

TEMA 4. EL MODEL DE REGRESSIÓ MÚLTIPLE: INFERÈNCIA

Joan Llull

Materials: <http://pareto.uab.cat/jllull>

Tutories: dijous de 11:00 a 13:00h
(concertar cita per email)
—Despatx B3-1132—

joan.llull [at] movebarcelona [dot] eu

T4. EL MODEL DE REGRESSIÓ MÚLTIPLE: INFERÈNCIA

- 1** Inferència estadística: un breu repàs
- 2** Contrast d'hipòtesi d'un coeficient i intervals de confiança
- 3** Contrast d'hipòtesi de múltiples coeficients
- 4** Aplicacions

T4. EL MODEL DE REGRESSIÓ MÚLTIPLE: INFERÈNCIA

- 1 Inferència estadística: un breu repàs
- 2 Contrast d'hipòtesi d'un coeficient i intervals de confiança
- 3 Contrast d'hipòtesi de múltiples coeficients
- 4 Aplicacions

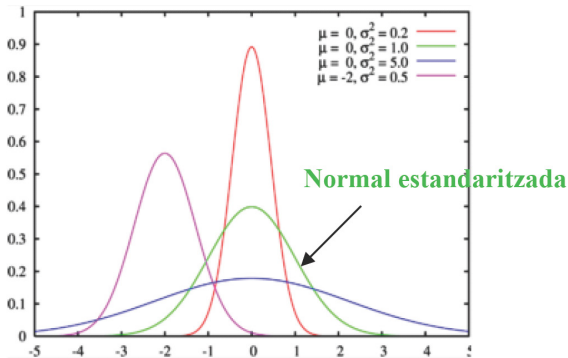
Objectiu

- El nostre objectiu és obtenir informació sobre β ...
...però no podem observar β (**població**).
- L'únic que tenim és una **mostra** de y i X de mida N que ha estat generada per $y = X\beta + u$ $u \sim \mathcal{N}(0, \sigma^2)$...
...i amb aquesta mostra hem calculat $\hat{\beta}$ i $\hat{\sigma}^2$.
- Tant $\hat{\beta}$ com $\hat{\sigma}^2$ són **variables aleatòries**.
- La **inferència** ens permetrà usar eines estadístiques per extreure conclusions sobre β .
- Aquestes eines seran: **contrastos d'hipòtesi i intervals de confiança**

(més val encertar aproximadament que equivocar-se exactament!)

Les distribucions rellevants (I): Normal

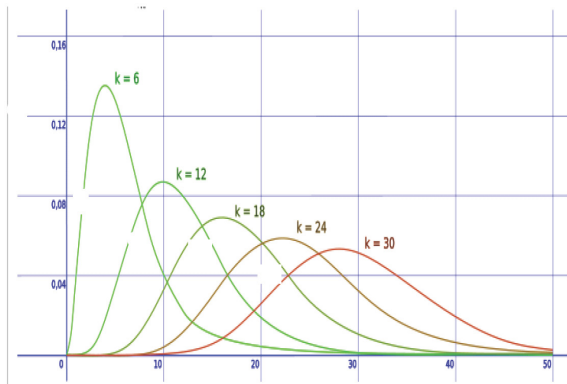
$$z \sim \mathcal{N}(\mu, \sigma^2)$$



Propietat important (estandaritzar): $\frac{z - \mu}{\sigma} \sim$

Les distribucions rellevants (II): χ^2

$z_1, z_2, \dots, z_\nu \sim \mathcal{N}(0, 1)$ independents $\Rightarrow z_1^2 + z_2^2 + \dots + z_\nu^2 \sim \chi^2(\nu)$



En el gràfic, k són els graus de llibertat (ν)

