Econometría II

Grado en Economía UB

Apuntes

0. REPASO ECONOMETRÍA I

EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

$Notación\ algebráica \quad y_i = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u_i \quad i = 1,2 \ldots n$

$Notación\ matricial \quad Y = X\beta + u$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{21} & \ldots & x_{k1} \\ 1 & x_{22} & & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{2n} & & x_{kn} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\hat{\beta}_{MQO} = (X'X)^{-1}(X'Y) \qquad Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

Modelo poblacional

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i$$

Modelo muestral

$$y_i = \widehat{\beta_1} + \widehat{\beta_2} x_{2i} + \widehat{\beta_3} x_{3i} + \cdots + \widehat{\beta_k} x_{ki} + \hat{u}_i$$

Error de predicción (residual, error o perturbación): $\hat{u}_i = e_i = y_i - \hat{y}_i$

Bondad de ajuste

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} \qquad \tilde{R}^2 = 1 - \left[\frac{N-1}{N-k} \cdot (1 - R^2)\right]$$

SUPUESTOS GENERALES DEL MODELO

- La relación entre las exógenas, la endógena y el término de perturbación debe ser **lineal.**
- Las variables explicativas (x) son independientes del término de perturbación: **E(u│x)=0**.
- Las variables x **no presentan errores** de medida
- El número de observaciones n debe ser igual o mayor que el número de regresores: **n≥k**
- Cada regresor debe ser linealmente independiente de los demás, no puede haber **multicolinealidad**
- Los parámetros $\boldsymbol{\beta_i}$ **son fijos** y deterministas (no aleatorios)
- No hay omisión de variables relevantes ni inclusión de variables irrelevantes

Supuestos del término de perturbación

- La perturbación uno tiene esperanza de 0 y sigue una **normal** independientemente distribuida

$$u \sim N(0, \sigma^2 \cdot Id) \qquad E(u) = 0$$

- La perturbación uno tiene una varianza constante (supuesto de **homocedasticidad**)

$$Var(u) = \sigma^2$$

- La perturbación u es aleatoria, no presenta **autocorrelación**. $Cov\,(u_i, u_j) = 0$

INTERPRETACIÓN ECONÓMICA DE LOS PARÁMETROS

Interpretaciones de los coeficientes en las diferentes formas funcionales:

1. $\hat{y}_i = \widehat{\beta_1} + \widehat{\beta_2}x_i$ el coef. $\widehat{\beta_2}$ nos indica el cambio medio esperado en y por cada unidad de x.

2. $\hat{y}_i = \widehat{\beta_1} + \widehat{\beta_2}\ln(x_i)$ el coef. $\widehat{\beta_2}$ nos indica que si x ↑ 1%, esperamos que y aumente en $\frac{\widehat{\beta_2}}{100}$ unidades

3. $\widehat{\ln(y_i)} = \widehat{\beta_1} + \widehat{\beta_2}x_i$ el coef. $\widehat{\beta_2}$ nos indica que si x ↑ 1u, esperamos que y aumente en $100 \cdot \hat{\beta}_2$%

4. $\widehat{\ln(y_i)} = \widehat{\beta_1} + \widehat{\beta_2}\ln(x_i)$ el coef. $\widehat{\beta_2}$ nos indica que si x ↑ 1%, esperamos que y aumente en $\hat{\beta}_2$%

DISTRIBUCIÓN DEL ESTIMADOR MCO Y PROPIEDADES

No sesgo: la esperanza del estimador es su parámetro $E(\widehat{\beta_i}) = \beta_i$

Eficiencia: la varianza del estimador es la mínima posible: $var(\widehat{\beta_{MCO}}) < var(\widehat{\beta_i})$

Consistencia: la estimación mejora al aumentar los datos $\lim_{n\to\infty}(\widehat{\beta_i}) = \beta_i$

Inferència sobre dels paràmetres del model

$$Intervalo\ de\ confianza \quad P\left(\widehat{\beta_j} - ee(\widehat{\beta_j}) \cdot t_{\substack{n-k \\ \alpha/2}} < \beta_j < \widehat{\beta_j} + ee(\widehat{\beta_j}) \cdot t_{\substack{n-k \\ \alpha/2}}\right) = 1 - \alpha$$

$$Contraste\ de\ significación\ individual \quad \begin{cases} Ho\ \beta_j = 0 \\ H_A\ \beta_j \neq 0 \end{cases} \quad EC = \frac{\widehat{\beta_j}}{ee(\widehat{\beta_j})} \sim t_{n-k} \ \rightarrow \ pv$$

$$Contraste\ de\ significación\ global \quad \begin{cases} Ho\ \beta_2 = \beta_3 = \cdots = 0 \\ H_A\ No\ Ho \end{cases} \quad EC = \frac{\frac{R^2}{k-1}}{\frac{1-R^2}{n-k}} = \sim F_{k-1,n-k} \ \rightarrow \ pv$$

