TEMA 1: MODELS ECONOMÈTRICS I DADES ECONÒMIQUES

1.1 Definició d'econometria

1.2 Etapes en la modelització economètrica

1.3 Dades econòmiques

Objectius:

- Definir i justificar l'estudi de l'econometria. ¿Per a què serveix l'econometria?
- Mostrar els elements d'un model economètric
- Descriure les etapes d'una anàlisi economètrica
- Introduir els diferents tipus de dades econòmiques

TEMA 1: MODELS ECONOMÈTRICS I DADES ECONÒMIQUES

1.1 Definició d'econometria

1.2 Etapes en la modelització economètrica

1.3 Dades econòmiques

Què és l'econometria?

- L'econometria es pot definir com la ciència social en la qual s'utilitzen les eines de la teoria econòmica, les matemàtiques i la inferència estadística per a l'anàlisi **dels fenòmens econòmics** (Goldberger).

- L'econometria s'ocupa de **formular** relacions entre variables econòmiques, **quantificar-les** i **valorar-ne** els resultats obtinguts (AFG).

Finalitat de l'econometria

- L'econometria és una disciplina que utilitza mètodes estadístics en l'anàlisi de les dades econòmiques amb la finalitat de:
  - establir i estimar relacions econòmiques
  - contrastar teories econòmiques
  - avaluar polítiques econòmiques / empresarials
  - predir

Exemples d'usos de l'econometria

- **Exemple 1**: Estimar relacions econòmiques

$$\text{Notes} = \beta_0 + \beta_1 \text{ Hores} + \beta_2 \text{ Intel·ligència}$$

$$\beta_0 \ \beta_1 \text{ i } \beta_2 \text{ són desconeguts } \Rightarrow \text{ Estimació}$$

- **Exemple 2**: Contrastar teories econòmiques

$$\text{Notes} = \beta_0 + \beta_1 \text{ Hores} + \beta_2 \text{ Intel·ligència}$$

$$\text{Hipòtesi: } \beta_2 > \beta_1 \Rightarrow \text{ Contrast}$$

- **Exemple 3.1**: Avaluar polítiques públiques

Programa de reciclatge de treballadors $\Rightarrow$ Ha augmentat la probabilitat de ser contractat?

- **Exemple 3.2**: Avaluar polítiques empresarials

  Campanya publicitària $\Rightarrow$ Quin n'ha estat l'impacte sobre les vendes?

- **Exemple 4**: Predicció

  Quant augmentarien els ingressos fiscals si el govern augmenta els impostos un 4%?

  Quant augmentaran les vendes si l'empresa redueix el preu un 10%?

En definitiva….

En definitiva, l'econometria intenta quantificar les relacions econòmiques per conèixer la realitat econòmica i facilitar així la presa de decisions. Per fer-ho, combina tres elements: teoria econòmica, dades i estadística.

TEMA 1: MODELS ECONOMÈTRICS I DADES ECONÒMIQUES

1.1 Definició d'econometria

1.2 Etapes en la modelització economètrica

1.3 Dades econòmiques

Model econòmic versus model economètric

- Una anàlisi economètrica empírica sol començar amb la formulació d'una pregunta. Per exemple: *Com incideixen els cursos de reciclatge en el salari?*

- O com afectaria la reducció del nombre d'estudiants dels grups en la universitat?

- Per tractar de respondre a la pregunta es planteja un **model econòmic**.

  Nota = f(hores, intel·ligència, grandària del grup)

- Partint del model teòric s'especifica un **model economètric**.

$$Nota = \beta_0 + \beta_1 H + \beta_2 I + \beta_2 G + u$$

- El model economètric incorpora:
  - una forma funcional específica entre les variables
  - paràmetres desconeguts a estimar (β)
  - un nou terme u, anomenat **pertorbació aleatòria**

Una vegada plantejat (o especificat) un model economètric

- Exemples de models:

  $\text{Nota} = \beta_0 + \beta_1 \text{ hores} + \beta_2 \text{ intel·ligència} + \beta_3 \text{ grandària} + u$

  $\text{Salari} = \beta_0 + \beta_1 \text{ educació} + \beta_2 \text{ experiència} + u$

  $\text{Vendes} = \beta_0 + \beta_1 \text{ preu} + \beta_2 \text{ publicitat} + u$

- Una vegada especificat el model economètric, s'obtenen les dades necessàries per a **estimar el model economètric.**

- Una vegada estimat el model economètric, se'l sotmet a unes proves mínimes de **validació** per assegurar-nos que el model té les propietats necessàries perquè tinguem confiança en les nostres estimacions, i que, per tant, el model estimat puga ser utilitzat per l'investigador, per exemple per a:

  - contrastar les hipòtesis d'interès ($\beta_3 = 0$ o $\beta_3 > 0$)
  - contrastar algun model econòmic formal
  - avaluar i quantificar els efectes d'alguna política
  - predir

Etapes de la modelització economètrica

Un treball economètric aplicat pot dividir-se o comporta tres etapes:

1) **Especificació**:
   - Es parteix d'un model econòmic (explícit o implícit) $Y = f(X_1, X_2)$
   - S'expressa aquest model econòmic en termes matemàtics (que inclouen variables aleatòries). És a dir, es planteja un model economètric:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

2) **Estimació**:
   - Es recullen dades referents al fet econòmic estudiat.
   - S'apliquen tècniques estadístiques per estimar i verificar les relacions entre les variables.

3) **Validació**:
   Una vegada estimat el model i abans de ser utilitzat per a diferents finalitats, s'ha de comprovar que el model estimat té "bones" propietats i que "ens podem fiar dels seus resultats".

TEMA 1: MODELS ECONOMÈTRICS I DADES ECONÒMIQUES

1.1 Definició d'econometria

1.2 Etapes en la modelització economètrica

1.3 Dades econòmiques

Tipus de dades

- **Dades de sèrie temporal**. Un conjunt d'observacions sobre una variable determinada en diferents moments del temps. Cada observació es refereix a un mateix "individu", en un moment donat del temps.

  Exemple: tipus d'interès interbancari de 1990 a 2005 d'un país, per exemple Espanya.

- **Dades de tall transversal** Un conjunt d'observacions de diferents individus o entitats referents a un mateix moment temporal.

  Exemple: tipus d'interès interbancari el 2005 en els països pertanyents a la UE.

- **Dades de panel**. Combinació de dades de sèrie temporal i de tall transversal. En un panel de dades hi ha observacions sobre diferents agents en diversos moments del temps. Els agents han de ser els mateixos.

  Exemple: tipus d'interès interbancari de 1990 a 2005 en els països pertanyents a la UE.

Exemples

- Exemple: dades de salaris i altres característiques individuals.

Obs. núm.	Salari	Educ.	Exper.	Civil
1	950	11	4	0
2	1400	12	17	1
3	1100	11	14	0
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
544	1300	14	8	1

- Exemple: dades macroeconòmiques espanyoles

Obs. núm.	Any	PIB	C	FBK
1	1980	46.789	30.228	4.700
2	1981	49.700	35.900	5.120
3	1982	55.100	40.900	5.200
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
26	2005	144.900	110.200	15.400

TEMA 2. EL MODEL DE REGRESSIÓ LINEAL (MRL)

2.1 El model de regressió lineal simple (MRLS)

2.2 El model de regressió lineal múltiple (MRLM)

2.3 Interpretació dels coeficients

2.4 Unitats de mesura i formes funcionals

The method of least squares is the automobile of modern statistical analysis; despite its limitations, occasional accidents, and incidental pollution, it and its numerous variations, extensions and related conveyances carry the bulk of statistical analysis, and are known and valued by all.

Stephen M. Stigler[1]

TEMA 2. EL MODEL DE REGRESSIÓ LINEAL (MRL)

2.1 El model de regressió lineal simple (MRLS)

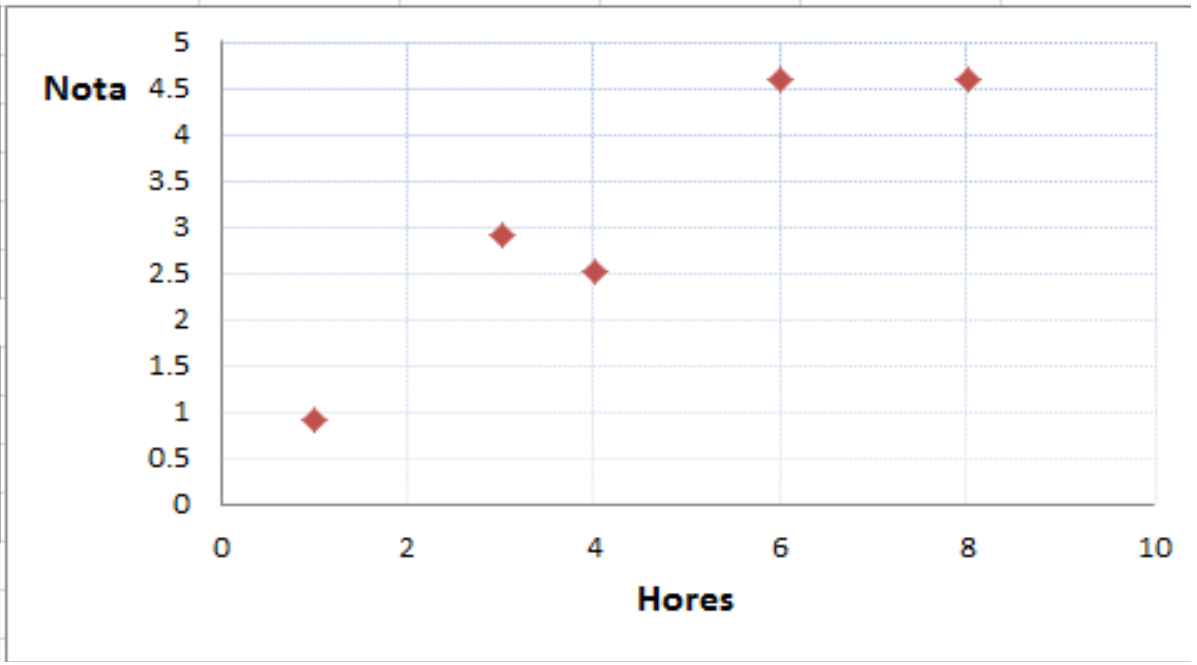
2.2 El model de regressió lineal múltiple (MRLM)

2.3 Interpretació de coeficients

2.4 Unitats de mesura i formes funcionals

- L'econometria s'ocupa de formular **relacions entre variables econòmiques**, quantificar-les i valorar els resultats obtinguts (AFG).
- Una anàlisi economètrica empírica sol començar amb la formulació d'una pregunta. Per exemple: **Com influeixen les hores d'estudi en la nota obtinguda en un examen?**
- Suposem que ens interessa analitzar la variable *Y*. Pensem que *Y* està relacionada amb la variable *X*; per tant, tractarem d'analitzar i quantificar la relació entre *X* i *Y*. Com? utilitzant les dades disponibles.
- *Recollim dades de X i Y. Ací els tenim… i ara què?*

i	Hores	Nota
Ariadna	3	2.9
Gregori	8	4.6
Albert	1	0.9
Joan	4	2.5
Maria	6	4.6
Mitjana	4.4	3.1
Variància	5.84	1.94
Desviació típica	2.41	1.39
Correlació	0.93	



- *Utilitzarem l'MRLS per a analitzar i quantificar la relació entre X i Y.*

- *L'MRLS parteix de suposar que una variable (Y) depèn d'una altra (X):* $Y = f(X)$

- Quina forma té aquesta relació? L'MRLS suposa que la relació entre *X* i *Y* és lineal:

$$Y_i = \beta_1 + \beta_2 X_i$$

- El model anterior és determinista. No obstant això, les relacions econòmiques no són deterministes, sempre hi ha un cert grau d'incertesa o aleatorietat. Per tant, en l'MRLS s'introdueix un terme addicional (u) anomenat pertorbació aleatòria, de manera que tindrem un model estocàstic:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \text{per a} \quad i = 1, \ldots, N$$

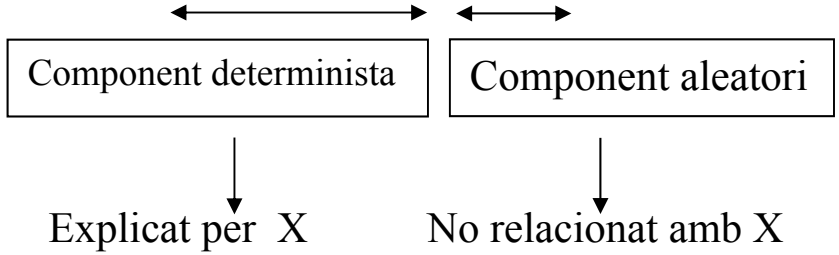
Terminologia

- El model de regressió lineal simple (MRLS) planteja que $Y_i = \beta_1 + \beta_2 X_i + u_i$
  - Ens referirem a la variable Y com a **variable dependent**, variable a explicar, regressand…
  - Ens referirem a la variable X com a variable independent, variable explicativa, **regressor**…
  - Ens referirem a la variable u com a terme d'error, **pertorbació** aleatòria, pertorbació estocàstica…
  - β_1, β_2 són els **paràmetres** que no coneixem i volem estimar.

- L'objectiu primordial de l'anàlisi de regressió és estimar els paràmetres poblacionals (β) partint d'una mostra de dades.

Interpretació de l'MRLS

- L'MRLS suposa que cada observació de Y és explicada per dues variables: X i la pertorbació aleatòria (u).
- De forma que $Y_i = \beta_1 + \beta_2 X_i + u_i$



- OBJECTIU: Quantificar la relació entre X i Y; és a dir, aproximar-nos o **estimar** els paràmetres desconeguts β_1 i β_2. Com? Utilitzant dades i mètodes estadístics (anàlisi de regressió).

Interpretació de u

- En l'MRLS: $Y_i = \beta_1 + \beta_2 X_i + u_i$ la pertorbació aleatòria (u) representa tots els altres factors (a banda de X) que afecten la variable Y

- Per exemple: $Salari_i = \beta_1 + \beta_2 Educació_i + u_i$

  - Quines variables poden estar darrere de u en aquest exemple?
  - Tot i així imaginat que s'introdueixen en el model totes aquestes variables que semblen importants, tindria sentit no incloure u en el nostre model?

- No hi ha models correctes, hi ha models útils.

Interpretació dels β

- Atès que la forma funcional és lineal, si els altres factors no canvien (*ceteris paribus*) aleshores X tindrà en el nostre model un efecte lineal en Y:

$$\Delta Y = \beta_x X \qquad \text{(sempre que } u \text{ es mantinga constant: } \Delta u = 0)$$

  - si X augmenta en 1-unitat, Y augmenta en β_x unitats

  - l'efecte d'augmentar X en 1 unitat és, segons el nostre MRLS, independent del valor inicial de X; és a dir, en el nostre exemple, l'efecte d'estudiar una hora més sempre és el mateix (β_x)

"Com que l'MRLS és una recta ...". El supòsit de linealitat implica … que un increment d'una unitat en X **augmenta el " valor esperat" de Y** en β_x unitats.

- Per exemple: *nota* = 0,5 + 2 *hores* + *u*

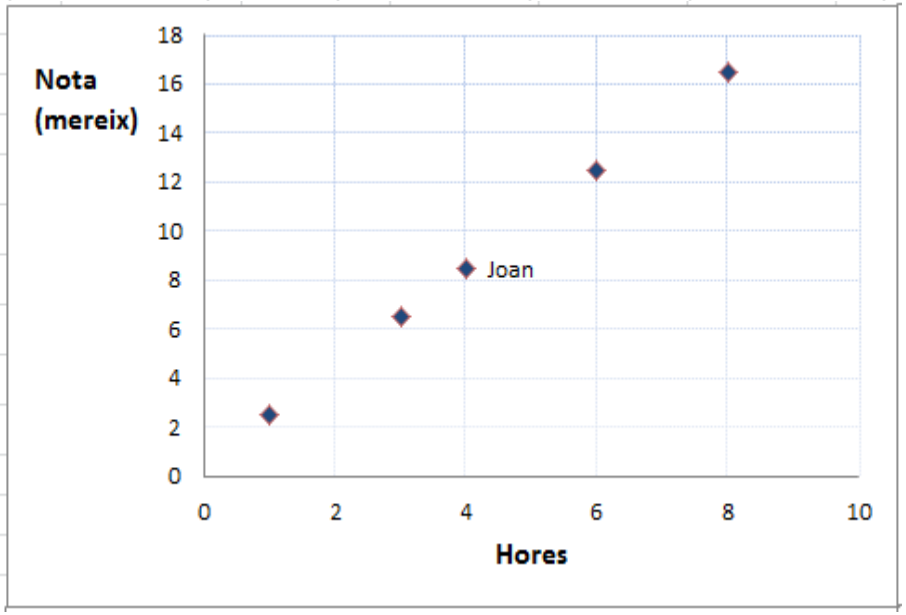
  - Com s'interpreta β_x en l'exemple anterior?
  - Com s'interpreta β_1?

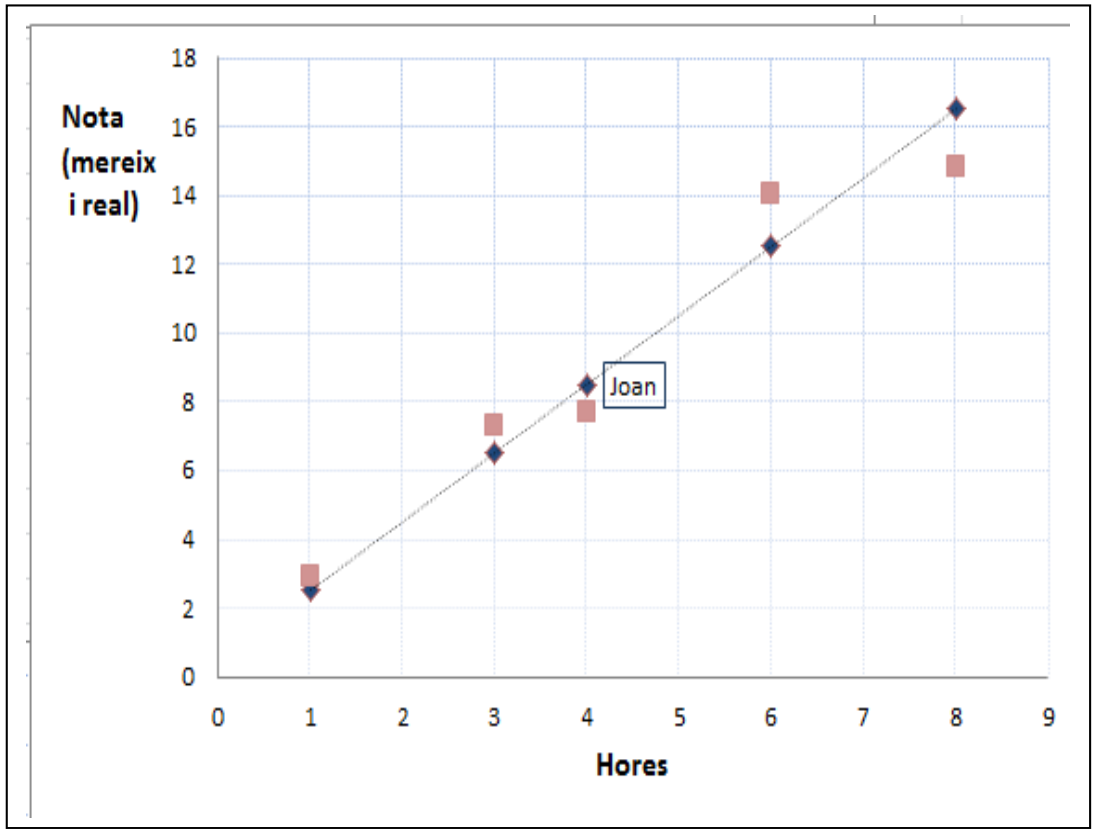
Model teòric (hipotetitzant… o racionalitzant la procedència de les dades)

- En l'MRLS, β_1 i β_2 són desconeguts, però ***suposem que jo els conec***.

$$Y_i = 0{,}5 + 2\,X_i + u_i$$

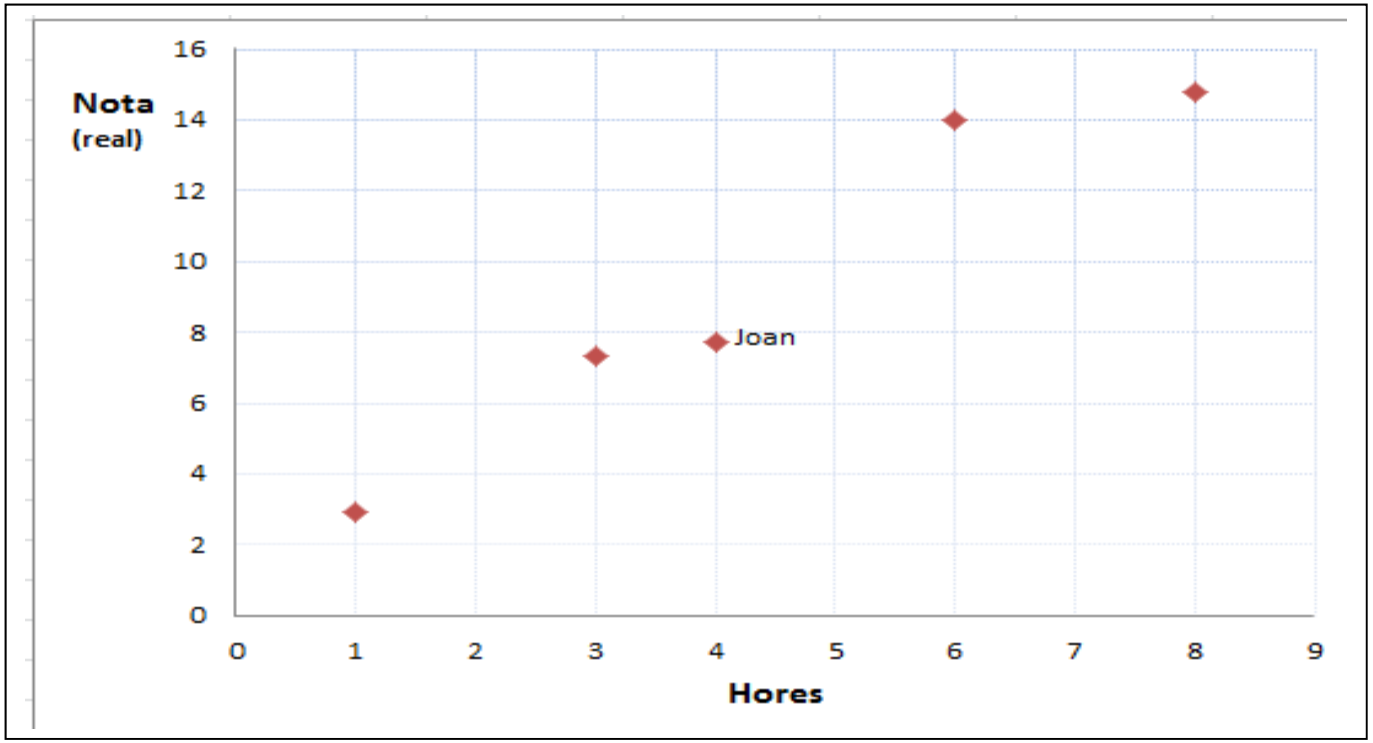
i	Hores	Nota (mereix)	u	Nota (real)
Ariadna	3	6.5	0.8	7.3
Gregori	8	16.5	-1.7	14.8
Albert	1	2.5	0.4	2.9
Joan	4	8.5	-0.8	7.7
Maria	6	12.5	1.5	14





Si tenim dades reals… ens permetran estimar els β

- En la realitat, si tenim dades del fenomen que volem analitzar, només observarem Y i X. No observarem de forma separada els seus dos components, ni observarem els u, ni els β

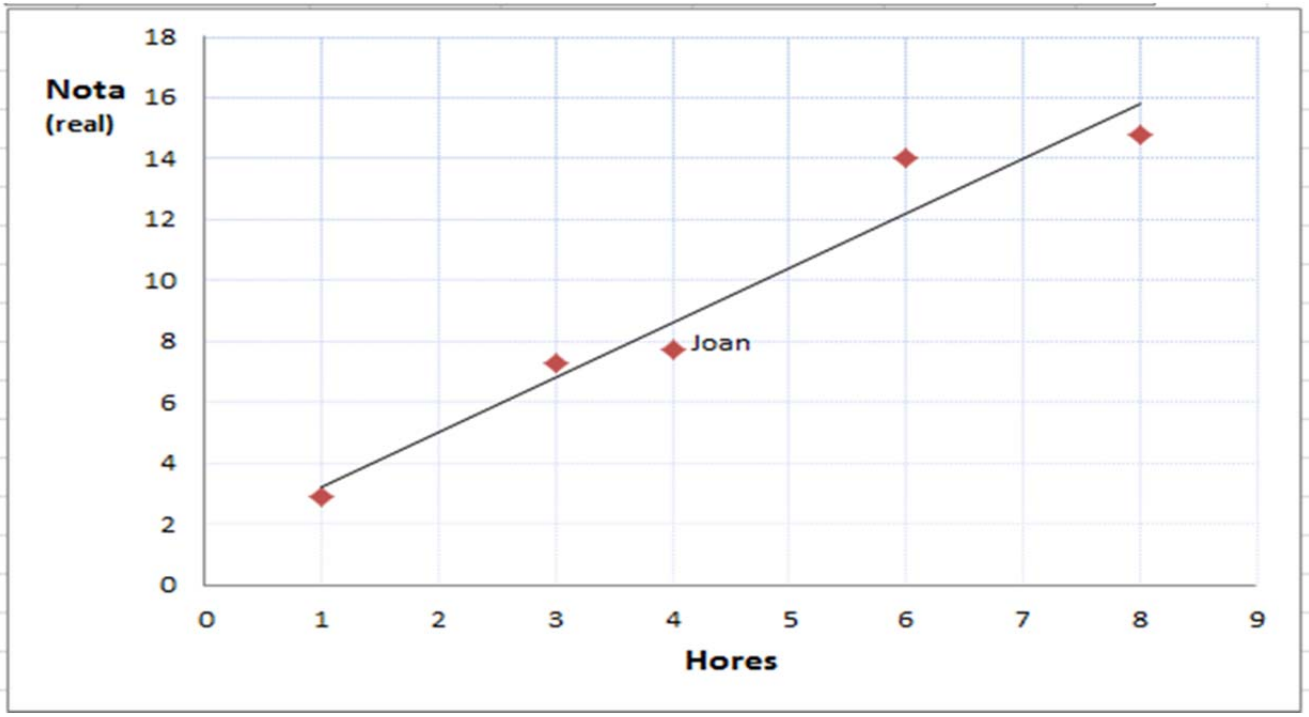


- Com podríem trobar/estimar els valors de β_1 i β_2? Recorda que són desconeguts.

