

Hoja de Ejercicios 6

Modelos con variables explicativas endógenas

Nota: En aquellos ejercicios en los que se incluyen estimaciones y referencia al archivo de datos utilizado, el estudiante debería comprobar los resultados obtenidos en Gretl.

1. Recuerde que en el modelo lineal simple

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

cuando X es endógena pero disponemos de un instrumento Z que puede estar correlacionado con ε , tenemos que

$$p \lim \tilde{\beta}_1 = \beta_1 + \frac{C(Z, \varepsilon)}{C(Z, X)} \quad (\text{estimador simple VI})$$

$$p \lim \hat{\beta}_1 = \beta_1 + \frac{C(X, \varepsilon)}{V(X)} \quad (\text{estimador MCO})$$

Suponga que $\sigma_X = \sigma_\varepsilon$, de modo que la variación poblacional del término de error es la misma que la de X . Suponga que la variable instrumental Z , presenta una cierta correlación con ε , $\rho(Z, \varepsilon) = 0.1$. Suponga también que Z y X tienen una correlación algo mayor, $\rho(X, Z) = 0.2$.

- (a) ¿Cuál es el sesgo asintótico del estimador de VI que utiliza Z como instrumento?
 - (b) ¿Cuánta correlación debería existir entre X y ε para que el estimador MCO tenga más sesgo asintótico que el estimador anterior de VI?
2. [Ejercicio 16.10 del libro de Wooldridge] Utilice los datos de `MROZ.gdt` y considere el ejemplo de oferta de trabajo de mujeres.

Sean *hours* el total de horas trabajadas en el año por la mujer, *wage* su salario-hora, *educ* sus años de educación, *age* su edad en años, *kidslt6* el número de hijos a su cargo menores de 6 años, *nwifeinc*, los ingresos de su hogar (excluidos los ingresos salariales de la mujer), *exper* sus años de experiencia laboral, *motheduc* y *fatheduc* los años de educación de su padre y de su madre.

- (a) Estime la ecuación de oferta de trabajo

$$\ln(\text{hours}) = \alpha_1 \ln(\text{wage}) + \beta_{10} + \beta_{11} \text{educ} + \beta_{12} \text{age} \\ + \beta_{13} \text{kidslt6} + \beta_{14} \text{nwifeinc} + u_1$$

por MC2E usando *exper* y *exper*² como instrumentos para $\ln(\text{wage})$, y compare el resultado con el obtenido con *hours* como variable dependiente.

- (b) En la ecuación de oferta de trabajo del apartado (a), es posible que *educ* sea endógena, dado que la habilidad está omitida. Utilice *motheduc* y *fatheduc* como instrumentos para *educ*. Ahora hay dos variables explicativas endógenas en esta ecuación.

- (c) Contraste las restricciones de sobre-identificación en la estimación MC2E del apartado (b). ¿Cuál es su conclusión acerca de los instrumentos?
3. Queremo estudiar los rendimientos salariales de la educación, esto es, el efecto que los años de educación (ED) tienen sobre el salario (W). Disponemos de una muestra de 3010 hombre jóvenes estadounidenses en 1976, procedentes de *National Longitudinal Survey of Young Men* (NLSYM). Para each individual, we observe ED (Years of education), EX (Experience, in years), EX^2 (Experience squared), $WHITE$ (Binary variable that takes the value of 1 if the young men is white and 0 otherwise).

Se considera el siguiente modelo para analizar el rendimiento de la educación:

$$\ln W = \beta_0 + \beta_1 ED + \beta_2 EX + \beta_3 EX^2 + \varepsilon_1 \quad (1)$$

Nótese que el rendimiento del término de error ε_1 puede incluir factores inobservables no incluidos en el modelo que afecten a los salarios. En concreto, puede incluir incluye $ABIL$ (habilidad), que es inobservable. Disponemos además información sobre la variable $NEAR$ que es una variable ficticia que toma valor 1 si el individuo vive cerca de una universidad, y cero en otro caso. Sabemos además que $C(NEAR, \varepsilon_1) = 0$.

- (a) Estime el modelo (S1) por MCO, usando la submuestra de hombres jóvenes blancos (es decir, restringiendo a las observaciones con $WHITE = 1$).
- (b) Suponga que $ABIL$ es una variable relevante, que $C(ABIL, ED) \neq 0$, y que no está correlacionada con el resto de las variables explicativas en (S1). ¿Será el estimador MCO de β_1 en (S1) un estimador consistente del efecto causal de la educación en los salarios? Justifique su respuesta.
- (c) Considere el uso de $NEAR$ como instrumento. Estime la ecuación auxiliar

$$ED = \pi_0 + \pi_1 EX + \pi_2 EX^2 + \pi_3 NEAR + v$$

Dados los supuestos y estimaciones anteriores, ¿podemos afirmar que $NEAR$ es un instrumento válido para ED ? Justifique su respuesta.

