Econometría 1 - Actividad 7

Análisis Empírico: Card (1993, 1995)

Ejercicio 3: Retornos a la Educación con Variables Instrumentales

División de Economía - CDE
Dr. Francisco Cabrera

1 Contexto: El Problema de Endogeneidad en Retornos a la Educación

1.1 La Ecuación de Mincer

El modelo estándar para estimar retornos a la educación es la **ecuación de Mincer**:

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + u_i \tag{1}$$

donde:

- $\log(\text{wage}_i)$ = logaritmo natural del salario por hora del individuo i
- educ_i = años de educación del individuo i
- exper_i = años de experiencia laboral potencial
- u_i = término de error (incluye habilidad, motivación, etc.)

Interpretación de β_1:

El coeficiente β_1 representa el **retorno porcentual** de un año adicional de educación:

$$\frac{\partial \log(\text{wage})}{\partial \text{educ}} = \beta_1 \quad \Rightarrow \quad \frac{\Delta \text{wage}}{\text{wage}} \approx \beta_1 \tag{2}$$

Por ejemplo, si $\beta_1 = 0.10$, un año adicional de educación aumenta el salario en aproximadamente 10%.

1.2 El Sesgo por Habilidad No Observada

El problema

La educación (educ_i) es probablemente **endógena** porque está correlacionada con el término de error:

$$\text{Cov}(\text{educ}_i, u_i) \neq 0 \tag{3}$$

Razón principal: Habilidad no observada ($\text{ability}_i \in u_i$)

$$\begin{aligned} \text{educ}_i &= f(\text{ability}_i, \text{familia}_i, \text{motivación}_i, \ldots) \\ \text{wage}_i &= g(\text{ability}_i, \text{familia}_i, \text{motivación}_i, \ldots, \text{educ}_i) \end{aligned} \tag{4}$$

Las personas más hábiles:

- Obtienen **más educación** (por elección y oportunidad)
- Ganan **salarios más altos** (incluso sin considerar la educación)

Dirección del sesgo

El estimador de MCO tiene el siguiente límite en probabilidad:

$$\text{plim}\, \hat{\beta}_1^{\text{OLS}} = \beta_1 + \underbrace{\frac{\text{Cov}(\text{educ}, u)}{\text{Var}(\text{educ})}}_{\text{Sesgo}} \tag{5}$$

Dado que típicamente $\text{Cov}(\text{educ}, u) > 0$ (habilidad aumenta tanto educación como salario):

$$\boxed{\text{plim}\, \hat{\beta}_1^{\text{OLS}} > \beta_1} \tag{6}$$

Conclusión: MCO **SOBREESTIMA** el verdadero retorno causal de la educación.

2 Ejercicio 3a: Evaluación de nearc4 como Variable Instrumental

2.1 El Instrumento Propuesto

Card (1993, 1995) propone usar:

$$\text{nearc4}_i = \begin{cases} 1 & \text{si creció cerca de una universidad de 4 años} \\ 0 & \text{si no creció cerca de una universidad} \end{cases} \tag{7}$$

2.2 Condiciones para un Instrumento Válido

Para que $z = \text{nearc4}$ sea un instrumento válido, debe cumplir:

Condición 1: Relevancia (Testeable)

$$\text{Cov}(z_i, \text{educ}_i) \neq 0 \tag{8}$$

El instrumento debe estar correlacionado con la variable endógena.

Condición 2: Exogeneidad / Restricción de Exclusión (NO testeable)

$$\text{Cov}(z_i, u_i) = 0 \tag{9}$$

El instrumento NO debe estar correlacionado con el error, excepto a través de su efecto sobre la variable endógena.

2.3 Prueba de Relevancia: Regresión de Primera Etapa

Modelo de primera etapa

$$\text{educ}_i = \pi_0 + \pi_1 \text{nearc4}_i + \pi_2 \text{exper}_i + \pi_3 \text{exper}_i^2 + \pi_4 \text{black}_i + \ldots + v_i \tag{10}$$

Hipótesis a probar

$$\begin{aligned} H_0 &: \pi_1 = 0 \quad \text{(el instrumento es irrelevante)} \\ H_1 &: \pi_1 \neq 0 \quad \text{(el instrumento es relevante)} \end{aligned} \tag{11}$$

Criterios de relevancia

Regla práctica (Stock y Yogo, 2005):
Un instrumento se considera **FUERTE** (no débil) si:

$$F_{\text{first stage}} > 10 \tag{12}$$

donde F es el estadístico F de la prueba de significancia de π_1 (o de todos los instrumentos si hay múltiples).

