PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
CIENCIAS SOCIALES
CICLO 2022-2

Fundamentos de Econometría
Práctica Dirigida 6

Profesor: Juan Palomino juan.palominoh@pucp.pe
Jefes de Práctica: Tania Paredes tania.paredes@pucp.edu.pe

Fecha: 22 – 10 – 2022

Parte I: Problema de Autocorrelación

1. El número de pequeños accidentes ocurridos en las calles de una ciudad (Y) y el número de coches matriculados en la misma (X) durante 10 años han sido los siguientes:

Y	X
25	510
27	520
28	528
32	540
33	590
36	650
38	700
40	760
41	800
45	870

Dado el modelo $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$, se pide:

a. Estimar la recta que exprese el número de accidentes ocurridos en función del número de coches matriculados.

Recordamos que dado el modelo regresión lineal $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$, podemos reexpresarlo como: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, en el que el valor de los parámetros estimados es $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$

Por lo que, procederemos a hallar los componentes de dicha expresión:

- $\boldsymbol{X'X} = \begin{pmatrix} 10 & 6\,468 \\ 6\,468 & 4\,335\,984 \end{pmatrix}$
- $(\boldsymbol{X'X})^{-1} = \begin{pmatrix} 2{,}843 & -0.0042 \\ -0.0042 & 0.0000065 \end{pmatrix}$

- $\boldsymbol{X'Y} = \begin{pmatrix} 345 \\ 230\ 674 \end{pmatrix}$
- $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} = \begin{pmatrix} 2.57 \\ 0.05 \end{pmatrix}$

Por lo que la ecuación de la recta estimada será:

$$\widehat{Y_t} = 2.5676 + 0.0494X_t$$

b. Calcular el estadístico Durbin-Watson y detectar la posible existencia de autocorrelación.

Calculamos los errores:

$$e = y - \hat{y} = y - X\hat{\beta}$$

y		$\hat{y}$		errores
25		27.7461989		-2.74619889
27		28.2398978		-1.2398978
28		28.6348569		-0.63485693
32		29.2272956		2.77270438
33		31.6957902		1.30420982
36	-	34.6579837	=	1.34201635
38		37.1264782		0.87352179
40		40.0886717		-0.08867168
41		42.0634673		-1.06346733
45		45.5193597		-0.51935971

Definimos el estadístico de Durbin-Watson

$$DW = \frac{\sum_{t=2}^{10}(e_t - e_{t-1})^2}{\sum_{t=1}^{10} e_t^2}$$

Hallamos los otros componentes pendientes

	e_t	e_t^2	e_{t-1}	$e_t - e_{t-1}$	$(e_t - e_{t-1})^2$
t=1	-2.74619889	7.54160832			
t=2	-1.2398978	1.53734655	-2.74619889	1.50630109	2.26894297
t=3	-0.63485693	0.40304332	-1.2398978	0.60504087	0.36607445
t=4	2.77270438	7.68788957	-0.63485693	3.40756131	11.6114741
t=5	1.30420982	1.70096325	2.77270438	-1.46849456	2.15647627
t=6	1.34201635	1.80100788	1.30420982	0.03780653	0.00142933
t=7	0.87352179	0.76304032	1.34201635	-0.46849456	0.21948715
t=8	-0.08867168	0.00786267	0.87352179	-0.96219347	0.92581628
t=9	-1.06346733	1.13096276	-0.08867168	-0.97479565	0.95022655
t=10	-0.51935971	0.26973451	-1.06346733	0.54410762	0.2960531
		22.8434591			18.7959802

Finalmente calculamos el estadístico

$$DW = \frac{\sum_{t=2}^{10}(e_t - e_{t-1})^2}{\sum_{t=1}^{10} e_t^2} = 0.8828$$

Y dado que

$$d_L = 0.8791$$
$$d_U = 1.3197$$

Hay autocorrelación serial positiva ya que

$$d < d_L$$

2. Dado un modelo lineal de consumo en función del PBI con los siguientes datos:

Yt	22	15	8	6	3	2	7
Xt	3	1	2	0	-2	-3	-1

Se pide que contraste la existencia de autocorrelación sabiendo que la regresión del modelo original por MCO produce los siguientes residuos:

et	4.63	3.21	-6.58	-3	-0.42	1.37	0.79

Teniendo en cuenta la información de la tabla

e_t	e_t^2	e_{t-1}	$e_t - e_{t-1}$	$(e_t - e_{t-1})^2$
4.63	21.44			
3.21	10.30	4.63	-1.42	2.01
-6.58	48.30	3.21	-9.79	95.84
-3	9	-6.58	3.58	12.82
-0.42	0.18	-3	2.58	6.66
1.37	1.88	-0.42	1.79	3.20
0.79	0.62	1.37	-0.58	0.34
	87.21			120.87

Se tiene que:

$$d = \frac{120.87}{82.21} = 1.39$$

Dado que d se aproxima a 2, se resuelve en que la perturbación aleatoria del modelo no tiene autocorrelación (no se rechaza la hipótesis nula).