- (d) Estime el modelo (S1) por MC2E, usando la submuestra de hombres jóvenes blancos, y $NEAR$ como instrumento para ED .
- (e) ¿Es ED exógena? Justifique su respuesta. Dados los resultados, elija la estimación apropiada del efecto causal de la educación en los salarios e interprete.
- (f) Considere ahora la muestra completa. Para tener en cuenta diferencias étnicas, se considera la ecuación

$$\ln W = \beta_0 + \beta_1 ED + \beta_2 EX + \beta_3 EX^2 + \beta_4 WHITE + \beta_5 (ED \times WHITE) + \varepsilon_2 \quad (2)$$

Proponga instrumentos para las dos variables potencialmente endógenas, ED y $(ED \times WHITE)$. Estime las ecuaciones de primera etapa correspondientes y compruebe la validez de los instrumentos.

- (g) Estime (S2) por MCO y por MC2E y compare los resultados. ¿Son ED y $(ED \times WHITE)$ exógenas? ¿Qué estimaciones elegiría? Justifique su respuesta.

4. Según la Teoría de la Renta Permanente de Friedman, se puede escribir

$$Y_i^* = \alpha + \beta X_i^* \quad (3)$$

donde Y_i^* es el consumo 'permanente' y X_i^* es la renta 'permanente'. En lugar de observar las variables 'permanentes', se observa

$$\begin{aligned} Y_i &= Y_i^* + u_i \\ X_i &= X_i^* + v_i \end{aligned}$$

donde Y_i , X_i son cantidades que podemos observar con errores de medida (u_i , v_i) en Y_i^* , X_i^* respectivamente.

Utilizando las cantidades observables, podemos escribir la función de consumo como

$$\begin{aligned} Y_i &= \alpha + \beta (X_i - v_i) + u_i \\ &= \alpha + \beta X_i + (u_i - \beta v_i) \end{aligned}$$

(a) Suponga que $E(u_i) = E(v_i) = 0$, $Var(u_i) = \sigma_u^2$, $Var(v_i) = \sigma_v^2$, $Cov(Y_i^* u_i) = 0$, $Cov(X_i^* v_i) = 0$, $Cov(u_i X_i^*) = Cov(v_i Y_i^*) = Cov(v_i u_i) = 0$. Demuestre que el estimador MCO de β converge en probabilidad a

$$\frac{\beta}{1 + (\sigma_v^2 / \sigma_{X^*}^2)}$$

(b) Comente el signo del sesgo asintótico de $\hat{\beta}$.

5. El argumento de que la inflación estimula el crecimiento ha sido desacreditado por estudios empíricos de sección cruzada. En estos estudios se realiza la regresión de la tasa de crecimiento de la renta real sobre la tasa de inflación. Sin embargo, las variables utilizadas para medir la inflación y la renta real, X_i e Y_i , respectivamente, están sujetas a error. Supongamos que, en realidad, existe una relación exacta entre la verdadera tasa de crecimiento de la renta real, Y_i^* y la verdadera tasa de inflación, X_i^* . Se supone que la tasa de crecimiento de la renta nominal, $W_i^* = X_i^* + Y_i^*$ se mide correctamente y se utiliza para obtener una medida de la tasa de crecimiento de la renta real las relaciones

$$\begin{aligned} Y_i &= W_i^* - X_i \\ X_i &= X_i^* + \varepsilon_i \quad \varepsilon_i \sim iid(0, \sigma_\varepsilon^2) \end{aligned}$$

(a) Derive el límite probabilístico del estimador de MCO de la regresión de Y sobre X .

(b) A partir de los resultados del apartado anterior, ¿qué puede decir sobre los futuros resultados de los estudios empíricos en contra de que la inflación estimula el crecimiento?.

6. Considere la siguiente especificación para las funciones de demanda y oferta de vino de un país

$$\begin{aligned} q_i^D &= \alpha_1 p_i + \alpha_2 y_i + u_{i1} \\ q_i^S &= \beta p_i + u_{i2} \\ q_i^D &= q_i^S = q_i \end{aligned}$$

donde, para un municipio i dado, q_i es el consumo de vino por hogar, p_i es el precio relativo del vino e y_i es la renta media del hogar. En el sistema se determinan por tanto conjuntamente precio y cantidad de equilibrio, mientras que la variable y_i es exógena. **Todas las variables están en logaritmos.** Para una muestra de tamaño 1000, se obtienen los siguientes estadísticos:

$$\begin{aligned} \sum_i p_i^2 &= 42 & \sum_i p_i q_i &= 5 & \sum_i p_i y_i &= 12 \\ \sum_i y_i^2 &= 10 & \sum_i y_i q_i &= 3 & & \\ \sum_i q_i^2 &= 11 & & & & \end{aligned}$$

Suponga que estamos interesados en estimar la ecuación de oferta.