Consecuencia de instrumentos débiles:

- Sesgo del estimador IV hacia el estimador MCO
- Inferencia no válida (intervalos de confianza incorrectos)
- Pérdida de precisión (errores estándar muy grandes)

Interpretación económica de la relevancia

Si $\pi_1 > 0$, significa que vivir cerca de una universidad **aumenta** los años de educación.

Mecanismos económicos:

1. **Reducción de costos:**
   - Menores costos de transporte
   - Posibilidad de vivir con los padres
   - Menores costos de búsqueda de información
2. **Reducción de barreras psicológicas:**

- Mayor familiaridad con el ambiente universitario
- Modelos a seguir (profesores, estudiantes)
- Reducción de incertidumbre sobre beneficios

3. **Flexibilidad de restricciones de crédito:**
   - Menores costos totales facilitan el financiamiento
   - Posibilidad de trabajar part-time localmente

2.4 Evaluación de Exogeneidad: Restricción de Exclusión

La restricción de exclusión formal

La restricción de exclusión requiere que:

$$\text{nearc4}_i \text{ afecta } \log(\text{wage}_i) \text{ SOLO a través de } \text{educ}_i \tag{13}$$

Formalmente:

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \text{educ}_i + \gamma \text{nearc4}_i + \text{otros controles} + u_i \tag{14}$$

La restricción de exclusión requiere:

$$\gamma = 0 \tag{15}$$

Amenazas a la exogeneidad

Posibles violaciones de la restricción de exclusión:

1. **Calidad de la región:**

$$\text{nearc4}_i \rightarrow \text{prosperidad regional} \rightarrow \log(\text{wage}_i) \tag{16}$$

   Áreas con universidades pueden ser más prósperas (más empleos, mejor infraestructura).

2. **Selección residencial:**

$$\text{familia educada} \rightarrow \text{nearc4}_i \quad \text{y} \quad \text{familia educada} \rightarrow \text{habilidad}_i \rightarrow \log(\text{wage}_i) \tag{17}$$

   Familias más educadas eligen vivir cerca de universidades Y transmiten ventajas a sus hijos.

3. **Efectos de red:**

$$\text{nearc4}_i \rightarrow \text{contactos/redes} \rightarrow \log(\text{wage}_i) \tag{18}$$

Vivir cerca de universidad da acceso a redes profesionales que aumentan salarios.

4. **Calidad educativa K-12:**

$$\text{nearc4}_i \rightarrow \text{mejores escuelas K-12} \rightarrow \text{habilidad}_i \rightarrow \log(\text{wage}_i) \tag{19}$$

Áreas con universidades tienen mejor educación primaria/secundaria.

Argumentos a favor de la exogeneidad

Por qué nearc4 puede ser exógeno (con controles apropiados):

1. **Variación histórica/idiosincrática:**
   - Las universidades fueron fundadas hace 50-100+ años
   - Su ubicación es en gran medida histórica/accidental
   - No fue planeada basándose en características económicas actuales
2. **Controles regionales:**

$$\text{nearc4}_i | \text{región}_i, \text{urbano}_i \perp\!\!\!\perp u_i \tag{20}$$

   Al controlar por:
   - Dummies regionales (reg662, ..., reg669)
   - Indicador de área urbana (smsa66)

   Comparamos personas **dentro** de la misma región y tipo de área, eliminando diferencias sistemáticas.
3. **Pruebas de balance:**

   Verificar que nearc4 no está correlacionado con características observables (como IQ) después de controlar por región.

3 Ejercicio 3b: Estimación OLS vs IV

3.1 Modelo OLS

Especificación completa:

$$\begin{aligned}\log(\text{wage}_i) =& \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 \\ &+ \beta_4 \text{black}_i + \beta_5 \text{smsa}_i + \beta_6 \text{south}_i + u_i\end{aligned} \tag{21}$$

Estimador:

$$\hat{\beta}_1^{\text{OLS}} = (\text{estimación estándar de MCO}) \tag{22}$$

Problema: Si $\text{Cov}(\text{educ}, u) > 0$, entonces $\hat{\beta}_1^{\text{OLS}}$ está sesgado hacia arriba.