"Podem pensar que les dades realment estan generades per l'MRL $(Y_i = \beta_1 + \beta_2 X_i + u_i)$ *i pensar que Y és explicat fonamentalment per X, mentre que el component aleatori (u) és de menor importància."*

Estimant els β… a ull!

- Com una primera aproximació es podria traçar una recta que passarà tan a prop de les dades reals, però... no és un criteri molt científic.



- Quina recta triem? (En realitat ens preguntem quins són els "millors" valors per als paràmetres β_1 i β_2?)

- Sembla lògic triar la recta que siga més pròxima a les dades. Més pròxima?

Com triem la recta? Com estimem β_1 i β_2?

- Triarem com a estimacions de β aquells valors que seleccionen la recta que s'aproxima més a les dades observades; és a dir, la que menys s'equivoca o menys s'allunya de les dades.

- De moment no estimarem, sinó que seleccionarem una recta, que representem com a:

$$\hat{Y}_i = \widehat{\beta_1} + \widehat{\beta_2} X_i$$

MODEL "ESTIMAT" (o recta de regressió)

Els valors estimats per a $\widehat{\beta_1}$ i $\widehat{\beta_2}$ s'anomenen **valors estimats o estimacions** dels paràmetres desconeguts β_1 i β_2

- Fixem-vos que tenim Y i $\hat{Y}$. Tenim dades reals (Y_i) i dades estimades pel model ($\widehat{Y_i}$)

- També podem veure en quant s'equivoca el nostre model estimat i definir:

$$\hat{u}_i = Y_i - \hat{Y}_i$$

RESIDU o error d'estimació

Recapitulant... Quants models tenim?

$$Y_i = \widehat{\beta_1} + \widehat{\beta_2} X_i + u_i$$ **MODEL TEÒRIC (MRL)**

$$\hat{Y}_i = \widehat{\beta_1} + \widehat{\beta_2} X_i$$ **MODEL "ESTIMAT"** (o recta de regressió)

$$\hat{u}_i = Y_i - \hat{Y}_i$$ **RESIDU** o error d'estimació

Per exemple: Suposa que amb les dades anteriors obtenim/estimem el següent model

$$\text{Nota}_i = \beta_1 + \beta_2\, \text{hores}_i + u_i$$

$$\widehat{\text{Nota}}_i = 0.6 + 2.1\, \text{hores}_i$$

i	Hores	Nota (real)	Nota estimada	Residus
Ariadna	3	7.3		
Gregori	8	14.8	17,4 (=0,6+2,1*8)	-2.4 (= 14,8-17,4)
Albert	1	2.9		
Joan	4	7.7		
Maria	6	14		

Cal "reconèixer" cada un dels elements del nostre esquema

- Model teòric *vs.* model estimat. Quin és observable? Quan és observable?
- Y *vs.* Y-estimat. Com s'interpreten? Són observables?
- Pertorbació *vs.* residu. Què són? Són observables?
- Paràmetres *vs*. estimacions/estimadors?

Una altra vegada... açò és bàsic per a entendre la matèria

- Per a cada observació (o individu "i") hi ha un valor predit pel model estimat ($\hat{Y}_i$) i un error d'estimació ($\hat{u}_i$).
- No s'han de confondre dades reals (X_i , Y_i) amb dades predites (X_i , $\hat{Y}_i$)
- No s'han de confondre pertorbacions (u_i) amb residus ($\hat{u}_i$).

Estimant el model (estimant els β)

- Com estimem els paràmetres del model teòric? $\boxed{Y_i = \widehat{\beta_1} + \widehat{\beta_2} X_i + u_i}$

- Seleccionant la recta (els $\hat{\beta}$) de manera que es minimitzen les equivocacions o residus.

- Quin criteri utilitzem?

- Criteri per a estimar: minimitzar la suma dels quadrats dels residus (MQO)

$$Min \sum_{i=1}^{n} \hat{u}_i^2 = Min \ (\hat{u}_1^2 + \hat{u}_2^2 + \cdots + \hat{u}_n^2)$$

Estimació de l'MRLS amb el mètode MQO (mínim quadràtic)

- Criteri per a estimar: $$Min \sum_{i=1}^{n} \hat{u}_i^2$$

- Com que $\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)$, l'expressió que cal minimitzar és:

- $$\underset{\{\hat{\beta}_1, \hat{\beta}_2\}}{Min} : \quad S = \sum_{i=1}^{N} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

- Derivant $$\frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{N}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) \qquad \frac{\partial S}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{N}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i$$

- Igualant a zero (condició de mínim) i aïllant arribem a les **equacions normals**

$$\sum_{i=1}^{N} Y_i = \hat{\beta}_1 N + \hat{\beta}_2 \sum_{i=1}^{N} X_i \qquad \sum_{i=1}^{N} Y_i X_i = \hat{\beta}_1 \sum_{i=1}^{N} X_i + \hat{\beta}_2 \sum_{i=1}^{N} X_i^2$$

Resolent el sistema d'equacions (obtenció dels estimadors MQO)

$$\sum_{i=1}^{N} Y_i = \hat{\beta}_1 N + \hat{\beta}_2 \sum_{i=1}^{N} X_i \qquad \sum_{i=1}^{N} Y_i X_i = \hat{\beta}_1 \sum_{i=1}^{N} X_i + \hat{\beta}_2 \sum_{i=1}^{N} X_i^2$$

- Aïllant β_1 de la primera equació obtenim:

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$$

- Substituint β_1 en la segona equació s'obté:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

- Les expressions matemàtiques per a $\hat{\beta}_1$ i $\hat{\beta}_2$ s'anomenen **<u>estimadors MQO</u>** dels paràmetres desconeguts.
- Els valors estimats per a $\hat{\beta}_1$ i $\hat{\beta}_2$ s'anomenen **valors estimats o <u>estimacions</u> per MQO** dels paràmetres desconeguts.

Exemple d'estimació (fent els càlculs a mà)

- Exercici 2.1 del qüestionari

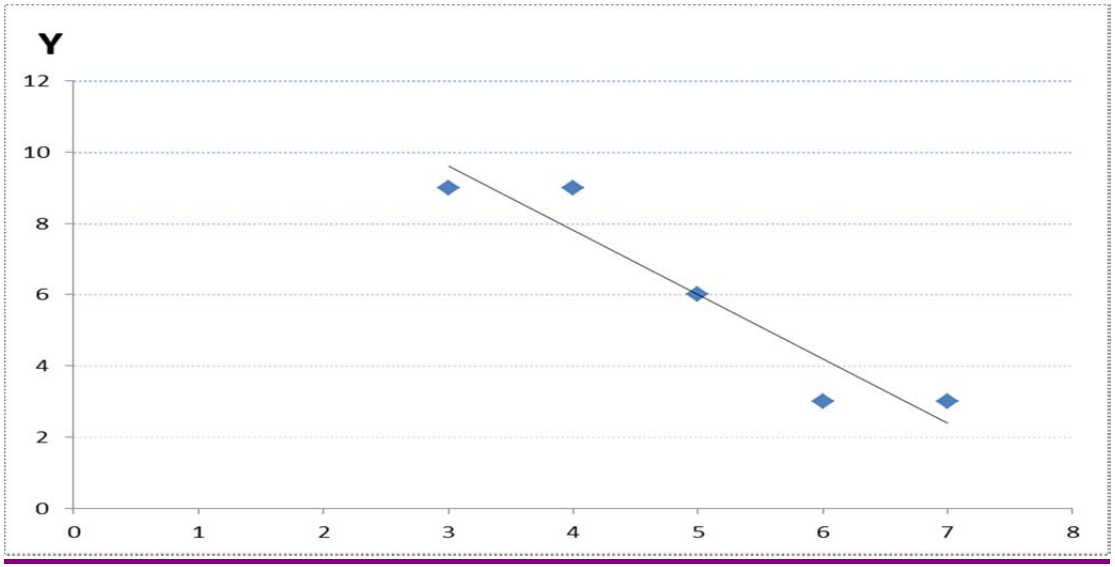
	X	**Y**	$(X-\bar{X})$	$(X-\bar{X})^2$	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
Juan	5	6	0	0	0	0	0
Carlos	7	3	2	4	-3	9	-6
Susana	4	9	-1	1	3	9	-3
Veronica	6	3	1	1	-3	9	-3
Andrea	3	9	-2	4	3	9	-6
Suma	*25*	*30*	*0*	*10*	*0*	*36*	*-18*
Mitjana	5	6		2		7.2	-3.6

- $$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{-18}{10} = -1{,}8 = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

- $$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X} = 6 - (1{,}8*6) = 15$$

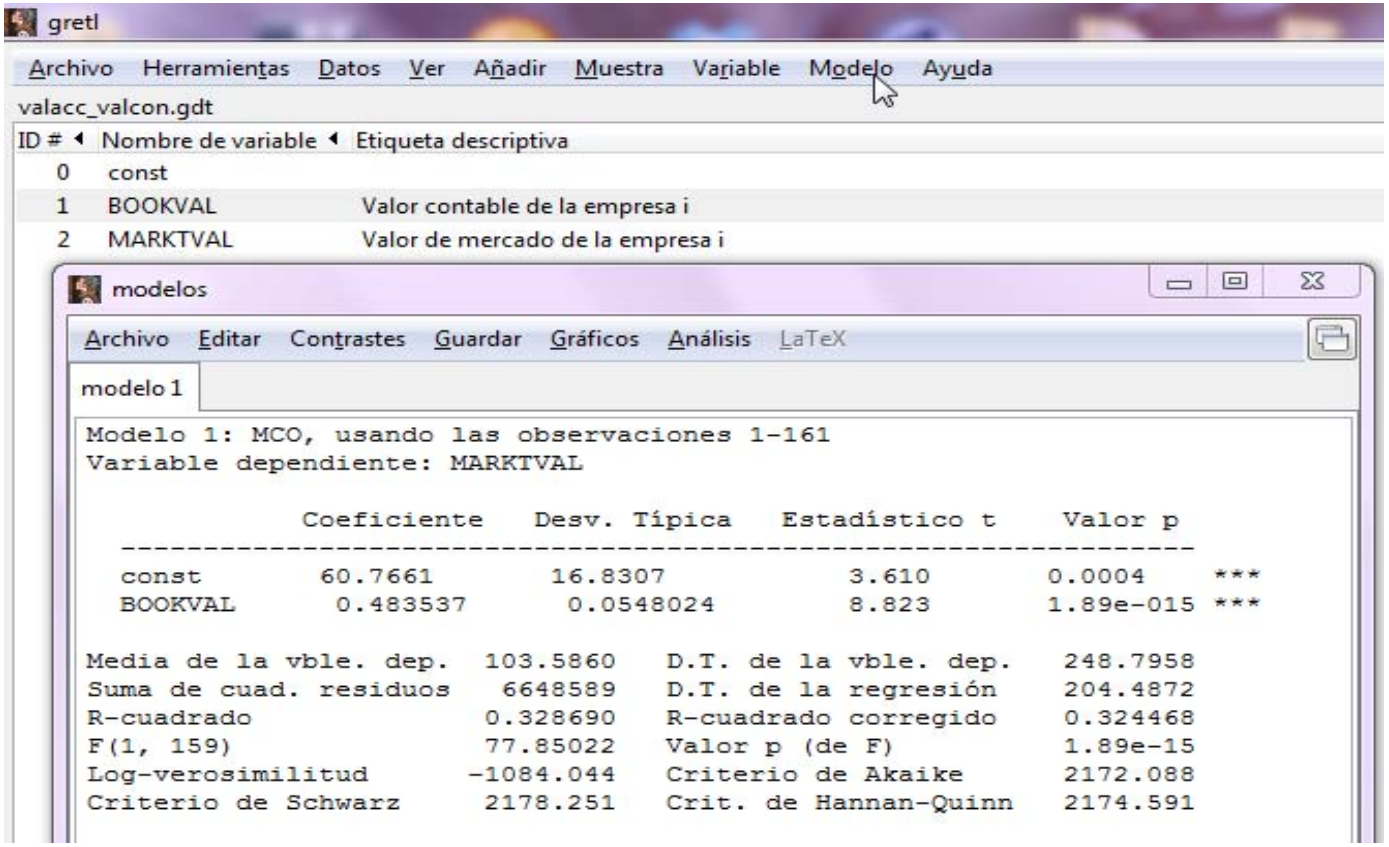
$$\hat{Y}_i = 15 - 1{,}8\, X_i$$

	X	Y	$\hat{Y}$	$\hat{u}$
Joan	5	6	6	0
Carles	7	3	2.4	0.6
Gregori	4	9	7.8	1.2
Anna	6	3	4.2	-1.2
Ariadna	3	9	9.6	-0.6

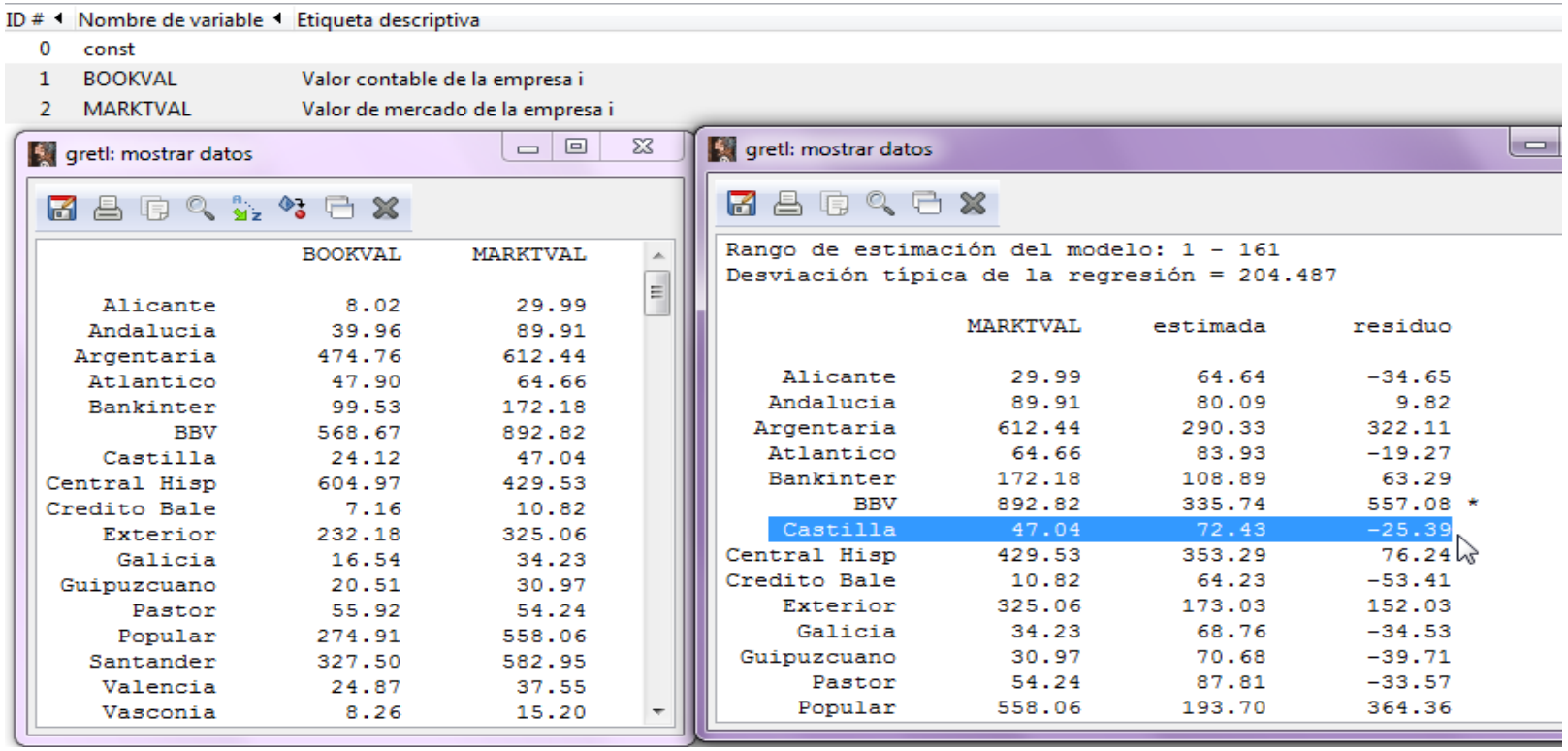


Estimació d'1 MRLS (amb programari)

- Carreguem les dades **bolmad95.gtd** a http://www.uv.es/~uriel/ficheroesgdt.htm
- O carreguem **Valacc_valcon.gdt** en l'aula virtual (és el mateix fitxer)



➢ Interpreteu les estimacions dels paràmetres



ID #	Nombre de variable	Etiqueta descriptiva
0	const	
1	BOOKVAL	Valor contable de la empresa i
2	MARKTVAL	Valor de mercado de la empresa i

gretl: mostrar datos

	BOOKVAL	MARKTVAL
Alicante	8.02	29.99
Andalucia	39.96	89.91
Argentaria	474.76	612.44
Atlantico	47.90	64.66
Bankinter	99.53	172.18
BBV	568.67	892.82
Castilla	24.12	47.04
Central Hisp	604.97	429.53
Credito Bale	7.16	10.82
Exterior	232.18	325.06
Galicia	16.54	34.23
Guipuzcuano	20.51	30.97
Pastor	55.92	54.24
Popular	274.91	558.06
Santander	327.50	582.95
Valencia	24.87	37.55
Vasconia	8.26	15.20

gretl: mostrar datos

Rango de estimación del modelo: 1 - 161
Desviación típica de la regresión = 204.487

	MARKTVAL	estimada	residuo	
Alicante	29.99	64.64	-34.65	
Andalucia	89.91	80.09	9.82	
Argentaria	612.44	290.33	322.11	
Atlantico	64.66	83.93	-19.27	
Bankinter	172.18	108.89	63.29	
BBV	892.82	335.74	557.08	*
Castilla	47.04	72.43	-25.39	
Central Hisp	429.53	353.29	76.24	
Credito Bale	10.82	64.23	-53.41	
Exterior	325.06	173.03	152.03	
Galicia	34.23	68.76	-34.53	
Guipuzcuano	30.97	70.68	-39.71	
Pastor	54.24	87.81	-33.57	
Popular	558.06	193.70	364.36	

- Obtingueu (a mà) el valor de mercat i el valor comptable d'Argentaria.
- Obtingueu (a mà) la Y estimada i el residu per a Argentaria. Interpreteu-ho.

TEMA 2. EL MODEL DE REGRESSIÓ LINEAL (MRL)

2.1 El model de regressió lineal simple (MRLS)

2.2 El model de regressió lineal múltiple (MRLM)

2.3 Interpretació de coeficients

2.4 Unitats de mesura i formes funcionals

Model de regressió lineal múltiple (MRLM)

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots \beta_k X_{ki} + u_i \qquad i = 1,2,....n$$

- En l'MRLM el regressand és funció lineal de k-1 variables explicatives i d'una pertorbació aleatòria. L'MRLM incorpora K regressors (k-1 variables explicatives més el terme constant)

Avantatges de la regressió múltiple

- És evident que *Y* pot ser influenciada o explicada per més d'una variable, per tant necessitem ampliar el model per poder controlar explícitament per altres variables i estimar amb més precisió l'efecte d'una *X* en *Y*.

- L'MRLM, com que introdueix més regressors, pot explicar una gran part de la variació en *Y*. D'aquesta manera també pot proporcionar millors prediccions de *Y*.

- Un altre avantatge addicional és que permet incorporar relacions funcionals entre *X* i *Y* molt generals. Per exemple:

$$consum = \beta_1 + \beta_2\, renda + \beta_3\, renda^2 + u$$

El model amb k regressors

- L'MRL**M**

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots \beta_k X_{ki} + u_i \qquad i = 1,2,....n$$

- L'MRLM té moltes similituds amb l'MRL**S**. En concret, s'utilitza la mateixa terminologia:

  - Ens referirem a la variable *Y* com a variable dependent.
  - Ens referirem a les variables X com a regressors.
  - Ens referirem a la variable *u* com a pertorbació aleatòria. La pertorbació aleatòria (*u*) continua representant **tots els altres factors** (a part de les *X*) que afecten la variable *Y*.
  - β_1, β_2 …, β_k són paràmetres desconeguts que volem estimar.
  - β_j, j = 2, 3, …, k es coneixen com a paràmetres de pendent i indiquen el canvi experimentat en *Y* quan varia X_j mantenint fixos els restants factors que afecten *Y*.

Interpretació dels β

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots \beta_k X_{ki} + u_i \qquad i = 1,2,....n$$

- Com que la forma funcional és lineal, si els altres factors no canvien (*ceteris paribus*), aleshores un determinat *X* presentarà un efecte lineal (o constant) en *Y*:

$$\Delta Y = \beta_3 \Delta X_3$$

  - si X_3 augmenta en 1-unitat, Y augmenta en β_3 unitats

  - l'efecte d'augmentar X_3 en 1-unitat és independent del valor inicial de *X; és a dir, és* constant.

- Per exemple: *nota* = 0,5 + 1,5 *hores* + 0,8 *assistència* + *u*

  - Com s'interpreta β_2 en l'exemple anterior?
  - Com s'interpretaria β_1?

Comparant estimacions: MRLS *vs.* MRLM

Modelo 14: MCO, usando las observaciones 1-526
Variable dependiente: salari

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	-0.904852	0.684968	-1.321	0.1871	
educacio	0.541359	0.0532480	10.17	2.78e-022	***

Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086
Suma de cuad. residuos	5980.682	D.T. de la regresión	3.378390
R-cuadrado	0.164758	R-cuadrado corregido	0.163164
F(1, 524)	103.3627	Valor p (de F)	2.78e-22
Log-verosimilitud	-1385.712	Criterio de Akaike	2775.423
Criterio de Schwarz	2783.954	Crit. de Hannan-Quinn	2778.764

Modelo 15: MCO, usando las observaciones 1-526
Variable dependiente: salari

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	-2.87273	0.728964	-3.941	9.22e-05	***
educacio	0.598965	0.0512835	11.68	3.68e-028	***
experiencia	0.0223395	0.0120568	1.853	0.0645	*
antiguitat	0.169269	0.0216446	7.820	2.93e-014	***

Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086
Suma de cuad. residuos	4966.303	D.T. de la regresión	3.084476
R-cuadrado	0.306422	R-cuadrado corregido	0.302436
F(3, 522)	76.87317	Valor p (de F)	3.41e-41
Log-verosimilitud	-1336.831	Criterio de Akaike	2681.662
Criterio de Schwarz	2698.723	Crit. de Hannan-Quinn	2688.342

- Quin efecte té l'educació en el salari?

Estimant els β en l'MRLM

- Com estimem els β?

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots \beta_k X_{ki} + u_i \qquad i = 1,2,....n$$

- Seleccionant la ~~recta~~ hiperplà (en realitat seleccionant els $\hat{\beta}$) de manera que es minimitzen les equivocacions o residus.

- Criteri per a estimar: MQO

$$Min \sum_{i=1}^{n} \hat{u}_i^2$$

- Com que $\hat{u}_i = Y_i - \hat{Y}_i$, l'expressió que cal minimitzar és:

$$Min\ S = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} \cdots - \hat{\beta}_k X_{ki} \right)^2$$

- (Derivant). Ara cal derivar respecte dels k estimadors, per això tindrem k equacions normals

k primeres derivades (de la funció objectiu)

- (Derivant). Ara cal derivar respecte dels k estimadors.

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} \cdots - \hat{\beta}_k X_{ki}\right)$$

$$\frac{\partial S}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} \cdots - \hat{\beta}_k X_{ki}\right)X_{2i}$$

$$\vdots$$

$$\frac{\partial S}{\partial \hat{\beta}_k} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} \cdots - \hat{\beta}_k X_{ki}\right)X_{ki}$$

➢ Com resolem un sistema de k equacions?

Sistema de k equacions normals

- Igualant a zero la primera derivada (1a condició de mínim) i aïllant, obtenim les k **equacions normals**

$$\sum_{i=1}^{n} Y_i = \hat{\beta}_1 n + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{ki}$$

$$\sum_{i=1}^{n} Y_i X_{2i} = \hat{\beta}_1 \sum_{i=1}^{n} X_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2 + \cdots \hat{\beta}_k \sum_{i=1}^{n} X_{2i} X_{ki}$$

$$\vdots$$

$$\sum_{i=1}^{n} Y_i X_{ki} = \hat{\beta}_1 \sum_{i=1}^{n} X_{ki} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i} X_{ki} + \cdots \hat{\beta}_k \sum_{i=1}^{n} X_{ki}^2$$

Resolent el sistema d'equacions (obtenció dels estimadors MQO)

- Sistema de k equacions normals. Quines són les incògnites?
- Com resolem un sistema de k equacions amb k incògnites? Gauss, Cramer, inversa.
- Podeu consultar-ho en la bibliografia.

Interpretant les estimacions

- Es planteja el model següent $nota = \beta_1 + \beta_2\, hores + \beta_3\, assistència + u$
  - Com s'interpreta β_2 en l'exemple anterior?
- Aconseguim dades i estimem el model anterior per MQO

$$nota = 0{,}4 + 1{,}3\, hores + 0{,}9\, assistència \quad (?)$$

  - Com s'interpreta $\hat{\beta}_2$ en l'exemple anterior?
  - Com s'interpretaria $\hat{\beta}_3$?

TEMA 2. EL MODEL DE REGRESSIÓ LINEAL (MRL)

2.1 El model de regressió lineal simple (MRLS)

2.2 El model de regressió lineal múltiple (MRLM)

2.3 Interpretació de coeficients

2.4 Unitats de mesura i formes funcionals

Què passa amb les estimacions si canvien les unitats de mesura?

A) Canvis d'escala

- Si X es **multiplica**/divideix per una constant ($c \neq 0$) el pendent estimat es **divideix**/multiplica per la mateixa constant, c. L'estimació de l'ordenada en l'origen no canvia.

- Si *Y* es **multiplica**/divideix per una constant ($c \neq 0$) tant l'ordenada en l'origen com els pendents estimats es **multipliquen**/divideixen per la mateixa constant, c.

B) Canvis d'origen (tant en X com en *Y*)

- **NO** afecten l'estimació dels pendents.

- **SÍ** que afecten l'estimació de l'ordenada.

