Nota 6
Variables Instrumentales

1 Introducción

Como describimos en la *Nota 2*, uno de los principales problemas de validez interna para los estimadores de MCO es el sesgo por variables omitidas. Esto provoca que los resultados de las estimaciones de MCO no puedan ser interpretadas de manera causal en la mayoría de las ocasiones. El sesgo por varibles omitidas se genera porque: (i) la variable de interés (X_1) esta correlacionada con alguna variable no observada o no incluida dentro de la estimación, y (ii) porque dicha variable no incluida está correlacionada con la variable dependiente. La primera condición implica que uno de nuestros supuestos utilizados para estimar el modelo de MCO ($E(X_{1i} * U_i) = 0$) no sea un supuesto valido, ya que la variable omitida implícitamente forma parte del error de la estimación (U_i).

El método de *variables instrumentales* es una alternativa para estimar el efecto de dicha variable de interés (X_1) sobre la variable dependiente. Intuitivamente, este método consiste en encontrar un instrumento (Z) que juegue el rol de la variable de interés (X_1) sin tener el problema que dicha variable de interés tiene.

2 Planteamiento

Empecemos por recordar el sesgo por variables omitidas. Supongamos que queremos estimar el efecto de la educación sobre los ingresos. Para llevar a cabo esto empezamos por estimar un modelo de mínimos cuadrados ordinarios donde nuestra variable dependiente es el ingreso mensual (Ing_i)[1] y nos interesa ver el efecto de los años de escolaridad ($Educ_i$):

$$Ing_i = \alpha_0 + \alpha_1 Educ_i + U_i \tag{1}$$

Un problema con esta estimación es que los años de educación pueden estar sesgados por

[1] Podriamos usar el logaritmo del ingreso también, pero para simplificar la exposición utilizamos solo ingreso.

omitir en esta estimación variables como educación de los padres, habilidad del individuo, mejores redes sociales, etc. Tomemos el ejemplo de habilidad. Si agregamos esta variable a nuestro modelo tendríamos:

$$Ing_i = \beta_0 + \beta_1 Educ_i + \beta_2 Habil_i + V_i \tag{2}$$

Y nuestro sesgo por variables omitidas estaría descrito por $\beta_2 * \gamma_1$ donde γ_1 tendrá el mismo signo que la correlación entre educación y habilidad. El problema en la estimación de [1] es que las variables omitidas implicitamente se encontraba en el error (U_i) y nuestra variable de interés ($Educ_i$) está correlacionada con ellas. Esto implica que se viola el supuesto de MCO $E(Educ_i * U_i) = 0$.

Una altenativa para estimar de manera consistente α_1 utilizando el modelo [1] consiste en utilizar el método de variables instrumentales. Este método consiste en identificar un instrumento (Z_i). Utilizando este instrumento y nuestro modelo [1] podemos calcular:

$$\begin{aligned} Cov(Ing_i, Z_i) &= Cov(\alpha_0 + \alpha_1 Educ_i + U_i, Z_i) \\ &= \alpha_1 Cov(Educ_i, Z_i) - Cov(U_i, Z_i) \end{aligned} \tag{3}$$

Por lo tanto obtenemos:[2]

$$\alpha_1 = \frac{Cov(Ing_i, Z_i)}{Cov(Educ_i, Z_i)} + \frac{Cov(U_i, Z_i)}{Cov(Educ_i, Z_i)} \tag{4}$$

Para tener un estimador consistente de α_1 se deben cumplir dos condiciones, que son los supuestos fundamentales de los modelos de variables instrumentales:

1. **Relevancia** ($Cov(Educ_i, Z_i) \neq 0$). Intuitivamente, este supuesto implica que dado que queremos utilizar al instrumento (Z_i) para representar a nuestra variable de interés ($Educ_i$), dichas variables deben estar fuertemente correlacionadas. Una manera de evaluar si está condicións se satisface es llevar a cabo una regresión de la variable de interés ($Educ_i$) contra el instrumento (Z_i):

$$Educ_i = \eta_0 + \eta_1 Z_i + U_i \tag{5}$$

[2]Nótese que la generalización de este modelo consiste en sustituir Ing_i con Y_i y $Educ_i$ con X_{1i}

