado que el uso de variables instrumentales entregan estimaciones asintóticamente insesgadas, este método se ha convertido en el estándar para inferenca causal en estudios con incumplimiento en ciencias sociales.

Incumplimiento

La idea de cumplimiento (*compliance*) refiere a si el tratamiento aplicado coincide con la asignación de este. Full compliance: 100% de los asignados al grupo de tratamiento reciben el tratamiento.

Incumplimiento (noncompliance) ocurre cuando el tratamiento aplicado y la asignación de este no coinciden. En otras palabras, algunos sujetos asignados al grupo de tratamiento en realidad no reciben el tratamiento. Ningún sujeto en el grupo de control es tratado.

- **Compliers**: Los individuos que toman el tratamiento son efectivamente los grupos asignados a este. La subpoblación cuyo estado de tratamiento se ve afectado por el instrumento en la dirección correcta.
- **Never Takers**: Independiente de su condición de asignación, los individuos nunca tomaran el tratamiento. La subpoblación de unidades que nunca toman el tratamiento independientemente del valor del instrumento.
- **Always Takers**: Independiente de su condición de asignación, los individuos siempre van a tomar el tratamiento. La subpoblación de unidades que siempre toman el tratamiento independientemente del valor del instrumento.
- **Defiers**: la subpoblación cuyo estado de tratamiento se ve afectado por el instrumento en la dirección incorrecta.

¿Cómo es la estimación cuando tenemos efectos heterogéneos del tratamiento?

δ_{IV} Es un efecto promedio del tratamiento para los cumplidores (CACE).

$$\delta_{IV} = \frac{E[Y_i|Z=1] - E[Y_i|Z=0]}{E[D_i|Z=1] - E[D_i|Z=0]}$$

$$\delta_{IV} = E[Y_{1i} - Y_{0i}|D_i(1) - D_i(0) = 1]$$

Supuestos relevantes a la hora de usar variables instrumentales

- Supuesto de valor de tratamiento unitario estable (SUTVA) que establece que los resultados potenciales para cada persona i no están relacionados con el estado de tratamiento de otros individuos. Con esto podemos estimar un efecto de tratamiento promedio.
- Independencia: "asignación tan buena como aleatoria". La variable instrumental es independiente de los posibles resultados y posibles asignaciones de tratamiento.
- Restricción de exclusión: cualquier efecto de Z sobre Y debe ser a través del efecto de Z sobre D
- En la primera etapa Z se correlacionará con la variable endógena
- Supuesto de monotonicidad: la variable instrumental (débilmente) opera en la misma dirección en todas las unidades individuales. En otras palabras, si bien el instrumento puede no tener efecto en algunas personas, todos los afectados se ven afectados en la misma dirección (es decir, positiva o negativamente, pero no ambas).