No linealitats i forma funcional

- En economia hi ha moltes relacions que poden ser no lineals. Les podem tractar amb el nostre MRLS?
- En molts casos sí, simplement redefinint la variable dependent i independent.

- Per exemple: $y = \beta_0 + \beta_1 x^2 + u$ o $\log(y) = \beta_0 + \beta_1 x + u$

  ➢ Com s'interpreta β_1 en els models anteriors?

Forma funcional i interpretació de β

- Per a interpretar les especificacions funcionals més habituals en econometria és útil:

Model	v. dependent	v. independent	**Interpretació de β**
nivell-nivell	y	x	$\Delta y = \beta\, \Delta x$
nivell-log	y	$\log(x)$	$\Delta y = (\beta/100)\Delta\%x$
log-nivell	$\log(y)$	x	$\Delta\%y = (100\ \beta)\Delta x$
log-log	$\log(y)$	$\log(x)$	$\Delta\%y = \beta\Delta\%x$

on $\Delta\%x = \frac{x_1 - x_0}{x_0}.100 = \frac{\Delta x}{x_0}.100$

Recordant que $\Delta\log(x) = \log(x_1) - \log(x_0) \approx \frac{x_1 - x_0}{x_0} = \frac{\Delta x}{x_0}$. Per tant, $100.\Delta\log(x) \approx \Delta\%x$

També cal recordar que $d\log(y) = \frac{1}{y}dy$

El significat de "model lineal"

- En aquest tema hem analitzat el model de regressió (lineal i múltiple); tanmateix, acabem de veure que el nostre esquema permet també analitzar relacions no lineals transformant adequadament les variables.

- La clau està en el fet que el model ha de ser lineal en els paràmetres (o linealitzable) i el terme de pertorbació ha de ser additiu.

- Per tant, és important recordar que encara que la mecànica de l'estimació per MQO no depèn de com estan definits X i Y, la interpretació dels coeficients **SÍ** que depèn de la forma funcional de les variables.

- Naturalment, hi ha models que no s'ajusten al format de l'MRLS. Caldria estimar aquests models per altres mètodes.

$$Y = \frac{1}{\beta_1 + \beta_2 X} + u \qquad\qquad Y = \beta_1 e^{\beta_2 X} + u$$

Tema 3. Propietats i hipòtesis en el model de regressió

3.1 Propietats descriptives de la regressió: implicacions algèbriques de l'estimació

3.2 Mesures de bondat de l'ajust: el coeficient de determinació i selecció de models (AIC)

3.3 Supòsits del model de regressió lineal clàssic

3.4 Propietats probabilístiques del model

TEMA 3. PROPIETATS I HIPÒTESIS EN EL MODEL DE REGRESSIÓ

3.1 Propietats descriptives de la regressió: implicacions algèbriques de l'estimació

3.2 Mesures de bondat de l'ajust: el coeficient de determinació i selecció de models (AIC)

3.3 Supòsits del model de regressió lineal clàssic

3.4 Propietats probabilístiques del model

Propietats descriptives

- Necessàriament, si estimes un MRL pel mètode de MQO es compleixen les propietats següents:

1) La suma dels residus MCO és zero: $\boxed{\sum_{i=1}^{n} \hat{u}_i = 0}$

Per tant, també es complirà que

a) La mitjana mostral dels residus és zero $\bar{\hat{u}} = 0$

b) $\bar{Y} = \bar{\hat{Y}}$

Demostració (mireu la primera equació normal):

$$\sum_{i=1}^{N} Y_i = \hat{\beta}_1 N + \hat{\beta}_2 \sum_{i=1}^{N} X_i \quad (1^{\text{a}} \text{ eq. normal})$$

que ve de...

$$-2\sum_{i=1}^{N} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\hat{u}_i = (Y_i - \hat{Y}_i) = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)$$

$$\sum \hat{u}_i = \sum Y_i - N\hat{\beta}_1 - \hat{\beta}_2 \sum x_i$$

$$\bar{\hat{u}} = \bar{Y} - \hat{\beta}_1 - \hat{\beta}_2 \bar{X} = \bar{Y} - (\bar{Y} - \hat{\beta}_2 \bar{X}) - \hat{\beta}_2 \bar{X} = 0$$

2) La covariància mostral entre regressor i residus MQO és zero. $\sum_{i=1}^{n} X_i \hat{u}_i = 0$

(Demostració: Mireu la segona equació normal abans d'aïllar)

$$-2\sum_{i=1}^{N}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)\, X_i = 0$$

$$\sum_{i=1}^{N} Y_i X_i = \hat{\beta}_1 \sum_{i=1}^{N} X_i + \hat{\beta}_2 \sum_{i=1}^{N} X_i^2 \quad (2^{a} \text{ eq. normal})$$

3) La recta de regressió MQO passa pel punt de mitjanes mostrals. $\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}$

(Demostració: Mireu la primera equació normal o l'expressió de $\hat{\beta}_1$)

$$\sum_{i=1}^{N} Y_i = \hat{\beta}_1 N + \hat{\beta}_2 \sum_{i=1}^{N} X_i \qquad (1^{a} \text{ eq. normal})$$

4) La covariància mostral entre valors ajustats i residus és zero. $\sum_{i=1}^{N} \hat{Y}_i \hat{u}_i = 0$

(Demostració: Substituïu $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ i utilitzeu les propietats 1 i 2 anteriors)

$$\sum_{i=1}^{N} \hat{u}_i \hat{Y}_i = \sum \hat{u}_i (\hat{\beta}_1 + \hat{\beta}_2 X_i) = \hat{\beta}_1 \sum \hat{u}_i + \hat{\beta}_2 \sum \hat{u}_i X_i = 0 + 0 = 0$$

Residus i regressors incorrelats. Què significa això?

- El MRL planteja que Y és explicat per les X i la pertorbació (u). El model aconsegueix explicar una part de Y, la $\hat{Y}$, i en deixa una altra part per explicar, la $\hat{u}$. La raó que una part de Y no siga explicada per X és que eixa part de Y no està relacionada amb X, o siga, que el residu no està correlacionat (en la mostra) amb X: $\sum_{i=1}^{N} \hat{Y}_i \hat{u}_i = 0$

Y explicada i residus incorrelats (part explicada enfront de part no explicada)

- De la definició de residu se segueix que: $\boxed{Y_i = \hat{Y}_i + \hat{u}_i}$

És a dir, podem veure que MQO descompon cada valor de Y en dues parts: un valor ajustat i un residu. (A més, $\hat{Y}_i$ i $\hat{u}_i$ estan incorrelats en la mostra, quarta propietat descriptiva.)

TEMA 3. PROPIETATS I HIPÒTESIS EN EL MODEL DE REGRESSIÓ

3.1 Propietats descriptives de la regressió: implicacions algèbriques de l'estimació

3.2 Mesures de bondat de l'ajust: el coeficient de determinació i selecció de models (AIC)

3.3 Supòsits del model de regressió lineal clàssic

3.4 Propietats probabilístiques del model

Bondat de l'ajust

Podem pensar que el propòsit de l'anàlisi de regressió és explicar el comportament de la variable dependent (Y).
Una vegada hem estimat per MQO un MRLS, ens interessaria tenir una mesura que ens informara del grau d'ajust entre el model i les dades: un estadístic que ens informe de la bondat de l'ajust.

Construcció de R^2. (Part explicada / part no explicada)

- El model estimat no explica o prediu perfectament les observacions de la variable Y, deixa un residu o error d'estimació:

$$Y_i = \hat{Y}_i + \hat{u}_i$$

- És a dir, podem descompondre cada valor de la variable endògena (Y) en dues parts, la part explicada pel model ($\hat{Y}$) i la part que el model no aconsegueix explicar: el residu ($\hat{u}$). A més, la part explicada i la no explicada estan incorrelades en la mostra (quarta propietat descriptiva).

La variància d'una suma …

- Y és la suma de dues variables : $Y_i = \hat{Y}_i + \hat{u}_i$

$$Var(Y) = Var(\hat{Y} + \hat{u}) = Var(\hat{Y}) + Var(\hat{u}) + 2Cov(\hat{Y}, \hat{u})$$

$$Var(Y) = Var(\hat{Y}) + Var(\hat{u})$$

Definició de SQT, SQE i SQR (numeradors de les variàncies)

- Suma de quadrats totals : $SQT = \sum_{i=1}^{n} (Y_i - \overline{Y}_i)^2$
- Suma de quadrats explicats : $SQE = \sum_{i=1}^{n} (\hat{Y}_i - \overline{\hat{Y}})^2$
- Suma de quadrats residuals : $SQR = \sum_{i=1}^{n} \hat{u}_i^2$

⟹ Es pot demostrar que **SQT = SQE+SQR**

Bondat de l'ajust

- Fins a quin punt s'ajusta la recta de regressió a les dades mostrals?
- Partint de que SQT = SQE+SQR; per tant, podem calcular la fracció de la SQT que és explicada pel model estimat i definir així el coeficient de determinació o R quadrat (R^2) de la regressió:

$$R^2 = \frac{SCE}{SCT} \qquad \text{ó} \qquad R^2 = 1 - \frac{SCR}{SQT}$$

Interpretació i propietats de R^2

- El valor de R^2 està fitat en [0,1] *(si el model té terme independent)*
  - Si $R^2 = 1$, la recta de regressió s'ajusta perfectament a les dades, per tant tots els residus són zero.
  - Si $R^2 = 0$, o està prop de zero, hi ha un ajust pobre: la variació de *Y* està poc representada per la recta de regressió.
- $100*R^2$ és el percentatge de la variació mostral de *Y* que és explicada pel model.

Més sobre R^2

- R^2 mesura fins a quin punt s'ajusta bé la recta de regressió a les dades.
- Un R^2 baix no significa que el model estimat no proporciona estimacions fiables dels paràmetres; pot simplement significar que en les ciències socials és difícil predir el comportament individual.
- Atès que volem models amb poder explicatiu alt, si hi ha igualtat d'altres factors triarem models amb R^2 elevats...
- ... però, la bondat de l'ajust no és l'única característica que hem de valorar en una equació de regressió.
- De vegades s'utilitza R^2 per a comparar models, però per poder comparar dos models sobre la base del R^2 aquests models **han de tenir estricament el mateix regressand**, la mateixa grandària mostral **i el mateix nombre de regressors**.

- R^2 no disminueix quan afegim variables explicatives, normalment augmenta.

- De vegades s'utilitza R^2 per a comparar models, però per poder comparar dos models sobre la base del R^2 aquests models **han de tenir estrictament el mateix regressand**, la mateixa grandària mostral **i el mateix nombre de regressors**.

R^2 corregit o ajustat

- R^2 no disminueix mai quan afegim regressors al model. Podem ajustar-lo per tenir en compte aquest fet.

- $$\overline{R}^2 = 1 - \frac{[SCR/(N-k)]}{[SCT/(n-1)]}$$

- $\overline{R}^2$ penalitza els models que afegeixen molts regressors. R^2 no pot disminuir quan introduïm en el model un nou regressor, però $\overline{R}^2$ sí. Vegem-ho…

- …quan introduïm un regressor més al model, SQR generalment cau, però els graus de llibertat (N-k) també, per la qual cosa el canvi en $\overline{R}^2$ no està determinat.

R^2 i R^2 corregit

- $\overline{R}^2$ i R^2 estan relacionats: $\overline{R}^2 = 1-(1-R^2)\frac{N-1}{N-k}$. Per tant:
  - Si k = 1, aleshores $\overline{R}^2 = R^2$
  - Si k>1, aleshores $\overline{R}^2 < R^2$
  - Si R^2= 1, aleshores $\overline{R}^2 = 1$
  - $\overline{R}^2$ pot ser negatiu (de fet, ho serà si k > 1 i R^2 = 0)

- $\overline{R}^2$ **es pot utilitzar per a comparar models amb distint nombre de regressors** (però, això sí, recordeu que el regressand dels models ha de ser el mateix). La raó és senzilla, la SQT a explicar serà distinta en models amb diferent regressand.

Altres mesures de bondat d'ajust (AIC)

Estadístic AIC (o criteri d'informació d'Akaike)

$$AIC = -\frac{2}{T}\ln L(y,\tilde{\beta}) + \frac{2k}{T} \qquad AIC = 1 + \ln(2\pi) + \ln\frac{SQR}{n} + \frac{2(k+1)}{n}$$

CRITERI: Un valor menor de l'AIC implica un millor ajust del model.

a) No és una mesura de caràcter relatiu com el R^2. El seu valor no indica per si mateix un ajust elevat o reduït, però **sí que permet comparar models**.

b) L'AIC es pot aplicar a models sense terme constant.

c) L'AIC penalitza la introducció de regressors addicionals, per tant permet comparar models amb diferent nombre de regressors.

d) L'AIC no es pot utilitzar tampoc per a comparar models amb diferent regressand. No obstant això, és possible trobar-ne una transformació que permeta la comparació.

Així doncs, malgrat que té certs avantatges respecte al R^2, l'AIC no és una solució definitiva per a jutjar quin model és l'adequat.

- Suposeu que l'AIC_1 = -2,32 i l'AIC_2 = -2,54. Quin d'ells seleccionaríeu?

TEMA 3. PROPIETATS I HIPÒTESIS EN EL MODEL DE REGRESSIÓ

3.1 Propietats descriptives de la regressió: implicacions algèbriques de l'estimació

3.2 Mesures de bondat de l'ajust: el coeficient de determinació i selecció de models (AIC)

3.3 Supòsits del model de regressió lineal clàssic

3.4 Propietats probabilístiques del model

MQO és un bon mètode d'estimació? Ens en podem fiar?

Recordem el nostre plantejament: Partim d'un model Y = f(X´s) i volem conèixer la influència de les variables X en Y, és a dir, volem estimar els paràmetres β. Per a això plantegem l'estimació del model (MRL) per MQO i obtenim els estimadors mínims quadràtics ($\hat{\beta}$).

Per a una mostra de dades concreta obtindrem una estimació dels paràmetres. Per exemple: suposeu que l'estimació de β_3 és 0,12. En aquest cas, 0,12 és una estimació puntual per a β_3. Però...

- Quina confiança tenim que β_3 és efectivament 0,12? És molt difícil que β_3 siga exactament 0,12. Com a molt, si MQO fóra un bon mètode podríem pensar que efectivament β_3 estarà prop de 0,12.

- Podem donar un rang de valors entre els quals es trobe β_3 amb una alta probabilitat? Al final del tema podrem; però abans, per veure la importància d'això, imagineu-vos que l'estimació puntual fóra 0,13, però:

    a) l'interval d'estimació fóra [0,11; 0,15]
    a) l'interval d'estimació fóra [-0,27; 0,33]

Probabilitat i inferència

- Com veieu, entrem en el terreny de la probabilitat i la inferència, però abans hem de tancar el nostre model incorporant-hi una sèrie de supòsits. Començarem amb els més bàsics o clàssics.

- Un resultat molt important que obtindrem consisteix que si es compleixen una sèrie de supòsits o hipòtesis estadístiques bàsiques (h.e.b.) el mètode de MQO és un bon mètode per a estimar els efectes de les X en *Y*. En concret, si es compleixen les h.e.b., els estimadors MQO del MRL són ELIO.

- Per contra, si alguna de les h.e.b. no es compleix, MQO pot deixar de ser ELIO.

Hipòtesis estadístiques bàsiques (h.e.b.) …

I) Hipòtesis sobre la forma funcional:

1) El model és lineal: $$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots \beta_k X_{ki} + u_i \qquad i = 1,2,....n$$

II) Hipòtesis sobre la pertorbació aleatòria:

2) u_i son v.a. no observables.

3) E(u_i) = 0 $\forall$i = 1, …… n

4) Var(u_i) = σ^2 $\forall$i = 1, …… n (HOMOCEDASTICITAT)

5) Cov(u_i , u_s) = 0 ; $\forall$ i $\neq$ s (NO AUTOCORRELACIÓ)

6) La pertorbació aleatòria segueix una distribució normal multivariant (NORMALITAT)

Les quatre hipòtesis sobre $\boldsymbol{u}$ es poden expressar conjuntament de la forma següent:

$$u_i \rightarrow N(0 \ , \sigma_i^2)$$

… Hipòtesis estadístiques bàsiques (h.e.b.)

I) Hipòtesi sobre els regressors:

7) Els regressors són no estocàstics, o siga, *els regressors són fixos*.

7*) Els regressors es distribueixen independentment del terme de pertorbació: $E(X'u) = 0$

8) Els regressors són linealment independents, la qual cosa implica que no hi ha relacions lineals exactes entre els regressors (NO COL·LINEALITAT PERFECTA).

9) Els regressors no tenen errors de mesura.

II) Hipòtesi sobre β:

10) Els paràmetres del model són fixos.

TEMA 3. PROPIETATS I HIPÒTESIS EN EL MODEL DE REGRESSIÓ

3.1 Propietats descriptives de la regressió: implicacions algèbriques de l'estimació

3.2 Mesures de bondat de l'ajust: el coeficient de determinació i selecció de models (AIC)

3.3 Supòsits del model de regressió lineal clàssic

3.4 Propietats probabilístiques del model

Hem presentat les h.e.b. De seguida veurem **que si es compleixen les h.e.b.** s'obtenen diversos resultats (propietats probabilístiques) que ens permetran fer inferències i prediccions sobre els β i la Y.

Dels resultats que obtindrem, en destaca un: demostrarem que, si es compleixen les h.e.b., el mètode de MQO és un "*bon"* mètode per a estimar els paràmetres d'un MRL. En concret, els ***estimadors MQO seran centrats i òptims*** (ELIO).

Propietats probabilístiques (I)

En l'MRL, si es compleixen totes les h.e.b.: $$Y_i \rightarrow N(\beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}, \sigma^2)$$

I.1) Distribució del regressand

"Atès que Y_i és funció de la pertorbació aleatòria, Y_i serà també una va. Si u_i segueix una distribució normal, llavors Y_i també es distribuirà com una normal".

I.2) L'esperança de Y_i serà:

$$E(Y_i) = E(\beta_1 + \beta_2 X_{2i} + \cdots + \beta_{ki} X_{ki} + u_i) = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_{ki} X_{ki} + E(u_i) =$$
$$= \beta_1 + \beta_2 X_{2i} + \cdots + \beta_{ki} X_{ki} = \mu_i$$

I.3) La variància i la covariància seran:

$$\text{var}(Y_i) = E[Y_i - \mu_i]^2 = E(u_i^2) = \sigma^2 \qquad \forall i$$

$$\text{cov}(Y_i Y_j) = E[(Y_i - \mu_i)(Y_j - \mu_j)] = E(u_i u_j) = 0 \qquad \forall i \neq j$$

Per tant: $$\boxed{Y_i \rightarrow N(\beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki}, \sigma^2)}$$

- L'expressió estadística anterior, què vol dir en paraules?

Propietats probabilístiques (II: $\hat{\beta}$ són ELIO)

- Analitzarem les **propietats estadístiques** dels estimadors MQO. Són esbiaixats? Són precisos?

- No té sentit parlar de les propietats estadístiques de les estimacions, obtingudes amb una mostra; són propietats dels estimadors.

I) Distribució dels estimadors MQO: $\hat{\beta}_k \rightarrow N(\beta_k, \sigma^2_{\hat{\beta}_k})$ (¿?)

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

✓ Els estimadors es poden expressar com una funció lineal de Y_i. Per tant, com que Y_i es distribueix com una normal, el vector d'estimadors també té una distribució normal. (També es poden expressar com una funció lineal de u_i. Per tant, si u_i, per hipòtesi, segueix una distribució normal, els estimadors també ho faran.)

(II: $\hat{\beta}$ son ELIO)

Gauss i Markow demostraren que en presència de les h.e.b.[1] els estimadors mínim quadràtics són estimadors lineals, centrats i òptims (ELIO):

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

1. Linealitat Els estimadors es poden expressar com una funció lineal del regressand (i per tant de *u*).

2. Falta de biaix: $E(\hat{\beta}_k) = \beta_k$

3. Dins de la classe d'estimadors lineals i centrats, $\hat{\beta}_k$ té **mínima variància**, és a dir, és un estimador **òptim** (teorema de Gauss-Markov). Qualsevol altre estimador (dins dels lineals i centrats) té una variància major, com a màxim igual, que l'estimador mínim quadràtic.

➢ Què significa que els estimadors MQO siguen ELIO? Cal entendre bé això.

[1] No és necessari el supòsit de normalitat perquè els estimadors siguen ELIO.

El resultat anterior… és molt important

Resulta que, si es compleixen les h.e.b., $\hat{\beta}_k \rightarrow N(\beta_k, \sigma^2_{\hat{\beta}_k})$ i, a més, com demostraren Gauss i Markov, són òptims.

- ✓ Els estimadors MQO són variables aleatòries, perquè depenen de u. Segons els valors concrets que prenga la u eixiran unes estimacions concretes. No tots els valors de u són igual de probables; per tant...
- ✓ D'acord, són aleatoris, però (si es compleixen les h.e.b.) en coneixem la distribució, per la qual cosa podrem fer inferències sobre ells.
- ✓ Se'n fa una distribució normal centrada en el veritable valor de β (falta de biaix). La falta de biaix no ens garanteix que les nostres estimacions l'encertaran, sinó que si disposàrem de moltes mostres podríem estimar moltes vegades i l'encertaríem "en mitjana". No sabem si amb una estimació concreta encertem, però sí que sabem que, si estimàrem moltes vegades, la mitjana de les estimacions tendiria a encertar-la.
- ✓ La propietat de falta de biaix "no és la bomba". Solament ens garanteix que l'estimador MQO, si tinguérem moltes mostres, tendiria a encertar-la. El que no seria massa desitjable és tenir un estimador no centrat; és a dir, que ni tan sols l'encertara en mitjana.
- ✓ D'acord. La falta de biaix és una propietat desitjable, "però tampoc no és la bomba". Recordeu que Gauss-Markov van demostrar que la variància de l'estimador MQO és la mínima. La falta de biaix juntament amb la variància mínima ja sí que faran que "ens fiem" dels estimadors MQO i de les seues estimacions.
- ✓ La variància mínima juntament amb la falta de biaix fan que els estimadors MQO siguen els que maximitzen la probabilitat d'encertar (amb una sola estimació).

Puntualitzacions sobre la falta de biaix

- Una estimació no pot ser no esbiaixada, ja que és un valor fixat.

- Quan diem que MQO és centrat, el que volem dir és que el procediment pel qual s'obtenen les estimacions és un procediment sense biaix, és a dir, que veiem el procediment aplicat a totes les mostres aleatòries possibles.

- No podem garantir que amb una mostra concreta obtinguem una estimació propera al paràmetre poblacional, només ho esperem.

D'acord, els estimadors MQO són ELIO… però

Gauss-Markov van demostrar que (si es compleixen les h.e.b.) els estimadors són ELIO; és a dir, la variància de l'estimador MQO és la mínima… PERÒ, que siga mínima no garanteix que siga xicoteta.

$$\hat{\beta}_k \rightarrow N(\beta_k, \sigma^2_{\hat{\beta}_k})$$

La variància (estimada) dels estimadors… amb Gretl

```
Modelo 15: MCO, usando las observaciones 1-526
Variable dependiente: salari

              Coeficiente  Desv. Típica  Estadístico t   Valor p
  ------------------------------------------------------------------
  const        -2.87273      0.728964       -3.941      9.22e-05  ***
  educacio      0.598965     0.0512835       11.68       3.68e-028 ***
  experiencia   0.0223395    0.0120568        1.853      0.0645    *
  antiguitat    0.169269     0.0216446        7.820      2.93e-014 ***

Media de la vble. dep.   5.896103   D.T. de la vble. dep.    3.693086
Suma de cuad. residuos   4966.303   D.T. de la regresión     3.084476
R-cuadrado               0.306422   R-cuadrado corregido     0.302436
F(3, 522)                76.87317   Valor p (de F)           3.41e-41
Log-verosimilitud       -1336.831   Criterio de Akaike       2681.662
Criterio de Schwarz      2698.723   Crit. de Hannan-Quinn    2688.342
```

- On és troba la variància dels estimadors ($\hat{\sigma}^2_{\hat{\beta}_k}$)? Concretament, quin valor pren *la variància estimada per a l'efecte de l educació* ($\hat{\sigma}^2_{\hat{\beta}_{educació}}$)?

- Per a què necessitem calcular aquestes variàncies?

Estimador de la variància dels estimadors MQO

- Podem demostrar que si es compleixen les h.e.b. :

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SCT_j(1-R_j^2)} \qquad \text{amb} \qquad SQT_j = \sum (X_{ij} - \bar{X}_j)^2$$

R_j^2 és el R^2 de la regressió de X_j enfront de les restants X

- Podem calcular-ho? Què ens falta per a poder fer-ho?

Estimació de la variància de les pertorbacions (σ^2)

- Per a què necessitem estimar la variància de les pertorbacions? Ho acabem de veure: si volem fer inferències sobre els β_k , necessitem estimar la variància dels estimadors, i per a això necessitem estimar prèviament la variància de les pertorbacions.

Les pertorbacions són variables aleatòries no observables. Això en dificulta l'estimació de la variància, però els residus mínim quadràtics constitueixen en les h.e.b. aproximacions adequades de les pertorbacions, i per tant, a partir dels residus podrem obtenir un estimador centrat de la variància de les pertorbacions. Per a això necessitem conèixer la distribució dels residus.

Estimació de la variància del terme de pertorbació

- Problema: σ^2 és desconeguda, però podem tractar d'estimar-la.

- No podem estimar Var(u), perquè les pertorbacions no són observables.

El que sí que observem són els residus: $\hat{u}_i$. Podem utilitzar els residus per a estimar la variància del terme de pertorbació σ^2? Sí.

- Si recordem que $\hat{u}_i = y_i - \hat{y}_i = (\beta_0 + \beta_1 x_i + u_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$, la relació entre la pertorbació i els residus és donada per: $\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - \sum_{j=1}^{k} (\hat{\beta}_1 - \beta_1) x_{ij}$.
- És evident que $\hat{u}_i$ és distint de u_i, tot i que $E[\hat{u}_i - u_i] = 0$

Propietats probabilístiques (III): Distribució dels residus

I) Distribució dels residus MQO: $\hat{u}_i \rightarrow N(0, \sigma^2_{\hat{u}_i})$ (***)

Els residus es poden expressar com una funció de la pertorbació aleatòria:

$$\hat{u}_i = Y_i - \hat{Y} = (\beta_1 - \hat{\beta}_1) + (\beta_2 - \hat{\beta}_2)X_{2i} + \cdots + (\beta_k - \hat{\beta}_k)X_{ki} + u_i$$

Per la qual cosa $\hat{u}_i$ es distribuirà com una normal per ser una combinació lineal de u_i, que es distribueix com una normal.

$$E(\hat{u}_i) = 0$$

$$\text{var}(\hat{u}) = E[\hat{u}]^2 \neq \sigma^2$$

Distribució dels residus

O siga: $\boxed{\hat{u}_i \rightarrow N(0, \sigma^2_{\hat{u}_i})}$

Les pertorbacions són homocedàstiques i no autocorrelacionades per hipòtesi, però **els residus són heterocedàstics i autocorrelacionats.**

Recordeu que volem obtenir un estimador centrat de σ^2 . Per obtenir-lo, partim de $SQR = \sum_{i=1}^{N} \hat{u}_i^2$.