En la literatura se sugiere que para tener un buen instrumento, el estadístico F que resulte de llevar a cabo la siguiente prueba de hipótesis[3] debe ser mayor a 10:
$H_0 : \eta_1 = 0$
$H_1 : \eta_1 \neq 0$

2. **Exogeneidad o restricciones de exclusión** ($Cov(U_i, Z_i) = 0$). Exogeneidad implica que nuestro instrumento no está correlacionado con el error (U_i), que es lo que causaba el problema de sesgo por variables omitidas. Cabe recordar que U_i incluye todas aquellas variables que no incluimos en el modelo [1], tales como educación de los padres, habilidad, redes sociales, etc. Generalmente, el supuesto de exogeneidad es el más difícil de justificar y en términos del modelo no se puede evaluar directamente, a menos que se cuenten con más instrumentos que variables endógenas (es decir, aquellas, cuyo estimador está sesgado). Implicitamente, este supuesto además implica que el instrumento no debe de afectar directamente a la variable dependiente (Ing_i). El único efecto que identificará el modelo es el efecto indirecto de la variable de interés que estamos instrumentando ($Educ_i$).[4]

Angrist y Krueger (1991) sugieren como posible instrumento en este caso la fecha de nacimiento de la persona. Utilizando la fecha de nacimiento identificaron a aquellos que nacen en diferentes trimestres del año. La motivación para su instrumento se basa en la idea de, que de acuerdo a las leyes vigentes en E.U., una persona es requerida a estudiar hasta el momento en que cumple 16 años. Sin embargo, las generaciones escolares conjuntan a los niños nacidos entre Agosto y Julio del siguiente año. Por lo tanto, una persona que cumple años en enero ya podrá trabajar y aun no habrá completado el año escolar, mientras que una persona que cumple años en julio ya habrá terminado el grado escolar en el momento en que puede empezar a trabajar. Por lo tanto, es muy posible que la fehca de nacimiento influya sobre los años de escolaridad completados de un individuo. Esto puede ser verificado como describimos en el inciso de relevancia. El mayor reto consiste en argumentar

[3] Recuerden que cuando se evalúa la hipótesis de un solo coeficiente, el cuadrado del estadístico t es igual que el estadístico F. Establecemos esta prueba en términos del estadístico F, ya que como veremos más adelante en la Nota, puede ser que tengamos más de un instrumento

[4] De no cumplirse esta condición, no será psible distinguir que tanto de la $Cov(Ing_i, Z_i)$ se debe al efecto directo de Z_i sobre Ing_i y que tanto por el efecto a través de $Educ_i$.

la exogeneidad. Puede argumentarse que la fecha de nacimiento no influye los ingresos del individuo, y que no está relacionado con variables como acceso a transporte, habilidad y redes sociales. El único problema potencial es que los padres reconozcan esto y padres de familia más sofisticados decidan tener hijos de manera tal que nazcan cerca del final del ciclo escolar (pero como pueden darse cuenta es un argumento mas dificil de establecer).

Si asumimos que la fecha de nacimiento es un instrumento satisfactorio, podríamos utilizar dummies de haber nacido en distintos trimetres para estimar el efecto de la edcuación sobre el ingreso. Para ilustrar esto tomaremos solo una dummy y posteriormente veremos cómo ampliarlo a más instrumentos. Sea $Q1_i$ una dummy que indica si el individuo nació en el ultimo triemstre del año. Utilizando este instrumento debemos estimar las siguientes dos ecuaciones. Estas ecuaciones se conocen como *ecuaciones de forma reducida* cuando solo incluyen variables exógenas como regresores:

$$\begin{aligned} Ing_i &= \gamma_0 + \gamma_1 Q1_i + \upsilon_i \\ Educ_i &= \eta_0 + \eta_1 Q1_i + \nu_i \end{aligned} \tag{6}$$

