

Examen Final

Profesor: Juan Palomino

December 10, 2022

Instrucciones

- (1) El examen se encuentra habilitado en el sistema PAIDEIA a partir de las 10:00am y tiene hasta las 12:00pm para entregarme los cuadernillos. No se recibirá ningún cuadernillo pasado esa hora.
- (2) Pueden subir la solución de su programación a PAIDEIA hasta la hora final del examen.
- (3) Redacción y ortografía **SERÁN TOMADAS EN CUENTA** para la calificación de la prueba. **SI EL ENVÍO DEL MANUSCRITO NO ES LEGIBLE RECIBIRÁN LA PENALIDAD DE MENOS UN PUNTO.**
- (4) Si se evidencia plagio en las respuestas, automáticamente la prueba será anulada y remitida a las autoridades correspondientes de la Facultad.
- (5) La pregunta sobre Heterocedasticidad lo pueden hacer en excel, Stata, RStudio, Colab u otro software. Pero solo quiero sus respuestas numéricas y su interpretación. En esta pregunta no se evaluará el procedimiento.

Teoría

Responder brevemente los siguientes conceptos:

(1) Explicar las causas de la endogeneidad. **(1 punto)**

La endogeneidad se debe a 3 problemas principales, si uno de estos problemas están presentes, los estimadores son inconsistentes, y no miden la magnitud y dirección de la causa, sino solo una simple correlación.

- **Variables Omitidas:** La omisión de variables relevantes provoca sesgo en las estimaciones de MCO cuando las variables omitidas no son ortogonales a las variables incluidas.
- **Error de Medición:** Si en nuestro modelo econométrico la variable X es difícil de medir o de observar, el investigador recurrirá a una proxy de dicha variable, la cual no es una medida perfecta, generando un sesgo de error de medida.
- **Sesgo por simultaneidad:** Se refiere al hecho de que X afecte a Y , y que Y afecta a X .

(2) Defina lo que es un instrumento y las propiedades que debe cumplir. **(1 punto)**

Se define a z una variable instrumental siempre y cuando satisfaga dos supuestos:

- **Exogeneidad del instrumento:** $cov(z, \epsilon) = 0$, entonces z no está correlacionada con ϵ ; es decir, z es exógena de la ecuación.
- **Relevancia del instrumento:** $cov(z, x) \neq 0$, entonces z está correlacionada, positiva o negativamente con x . Por tanto, z es relevante para explicar la variación en x .

Modelo de Ecuaciones Simultáneas

Considere el siguiente modelo de demanda y oferta de dinero:

$$\begin{aligned}\text{Demanda de dinero } M_t^d &= \beta_0 + \beta_1 Y_t + \beta_2 R_t + \beta_3 P_t + \epsilon_{1t} \\ \text{Oferta de dinero } M_t^s &= \alpha_0 + \alpha_1 Y_t + \epsilon_{2t}\end{aligned}$$

en donde M =dinero, Y =ingreso, R =tasa de interés, P =precio. Suponga que R y P están predeterminados.

(a) Halle la forma reducida del modelo. **(1 punto)**

$$\begin{aligned}M_t^d - \beta_1 Y_t &= \beta_0 + \beta_2 R_t + \beta_3 P_t + \epsilon_{1t} \\ M_t^s - \alpha_1 Y_t &= \alpha_0 + \epsilon_{2t}\end{aligned}$$

La forma estructural matricial es:

$$\begin{bmatrix} M_t & Y_t \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -\beta_1 & -\alpha_1 \end{bmatrix} = \begin{bmatrix} 1 & R_t & P_t \end{bmatrix} \begin{bmatrix} \beta_0 & \alpha_0 \\ \beta_2 & 0 \\ \beta_3 & 0 \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} & \epsilon_{2t} \end{bmatrix}$$

Esto es igual a:

$$y_t \underbrace{\Gamma}_{2 \times 2} = X_t \underbrace{B}_{3 \times 2} + \epsilon_t$$

Multiplicando por Γ^{-1}

$$y_t = X_t \underbrace{B}_{3 \times 2} \underbrace{\Gamma^{-1}}_{2 \times 2} + \epsilon_t \Gamma^{-1}$$

Obtenemos la forma reducida matricial:

$$y_t = X_t \underbrace{\Pi}_{3 \times 2} + \nu_t$$

donde $\Pi = B\Gamma^{-1}$ y $\nu = \epsilon\Gamma^{-1}$. Equivalente a:

$$\begin{bmatrix} M_t & Y_t \end{bmatrix} = \begin{bmatrix} 1 & R_t & P_t \end{bmatrix} \begin{bmatrix} \pi_1 & \pi_4 \\ \pi_2 & \pi_5 \\ \pi_3 & \pi_6 \end{bmatrix} \begin{bmatrix} \nu_{1t} & \nu_{2t} \end{bmatrix}$$

