Tema 3. Extensiones del Modelo Lineal General.

Econometría I

Tema 3. Extensiones del MLG

- 1. Prácticas con los programas TSP / EViews
- 2. Transformaciones lineales
- 3. Multicolinealidad
- 4. Variables ficticias
- 5. Omisión / inclusión de variables explicativas relevantes / irrelevantes

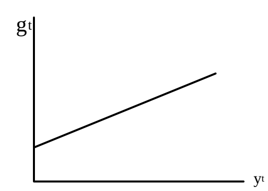
4.Variables ficticias:

- Dummy-binarias-categóricas-intervención
- Diferentes categorías no numéricas en la muestra.
- Diferentes:
  - días de la semana
  - trimestres, meses
  - razas
  - sexos

3

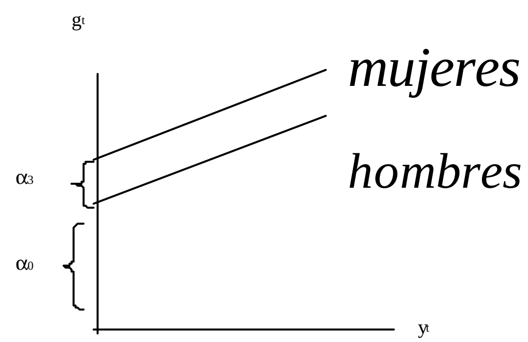
Ejemplo: gasto en ropa

$$g_t = a_0 + a_1 y_t + a_2 p_t + v_t$$



4

Si en la muestra hay personas de ambos sexos, podemos querer distinguir entre ellos (utilizar más información para mejorar el modelo).



5

$D_t = 1$ si y_t mujer

- $D_t = 0$ si y_t hombre

$$g_t = a_0 + a_1 y_t + a_2 p_t + a_3 D_t + w_t$$

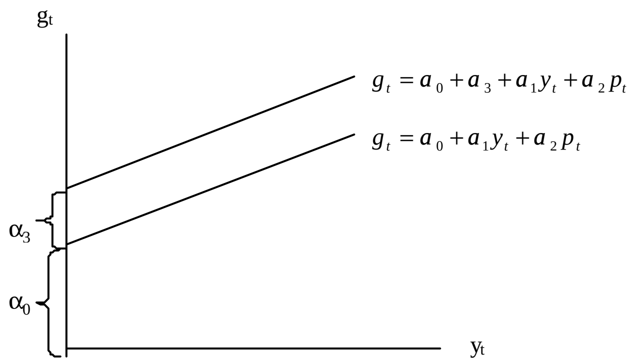
- equivalente a:

$$g_t = a_0 + a_3 + a_1 y_t + a_2 p_t + u_t$$

$$g_t = a_0 + a_1 y_t + a_2 p_t + u_t$$

6

Modelos restringido y sin restringir:



7

Con una ficticia permitimos diferente constante, pero forzamos la restricción de que el resto de los coeficientes y la varianza del error son iguales.

α_0 : común a los dos modelos

α_1 : común a los dos modelos.

α_2 : común a los dos modelos

σ^2_u : común a los dos modelos

8

Ejemplo uso ficticias para contrastar homogeneidad término constante:

$$y_t = a_1 D_{1t} + a_2 D_{2t} + x_t' b + u_t$$

$t=1, 2, ..., T_1$

$t=T_1+1, T_1+2, ..., T \qquad (T - T_1 = T_2)$

$D_{1t} = 1,\ D_{2t} = 0, \quad \text{si } 1 \leq t \leq T_1$

$D_{1t} = 0,\ D_{2t} = 1, \quad \text{si } T_1 < t \leq T$

9

Matriz observaciones:

$$X = \begin{pmatrix} 1_{T1} & 0_{T1} & X_1 \\ 0_{T2} & 1_{T2} & X_2 \end{pmatrix}$$

10

Contraste de homogeneidad:

$$\left.\begin{array}{l} H_0 : \alpha_1 = \alpha_2 \\ H_1 : \alpha_1 \neq \alpha_2 \end{array}\right\} \quad \text{Contraste F}$$