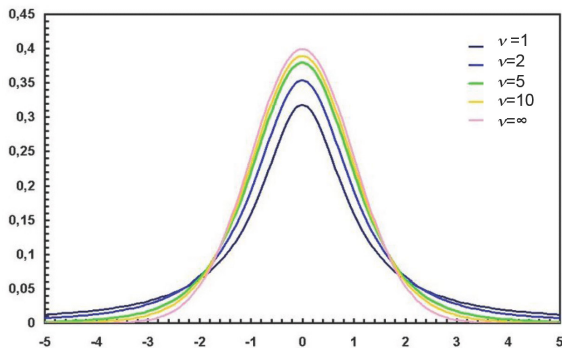
Esperança = ν

Variància = 2ν

Les distribucions rellevants (III): t-Student

w_1, w_2 independents, $w_1 \sim \mathcal{N}(0, 1)$, $w_2 \sim \chi^2(\nu)$

$$\Rightarrow \frac{w_1}{\sqrt{w_2/\nu}} \sim t(\nu)$$

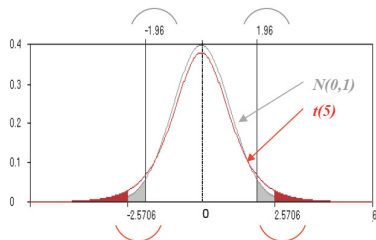


Esperança=0

Variància= $\frac{\nu}{\nu-2}$

Distribució normal vs Distribució t-Student

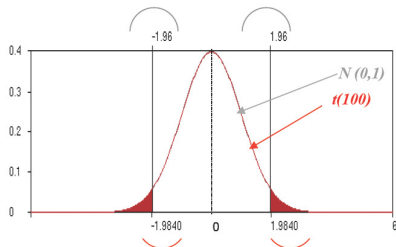
area total = 5%
area total = 5%



distribució $N(0,1)$

distribució $t(5)$

area total = 5%
area total = 5%



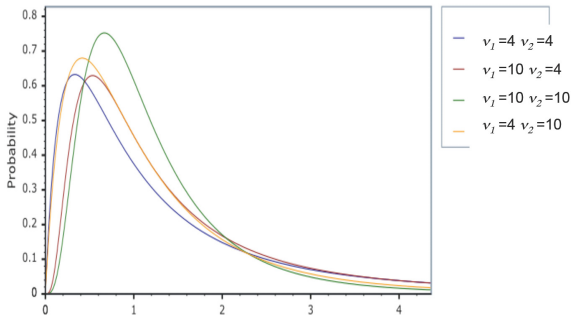
distribució $N(0,1)$

distribució $t(100)$

Les distribucions rellevants (IV): F

x_1, x_2 independents, $x_1 \sim \chi^2(\nu_1)$, $x_2 \sim \chi^2(\nu_2)$

$$\Rightarrow \frac{x_1/\nu_1}{x_2/\nu_2} \sim F(\nu_1, \nu_2)$$



$$\text{Esperança} = \frac{\nu_2}{\nu_2 - 2}$$

$$\text{Variança} = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$$

Contrastos d'hipòtesi

1. Establir hipòtesi **nul·la** H_0 i hipòtesi **alternativa** H_A .
2. Determinar què és probable observar i què no si H_0 és **certa**.
3. Agafar la **mostra** i determinar si el que observem és molt o poc probable si H_0 fos certa:
 - Si el que observem és probable sota $H_0 \Rightarrow$ **No rebutgem** H_0 (“ens quedem” amb H_0)
 - Si el que observem és poc probable sota $H_0 \Rightarrow$ **Rebutgem** H_0 (“ens quedem” amb H_A)

Eines de contrast

La variable que ens permet fer el contrast es coneix com a **estadístic de contrast**.

- Hem de conèixer la seva distribució sota H_0
- S'ha de poder calcular a partir de la mostra (pas 3)

Donat que rebutgem quan una cosa és poc probable sota H_0 i no rebutgem quan és molt probable, podem cometre **errors**:

	No rebutgem H_0	Rebutgem H_0
H_0 és certa	OK	Error tipus I
H_0 no és certa	Error tipus II	OK

T4. EL MODEL DE REGRESSIÓ MÚLTIPLE: INFERÈNCIA

- 1 Inferència estadística: un breu repàs
- 2 Contrast d'hipòtesi d'un coeficient i intervals de confiança
- 3 Contrast d'hipòtesi de múltiples coeficients
- 4 Aplicacions

Distribucions

Com ja sabem del tema anterior:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1}) \Rightarrow \hat{\beta}_k \sim \mathcal{N}(\beta_k, \sigma^2(X'X)^{-1}_{(k+1)(k+1)})$$

Per tant:

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2(X'X)^{-1}_{(k+1)(k+1)}}} \sim$$

Imaginem que volem contrastar si $\beta_k = 0$ (H_0). Llavors, sota la hipòtesi nul·la:

$$\frac{\hat{\beta}_k - 0}{\sqrt{\sigma^2(X'X)^{-1}_{(k+1)(k+1)}}} \underset{\text{sota } H_0}{\sim}$$

Coneixem la distribució? **Podem calcular** aquest estadístic?

Una **temptació** podria ser substituir σ^2 per $\hat{\sigma}^2$, però:

- $\hat{\sigma}^2$ és una **variable aleatòria**.
- Per tant, $\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{(k+1)(k+1)}}} \approx \mathcal{N}(0, 1)$!

