Two Stage Least Squares

Definición 1. Endogeneidad: Una variable explicativa X se considera endógena cuando está correlacionada con el término de error u, es decir: $\mathbb{E}[Xu] \neq 0$.

Definición 2. Exogeneidad: Una variable X es exógena si es independiente del término de error u, formalmente: $\mathbb{E}[Xu] = 0$.

Definición 3. Variable Instrumental: Un instrumento Z es válido si cumple:
i) Relevancia: $\mathbb{E}[ZX] \neq 0$.
ii) Exogeneidad: $\mathbb{E}[Zu] = 0$.

Definición 4. Consistencia: Un estimador $\hat{\theta}$ es consistente si: $plim(\hat{\theta}) = \theta$ cuando $n \to \infty$.[1]

Definición 5. Proxy: Variable observable que aproxima una característica no directamente medible en el modelo.

Supongamos que estamos interesados en estimar el efecto causal de una variable X sobre una variable Y. El problema es que X es endógena, es decir, está correlacionada con el término de error u en la ecuación:

$$Y = \beta_0 + \beta_1 X + u \tag{1}$$

Esto viola el supuesto de exogeneidad necesario para que el estimador de Mínimos Cuadrados Ordinarios (MCO) sea consistente y nos dé estimaciones insesgadas de β_1.

Demostración

Partimos de la ecuación:

$$Y = \beta_0 + \beta_1 X + u \tag{2}$$

Donde existe endogeneidad, es decir:

$$\mathbb{E}[Xu] \neq 0 \tag{3}$$

El estimador MCO viene dado por:

$$\hat{\beta}_{1,MCO} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{4}$$

Sustituyendo la ecuación original:

$$\hat{\beta}_{1,MCO} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\beta_0 + \beta_1 X_i + u_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{5}$$

Tomando el límite en probabilidad:

$$plim(\hat{\beta}_{1,MCO}) = \beta_1 + \frac{plim(\frac{1}{n}\sum_{i=1}^{n} X_i u_i)}{plim(\frac{1}{n}\sum_{i=1}^{n} X_i^2)} \tag{6}$$

[1] La notación *plim* hace referencia al límite en probabilidad.

$$plim(\hat{\beta}_{1,MCO}) = \beta_1 + \frac{\sigma_{Xu}}{\sigma_X^2} \tag{7}$$

Donde σ_{Xu} es la covarianza entre X y u, y σ_X^2 es la varianza de X. Como $\sigma_{Xu} \neq 0$ por la endogeneidad:

$$plim(\hat{\beta}_{1,MCO}) \neq \beta_1 \tag{8}$$

Por lo tanto, el estimador MCO no converge al verdadero valor del parámetro cuando el tamaño de la muestra tiende a infinito, lo que demuestra que no es consistente.

La idea central de 2SLS es encontrar una variable instrumental Z que cumpla dos condiciones:

1. Relevancia: Z está correlacionada con la variable endógena X. 2. Exogeneidad: Z no está correlacionada con el término de error u.

Si tenemos un instrumento válido Z, podemos usarlo para obtener una estimación consistente de β_1 en dos etapas:

Primera etapa: Regresamos X sobre Z y obtenemos los valores predichos $\hat{X}$:

$$X = \pi_0 + \pi_1 Z + v \tag{9}$$

Aquí, π_1 captura la correlación entre Z y X, y v es un nuevo término de error.

Segunda etapa: Regresamos Y sobre los valores predichos $\hat{X}$ de la primera etapa:

$$Y = \beta_0 + \beta_1 \hat{X} + u \tag{10}$$

Intuitivamente, al usar $\hat{X}$ en lugar de X, estamos usando sólo la parte de la variación en X que viene de la variación en Z, que por construcción es exógena. Esto nos permite obtener una estimación consistente de β_1.

Ejemplo

Supongamos que queremos estimar el efecto de la educación (X) sobre los salarios (Y), pero sospechamos que la educación es endógena debido a variables omitidas como la habilidad. Podríamos usar la distancia a la universidad más cercana como un instrumento (Z).

¿Por qué? Es probable que la distancia a la universidad esté correlacionada con la educación (relevancia), ya que las personas que viven más cerca de una universidad pueden tener más probabilidades de asistir. Pero es poco probable que la distancia en sí misma afecte los salarios, excepto a través de su efecto sobre la educación (exogeneidad).