Cal tenir en compte que tenim N residus, però només N-k són linealment independents. L'estimador centrat de σ^2 serà:

$$\hat{\sigma}_i^2 = \frac{\sum_{i=1}^{N} \hat{u}_i^2}{N-k}$$

(N-k) es coneix amb el nom de graus de llibertat i és igual al nombre total d'observacions en la mostra menys el nombre de restriccions lineals imposades per les equacions normals (coincideix amb el nombre de paràmetres β a estimar).

Tornem a la variància dels estimadors MQO

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{SQT_j\left(1 - R_j^2\right)} \qquad \text{on} \qquad SQT_j = \sum\left(X_{ij} - \overline{X}_j\right)^2$$

R_j^2 és el R^2 de la regressió de X_j enfront de les restants X

Estimador de la variància dels estimadors MQO

$$\hat{Var}\left(\hat{\beta}_j\right) = \frac{\hat{\sigma}^2}{SQT_j\left(1 - R_j^2\right)}$$

Tornem a la variància dels estimadors MQO

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SCT_j(1 - R_j^2)} \qquad \text{on} \qquad SCT_j = \sum (x_{ij} - \bar{x}_j)^2$$

R_j^2 és el R^2 de la regressió de X_j enfront de les restants X

- La grandària de la variància és important. A major variància tenim un estimador menys precís i per tant estimacions menys fiables.

- La grandària de la variància de MQO depèn de:

  - La variància de l'error: major σ^2 → major variància de l'estimador

  - La variació mostral: major SQT_j → menor variància

  - Relacions lineals entre regressors: major R_j^2 → major variància

Més sobre la variància dels estimadors MQO

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SQT_j(1-R_j^2)} = \frac{\sigma^2}{N * Var(x_j)\,(1-R_j^2)}$$

- Com més soroll hi haja (σ^2) més difícil serà estimar els paràmetres. Més difícil d'estimar-la de forma precisa.

- Si augmenta la grandària mostral disminueix la variància de l'estimador.

- A igualtat dels altres factors, és preferible una major variabilitat en la mostra. Si augmenta la variància del regressor, augmentarà la SQT_j. Per tant, si $\uparrow N \Rightarrow \downarrow Var(\hat{\beta}_j)$

- R_j^2 és la proporció de la variació total en X_j que és explicada per la resta de regressors. A igualtat dels altres factors, la variància de l'estimador MQO serà menor com menor siga la relació entre els regressors. (En veurem més en el tema sobre multicol·linealitat.)

- Si $R_j^2 = 1$, estem en un cas de multicol·linealitat perfecta entre els regressors.

Estimació de la variància del terme de pertorbació

- Problema: σ^2 és desconeguda, però podem tractar d'estimar-la.

- No podem estimar Var(u) utilitzant els errors, perquè aquests no són observables.

- El que sí que observem són els residus: $\hat{u}_i$. Podem utilitzar els residus per a estimar la variància del terme de pertorbació σ^2?

- Si recordem que $\hat{u}_i = y_i - \hat{y}_i = (\beta_0 + \beta_1 x_i + u_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$, la relació entre la pertorbació i els residus és donada per: $\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - \sum_{j=1}^{k} (\hat{\beta}_1 - \beta_1) x_{ij}$.

- És evident que $\hat{u}_i$ és distint de u_i, tot i que $E[\hat{u}_i - u_i] = 0$

Ús de $\hat{\sigma}^2$ per a estimar la desviació típica dels estimadors MQO

- Recordeu que $\hat{\sigma}^2 = \frac{\sum_{n=1}^{n} \hat{u}_i^2}{n-k}$

- Si introduïm $\hat{\sigma}^2$ en $Var(\hat{\beta}_j) = \frac{\sigma^2}{SQT_j\,(1-R_j^2)}$, tindrem: $\hat{V}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{SCT_j\,(1-R_j^2)}$

- L'estimador de l'error estàndard de $\hat{\beta}_1$ serà: $s\hat{d}(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SQT_j\,(1-R_j^2)}}$

- Per a una mostra concreta obtindrem una estimació de $sd(\hat{\beta}_j)$, que ens proporciona una idea de la precisió de l'estimador.

  ➢ Imagineu que l'estimació puntual d'un paràmetre fóra 0,13. Seria la mateixa si la desviació típica fóra 0,02 o 0,13.

  a) L'interval d'estimació seria aprox. [0,09; 0,17]
  b) L'interval d'estimació seria aprox. [-0,13; 0,39]

Tema 4. Contrast d'hipòtesis en el model de regressió múltiple

4.1 Introducció al contrast d'hipòtesis

4.2 Contrast d'hipòtesis sobre un únic paràmetre: l'estadístic *t*

4.3 Contrastos d'hipòtesis sobre un conjunt de paràmetres: l'estadístic *F*

4.4 Contrastos d'hipòtesis mitjançant sumes de quadrats de residus

4.5 Contrastos d'estabilitat estructural

4.6 Predicció

TEMA 4. CONTRAST D'HIPÒTESIS EN EL MODEL DE REGRESSIÓ MÚLTIPLE

4.1 Introducció al contrast d'hipòtesis

4.2 Contrast d'hipòtesis sobre un únic paràmetre: l'estadístic t

4.3 Contrastos d'hipòtesis sobre un conjunt de paràmetres: l'estadístic F

4.4 Contrastos d'hipòtesis mitjançant sumes de quadrats de residus

4.5 Contrastos d'estabilitat estructural

4.6 Predicció

Inferència sobre β

En els temes anteriors hem vist com estimar per MQO (T.2). També hem vist que els estimadors MQO si es compleixen els h.e.b. són ELIO (T.3). També coneixem quina és la distribució dels estimadors MQO. Ara ja podem fer inferència sobre β.

Archivo Herramientas Datos Ver Añadir Mue

dades_salari_1.gdt

ID #	Nombre de variable	Etiqueta descriptiva
0	const	
1	salari	salari per hora en euros
2	educacio	anys d'educacio
3	experiencia	anys d'experiència
4	antiguitat	anys d'antiguitat
5	negre	(1=negre)
6	dona	(1=dona)
7	casada	(1=casada)

gretl: modelo 2

Archivo Editar Contrastes Guardar Gráficos Análisis LaTeX

Modelo 2: MCO, usando las observaciones 1-526
Variable dependiente: salari

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	−1.60221	0.731067	−2.192	0.0289	**
educacio	0.555099	0.0500552	11.09	8.57e−026	***
experiencia	0.0187467	0.0120402	1.557	0.1201	
antiguitat	0.138829	0.0211572	6.562	1.29e−010	***
dona	−1.74241	0.266818	−6.530	1.57e−010	***
negre	−0.0658096	0.426567	−0.1543	0.8775	
casada	0.556565	0.286742	1.941	0.0528	*

Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086
Suma de cuad. residuos	4523.823	D.T. de la regresión	2.952359
R-cuadrado	0.368218	R-cuadrado corregido	0.360914
F(6, 519)	50.41424	Valor p (de F)	8.27e-49
Log-verosimilitud	−1312.288	Criterio de Akaike	2638.576
Criterio de Schwarz	2668.433	Crit. de Hannan-Quinn	2650.267

- L'educació té un efecte positiu en els salaris?
- L'experiència té efecte en els salaris?
- Hi ha discriminació salarial en contra de les dones? Hi ha discriminació per raça?

Totes tres són qüestions sobre el fenomen econòmic analitzat; però, en termes del model economètric, són preguntes sobre els β. Com responem aquestes preguntes?

Contrast d'hipòtesis (rudiments): els vau veure en Introducció a la Inferència Estadística

- Els contrastos sempre són sobre els paràmetres poblacionals. "Es pren una decisió" sobre alguna característica de la població en funció dels valors d'una mostra.
- Les restriccions que es contrasten es recullen en la hipòtesi nul·la (H_0).
- També es defineix una hipòtesi alternativa (H_1), que serà la conclusió si el contrast indica que H_0 és falsa.
- Per contrastar H_0 enfront de H_1, es necessita un estadístic (amb distribució coneguda sota H_0) i una regla de decisió sobre si rebutjar o no rebutjar H_0 .
- Habitualment s'especifica un nivell de significació (α) que indica el nostre marge de tolerància cap a l'error de tipus I (rebutjar H_0 quan en realitat és certa) i que, juntament amb l'alternativa, definirà la regió de rebuig.
- Si el valor de l'estadístic en la mostra pren un valor que pertany a la regió crítica, es rebutjarà H_0 per al nivell de significació α .

Contrast d'hipòtesis (repàs)

Per respondre les preguntes anteriors, hem de realitzar contrastos d'hipòtesis sobre els β. Per realitzar qualsevol contrast d'hipòtesis, és necessari:

1. Establir clarament la hipòtesi nul·la, que es vol contrastar, i l'alternativa.
2. Disposar un estadístic per contrastar la hipòtesi formulada.
3. Definir una regla de decisió (que ens permetrà **rebutjar** o **no rebutjar** la hipòtesi nul·la).

1) **Formulació de H_0 i de H_1** (H_1 serà la teua conclusió si el contrast t'indica que H_0 és "falsa", açò és, si pots rebutjar H_0). P. ex.: $H_0: \beta_{EDu} = 0$ enfront de $H_1: \beta_{EDU} \neq 0$. En efectuar un contrast, tractem de determinar la credibilitat de H_0, veure si l'evidència empírica és suficient per rebutjar H_0 o si, per contra, no permet rebutjar-la.

2) **Disposar d'un estadístic apropiat per efectuar el contrast.**
Es tracta de trobar un estadístic que puguem calcular i que sapiguem com es distribueix si H_0 és certa (és a dir, sota H_0). Del T.3 sabem que sí que es compleixen les h.e.b.:

$$\hat{\beta}_j \rightarrow N\left(\beta_j,\ \sigma^2_{\hat{\beta}_j}\right)$$

amb

$$\sigma^2_{\hat{\beta}_j} = \frac{\sigma^2}{SCT_j\left(1-R_j^2\right)}$$

Per tant, en el nostre exemple, si realment β_{EDU} fóra zero; es à dir, si la H_0 fóra certa, llavors: $\hat{\beta}_{EDU} \rightarrow N\left(0,\ \sigma^2_{\hat{\beta}_{EDU}}\right)$.

És a dir, sabem que si la H_0 fóra certa, la distribució de l'estimador de l'efecte de l'educació ($\hat{\beta}_{EDU}$) estaria centrada en 0 i amb una variància determinada.

- Podem calcular la variància de l'estimador de β_{EDU} ? Sí. Ho hem vist en T.3 $\hat{Var}\left(\hat{\beta}_j\right) = \frac{\hat{\sigma}^2}{SCT_j\left(1-R_j^2\right)}$

  Encara que el més habitual és que ho calcule per nosaltres un programa economètric (Gretl).

gretl: modelo 2

Archivo Editar Contrastes Guardar Gráficos Análisis LaTeX

Modelo 2: MCO, usando las observaciones 1-526
Variable dependiente: salari

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	-1.60221	0.731067	-2.192	0.0289	**
educacio	0.555099	0.0500552	11.09	8.57e-026	***
experiencia	0.0187467	0.0120402	1.557	0.1201	
antiguitat	0.138829	0.0211572	6.562	1.29e-010	***
dona	-1.74241	0.266818	-6.530	1.57e-010	***
negre	-0.0658096	0.426567	-0.1543	0.8775	
casada	0.556565	0.286742	1.941	0.0528	*

Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086
Suma de cuad. residuos	4523.823	D.T. de la regresión	2.952359
R-cuadrado	0.368218	R-cuadrado corregido	0.360914
F(6, 519)	50.41424	Valor p (de F)	8.27e-49
Log-verosimilitud	-1312.288	Criterio de Akaike	2638.576
Criterio de Schwarz	2668.433	Crit. de Hannan-Quinn	2650.267

Recapitulant: si una H_0 (per exemple, $H_0 : \beta_j = \beta_j^*$) fóra certa, llavors:

$$\hat{\beta}_j \rightarrow N(\beta_j^*, \sigma^2_{\hat{\beta}_j})$$

Per tant,

$$\frac{\hat{\beta}_j - \beta_j^*}{\sigma_{\hat{\beta}_j}} \rightarrow N(0,1)$$

El problema que tenim és que no podem calcular $\sigma^2_{\hat{\beta}_j} = \frac{\sigma^2}{SCT_j(1-R_j^2)}$, perquè σ^2 és desconegut

Però el que sí que podem fer és substituir σ^2 per un estimador $\hat{\sigma}^2 = \frac{SCR}{(N-k)}$

PERÒ, en reemplaçar σ^2 pel seu estimador, tindrem (**)

$$\frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \rightarrow t_{N-k}$$

que és l'**estadístic *t* o *t*-ràtio**

Regla de decisió (rebutjar o no rebutjar)...

1. Establir clarament la hipòtesi nul·la, que es vol contrastar, i l'alternativa.
2. Construir un estadístic per contrastar la hipòtesi formulada.
3. Definir una **regla de decisió** (que ens permetrà **rebutjar** o **no rebutjar** la hipòtesi nul·la).

- Habitualment, s'especifica un nivell de significació (α) que indica el nostre marge de tolerància cap a l'error de tipus I (rebutjar H_0 quan en realitat és certa) i que, juntament amb l'alternativa, definirà la regió de rebuig.
- Si el valor de l'estadístic en la mostra pren un valor que pertany a la regió crítica, rebutge H_0 per al nivell de significació α.

En la pràctica, es divideix l'espai en dues regions (regió d'acceptació i regió crítica); si el valor de l'estadístic cau a la regió crítica, rebutge la hipòtesi nul·la. (Si el valor de l'estadístic no cau en la regió crítica, no podré rebutjar la H_0.)

Com es fixen els límits de la regió crítica?

Nivell de significació (α). Probabilitat de rebutjar H_0 encara que siga certa ("marge d'error").

Una vegada fixat el nivell de significació (α), la regla de decisió es tradueix en ……. (dibuix)

- Si el valor de l'estadístic cau en la regió crítica, rebutge H_0 al α%. Si el valor mostral de l'estadístic és major que el valor crític (taules), rebutge H_0 al α%.

- Si el valor de l'estadístic NO cau en la regió crítica, no podré rebutjar H_0 al α%. Si el valor mostral de l'estadístic és menor que el valor crític (taules), no podré rebutjar H_0 al α%.

α o nivell de significació: indica la probabilitat de rebutjar H_0 encara que H_0 siga efectivament certa. Generalment, α es fixa en el 0,05 o 0,10 (5% o 10%).

- Si rebutges H_0, això vol dir que hi ha prou evidència mostral en contra de H_0, per això la rebutges.
- Si no pots rebutjar H_0, això no significa que la hipòtesi siga certa, sinó que l'evidència mostral no és suficient per rebutjar-la: "*sembla certa*"

Nivell de significació crític o valor *p* ("enfocament alternatiu")

Enfocament alternatiu: els ordinadors, quan ens ofereixen un contrast, solen calcular el valor de l'estadístic (amb la meua mostra, amb les dades que tenim) i, perquè no hàgem de mirar taules estadístiques, ens ofereixen el nivell de significació crític o valor *p* (*p-value*) associat al contrast.

Si l'ordinador ens ofereix el valor *p*, segons la nostra regla de decisió:

- Rebutjarem H_0 si valor $p < \alpha$
- No rebutjarem H_0 si valor $p > \alpha$

El nivell de significació crític és un indicador del nivell d'admissibilitat de H_0. Com més gran és el valor *p*, "major confiança tenim que H_0 és certa".

- Suposa que s'ha estimat per MQO un MRL i l'estimació per β_2 és 0,5. És $\beta_2 = 0{,}5$?
- Suposa que s'ha estimat per MQO un MRL i l'estimació per β_2 és 0,5. Rebutjaria H_0: $\beta_2 = 0{,}7$
- Suposa que s'ha estimat per MQO un MRL i s'ha rebutjat H_0: $\beta_2 = 0{,}7$. Estàs completament segur que β_2 és diferent de 0,7?
- Suposa que s'ha estimat per MRL un MLB i el *t*-ràtio per contrastar H_0: $\beta_2 = 0{,}7$ és 1,5. Ho rebutgem?
- Suposa que s'ha estimat per MRL un MLB i el valor *p* per al *t*-ràtio és 0,00001. Rebutgem la hipòtesi?

TEMA 4. CONTRAST D'HIPÒTESIS EN EL MODEL DE REGRESSIÓ MÚLTIPLE

4.1 Introducció al contrast d'hipòtesis

4.2 Contrast d'hipòtesis sobre un únic paràmetre: l'estadístic t

4.3 Contrastos d'hipòtesis sobre un conjunt de paràmetres: l'estadístic F

4.4 Contrastos d'hipòtesis mitjançant sumes de quadrats de residus

4.5 Contrastos d'estabilitat estructural

4.6 Predicció

Contrastos d'hipòtesis sobre un únic paràmetre

- Si plantegem H_0: $\beta_j = a$, utilitzarem l'estadístic : $t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - a}{\hat{\sigma}_{\hat{\beta}_j}}$
- P. ex.: H_0: $\beta_j = 0$ (contrast de significativitat individual). En aquest cas: $t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$
- Si "acceptem" H_0, acceptaríem que X_j no té efecte sobre el *regressand*, després d'haver controlat pels efectes de les altres X.
- Utilitzarem l'estadístic t juntament amb la regla de rebuig corresponent per determinar si rebutgem o no rebutgem la hipòtesi nul·la, H_0 .
- A més de la hipòtesi nul·la (H_0) necessitem una alternativa (H_1) i un nivell de significació (α).

L'estadístic t: alternatives d'una cua

- H_1 pot ser d'una o dues cues:
  - $H_1: \beta_j > 0$ i $H_1: \beta_j < 0$ són d'una cua
  - $H_1: \beta_j \neq 0$ és una alternativa de dues cues

- Suposem que el nostre contrast planteja una alternativa a una cua

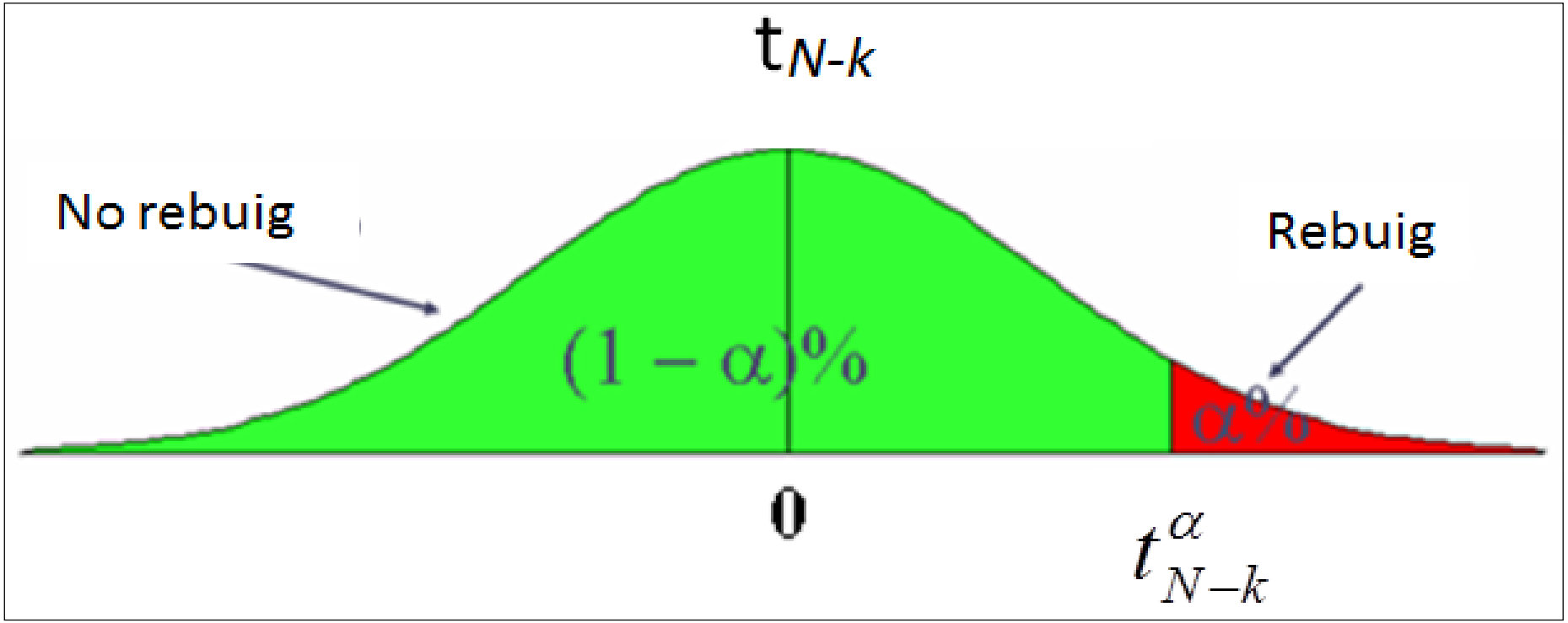
  $H_0: \beta_j = 0$ enfront de $H_1: \beta_j > 0$

- Com queda la regla de decisió? La rebutjarem si observem un valor de l'estadístic $t_{\hat{\beta}_j}$ "suficientment" allunyat de zero per la dreta. Valors negatius de $t_{\hat{\beta}_j}$ no proveeixen evidència a favor de H_1 .

- Cal fixar el nivell de significació (α) o probabilitat de rebutjar H_0 quan en realitat és certa. Habitualment α es fixa en el 5%.

Alternatives d'una cua

- Després de seleccionar un nivell de significació, α , busquem el percentil $(1 - \alpha)$-èsim en les taules de la distribució apropiada (en aquest cas, una *t de student amb (N-k) graus de llibertat*) i el denominem valor crític.
- Rebutgem la H_0 si el valor de l'estadístic *t* és major que el valor crític. Si l'estadístic *t* és menor que el valor crític, no rebutgem la H_0.



Alternatives d'una cua

- Com que la distribució *t* és simètrica, contrastar H_0: $\beta_j = 0$ enfront de H_1: $\beta_j < 0$ és senzill, ja que el valor crític (el de les taules) és l'anterior però canviat de signe.

- Evidentment, rebutjaríem la nul·la si l'estadístic *t* en la mostra pren un valor menor que $-t^{\alpha}_{N-k}$

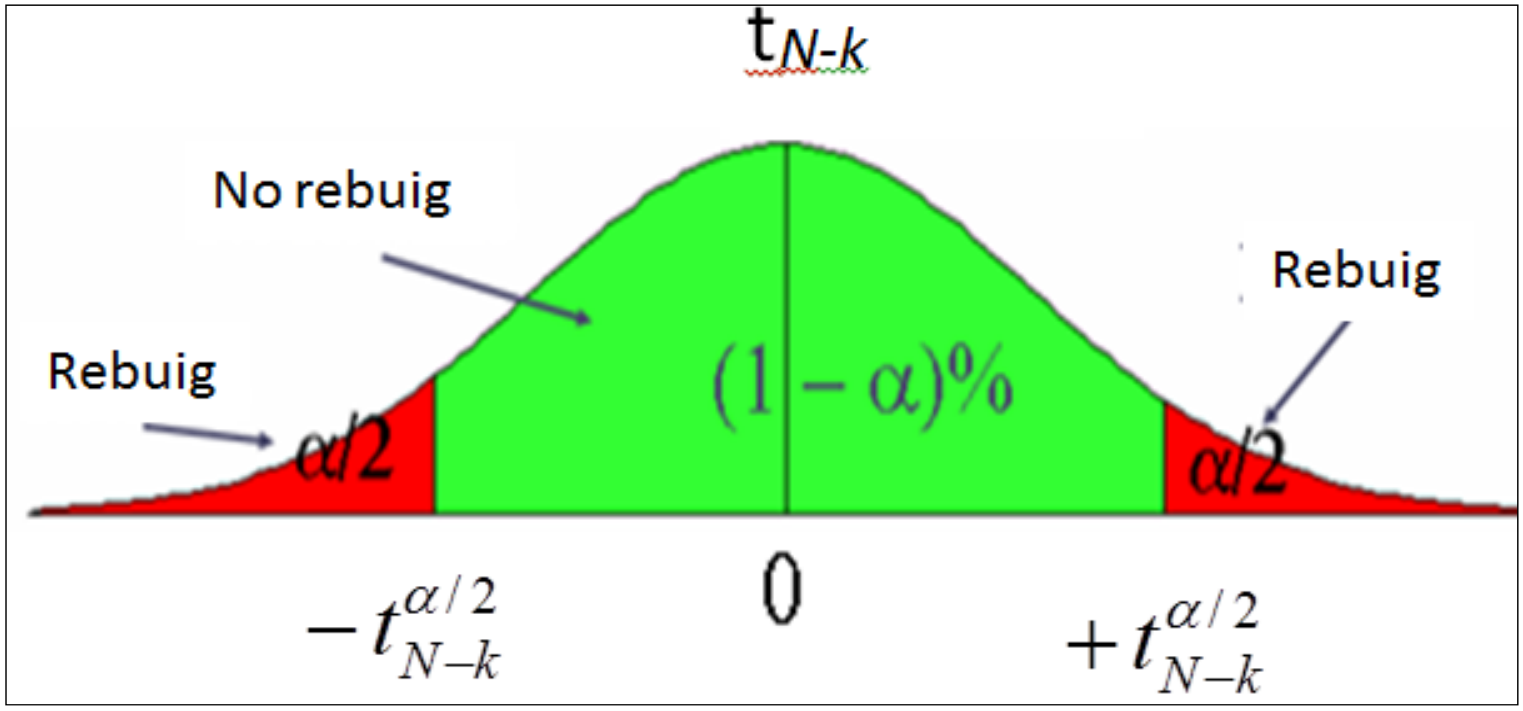
- No es rebutjaria H_0 si el valor de l'estadístic en la meua mostra és major que $-t^{\alpha}_{N-k}$

- Dibuixa tu mateix la zona de rebuig i de no rebuig.