Per conèixer quina és la distribució de $\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{(k+1)(k+1)}}$ primer hem de saber quina és la distribució de $\hat{\sigma}^2$:

$$\frac{(N - K)\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^N \hat{u}_i^2}{\sigma^2} \sim$$

Per tant:

$$T \equiv \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 (X'X)^{-1}_{(k+1)(k+1)}}}}{\sqrt{(N - K)\hat{\sigma}^2 / \sigma^2} / (N - K)} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{(k+1)(k+1)}}} = \frac{\hat{\beta}_k - \beta_k}{s.e.(\hat{\beta}_k)} \sim$$

Contrast de dues cues

Volem contrastar:

$$H_0 : \beta_k = 0$$

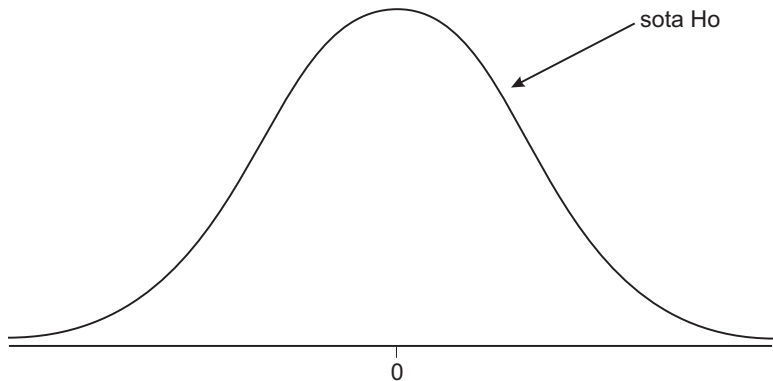
$$H_A : \beta_k \neq 0.$$

Hem de **calcular** l'estadístic T :

$$T = \frac{\hat{\beta}_k - 0}{s.e.(\hat{\beta}_k)} \underset{\text{sota } H_0}{\sim} t(N - K).$$

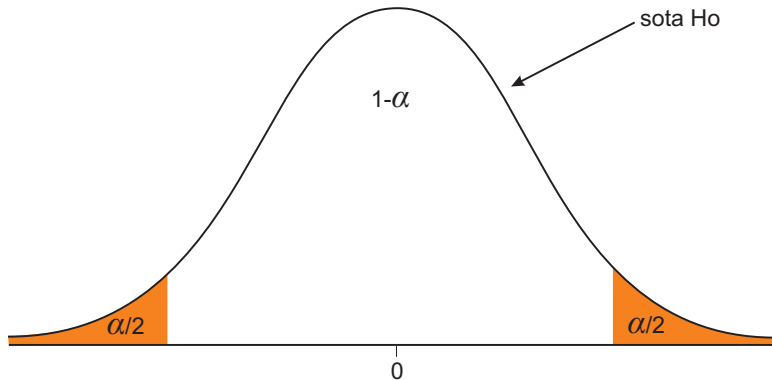
I **comprovar** si el valor d'aquest estadístic seria “normal” o “extrany” si H_0 fos certa.

Zona d'acceptació i zona crítica (dues cues)



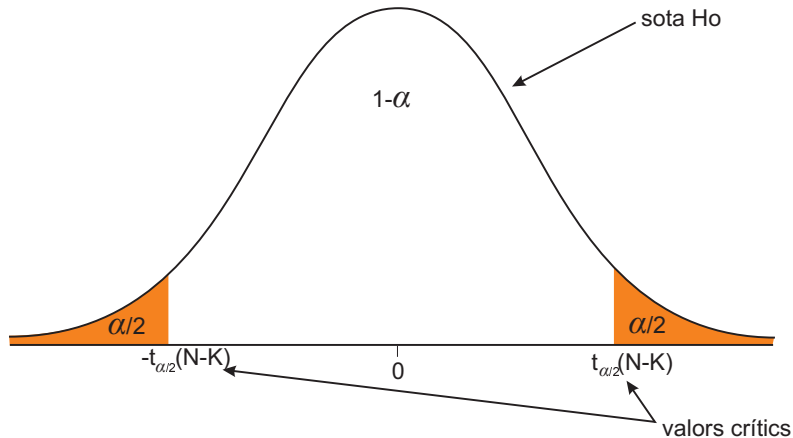
Zona d'acceptació i zona crítica (dues cues)

Nivell de significativitat: α



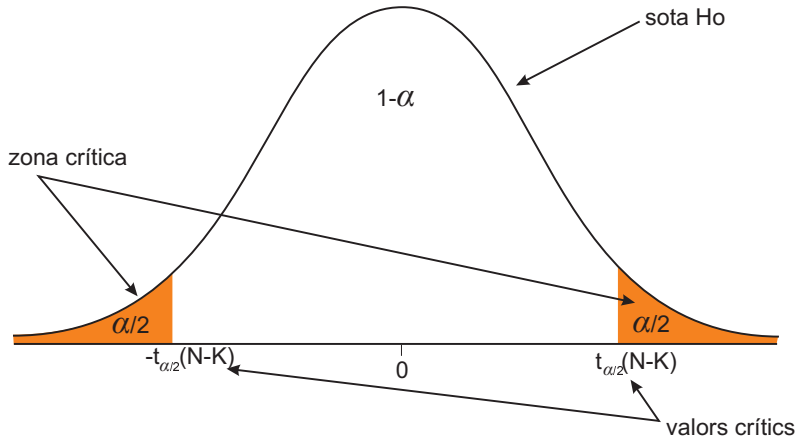
Zona d'acceptació i zona crítica (dues cues)