En este caso:

$$\hat{\alpha}_1 = \hat{\gamma}_1 / \hat{\eta}_1 \tag{7}$$

3 Agregar controles

El modelo de variables instrumentales puede incluir otras variables de control. Para llevar a cabo la estimación veamos como desarrollar el caso generalizado. Supongamos que tenemos k variables que queremos incluir en el modelo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + U_i \tag{8}$$

Supongamos que en este caso nos interesa estimar el efecto causal de X_1 sobre Y. Estimar el modelo [8] utilizando MCO genera un estimador sesgado de β_1 por haber sesgo por variables omitidas. Por lo tanto, incluimos un instrumento (Z) para X_1 que cumpla con las condiciones antes descritas. Si estimamos un modelo de forma reducida para estimar X_1 utilizando nuestro instrumento y las demás variables del modelo tendremos:

$$X_{1i} = \phi_0 + \phi_2 X_{2i} + \cdots + \phi_k X_{ki} + \eta Z_i + V_i \tag{9}$$

Sustituyendo [9] en [8] obtenemos:

$$Y_i = \beta_0 + \beta_1[\phi_0 + \phi_2 X_{2i} + \cdots + \phi_k X_{ki} + \eta Z_i + V_i] + \cdots + \beta_k X_{ki} + U_i \tag{10}$$

Reordenando los términos obtenemos:

$$Y_i = \psi_0 + \psi_2 X_{2i} + \cdots + \psi_k X_{ki} + \gamma Z_i + W_i \tag{11}$$

donde:

$$\begin{aligned} \psi_j =& \beta_j + \beta_1 \phi_j \\ \gamma =& \beta_1 \eta \\ W_i =& U_i + \beta_1 V_i \end{aligned}$$

Por lo tanto, para obtener estimadores insesgados utilizando MCO en [11] tendremos las siguientes condiciones de primer órden:

$$\sum_{i=1}^{N} (Y_i - \hat{\psi}_0 - \hat{\psi}_2 X_{2i} - \ldots \hat{\psi}_k X_{ki} - \hat{\gamma} Z_i) = 0 \tag{12}$$

$$\sum_{i=1}^{N} Z_i (Y_i - \hat{\psi}_0 - \hat{\psi}_2 X_{2i} - \ldots \hat{\psi}_k X_{ki} - \hat{\gamma} Z_i) = 0 \tag{13}$$

Y para $j = 2, \ldots, k$:

$$\sum_{i=1}^{N} X_{ji} (Y_i - \hat{\psi}_0 - \hat{\psi}_2 X_{2i} - \ldots \hat{\psi}_k X_{ki} - \hat{\gamma} Z_i) = 0 \tag{14}$$

Por lo tanto, para que los estimadores $\psi_0, \psi_2, \ldots \psi_k$ y γ sean insesgados, tendrá que cumplirse que la covarianza de $X_2, \ldots, X_k, Z$ con el error W del modelo [11] sea cero en cada caso. Esto se cumplirá si la covarianza de $X_2, \ldots, X_k, Z$ con los errores U y V de los modelos [8] y [9] son cero, respectivamente. En el caso de Z, este requisito es el supuesto de exogeneidad. Para el resto de los controles, este supuesto está imponiendo el requisito de exogeneidad. Recordemos que nuestro interés radica en obtener un estimador insesgado de β_1. Si removemos alguno de los controles porque nos preocupa que no cumple con los

supuestos de exogeneidad, el requisito adicional que estamos imponiendo por no incluir dicho control es que el instrumento no deberá estar correlacionado con éste control, ya que el control pasará a formar parte del error U del modelo [8]. Si el control no es relevante para explicar la variable dependiente, es mejor no incluirlo en la estimación.