Usando la distancia como instrumento, primero podemos predecir la educación en función de la distancia (primera etapa), y luego usaríamos esos valores predichos de educación para predecir los salarios (segunda etapa). El coeficiente resultante en la educación predicha nos daría una estimación del efecto causal de la educación sobre los salarios.

1 Forma reducida

Recuerda que nuestro modelo estructural es:

$$Y = \beta_0 + \beta_1 X + u \tag{11}$$

Donde X es endógena, es decir, $\text{cov}(X, u) \neq 0$. Para resolver este problema, introducimos una variable instrumental Z que está correlacionada con X pero no con u.

Ahora, podemos escribir la ecuación de la primera etapa, que relaciona la variable endógena X con el instrumento Z:

$$X = \pi_0 + \pi_1 Z + v \tag{12}$$

Aquí, π_0 y π_1 son parámetros desconocidos, y v es un término de error. Esta ecuación descompone X en dos partes:

1. $\pi_0 + \pi_1 Z$, que captura la parte de X que está relacionada con Z (y por lo tanto es exógena).

2. v, que captura todo lo demás de X, incluyendo la parte que está correlacionada con u.

Podemos pensar en π_1 como la fuerza de la relación entre Z y X. Si $\pi_1 = 0$, entonces Z no está correlacionada con X, y no sería un buen instrumento. Este es el requisito de relevancia para un instrumento válido.

Ahora, sustituyamos la ecuación de la primera etapa en nuestro modelo estructural:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + u \\ &= \beta_0 + \beta_1(\pi_0 + \pi_1 Z + v) + u \\ &= (\beta_0 + \beta_1 \pi_0) + (\beta_1 \pi_1) Z + (\beta_1 v + u) \end{aligned} \tag{13}$$

Podemos simplificar esta ecuación definiendo nuevos términos:

Ahora, nuestra ecuación se ve así:
$Y = \alpha_0 + \alpha_1 Z + e$

Esta es la forma reducida de nuestro modelo. Observa que hemos eliminado la variable endógena X por completo. En su lugar, tenemos el instrumento Z, que es exógeno por construcción.

Para que esto funcione, necesitamos que $\text{cov}(Z, u) = 0$, lo que significa que Z no está correlacionada con el término de error original u. Este es el requisito de exogeneidad para un instrumento válido.

Si se cumplen las condiciones de relevancia y exogeneidad, podemos estimar consistentemente α_1 usando MCO en la forma reducida. Luego, dado que $\alpha_1 = \beta_1 \pi_1$, podemos recuperar nuestro parámetro de interés β_1 dividiendo α_1 por π_1 (que obtenemos de la primera etapa).

Añadiendo más regresores exógenos

Consideremos un modelo con un regresor endógeno y_2 y varios regresores exógenos $z_1, \ldots, z_{k-1}$:

$$\begin{gathered} y_1 = \beta_0 + \beta_k y_2 + \beta_1 z_1 + \ldots + \beta_{k-1} z_{k-1} + u \\ \text{cov}(z_j, u) = 0, \quad j = 1, \ldots, k-1 \\ \text{cov}(y_2, u) \neq 0 \end{gathered} \tag{14}$$

En este contexto, la ecuación de forma reducida de y_2 es:

$$y_2 = \pi_0 + \pi_1 z_1 + \ldots + \pi_k z_k + v \tag{15}$$

Para que z_k sea un buen instrumento para y_2, debe cumplir dos condiciones:

1. Exogeneidad: $\text{cov}(z_k, u) = 0$. Esto significa que z_k no está correlacionada con el término de error u en la ecuación estructural.

2. Relevancia: $\pi_k \neq 0$. Esto implica que z_k está correlacionada con la variable endógena y_2 después de controlar por los otros regresores exógenos.

Intuitivamente, un buen instrumento debe ser una fuente de variación exógena en y_2 que no esté relacionada con los factores no observados que afectan a y_1.

Añadiendo más instrumentos

Supongamos ahora que tenemos dos potenciales instrumentos para y_2: z_k y z_{k+1}. Ambos son exógenos, es decir, $\text{cov}(z_k, u) = \text{cov}(z_{k+1}, u) = 0$. Además, al menos uno de ellos es relevante: en la ecuación de forma reducida $y_2 = \pi_0 + \pi_1 z_1 + \ldots + \pi_k z_k + \pi_{k+1} z_{k+1} + v$, tenemos que $\pi_k \neq 0$ o $\pi_{k+1} \neq 0$, o ambos.