Parte II: Problema de Endogeneidad

3. **El problema de endogeneidad**

a. Si se tiene un modelo $Y_i = \beta_1 + \beta_2 X_i + u_i$, en cual $\mathrm{Cov}(X_i, u_i) \neq 0$, demostrar que el estimado $\widehat{\beta_2}$ es sesgado e inconsistente.

(1) (a) Si tenemos el modelo $Y_i = \beta_1 + \beta_2 X_i + u_i$, su solución, estimando por desviaciones (lo cual hicimos en la PD1), sería:

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \beta_2 + \frac{\sum x_i u_i}{\sum x_i^2}$$

Y tomamos en consideración que la $\mathrm{Cov}(X_i, u_i) \neq 0$

✓ Demostración de sesgo:

Aplicando la ley de expectativas totales $E[E[a|b]] = E[a]$, tenemos:

$$E[\hat{\beta}_2] = E[E[\hat{\beta}_2|X]]$$

$$E[\hat{\beta}_2] = E\left[\beta_2 + E\left[\frac{\sum x_i u_i}{\sum x_i^2} \mid X\right]\right]$$

$$E[\hat{\beta}_2] = E\left[\beta_2 + \frac{1}{\sum x_i^2}\left(\sum x_i E[u_i|X]\right)\right]$$

$\underbrace{\quad}_{\neq 0}$

Ello porque si la $Cov(X_i, u_i) \neq 0$ entonces $E[u_i|X_i] \neq 0$

$\therefore E[\hat{\beta}_2] \neq \beta_2$

✓ Para demostrar la inconsistencia:

- Recordamos lo que dice la ley (L.G.N.) de los grandes números, que observa la convergencia en los promedios simples:
- Multiplicamos y dividimos el segundo término de $\hat{\beta}_2$ por $1/n$, para tener los promedios y tomamos Plim:

$$\hat{\beta}_2 = \beta_2 + \frac{\frac{1}{n}\sum x_i u_i}{\frac{1}{n}\sum x_i^2}$$

$$Plim(\hat{\beta}_2) = \beta_2 + \underbrace{\frac{Plim\left[\frac{1}{n}\sum x_i u_i\right]}{Plim\left[\frac{1}{n}\sum x_i^2\right]}}$$

Evaluaremos si esto es igual a cero

→ En el numerador:

$$\text{Plim}\left(\frac{1}{n}\sum x_i u_i\right) = \text{Plim}\left(\frac{1}{n}\sum (x_i-\bar{X})u_i\right)$$
$$= \text{Plim}\left(\frac{1}{n}\sum (x_i u_i - \bar{X}u_i\right)$$
$$= \text{Plim}\left(\frac{1}{n}\sum x_i u_i - \bar{X}\frac{1}{n}\sum u_i\right)$$
$$= \text{Plim}\left(\frac{1}{n}\sum x_i u_i\right) - \text{Plim}(\bar{X})\,\text{Plim}\left(\frac{1}{n}\sum u_i\right)$$
$$= E[x_i u_i] - E[x_i].E[u_i] \quad \text{por I.G.N.}$$
$$= \text{Cov}(x_i, u_i) \neq 0$$

→ En el denominador:

$$\text{Plim}\left(\frac{1}{n}\sum x_i^2\right) = \text{Plim}\left(\frac{(n-1)}{n}\cdot\frac{\sum (x_i-\bar{X})^2}{(n-1)}\right)$$
$$= \text{Plim}\left(\left(\frac{n-1}{n}\right) S_X^2\right)$$ → Varianza muestral

Se sabe que $\text{Plim}\, S_X^2 = \text{Var}(x_i)$ y $\text{Plim}\left(\frac{n-1}{n}\right) = \lim_{n\to\infty}\frac{n-1}{n} = 1$

→ Finalmente:

$$\text{Plim}(\hat{\beta}_2) = \beta_2 + \frac{\text{Cov}(x_i,u_i)}{\text{Var}(x_i)} \neq \beta_2 \Rightarrow \text{El estimador es inconsistente.}$$

b. En el siguiente modelo macroeconómico $Y_t = C_t + I_t$, donde $C_t = \beta_1 + \beta_2 Y_t + u_t$, donde Y_t=ingreso nacional, C_t= consumo, I_t= Inversión. Asumimos que la inversión no está correlacionada con u_t, por lo que $Cov(I_t, u_t) = 0$. Determine en este caso por qué se da la endogeneidad.