A partir de estimar los modelo [8] y [9], se puede obtener un estimador de $\beta_0, \beta_1, \ldots, \beta_k$. Dadas las derivaciones previas tenemos que:

$$\beta_1 = \gamma/\eta \tag{15}$$

Por lo tanto, únicamente dividimos el coeficiente que resulta de estimar [11] entre el que resulta de estimar [9].

4 Mínimos Cuadrados en 2 Etapas (Two-Stage Least Squares, 2SLS)

Supongamos ahora que queremos estimar el modelo [8] y que tenemos dos instrumentos $(Z1, Z2)$ que cumplen con los supuestos de relevancia y las restricciones de exclusión. En este caso, podríamos llevar a cabo dos estimaciones de las ecuaciones [9] y [11] para obtener dos valores estimados insesgados de β_1. En el caso del primer estimador utilizaríamos las condiciones de primer órden dadas por [12], [14] y:

$$\sum_{i=1}^{N} Z1_i(Y_i - \hat{\psi}_0 - \hat{\psi}_2 X_{2i} - \ldots \hat{\psi}_k X_{ki} - \hat{\eta} Z1_i) = 0 \tag{16}$$

En el segundo caso utilizaríamos las condiciones de primer órden dadas por [12], [14] y:

$$\sum_{i=1}^{N} Z2_i(Y_i - \hat{\psi}_0 - \hat{\psi}_2 X_{2i} - \ldots \hat{\psi}_k X_{ki} - \hat{\eta} Z2_i) = 0 \tag{17}$$

Sin embargo, existe una manera de agrupar la información de manera eficiente para producir un solo estimador. Para esto es útil el método de 2SLS. Este método lleva a cabo la estimación en dos etapas, donde la primera etapa combina los instrumentos de manera eficiente y la segunda utiliza el supuesto de exogeneidad para derivar coeficientes insesgados del modelo [8]. (Cabe señalar que el métdo de 2SLS puede ser aplicado también en el

caso que tengamos un solo instrumento y un estimador y resultará en el mismo coeficiente estimado que el derivado utilizando el método antes descrito)

4.1 Primera Etapa (First Stage)

La primera etapa esta relacionada con el supuesto de relevancia. Esta etapa consiste en utilizar los instrumentos para predecir el valor de la variable de interés (X_1). Este paso es el descrito en la ecuación de forma reducida [9], pero incluyendo todos los instrumentos válidos disponibles. Para llevar a cabo esto utilizamos un modelo de MCO:

$$X_{1i} = \phi_0 + \eta_1 Z1_i + \eta_2 Z2_i + \phi_2 X_{2i} + \cdots + \phi_k X_{ki} + V_i \tag{18}$$

En este caso, para evaluar si los instrumentos son relevantes, calculamos el estadístico F que se relaciona con la siguiente prueba de hipótesis:

$$\begin{aligned} H_0 :& \eta_1 = 0 \\ & \eta_2 = 0 \\ H_1 :& \eta_1 \neq 0 | \eta_2 \neq 0 \end{aligned}$$

Utilizando los resultados de esta estimación podemos predecir el valor de X_{1i} basado únicamente en la información que proporcionan los instrumentos y las variables exógenas:

$$\widehat{X}_{1i} = \hat{\phi}_0 + \hat{\eta}_1 Z1_i + \hat{\eta}_2 Z2_i + \hat{\phi}_2 X_{2i} + \cdots + phi\eta_k X_{ki} \tag{19}$$

4.2 Segunda Etapa (Second Stage)

La segunda etapa consiste en utilizar el supuesto de exogeneidad de los instrumentos para derivar estimadores insesgados de los coeficientes del modelo [8]. Para esto utilizaremos las condiciones de primer orden dadas por:

$$\sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \widehat{X}_{1i} - \hat{\beta}_2 X_{2i} - \ldots \hat{\beta}_k X_{ki}) = 0 \tag{20}$$

$$\sum_{i=1}^{N} \widehat{X}_{1i} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \widehat{X}_{1i} - \hat{\beta}_2 X_{2i} - \ldots \hat{\beta}_k X_{ki}) = 0 \tag{21}$$