Especificación alternativa:

$$y_t = a_0 + a_2 D_{2t} + x_t^{\prime} b + u_t$$

$$\left.\begin{array}{l} H_0 : \alpha_2 = 0 \\ H_1 : \alpha_2 \neq 0 \end{array}\right\} \quad \text{Contraste t mucho más sencillo}$$

11

$$y_t = a_0 + a_2 D_{1t} + x_t^{\prime} b + u_t$$

$$\left.\begin{array}{l} H_0 : \alpha_2 = 0 \\ H_1 : \alpha_2 \neq 0 \end{array}\right\} \quad \text{Contraste t}$$

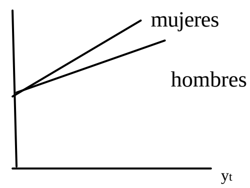
Nota:
no usar constante, D_1 y D_2 por multicolinealidad perfecta.

12

Heterogeneidad de pendientes (con variables ficticias)

$$g_t = a_0 + a_1 y_t + a_2 p_t + u_t$$

$$g_t = a_0 + a_4 y_t + a_2 p_t + u_t$$



13

$H_0 : \alpha_1 = \alpha_4$

$H_1 : \alpha_1 \neq \alpha_4$

F general, complicada

Más sencillo usar una variable ficticia de pendiente.

$$g_t = a_0 + a_1 y_t + a_2 p_t + a_3 D_t y_t + v_t$$

$H_0 : \alpha_3 = 0$

$H_1 : \alpha_3 \neq 0$

$$D_t = \begin{cases} 1 \text{ mujer} \\ 0 \text{ hombre} \end{cases}$$

14

Contraste: de significatividad t-student

$$\text{Datos} = \begin{pmatrix} 1_{T1} & y_{T1} & p_{T1} & 0_{T1} \\ 1_{T2} & y_{T2} & p_{T2} & y_{T2} \end{pmatrix} \begin{array}{l} \longrightarrow \text{Hombres} \\ \longrightarrow \text{Mujeres} \end{array}$$

15

Comentarios sobre variables ficticias:

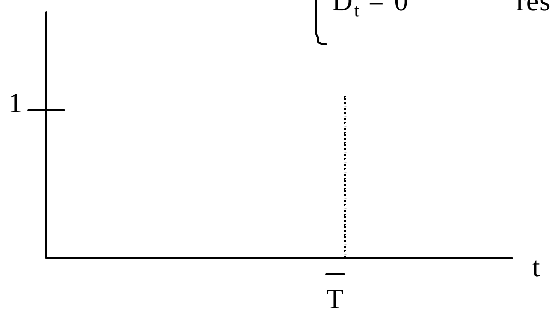
1. Estimación, igual que con variables continuas (MCO)
2. Inferencia (contraste hipótesis), igual que con variables continuas: t, F y sus variantes.
3. Predicción (puntual, intervalos), igual que con variables continuas.
4. Variables ficticias frecuentes: algunas encuestas se codifican exclusivamente con variables ficticias.
5. Otros tipos de variables ficticias: (a medida)

16

Otros tipos de variables ficticias: (a medida)

A) Impulso :

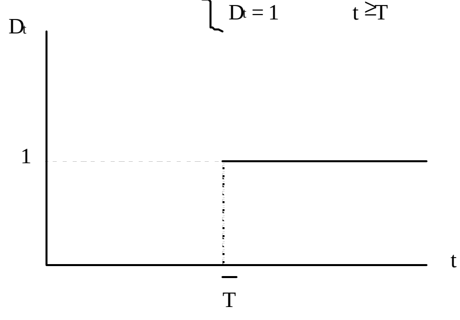
$$\begin{cases} D_t = 1 & t = \overline{T} \\ D_t = 0 & \text{resto} \end{cases}$$



