TEMA 1: EXTENSIONES DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE: VARIABLES FICTICIAS Y CAMBIO ESTRUCTURAL.

Wooldridge: Capítulos 6 (apartado 6.1) y 7

Gujarati: Capítulos 9 (apartado 9.8), 10 y 12

---

1. VARIABLES FICTICIAS

- **Las variables ficticias (o binarias o "dummy") las empleamos para recoger información cualitativa: ser hombre o mujer, estar o no estar casado, que una empresa sea del sector industrial o del sector servicios, que cotice o no en bolsa, etc.**

- **Las variables ficticias se emplean en los modelos de regresión cuando queremos ver si el efecto de alguna/s de las X sobre *Y* varía según alguna característica de la población (sexo, raza, tamaño de la empresa, tipo de período temporal, etc.).**
- **En definitiva, las variables ficticias se emplean para analizar y modelizar "cambios estructurales" en los parámetros de los modelos.**

- **Las variables ficticias toman valor 1 en una categoría y valor 0 en el resto. Por ejemplo:**

$$Mujer = \begin{cases} 1 \ si \ \ individuo \ \ es \ \ mujer \\ 0 \ si \ \ \ es \ \ \text{hom}bre \end{cases}$$

$$Hombre = \begin{cases} 1 \ si \ \ individuo \ \ es \ \ \text{hom}bre \\ 0 \ si \ \ \ es \ \ mujer \end{cases}$$

$$Pequeña = \begin{cases} 1 \ si \ \ empresa \ \ es \ \ pequeña \\ 0 \ \ en \ \ caso \ \ contrario \end{cases}$$

$$Mediana = \begin{cases} 1 \ si \ \ empresa \ \ es \ \ mediana \\ 0 \ \ en \ \ caso \ \ contrario \end{cases}$$

$$Grande = \begin{cases} 1 \ si \ \ empresa \ \ es \ \ grande \\ 0 \ \ en \ \ caso \ \ contrario \end{cases}$$

CAMBIO ESTRUCTURAL

- **Denominamos cambio estructural a las modificaciones en el valor de los parámetros del modelo para diferentes subpoblaciones.**

- **Los cambios estructurales se pueden deber a las más diversas circunstancias: diferencias en los gustos de los individuos, cambios en la política económica, cambios en las condiciones socioeconómicas, etc.**

MODELIZANDO EL CAMBIO ESTRUCTURAL

A) Efecto aditivo: Empleamos las variables ficticias para modelizar cambios en el término constante del modelo.

Veámoslo con un ejemplo. Consideremos el modelo de regresión múltiple:

$$w_i = \beta_0 + \beta_1 edu_i + \beta_2 mujer_i + \varepsilon_i \quad i = 1,..,n$$

donde:

w_i = salario

edu_i = años de educación

$$\textbf{mujer}_i = \begin{cases} 1 & si \quad i \quad es \quad mujer \\ 0 & si \quad i \quad es \quad \text{hom}bre \end{cases}$$

Tenemos que:

$$E(w_i | edu_i, mujer_i) = \beta_0 + \beta_1 edu_i + \beta_2 mujer_i$$

con lo cual:

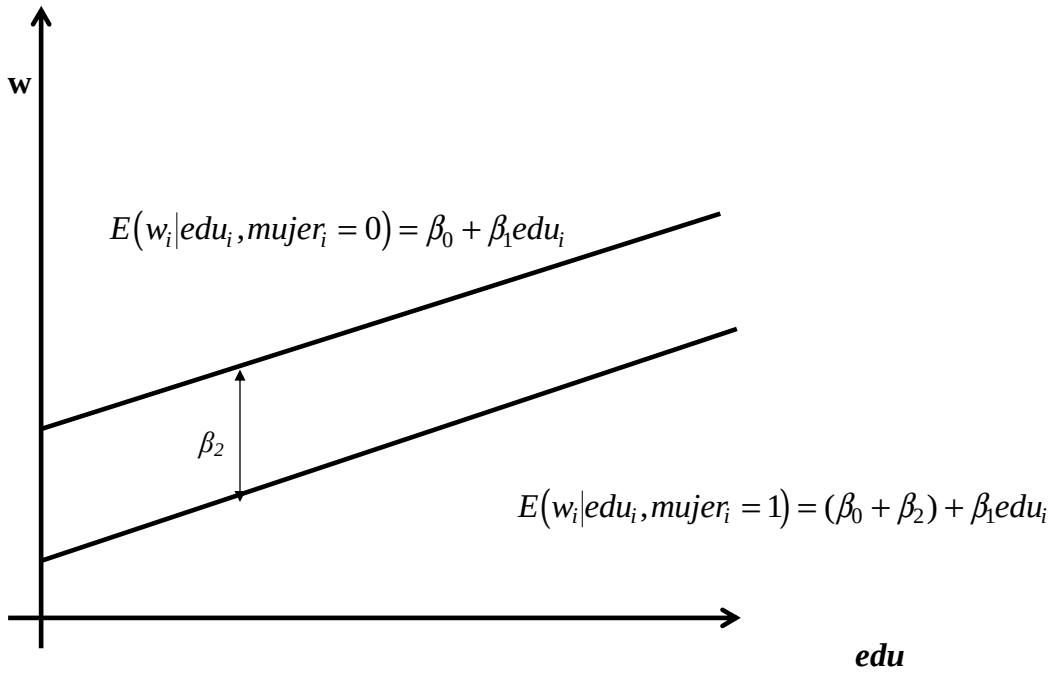
$$E\left(w_i \middle| edu_i, mujer_i = mujer\right) = E\left(w_i \middle| edu_i, mujer_i = 1\right) = (\beta_0 + \beta_2) + \beta_1 edu_i$$

$$E\left(w_i \middle| edu_i, mujer_i = \text{hom}bre\right) = E\left(w_i \middle| edu_i, mujer_i = 0\right) = \beta_0 + \beta_1 edu_i$$

y

$\beta_2 = E(w_i | edu_i, mujer) - E(w_i | edu_i, \text{hom}bre) \rightarrow$ **es la diferencia, en media, entre el salario de una mujer y el de un hombre, para un mismo nivel educativo.**

Suponiendo $\beta_2 < 0$



OTRAS DOS FORMULACIONES ALTERNATIVAS DE ESTE MISMO MODELO SERÍAN:

1.- $w_i = \alpha_0 + \alpha_1 edu_i + \alpha_2 \,\text{hom}bre_i + \varepsilon_i \quad i = 1,..,n$ **con** $\boldsymbol{hombre_i} = \begin{cases} 1 & si \;\; i \;\; es \;\; \text{hom}bre \\ 0 & si \;\; i \;\; es \;\; mujer \end{cases}$

Tenemos que: $E(w_i|edu_i, \text{hom}bre_i) = \alpha_0 + \alpha_1 edu_i + \alpha_2 \,\text{hom}bre_i$

con lo cual:

$E(w_i|edu_i, mujer) = E(w_i|edu_i, \text{hom}bre_i = 0) = \alpha_0 + \alpha_1 edu_i$

$E(w_i|edu_i, \text{hom}bre) = E(w_i|edu_i, \text{hom}bre_i = 1) = (\alpha_0 + \alpha_2) + \alpha_1 edu_i$

$\alpha_2 = E(w_i|edu_i, \text{hom}bre) - E(w_i|edu_i, mujer) \rightarrow$ **es la diferencia, en media, entre el salario de un hombre y el de una mujer, para un mismo nivel educativo.**

Obviamente:

$\alpha_1 = \beta_1$

$\alpha_0 = \beta_0 + \beta_2$

$\alpha_0 + \alpha_2 = \beta_0$

2.- $w_i = \delta_1 edu_i + \delta_2 mujer_i + \delta_3 \text{hombre}_i + \varepsilon_i \quad i = 1,..,n$

Tenemos que: $E(w_i|edu_i, mujer_i, \text{hombre}_i) = \delta_1 edu_i + \delta_2 mujer_i + \delta_3 \text{hombre}_i$

con lo cual:

$E(w_i|edu_i, mujer) = E(w_i|edu_i, mujer_i = 1, \text{hombre}_i = 0) = \delta_2 + \delta_1 edu_i$