Nivell de significativitat: α



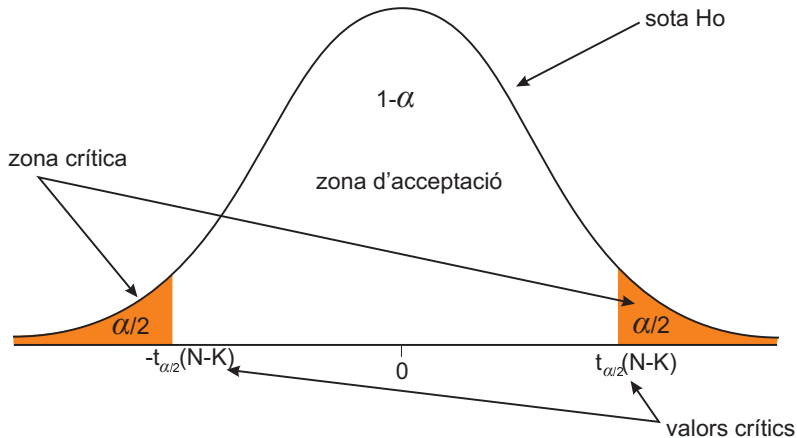
Zona d'acceptació i zona crítica (dues cues)

Nivell de significativitat: α

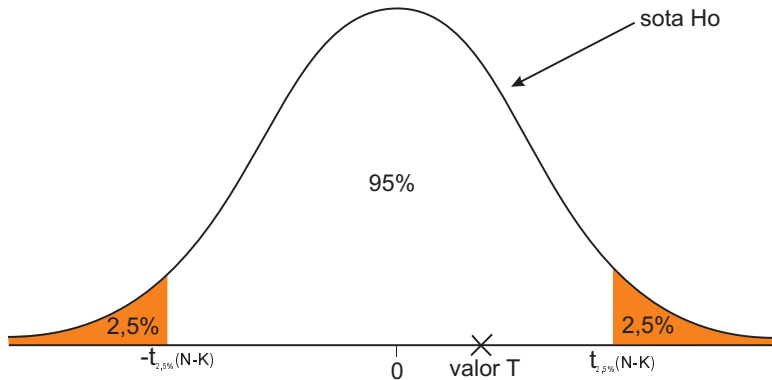


Zona d'acceptació i zona crítica (dues cues)

Nivell de significativitat: α

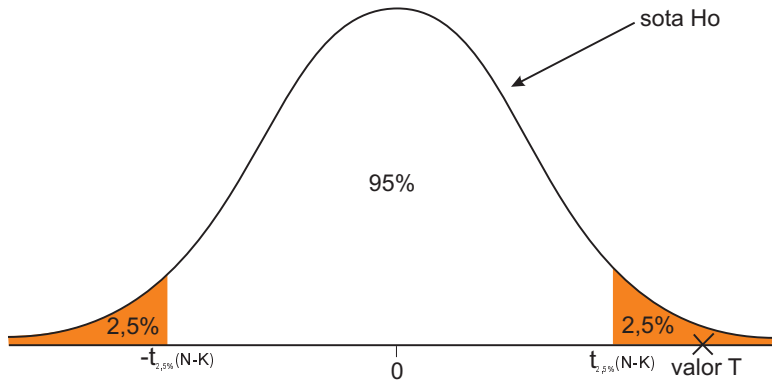


No rebutgem H_0



$$|T| < |t_{\alpha/2}(N - K)|$$

Rebutgem H_0



$$|T| > |t_{\alpha/2}(N - K)|$$

Contrast d'una cua

Ara volem contrastar:

$$H_0 : \beta_k = 0$$

$$H_A : \beta_k > 0$$

o

$$H_0 : \beta_k = 0$$

$$H_A : \beta_k < 0.$$

L'estadístic T (i la seva distribució sota H_0) serà **igual que abans**, ja que H_0 no ha canviat.

El que canviarà ara és la **zona crítica**: ara acumulem tot α en una de les cues (en lloc de la meitat a cada cua).

Quin serà ara el **valor crític**?

Valors crítics

Per trobar els valors crítics podem utilitzar les taules de la distribució o algun “t-calculator” com el vist a pràctiques.