17

A) Escalón

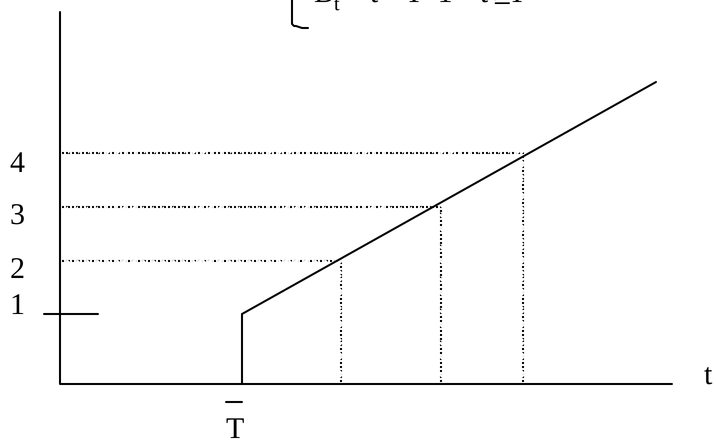
$$\begin{cases} D_t = 0 & t < \overline{T} \\ D_t = 1 & t \geq \overline{T} \end{cases}$$



18

A) Rampa

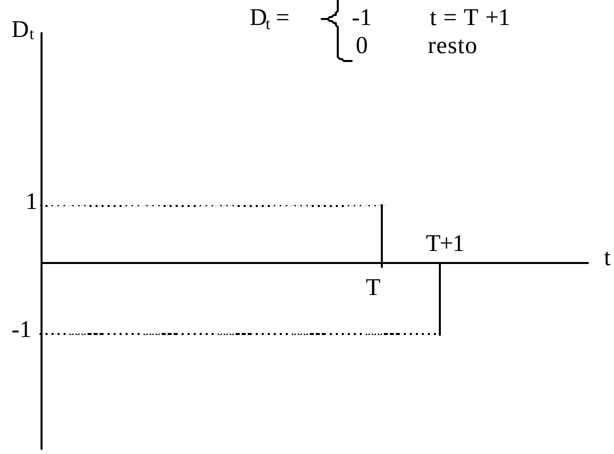
$$D_t = \begin{cases} 0 & t < \overline{T} \\ t - T + 1 & t \geq \overline{T} \end{cases}$$



19

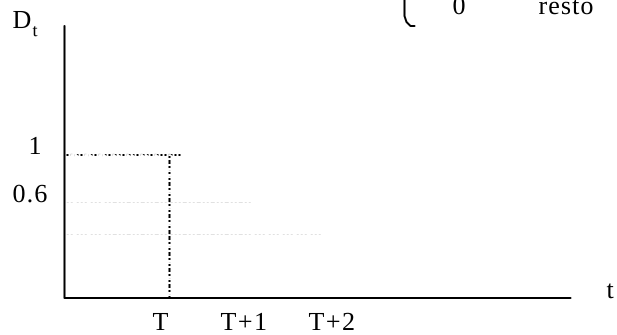
A) Impulso compensado

$$D_t = \begin{cases} 1 & t = T \\ -1 & t = T + 1 \\ 0 & \text{resto} \end{cases}$$



20

$$D_t = \begin{cases} 1.0 & t = T \\ 0.6 & t = T+1 \\ 0.3 & t = T+2 \\ 0.2 & t = T+3 \\ 0.15 & t = T+4 \\ 0 & \text{resto} \end{cases}$$



21

1) <u>Variables ficticias para permitir diferentes medias</u>

$$y_t = a_0 + u_t \qquad\qquad \hat{a}_0 = \sum_{t=1}^{T} \frac{y_t}{T} = \bar{y}$$

Y_t = salario;   muestra: negros y blancos.

$$y_t = a_0 + a_1 D_t + u_t$$

$$D_t = \begin{cases} 1 \text{ negros} \\ 0 \text{ blancos} \end{cases}$$

α_1 = diferencia de salario medio entre negros y blancos.

22

Tres categorías : negros, blancos y orientales.

$$D_{1t} = \begin{cases} 1 \text{ negro} \\ 0 \text{ no negro} \end{cases} \qquad D_{2t} = \begin{cases} 1 \text{ oriental} \\ 0 \text{ no oriental} \end{cases}$$

$$y_t = a_0 + a_1 D_{1t} + a_2 D_{2t} + u_t$$

α_0 = salario medio blancos

α_1 = diferencia salario medio negros

α_2 = diferencia salario medio orientales con respecto al salario medio blancos.