En ecuaciones, la forma reducida es:

$$\begin{aligned} M_t &= \pi_1 + \pi_2 R_t + \pi_3 P_t + \nu_{1t} \\ Y_t &= \pi_4 + \pi_5 R_t + \pi_6 P_t + \nu_{2t} \end{aligned}$$

Los parámetros reducidos están en función de los estructurales $\Pi = B\Gamma^{-1}$, entonces:

$$\begin{bmatrix} \pi_1 & \pi_4 \\ \pi_2 & \pi_5 \\ \pi_3 & \pi_6 \end{bmatrix} = \begin{bmatrix} \beta_0 & \alpha_0 \\ \beta_2 & 0 \\ \beta_3 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -\beta_1 & \alpha_1 \end{bmatrix}^{-1}$$

(b) ¿Está identificada la función de demanda? **(1 punto)**

La demanda no está identificada porque no hay variables de la oferta que provoquen desplazamientos en la oferta que permitan identificarla.

Usando la condición de orden donde $k = 3$ para la ecuación de demanda:

- $k - k_j = 3 - 3 = 0$
- $g_j - 1 = 2 - 1 = 1$

Como $k - k_j < g_j - 1$, la ecuación de demanda está subidentificada.

(c) ¿Está identificada la función de oferta? **(1 punto)**

La oferta si está identificada porque hay variables en la demanda que provoca desplazamientos en la demanda, que permite identificar a la oferta.

Usando la condición de orden donde $k = 3$ para la ecuación de oferta:

- $k - k_j = 3 - 1 = 2$
- $g_j - 1 = 2 - 1 = 1$

Como $k - k_j > g_j - 1$, la ecuación de oferta está sobreidentificada.

(d) ¿Cuál método se utilizaría para estimar los parámetros de la(s) ecuación(es) identificada(s)? Por qué? Explicar los pasos de la estimación. **(1.5 puntos)**

Se estima por MC2E dado que la sobreidentificación no la afecta. Los pasos son los siguientes:

(i) En la primera etapa se estima la forma reducida de Y por MCO y se calcula la predicción \hat{Y}_t :

$$\hat{Y}_t = \hat{\pi}_4 + \hat{\pi}_5 R_t + \hat{\pi}_6 P_t$$

(ii) Y en la segunda etapa se estima la oferta de dinero por MCO:

$$\hat{M}_t = \alpha_0 + \alpha_1 \hat{Y}_t + \eta_t$$

Así se obtiene una estimación única de los parámetros de la oferta de dinero: $\hat{\alpha}_0$ y $\hat{\alpha}_1$.

- (e) Suponga que se modifica la función de oferta agregando las variables explicativas Y_{t-1} y M_{t-1} ¿Qué sucede con el problema de la identificación? ¿Se utilizará aún el método que utilizó en (c)? ¿Por qué sí o por qué no? Explicar los pasos de la estimación. **(1.5 puntos)**

El sistema sería:

$$\begin{aligned} \text{Demanda de dinero } M_t^d &= \beta_0 + \beta_1 Y_t + \beta_2 R_t + \beta_3 P_t + \epsilon_{1t} \\ \text{Oferta de dinero } M_t^s &= \alpha_0 + \alpha_1 Y_t + \alpha_2 Y_{t-1} + \alpha_3 M_{t-1} + \epsilon_{2t} \end{aligned}$$

En este caso el número de exógenas es $k = 5$.

Usando la condición de orden para la ecuación de demanda:

- $k - k_j = 5 - 3 = 2$
- $g_j - 1 = 2 - 1 = 1$

Como $k - k_j > g_j - 1$, la ecuación de demanda está sobreidentificada. Usando la condición de orden para la ecuación de oferta:

- $k - k_j = 5 - 3 = 2$
- $g_j - 1 = 2 - 1 = 1$

Como $k - k_j > g_j - 1$, la ecuación de oferta está sobreidentificada.

Se sigue utilizando MC2E porque la sobreidentificación no le afecta. Los pasos son similares al caso anterior para cada ecuación.