¿Qué instrumento debemos usar para y_2? Para z_k y z_{k+1} podemos calcular un estimador de Variables Instrumentales (IV). Cada uno de estos estimadores explota información importante, pero también omite alguna información. Por ejemplo, al usar z_k, no aprovechamos el hecho de que $\text{cov}(z_{k+1}, u) = 0$.

Además, cualquier combinación lineal $z = \alpha_1 z_k + \alpha_2 z_{k+1}$ también es un buen instrumento para y_2:

1. Es exógena: $\text{cov}(z, u) = \alpha_1 \text{cov}(z_k, u) + \alpha_2 \text{cov}(z_{k+1}, u) = 0$.
2. Es relevante: $\text{cov}(z, y_2) \neq 0$ si $\pi_k \neq 0$ o $\pi_{k+1} \neq 0$.

El mejor instrumento es la combinación lineal que está más altamente correlacionada con y_2. Esta combinación lineal óptima está dada por la proyección lineal de y_2 sobre todos los instrumentos exógenos:

$$y_2^* = \pi_0 + \pi_1 z_1 + \ldots + \pi_k z_k + \pi_{k+1} z_{k+1} \tag{16}$$

Intuitivamente, y_2^* captura toda la información relevante en $z_1, \ldots, z_{k+1}$ para predecir y_2, mientras que el residuo $v = y_2 - y_2^*$ captura la variación en y_2 que no está relacionada con los instrumentos exógenos.

En resumen, al añadir más regresores exógenos y más instrumentos, tenemos más información para identificar y estimar el efecto causal de y_2 sobre y_1. La clave es encontrar instrumentos que sean exógenos y relevantes. Cuando tenemos múltiples instrumentos válidos, el mejor enfoque es usar una combinación lineal óptima de todos ellos, que se puede obtener mediante la proyección lineal de la variable endógena sobre todos los instrumentos exógenos.

2 Multicolinealidad

La multicolinealidad en el contexto de Mínimos Cuadrados en Dos Etapas (2SLS) se refiere a la situación en la que los regresores exógenos están altamente correlacionados entre sí. Esto puede llevar a estimaciones imprecisas de los coeficientes. La varianza asintótica del estimador 2SLS de β_k puede ser aproximada por:

$$\frac{\sigma^2}{S\hat{S}T_{\hat{y}}(1 - R_2^2)} \tag{17}$$

donde R_2^2 es el R^2 de regresar la variable endógena y_2 sobre todos los regresores exógenos. Esta fórmula indica que 2SLS es menos preciso que Mínimos Cuadrados Ordinarios (OLS) por dos razones:

1. Los valores predichos de la variable endógena en la primera etapa, $\hat{y}_2$, tienen menos variación muestral que la variable endógena original y_2.

2. $\hat{y}_2$ tiene más correlación con todos los regresores exógenos que y_2, ya que $\hat{y}_2$ se construye específicamente para estar correlacionado con los instrumentos, que son regresores exógenos.

Respecto a los errores en las variables, considere la siguiente ecuación de ahorro:

$$sav = \beta_0 + \beta inc^* + u \tag{18}$$

donde inc^* es el ingreso verdadero. En la práctica, a menudo se observa inc en lugar de inc^*, donde $inc = inc^* + e$, siendo e el error de medición. Al usar inc en lugar de inc^* en la regresión, se estima:

$$sav = \beta_0 + \beta inc + (u - \beta e) \tag{19}$$

Si el error de medición no está correlacionado con el ingreso verdadero, entonces:

$$cov(inc, e) = var(e) \neq 0 \Rightarrow cov(inc, u - \beta e) = -\beta var(e) \tag{20}$$

Esto implica que *inc* está correlacionado con el término de error $(u - \beta e)$, violando el supuesto clave de exogeneidad y llevando a estimaciones inconsistentes de β si se usa OLS. El límite de probabilidad (plim) del estimador OLS será:

$$plim(\hat{\beta}_{OLS}) = \beta \left(1 - \frac{var(e)}{var(inc)}\right) < \beta \tag{21}$$

Este fenómeno se conoce como sesgo de atenuación.