Y para $j = 2, \ldots, k$:

$$\sum_{i=1}^{N} X_{ji}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \widehat{X}_{1i} - \hat{\beta}_2 X_{2i} - \ldots \hat{\beta}_k X_{ki}) = 0 \tag{22}$$

Esto resultará en los mismos estimadores que los obtenidos por estimar el siguiente modelo utilizando MCO:

$$Y_i = \beta_0 + \beta_1 \widehat{X}_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + W_i \tag{23}$$

Nótese que dado que $\widehat{X}_{1i}$ es una función de $Z1_i$ y $Z2_i$, no tendremos el problema de sesgo por variables omitidas y los coeficientes que resulten de esta estimación serán insesgados si $Z1_i$ y $Z2_i$ cumplen con los supuestos para ser instrumentos válidos.

5 Inferencia - Errores estándar

Tomando notación matricial, sea el modelo [8]:

$$Y_i = X_i'\beta + U_i \tag{24}$$

donde $X_i = [\ 1 \quad X_{1i} \quad X_{2i} \quad \ldots \quad X_{ki}\ ]'$

Y sea:

$$X_i = Z_i'\Pi + W_i \tag{25}$$

donde $Z_i = [\ 1 \quad Z1_i \quad Z2_i \quad X_{2i} \quad \ldots \quad X_{ki}\ ]'$; la segunda columna de Π es la primera etapa: $\Pi(., 2) = [\ \eta_0 \quad \phi_1 \quad \phi_2 \quad \eta_2 \quad \ldots \quad \eta_k\ ]'$ y el resto de las columnas tienen un coeficiente de 1 en la columna correspondiente a las variables exógenas del modelo original (i.e. $X_2, \ldots, X_k$) dado que éstas variables están en X_i y en Z_i; y $W_i = [\ 0 \quad V_i \quad 0 \quad \ldots \quad 0\ ]'$ donde V_i es el error de la primera etapa.

Las ecuaciones [24] y [25] en términos matriciales se convierten en:

$$Y = X\beta + U \tag{26}$$

donde Y es el vector de variables dependientes que tiene una dimensión de $(n * 1)$; X es la matriz de variables independientes o regresores que tiene una dimensión de $(n * k)$; β es un vector de coeficientes de dimensión $(k * 1)$; y U es un vector de errores del modelo estructural con dimensión $(n * 1)$.

$$X = Z\Pi + W \tag{27}$$

donde Z es una matriz que incluye los instrumentos y variables exógenas de X y tiene una dimensión $(n * L)$ (L es el número de variables exógenas más el número de instrumentos); Π es la matriz descrita antes y tiene dimensión $(L * k)$; y W es una matriz de $(n * k)$ que conjunta a las W_i antes descritas ($W = [\ 0 \quad V \quad 0 \quad \dots \quad 0\ ]'$).

Partiendo de [27] tenemos:

$$\begin{aligned} X &= Z\Pi + W \\ Z'X &= Z'Z\Pi + Z'W \end{aligned} \tag{28}$$

y bajo el supuesto de que $E(Z'W) = E(Z'V) = 0$ por exogeneidad de los instrumentos y las variables exógenas en la primera etapa:

$$\Pi = E(Z'Z)^{-1}E(Z'X) \tag{29}$$

Por lo tanto, el estimador será:

$$\widehat{\Pi} = (Z'Z)^{-1}Z'X \tag{30}$$

Para simplificar la notación subsecuente utilizaremos la matriz de proyección $P_Z = Z(Z'Z)^{-1}Z'$ que nos permite seguir la metodología descrita en el método de *2SLS*. Utilizando el resultado de la primera etapa generamos una matriz X^* que corresponde a la parte de X explicada por los instrumentos y las variables exógenas de X (i.e. $X^* = Z\Pi$). Esta matriz deja intactas las variables exógenas de X y sustituye la variable endógena con el valor predicho por la primera etapa utilizando los instrumentos y las variables exógenas de X. La contraparte muestral de X^* será:

$$\widehat{X} = Z\widehat{\Pi} = P_Z X \tag{31}$$

Utilizando X^* en [26] obtenemos:

$$\begin{aligned} Y &= X\beta + U \\ X^{*\prime}Y &= X^{*\prime}X\beta + X^{*\prime}U \end{aligned} \tag{32}$$