Heterocedasticidad

Considere la base de datos `base_anemia.dta` que se encuentra en PAIDEIA. Esta base de datos cuenta con información sobre la anemia en el Perú, construido a partir de la Encuesta Nacional Demográfica de Salud (ENDES). A partir de esta base de datos, se pide estimar el siguiente modelo:

$$\text{hemoglobina} = \beta_0 + \beta_1 \text{sexo} + \beta_2 \text{edadmeses} + \beta_3 \text{pesoalnacer} + \beta_4 \text{educmadre} + \beta_5 \text{area} + \epsilon \quad (1)$$

donde:

Variables	Valor
sexo	sexo del niño 1: Mujer 0: Hombre
edadmeses	Edad del niño en meses
pesoalnacer	Peso del niño al nacer (kilos)
educmadre	Nivel Educativo de la madre 0: Primaria o menos 1: Secundaria 2: Superior
area	1: Urbano 0: Rural

Utilizando cualquier software (STATA, RStudio o Colab), se le pide:

1. Realizar el test de White y el test de Breusch-Pagan para detectar heterocedasticidad. ¿Existe heterocedasticidad en el modelo? **(4 puntos)**

Pista: Recuerden que si una variable tiene n categorías, debe haber al menos $n - 1$ parámetros estimados para esa variable.

Se rechaza la hipótesis nula al 1% de significancia en ambos test. Por lo tanto, existe heterocedasticidad en el modelo.

Ver solucionario en <https://colab.research.google.com/drive/1amNGHFUGqKd9aR40jEI91zHqJjYHavfF?usp=sharing> o en el código de Stata.

Variables Instrumentales

Card (1994) utilizó datos de salarios y educación para una muestra de hombres en 1976 para estimar el retorno a la educación utilizando la siguiente ecuación de Mincer:

$$\ln w_i = \beta_1 + \beta_2 S_i + \beta_3 E_i + \beta_4 E_i^2 + h_i' \delta + \epsilon_i$$

donde $\ln w_i$ denota el logaritmo de salarios individuales, S_i denota años de escolaridad y E_i denota años de experiencia. En ausencia de información sobre la experiencia real, E_i a veces se reemplaza por “experiencia potencial”, medida como $\text{age}_i - S_i - 6$, suponiendo que las personas empiezan la escuela a la edad de 6 años. Esta especificación generalmente se aumenta con variables explicativas adicionales, h_i , que uno quiere controlar, como dummies regionales y raciales.

Se usa los datos de 3010 hombres de la Encuesta Longitudinal Nacional de Hombres Jóvenes de EE. UU. (NLS-Y), empleados en Card (1994). En esta encuesta, se sigue a un grupo de individuos desde 1966 cuando tenían entre 14 y 24 años, y se los entrevistaba en varios años consecutivos. La información del mercado laboral que se utiliza abarca 1976. Varios resultados se presentan en la Tabla 2.1. Los controles utilizados son los siguientes:

- *experience*: experiencia potencial en años
- *experience*²: experiencia potencial al cuadrado
- *black*: dummy para raza negra (1 si el individuo es de raza negra)
- *south*: dummy si individuo vive en el sur
- *smsa*: dummy para residencia en area metropolitana
- *smsa66*: dummy para residencia en area metropolitana en 1966
- *reg662 – reg668*: un set de variables dummies regionales

1. La columna (1) de la Tabla 2.1 muestra el resultado de una regresión MCO. Interprete el coeficiente de escolaridad (Educ), que mide los años de escolaridad. ¿Es este coeficiente potencialmente endógeno? ¿Por qué? **(1 punto)**

Hay una variedad de razones por las cuales la escolarización puede estar correlacionada con el término de error. Uno de ellos es el “sesgo de habilidad”. En este caso, β estará sesgado hacia arriba. Para ver esto, considere el siguiente modelo verdadero:

$$\ln w_i = \alpha + \beta S_i + \gamma A_i + h_i' \delta + \epsilon_i$$

donde A_i es una medida de habilidad, y h ahora incluye experiencia y experiencia al cuadrado. ¿Cuál sería la consecuencia de omitir A de la ecuación salarial? Se puede demostrar que si A_i se omite de la ecuación, entonces el estimador MCO de β :

$$\hat{\beta}_{OLS} \xrightarrow{p} \beta + \gamma \frac{\text{Cov}(S_i, A_i)}{\text{Var}(S_i)}$$

Luego, la estimación de MCO del coeficiente de escolaridad β incluye el efecto indirecto de la habilidad sobre el log de salarios a través de la escolarización $\gamma \frac{\text{Cov}(S_i, A_i)}{\text{Var}(S_i)}$, así como el efecto directo de la escolarización β . Si la correlación entre S y A es positiva, $\frac{\text{Cov}(S_i, A_i)}{\text{Var}(S_i)} > 0$, entonces $\hat{\beta}_{OLS}$ está sesgado asintóticamente hacia arriba.