3.2 Modelo IV usando 2SLS

Two-Stage Least Squares (2SLS)

Primera etapa:

$$\begin{aligned}\text{educ}_i =& \pi_0 + \pi_1 \text{nearc4}_i + \pi_2 \text{exper}_i + \pi_3 \text{exper}_i^2 \\ &+ \pi_4 \text{black}_i + \pi_5 \text{smsa}_i + \pi_6 \text{south}_i + v_i\end{aligned} \tag{23}$$

Obtener valores ajustados:

$$\widehat{\text{educ}}_i = \hat{\pi}_0 + \hat{\pi}_1 \text{nearc4}_i + \hat{\pi}_2 \text{exper}_i + \dots \tag{24}$$

Segunda etapa:

$$\begin{aligned}\log(\text{wage}_i) =& \beta_0 + \beta_1 \widehat{\text{educ}}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 \\ &+ \beta_4 \text{black}_i + \beta_5 \text{smsa}_i + \beta_6 \text{south}_i + \varepsilon_i\end{aligned} \tag{25}$$

Estimador:

$$\hat{\beta}_1^{\text{IV}} = (\text{coeficiente de } \widehat{\text{educ}}_i \text{ en segunda etapa}) \tag{26}$$

Propiedades del estimador IV

Consistencia:

$$\text{plim}_{n\to\infty} \hat{\beta}_1^{\text{IV}} = \beta_1 \tag{27}$$

siempre que:

- $\text{Cov}(\text{nearc4}, \text{educ}) \neq 0$ (relevancia)
- $\text{Cov}(\text{nearc4}, u) = 0$ (exogeneidad)

Eficiencia:
Cuando la educación NO es endógena ($\text{Cov}(\text{educ}, u) = 0$):

$$\text{Var}(\hat{\beta}_1^{\text{IV}}) > \text{Var}(\hat{\beta}_1^{\text{OLS}}) \tag{28}$$

IV es menos eficiente que OLS. Por eso solo usamos IV cuando hay endogeneidad.

3.3 Interpretación de la Comparación OLS vs IV

Caso 1: $\hat{\beta}_1^{\textbf{OLS}} > \hat{\beta}_1^{\textbf{IV}}$

$$\hat{\beta}_1^{\text{OLS}} - \hat{\beta}_1^{\text{IV}} > 0 \tag{29}$$

Interpretación:

- Consistente con sesgo positivo por habilidad no observada

- OLS sobreestima el retorno de la educación
- El sesgo estimado es: Sesgo $\approx \hat{\beta}_1^{\text{OLS}} - \hat{\beta}_1^{\text{IV}}$

Caso 2: $\hat{\beta}_1^{\mathbf{IV}} > \hat{\beta}_1^{\mathbf{OLS}}$

$$\hat{\beta}_1^{\text{IV}} - \hat{\beta}_1^{\text{OLS}} > 0 \tag{30}$$

Posibles interpretaciones:

1. **Heterogeneidad de efectos y LATE:**
   IV identifica el Local Average Treatment Effect (LATE):

   $$\text{LATE} = E[\text{wage}_i(\text{educ}+1) - \text{wage}_i(\text{educ}) \mid \text{complier}_i] \tag{31}$$

   donde los **compliers** son personas cuya educación fue afectada por nearc4.

   Los compliers pueden tener retornos mayores porque:

   - Vienen de familias con menos recursos (restricciones de crédito)
   - La educación universitaria es más transformativa para ellos
   - Tienen mayores costos de acceso (que nearc4 reduce)

2. **Error de medición en educación:**
   Si educación se mide con error:

   $$\text{educ}_i^{\text{obs}} = \text{educ}_i^{\text{true}} + \eta_i \tag{32}$$

   entonces OLS está sesgado hacia cero (sesgo de atenuación), mientras que IV corrige este sesgo.

4 Ejercicio 3c: Correlación entre nearc4 e IQ

4.1 Objetivo del Análisis

Evaluar si el instrumento nearc4 está correlacionado con habilidad observada (IQ).

Lógica: Si nearc4 está correlacionado con IQ, es probable que también esté correlacionado con habilidad no observada, violando la exogeneidad.