Ejemplo Gretl:

```
Modelo 2: MCO, usando las observaciones 1-35
Variable dependiente: Notas_final

             coeficiente  Desv. típica  Estadístico t  valor p
  ------------------------------------------------------------
  const       0.537561     2.67569        0.2009       0.8421
  Edad        0.0156761    0.102044       0.1536       0.8789
  Academia    3.38150      0.659482       5.128        1.49e-05 ***
  Asistencia  0.0356565    0.0398292      0.8952       0.3776

Media de la vble. dep.   6.057143   D.T. de la vble. dep.   2.472708
Suma de cuad. residuos   107.2981   D.T. de la regresión    1.860438
R-cuadrado               0.483860   R-cuadrado corregido    0.433911
F(3, 31)                 9.687085   Valor p (de F)          0.000115
Log-verosimilitud       -69.26745   Criterio de Akaike      146.5349
Criterio de Schwarz      152.7563   Crit. de Hannan-Quinn   148.6825
```

1. VARIABLES EXPLICATIVAS ENDÓGENAS

INTRODUCCIÓN

Bajo el cumplimiento de las hipótesis básicas, los estimadores por MCO son no sesgados, consistentes y eficientes. Una hipótesis que podría incumplirse es la de que las **variables explicativas no deben estar correlacionadas con el término de perturbación**.

$$E(x_i, u_i) = E(X'U) = 0 \quad Variables\ explicativas\ exógenas$$
$$E(x_i, u_i) = E(X'U) \neq 0 \quad Variables\ explicativas\ endógenas$$

Si tenemos variables explicativas endógenas tendremos estimadores **inconsistentes** (**teorema de Mann- Wald**):

$$E(X'U) \neq 0$$
$$\Rightarrow \underset{N\to\infty}{p\lim}\left[\frac{1}{N}(X'U)\right] \neq 0$$
$$\Rightarrow \underset{N\to\infty}{p\lim}\,\hat{\beta}_{MQO} \neq \beta$$

CAUSAS DE LA CORRELACIÓN ENTRE VARIABLES EXPLICATIVAS (X) Y LA PERTURBACIÓN (U)

1.- Omisión de variable relevante no observable correlacionada con variables del modelo.

$$modelo\ real\ \rightarrow \quad Salario_i = \beta_1 + \beta_2 educación_i + \beta_3 capacidad_i + u_i$$

$$modelo\ estimado\ \rightarrow \quad Salario_i = \beta_1 + \beta_2 educación_i + v_i \quad on \quad v_i = \beta_3 capacidad_i + u_i$$

2.- Errores de medida a la hora de registrar una de las variables explicativas

$$modelo\ real\ \rightarrow\ Salario_i = \beta_1 + \beta_2 experiencia_i + u_i$$

$$modelo\ estimado \rightarrow \quad Salario_i = \beta_1 + (\beta_2 experiencia_i + e_i) + v_i$$
$$Salario_i = \beta_1 + \beta_2 experiencia_i + \beta_2 e_i + v_i$$

ESTIMACIÓN POR VARIABLES INSTRUMENTALES

Definiremos una **matriz Z** de dimensión N*K como la matriz de variables instrumentales. Dicha matriz debe cumplir que esté:

1.- No correlacionada con el término de error: $E(z_i, u_i) = 0$

2.- Correlacionada con las variables de la matriz X: $E(z_i, x_i) \neq 0$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{pmatrix} \quad X = \begin{pmatrix} 1 & X_{21} & X_{31} \cdots & X_{k1} \\ 1 & X_{22} & X_{32} \cdots & X_{k2} \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & \vdots \\ 1 & X_{2N} & X_{3N} \cdots & X_{kN} \end{pmatrix} \quad \text{Sigui } Z = \begin{pmatrix} 1 & Z_1 & X_{31} \cdots & X_{k1} \\ 1 & Z_2 & X_{32} \cdots & X_{k2} \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & \vdots \\ 1 & Z_N & X_{3N} \cdots & X_{kN} \end{pmatrix}$$

Así, imaginamos un modelo donde la variable x_3 tiene propiedades de endogeneidad:

$$y_i = \beta_1 + \beta_2 x_{2i} + \boldsymbol{\beta_3 x_{3i}} + u_i$$

$$X = [1, x_{2i}, x_{3i}] \qquad Z = [1, x_{2i}, z_{1i}]$$

Deberemos buscar z_{1i}, una variable instrumental de x_{3i}, que cumpla las siguientes propiedades:

$$E(z_{1i}, u_i) = E(Z'U) = 0 \ \ (\mathit{condición\ de\ exogeneidad})$$
$$E(z_{1i}, x_{3i}) = E(Z'X) \neq 0 \ \ (\mathit{condición\ de\ relevancia})$$