23

1) Contraste de cambio estructural usando variables ficticias

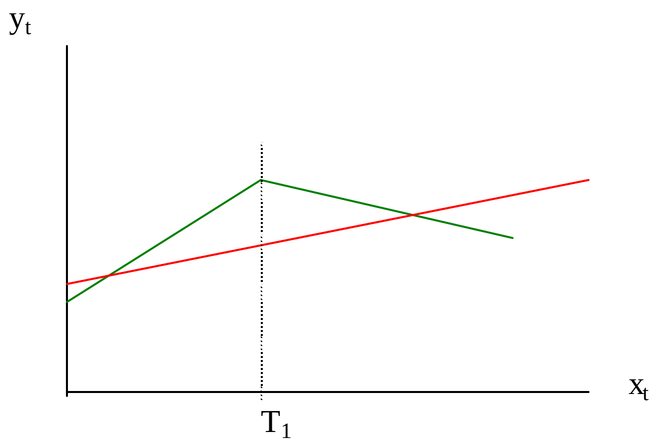
- Importante en trabajos saber si hay estabilidad de coeficientes
- en distintas submuestras o períodos de tiempo.

  Si hay estabilidad: mayor eficiencia (ausencia de cambio estructural)

  Si no hay estabilidad: necesario partir la muestra, para evitar sesgo (cambio estructural).

24

Ejemplo:



25

- Si no hay cambio estructural:

$t = 1, ..., T$

$$y_t = x_t' b + e_t$$

Estimación más eficiente. Se usa toda la muestra.

- Si hay cambio estructural:

$$y_t = x_t' b_1 + u_t \qquad\qquad y_t = x_t' b_2 + v_t$$

$t = 1, ..., T_1$ $\qquad\qquad$ $t = T_1+1,...,T_2$

26

Dos submuestras, la estimación será insesgada en cada una de las dos. (Se puede plantear en varias submuestras).

- Tradicionalmente se haría con contraste de cambio
- estructural de Chow o bien, usando variables ficticias,
- por ejemplo:

27

$$y_t = a_0 + a_1 x_{1t} + a_2 x_{2t} + u_t$$

(modelo restringido) (1)

$$y_t = a_0 + a_1 x_{1t} + a_2 x_{2t} + a_3 D_t + a_4 D_t x_{1t} + a_5 D_t x_{2t} + v_t$$

(2)

$$D_t = \begin{cases} 1 \text{ en } 2^a \text{ parte de la muestra} \\ 0 \text{ en } 1^a \text{ parte de la muestra} \end{cases}$$

28

Comentarios:

1) La varianza del error se supone constante: $\sigma^2_u = \sigma^2_v$
2) Generalizable a varios grupos (más de dos), usando tantas variables ficticias adicionales como números de grupos adicionales por número de parámetros del modelo restringido.
3) Fácil implementación con paquetes de regresión.

5. Omisión/inclusión de variables relevantes/irrelevantes

- Fallo de S.2.
-
- Caso 1: omisión de variables relevantes
- Caso 2: inclusión de variables irrelevantes.

31

CASO 1: Omisión de variables relevantes

Mecanismo que genera los datos: mgd

$y = X\beta + u$

$E(u) = 0$

$V(u) = \sigma^2_u * I_T$

Estimamos

$$y = X_0 b^0 + v$$

32

$$X = (X_0 ; Z)$$

- $X_0 : T \times r$

$Z : T \times (k - r)$

$$\hat{b}^0 = (X_0' X_0)^{-1} X_0' y$$

33

$$E(\hat{b}^0) = b^0 + (X_0' X_0)^{-1} X_0' Z b_z$$

- Sesgado, excepto si
-
- a) $X'_0 Z = 0$: ortogonales
- b) $\beta_Z = 0$

- Sesgo: $(X_0' X_0)^{-1} X_0' Z b_z$

34

CASO 2: Inclusión de variables irrelevantes

- mgd
- $y = X\beta' + u$

Estimamos

$$y = X_0 b^0 + v$$

$$y = Xb^1 + Z b^2 + v$$

35

$\beta^2 \equiv 0$, $X_0 = (X, Z)$

$$\hat{b}^0 = (X_0' X_0)^{-1} X_0' y$$

$$E(\hat{b}^0) = E\begin{pmatrix} \hat{b}^1 \\ \hat{b}^2 \end{pmatrix} = \begin{pmatrix} b^1 \\ 0_s \end{pmatrix}$$

36

MCO : insesgado pero <u>ineficiente</u>

<u>(varianza grande)</u>

$\Rightarrow$ pueden resultar no significativas

las variables que si lo son.