Una cua *vs.* dues cues

- Quan l'alternativa és a una cua, la regió de rebuig es concentra en una cua de la distribució. A més, el signe de l'estadístic *t* és important.

- Si H_1 s'especifica a dues cues (H_1: $\beta_j \neq 0$), encara que el contrast és al α%, el valor crític (el de taules) es basarà en $\alpha/2$. Rebutjarem H_0: $\beta_j = 0$ enfront de H_1: $\beta_j \neq 0$ si el valor de l'estadístic en valor absolut supera el valor crític ($|t| > t^{\alpha}_{N-k}$).



Càlcul de *p*-valors per a contrastos *t*

- Hem vist l'enfocament clàssic de contrastament d'hipòtesis, que, després d'especificar les hipòtesis nul·la i alternativa, es basa a triar un nivell de significació (α) que determina la regió crítica, per comparar després el valor mostral de l'estadístic amb el valor crític (de taules) i concloure que H_0 es rebutja o no es rebutja al α%.

- En algun sentit, l'enfocament clàssic és arbitrari, ja que s'ha de fixar α.

- Una vegada fixat α, H_0 és o no és rebutjada, però no sabem si el rebuig o no rebuig és fort o feble.

- En comptes de fixar α, considerem la qüestió següent: donat el valor de l'estadístic, quin és el menor nivell de significació a què rebutjaríem la nul·la?

- Aquest nivell s'anomena *p*-valor del contrast ("probabilitat de trobar un valor que siga major que l'estadístic estimat").

- Una vegada s'ha calculat el *p*-valor, és senzill realitzar un contrast clàssic, per a qualsevol nivell de significació: es rebutjarà H_0 si *p*-valor $< \alpha$.

Contrastos que fa automàticament Eviews o Gretl: exemple

Modelo 1: MCO, usando las observaciones 1-526
Variable dependiente: salari

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	-1.71453	0.761686	-2.251	0.0248	**
educacio	0.601749	0.0513538	11.72	2.61e-028	***
experiencia	0.0642164	0.0104108	6.168	1.39e-09	***
negre	-0.0838851	0.444298	-0.1888	0.8503	
dona	-2.15649	0.270605	-7.969	1.01e-014	***

Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086
Suma de cuad. residuos	4945.334	D.T. de la regresión	3.080910
R-cuadrado	0.309351	R-cuadrado corregido	0.304048
F(4, 521)	58.34070	Valor p (de F)	1.09e-40
Log-verosimilitud	-1335.718	Criterio de Akaike	2681.436
Criterio de Schwarz	2702.762	Crit. de Hannan-Quinn	2689.786

Dependent Variable: SALARI
Method: Least Squares
Date: 10/11/13 Time: 13:26
Sample: 1 526
Included observations: 526

	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.714526	0.761686	-2.250963	0.0248
EDUCACIO	0.601749	0.051354	11.71771	0.0000
EXPERIENCIA	0.064216	0.010411	6.168238	0.0000
NEGRE	-0.083885	0.444298	-0.188804	0.8503
DONA	-2.156494	0.270605	-7.969158	0.0000

R-squared	0.309351	Mean dependent var	5.896103
Adjusted R-squared	0.304048	S.D. dependent var	3.693086
S.E. of regression	3.080910	Akaike info criterion	5.097787
Sum squared resid	4945.334	Schwarz criterion	5.138332
Log likelihood	-1335.718	Hannan-Quinn criter.	5.113662
F-statistic	58.34070	Durbin-Watson stat	1.813263
Prob(F-statistic)	0.000000		

Significativitat econòmica *vs* significativitat estadística

- La significativitat estadística es determina pel valor del *t*-ràtio mentre que la significativitat econòmica es relaciona amb la magnitud (i signe) de $\hat{\beta}_j$.

- Posar massa èmfasi en la significativitat estadística pot portar a concloure que una variable és "important" per explicar el regressand, fins i tot encara que l'efecte estimat siga molt modest.

- Amb grandàries de mostra gran, els paràmetres se solen estimar de forma precisa: els errors estàndards solen ser petits, cosa que sol resultar significativa estadísticament.

- Encara que una variable siga estadísticament significativa, també cal analitzar el valor estimat del coeficient per donar una idea de la seua importància pràctica o econòmica.

Fins ara només hem considerat hipòtesis amb una sola restricció, però sovint cal plantejar hipòtesis amb diverses restriccions. Vegem-ho...

TEMA 4. CONTRAST D'HIPÒTESIS EN EL MODEL DE REGRESSIÓ MÚLTIPLE

4.1 Introducció al contrast d'hipòtesis

4.2 Contrast d'hipòtesis sobre un únic paràmetre: l'estadístic t

4.3 Contrastos d'hipòtesis sobre un conjunt de paràmetres: l'estadístic F

4.4 Contrastos d'hipòtesis mitjançant sumes de quadrats de residus

4.5 Contrastos d'estabilitat estructural

4.6 Predicció

Restriccions lineals múltiples: restriccions d'exclusió

- Un cas típic (i el més senzill) de restriccions múltiples són les "restriccions d'exclusió".
- Volem saber si un grup de variables independents no tenen efecte parcial en el regressand.
- Per exemple: considerem el model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + u$$ i plantegem

$$H_0: \beta_4 = \beta_5 = \beta_6 = 0$$

Plantegem si el conjunt de paràmetres són conjuntament no significatius, ja que els seus efectes parcials serien nuls.

Restriccions conjuntes d'exclusió: quina forma pren l'alternativa?

- H_1: existeix algun β_i ($i = 4,5,6$) diferent de zero. (No cal que tots siguen diferents de zero.)
- És a dir, H_1 es defineix com la negació de la nul·la (H_1: H_0 no és certa); açò és, el contrast es construeix de manera que detecte qualsevol allunyament de la nul·la.
- Pot ser temptador contrastar l'anterior H_0 mitjançant una successió de contrastos individuals amb l'estadístic t, però aquesta opció no és apropiada, necessitem un mitjà **de contrastar les restriccions conjuntament**. Hi ha la possibilitat que cap de les tres variables no siguen individualment significatives però sí que ho siguen conjuntament.

Restriccions conjuntes d'exclusió: quin estadístic usem?

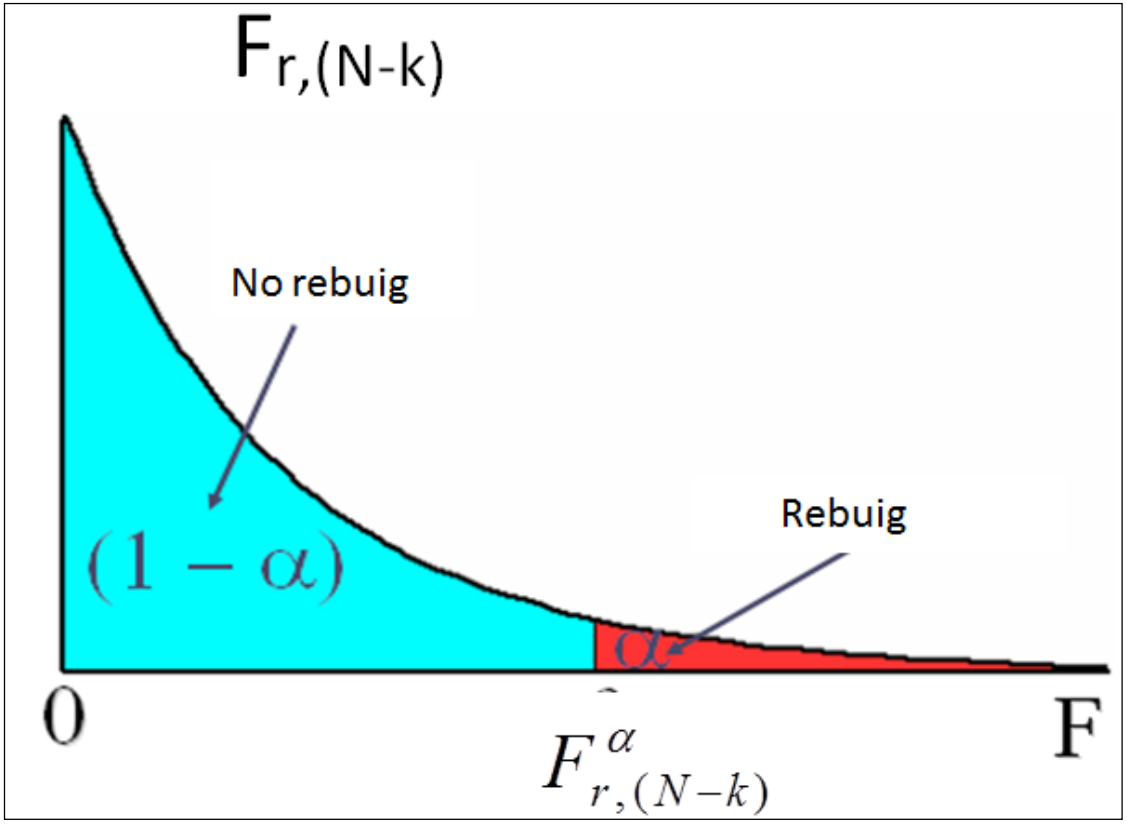
- Quin estadístic usem? **L'estadístic *F***, del qual ni tan sols us mostraré l'expressió, ja el calcularà Gretl per nosaltres.

- L'estadístic *F*, sota H_0 (i suposant que es compleixen les h.e.b.), es distribueix:

$$F \sim F_{r,(N-k)}$$

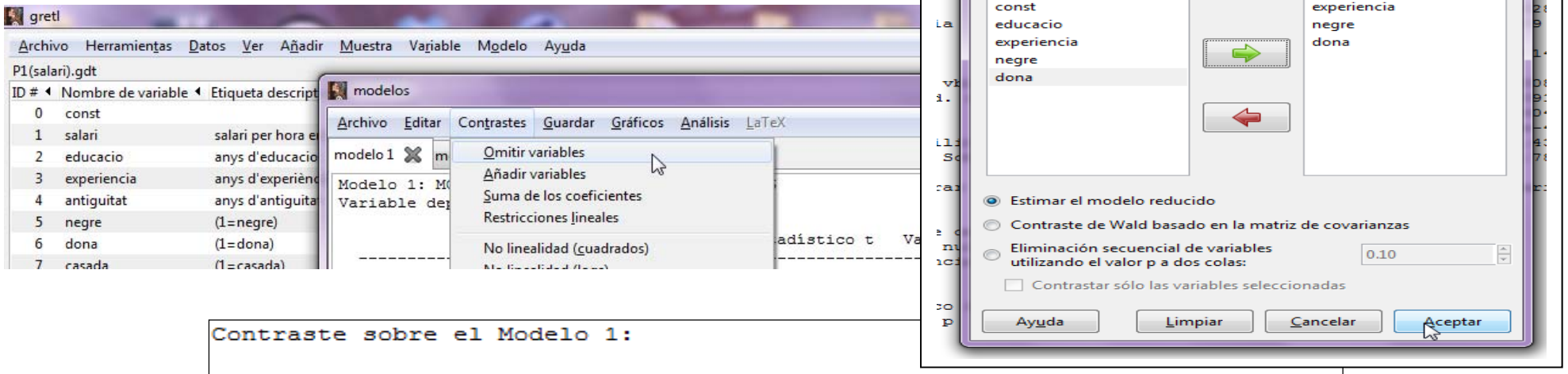
L'estadístic *F*

- De manera que (per a un α donat) es rebutjarà H_0 si el valor mostral de l'estadístic *F* excedeix el valor crític de les taules (el α centil corresponent de la distribució $F_{r,(N-k)}$).



- Si es rebutja la H_0, es dirà que les variables són estadísticament significatives de forma conjunta al nivell de significativitat α.

Restriccions d'exclusió amb Gretl




```
Contraste sobre el Modelo 1:

  Hipótesis nula: los parámetros de regresión son cero para las variables
    experiencia, negre, dona
  Estadístico de contraste: F(3, 521) = 36.3586, Valor p 2.39033e-021
  Al omitir variables mejoraron 0 de los 3 estadísticos de selección de
modelos considerados.

Modelo 2: MCO, usando las observaciones 1-526
Variable dependiente: salari

             Coeficiente   Desv. Típica   Estadístico t    Valor p
  ------------------------------------------------------------------
  const       -0.904852     0.684968         -1.321        0.1871
  educacio     0.541359     0.0532480        10.17         2.78e-022 ***

Media de la vble. dep.   5.896103   D.T. de la vble. dep.    3.693086
Suma de cuad. residuos   5980.682   D.T. de la regresión     3.378390
R-cuadrado               0.164758   R-cuadrado corregido     0.163164
F(1, 524)                103.3627   Valor p (de F)           2.78e-22
Log-verosimilitud       -1385.712   Criterio de Akaike       2775.423
Criterio de Schwarz      2783.954   Crit. de Hannan-Quinn    2778.764
```

Relació entre els estadístics *t* i *F*

- Acabem de veure com usar l'estadístic F per contrastar si un grup de variables explicatives són conjuntament significatives. Però, què passa si utilitzem l'estadístic *F* per contrastar la significativitat estadística d'una sola variable explicativa?

- És a dir, utilitzar l'estadístic *F* per al cas d'una única restricció ($r = 1$). Per exemple, H_0: $\beta_k = 0$. Ja sabem que en aquest cas podem utilitzar el *t*-ràtio.

- Llavors, hi ha dues formes de contrastar la mateixa hipòtesi? La resposta és **NO**.

- Es pot demostrar que, quan es contrasta H_0: $\beta_k = 0$, l'estadístic *F* és exactament el quadrat del corresponent *t* estadístic. Per tant, tots dos portarien al mateix resultat (sempre que l'alternativa siga a dues cues).

- Com que l'estadístic *t* és més flexible (es pot utilitzar per contrastar alternatives d'una i de dues cues) i és més fàcil de calcular, no hi ha cap raó per usar l'estadístic *F* quan es vol contrastar hipòtesis amb una única restricció.

Conclusió: per contrastar una única restricció, és millor utilitzar el *t*-ràtio

Contrast de significativitat global

- Un cas particular de restriccions d'exclusió és: H_0: $\beta_2 = \beta_3 = \ldots = \beta_k = 0$.
- És a dir, cap de les variables explicatives no afecta el regressand.
- L'alternativa consisteix que almenys un dels regressors és significatiu:

  H_1: $\beta_j \neq 0$ (per a algun $j = 2, \ldots, k$) no esta inclòs el terme independent

- Per a aquest cas particular, l'estadístic F se simplifica molt i queda:

$$F = \frac{R^2 / k}{\left(1 - R^2\right) / \left(N - k\right)}$$

i es distribueix com una $\boldsymbol{F_{r,(N-k)}}$

En aquest cas, el nombre de restriccions **r =(k-1);** és a dir, tots el paràmetres excepte la constant (β_1)

Contrast de significativitat global (*output* de regressió)

- El contrast de significativitat global sol ser ofert automàticament per tots els paquets economètrics quan s'efectua una regressió per MQO.

```
Modelo 1: MCO, usando las observaciones 1-526
Variable dependiente: salari

             Coeficiente  Desv. Típica  Estadístico t  Valor p
  ------------------------------------------------------------
  const       -1.71453     0.761686       -2.251      0.0248    **
  educacio     0.601749    0.0513538       11.72       2.61e-028 ***
  experiencia  0.0642164   0.0104108        6.168      1.39e-09  ***
  negre       -0.0838851   0.444298        -0.1888     0.8503
  dona        -2.15649     0.270605       -7.969      1.01e-014 ***

Media de la vble. dep.  5.896103   D.T. de la vble. dep.   3.693086
Suma de cuad. residuos  4945.334   D.T. de la regresión    3.080910
R-cuadrado              0.309351   R-cuadrado corregido    0.304048
F(4, 521)               58.34070   Valor p (de F)          1.09e-40
Log-verosimilitud      -1335.718   Criterio de Akaike      2681.436
Criterio de Schwarz     2702.762   Crit. de Hannan-Quinn   2689.786
```

Dependent Variable: SALARI
Method: Least Squares
Date: 10/11/13 Time: 13:26
Sample: 1 526
Included observations: 526

	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.714526	0.761686	-2.250963	0.0248
EDUCACIO	0.601749	0.051354	11.71771	0.0000
EXPERIENCIA	0.064216	0.010411	6.168238	0.0000
NEGRE	-0.083885	0.444298	-0.188804	0.8503
DONA	-2.156494	0.270605	-7.969158	0.0000

R-squared	0.309351	Mean dependent var	5.896103
Adjusted R-squared	0.304048	S.D. dependent var	3.693086
S.E. of regression	3.080910	Akaike info criterion	5.097787
Sum squared resid	4945.334	Schwarz criterion	5.138332
Log likelihood	-1335.718	Hannan-Quinn criter.	5.113662
F-statistic	58.34070	Durbin-Watson stat	1.813263
Prob(F-statistic)	0.000000		

TEMA 4. CONTRAST D'HIPÒTESIS EN EL MODEL DE REGRESSIÓ MÚLTIPLE

4.1 Introducció al contrast d'hipòtesis

4.2 Contrast d'hipòtesis sobre un únic paràmetre: l'estadístic t

4.3 Contrastos d'hipòtesis sobre un conjunt de paràmetres: l'estadístic F

4.4 Contrastos d'hipòtesis mitjançant sumes de quadrats de residus

4.5 Contrastos d'estabilitat estructural

4.6 Predicció

Contrastar restriccions conjuntes (d'exclusió) mitjançant *SCR*

- Una forma senzilla de contrastar restriccions d'exclusió és a través de les *SCR*.
- Necessitarem estimar el model general (sense les restriccions incloses) i el model restringit (model que incorpora les restriccions H_0: $\beta_3 = \beta_4 = \beta_5 = 0$).

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + u \quad \text{(general)}$$

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u \quad \text{(restringit)}$$

- És evident que sempre ocorrerà que $SCR_R \geq SCR_G$. Per què?
- Es tracta d'avaluar si el canvi en *SCR* és suficientment gran per no excloure les variables $x_3\ x_4\ x_5$ del model.
  - Si, en incloure-hi les restriccions (model restringit), es produeix un gran increment en *SCR*, això és evidència en contra de H_0.
  - Si SCR_R és proper a SCR_G, això indica que les restriccions són aproximadament certes.

Comparant els *SCR* dels models general i restringit

- Per tant, fins i tot encara que *SCR* per si mateix no ens diu res sobre la validesa de H_0, l'increment de *SCR* quan s'imposen les restriccions ens pot ajudar a decidir sobre la validesa de H_0 .

- Cal valorar si l'increment observat en *SCR* (en imposar les restriccions) és suficientment gran amb relació a la *SCR* del model general per decidir no rebutjar H_0.

- Igual que en tot contrast d'hipòtesis, la resposta depèn del nivell de significació (α).

- Però també necessitarem un estadístic amb distribució coneguda sota la nul·la (si la nul·la fóra certa).

L'estadístic *F* (en termes de *SCR*)

- L'estadístic *F* pren la forma $$F \equiv \frac{\left(SCR_R - SGR_G\right)/r}{SCR_G/\left(N-k\right)}$$

  SCR_G: suma de quadrats residual del model general
  SCR_R: suma de quadrats residual del model restringit
  $(N-k)$: graus de llibertat del model general
  r: nombre de restriccions

- Fixa't que el denominador de l'estadístic *F* és precisament l'estimador de σ^2 en el model general.

- Sota H_0 (i suposant que es compleixen les h.e.b.), l'estadístic *F* es distribueix:

$$\boldsymbol{F \sim F_{r,(N-k)}}$$

L'estadístic F en termes de R^2

- Hi ha una formulació de l'estadístic F molt fàcil de calcular i, per tant, és molt convenient conèixer-la.

- Aprofitant que $SCR = SCT(1 - R^2)$, es pot reformular l'estadístic F d'aquesta manera:

$$F = \frac{\left(R_G^2 - R_R^2\right)/r}{\left(1 - R_G^2\right)/\left(N - k\right)}$$

- Aquest estadístic és molt convenient per contrastar restriccions d'exclusió, però no pot ser aplicat a tots els tipus de restriccions lineals.

- Feu el contrast de significativitat conjunta del model vosaltres mateixos de dues formes: a) usant l'estadístic *F* en funció del *SCR*; i b) estadístic *F* en funció de R^2:

Modelo 1: MCO, usando las observaciones 1-526
Variable dependiente: salari

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	-1.71453	0.761686	-2.251	0.0248	**
educacio	0.601749	0.0513538	11.72	2.61e-028	***
experiencia	0.0642164	0.0104108	6.168	1.39e-09	***
negre	-0.0838851	0.444298	-0.1888	0.8503	
dona	-2.15649	0.270605	-7.969	1.01e-014	***

Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086
Suma de cuad. residuos	4945.334	D.T. de la regresión	3.080910
R-cuadrado	0.309351	R-cuadrado corregido	0.304048
F(4, 521)	58.34070	Valor p (de F)	1.09e-40
Log-verosimilitud	-1335.718	Criterio de Akaike	2681.436
Criterio de Schwarz	2702.762	Crit. de Hannan-Quinn	2689.786

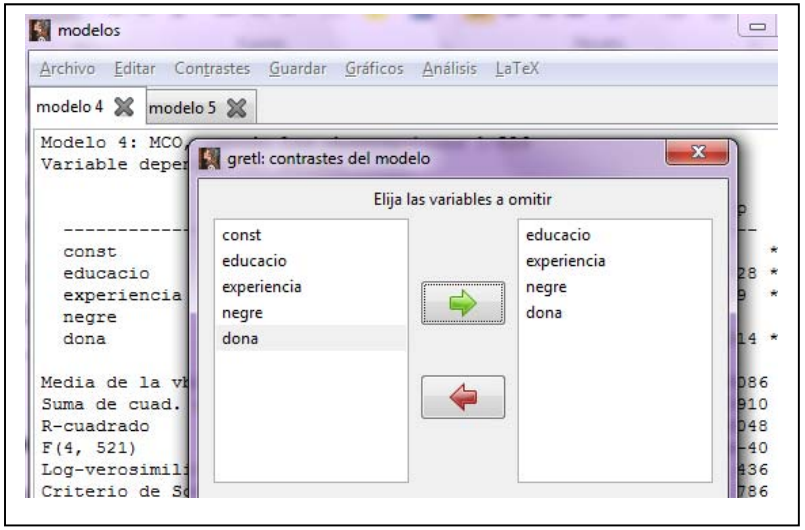
Modelo 1: MCO, usando las observaciones 1-526
Variable dependiente: salari

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	5.89610	0.161026	36.62	1.14e-146	***

Media de la vble. dep.	5.896103	D.T. de la vble. dep.	3.693086
Suma de cuad. residuos	7160.414	D.T. de la regresión	3.693086
R-cuadrado	0.000000	R-cuadrado corregido	0.000000
Log-verosimilitud	-1433.060	Criterio de Akaike	2868.121
Criterio de Schwarz	2872.386	Crit. de Hannan-Quinn	2869.791

- També ho podem demanar a Gretl:

```
Contraste sobre el Modelo 4:

  Hipótesis nula: los parámetros de regresión son cero para las variables
    educacio, experiencia, negre, dona
  Estadístico de contraste: F(4, 521) = 58.3407, Valor p 1.09191e-040
  Al omitir variables mejoraron 0 de los 3 estadísticos de selección de
modelos considerados.

Modelo 5: MCO, usando las observaciones 1-526
Variable dependiente: salari

             Coeficiente   Desv. Típica   Estadístico t    Valor p
  ------------------------------------------------------------------
  const        5.89610       0.161026         36.62       1.14e-146 ***

Media de la vble. dep.  5.896103   D.T. de la vble. dep.   3.693086
Suma de cuad. residuos  7160.414   D.T. de la regresión    3.693086
R-cuadrado              0.000000   R-cuadrado corregido    0.000000
Log-verosimilitud      -1433.060   Criterio de Akaike      2868.121
Criterio de Schwarz     2872.386   Crit. de Hannan-Quinn   2869.791
```



Contrast de restriccions lineals generals

- Moltes vegades és interessant contrastar restriccions múltiples, no totes elles d'exclusió.
- La forma bàsica de l'estadístic F en termes de SCR és possible usar-la per a qualsevol conjunt de restriccions lineals.
- Caldrà estimar el model general i el model restringit, i calcular-ne els SCR.
- Definir el model restringit és el més complicat.
- Recordar que l'estadístic F mesura l'augment relatiu en SCR en moure'ns del model general al model restringit.

Contrast de restriccions lineals generals: exemple

- Per exemple: $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + u$ i plantegem:

  H_0: $\beta_2 = \beta_3$, $\beta_4 = 1$, $\beta_5 = 0$, $\beta_6 = 0$ enfront de **H_1**: H_0 no és certa.

- El model restringit és $y = \beta_1 + \beta_2 x_2 + \beta_2 x_3 + x_4 + u$

- Però, per estimar-lo, l'hem d'escriure: $(y - x_4) = \beta_1 + \beta_2 (x_2 + x_3) + u$

- És a dir, hem de tornar la variable $(y - x_4)$ enfront d'una constant i el regressor $(x_2 + x_3)$

- En aquest exemple no es podria utilitzar l'estadístic F en termes de R^2, ja que la variable dependent és diferent en els models (general i restringit). Això fa que siga distinta la *SCT* de tots dos models.

TEMA 4. CONTRAST D'HIPÒTESIS EN EL MODEL DE REGRESSIÓ MÚLTIPLE

4.1 Introducció al contrast d'hipòtesis

4.2 Contrast d'hipòtesis sobre un únic paràmetre: l'estadístic t

4.3 Contrastos d'hipòtesis sobre un conjunt de paràmetres: l'estadístic F

4.4 Contrastos d'hipòtesis mitjançant sumes de quadrats de residus

4.5 Contrastos d'estabilitat estructural

4.6 Predicció

Contrast d'estabilitat estructural: hi ha diferències entre els diferents grups?

- A vegades, l'objectiu de l'anàlisi consisteix a contrastar si les dues "poblacions" o grups segueixen la mateixa funció de regressió.

- Això porta a contrastar la significativitat conjunta de totes les *dummies*, tant additives com multiplicatives.

- Evidentment, això es pot fer estimant el model amb *dummies* (model general) i el model sense cap *dummy* (model restringit) i construir l'estadístic *F* corresponent.

- PERÒ, com que encara no hem vist què són i per a què serveixen les variables *dummies*, ajornem el contrast d'estabilitat estructural fins que vegem, en el tema 5, les variables *dummies*.

- Hi ha un procediment alternatiu en el qual no s'ha d'estimar el model general. És el test de Chow.