- (a) Calcule el estimador de mínimos cuadrados ordinarios de β e interprete el coeficiente. ¿Es este estimador consistente? Justifique la respuesta.
 - (b) ¿Es y_i es un instrumento válido para p_i ? Justifique su respuesta. Calcule el estimador de variables instrumentales de β basado en este instrumento.
7. Una empresa que vende artículos deportivos quiere evaluar el impacto de la renta de sus clientes en sus ventas. Para ello, realiza una encuesta para una muestra de individuos que compran artículos deportivos y considera la siguiente especificación:

$$\begin{aligned} cons &= \beta_0 + \beta_1 renta + \beta_2 edad + \beta_3 edad^2 + \beta_4 sexo + \beta_5 civil + \beta_6 sur \\ &+ \beta_7 peso + \beta_8 peso \times sexo + \beta_9 sur \times sexo + \varepsilon \end{aligned} \quad (C)$$

donde *cons* es el consumo anual de artículos deportivos del individuo (medido en miles de euros),

renta es su ingreso anual (medido en miles de euros),

edad su edad medida en años,

sexo es una variable binaria que toma el valor 1 si el individuo es una mujer y 0 si es un hombre,

civil representa al estado civil del individuo, es una variable binaria que toma el valor 1 si el individuo es soltero y 0 si no lo es,

sur es una variable binaria que toma el valor 1 si el individuo vive en el sur del país y 0 si no,

peso es el peso del individuo medido en kilogramos.

Además, el analista piensa que la renta puede estar correlacionada con características no observables que afectan a la vez al gasto en artículos deportivos. Si eso fuese así, es de esperar que $C(renta, \varepsilon) \neq 0$. El resto de las variables explicativas del modelo no están correlacionadas con el error. Además de las variables incluidas en el modelo anterior, se dispone de información acerca de los años de educación del individuo (*educ*) y de su padre (*educp*). Para el análisis, supondrá que estas variables no están correlacionadas con características no observables que afectan al gasto en artículos deportivos (ε). Cabe esperar que individuos más educados o con padres más educados.

Se han obtenido las siguientes estimaciones con Gretl (donde algunos valores pueden estar omitidos):

SALIDA 1: OLS estimates using the 935 observations 1–935

Dependent variable: *cons*

	Coefficient	Std. Error	<i>t</i> -ratio	p-value
<i>const</i>	37.7486	31.5318	1.1972	0.2316
<i>renta</i>	0.6671	0.6138	1.0869	0.2774
<i>edad</i>	0.3401	1.9090	0.1782	0.8586
<i>edad2</i>	-0.0036	0.0014		
<i>sexo</i>	-2.2428	0.6031	-3.7190	0.001
<i>civil</i>	0.7775			0.002
<i>sur</i>	-0.1803	0.5536	-0.3257	0.7447
<i>peso</i>	-0.1025	0.0512	-2.0023	0.0455
<i>peso</i> × <i>sexo</i>	-0.0323	0.1506	-0.2144	0.8303
<i>sur</i> × <i>sexo</i>	-0.0692	1.4793	-0.0468	0.9627

Mean dependent var	0.043929	S.D. dependent var	0.007224
Sum squared resid	0.047795	S.E. of regression	0.007188
R^2	0.019495	Adjusted R^2	0.009955
$F(9, 925)$	2.043539	P-value(F)	0.032059
Log-likelihood	3292.838	Akaike criterion	-6565.675
Schwarz criterion	-6517.270	Hannan-Quinn	-6547.218

NOTA: El R^2 de una estimación similar a la anterior omitiendo *sexo*, *peso* × *sexo* y *sur* × *sexo* es 0.009495

SALIDA 2: OLS estimates using the 935 observations 1–935

Dependent variable: *renta*

	Coefficient	Std. Error	<i>t</i> -ratio	p-value
<i>const</i>	1.2033	1.8148	0.6631	0.5075
<i>edad</i>	-0.0577	0.1103	-0.5233	0.6010
<i>edad2</i>	0.0006	0.0017	0.3597	0.7192
<i>sexo</i>	0.1523	0.0995	1.5305	0.1263
<i>civil</i>	-0.1739	0.0436	-3.9930	0.0001
<i>sur</i>	0.0589	0.0315	1.8688	0.0620
<i>peso</i>	-0.0035	0.0029	-1.2100	0.2267
<i>peso</i> × <i>sexo</i>	-0.0165	0.0102	-1.6209	0.1055
<i>sur</i> × <i>sexo</i>	0.0977	0.0966	1.0111	0.3123
<i>educp</i>	0.0138	0.0047	2.9464	0.0033
<i>educ</i>	0.0470	0.0068	6.9315	0.0000