Alguns exemples de valors crítics de la distribució t :

$$t_{5\%}(40) = 1,68$$

$$t_{5\%}(60) = 1,67$$

$$t_{5\%}(100) = 1,66$$

$$t_{5\%}(\infty) = 1,64$$

$$t_{2.5\%}(40) = 2,02$$

$$t_{2.5\%}(60) = 2,00$$

$$t_{2.5\%}(100) = 1,98$$

$$t_{2.5\%}(\infty) = 1,96$$

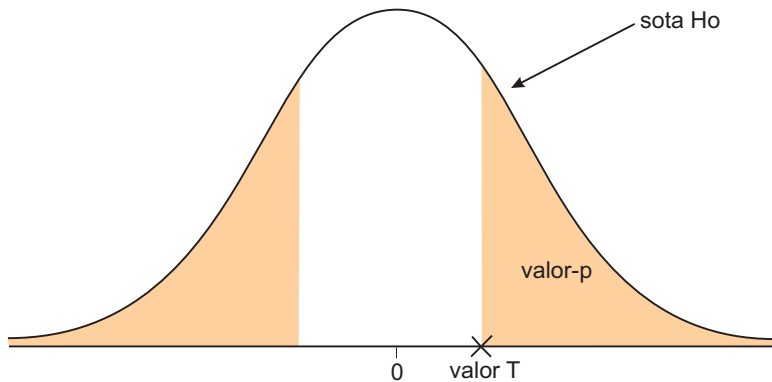
$$t_{0.5\%}(40) = 2,70$$

$$t_{0.5\%}(60) = 2,66$$

$$t_{0.5\%}(100) = 2,63$$

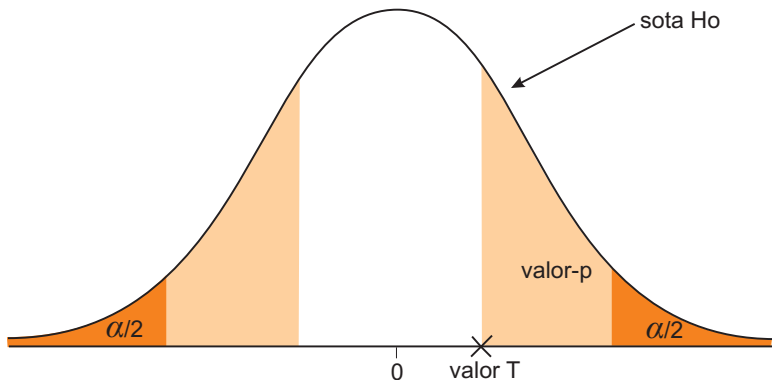
$$t_{0.5\%}(\infty) = 2,58$$

Valor-p (dues cues)



Valor-p (dues cues)

Nivell de significativitat: α



$valor - p > \alpha \Rightarrow$ No rebutgem H_0

$valor - p < \alpha \Rightarrow$ Rebutgem H_0

Intèrvals de confiança

El nostre **objectiu** és donar un interval dins el que β_k es troba amb un $1 - \alpha$ (p.ex. 95%) de probabilitat:

$$\Pr \left[\underline{\hat{\beta}}_k < \beta_k < \overline{\hat{\beta}}_k \right] = 95\%$$

El que nosaltres sabem és:

$$\Pr \left[-t_{2.5\%}(N - K) < \frac{\hat{\beta}_k - \beta_k}{s.e.(\hat{\beta}_k)} < t_{2.5\%}(N - K) \right] = 95\%$$

Per tant, (demostració pissarra)

$$\Pr \left[\hat{\beta}_k - t_{2.5\%}(N - K)s.e.(\hat{\beta}_k) < \beta_k < \hat{\beta}_k + t_{2.5\%}(N - K)s.e.(\hat{\beta}_k) \right] = 95\%$$

Aleshores, l'**interval de confiança** al 95% de $\hat{\beta}_k$ ve donat per:

$$\hat{\beta}_k \pm t_{2.5\%}(N - K)s.e.(\hat{\beta}_k)$$

T4. EL MODEL DE REGRESSIÓ MÚLTIPLE: INFERÈNCIA

- 1 Inferència estadística: un breu repàs
- 2 Contrast d'hipòtesi d'un coeficient i intervals de confiança
- 3 Contrast d'hipòtesi de múltiples coeficients
- 4 Aplicacions

Contrast t d'una combinació lineal de coeficients (I)

Volem contrastar **una** combinació lineal de paràmetres. Per exemple:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{K-1} x_{iK-1} + u_i \quad u_i \sim i.i. \mathcal{N}(0, \sigma^2)$$

$$H_0 : \beta_1 + \beta_2 = 1$$

$$H_A : \beta_1 + \beta_2 \neq 1.$$

Coneixem les **distribucions** de cada un dels estimadors:

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma^2 (X'X)_{22}^{-1})$$

$$\hat{\beta}_2 \sim \mathcal{N}(\beta_2, \sigma^2 (X'X)_{33}^{-1})$$

$$\hat{\beta}_1 + \hat{\beta}_2 \sim ?$$

Contrast t d'una combinació lineal de coeficients (II)

Per tant, podem escriure un **estadístic** T com hem fet fins ara:

$$T = \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{\sqrt{\sigma^2(X'X)_{22}^{-1} + \sigma^2(X'X)_{33}^{-1} + 2\sigma^2(X'X)_{23}^{-1}}} = \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{\sqrt{\frac{(N-K)\hat{\sigma}^2}{\sigma^2} / N - K}} \underset{\text{sota } H_0}{\sim} t(N-K)$$

I fer el contrast com hem fet abans.