TEMA 4. CONTRAST D'HIPÒTESIS EN EL MODEL DE REGRESSIÓ MÚLTIPLE

4.1 Introducció al contrast d'hipòtesis

4.2 Contrast d'hipòtesis sobre un únic paràmetre: l'estadístic t

4.3 Contrastos d'hipòtesis sobre un conjunt de paràmetres: l'estadístic F

4.4 Contrastos d'hipòtesis mitjançant sumes de quadrats de residus

4.5 Contrastos d'estabilitat estructural

4.6 Predicció

Predicció

- En aquest tema 4 hem vist un dels usos que podem donar al nostre model de regressió: fer preguntes o plantejar hipòtesis sobre el fenomen econòmic que estem analitzant.
- Aquestes preguntes, les fem mitjançant els contrastos d'hipòtesis que hem descrit en aquest tema.
- Un altre dels usos que podem donar al nostre model de regressió estimat és la predicció; és a dir, tractar de predir el valor "futur" del regressand.

- En Economia, i també al món de l'empresa, interessa moltes vegades disposar de tècniques per predir el valor futur d'una variable (I). Hi ha múltiples mètodes de predicció (qualitatius i quantitatius).
- Fins ara ens hem centrat a obtenir i contrastar hipòtesis sobre els paràmetres de l'MRL. En aquest tema utilitzarem el nostre MRL per anticipar el comportament futur de la variable endògena, és a dir PREDIR Y_i
- Cal tenir en compte que la predicció puntual és una "simple" extrapolació mentre que la predicció per intervals també recull una mesura de la precisió de la predicció (mesura de fiabilitat de la predicció).
- D'altra banda també, a posteriori, interessarà avaluar quina ha estat la capacitat predictiva del model. És el model adequat per predir?

Predicció puntual

Suposem que hem plantejat el següent model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots \beta_k X_{ki} + u_i = \theta_i + u_i$$

i en estimar obtenim:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots \hat{\beta}_k X_{ki}$$

Si els valors de les variables explicatives són conegudes per a l'observació $i = 0$, llavors podem utilitzar el model estimat per obtenir una predicció d'Y per a $i = 0$

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_{20} + \hat{\beta}_3 X_{30} + \cdots \hat{\beta}_k X_{k0} = \hat{\theta}_0$$

El predictor o predicció puntual $\hat{Y}_0$ és, si es compleixen les h.e.b., un estimador lineal, sense biaix i òptim de Y_0.

Predicció per intervals del valor individual

En comparar la predicció de Y_0 amb el vertader Y_0 vegem que hi ha tres possibles fonts d'error:

$$Y_0 = \theta_0 + u_0 \qquad \hat{Y}_0 = \hat{\theta}_0$$

- Hi ha tres possibles fonts d'error en utilitzar la predicció puntual per predir Y_0 en l'MRL:

  1) S'ha utilitzat el model estimat, no el model teòric; és a dir, no s'han utilitzat els vertaders valors dels paràmetres.
  2) Pertorbació aleatòria. En el valor d'Y_0 hi ha una part aleatòria.
  3) Els valors de les variables explicatives no es coneixen exactament i aquests valors sota els quals s'ha condicionat la predicció poden ser erronis.

- Per tant, en general, les prediccions no coincidiran amb el vertader valor Y_0

Interval de predicció per al valor individual

Error de predicció: $$Y_0 - \hat{Y}_0 = \theta_0 + u_0 - \hat{\theta}_0$$

Sota les h.e.b. $$E\left[Y_0 - \hat{Y}_0\right] = E\left[\theta_0 + u_0 - \hat{\theta}_0\right] = \theta_0 - \theta_0 = 0$$

Per tant, sota les h.e.b., l'error de predicció tindrà la següent distribució:

$$Y_0 - \hat{Y}_0 \rightarrow N\left(0, \sigma^2_{Y_0 - \hat{Y}_0}\right)$$

Per la qual cosa, $$\frac{\left(Y_0 - \hat{Y}_0\right)}{\sqrt{\hat{\sigma}^2_{Y_0 - \hat{Y}_0}}} \rightarrow t_{n-k}$$

Per a un nivell de confiança (1-α) s'haurà de: $$\Pr ob[-t^{\alpha/2}_{n-k} \leq \frac{\left(Y_0 - \hat{Y}_0\right)}{\hat{\sigma}_{Y_0 - \hat{Y}_0}} \leq t^{\alpha/2}_{n-k}] = 1 - \alpha$$

Per tant, un interval de probabilitat per al valor individual queda definit per:

$$\left[\hat{Y}_0 - t^{\alpha/2}_{n-k}\hat{\sigma}_{Y_0 - \hat{Y}_0} ; \hat{Y}_0 + t^{\alpha/2}_{n-k}\hat{\sigma}_{Y_0 - \hat{Y}_0}\right]$$

Exemple

(Examen 29/01/2008)

1. Per a una mostra de 61 individus es disposa de les següents variables:

LSALARIO= Logaritme del salari
EDUCAC= Anys d'educació
EXPLAB= Anys d'experiència laboral
SEXE= Variable dicotòmica que pren valor 1 si l'individu és home i 0 si no ho és
SEXOEXP = SEXE*EXPLAB
SEXOEDU= SEXE*EDUCAC

Es pretén analitzar els diferents factors que determinen el salari dels individus. Utilitzant els resultats del QUADRE 1:

a) Estimeu el sou que cobraria un home amb 30 anys d'edat, 4 anys d'experiència i 10 anys d'educació.
b) Proporcioneu un interval de predicció al 90% per al logaritme del salari (per a la mateixa persona de l'apartat anterior). [Nota: suposeu que la desviació típica de l'error de predicció és 2.]

QUADRE 1

Variable Dep.: LSALARIO
Regresores : 1,EDUCAC,SEXO,EXPLAB
Muestra : 1 - 61 Nº Observaciones : 61

Regresores	Coeficiente	Desv. Típica	Estadís. t	Prob>|t|
1	4.351598	0.091895	47.35	0.0000
EDUCAC	0.074941	0.015102	4.96	0.0000
SEXO	0.272021	0.090038	3.02	0.0038
EXPLAB	0.012982	0.004183	3.10	0.0030

Media Var. Dependiente:	5.0309	Des. Típ. Var. Depen.:	0.4561
Error Típico Regresión:	0.3147	Suma Cuadrados Resid.:	5.6437
R Cuadrado :	0.5479	R Cuadrado Corregido :	0.5241
Logaritmo de Verosim. :	-13.9552	Criterio AIC :	0.5887
Estadístico F(3, 57):	23.0271	Prob > F :	0.0000
Estadís. Durbin-Watson:	2.0716	Est. Autocorrelación :	-0.0358

a) Estimeu el sou que cobraria un home amb 30 anys d'edat, 4 anys d'experiència i 10 anys d'educació.
b) Proporcioneu un interval de predicció al 90% per al logaritme del salari (per a la mateixa persona de l'apartat anterior). [Nota: suposeu que la desviació típica de l'error de predicció és 2.]

Exemple

(Examen 4/7/2006)

2. S'ha estimat el següent model de demanda d'habitatge amb observacions anuals corresponents al període 1970-2004:

$$\ln \hat{V}_t = \underset{(0,15)}{-0,39} + \underset{(0,05)}{0,31} \ln R_t - \underset{(0,02)}{0,67} \ln P_t + \underset{(0,04)}{0,70} \ln V_{t-1};$$

$$R^2 = 0,99; DW = 0,52$$

on V és la despesa en habitatge, R és la renda disponible, P és el preu de l'habitatge. Entre parèntesi apareixen les desviacions típiques. A més es disposa de la següent informació:

$$\ln R_{2005} = 10,2; \quad \ln P_{2005} = 2,7; \quad \ln V_{2004} = 5,8$$

a) Obteniu el valor del predictor puntual de la despesa en habitatge per a l'any 2005. Obteniu l'interval de predicció, sabent que la desviació típica de l'error de predicció és 0,45.

Solució

$$\ln \hat{V}_{05} = 0,39 + 0,31 * 10,2 - 0,67 * 2,7 + 0,7 * 5,8 = 5,023$$

Interval de predicció: $\left(5,023 \pm t_{31}^{0,025} 0,45\right) = \left(5,023 \pm 2,042 * 0,45\right) = \left(4,104 \quad 5,942\right)$

Tema 5. Anàlisi de regressió múltiple amb informació qualitativa

5.1. Les variables fictícies
5.2. Interpretació de coeficients de variables fictícies
5.3. Múltiples categories
5.4. Interaccions de variables fictícies

TEMA 5. ANÀLISI DE REGRESSIÓ MÚLTIPLE AMB INFORMACIÓ QUALITATIVA

5.1 Les variables fictícies

5.2 Interpretació de coeficients de variables fictícies

5.3 Múltiples categories

5.4 Interaccions de variables fictícies

Variables fictícies

- Fins ara les variables que hem analitzat han tingut un significat quantitatiu (salari, educació, etc.).

- Però, en el treball empíric, sovint cal incorporar factors qualitatius en el model de regressió (per exemple: el gènere, la grandària, la situació geogràfica, l'estació de l'any...).

- Com? Mitjançant la creació i la introducció en el model d'una sèrie de variables, anomenades variables fictícies, també dites variables artificials o variables *dummy*.

- Aquestes variables fictícies prenen el valor 1 si l'individu posseeix una determinada característica, i 0 si no la posseeix.

- Per exemple: podem definir la variable *home* com una variable binària que prenga el valor 1 si l'individu és home i zero si és dona. També podem definir la variable *dona* o la variable *casat*, o la variable *gran*, o la variable "*a la costa*"…

Dades: variables quantitatives i qualitatives

gretl: mostrar datos

	salari	educacio	experiencia	dona	casada
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
6	8.75	16	9	0	1
7	11.25	18	15	0	0
8	5.00	12	5	1	0
9	3.60	12	26	1	0
10	18.18	17	22	0	1
11	6.25	16	8	1	0

- *Dona* i *casada* són dues variables fictícies (o *dummies*); permetran incorporar al nostre MLR informació qualitativa (en aquest cas, el gènere i/o l'estat civil).
- La variable *dona* s'ha definit de la manera següent: pren el valor 1 si l'individu és dona i pren el valor 0 si l'individu no és dona (en aquest cas, si és home).
- Com es defineix la variable *casada*?

TEMA 5. ANÀLISI DE REGRESSIÓ MÚLTIPLE AMB INFORMACIÓ QUALITATIVA

5.1 Les variables fictícies

5.2 Interpretació de coeficients de variables fictícies

5.3 Múltiples categories

5.4 Interaccions de variables fictícies

Com incorporar informació qualitativa en el model de regressió?

- Per incorporar informació qualitativa en el model regressió, senzillament introduirem les variables fictícies com si foren una variable més del model.

- Quantes fictícies incorporar i com incorporar-les, dependrà del fenomen econòmic concret que es vulga analitzar.

Exemple: una característica amb dues categories

- **Exemple**: es vol contrastar si hi ha discriminació per gènere en la determinació dels salaris. Una possibilitat és plantejar el model següent:

$$\text{Salari}_i = \beta_1 + \delta_1\, \text{dona}_i + \beta_2\, \text{educ}_i + u_i$$

- Quina és la interpretació de δ_1? δ_1 és la diferència en el salari (de mitjana) entre homes i dones, donat el mateix nivell d'educació (i el mateix *u*). És a dir, si δ_1 és diferent de zero, hi haurà discriminació salarial. Vegem-ho amb detall:

Exemple: una característica amb dues categories (discriminació laboral)

- Hem plantejat: $Salari_i = \beta_1 + \delta_1\, dona_i + \beta_2\, educ_i + u_i$

- En termes d'esperances: si E(u | educ, dona) = 0, llavors:

$$E(salari|\ educ,\ \mathbf{dona = 1}) = \beta_1 + \delta_1 + \beta_2\, educ_i$$

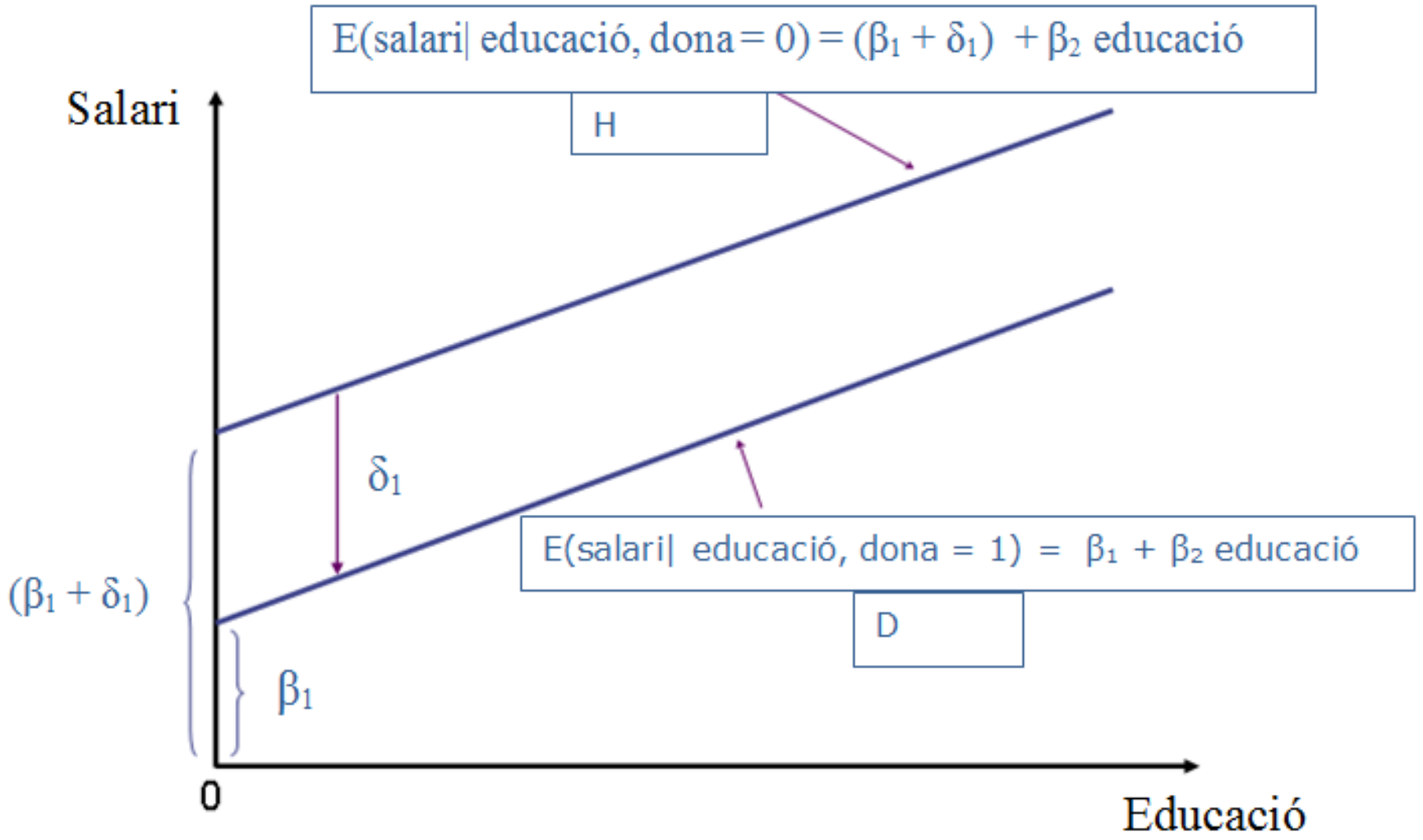
$$E(salari|\ educ,\ \mathbf{dona = 0}) = \beta_1 + \quad + \beta_2\, educ_i$$

$$\boldsymbol{\delta_1} = E(salari|\ educ,\ \mathbf{dona = 1}) - E(salari|\ educ,\ \mathbf{dona = 0})$$

- Gràficament, l'ordenada en l'origen serà diferent per a homes i dones.

Hi haurà discriminació en contra de les dones si $\delta_1<0$. Les dones (per al mateix nivell dels altres factors) guanyaran menys de mitjana.

Exemple amb $\delta_1 < 0$



Contrastos sobre fictícies. Realment hi ha discriminació?

- Introduir fictícies no canvia res en la mecànica d'estimació per MCO ni en la forma d'efectuar els contrastos. L'única diferència respecte als regressors quantitatius és la interpretació del coeficient.

modelo 2

```
Modelo 2: MCO, usando las observaciones 1-526
Variable dependiente: salari

              Coeficiente  Desv. Típica  Estadístico t   Valor p
  ------------------------------------------------------------
  const        -1.73448     0.753620       -2.302      0.0218    **
  educacio      0.602580    0.0511174       11.79       1.33e-028 ***
  experiencia   0.0642417   0.0104003        6.177      1.32e-09  ***
  dona         -2.15552     0.270305        -7.974      9.74e-015 ***

Media de la vble. dep.  5.896103   D.T. de la vble. dep.   3.693086
Suma de cuad. residuos  4945.672   D.T. de la regresión    3.078062
R-cuadrado              0.309304   R-cuadrado corregido    0.305334
F(3, 522)               77.91966   Valor p (de F)          1.15e-41
Log-verosimilitud      -1335.736   Criterio de Akaike      2679.472
Criterio de Schwarz     2696.533   Crit. de Hannan-Quinn   2686.152
```

Quantes fictícies?: parany de les fictícies

- En l'exemple de la discriminació salarial hem introduït la variable fictícia "dona". Per què no hem introduït les dues fictícies, "home" i "dona" al mateix temps?

- Intuïtivament, perquè totes dues variables proporcionen la mateixa informació i, més tècnicament, perquè si introduíem una fictícia per a cada categoria (home/dona), es crea un problema de multicol·linealitat perfecta en el model de regressió, ja que home + dona = 1.

- Per tant, si el model té constant, només es poden introduir en el model tantes fictícies com categories menys una.

- Si s'incorporen al model tantes fictícies com categories, es genera multicol·linealitat perfecta. Aquesta situació és coneguda amb el nom de parany de les variables fictícies.

Quina fictícia cal introduir en el model?: categoria de referència

- Ja sabem que si no volem caure en el parany de les fictícies, cal introduir una *dummy* menys que categories; però, quina fictícia introduïsc en el model, home o dona?

- La categoria que no tindrà *dummy* és elecció de l'investigador, no afecta els resultats, encara que sí la interpretació dels coeficients de les variables fictícies.

- **La categoria que no té *dummy*** s'anomena grup o **categoria de referència**.

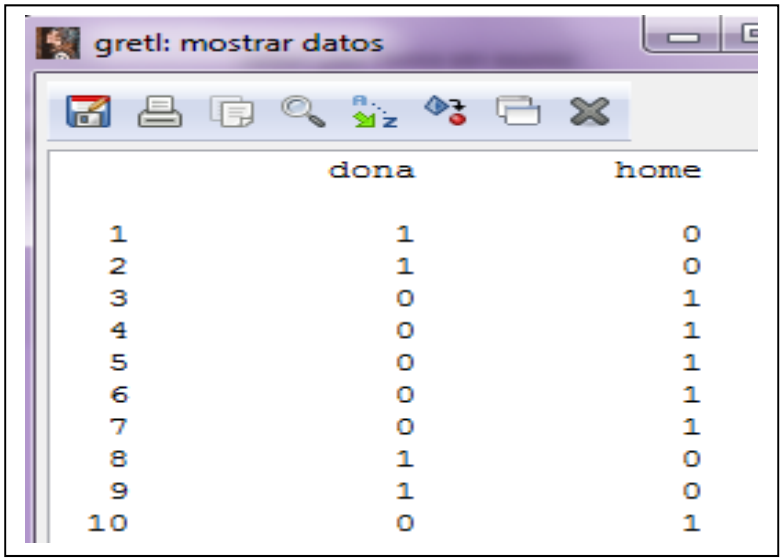
- El coeficient que acompanya una *dummy* indica la diferència en el (valor esperat del) *regressant* entre aquesta categoria i la categoria de referència.

- En el nostre exemple, la variable introduïda és "dona", cosa que fa que la categoria de referència siguen els homes. Per tant, el coeficient que acompanya "dona" indica la diferència en la constant entre les dones i la categoria de referència (homes).

Exemple: canviant la categoria de referència ("dona")

- **Exemple**: suposem que, per contrastar l'existència de discriminació salarial, s'especifica el model següent:

$$Salari_i = \beta_1 + \gamma_1\, home_i + \beta_2\, educ_i + u_i$$

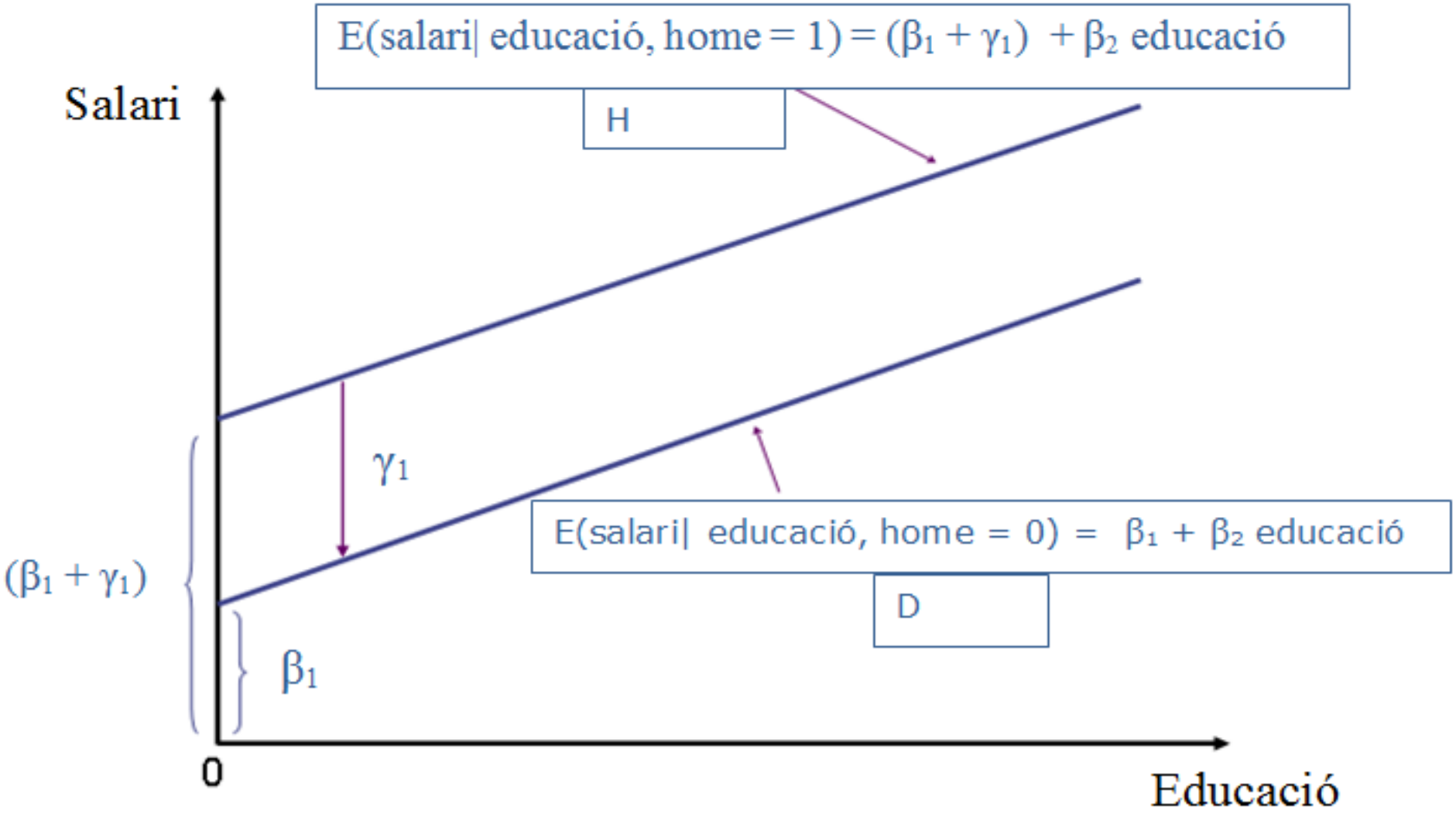


	dona	home
1	1	0
2	1	0
3	0	1
4	0	1
5	0	1
6	0	1
7	0	1
8	1	0
9	1	0
10	0	1

- Ara la categoria de referència és "dona". En aquest cas, β_1 és l'ordenada en l'origen per a les dones (la categoria de referència); mentre que el paràmetre que acompanya la *dummy* (γ_1) és la diferència en el salari entre els homes i la categoria de referència (dones).

- Ara, perquè hi haja discriminació (en contra de les dones), γ_1 ha de ser positiu.

Exemple: hi ha discriminació ($\gamma_1 > 0$) sent "dona" la categoria de referència



Canvi en la categoria de referència

```
Modelo 1: MCO, usando las observaciones 1-526
Variable dependiente: salari

              Coeficiente  Desv. Típica  Estadístico t   Valor p
  ------------------------------------------------------------------
  const        -3.89000      0.727144        -5.350      1.32e-07  ***
  educacio      0.602580     0.0511174       11.79       1.33e-028 ***
  experiencia   0.0642417    0.0104003        6.177      1.32e-09  ***
  home          2.15552      0.270305         7.974      9.74e-015 ***

Media de la vble. dep.  5.896103   D.T. de la vble. dep.   3.693086
Suma de cuad. residuos  4945.672   D.T. de la regresión    3.078062
R-cuadrado              0.309304   R-cuadrado corregido    0.305334
F(3, 522)               77.91966   Valor p (de F)          1.15e-41
Log-verosimilitud      -1335.736   Criterio de Akaike      2679.472
Criterio de Schwarz     2696.533   Crit. de Hannan-Quinn   2686.152
```

Els pendents han de ser iguals entre categories?

- En l'exemple que hem usat hem plantejat un model que permetia diferents interceptes (ordenades) per categories, però res no impedeix que també hi pugui haver diferències en el pendent.

- Per introduir diferències en l'intercepte, hem introduït les ***dummys* en forma additiva** (elles a soles, només acompanyades del seu paràmetre).