$E(w_i|edu_i, \text{hombre}) = E(w_i|edu_i, mujer_i = 0, \text{hombre}_i = 1) = \delta_3 + \delta_1 edu_i$

$\delta_3 - \delta_2 = E(w_i|edu_i, \text{hombre}) - E(w_i|edu_i, mujer) \rightarrow$ **es la diferencia, en media, entre el salario de un hombre y el de una mujer, para un mismo nivel educativo.**

Obviamente:

$$\delta_1 = \alpha_1 = \beta_1$$

$$\delta_2 = \alpha_0 = \beta_0 + \beta_2$$

$$\delta_3 = \alpha_0 + \alpha_2 = \beta_0$$

<u>IMPORTANTE</u>: No sería válido (habría multicolinealidad exacta) un modelo como:

$$w_i = \gamma_0 + \gamma_1 edu_i + \gamma_2 mujer_i + \gamma_3 \text{hom}bre_i + \varepsilon_i$$

ya que $mujer_i + \text{hom}bre_i = 1 \ \forall i$

¿Cómo contrastaríamos si existen diferencias en media entre el salario-hora de un hombre y de una mujer, para un mismo nivel educativo? Estimando los modelos y contrastando de la manera tradicional las siguientes hipótesis:

$$w_i = \beta_0 + \beta_1 edu_i + \beta_2 mujer_i + \varepsilon_i \qquad \rightarrow H_0: \beta_2 = 0$$

$$w_i = \alpha_0 + \alpha_1 edu_i + \alpha_2 \text{hom}bre_i + \varepsilon_i \qquad \rightarrow H_0: \alpha_2 = 0$$

$$w_i = \delta_1 edu_i + \delta_2 mujer_i + \delta_3 \text{hom}bre_i + \varepsilon_i \rightarrow H_0: \delta_2 = \delta_3$$

<u>B) Efecto interacción</u>: Empleamos las variables ficticias para modelizar cambios en el efecto de las X sobre *Y* (en las pendientes del modelo).

Veamos un ejemplo con efectos aditivos e interacción:

$$w_i = \beta_0 + \beta_1 edu_i + \beta_2 mujer_i + \beta_3 edu_i \times mujer_i + \varepsilon_i \quad i = 1,..,n$$

donde:

$$\textbf{\textit{mujer}}_i = \begin{cases} 1 & si \quad i \quad es \quad mujer \\ 0 & si \quad i \quad es \quad \text{hom}bre \end{cases}$$

$$\textbf{\textit{edu}}_i \textbf{ x } \textbf{\textit{mujer}}_i = \begin{cases} edu_i & si \quad i \quad es \quad mujer \\ 0 & si \quad i \quad es \quad \text{hom}bre \end{cases}$$

Tenemos que: $E\left(w_i | edu_i, mujer_i, edu_i \times mujer_i\right) = \beta_0 + \beta_1 edu_i + \beta_2 mujer_i + \beta_3 edu_i \times mujer_i$

con lo cual:

$$E\left(w_i | edu_i, mujer\right) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) edu_i$$

$$E\left(w_i | edu_i, \text{hom}bre\right) = \beta_0 + \beta_1 edu_i$$

$\beta_2 \rightarrow$ **mide la diferencia en el término constante entre hombres y mujeres.**