4.2 Regresión

$$\text{IQ}_i = \gamma_0 + \gamma_1 \text{nearc4}_i + \varepsilon_i \tag{33}$$

Hipótesis:

$$\begin{aligned} H_0 &: \gamma_1 = 0 \quad \text{(nearc4 no correlacionado con habilidad)} \\ H_1 &: \gamma_1 \neq 0 \quad \text{(nearc4 correlacionado con habilidad)} \end{aligned} \tag{34}$$

4.3 Interpretación de Resultados

Si γ_1 es estadísticamente significativo:

$$\gamma_1 \neq 0 \quad \Rightarrow \quad \text{Cov}(\text{nearc4}, \text{IQ}) \neq 0 \tag{35}$$

Implicación:

- ADVERTENCIA: nearc4 está correlacionado con habilidad
- Sugiere que nearc4 podría estar correlacionado con u (habilidad no obs.)
- Pone en duda la validez del instrumento

Posibles explicaciones:

1. Selección residencial: familias más educadas viven cerca de universidades
2. Calidad regional: áreas con universidades tienen mejor educación K-12
3. Efectos ambientales: crecer cerca de universidad aumenta aspiraciones/habilidad

Si γ_1 NO es estadísticamente significativo:

$$\gamma_1 \approx 0 \quad \Rightarrow \quad \text{Cov}(\text{nearc4}, \text{IQ}) \approx 0 \tag{36}$$

Implicación:

- Evidencia a favor de la exogeneidad de nearc4
- No hay selección aparente por habilidad
- Apoya la credibilidad del instrumento

5 Ejercicio 3d: nearc4 e IQ con Controles Regionales

5.1 Regresión con Controles

$$\text{IQ}_i = \gamma_0 + \gamma_1 \text{nearc4}_i + \gamma_2 \text{smsa66}_i + \sum_{j=2}^{9} \delta_j \text{reg66}_j + \varepsilon_i \tag{37}$$

donde:

- smsa66 = indicador de vivir en área metropolitana en 1966
- reg66_j = dummies regionales de residencia en 1966

5.2 Interpretación

Si γ_1 era significativo sin controles pero NO lo es con controles

Conclusión:

La correlación entre nearc4 e IQ se debe completamente a diferencias **regionales**.

Implicación:

- CONDICIONAL en región, nearc4 es exógeno respecto a habilidad
- La estrategia de identificación es válida **con controles regionales**
- Los controles son ESENCIALES para la validez del instrumento

Interpretación de la estrategia:

La identificación NO proviene de comparar personas en diferentes regiones, sino de comparar personas **dentro** de la misma región que viven a diferentes distancias de universidades:

$$\beta_1 = E[\log(\text{wage}_i) \mid \text{educ}_i + 1, \text{región}_i, \text{urbano}_i] - E[\log(\text{wage}_i) \mid \text{educ}_i, \text{región}_i, \text{urbano}_i] \tag{38}$$

donde la variación en educación es inducida por nearc4 **dentro** de región.

Si γ_1 sigue siendo significativo con controles

Conclusión:

La correlación entre nearc4 e IQ persiste incluso después de controlar por región.

Implicación:

- La exogeneidad del instrumento es más cuestionable
- Incluso dentro de región, hay selección por habilidad
- Posibles violaciones de la restricción de exclusión

Posibles explicaciones:

1. Selección residencial a nivel micro (dentro de región)
2. Efectos causales de la proximidad sobre habilidad (efectos de red, ambiente)
3. Calidad escolar altamente localizada

Opciones:

- Buscar instrumentos adicionales o alternativos

- Incluir más controles (IQ mismo, si está disponible)
- Reconocer las limitaciones de la estrategia

6 Ejercicio 3e: Importancia de los Controles Regionales

6.1 Síntesis de Evidencia

Del análisis de los incisos 3c y 3d, concluimos:

Lección clave:
La **credibilidad** de un instrumento a menudo depende de qué **controles** se incluyen en el modelo.

Un instrumento que parece endógeno sin controles puede ser válido una vez que se controla por las variables apropiadas.

