Solutions Problem Set 4 (*Endogeneidad*)

Tatiana Rosá Alejo Eyzaguirre

PUC

24th November 2021

Outline

Pregunta 1: Retornos a la Educación

Sea y_i el logaritmo del salario (por hora) de un individuo i, y sea x_i su nivel educativo (en años). Queremos saber el efecto causal de x_i en y_i.

Empezamos por asumir el siguiente modelo:

$$y_i = \delta + \alpha x_i + u_i$$
$$x_i = \mu + \beta z_i + \upsilon_i$$
$$u_i = z_i + \eta_i$$

En este sistema z_i es inobservable, *i.i.d.* y con varianza σ_z^2. Los errores υ_i y η_i son *i.i.d.*, incorrelacionados, tienen media cero y varianza σ_υ^2 y σ_η^2 respectivamente. Finalmente, $\mathbb{E}(\upsilon_i z_i) = \mathbb{E}(\eta_i z_i) = 0$.

Pregunta 1.1: Interpretación

Interprete el modelo estructural. ¿Qué interpretación le da a z_i?

Solución: El salario depende en el nivel de educación x_i y otras características más: algunas están correlacionadas con la educación y otras no.

Por ejemplo si z_i –que es inobservable– corresponde a la habilidad, tendremos que esta variable que está en el error u_i, y que a la vez esta correlacionada con el nivel de educación ¿Por qué?

- Una persona i con más habilidades probablemente se educará más años. Tendremos entonces que la variable z_i correlacionará con x_i.
- Una persona i con más habilidades probablemente tendrá también un mayor salario, por razones distintas a la educación extra que ganó por ser más "habiloso". Tendremos entonces que la variable z_i estará en el término de error u_i.

Pregunta 1.2: Proyecciones Lineales y *plim*

Sea $\hat{\alpha}$ el coeficiente de la proyección lineal de y_i sobre x_i. Compute:

$$\operatorname*{plim}_{N\to\infty} \hat{\alpha} - \alpha$$

Piense en $\hat{\alpha}$ como $\frac{Cov(x_i,y_i)}{Var(x_i)}$ ¿Qué signo esperaría para esta diferencia? Discuta.

Solución: Aplicando las fórmulas básicas de varianzas y covarianzas tenemos:

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty} \hat{\alpha} &= [Var(x_i)]^{-1} Cov(y_i, x_i) \\
&= [\beta^2\sigma_z^2 + \sigma_v^2]^{-1}[\beta(\alpha\beta+1)\sigma_z^2 + \alpha\sigma_v^2] \\
&= \alpha + \beta\frac{\sigma_z^2}{\beta^2\sigma_z^2 + \sigma_v^2}
\end{aligned}$$

Entonces tendremos que $\operatorname{plim}_{N\to\infty}(\hat{\alpha} - \alpha)$ tendrá el mismo signo que β. Dado que esperamos que los inobservables (habilidad, etc.) tengan un efecto positivo sobre el nivel educativo (i.e., $\beta > 0$) entonces se espera que el sesgo de $\hat{\alpha}$ sea positivo; es decir, estaríamos sobre estimando el efecto de la educación sobre los salarios.

Pregunta 1.2: Proyecciones Lineales y *plim*

Pregunta 1.3: Posibles Instrumentos

De la lista de variables que se presenta a continuación, discuta cuáles son potenciales instrumentos válidos y por qué.

Solución: Un instrumento es válido cuando es relevante y es ortogonal a u_i.

1. Coeficiente Intelectual (✗): No cumple con la restricción de exclusión, explica salario no solo por el efecto que tiene sobre la educación.
2. Distancia a la Universidad del hogar (✓): Este instrumento ha usado previamente por Card (1989). En teoría no debiera afectar el salario de una persona fuera del efecto que tiene vía la educación.
3. Profesión del Padre (✗): Puede afectar salario por otro canal que no sea nivel educativo (redes).
4. Salario Mensual (✗): Pésimo instrumento.
5. Horas Trabajadas (✗): Lo mismo.
6. Mes de Nacimiento (✓): Instrumento creativo usado por Angrist & Krueger. Sin embargo, resulta no ser muy relevante en la práctica.

Pregunta 1.4: Mínimos Cuadrados en 2 Etapas

Utilizando un solo instrumento, explique como estimar consistentemente α. Demuestre la consistencia del estimador propuesto.

Solución: Usamos el bien conocido β_{2SLS} y llamando el instrumento utilizado como W tenemos que:

$$\beta_{2SLS} = (\hat{X}'\hat{X})^{-1}(\hat{X}'Y)$$

Donde,

$$\hat{X} = \hat{\gamma}W$$

y el $\hat{\gamma}$ se obtiene de la siguiente estimación (Primera Etapa),

$$x_i = \gamma w_i + \varepsilon_i$$

Note que en este caso estamos asumiendo que no tenemos variables exógenas en la ecuación estructural (extras) sino tendríamos que agregarlas a la estimación de la primera etapa y como además usamos un solo instrumento tenemos que w_i es un *singleton*.