Contrast F d'una o vàries combinacions lineals de coeficients (I)

Suposem que volem **contrastar** hipòtesis que consten de **més d'una combinació** lineal de paràmetres:

$$H_0 : R\beta = r$$

$$H_A : R\beta \neq r.$$

Ja **no** podem calcular l'estadístic T : el numerador seria una normal **multivariada** (un vector de restriccions).

Haurem de cercar **una altra característica** de la que coneguem la distribució sota H_0 .

Contrast F d'una o vàries combinacions lineals de coeficients (II)

Si estimem el **model restringit**, sabem que:

$$\frac{SQR_R}{\sigma^2} \underset{\text{sota } H_0}{\sim} \chi^2(N - K + q),$$

i, per altra banda:

$$\frac{SQR_R - SQR}{\sigma^2} \underset{\text{sota } H_0}{\sim}$$

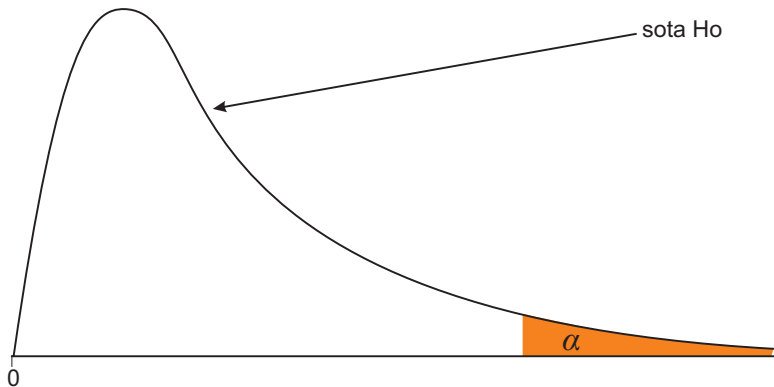
Així que el nostre **estadístic de contrast** serà:

$$F \equiv \frac{\frac{(SQR_R - SQR)}{\sigma^2}/q}{\frac{SQR}{\sigma^2}/(N - K)} = \frac{(SQR_R - SQR)/q}{SQR/(N - K)} \underset{\text{sota } H_0}{\sim}$$

Una **forma alternativa** de escriure-ho seria:

$$F = \frac{(R^2 - R_R^2)/q}{(1 - R^2)/(N - K)}.$$

Contrast F d'una o vàries combinacions lineals de coeficients (III)



Cas particular: significativitat conjunta

Un **contrast** que habitualment fem és el següent:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{K-1} = 0$$

$$H_A : \text{no } H_0.$$

En aquest cas:

$$F = \frac{R^2/(K-1)}{(1-R^2)/(N-K)} \underset{\text{sota } H_0}{\sim} F(K-1, N-K).$$

T4. EL MODEL DE REGRESSIÓ MÚLTIPLE: INFERÈNCIA

- 1 Inferència estadística: un breu repàs
- 2 Contrast d'hipòtesi d'un coeficient i intervals de confiança
- 3 Contrast d'hipòtesi de múltiples coeficients
- 4** Aplicacions

Demanda de llet

gretl: model 2

Model 2: OLS, using observations 1-50
Dependent variable: Vendes

	coefficient	std. error	t-ratio	p-value	
const	1186,59	130,374	9,101	9,12e-12	***
Preu	-418,495	72,7365	-5,754	7,24e-07	***
Publicitat	32,9186	8,54344	3,853	0,0004	***
Establiments	10,7193	1,49730	7,159	5,92e-09	***
Calci	-0,131206	1,05820	-0,1240	0,9019	

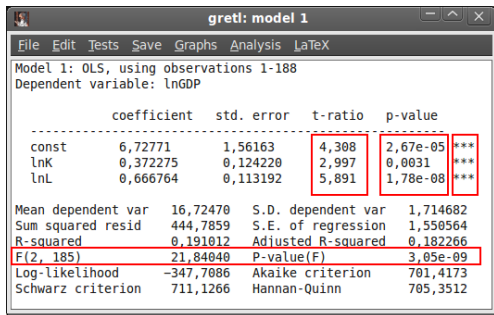
Mean dependent var	1290,220	S.D. dependent var	501,8181
Sum squared resid	3645967	S.E. of regression	284,6427
R-squared	0,704523	Adjusted R-squared	0,678258
F(4, 45)	26,82400	P-value(F)	2,06e-11
Log-likelihood	-350,8747	Akaike criterion	711,7493
Schwarz criterion	721,3094	Hannan-Quinn	715,3899