(b)

Del modelo $Y_t = C_t + I_t$, donde $C_t = \beta_1 + \beta_2 Y_t + u_t$

Datos:

- ✓ $Cov(I_t, u_t) = 0$
- ✓ Y_t = Ingreso nacional
- ✓ C_t = Consumo
- ✓ I_t = Inversión

→ En este caso se demostrará que hay correlación entre Y_t y u_t por causalidad simultánea.

$$Cov(Y_t, u_t) = Cov(C_t + I_t, u_t)$$

$$Cov(Y_t, u_t) = Cov(C_t, u_t) + Cov(I_t, u_t)$$ ($Cov(I_t, u_t) = 0$)

$$Cov(Y_t, u_t) = Cov(\beta_1 + \beta_2 Y_t + u_t, u_t)$$

$$Cov(Y_t, u_t) = Cov(\beta_1, u_t) + \beta_2 Cov(Y_t, u_t) + Var(u_t)$$

$$(1 - \beta_2) Cov(Y_t, u_t) = \underbrace{Cov(\beta_1, u_t)}_{=0} + \underbrace{Var(u_t)}_{\sigma_u^2}$$

$$Cov(Y_t, u_t) = \frac{\sigma_u^2}{(1 - \beta_2)}$$

Por lo que en la ecuación de consumo hay endogeneidad en el regresor.

c. Mencione otros dos ejemplos en los cuales pueda darse este mismo problema.

© Otras 2 formas en que se puede dar el problema de endogeneidad

✓ Error en la medición de la variable:

Por ejemplo, esto se puede dar al estimar un modelo sin intercepto

$$Y_i = \beta X_i^* + v_i \quad ; \; v_i \sim N(0, \sigma^2)$$

- $Y_i \rightarrow$ salario
- $X_i^* \rightarrow$ la habilidad del individuo i no observable
- Y el investigador propone una variable proxy X_i: educación

$$X_i = X_i^* + \varepsilon_i \quad ; \; \varepsilon_i: \text{error de medición}, \; \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

Reemplazando X_i en Y_i:

$$Y_i = \beta \underbrace{(X_i - \varepsilon_i)}_{X^*} + v_i = \beta X_i + \underbrace{v_i - \beta \varepsilon_i}_{u_i}$$

Se cumplirá que $E[u_i] = 0$ y $Var(u_i) = \sigma_v^2 + \beta^2 \varepsilon_i$

El modelo que incorpora la variable proxy cumplirá con todos los supuestos de MRLC, excepto que la $Cov(X_i, u_i) = 0$

Calculando $Cov(X_i, u_i)$:

$$\boxed{Cov(X_i, u_i)} = Cov(X_i^* + \varepsilon_i, \; v_i - \beta \varepsilon_i)$$

$$= \underbrace{Cov(X_i^*, v_i)}_{=0} + \underbrace{Cov(\varepsilon_i, v_i)}_{=0} - \beta \underbrace{Cov(X_i^*, \varepsilon_i)}_{=0}$$

$$- \beta \underbrace{Cov(\varepsilon_i, \varepsilon_i)}_{Var(\varepsilon_i)} = \boxed{-\beta \sigma_\varepsilon^2}$$

✓ Variables omitidas:

Por ejemplo, cuando se plantea estimar el salario y este dependerá de variables como educación y habilidad. No obstante, al no ser observable la habilidad, se omite y puede generarse endogeneidad porque la Covarianza omite educación y el término de error será diferente de cero, dado que la educación está correlacionada con la variable no observada de habilidad que está en el error.

4. Variables instrumentales

a. Para el caso del modelo de la parte 4.a, halle el estimador de variables instrumentales que soluciona el problema de endogeneidad.

a) Dado el caso de endogeneidad identificado en la parte 1.a, se propondrá el uso de una variable Z_i, la cual tendrá que cumplir dos condiciones y para poder obtener un estimador insesgado y consistente, a pesar del problema de endogeneidad.