Vimos resultado general de que :

$$V(\hat{b}_R) < V(\hat{b})$$

37

- Conclusión: <u>ambos</u> errores
- de especificación son muy
- perjudiciales, y han
- de evitarse si es posible.

38

3. Multicolinealidad :

- (Cap. 10, p. 344 Novales)

  - Naturaleza y causas
  - Efectos sobre MCO
  - Detección
  - Posibles remedios

39

SUPUESTOS del MLG

- S.1. Relación estocástica
- S.2. Ausencia error especificación
- S.3. E(u) = 0
- S.4. V(u) = F^2_u I
- S.5. Linealidad en parámetros
- S.6. x deterministas
- S.7. $x_1 \ldots x_k$ linealmente independientes.

40

- (Problema datos)
- Multicolinealidad exacta:
- Fallo S.7.
- Ejemplo:

$$P_i = b_0 + b_1 K_i + b_2 NOF_i + b_3 NOTE_i + b_4 NOTO_i + u_i$$

41

P_i : Producción total

K_i: Cantidad de capital

NOF_i: número de obreros fijos

$NOTE_i$: número obreros temporales

$NOTO_i$: número total de obreros.

$NOTOi = NOF_i + NOTE_i$

42

- $\Rightarrow x_1, ..., x_k$ linealmente dependientes
- $\Rightarrow$ $|X'X| = 0$ $\Rightarrow$ infinitas soluciones para

$$\hat{b} = (X'X)^{-1}X'y$$

43

Multicolinealidad exacta:

- No se pueden estimar todos los parámetros (no hay información suficiente en la muestra),
- pero sí combinaciones lineales de ellos,
- o se puede eliminar algún parámetro (y su correspondiente variable).

44

Ejemplo:

$$y_t = b_1 + b_2 x_{2t} + b_3 x_{3t} + u_t$$

Si $x_{3t} = 8x_{2t}$

$$y_t = b_1 + b_2 x_{2t} + b_3 l x_{2t} + u_t$$

$$y_t = b_1 + (b_2 + b_3 l) x_{2t} + u_t$$

45

- 1.) Eliminamos x_{3t}
- 2.) Estimamos por MCO
- $\$_2 + \$_3 8$

46

Multicolinealidad exacta:

- Se suele deber a la existencia de una restricción entre los datos (p.ej.: identidad contable).
- Se puede resolver eliminando una variable explicativa (o más) o aumentando la muestra.

47

Multicolinealidad aproximada:

- Muy frecuente en datos económicos.

$$x_{it} \approx d_1 + d_2 x_{2t} + \dots + d_{i-1} x_{i-1t} + d_{i+1} x_{i+1,t} + \dots + d_{k-1} x_{kt}$$

48

- Se nota por R^2_i de esta regresión AUXILIAR alto

- $(R^2_i \geq 0.9) \Rightarrow \quad |X´X| \approx 0,$
-
- aunque $\quad |X´X| \neq 0,$

- pero pequeño $\Rightarrow (X´X)^{-1}$ grande.

49

Efectos de la multicolinealidad sobre los estimadores MCO:

- 1) Exacta
- $$\left| X´X \right| = 0 \Rightarrow \hat{b} : \infty \ soluciones$$
- 2) Aproximada:
- Inestable numéricamente:
- Pequeños cambios en los valores numéricos de los datos hacen cambiar mucho los valores numéricos de las estimaciones

50

Impreciso :

$$V(\hat{b}) = s_u^2 (X'X)^{-1}$$

Grande.