Los resultados para el MCO implican que un aumento de 1 año de escolaridad se correlaciona con un aumento de $= 7.5\%$ en los salarios.

2. En la columna (2) controlamos por la habilidad (IQ). ¿Cómo interpretas el resultado para el coeficiente de escolaridad (Educ)? **(1 punto)**

El hecho de que las personas con test scores más altos (en IQ o prueba de rendimiento) tienden a tener mayores ganancias y más escolaridad, a menudo se interpreta como evidencia de sesgo de habilidad. El coeficiente de escolaridad estimado es menor, lo que confirma nuestra predicción de que el coeficiente de escolaridad estimado incluye el efecto de la habilidad en la tasa salarial cuando la habilidad se omite de la regresión.

3. La columna (3) presenta una estimación MC2E donde el instrumento para Educ es Nearc4, que es una variable dummy para determinar si alguien creció cerca de una universidad durante cuatro años como una variable instrumental para la educación. ¿Este modelo está subidentificado, identificado o sobreidentificado? Comente el coeficiente para Educ. **(2 puntos)**

El modelo es identificado. Curiosamente, la estimación IV del retorno a la educación es casi dos veces ($0.132/0.075 = 2$) tan grande como la estimación MCO de la columna 2, pero el error estándar de la estimación IV es más de 18 veces ($0.055/0.003 = 18$) mayor que el error estándar de MCO. El intervalo de confianza del 95% de la estimación IV tiene un rango muy amplio. Los intervalos de confianza más grandes son un precio que debemos pagar para obtener un estimador consistente del retorno a la educación cuando pensamos que la educación es endógena.

4. La columna (4) muestra la estimación de la ecuación de forma reducida. ¿Nearc4 es un instrumento válido? **(1 punto)**

Recuerde que uno de los requisitos es que el instrumento está parcialmente correlacionado con la educación una vez que se han eliminado otras variables exógenas (south, smsa, etc.). Para verificar este requisito, regresionamos educ en nearc4 y todas las variables exógenas que aparecen en la ecuación. Es decir, estimamos la forma reducida de educ.

Los resultados para la forma reducida muestran que, en igualdad de condiciones, las personas que vivían cerca de una universidad en 1966 tenían, en promedio, alrededor de un tercio (el coeficiente es 0.32) de un año más de educación que aquellos que no crecieron cerca de una universidad. ¿Tenemos un problema de instrumento débil? Recuerde que el estadístico F para la significancia (conjunta) de los instrumentos en la primera etapa debe exceder 10. En este caso, dado que tenemos solo un instrumento, el estadístico F es $(0.320/0.088)^2 = 13.22$, que es mayor que 10. Entonces, nuestro instrumento es lo suficientemente fuerte.

5. Explica por qué Nearc4 podría correlacionarse con la habilidad (que está en el término de error). **(1 punto)**

El segundo requisito es que el instrumento no esté correlacionado con el término de error en la ecuación salarial. Sin embargo, esto no puede ser testeado. La validez de los instrumentos solo puede probarse, hasta cierto punto, si el modelo está sobreidentificado. Por lo tanto, para que la proximidad a la universidad sirva como un instrumento legítimo para la educación completa, debe afectar las decisiones individuales de escolarización pero no tener un efecto directo en los ingresos. Sin embargo, hay al menos tres razones por las cuales los hombres que crecieron cerca de una universidad pueden tener mayores ganancias que otros hombres, controlando la educación, la información geográfica y los antecedentes de los padres.

- Las familias que ponen un fuerte énfasis en la educación pueden optar por vivir cerca de una universidad. Los niños de estas familias pueden tener una mayor habilidad o pueden estar más motivados para lograr el éxito en el mercado laboral. En cualquier caso, podríamos tener una correlación positiva entre la proximidad a la universidad y ϵ .
- La presencia de una universidad puede estar asociada con una mayor calidad escolar en las escuelas primarias y secundarias cercanas.
- Si los hombres que crecieron en áreas con una universidad cercana tienden a vivir en áreas con salarios más altos, entonces la proximidad a la universidad puede estar correlacionada con primas salariales geográficas no observadas.