Del modelo: $Y_i = \beta_1 + \beta_2 X_i + u_i \quad ; \quad Cov(X_i, u_i) \neq 0$

Y tenemos Z_i que cumple:

(1) $Cov(Z_i, X_i) \neq 0$ "condición de Relevancia"

(2) $Cov(Z_i, u_i) = 0$ "Condición de exogeneidad o exclusión"

Z_i será la variable instrumental

- Si calculamos $Cov(Z_i, Y_i)$ para hallar β_2

$$Cov(Z_i, Y_i) = Cov(Z_i, \beta_1 + \beta_2 X_i + u_i)$$

$$= \underbrace{Cov(Z_i, \beta_1)}_{=0} + \beta_2 \underbrace{Cov(Z_i, X_i)}_{\neq 0 \ (1)} + \underbrace{Cov(Z_i, u_i)}_{=0 \ (2)}$$

Despejando:

$$\beta_2 = \frac{Cov(Z_i, Y_i)}{Cov(Z_i, X_i)}$$

Para estimar β_2 usamos los siguientes estimadores consistentes de las covarianzas muestrales.

$$S_{XZ} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Z_i - \bar{Z})$$

$$S_{YZ} = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(Z_i - \bar{Z})$$

Finalmente, obtenemos el estimador consistente:

$$\hat{\beta}_{VI} = \frac{S_{YZ}}{S_{XZ}}$$

Continuado por análogos muestrales.

b. ¿Por qué es importante que el instrumento cumpla con las condiciones de relevancia y exogeneidad?

Es importante que se cumplan los supuestos de relevancia y exogeneidad para garantizar que los estimadores hallados sean insesgados y consistentes. Sin el cumplimiento de alguno de estos supuestos en el instrumento, podríamos generar estimadores con sesgos aún mayores que los obtenidos por MCO.

c. ¿A qué se le denomina como variables instrumentales débiles? ¿En qué situaciones podría darse este problema?

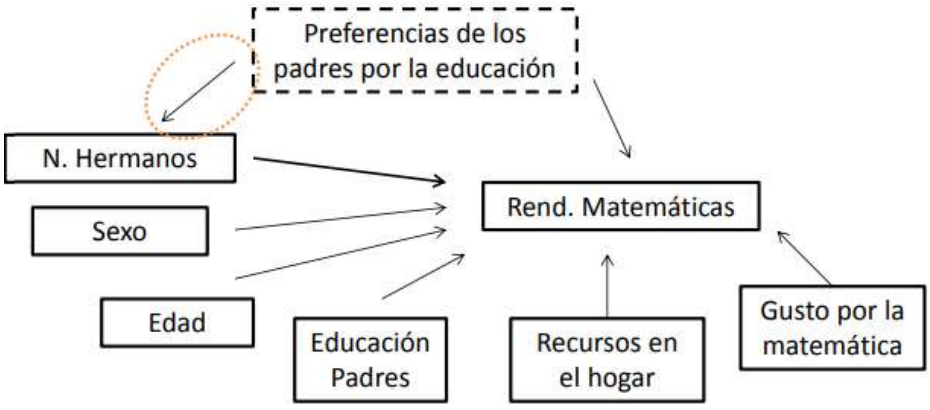
La efectividad del método de variables instrumentales recae en el cumplimiento de las condiciones de relevancia y exogeneidad. No obstante, puede darse el caso que cuando se evalué la condición de relevancia se observe una baja correlación (aunque diferente de cero) entre *Z* y *X*.

d. En el caso del modelo:

$$\text{rend\_mat}_i = \beta_1 + \beta_2 \text{nhermanos}_i + \beta_3 \text{Sexo}_i + \beta_4 \text{Edad}_i + u_i$$

Proponga dos instrumentos que podrían solucionar el problema de endogeneidad entre las variables nhermanos_i y variables no observadas

En este modelo podría darse un posible problema de endogeneidad de la variable número de hermanos y el término de variables no observables u_i, el cual puede incluir la preferencia de los padres por la educación.



Por lo que, puede darse el caso que la preferencia de los padres por la educación, determine el número de hijos que estos tengan, la cual también puede estar afectada también por las preferencias locales o regionales. Por lo que, para aliviar este problema pueden proponerse dos instrumentos: la diferencia de tasa global de fecundidad por distrito y el uso de radio, el cual indicaría la proporción de las mujeres entre 15-49 que se informaron de métodos de anticoncepción.

El primer instrumento buscará capturar algunas preferencias locales por fecundidad (que podría tener relación con el número de hermanos) y el segundo instrumento buscaría capturar los efectos de la planificación familiar.