Excluding the constant, p-value was highest for variable 4 (Calci)

$$t_{5\%}(45) = 1,68; \quad t_{2,5\%}(45) = 2,01; \quad t_{0,5\%}(45) = 2,69$$

$$F_{10\%}(4, 45) = 2,07; \quad F_{5\%}(4, 45) = 2,58; \quad F_{1\%}(4, 45) = 3,77$$

Cobb-Douglas



The screenshot shows the gretl software interface for 'Model 1'. The dependent variable is lnGDP. The regression results are as follows:

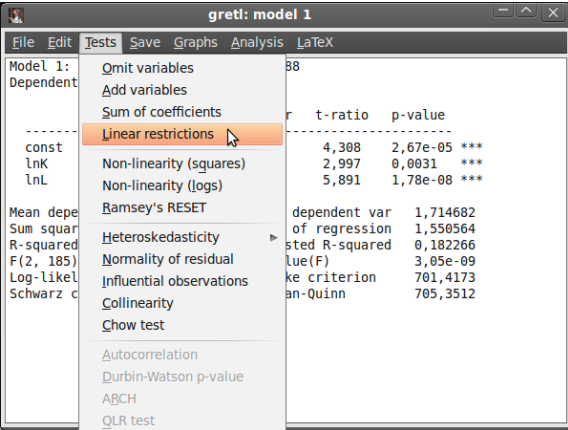
	coefficient	std. error	t-ratio	p-value	
const	6,72771	1,56163	4,308	2,67e-05	***
lnK	0,372275	0,124220	2,997	0,0031	***
lnL	0,666764	0,113192	5,891	1,78e-08	***
Mean dependent var	16,72470	S.D. dependent var	1,714682		
Sum squared resid	444,7859	S.E. of regression	1,550564		
R-squared	0,191012	Adjusted R-squared	0,182266		
F(2, 185)	21,84040	P-value(F)	3,05e-09		
Log-likelihood	-347,7086	Akaike criterion	701,4173		
Schwarz criterion	711,1266	Hannan-Quinn	705,3512		

$$t_{5\%}(185) = 1,65; \quad t_{2,5\%}(185) = 1,97; \quad t_{0,5\%}(185) = 2,60$$

$$F_{10\%}(2, 185) = 2,33; \quad F_{5\%}(2, 185) = 3,05; \quad F_{1\%}(2, 185) = 4,72$$

Com contrastaríem $\beta_1 = 0,3$? I $\beta_1 + \beta_2 = 1$?

Cobb-Douglas: contrasts amb l'estadístic F (I)

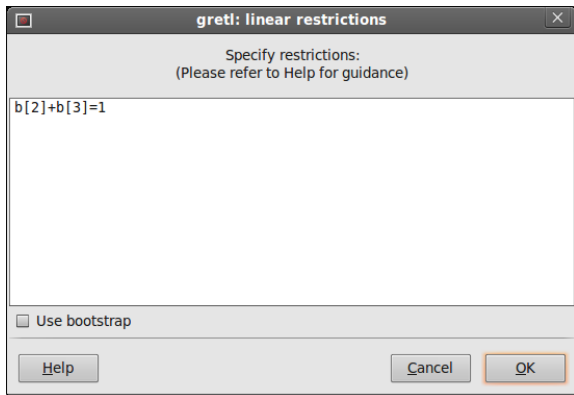


The screenshot shows the 'gretl: model 1' window. The 'Tests' menu is open, listing various diagnostic tests. The 'Linear restrictions' option is highlighted. In the background, a table of regression statistics is visible, including t-ratios and p-values for coefficients, and F-statistics for various tests.

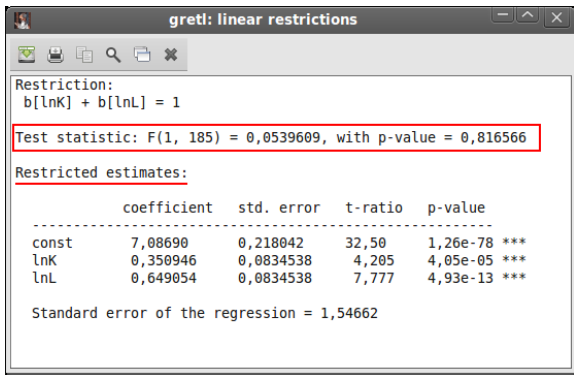
	t-ratio	p-value
const	4,308	2,67e-05 ***
lnK	2,997	0,0031 ***
lnL	5,891	1,78e-08 ***

Mean dependent var	1,714682
Sum of squares of regression	1,550564
R-squared	0,182266
F(2, 185)	3,05e-09
Log-likelihood criterion	701,4173
Schwarz criterion	705,3512

Cobb-Douglas: contrasts amb l'estadístic F (I)