2SLS puede resolver este problema si se encuentra una variable que esté correlacionada con el ingreso verdadero pero no con el error de medición en el ingreso observado. Esta variable servirá como un instrumento válido. En la práctica, cuando se tienen múltiples proxies para el ingreso, estas pueden ser usadas como instrumentos para computar el estimador 2SLS. Al hacer esto, se aísla la variación en *inc* que está relacionada con inc^* pero no con e, permitiendo obtener estimaciones consistentes de β a pesar del error de medición.

En resumen, 2SLS es una herramienta poderosa para lidiar con problemas de endogeneidad y errores de medición, siempre y cuando se cuente con instrumentos válidos. Sin embargo, la multicolinealidad entre los regresores exógenos puede llevar a estimaciones menos precisas en comparación con OLS.

Cuando se estima la relación entre el ahorro (sav) y el ingreso (inc), un problema común es que el ingreso verdadero (inc*) a menudo no es directamente observable. En cambio, típicamente medimos el ingreso observado (inc), que es el ingreso verdadero más algún error de medición (e):

$$inc = inc* + e \tag{22}$$

Si usamos el ingreso observado en lugar del ingreso verdadero en nuestra regresión, en realidad estamos estimando:

$$sav = \beta_0 + \beta inc + (u - \beta e) \tag{23}$$

Si el error de medición no está correlacionado con el ingreso verdadero, entonces:

$$cov(inc, e) = var(e) \neq 0 \Rightarrow cov(inc, u - \beta e) = -\beta var(e) \tag{24}$$

Esto implica que el ingreso observado (inc) está correlacionado con el término de error compuesto $(u - \beta e)$, violando el supuesto clave de exogeneidad. Como resultado, el estimador de Mínimos Cuadrados Ordinarios (OLS) será inconsistente, con un límite de probabilidad (plim) de:

$$plim(\hat{\beta}_{OLS}) = \beta \left(1 - \frac{var(e)}{var(inc)}\right) < \beta \tag{25}$$

Este fenómeno se conoce como sesgo de atenuación, ya que el error de medición en el ingreso observado "atenúa" nuestra estimación de β hacia cero.

La solución a este problema es encontrar una variable instrumental que esté correlacionada con el ingreso verdadero pero no con el error de medición en el ingreso observado. Cualquier variable que cumpla estas condiciones será un instrumento válido.

En la práctica, a menudo tenemos varios *proxies* para el ingreso, ninguno de los cuales es perfecto. En este caso, podemos usar estos *proxies* como instrumentos y computar el estimador de Mínimos Cuadrados en Dos Etapas (2SLS).

La intuición detrás de este enfoque es que, al usar los *proxies* como instrumentos, estamos aislando la variación en el ingreso observado que está relacionada con el ingreso verdadero, pero no con el error de medición. Esto nos permite obtener estimaciones consistentes de β a pesar del error de medición.

Matemáticamente, la primera etapa de 2SLS regresará el ingreso observado sobre los proxies:

$$inc = \pi_0 + \pi_1 proxy_1 + \pi_2 proxy_2 + ... + v \tag{26}$$

Aquí, $\pi_1, \pi_2, ...$ capturan la relación entre los proxies y el ingreso verdadero, mientras que v captura la variación en el ingreso observado que no está relacionada con el ingreso verdadero (es decir, el error de medición).

En la segunda etapa, regresamos el ahorro sobre los valores predichos del ingreso de la primera etapa:

$$sav = \beta_0 + \beta \widehat{inc} + u \tag{27}$$

Donde $\widehat{inc}$ son los valores predichos del ingreso de la primera etapa.

Intuitivamente, $\widehat{inc}$ captura la parte del ingreso observado que está relacionada con el ingreso verdadero, pero no con el error de medición. Al usar $\widehat{inc}$ en lugar de inc en nuestra regresión, obtenemos estimaciones consistentes de β.

En resumen, los errores de medición pueden llevar a estimaciones sesgadas e inconsistentes cuando se usa OLS. Sin embargo, si tenemos variables instrumentales que están correlacionadas con la variable verdadera pero no con el error de medición, podemos usar 2SLS para obtener estimaciones consistentes a pesar de los errores de medición.