Pregunta 2: Endogeneidad y Error de Medición

Un investigador que también esta intentando estimar el efecto causal de la educación sobre los salarios, utilizando un instrumento válido encuentra:

$$\tilde{\alpha} = 1.2\hat{\alpha}$$

Para intentar explicar este resutlado, consideremos este segundo modelo aumentad

$$\begin{aligned} y_i &= \delta + \alpha x_i + u_i \\ x_i &= \mu + \beta z_i + \upsilon_i \\ u_i &= z_i + \eta_i \\ \tilde{y_i} &= y_i + \epsilon_i \\ \tilde{x_i} &= x_i + \nu_i \end{aligned}$$

Donde $\tilde{y_i}$ y $\tilde{x_i}$ son las **únicas variables observables del modelo**.

Además de los supuestos del ejercicio anterior, asumimos que ϵ_i y ν_i son *i.i.d.*, tienen media cero, están incorrelacionados y que $\mathbb{E}(\epsilon_i z_i) = \mathbb{E}(\epsilon_i \upsilon_i) = \mathbb{E}(\nu_i z_i) = \mathbb{E}(\nu_i \upsilon_i) = \mathbb{E}(\nu_i \eta_i) = 0$. Asumimos además que ν_i tiene varianza σ_ν^2.

Pregunta 2.1: Interpretación Modelo

Interprete este nuevo modelo. ¿Cuál es la diferencia con el modelo presentado en el ejercicio 1?

Este modelo tiene la misma interpretación que el modelo del acápite anterior, solo que el salario (y_i) y la educación (x_i) están medidos con un cierto error, ϵ_i y ν_i respectivamente.

Cabe destacar que no es raro tener problemas de medición en estas variables. Es común que en encuestas de hogares el nivel de educación sea reportado con error. El ingreso para personas que tienen más de una fuente por ingreso puede ser también no muy claro, por lo que es probable que si este es autoreportado también sufra de error de medición.

Pregunta 2.2: Examinando Consistencia

Sea $\hat{\alpha}$ el coeficiente de la proyección lineal de $\tilde{y}_i$ sobre $\tilde{x}_i$. Compute:

$$\operatorname*{plim}_{N\to\infty}(\hat{\alpha}-\alpha)$$

Solución: Siguiendo lo realizado en la sección 1.2:

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty}\hat{\alpha} &= [Var(\tilde{x}_i)]^{-1}Cov(\tilde{y}_i,\tilde{x}_i) \\
&= (Var(x_i)+\sigma_\nu^2)^{-1}Cov(y_i,x_i) \\
&= (\beta^2\sigma_z^2+\sigma_\upsilon^2+\sigma_\nu^2)^{-1}(\beta(\alpha\beta+1)\sigma_z^2+\alpha\sigma_\upsilon^2) \\
&= \alpha+\frac{\beta\sigma_z^2-\alpha\sigma_\nu^2}{\beta^2\sigma_z^2+\sigma_\upsilon^2+\sigma_\nu^2}
\end{aligned}$$

Por lo tanto tenemos que:

$$\operatorname*{plim}_{N\to\infty}(\hat{\alpha}-\alpha)=\frac{\beta\sigma_z^2-\alpha\sigma_\nu^2}{\beta^2\sigma_z^2+\sigma_\upsilon^2+\sigma_\nu^2}$$

Pregunta 2.2: Examinando Consistencia

Pregunta 2.3: MC2E con Error de Medición

En este nuevo modelo, ¿estimar por dos MC2E, nos permite estimar consistentemente α? Comente que condiciones adicionales deberían cumplirse (si existe alguna) respecto al caso del ejercicio 1 (donde tenemos solamente endogeneidad por omisión de variables relevantes).

Solución: En el caso en que el instrumento a utilizar no esté correlacionado con los errores ϵ_i y ν_i, mínimos cuadrados en 2 etapas arrojará estimaciones consistentes para α. Esto por dos razones:

1. El instrumento no estaría correlacionado con el residuo de la regresión de $\tilde{y}_i$ sobre $\tilde{x}_i$ (exogeneidad), pese al ruido en la medición de ambas variables.
2. El instrumento estaría correlacionado a la vez con $\tilde{x}_i$ o mejor dicho con x_i, dado que no correlaciona con ν_i (relevancia).

Pregunta 2.4: Comente

Se cree que en los datos utilizados por el investigador para este estimación el error de medida de los años de educación representa el 10% de la varianza de esta variable. ¿Cómo relaciona usted esto con el sesgo encontrado por el investigador en el estimador por MCO de α ($\hat{\alpha}$). Discuta.

Solución: El sesgo de atenuación en este caso sería:

$$\frac{\text{Varianza Error } \nu}{\text{Varianza de } \tilde{x}} = \frac{\sigma_\nu^2}{\beta^2\sigma_z^2 + \sigma_\upsilon^2 + \sigma_\nu^2} \approx 10\%$$

Pero en el caso en que creamos que la estrategia de IV arroja estimaciones consistentes y que se cumplen las condiciones mencionadas en el acápite anterior, el verdadero sesgo de OLS sería de -20% ($\tilde{\alpha}_{IV} = 1.2\hat{\alpha}_{OLS}$). Sin embargo, el sesgo de atenuación nos sugiere que las estimaciones de OLS debieran ser solo 10% menores que las verdaderas (i.e., las de IV).

Por esta razón, el error de medición detectado no puede ser la única fuente de endogeneidad (y sesgo) en la estimación de OLS. Deben haber otras explicaciones, como variables omitidas, que estén haciendo que OLS arroje estimaciones más pequeñas que las reales.

Gracias!

jeeyzaguirre@uc.cl