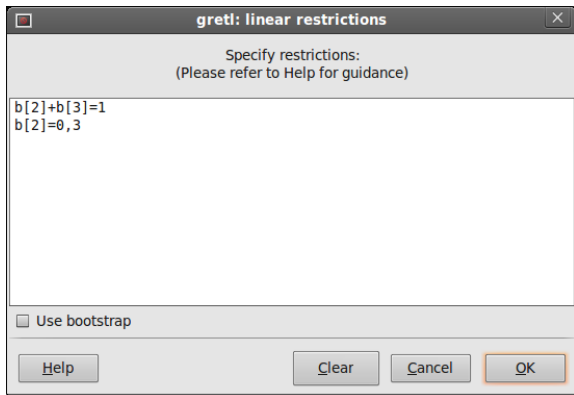


Cobb-Douglas: contrasts amb l'estadístic F (I)

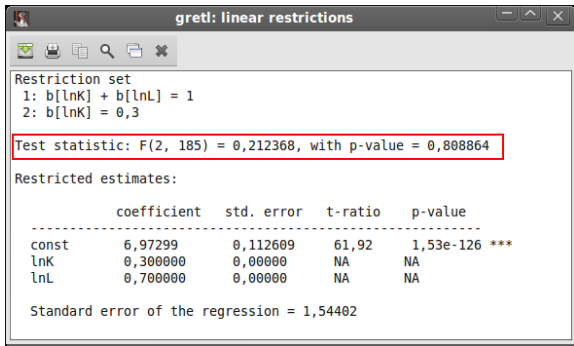


$$F_{10\%}(1, 185) = 2,73; \quad F_{5\%}(1, 185) = 3,89; \quad F_{1\%}(1, 185) = 6,77$$

Cobb-Douglas: contrasts amb l'estadístic F (II)



Cobb-Douglas: contrasts amb l'estadístic F (II)



$$F_{10\%}(2, 185) = 2,33; \quad F_{5\%}(2, 185) = 3,05; \quad F_{1\%}(2, 185) = 4,72$$

Salaris (I)

gretl: model 1

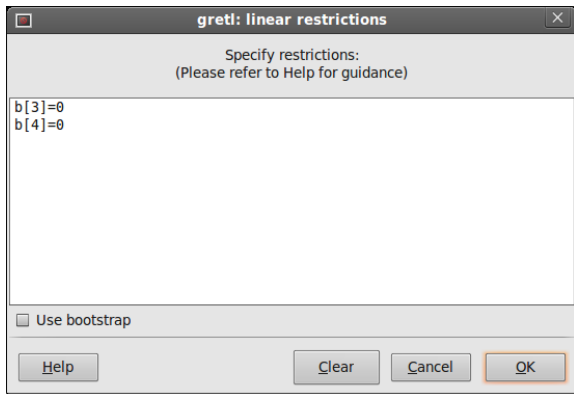
File Edit Tests Save Graphs Analysis LaTeX

Model 1: OLS, using observations 1-50
Dependent variable: LogSalari

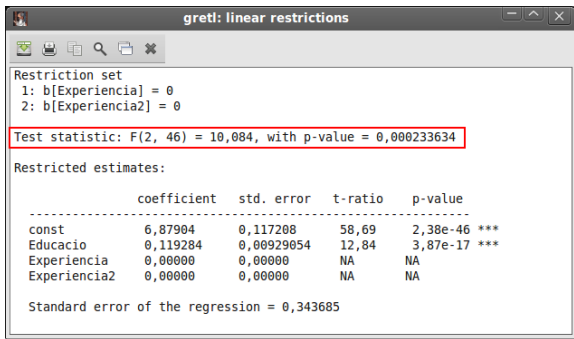
	coefficient	std. error	t-ratio	p-value	
const	6,30299	0,171197	36,82	8,85e-36	***
Educacio	0,121713	0,00798800	15,24	1,35e-19	***
Experiencia	0,0886850	0,0197774	4,484	4,85e-05	***
Experiencia2	-0,00262955	0,000613051	-4,289	9,11e-05	***
Mean dependent var	8,248416	S.D. dependent var	0,716302		
Sum squared resid	3,941596	S.E. of regression	0,292723		
R-squared	0,843223	Adjusted R-squared	0,832998		
F(3, 46)	82,46996	P-value(F)	1,57e-18		
Log-likelihood	-7,435993	Akaike criterion	22,87199		
Schwarz criterion	30,52008	Hannan-Quinn	25,78442		

$$t_{10\%}(44) = 1,30; t_{5\%}(44) = 1,68; t_{2,5\%}(44) = 2,02; t_{1\%}(44) = 2,41; t_{0,5\%}(185) = 2,69$$

Salaris (II)



Salaris (II)



$$F_{10\%}(2, 46) = 2,42; \quad F_{5\%}(2, 46) = 3,20; \quad F_{1\%}(2, 46) = 5,10$$