Buscamos la variable externa que más correlacione con x_3 y estimamos por variables instrumentales:

$$\hat{\beta}_{VI} = (Z'X)^{-1}(Z'Y)$$

La nueva estimación sí será **consistente:**

$$\hat{\beta}_{VI} = (Z'X)^{-1}(Z'Y) = (Z'X)^{-1}(Z'(X\beta + U)) = \beta + (Z'X)^{-1}(Z'U)$$

$$\underset{N\to\infty}{p\lim}\,\hat{\beta}_{VI} = \beta + \underset{N\to\infty}{p\lim}\left[\left(\frac{Z'X}{N}\right)^{-1}\right]\underset{N\to\infty}{p\lim}\left[\left(\frac{Z'U}{N}\right)\right] \qquad \underset{N\to\infty}{p\lim}\left[\frac{1}{N}(Z'U)\right] = 0 \ \Rightarrow \ \underset{N\to\infty}{p\lim}\,\hat{\beta}_{VI} = \beta$$

Sin embargo, **no será eficiente** ya que no tendrá la mínima varianza. Cuanto mayor sea la correlación entre la variable original y su instrumento, más eficiente será el estimador por VI. Si nos equivocamos en que había endogeneidad y estimamos por VI, tendremos estimaciones ineficientes, peores que con MCO. Debemos tener en cuenta dos cosas:

1.- Para comprobar si z_{1i} es un buen instrumento para x_{3i} habría que realizar la siguiente regresión auxiliar:

$$x_{3i} = \gamma_1 + \gamma_2 x_{2i} + \gamma_3 z_{1i} + u_i \qquad \mathit{con}\ \gamma_3\ \mathit{significativo}$$

2.- El cálculo de la R^2 después de la estimación por VI no tiene sentido y no puede valorarse.

Exemple: $\boxed{salari_i = \beta_1 + \beta_2 nivell\ educatiu_i + \beta_3 capacitat_i + v_i}$

Però *Capacitat* no es pot quantificar, de manera que es treu del model i s'especifica el següent model:

$\boxed{salari_i = \beta_1 + \beta_2 nivell\ educatiu_i + u_i}$ on $u_i = \beta_3 capacitat_i + v_i$

Model 1: OLS, using observations 1-428
Dependent variable: l_WAGE

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	-0.185197	0.185226	-0.9998	0.31795	
EDUC	0.108649	0.0143998	7.5451	<0.00001	***

On WAGE: Salari $ per hora
EDUC: Anys d'educació

En aquest cas, com molt probablement la variable inclosa *nivell educatiu* estigui correlacionada amb la variable exclosa *capacitat*, els estimadors per MQO que hem obtingut a l'exemple anterior seran inconsistents ja que :

$$E[nivell\ educatiu_i, u_i] \neq 0$$

$$\Downarrow$$

$X = [1 \ \ nivell\ educatiu_i]$ ➡ Variable *nivell educatiu* amb problemes d'endogenitat que caldrà instrumentalitzar

Caldrà definir una **variable instrumental z_{1i}** per la variable *nivell educatiu* que estigui altament correlacionada amb aquesta i que estigui incorrelacionada amb el terme de pertorbació....Per exemple, **nivell educatiu del pare. Per a què aquesta variable sigui un instrument vàlid, caldrà que**:

1º) $E[nivell\ educatiu\ pare_i, u_i] = 0$ ➡ Difícil de contrastar ja que u_i no és observable...Però podem suposar que aquesta variable no està relacionada amb la capacitat de l'individu...

2º) $E[nivell\ educatiu_i, nivell\ educatiu\ pare_i] \neq 0$

Aquesta segona condició és fàcilment contrastable...

$\boxed{nivell\ educatiu_i = \gamma_1 + \gamma_2 nivell\ educatiu\ pare_i + v_i}$

Model 2: OLS, using observations 1-428
Dependent variable: EDUC

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	10.2371	0.275936	37.0993	<0.00001	***
EDUC_FATHER	0.269442	0.0285863	9.4255	<0.00001	***

Mean dependent var	12.65888	S.D. dependent var	2.285376
Sum squared resid	1845.354	S.E. of regression	2.081302
R-squared	0.172560	Adjusted R-squared	0.170617
F(1, 426)	88.84076	P-value(F)	2.76e-19
Log-likelihood	-920.0246	Akaike criterion	1844.049
Schwarz criterion	1852.167	Hannan-Quinn	1847.255

Donat que la variable nivell educatiu del pare ha resultat ser rellevant a la regressió auxiliar efectuada, **podem dir que la variable nivell educatiu del pare és un instrument vàlid per la variable nivell educatiu!!!**

Este dosier está hecho para seguir la clase de prueba.

Si te apuntas al curso te enviaremos por correo el dosier entero con todos los temas que faltan, ejercicios y exámenes de años anteriores

Más información en:

www.estadistix.com

Y si tienes cualquier consulta,
escríbenos un whatsapp al 644310902