Por lo tanto, si se cumple el supuesto de exogeneidad ($E(X^{*\prime}U) = \Pi' E(Z'U) = 0$):

$$\beta = E(X^{*\prime}X)^{-1}E(X^{*\prime}Y) \tag{33}$$

Y el estimador será:[5]

$$\begin{aligned} \widehat{\beta} &= (\widehat{X}'X)^{-1}\widehat{X}'Y \\ &= (X'P_Z'X)^{-1}\widehat{X}'Y \\ &= (X'P_Z'P_ZX)^{-1}\widehat{X}'Y \\ &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'Y \end{aligned} \tag{34}$$

Con esto hemos demostrado que el estimador de β resulta de hacer una regresión de Y como variable dependiente y $\widehat{X}$ como variables independientes.

La derivación de los errores estándar bajo los supuestos de homocedasticidad y heterocedasticidad sigue los mismos pasos que los descritos en la *Nota 2* tan solo sustituyendo X por $\widehat{X}$. En este caso tendremos los siguientes estimadores y convergencias en probabilidad:

$$\hat{\alpha}_{IV} = \left(\frac{1}{N}\sum_{i=1}^{N}\widehat{X}_i\widehat{X}_i'\right)^{-1} \longrightarrow E(X_i^*X_i^{*\prime})^{-1} = \alpha_{IV} \tag{35}$$

$$\hat{\Sigma}_{IV} = \left(\frac{1}{N}\sum_{i=1}^{N}\widehat{U}_i^2\widehat{X}_i\widehat{X}_i'\right)^{-1} \longrightarrow E(U_i^2X_i^*X_i^{*\prime})^{-1} = \Sigma_{IV} \tag{36}$$

donde $\widehat{U}_i = Y_i - X_i'\hat{\beta}$.

Y de la misma forma que la *Nota 2* en el caso de muestras grandes (teoría sintótica) tendremos convergencia en distribución para el estimador de β:

$$\sqrt{N}(\hat{\beta} - \beta) \longrightarrow N(0, \alpha_{IV}\Sigma_{IV}\alpha_{IV}') \tag{37}$$

[5]Estos pasos asumen que la matriz P_Z es idempotente ($P_Z = P_ZP_Z$) y simétrica ($P_Z = P_Z'$).

6 Problemas de instrumentos débiles

Los principales problemas del método de variables instrumentales (además de lograr encontrar un instrumento que cumpla con los supuestos establecidos) son:

1. **Sesgo.** A partir del resultado mostrado en [4] podemos ver que si el supuesto de exogeneidad falla (i.e. $Cov(U_i, Z_i) \neq 0$) y nuestro instrumento es débil (i.e. $Cov(Educ_i, Z_i)$ es pequeño) el sesgo que resultaría en el estimador podría ser peor que en el caso de MCO.

2. **Errores estándar.** Los instrumentos débiles provocan que los errores estándar estimados del coeficiente sean grandes. Por lo tanto, el intervalo de confianza será amplio y la capacidad de determinar que un coeficiente es significativo será menor. Para una explicación de por qué los errores estándar son aumentan con instrumentos débiles se recomienda consultar (Wooldridge 2002, pp. 101-105).

Por útlimo, una cualidad adicional que comúnmente se otorga al método de variables instrumentales es que evita el sesgo de atenuación causado por errores de medición como los discutidos en la *Nota 2, sección 9.2*.