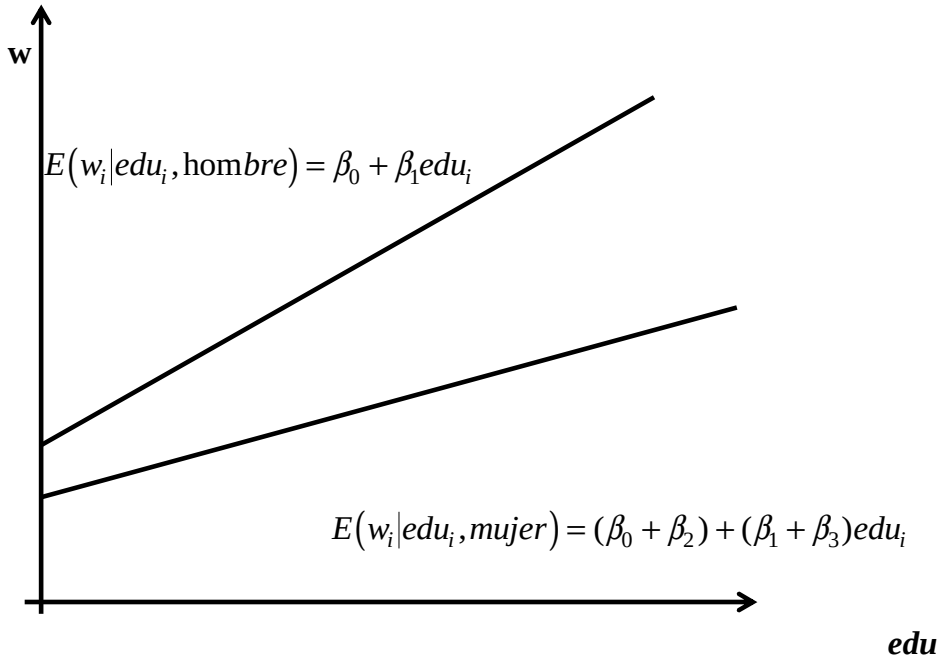
$\beta_3 \rightarrow$ **mide la diferencia en la pendiente entre hombres y mujeres:**

Si la educación (edu) varía en una unidad el salario-hora varía en media en:

$\beta_1 + \beta_3$ **unidades en las mujeres**

β_1 **unidades en los hombres**

Suponiendo $\beta_2 < 0$ y $\beta_3 < 0$



¿Cómo se contrastaría si las variaciones unitarias en la educación generan el mismo efecto medio sobre el salario-hora en hombres y en mujeres? Estimando el modelo y contrastando de la manera habitual la hipótesis:

$$H_0: \beta_3 = 0$$

¿Cómo se contrastaría si el término constante es el mismo para hombres y para mujeres? Estimando el modelo y contrastando de la manera habitual la hipótesis:

$$H_0: \beta_2 = 0$$

¿Cómo se contrastaría si el modelo de determinación salarial es el mismo en hombres y en mujeres? Estimando el modelo y contrastando de la manera habitual la hipótesis:

$$H_0: \beta_2 = \beta_3 = 0$$

<u>Comentarios</u>:

- **Igual que hemos visto con el efecto aditivo, existen otras formulaciones alternativas de este mismo modelo. Por ejemplo:**

$$w_i = \alpha_0 + \alpha_1 edu_i + \alpha_2 \,\text{hom}bre_i + \alpha_3 edu_i \times \text{hom}bre_i + \varepsilon_i$$

$$\textbf{\textit{hombre}}_i = \begin{cases} 1 & si \;\; i \;\; es \;\; \text{hom}bre \\ 0 & si \;\; i \;\; es \;\; mujer \end{cases} \qquad \textbf{\textit{edu}}_i \textbf{ x } \textbf{\textit{hombre}}_i = \begin{cases} edu_i & si \;\; i \;\; es \;\; \text{hom}bre \\ 0 & si \;\; i \;\; es \;\; mujer \end{cases}$$

o alternativamente

$$w_i = \delta_1 mujer_i + \delta_2 \,\text{hom}bre_i + \delta_3 edu_i \times mujer_i + \delta_4 edu_i \times \text{hom}bre_i + \varepsilon_i$$

- **<u>IMPORTANTE</u>: No sería válido un modelo (ya que habría multicolinealidad exacta) como:**

$$w_i = \gamma_1 mujer_i + \gamma_2 \,\text{hom}bre_i + \gamma_3 edu_i \times mujer_i + \gamma_4 edu_i \times \text{hom}bre_i + \gamma_5 edu_i + \varepsilon_i$$

ya que $edu_i \times mujer_i + edu_i \times \text{hom}bre_i = edu_i \quad \forall\, i$

● **Podríamos tener más de dos categorías. Por ejemplo:**

$$V_i = \beta_0 + \beta_1 S1_i + \beta_2 S2_i + \beta_3 P_i + \beta_4 (P_i \times S1_i) + \beta_5 (P_i \times S2_i) + \varepsilon_i$$

donde:

***V_i*: ventas de la empresa**

***P_i*: gastos en publicidad de la empresa**

$$\boldsymbol{S1_i} = \begin{cases} 1 & si \;\; i \;\; es \;\; del \;\; \text{sector } 1 \\ 0 & si \;\; i \;\; es \;\; del \;\; \text{sector } 2 \; o \; 3 \end{cases}$$

$$\boldsymbol{S2_i} = \begin{cases} 1 & si \;\; i \;\; es \;\; del \;\; \text{sector } 2 \\ 0 & si \;\; i \;\; es \;\; del \;\; \text{sector } 1 \; o \; 3 \end{cases}$$

Entonces:

$$E(V_i | P_i, Sector\ 1) = (\beta_0 + \beta_1) + (\beta_3 + \beta_4) P_i$$

$$E(V_i | P_i, Sector\ 2) = (\beta_0 + \beta_2) + (\beta_3 + \beta_5) P_i$$

$$E(V_i | P_i, Sector\ 3) = \beta_0 + \beta_3 P_i$$

2. CONTRASTE DE CHOW

- **Tal y como hemos visto para efectuar contrastes de cambio estructural es válida la inferencia habitual en el contexto del Modelo Lineal General empleando variables ficticias.**

- **Sin embargo, cuando se desea analizar la existencia de cambio estructural en todos los parámetros de un modelo se suele emplear una expresión particular del contraste F conocida como "contraste de Chow".**

- **Modelo sin restringir (con cambio estructural)**

$$Y_i = \beta_0^A + \beta_1^A X_{1i} + \beta_2^A X_{2i} + ... + \beta_k^A X_{Ki} + \varepsilon_i \quad Submuestra \quad A$$

$$Y_i = \beta_0^B + \beta_1^B X_{1i} + \beta_2^B X_{2i} + ... + \beta_k^B X_{Ki} + \varepsilon_i \quad Submuestra \quad B$$

- **Modelo restringido (sin cambio estructural)**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_K X_{Ki} + \varepsilon_i$$

Restricción $H_0: \beta_j^A = \beta_j^B \quad j = 0,1,...,k$

- **Sea SRS la suma al cuadrado de los residuos MCO del modelo sin restringir. Que se puede obtener estimando el modelo por separado con cada una de las dos submuestras (*SRS=SR$_A$+SR$_B$*)**
- **Sea SRR la suma al cuadrado de los residuos MCO del modelo restringido.**

- **Si suponemos que:**

Subpoblación A: $Y|X_{i1},...,X_{Ki} \sim N\left(\beta_0^A + \beta_1^A X_{1i} + ... + \beta_K^A X_{Ki}, \sigma^2\right)$

Subpoblación B: $Y|X_{1i},...,X_{Ki} \sim N\left(\beta_0^B + \beta_1^B X_{1i} + ... + \beta_K^B X_{Ki}, \sigma^2\right)$

entonces bajo $H_0: \beta_j^A = \beta_j^B \quad j = 0,1,...,K$ **[*"K+1"* hipótesis lineales]**

$$F = \frac{(SRR - SRS)}{(SRS)} \times \frac{(n - 2(K+1))}{K+1} \sim \mathbf{F}_{K+1,n-2(K+1)}$$ **con *SRS=SR$_A$+SR$_B$***

- **Empleando la aproximación asintótica tenemos bajo** $H_0: \beta_j^A = \beta_j^B \quad j = 0,1,...,K$ **[*"K+1"* hipótesis lineales]:**

$$W^0 = \frac{(SRR - SRS)}{(SRS)} \times (n - 2(K+1)) = (K+1)F \underset{a}{\sim} \chi^2_{k+1}$$ ***con SRS=SRA+SRB***