51

Definiendo:

$$ST_i = \sum_1^T (x_{it} - \bar{x}_i)^2$$

$R_i^2 = R^2$ de la regresión de x_i sobre las demás x.

$$V(\hat{b}_i) = \frac{s_u^2}{ST_i(1-R_i^2)}$$

52

- Menor cuanto : más varíe x_i (mayor ST_i)
- menor colinealidad (menor R^2_i)
- Todavía ELIO

53

Detección de la multicolinealidad:

- 1) Exacta
- $X'X = 0$, fácil; programa nos informa.
- 2) Aproximada
- Más difícil de detectar.
- Síntoma típico: R^2 alto, F alto y t-student bajos.

54

Detección

- 1.) Regresiones de cada variable explicativa sobre las demás. Habrá multicolinealidad si algún $R^2_i \geq 0.9$.
- 2.) Métodos basados en el tamaño de X´X
- Examinar determinante: sensible a unidades de medida

55

- Número de condición: como determinante de una matriz simétrica es igual al producto de sus valores propios $\Rightarrow$ examinar valores propios, pero dependen de unidades $\Rightarrow$ necesario normalizar antes de calcularlos. Dividir cada variable por su desviación típica.

56

$$\sqrt{\sum_{t=1}^{T} (x_{it} - \overline{x}_i)^2}$$

Valor propio:

Sea A matriz cuadrada nxn, entonces una constante λ y un vector x, nx1 no nulo, que satisfagan el sistema de ecuaciones $Ax = \lambda x$ se llaman, respectivamente, valor propio y vector propio de la matriz A

57

- Si examinamos el cociente entre el valor propio más grande λ_{max} y el menor λ_{min}, y hay gran diferencia, problema de multicolinealidad aproximada.

$$n^{o}\ condición = \sqrt{\frac{l_{\max}}{l}}$$

- Si nº condición $\geq 25 \Rightarrow$ Multicolinealidad aproximada.

58

2. Cambios de escala y origen.

$$y_t = b_1 + b_2 x_{2t} + \ldots + b_k x_{kt} + u_t$$

$$y_t^* = b_1^* + b_2^* x_{2t}^* + \ldots + b_k^* x_{kt}^* + u_t^*$$

– Cambiar unidades de medida y/o origen de medida. ¿Cómo afecta a estadísticos relevantes?

$$\hat{b}\,,\ \hat{u}\,,\ \hat{u}'\hat{u}\,,\ \hat{R}^2\,,\ \overline{R}^2$$

59

- CASO 1: <u>Cambios de escala</u>
- En un modelo lineal general se multiplican todas las variables por una constante, posiblemente diferente.

$l, l_i \neq 0, \prec \infty$

$$X^* = \begin{pmatrix} l_1 x_{11} & l_2 x_{21} & . & . & . & l_k x_{k1} \\ l_1 x_{12} & l_2 x_{22} & . & . & . & l_k x_{k2} \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ l_1 x_{1T} & l_2 x_{2T} & . & . & . & l_k x_{kT} \end{pmatrix}$$

60

$$y^* = \begin{pmatrix} l y_1 \\ l y_2 \\ . \\ . \\ . \\ l y_T \end{pmatrix}$$

$$y^* = x^* \, \$ + u^*$$

61

$$\hat{b}_i^{\,*} = \left(\frac{l}{l_i} \right) \hat{b}_i$$

- $\hat{b}_i^{\,*} = l\, \hat{b}_i$ cambio escala var. dependiente
- $\hat{b}_i^{\,*} = \frac{1}{l_i} \hat{b}_i$ cambio escala vars. explicativas
- ¿Qué sucede con el resto de los estadísticos?