Mean dependent var	0.975920	S.D. dependent var	0.405896
Sum squared resid	99.17078	S.E. of regression	0.368579
R^2	0.186567	Adjusted R^2	0.175424
$F(10, 730)$	16.74310	P-value(F)	1.79e-27
Log-likelihood	-306.2997	Akaike criterion	634.5994
Schwarz criterion	685.2874	Hannan-Quinn	654.1416

NOTA: El R^2 de una estimación similar a la anterior omitiendo *educp* y *educ* es 0.1717

SALIDA 3: OLS estimates using the 935 observations 1-935

Dependent variable: *cons*

	Coefficient	Std. Error	<i>t</i> -ratio	p-value
const	48.0705	35.4399	1.3564	0.1754
<i>renta</i>	-0.4594	0.2069	-2.2199	0.0267
<i>edad</i>	-0.4097	2.1563	-0.1900	0.8494
<i>edad2</i>	0.055	0.0324	1.6959	0.0899
<i>sexo</i>	-0.5713	1.9915	-0.2869	0.7743
<i>civil</i>	-0.2548	0.9057	-0.2813	0.7786
<i>sur</i>	0.1037	0.6348	0.1634	0.8703
<i>peso</i>	-0.1139	0.0568	-2.0047	0.0454
<i>peso</i> × <i>sexo</i>	-0.2221	0.2004	-1.1082	0.2681
<i>sur</i> × <i>sexo</i>	1.6185	1.8847	0.8587	0.3908
<i>resid</i> × <i>renta</i>	6.3914	2.1915	2.9164	0.0036

Mean dependent var	0.043929	S.D. dependent var	0.007224
Sum squared resid	0.037615	S.E. of regression	0.007178
R^2	0.028753	Adjusted R^2	0.015448
$F(10, 730)$	2.161070	P-value(F)	0.018389
Log-likelihood	-2506.447	Akaike criterion	5034.894
Schwarz criterion	5085.582	Hannan-Quinn	5054.436

SALIDA 4: TSLS, using the 935 observations 1-935

Dependent variable: *cons*

Instrumented: *renta*

Instruments: const *edad* *edad2* *sexo* *civil* *sur* *peso* *peso* × *sexo* *sur* × *sexo*
feduc *educ*

	Coefficient	Std. Error	z	p-value
<i>const</i>	48.0705	37.2741	1.2896	0.1972
<i>renta</i>	4.5944	2.1767	2.1107	0.0348
<i>edad</i>	-0.4097	2.2679	-0.1806	0.8566
<i>edad2</i>	0.0055	0.0341	0.1620	0.8713
<i>sexo</i>	-0.5713	2.0946	-0.2727	0.7850
<i>civil</i>	-0.2548	0.9526	-0.2675	0.7891
<i>sur</i>	0.1037	0.6676	0.1553	0.8766
<i>peso</i>	-0.1139	0.0598	-1.9060	0.0566
<i>peso</i> × <i>sexo</i>	-0.2221	0.2107	-1.0537	0.2920
<i>sur</i> × <i>sexo</i>	1.6185	1.9822	0.8165	0.4142

Mean dependent var	0.043929	S.D. dependent var	0.007224
Sum squared resid	0.041666	S.E. of regression	0.007549
R^2	0.000006	Adjusted R^2	-0.012306
$F(9, 731)$	1.546437	P-value(F)	0.127654

(a) En el modelo de referencia, ¿podemos afirmar que

$$E(\text{cons} | \text{renta}, \text{edad}, \text{sexo}, \text{civil}, \text{sur}, \text{peso}) = L(\text{cons} | \text{renta}, \text{edad}, \text{sexo}, \text{civil}, \text{sur}, \text{peso})?$$

Justifique la respuesta.

- (b) Dada la evidencia disponible y los supuestos establecidos en el problema, ¿podemos afirmar que *educp* y *educ* son instrumentos válidos para *renta*? Justifique la respuesta, indicando la evidencia en que ésta se basa.
- (c) ¿Podemos afirmar que la variable *renta* es endógena? Justifique la respuesta.
- (d) Queremos contrastar que el gasto en artículos deportivos es independiente del sexo del individuo. Escriba la hipótesis nula y la hipótesis alternativa en términos de los parámetros del modelo. Indique cómo construiría el estadístico de contraste y realice dicho contraste si dispone de la información necesaria. En caso contrario, indique qué información adicional precisaría.