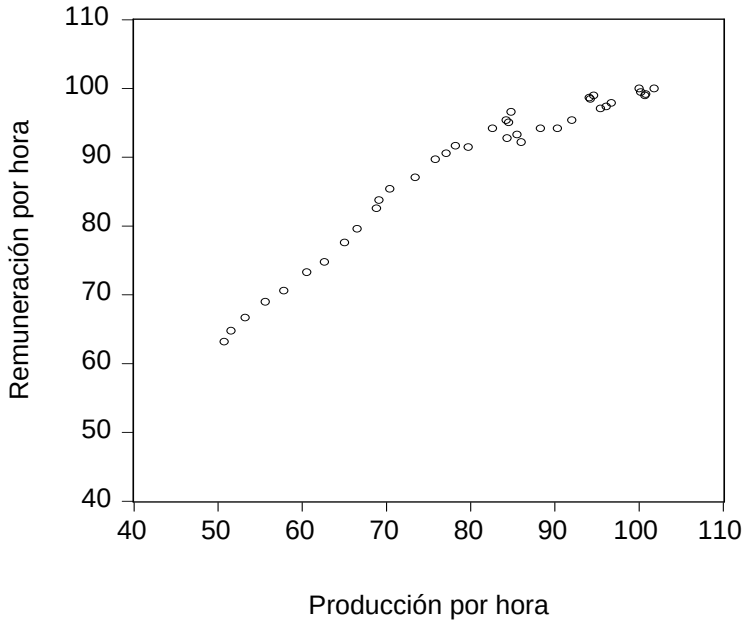
Ejemplo:

Datos de Estados Unidos de 1959-1996.

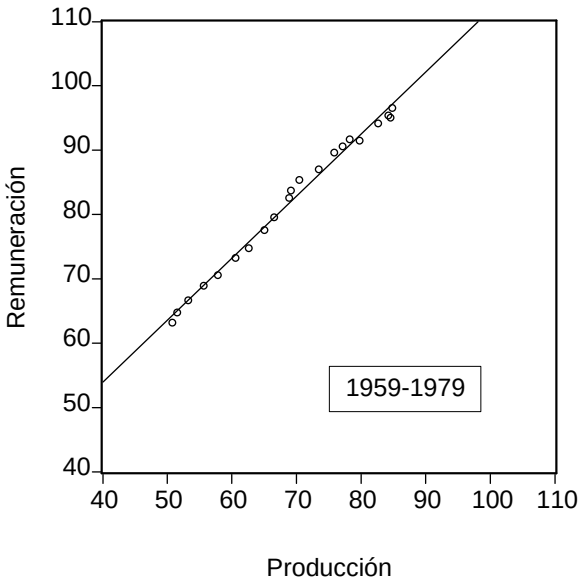
Relación salario/hora versus producción/hora