- Per introduir diferents pendents, les variables fictícies han d'interactuar amb els altres regressors; és a dir, s'ha d'introduir en el model multiplicant algun regressor (**dummy multiplicativa**). Per exemple:

$$\text{salari} = \beta_1 + \beta_2\,\text{educ} + \delta_2(\text{educ} \times \text{dona}) + u$$

- Si es vol especificar un model que permeta diferències entre grups tant en l'ordenada en l'origen com en el pendent, cal introduir la fictícia tant en forma additiva com multiplicativa. Per exemple:

$$\text{salari} = \beta_1 + \delta_1\,\text{dona} + \beta_2\,\text{educ} + \gamma_2(\text{educ} \times \text{dona}) + u$$

Exemple de fictícies multiplicatives

- Hi ha diferències en la rendibilitat o efecte de l'educació entre dones i homes?

```
Modelo 3: MCO, usando las observaciones 1-526
Variable dependiente: salari
Desviaciones típicas robustas ante heterocedasticidad, variant

             Coeficiente   Desv. Típica   Estadístico t    Val
  ------------------------------------------------------------
  const       0.200496      0.871723         0.2300       0.81
  dona       -1.19852       1.46081         -0.8205       0.41
  educacio    0.539476      0.0733959        7.350        7.69
  edu_dona   -0.0859990     0.123818        -0.6946       0.48

Media de la vble. dep.   5.896103   D.T. de la vble. dep.   3.6
Suma de cuad. residuos   5300.170   D.T. de la regresión    3.1
R-cuadrado               0.259796   R-cuadrado corregido    0.2
F(3, 522)                49.07709   Valor p (de F)          5.9
Log-verosimilitud       -1353.942   Criterio de Akaike      271
Criterio de Schwarz      2732.946   Crit. de Hannan-Quinn   272
```

gretl: mostrar datos

	educacio	dona	edu_dona
1	11	1	11
2	12	1	12
3	11	0	0
4	8	0	0
5	12	0	0
6	16	0	0
7	18	0	0
8	12	1	12
9	12	1	12
10	17	0	0
11	16	1	16
12	13	1	13

TEMA 5. ANÀLISI DE REGRESSIÓ MÚLTIPLE AMB INFORMACIÓ QUALITATIVA

5.1 Les variables fictícies

5.2 Interpretació de coeficients de variables fictícies

5.3 Múltiples categories

5.4 Interaccions de variables fictícies

Una fictícia amb múltiples categories

- **Exemple** (discriminació salarial per raça): En els EUA és habitual dividir la població en tres (o més) grups: raça blanca, negra i hispana. Per incorporar aquesta informació qualitativa en un model de regressió, cal definir les corresponents variables fictícies.
- Cal recordar que només s'ha d'introduir en el model una fictícia menys que categories.
- La categoria que no tinga *la seua "dummy"* en el model serà el grup de referència.
- Les fictícies es poden introduir de forma additiva i/o multiplicativa.
- Vegem un exemple...

Exemple: una fictícia amb múltiples categories (solament additives)

```
Modelo 4: MCO, usando las observaciones 1-526
Variable dependiente: salari
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

             Coeficiente   Desv. Típica   Estadístico t    Valor p
  ------------------------------------------------------------------
  const       -0.175746     0.733805         -0.2395      0.8108
  negre       -0.622091     0.453779         -1.371       0.1710
  hisp        -1.47274      0.293803         -5.013       7.36e-07  ***
  educacio     0.528523     0.0603190         8.762       2.69e-017 ***

Media de la vble. dep.  5.896103   D.T. de la vble. dep.   3.693086
Suma de cuad. residuos  5739.640   D.T. de la regresión    3.315943
R-cuadrado              0.198421   R-cuadrado corregido    0.193814
F(3, 522)               34.08791   Valor p (de F)          3.92e-20
Log-verosimilitud      -1374.892   Criterio de Akaike      2757.785
Criterio de Schwarz     2774.846   Crit. de Hannan-Quinn   2764.465
```

Exemple: una fictícia amb múltiples categories (solament multiplicatives)

```
Modelo 6: MCO, usando las observaciones 1-526
Variable dependiente: salari
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

              Coeficiente  Desv. Típica  Estadístico t   Valor p
  ------------------------------------------------------------------
  const        -0.746556     0.713165        -1.047      0.2957
  educacio      0.458631     0.0629836        7.282      1.22e-012 ***
  edu_x_negre   0.0545374    0.0386720        1.410      0.1591
  edu_x_blanc   0.115018     0.0255442        4.503      8.28e-06  ***

Media de la vble. dep.  5.896103   D.T. de la vble. dep.   3.693086
Suma de cuad. residuos  5735.617   D.T. de la regresión    3.314781
R-cuadrado              0.198983   R-cuadrado corregido    0.194379
F(3, 522)               33.44478   Valor p (de F)          8.73e-20
Log-verosimilitud      -1374.708   Criterio de Akaike      2757.416
Criterio de Schwarz     2774.477   Crit. de Hannan-Quinn   2764.096
```

Exemple: una fictícia amb múltiples categories (additives i multiplicatives)

```
Modelo 7: MCO, usando las observaciones 1-526
Variable dependiente: salari

              Coeficiente  Desv. Típica  Estadístico t   Valor p
  ------------------------------------------------------------------
  const         1.09065      1.69014          0.6453     0.5190
  educacio      0.582259     0.0722663        8.057      5.38e-015 ***
  blanc        -1.95190      1.93497         -1.009      0.3136
  hisp         -2.54481      2.05753         -1.237      0.2167
  edu_x_negre  -0.212829     0.155081        -1.372      0.1705
  edu_x_hisp   -0.0693380    0.117062        -0.5923     0.5539

Media de la vble. dep.  5.896103   D.T. de la vble. dep.   3.693086
Suma de cuad. residuos  5718.460   D.T. de la regresión    3.316178
R-cuadrado              0.201379   R-cuadrado corregido    0.193700
F(5, 520)               26.22441   Valor p (de F)          1.19e-23
Log-verosimilitud      -1373.920   Criterio de Akaike      2759.840
Criterio de Schwarz     2785.432   Crit. de Hannan-Quinn   2769.860
```

TEMA 5. ANÀLISI DE REGRESSIÓ MÚLTIPLE AMB INFORMACIÓ QUALITATIVA

5.1 Les variables fictícies

5.2 Interpretació de coeficients de variables fictícies

5.3 Múltiples categories

5.4 Interaccions de variables fictícies

Múltiples fictícies

- Res no impedeix que el nostre model incorpore diversos tipus d'informació qualitativa o característiques.
- El mecanisme és el mateix que amb una característica: definir les corresponents fictícies i introduir per a cada característica tantes *dummys* com categories menys una. Per a cada atribut tindrem una categoria de referència.
- **Exemple**: discriminació salarial per sexe (home/dona) i estat civil (solter/casat).
- Res no canvia quant a la mecànica, només que alerta amb la multicol·linealitat?
- Quan hi ha múltiples fictícies, sorgeix la possibilitat que les dues característiques interactuen (efecte interacció).

Exemple: múltiples fictícies (sexe i estat civil)

```
Modelo 8: MCO, usando las observaciones 1-526
Variable dependiente: salari
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

             Coeficiente   Desv. Típica   Estadístico t    Valor p
  ------------------------------------------------------------------
  const       -2.12781      0.715292         -2.975        0.0031    ***
  educacio     0.494954     0.0595780         8.308        8.46e-016 ***
  home         2.08699      0.257623          8.101        3.88e-015 ***
  casada       1.18153      0.257484          4.589        5.59e-06  ***

Media de la vble. dep.  5.896103   D.T. de la vble. dep.   3.693086
Suma de cuad. residuos  5137.567   D.T. de la regresión    3.137209
R-cuadrado              0.282504   R-cuadrado corregido    0.278381
F(3, 522)               49.97140   Valor p (de F)          2.10e-28
Log-verosimilitud      -1345.748   Criterio de Akaike      2699.495
Criterio de Schwarz     2716.556   Crit. de Hannan-Quinn   2706.175
```

Exemple: múltiples fictícies i efecte interacció (dona casada?)

```
Modelo 9: MCO, usando las observaciones 1-526
Variable dependiente: salari
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

              Coeficiente  Desv. Típica  Estadístico t   Valor p
  ------------------------------------------------------------------
  const         -1.02442     0.777706       -1.317      0.1883
  educacio       0.493559    0.0583005       8.466      2.60e-016 ***
  dona          -0.368964    0.371298       -0.9937     0.3208
  casada         2.64107     0.401705        6.575      1.19e-010 ***
  dona_casada   -2.82883     0.497760       -5.683      2.20e-08  ***

Media de la vble. dep.  5.896103   D.T. de la vble. dep.   3.693086
Suma de cuad. residuos  4894.020   D.T. de la regresión    3.064884
R-cuadrado              0.316517   R-cuadrado corregido    0.311270
F(4, 521)               39.35195   Valor p (de F)          8.34e-29
Log-verosimilitud      -1332.975   Criterio de Akaike      2675.950
Criterio de Schwarz     2697.276   Crit. de Hannan-Quinn   2684.300
```

Hi ha diferències entre els diversos grups?

- A vegades, l'objectiu de l'anàlisi consisteix a contrastar si les dues poblacions o grups segueixen la mateixa funció de regressió.

- Això implica contrastar la significativitat conjunta de totes les *dummys*, tant additives com multiplicatives.

- Evidentment, això es pot fer estimant el model amb *dummies* (model general) i el model sense cap *dummy* (model restringit) i construir el corresponent estadístic *F*.

Tema 6. Incompliment de les hipòtesis bàsiques

6.1 Multicol·linealitat

6.2 Normalitat

6.3 Heterocedasticitat

6.4 Autocorrelació

- Un resultat molt important que vam obtenir en el tema 3 consisteix que si es compleixen les hipòtesis estadístiques bàsiques (h.e.b.) podem estimar un MRLM per MCO ja que els estimadors MCO, si es compleixen les h.e.b., són ELIO: no tenen biaix i òptims (els de menor variància).

- Per contra, si alguna de les h.e.b. no es compleix, MCO **pot** deixar de ser ELIO.

- Si es compleixen les h.e.b., vam obtenir els següents resultats:

  1) Els estimadors MCO són ELIO.

  2) Els estimadors MCO es distribueixen com:

$$\hat{\beta}_k \rightarrow N(\beta_k, \sigma^2_{\hat{\beta}_k})$$

amb

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SCT_j(1-R_j^2)} = \frac{\sigma^2}{N * Var(x_j)\,(1-R_j^2)}$$

Recordem les hipòtesis estadístiques bàsiques (h.e.b.)

I) Hipòtesi sobre la forma funcional

1) El model és lineal: $$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots \beta_k X_{ki} + u_i \qquad i = 1,2,....n$$

II) Hipòtesi sobre la pertorbació aleatòria

2) u_i són v.a. no observables.

3) $E(u_i) = 0$ $\forall$i = 1, ... n

4) $Var(u_i) = \sigma^2$ $\forall$i = 1, ... n (HOMOCEDASTICITAT) (apartat 6.3)

5) Cov(u_i , u_s) = 0 ; $\forall\ i \neq s$ (NO AUTOCORRELACIÓ) (apartat 6.4)

6) La pertorbació aleatòria segueix una distribució normal multivariant

(NORMALITAT) (apartat 6.2)

Les quatre hipòtesis sobre u es poden expressar conjuntament de la següent forma:

$$u_i \rightarrow N(0\ ,\ \sigma_i^2\ )$$

… Hipòtesis estadístiques bàsiques (h.e.b.)

III) Hipòtesi sobre els regressors

7) Els regressors són no estocàstics, o sigui, *els regressors són fixos*.

7*) Els regressors es distribueixen independentment del terme de pertorbació: $E(X'u) = 0$

8) Els regressors són linealment independents, la qual cosa implica que no existeixen relacions lineals exactes entre els regressors.

(NO COLINEALITAT PERFECTA) (apartat 6.1)

9) Els regressors no tenen errors de mesura.

IV) Hipòtesi sobre β

10) Els paràmetres del model són fixos.

Resum dels resultats que obtindrem

- Si es compleixen les h.e.b. els estimadors MCO són ELIO. Perquè siguin ELIO no és necessària la hipòtesi de normalitat de les pertorbacions. Si les pertorbacions es distribueixen normalment, els estimadors també i llavors podem fer-ne inferència perquè sabem com es distribueixen.

Per contra, si alguna de les h.e.b. no es compleix, MCO poden deixar de ser ELIO.

- (6.1 col·linealitat o multicol·linealitat). Si hi ha una elevada col·linealitat entre els regressors (X's), els estimadors MCO segueixen sent ELIO, però la seua variància és elevada, és a dir són **poc precisos**.

- (6.2 no normalitat). Si les pertorbacions no segueixen una distribució normal, els estimadors continuen sent ELIO, **però ja no seguiran una distribució normal**; per tant, tindrem problemes per fer contrastos amb els estadístics habituals (t, F…).

- (6.3 heterocedasticitat). Si les pertorbacions no són homocedàstiques. Els estimadors MCO continuen essent sense biaix, però **ja no seran òptims**.

- (6.4 autocorrelació). Si les pertorbacions estan correlacionades, els estimadors MCO continuen essent sense biaix, però **ja no seran òptims**.

TEMA 6. INCOMPLIMENT DE LES HIPÒTESIS BÀSIQUES

6.1 Multicol·linealitat

6.2 Normalitat

6.3 Heterocedasticitat

6.4 Autocorrelació

Què és la col·linealitat o multicol·linealitat?

- Hi ha col·linealitat quan els regressors (X) estan correlacionats.

- Un dels objectius de l'MRL és explicar el comportament d'una variable (Y) en funció d'una sèrie de variables explicatives (X_1 ..... X_k). Per a això s'han de separar els efectes de cadascun dels regressors sobre Y.

- Si les variables explicatives tendeixen a moure's conjuntament (és a dir, estan correlacionades), el model presentarà cert grau de multicol·linealitat i la separació dels efectes individuals de cada X sobre Y es veurà dificultada. (Exemple: l'experiència laboral i l'edat com a variables explicatives dels salaris.)

Multicol·linealitat PERFECTA

- Es tracta d'un cas teòric ja que en la pràctica no sol produir-se aquest tipus de multicol·linealitat.

- Es produeix solament quan hi ha un regressor que és C.L. exacta d'altres regressors del model (incompliment de la hipòtesi 8). Per exemple: parany de les fictícies, despesa mesurada en euros i en dòlars...

- Si hi ha col·linealitat perfecta no és possible efectuar estimacions dels paràmetres, els estimadors no estan definits i per això ni tan sols cal plantejar-se quines propietats tenen els estimadors.

- Recorda que la col·linealitat teòrica és una "curiositat" teòrica. En la pràctica no sol donar-se i si es donés no podríem estimar, la qual cosa ens avisaria de la seua existència.

Multicol·linealitat no perfecta (o senzillament multicol·linealitat)

- En la pràctica no solen donar-se situacions en les quals es presenti multicol·linealitat perfecta. No obstant això, **sí que és habitual que les variables explicatives presenten cert grau de col·linealitat**.

- Com més alta siga la correlació entre els regressors, més difícil serà separar els seus efectes, fent que augmenten les variàncies dels estimadors MCO, sent per tant major el risc d'obtenir estimacions imprecises.

- Si l'elevada correlació entre els regressors fa que els resultats de l'estimació siguen "insatisfactoris", llavors direm que el model sofreix de multicol·linealitat.

- La multicol·linealitat és un problema de grau: tota regressió sofreix d'aquest problema, per la qual cosa només es diu que existeix multicol·linealitat quan es creu que està afectant seriosament els resultats de regressió.

La col·linealitat és part d'un problema més general que és la precisió dels estimadors

- En el tema 3 vam veure que la variància dels estimadors depenia de 4 factors:

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{N * Var(x_j)\ (1 - R_j^2)}$$

- Com menor siga la variància, més precisos seran els estimadors i per tant més fiables seran les estimacions.

1) Si augmenta la grandària mostral (N) disminueix la variància de l'estimador.

2) Si augmenta la variabilitat del regressor (*Var(x_j)*) disminueix la variància de l'estimador.

3) Si augmenta la variància de les pertorbacions (σ^2) augmenta la variància de l'estimador.

4) El quart factor ($(1 - R_j^2)$) està relacionat amb la col·linealitat (dels regressors).

Vegem-ho…..

Com afecta la col·linealitat a la variància dels estimadors?

- (6.1 col·linealitat o multicol·linealitat). Si hi ha una elevada col·linealitat entre els regressors (X's), els estimadors MCO segueixen sent ELIO, però la seva variància és elevada, és a dir són **poc precisos**.

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{N * Var(x_j)\ (1 - R_j^2)}$$

- Si hi ha un elevat grau de col·linealitat entre els regressors, estem dient que X_j està molt correlacionat amb un o més dels altres regressors del model; i si això ocorre… és com si els altres regressors poguessin explicar el comportament de X_j

- El quart factor (R_j^2) representa la proporció de la variació total en X_j que és explicada per la resta de regressors. Si la col·linealitat és elevada, R_j^2 serà elevat.

- **Si R_j^2 augmenta, la variància dels estimadors també augmenta**.

- Si $R_j^2 = 1$ estaríem en el cas de multicol·linealitat perfecta. En aquest cas, no es poden obtenir els estimadors MCO, i la variància dels estimadors és com si fos infinita.

- Ara bé, Què ocorre si R_j^2 és elevat?

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{N * Var(x_j)\ (1 - R_j^2)}$$

- La variància de MCO serà elevada, la qual cosa és un problema per a la precisió i fiabilitat del nostre estimador.

- No obstant això, no hi ha un valor que R_j^2 ens diga inequívocament quan la multicol·linealitat constitueix un problema greu.

- En realitat el problema de la multicol·linealitat ($R_j^2 \rightarrow 1$) és similar a tenir una mostra petita (N petit) o amb poca variabilitat de X_j ($Var(x_j)$), o un fenomen amb molt de soroll (σ^2) .

- És a dir, la multicol·linealitat és un dels factors que ens poden portar a tenir estimadors amb insuficient precisió.
- Una gran correlació entre dos regressors generalment és irrellevant en l'estimació dels efectes dels altres regressors.

Com podem detectar si el nostre model té problemes de col·linealitat?

- Com que la multicol·linealitat és un problema essencialment mostral associat a les dades de les variables explicatives, no es compta amb un mètode únic per detectar quan la multicol·linealitat constitueix un problema seriós. El que tenim en realitat són unes regles generals (algunes formals i altres informals), com ara:

1) Altes correlacions entre parelles de regressors.
2) Petits canvis en la mostra provoquen grans canvis en les estimacions.
3) Un R^2 elevat (que significa que els regressors expliquen un alt % de la variància de Y, per la qual cosa els regressors seran conjuntament significatius) però poques variables significatives individualment.
4) FAV

Si tenim un problema de col·linealitat, podem solucionar-ho?

- Les dades en Economia es recullen per recopilació passiva, poc podem fer si resulta que dos regressors estan correlacionats en la meva mostra. Algunes vegades hi ha "solucions" imaginatives…

  a) Tractar d'obtenir una mostra en què les variables explicatives estiguen menys correlacionades. Millora del disseny mostral, p. ex. si es vol veure com afecta l'edat i l'experiència en el salari, sembla que en les dones aquestes dues variables estan menys correlacionades.

  b) Eliminar el regressor col·lineal. El problema de la multicol·linealitat essencialment és un problema d'insuficiència de la informació mostral per estimar de forma precisa els efectes individuals. De vegades, estem més interessats en uns paràmetres que en uns altres però si optem per eliminar el regressor col·lineal podem incórrer en un biaix d'especificació que dóna lloc a estimadors esbiaixats, i per això la solució pot implicar problemes fins i tot més greus que els que genera la multicol·linealitat.

  c) Utilitzar informació extramostral. Per exemple, conèixer el valor de certs paràmetres per altres recerques. Establint restriccions sobre els paràmetres o combinant les variables si són conceptualment semblants, d'aquesta forma es redueixen els paràmetres a estimar i es pal·lien possibles deficiències mostrals (p. ex.: la despesa en publicitat o l'educació del pare i de la mare)

  d) Transformar les variables (ràtios, taxes de creixement, primeres diferències, desviacions respecte a una tendència). Ha de tenir-se en compte la teoria econòmica per veure si tenen sentit.

- De vegades la multicol·linealitat la provoca el mateix investigador en demanar-los massa a les dades. Per exemple: volem estimar l'efecte de la despesa escolar de les escoles en el rendiment dels estudiants. Per a això s'introdueixen en el model tres variables de despesa, la despesa en salaris, la despesa en ordinadors i la despesa en activitats complementàries.

Si tenim un problema de col·linealitat que porta a estimadors imprecisos podem solucionar-ho o alleujar-ho?

- Ja hem vist que solucionar o alleujar la col·linealitat és complicat i no sempre és possible.

- També podríem intentar alleujar els efectes que provoca la col·linealitat intentant que els altres 3 factors que afecten la precisió de les meues estimadores contribueixin a reduir la variància:

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{N * Var(x_j)\ (1 - R_j^2)}$$

1) Recollir més dades (↑N). Evidentment, això no sempre és factible.

2) Augmentar la variabilitat del regressor del qual vull estimar l'efecte (↑Var(x_j))

3) Reduir σ^2. És possible?

Per finalitzar dues reflexions:

1) Als investigadors ens agradaria entendre tots els fenòmens i estimar de forma precisa els efectes de tots els regressors, però això no és sempre possible.

2) Finalment, cal preguntar-se: la multicol·linealitat és necessàriament un problema? Si l'objecte de la regressió no és la interpretació o estimació dels efectes individuals dels diferents regressors sinó tan sols la predicció del regressand, en aquest cas la col·linealitat no és un problema greu. Si l'objectiu del meu model és la predicció, llavors un elevat R^2 és suficient perquè el model compleixi amb el seu objectiu de ser adequat per a la predicció.

Exemple (Examen 4/7/2006):
Una multinacional desitja analitzar els factors que determinen els salaris dels seus treballadors i per a això es disposa d'una mostra per la qual es coneix:

SALARI: salari brut anual del treballador en milers d'euros.
EXPLAB: experiència laboral del treballador en anys.
SEXE: variable fictícia que pren valor 1 si el treballador és home i 0 en cas contrari.
TAMSUC: grandària de la sucursal mesurada pel nombre de treballadors.

Cuadro 2

Regresores : 1,TAMSUC,EXPLAB,EDAD,SEXO
Muestra : 1 - 150 Nº Observaciones : 150

Regresores	Coeficiente	Desv. Típica	Estadís. t	Prob>|t|
1	11.483656	0.916949	12.52	0.0000
TAMSUC	0.801402	0.001716	467.03	0.0000
EXPLAB	-1.658855	3.813179	-0.44	0.6642
EDAD	2.937429	3.773577	0.78	0.4376
SEXO	1.364399	0.184895	7.38	0.0000

Media Var. Dependiente:	462.7805	Des. Típ. Var. Depen.:	45.9906
Error Típico Regresión:	1.1257	Suma Cuadrados Resid.:	183.7580
R Cuadrado :	0.9994	R Cuadrado Corregido :	0.9994
Logaritmo de Verosim. :	-228.0646	Criterio AIC :	3.1075
Estadístico F(4, 14):	62134.389	Prob > F :	0.0000
Estadís. Durbin-Watson:	1.9727	Est. Autocorrelación :	0.0136

Cuadro 1

Coeficientes de correlación

EDAD	EXPLAB	0.999
	SEXO	0.071
	TAMSUC	0.040
	SALARIO	0.346
EXPLAB	SEXO	0.071
	TAMSUC	0.040
	SALARIO	0.346
SEXO	TAMSUC	-0.00
	SALARIO	0.033
TAMSUC	SALARIO	0.950

Amb la informació proporcionada pels quadres 1 a 2, es demana:

a) Interpreteu els coeficients i analitzeu la bondat d'ajust.

b) Influeix l'experiència laboral en el salari percebut? Raoneu la resposta.

c) Hi ha discriminació per raó de sexe? Raoneu la resposta.

d) Plantegeu un model de regressió que permeti analitzar si la diferència salarial entre homes i dones augmenta amb l'experiència laboral i indiqueu com realitzaríeu el contrast pertinent.

e) Detecteu algun problema en l'estimació del model? Com ho solucionaríeu? Raoneu la resposta.

TEMA 6. INCOMPLIMENT DE LES HIPÒTESIS BÀSIQUES

6.1 Multicol·linealitat

6.2 Normalitat

6.3 Heterocedasticitat

6.4 Autocorrelació

Que ocorre si la pertorbació no segueix una distribució Normal?

- (6.2 no normalitat). Si les pertorbacions no segueixen una distribució normal, els estimadors continuen sent ELIO, **però ja no seguiran una distribució normal**; per tant, tindrem problemes per fer contrastos amb els estadístics habituals (t, F…).

- Si la pertorbació no es distribueix Normal, els estimadors tampoc es distribuiran Normal i el ratio t o t-estadístic no sabem quina distribució segueix; per tant no podem fer contrastos d'hipòtesi amb els estadístics habituals perquè no sabem com es distribueixen

Com podem contrastar si les pertorbacions segueixen una distribució Normal?

- Com podem contrastar alguna cosa sobre les pertorbacions (u) si no són observables?

  - Per a contrastar alguna propietat de les u utilitzarem els residus (que sí que són observables una vegada estimat el model).

- Per a contrastar si les u son Normals mirarem si els residus són Normals.

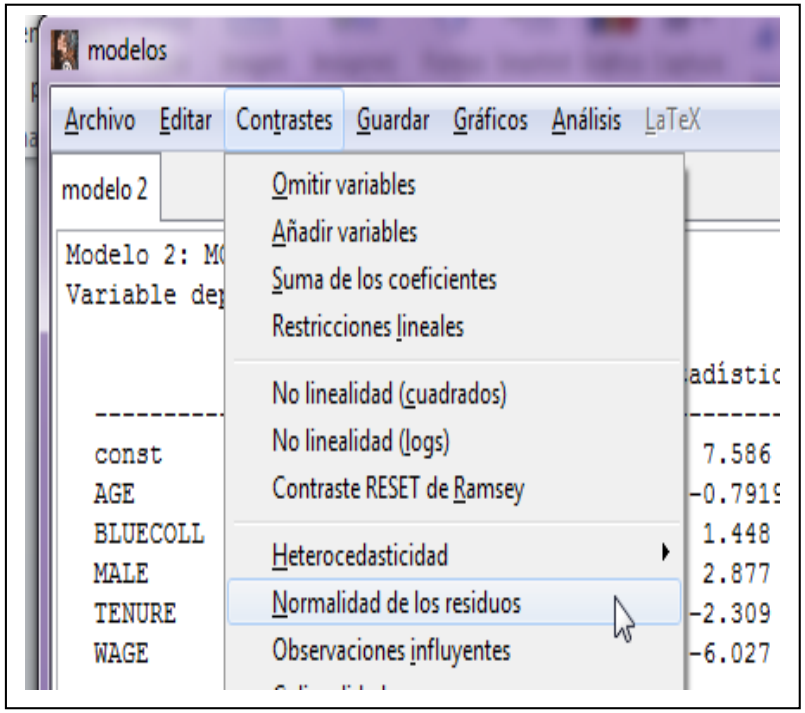
- La distribució Normal és molt coneguda i es caracteritza per ser simètrica i tenir una forma típica (coeficient d'asimetria ($\hat{\gamma}_1$ =0), coeficient de curtosi ($\hat{\gamma}_2 = 3$)).