62

$$\hat{u}^{*} = l\,\hat{u}$$

$$SR^{*} = l^{2} SR$$

$$SR^{*} = \hat{u}^{*\prime}\hat{u}^{*} = (l\,\hat{u})'(l\,\hat{u}) = l^{2}\,\hat{u}'\hat{u} = l^{2} SR$$

$$ST^{*} = l^{2} ST$$

63

$$R^{2*} = R^{2}$$

$$\bar{R}^{2*} = \bar{R}^{2}$$

$$R^{2*} = 1 - \frac{SR^{*}}{ST^{*}} = 1 - \frac{l^{2} SR}{l^{2} ST} = R^{2}$$

$$\hat{s}_{u}^{2*} = l^{2}\,\hat{s}_{u}^{2}$$

64

- CASO 2: Traslación de las variables explicativas (cambios origen medida)

$$y = b_1^* + X_2^* b_2^* + u^*$$

- $X^*_2 = X_2 + B$

65

$$B = \begin{pmatrix} b_2 & b_3 & . & . & . & b_k \\ b_2 & b_3 & . & . & . & b_k \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ b_2 & b_3 & . & . & . & b_k \end{pmatrix} = 1_T b'$$

66

<u>Resultado</u>:
Todos los parámetros iguales, excepto la constante.

$$\hat{b}_2^* = \hat{b}_2 \; pero \; \hat{b}_1^* \neq \hat{b}_1$$

$$\hat{u}^* = \hat{u}$$

$$SR^* = SR$$

$$R^{2*} = R^2$$

$$\overline{R}^{2*} = \overline{R}^2$$

$$\hat{s}_u^{2*} = \hat{s}_u^2$$

(Similitud con desviaciones respecto de las medias)

67

- **6. <u>Estimación MCO con datos en desviaciones con respecto a la media</u>:**
  Partimos del MLG con término independiente

$$y_t = b_1 + b_2 x_{2t} + b_3 x_{3t} + ... + b_k x_{kt} + u_t$$

- Y, tras estimarlo:

$$y_t = \hat{b}_1 + \hat{b}_2 x_{2t} + \hat{b}_3 x_{3t} + ... + \hat{b}_k x_{kt} + \hat{u}_t$$

- Calculamos las medias muestrales

- $$\overline{y} = \hat{b}_1 + \hat{b}_2 \overline{x}_2 + \hat{b}_3 \overline{x}_3 + ... + \hat{b}_k \overline{x}_k$$

(La media de los residuos es cero, debido a la existencia de término independiente)

68

- Restando estas dos últimas expresiones:

$$y_t - y = \hat{b}_2(x_{2t} - x_2) + \hat{b}_3(x_{3t} - x_3) + \ldots + \hat{b}_k(x_{kt} - x_k) + \hat{u}_t$$

Obtenemos un nuevo modelo, que carece de término independiente y en el que las variables están medidas en diferencias con respecto a sus promedios muestrales. Aunque, los residuos y los coeficientes estimados coinciden con los del modelo inicial.

69

- INTUICIÓN

➡ El interés se centra en analizar el impacto que las variables explicativas tienen sobre la variable endógena, en cuyo caso la constante nos sirve para garantizar que los promedios muestrales de ambos miembros del modelo econométrico coincidan. En definitiva, el modelo en desviaciones respecto a la media es todo cuanto se requiere.

➡ Una vez estimado éste, recuperamos la constante utilizando

$$\hat{b}_1 = y - \hat{b}_2 x_2 - \hat{b}_3 x_3 - \ldots - \hat{b}_k x_k$$

70

- <u>Coeficiente de determinación</u>

  En un modelo $y_t = x_t\beta + u_t$, donde las variables se hallan en desviaciones respecto a la media, R^2 puede escribirse como:

$$R^2 = \frac{\hat{b}'\tilde{X}_2'\tilde{y}}{\tilde{y}'\tilde{y}} = \frac{y'X(X'X)^{-1}X'y}{y'y} = \frac{\left(\sum_1^T y_t x_t'\right)\left(\sum_1^T x_t x_t'\right)\left(\sum_1^T x_t y_t\right)}{\sum_1^T y_t^2}$$

Siendo $\tilde{X}_2$ la matriz cuyas columnas son las observaciones de las variables explicativas (excepto la constante) en desviaciones respecto a sus medias y $\tilde{y}$ el vector de desviaciones de la variable endógena con respecto a su media muestral.

Coeficiente de determinación ajustado, igual.

71

- varianza residual, igual.
- Coeficientes estimados y residuos estimados, numéricamente idénticos.

72