6.2 Implicaciones para la Especificación del Modelo

Modelo preferido (con controles regionales)

Primera etapa:

$$\begin{aligned}\text{educ}_i =& \pi_0 + \pi_1 \text{nearc4}_i + \pi_2 \text{exper}_i + \pi_3 \text{exper}_i^2 + \pi_4 \text{black}_i \\ &+ \pi_5 \text{smsa}_i + \pi_6 \text{south}_i + \pi_7 \text{smsa66}_i + \sum_{j=2}^{9} \gamma_j \text{reg66}_j + v_i\end{aligned} \tag{39}$$

Segunda etapa:

$$\begin{aligned}\log(\text{wage}_i) =& \beta_0 + \beta_1 \widehat{\text{educ}}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \beta_4 \text{black}_i \\ &+ \beta_5 \text{smsa}_i + \beta_6 \text{south}_i + \beta_7 \text{smsa66}_i + \sum_{j=2}^{9} \delta_j \text{reg66}_j + u_i\end{aligned} \tag{40}$$

Por qué estos controles son esenciales

1. **Validez de la restricción de exclusión:**

$$E[u_i \mid \text{nearc4}_i, \mathbf{X}_i] = E[u_i \mid \mathbf{X}_i] \tag{41}$$

donde $\mathbf{X}_i$ incluye los controles regionales.

Sin los controles, nearc4 podría capturar diferencias regionales en salarios.

2. **Estrategia de identificación:**

La variación identificadora proviene de diferencias en proximidad a universidad **dentro** de región:

$$\beta_1 = \frac{\text{Cov}(\text{nearc4}, \log(\text{wage}) \mid \text{región})}{\text{Cov}(\text{nearc4}, \text{educ} \mid \text{región})} \tag{42}$$

3. **Credibilidad del supuesto:**

   Es más creíble que:

$$\text{Cov}(\text{nearc4}, u \mid \text{región}) = 0 \tag{43}$$

   que:

$$\text{Cov}(\text{nearc4}, u) = 0 \tag{44}$$

   Controlando por región, eliminamos fuentes importantes de endogeneidad.

6.3 Lecciones Generales para Investigación con IV

Mejores prácticas al usar Variables Instrumentales:

1. **Evaluar la relevancia:**
   - Reportar F-estadístico de primera etapa
   - Verificar que $F > 10$ (regla de Stock-Yogo)
2. **Argumentar la exogeneidad:**
   - Explicar por qué el instrumento es "as-if random"
   - Discutir posibles amenazas a la restricción de exclusión
   - Probar correlación con características observables
3. **Identificar controles necesarios:**
   - ¿Qué variables hacen creíble la exogeneidad condicional?
   - Incluir estos controles en AMBAS etapas
   - Discutir la estrategia de identificación condicional
4. **Interpretar cuidadosamente:**
   - IV identifica LATE, no necesariamente ATE
   - Discutir quiénes son los "compliers"
   - Explicar por qué el efecto puede diferir de OLS
5. **Pruebas de robustez:**
   - Mostrar resultados con/sin controles

- Probar instrumentos alternativos si están disponibles
- Realizar pruebas de sobreidentificación si hay múltiples instrumentos

7 Resumen: Contribución de Card (1993, 1995)

7.1 Hallazgos Principales

Estimaciones típicas:

- OLS: $\hat{\beta}_1 \approx 0.07$ (7% retorno por año)
- IV: $\hat{\beta}_1 \approx 0.10 - 0.13$ (10-13% retorno por año)

Interpretación:
En la muestra de Card, las estimaciones IV son **mayores** que OLS, sugiriendo que:

1. Los "compliers" (personas al margen de ir a universidad) tienen retornos altos
2. Posible corrección de error de medición en educación
3. El sesgo por habilidad puede ser compensado por estos efectos

7.2 Impacto en la Literatura

El trabajo de Card es fundamental porque:

1. **Estableció el uso de IV en economía laboral**

   Demostró que es posible encontrar fuentes de variación exógena en datos observacionales.

2. **Popularizó el concepto de LATE**

   Clarificó que IV identifica efectos para "compliers", no efectos promedio poblacionales.

3. **Desarrolló estrategias de validación**

   Mostró la importancia de probar la exogeneidad indirectamente (correlación con observables).

4. **Estimación creíble de retornos causales**

   Proporcionó estimaciones más confiables del valor de la educación para política pública.

7.3 Limitaciones y Extensiones

Limitaciones:

- LATE específico a compliers (no generalizable a toda la población)

- Depende crucialmente de controles regionales
- Puede no aplicar a contextos/periodos diferentes

Extensiones posteriores:

- Uso de reformas educativas como instrumentos
- Diseños de discontinuidad en regresión
- Métodos de matching y propensity scores
- Análisis de heterogeneidad en retornos