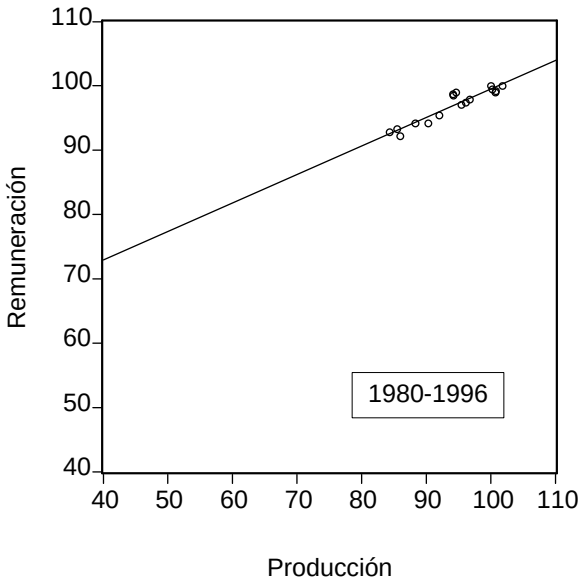
Datos en dólares constantes

Posible cambio estructural en 1980



Nota: Índice =100 en 1992



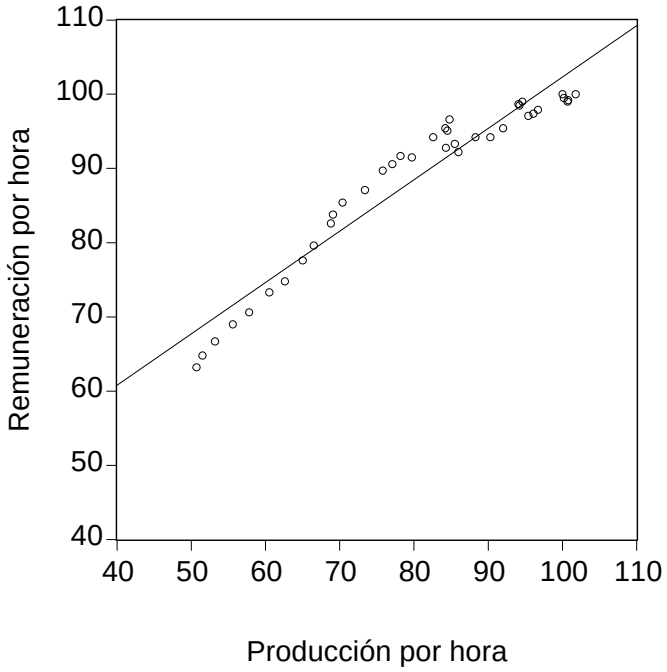


Ejemplo
Y=Remuneración
X=Producción

Dependent Variable: REMUNERAC
Method: Least Squares
Sample: 1959 1996
Included observations: 38
REMUNERAC=C(1)+C(2)*PRODUCC

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	33.14154	2.540806	13.04372	0.0000
C(2)	0.691884	0.031062	22.27462	0.0000

R-squared	0.932351	Mean dependent var	88.72895
Adjusted R-squared	0.930472	S.D. dependent var	11.16115
S.E. of regression	2.942994	Akaike info criterion	5.047928
Sum squared resid	311.8037	Schwarz criterion	5.134116
Log likelihood	-93.91063	F-statistic	496.1585
Durbin-Watson stat	0.115363	Prob(F-statistic)	0.000000



Y=Remuneración X=Producción

Submuestra A=1959-1979

Dependent Variable: REMUNERAC
Method: Least Squares
Sample: 1959 1979
Included observations: 21
REMUNERAC=C(1)+C(2)*PRODUCC

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	15.43406	1.461802	10.55825	0.0000
C(2)	0.963626	0.020876	46.15875	0.0000

R-squared	0.991161	Mean dependent var	82.06190
Adjusted R-squared	0.990696	S.D. dependent var	10.97003
S.E. of regression	1.058134	Akaike info criterion	3.041284
Sum squared resid	21.27332	Schwarz criterion	3.140763
Log likelihood	-29.93349	F-statistic	2130.630
Durbin-Watson stat	0.450144	Prob(F-statistic)	0.000000

Y=Remuneración X=Producción

Submuestra B=1980-1996

Dependent Variable: REMUNERAC
Method: Least Squares
Simple: 1980 1996
Included observations: 17
REMUNERAC=C(1)+C(2)*PRODUCC

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	55.24681	3.947315	13.99605	0.0000
C(2)	0.442976	0.041842	10.58690	0.0000

R-squared	0.881966	Mean dependent var	96.96471
Adjusted R-squared	0.874098	S.D. dependent var	2.690247
S.E. of regression	0.954573	Akaike info criterion	2.855026
Sum squared resid	13.66814	Schwarz criterion	2.953051
Log likelihood	-22.26772	F-statistic	112.0825
Durbin-Watson stat	1.157740	Prob(F-statistic)	0.000000

Contraste de Chow

- **Si suponemos que:**

$$Y|X \sim N\left(\beta_0^A + \beta_1^A X, \sigma^2\right) \quad A$$

$$Y|X_1 \sim N\left(\beta_0^B + \beta_1^B X, \sigma^2\right) \quad B$$

entonces bajo $H_0: \beta_j^A = \beta_j^B \quad j = 0,1$ **[“2” hipótesis lineales]**

$$F = \frac{(SRR - SRS)}{(SRS)} \times \frac{(38 - 2(2))}{2} \sim \mathbf{F}_{2,34}$$

Tenemos que:

$$F = \frac{(311{,}8037 - 34{,}94146)}{(34{,}94146)} \times \frac{(34)}{2} = 134{,}701 > F_{2,34}(5\%) \approx 3{,}32$$

con SRS=SR_A+SR_B=21,27332+13,66814=34,94146

Se rechaza $H_0: \beta_j^A = \beta_j^B \quad j = 0,1 \rightarrow$ **Hay evidencia de cambio estructural**