6. Explique la idea de las columnas 5 y 6 de la tabla de acuerdo con su respuesta en (5). ¿Qué concluye sobre estos resultados? **(2 puntos)**

Recuerde que no controlamos la habilidad, por lo que esta variable va al término de error. Sin embargo, como explicamos anteriormente, podría suceder que las familias elijan vivir cerca de una universidad y que sus hijos tengan una mayor habilidad. Entonces, una pregunta interesante es: ¿podría nearc4 estar correlacionado con algo en el término de error, como la habilidad no observada? Se sabe que los scores de IQ varían según la región geográfica, y también la disponibilidad de la universidad. Podría ser que, por una variedad de razones, las personas con habilidades superiores crecen en áreas con universidades cercanas.

La columna 5 de la Tabla 2.1 muestra el resultado de la regresión de IQ sobre nearc4. El score de IQ predicho es aproximadamente 2.6 puntos más alto para un hombre que creció cerca de una

universidad. La diferencia es estadísticamente significativa. En otras palabras, el instrumento está correlacionado con alguna variable que sabemos que afecta las ganancias y está en el término de error. Por lo tanto, la condición $E(z_i \cdot \epsilon_i) = 0$ no se cumple. Pero, ¿qué sucede si controlamos por otras dummies regionales? La columna 6 de la Tabla 2.1 es la misma regresión de la columna 5, pero ahora controlamos para `smsa`, `smsa66` y variables dummies regionales. Ahora, la relación entre `IQ` y `nearc4` es mucho más débil y estadísticamente insignificante. En otras palabras, una vez que controlamos por región y el medio ambiente mientras crecemos, no hay un vínculo aparente entre el score de `IQ` y vivir cerca de una universidad. Esto implica que es importante incluir estas variables en la ecuación salarial para controlar las diferencias en el acceso a las universidades que también podrían estar correlacionadas con la habilidad.

Tabla 2.1. Retornos a la Educación

Variables	Dependent variable:					
	log(wage) (1)	log(wage) (2)	log(wage) (3)	educ (4)	IQ (5)	IQ (6)
Educ	0.075*** (0.003)	0.070*** (0.005)	0.132** (0.055)			
Nearc4				0.320*** (0.088)	2.596*** (0.745)	0.348 (0.814)
Experience	0.085*** (0.007)	0.095*** (0.009)	0.108*** (0.024)	-0.413*** (0.034)		
<i>Experience</i> ²	-0.002*** (0.0003)	-0.003*** (0.0005)	0.002*** (0.0003)	0.001 (0.002)		
Black	-0.199*** (0.018)	-0.148*** (0.027)	-0.147*** (0.054)	-0.936*** (0.094)		
South	-0.148*** (0.026)	-0.100*** (0.032)	-0.145*** (0.027)	-0.052 (0.135)		
SMSA	0.136*** (0.020)	0.123*** (0.024)	0.112*** (0.032)	0.402*** (0.105)		
reg661	-0.119*** (0.039)	-0.122*** (0.044)	-0.108*** (0.042)	-0.210 (0.202)		2.892 (1.797)
reg662	-0.022 (0.028)	-0.019 (0.032)	-0.007 (0.033)	-0.289** (0.147)		3.991*** (1.294)
reg663	0.026 (0.027)	0.004 (0.031)	0.040 (0.032)	-0.238* (0.143)		1.333 (1.259)
reg664	-0.063* (0.036)	-0.073* (0.040)	-0.058 (0.038)	-0.093 (0.186)		2.349 (1.635)
reg665	0.009 (0.036)	0.002 (0.043)	0.038 (0.047)	0.483** (0.188)		-5.584*** (1.322)
reg666	0.022 (0.040)	0.037 (0.051)	0.055 (0.053)	-0.513** (0.210)		-4.529*** (1.698)
reg667	-0.001 (0.039)	-0.031 (0.047)	0.027 (0.049)	-0.427** (0.206)		-5.502*** (1.520)
reg668	-0.175*** (0.046)	-0.170*** (0.052)	-0.191*** (0.051)	0.314 (0.242)		-0.033 (2.111)
smsa66	0.026 (0.019)	0.036 (0.023)	0.019 (0.022)	0.025 (0.106)		1.089 (0.809)
IQ		0.002*** (0.001)				
Constant	4.739*** (0.072)	4.501*** (0.106)	3.774*** (0.935)	16.849*** (0.211)	100.611*** (0.627)	101.882*** (1.292)
Observations	3010	2061	3010	3010	2061	2061
R^2	0.300	0.237	0.238	0.477	0.006	0.063
Adjusted R^2	0.296	0.231	0.234	0.474	0.005	0.058

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$