- Per a veure si els residus es distribueixen aproximadament com una Normal utilitzarem l'estadístic proposat per Jarque i Bera:

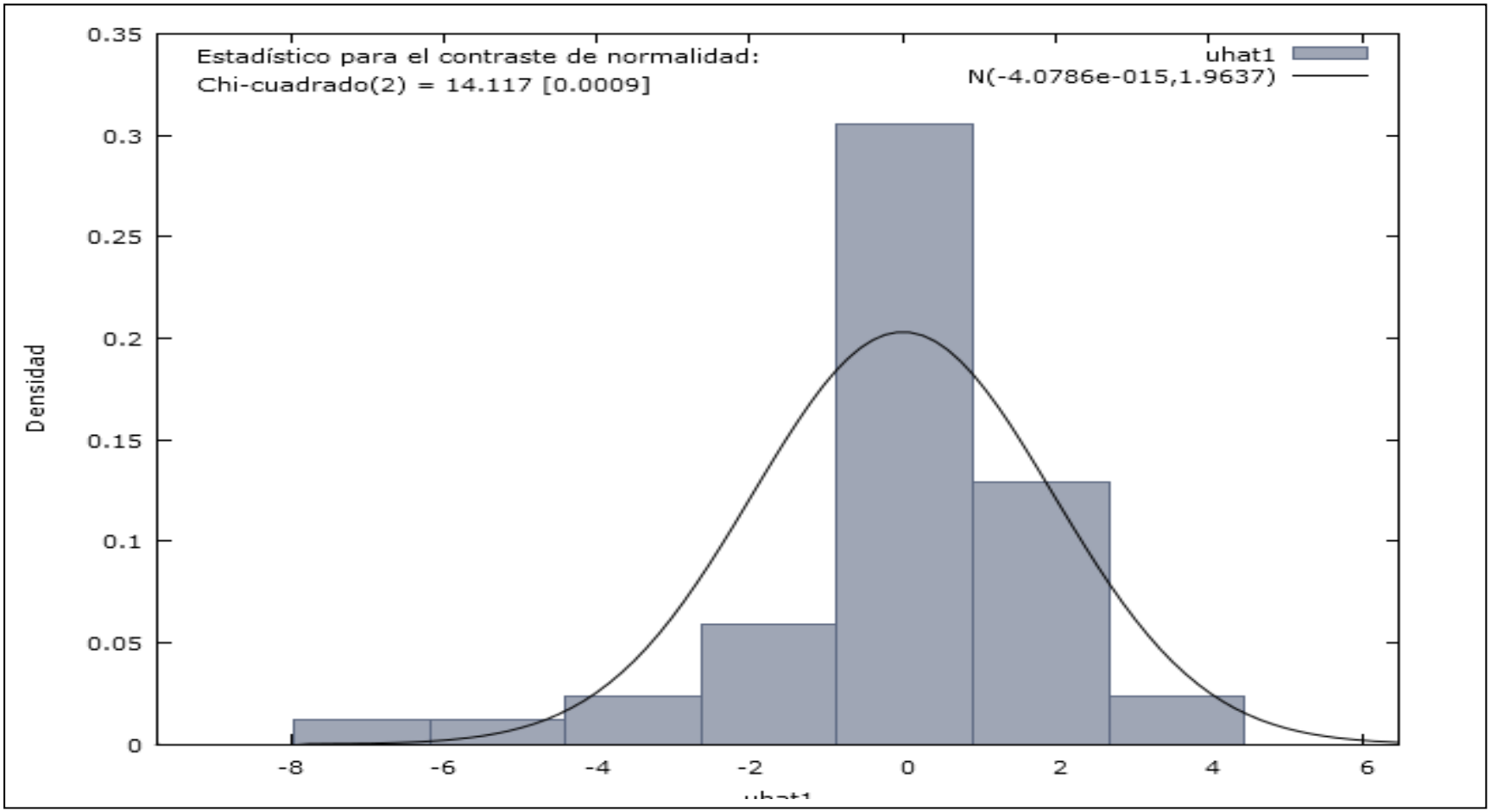
$$JB = [\ \frac{T}{6}(\hat{\gamma}_1)^2 + \frac{T}{24}(\hat{\gamma}_2 - 3)^2\ ] \rightarrow \chi^2_{(2)}$$ (sota la H_0: u es distribueix Normal)

- Cal recordar que JB és un contrast asimptòtic.

```
Modelo 2: MCO, usando las observaciones 1-48
Variable dependiente: ABSENT

            Coeficiente   Desv. Típica   Estadístico t   Valor p
  ------------------------------------------------------------
  const      12.4436        1.64043          7.586       2.14e-09 ***
  AGE        -0.0372111     0.0469913       -0.7919      0.4329
  BLUECOLL    0.968460      0.668824         1.448       0.1550
  MALE        2.04929       0.712224         2.877       0.0063   ***
  TENURE     -0.150770      0.0652833       -2.309       0.0259   **
  WAGE       -0.0442879     0.00734790      -6.027       3.63e-07 ***

Media de la vble. dep.  4.500000   D.T. de la vble. dep.   3.786875
Suma de cuad. residuos  161.9505   D.T. de la regresión    1.963661
R-cuadrado              0.759717   R-cuadrado corregido    0.731112
F(5, 42)                26.55884   Valor p (de F)          5.28e-12
Log-verosimilitud      -97.29520   Criterio de Akaike      206.5904
Criterio de Schwarz     217.8176   Crit. de Hannan-Quinn   210.8332
```



- Es compleix la hipòtesi de Normalitat?
- Mireu els residus si són Normals. Podeu fer un gràfic però també fer el contrast de JB.


Estadístico para el contraste de normalidad:
Chi-cuadrado(2) = 14.117 [0.0009]
uhat1
N(-4.0786e-015,1.9637)
Densidad
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-8
-6
-4
-2
0
2
4
6

TEMA 6. INCOMPLIMENT DE LES HIPÒTESIS BÀSIQUES

6.1 Multicol·linealitat

6.2 Normalitat

6.3 Heterocedasticitat

6.4 Autocorrelació

- (6.3 heterocedasticitat). Si les pertorbacions no són homocedàstiques, els estimadors MCO continuen sent sense biaix, però **ja no seran òptims**.

Recordeu què significava homocedasticitat? … Sí, és clar!

Hipòtesi 4) $\text{Var}(u_i) = \sigma^2 \quad \forall i = 1, \ldots\ldots n$ (HOMOCEDASTICITAT)

- Si hi ha homocedasticitat la variabilitat o volatilitat és la mateixa per a tots els individus.

  Sota la hipòtesi d'homocedasticitat s'han d'estimar k+1 paràmetres: k coeficients (β) més la variància de les pertorbacions (σ^2).

- Si no es compleix la hipòtesi d'homocedasticitat; és a dir, **si hi ha heterocedasticitat** ($\text{Var}(u_i) = \sigma_i^2$) la volatilitat serà diferent per a alguns dels individus.

  En aquest cas caldria estimar (k+N) paràmetres: k coeficients (β) més una variància per a cada individu. Atès que la grandària mostral és N, el nombre de paràmetres a estimar és major que el nombre d'observacions disponibles. Evidentment això no és possible i serà necessari suposar que l'heterocedasticitat segueix un determinat esquema de comportament.

Heterocedasticitat. Naturalesa del problema

- Exemples: en l'exemple de l'examen i la nota podria haver-hi heterocedasticitat si els individus no eren iguals de nerviosos, o si l'examen era corregit per 2 professors diferents.

- Exemples de llibres: pàgines …

- Per exemple, si analitzem la recerca en les empreses: les empreses petites tenen poques possibilitats d'invertir quantitats elevades en R+D i les oscil·lacions de la despesa entre elles és poc rellevant. No obstant això, les empreses grans poden invertir gran quantitat dels seus ingressos en R+D o no invertir gens, raó per la qual hi haurà una elevada volatilitat.

- Un altre exemple seria la demanda d'un bé de luxe entre diferents consumidors. cal esperar que conforme augmenti la renda dels consumidors augmenta la variabilitat del comportament dels consumidors.

- L'heterocedasticitat és un problema que és més probable que aparegue amb dades de tall transversal. Variables como la renda o la grandària porten a —o estan gairebé sempre associades a— la presència d'heterocedasticitat. (*Però també pot ocórrer amb dades de sèrie temporal. Per exemple, la volatilitat de les accions no és la mateixa en temps de boom que en recessió.*)

- Podem pensar que la causa última que existeixi heterocedasticitat és "simplement" que els individus són heterogenis i és impossible recollir en el model tots els aspectes o variables que afecten el comportament dels diferents individus.

Quines conseqüències té l'heterocedasticitat?

- Els estimadors MCO continuen sent sese biaix, però **ja no seran òptims**. Hi ha un mètode millor per estimar els paràmetres del model: MCG.
- Per la seua banda, els intervals de confiança i els contrastos que es construeixen suposant homocedasticitat no seran vàlids, ja que la variància dels estimadors ja no segueix l'expressió habitual:

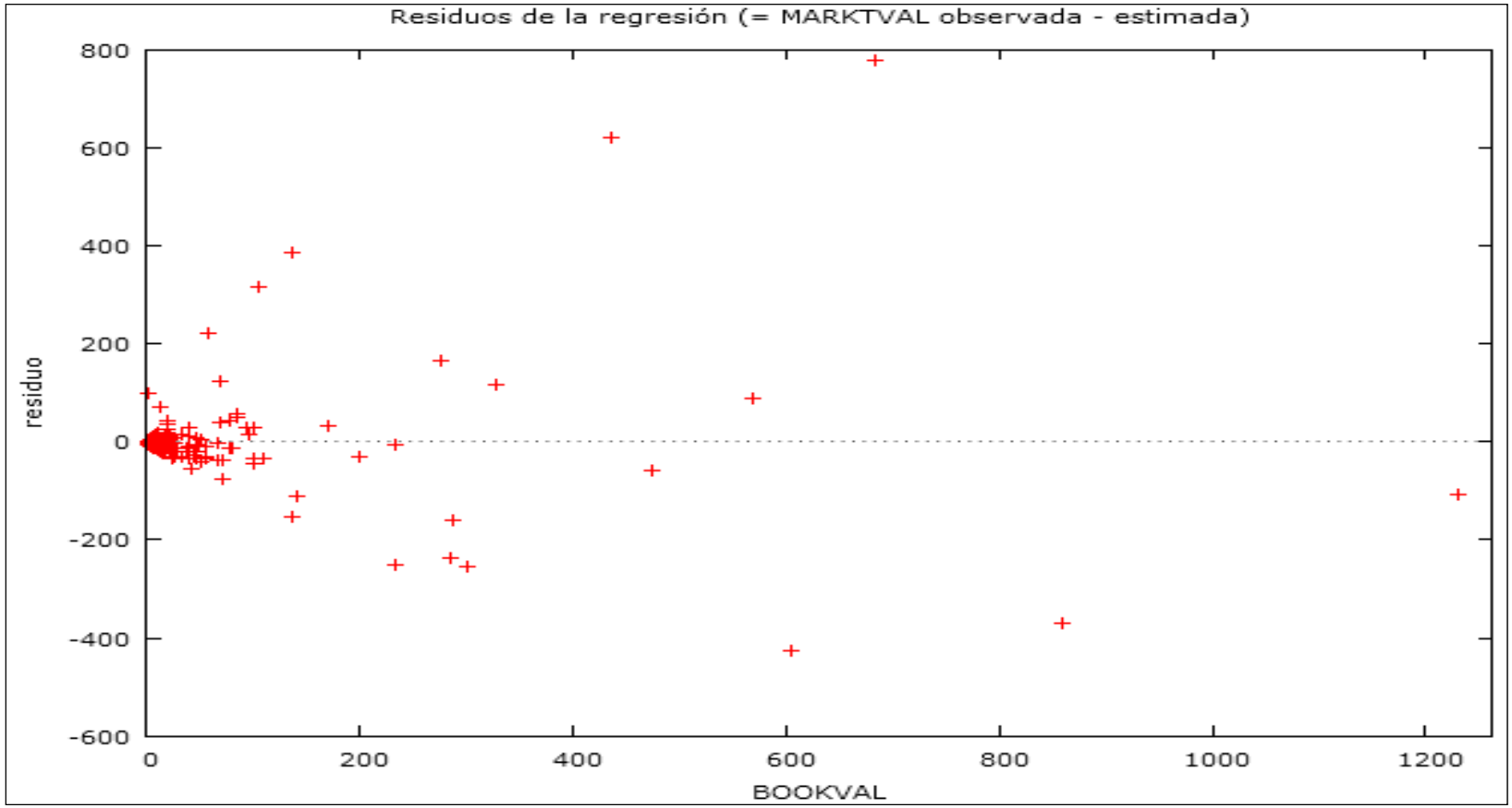
$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{N * Var(x_j)\ (1 - R_j^2)}$$

Si el meu model presenta heterocedasticitat… Què faig?

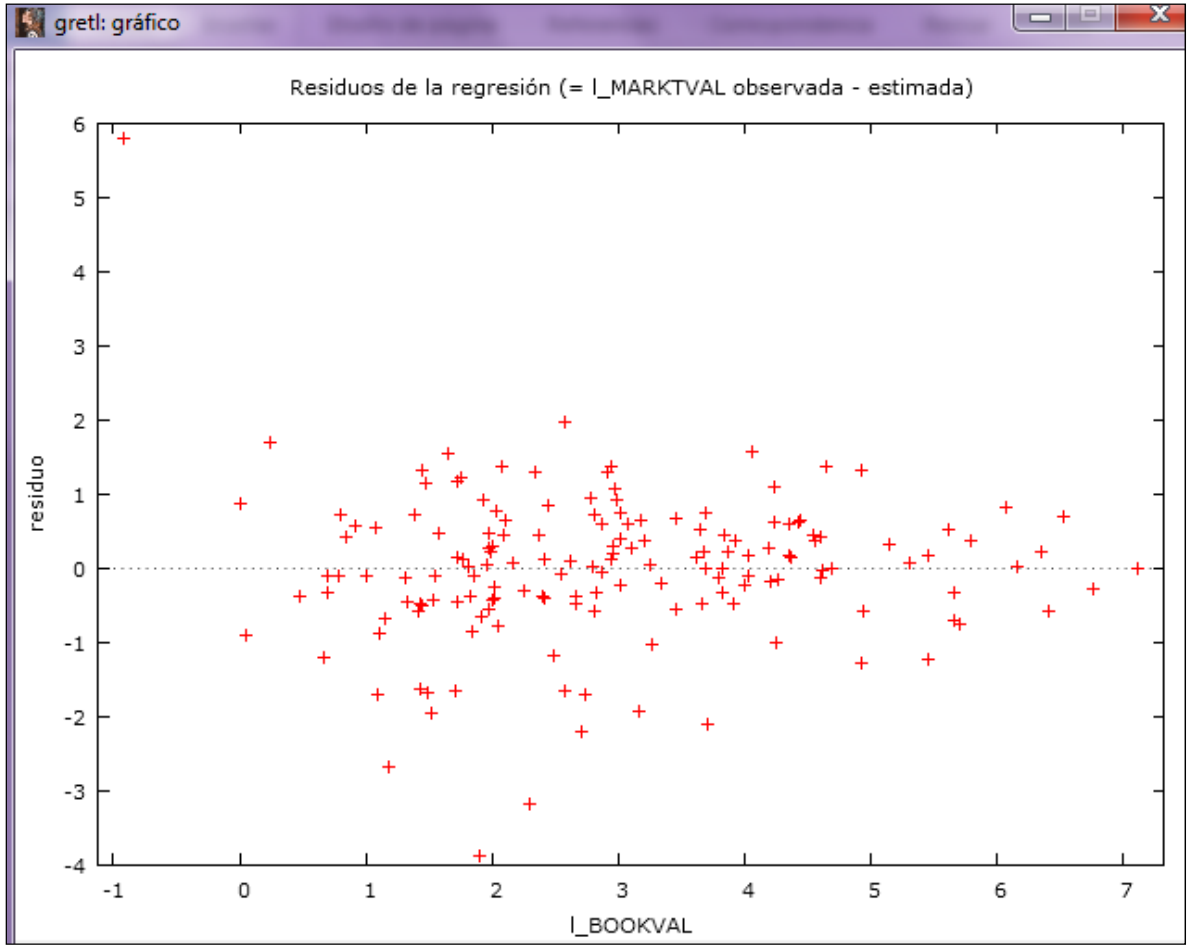
- Els estimadors MCO no són òptims, ja no són de menor variància, ja no són els més precisos. Hi ha un mètode millor per estimar els paràmetres del model: MCG. Lògicament si hi ha un mètode millor hauria d'utilitzar-lo; o sigui, **si el teu model té heterocedasticitat cal estimar per MCG.**
- Però això no sempre és "factible".
- Si continues estimant el model per MCO, malgrat saber que no és òptim, almenys s'han de calcular les variàncies de forma correcta (robustes al problema d'heterocedasticitat).

Com es detecta si hi ha heterocedasticitat? Contrastos d'heterocedasticitat

- Com una primera aproximació, pot analitzar-se el gràfic dels residus mínims quadràtics enfront de la Y-estimada o enfront del regressor que creiem que està generant l'heterocedasticitat.
- Al principi del curs, en la pràctica P2-0 vam veure el següent gràfic: hi ha heterocedasticitat?



- Més tard, en la mateixa pràctica P2-0, es va estimar el mateix model però en logaritmes i es va obtenir el següent gràfic: hi ha ara heterocedasticitat?



Contrastos d'heterocedasticitat: contrast de White

- S'estima un model i se sospita que hi ha heterocedasticitat: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
- Les hipòtesis nul·la i alternativa del contrast de White són:

$$H_0: \sigma_i^2 = \sigma^2 \quad \forall i$$
$$H_1: \sigma_i^2 = f(X's)$$

- Els passos per efectuar el contrast són els següents:

1) S'estima el model original i s'obtenen els residus mínims quadràtics.
2) S'efectua una regressió auxiliar. En aquesta regressió auxiliar el tornant seran els residus al quadrat del model original, i els regressors seran els regressors del model original, els regressors originals al quadrat i opcionalment els productes creuats dels regressors del model original. En el nostre exemple (incloent els productes creuats) la regressió auxiliar serà:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i}^2 + \alpha_3 X_{3i}^2 + \alpha_4 X_{2i} + \alpha_5 X_{3i} + \alpha_6 X_{2i} X_{3i} + \eta_i$$

En la regressió auxiliar sempre ha d'haver-hi terme independent i cal eliminar les redundàncies.

3) Es calcula el R^2 de la regressió auxiliar (R_{RA}^2)

- L'estadístic de White és nR_{RA}^2
- Aquest estadístic, ball la hipòtesi nul·la d'homocedasticitat, es distribueix asimptòticament com una khi-quadrat amb **m** graus de llibertat, sent **m** el nombre de regressors en la regressió auxiliar exclòs el terme independent (nombre de regressors de la RA menys 1).

Tractament de l'heterocedasticitat:
Què faig si el contrast de White detecta que existeix un problema d'heterocedasticitat

Què fem si hem estimat un model per MCO i els contrastos d'heterocedasticitat ens indiquen que hi ha problemes d'heterocedasticitat? Recordeu que si existeix heterocedasticitat, els estimadors mínims quadràtics són no òptims i la inferència tampoc és vàlida, **perquè les desviacions típiques de l'estimador s'han de calcular d'una altra manera.**

Hi ha dues possibles formes d'actuar:

1. Estimar el model per MCG (o de forma equivalent transformar el model i estimar el "model transformat" per MCO). Per a això és necessari conèixer o estimar la variància de les pertorbacions. En aquest cas obtenim estimadors òptims i la inferència és vàlida.

2. Si, havent-hi encara heterocedasticitat, es continua estimant els paràmetres per MCO, almenys haurem de corregir l'estimació de les variàncies-covariàncies dels estimadors MCO. És a dir, haurem d'obtenir una **estimació consistent de les variàncies-covariàncies** dels estimadors**.** En aquest cas la inferència és vàlida però els estimadors mínims quadràtics no són òptims.

MCG (Estimació per mínims quadrats generalitzats)

- Per estimar de manera òptima un model amb pertorbacions heterocedàstiques és necessari conèixer o estimar l'esquema d'heterocedasticitat. La forma més habitual d'implementar MCG consisteix en: una vegada conegut o aproximat el patró d'heterocedasticitat, l'aplicació de MCG es realitza en dues etapes:

    1. Transformem el model de manera que compleixa les h.e.b.
    2. Apliquem MCO al model transformat.

- Exemple: suposem que tenim un MRLM ($Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$) però que u_i és heterocedàstica $Var(u_i) = \sigma_i^2 = f(X_{ji})$

- *Model transformat* (model original dividit per la desviació típica d'u_i):

$$\frac{Y_i}{\sqrt{f(X_{ji})}} = \beta_1 \frac{1}{\sqrt{f(X_{ji})}} + \beta_2 \frac{X_{2i}}{\sqrt{f(X_{ji})}} + \cdots + \beta_k \frac{X_{ki}}{\sqrt{f(X_{ji})}} + \frac{u_i}{\sqrt{f(X_{ji})}}$$

Es pot comprovar que les pertorbacions d'aquest model són homocedàstiques. Per això en la segona etapa s'estima per MCO el model transformat i s'obtenen estimadors ELIO. Observeu que en aquest model s'està ponderant cada observació per l'invers del valor de la desviació típica de la pertorbació, per això el procediment anterior es denomina MCP (mínims quadrats ponderats).

Exemple (Examen 6/2/2007)

Amb una mostra de 100 individus s'ha estimat la següent funció de demanda del ben Y:

$$LY_i = \beta_1 + \beta_2 LI_i + \beta_3 NE_i + \beta_4 SEX_i + \beta_5 SINGLE_i + \beta_6 HIJO_i + \beta_7 E_{i_I} + \beta_8 URB_i + u_t$$

on: LY= Logaritme de la despesa anual en el ben Y
LI= Logaritme de l'ingrés anual
NE= Nivell d'estudis mesurat pel nombre d'anys dedicats a la formació
SEX= Variable dicotòmica que pren valor 1 si l'individu és un home i 0 si és una dona
SINGLE= Variable dicotòmica que pren valor 1 si l'individu té parella i 0 si no la té
FILL= Nombre de fills
I= Edat en anys
URB= Variable dicotòmica que pren valor 1 si l'individu viu en un nucli urbà i 0 si viu en un nucli rural

A la vista dels resultats proporcionats en els quadres 1 i 2:

a) Expliqueu i contrasteu l'heterocedasticitat en aquest model.
b) Expliqueu la raó per la qual creieu que el model ha estat estimat amb una matriu consistent amb heterocedasticitat i quines conseqüències se'n deriven.

Quadre 1

Mínimos C.: Matriz de covarianzas consistente bajo heterocedasticidad

○ Estándard ◉ Resultados de la regresión

Regresores : 1,LI,NE,SEX,SINGLE,HIJO,E,URB
Muestra : 1 - 100 Nº Observaciones : 100

Regresores	Coeficiente	Desv. Típica	Estadís. t	Prob>|t|
1	-9.897638	1.247063	-7.94	0.0000
LI	2.335210	0.121779	19.18	0.0000
NE	1.528719	0.036920	41.41	0.0000
SEX	0.173770	0.177072	0.98	0.3290
SINGLE	0.466186	0.099417	4.69	0.0000
HIJO	-0.150424	0.035316	-4.26	0.0000
E	0.350672	0.003889	90.18	0.0000
URB	0.375061	0.165948	2.26	0.0262

Media Var. Dependiente: 50.3674 Des. Típ. Var. Depen.: 4.9057
Error Típico Regresión: 0.4670 Suma Cuadrados Resid.: 20.0666
R Cuadrado : 0.9916 R Cuadrado Corregido : 0.9909

Quadre 2. Contrast de White

Estadístico Chi-cuadrado (11): 48.478 Prob > Chi-cuadrado : 0.0000

TEMA 6. INCOMPLIMENT DE LES HIPÒTESIS BÀSIQUES

6.1 Multicol·linealitat

6.2 Normalitat

6.3 Heterocedasticitat

6.4 Autocorrelació

Si es compleixen les h.e.b.

- Si es compleixen les h.e.b., obtenim els següents resultats:

3) Els estimadors MCO són ELIO.

4) Els estimadors MCO es distribueixen com:

$$\hat{\beta}_k \rightarrow N(\beta_k, \sigma^2_{\hat{\beta}_k})$$

amb

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SCT_j(1-R_j^2)} = \frac{\sigma^2}{N * Var(x_j)\,(1-R_j^2)}$$

Entre les h.e.b. aquesta la hipòtesi de no autocorrelació en les pertorbacions

5) $Cov(u_i, u_s) = 0 \quad ; \quad \forall\, i \neq s$ (NO AUTOCORRELACIÓ) (apartat 6.4)

Si tenim autocorrelació en les pertorbacions….

Si les pertorbacions estan correlacionades, els estimadors MCO continuen sent sense biaix, però **ja no seran òptims**…..

Autocorrelació

- Si es compleix la h.eb. de no autocorrelació $E[u_t u_s] = Cov(u_t\, u_s) = 0 \quad \forall t \neq s$. La pertorbació corresponent a una observació és independent de la pertorbació de qualsevol altra observació.
- Aquesta hipòtesi no sol ser molt realista amb dades de sèries temporals.
- Amb dades de sèries temporals la situació habitual és d'autocorrelació

$$E[u_t u_s] = Cov(u_t\, u_s) \neq 0 \quad para\ algun\ t \neq s$$

Causes

Les causes de la presència d'autocorrelació en les pertorbacions d'un model són:

1. La naturalesa de les dades. Les dades de sèrie temporal presenten inèrcia o tendència, cosa que pot determinar que les pertorbacions estiguin correlacionades entre si.
2. Errors d'especificació, bé perquè s'ometen variables explicatives rellevants o bé s'especifica una forma funcional incorrecta.
3. Desfasaments en els efectes dels *X*.
4. Manipulació de les dades. Per exemple, dades trimestrals interpolades o per mitjanes, desestacionalització, allisats, etc.

Conseqüències

Les conseqüències de l'autocorrelació són:

1. Els estimadors de MCO seguiran sent lineals i sense biaix però no òptims.

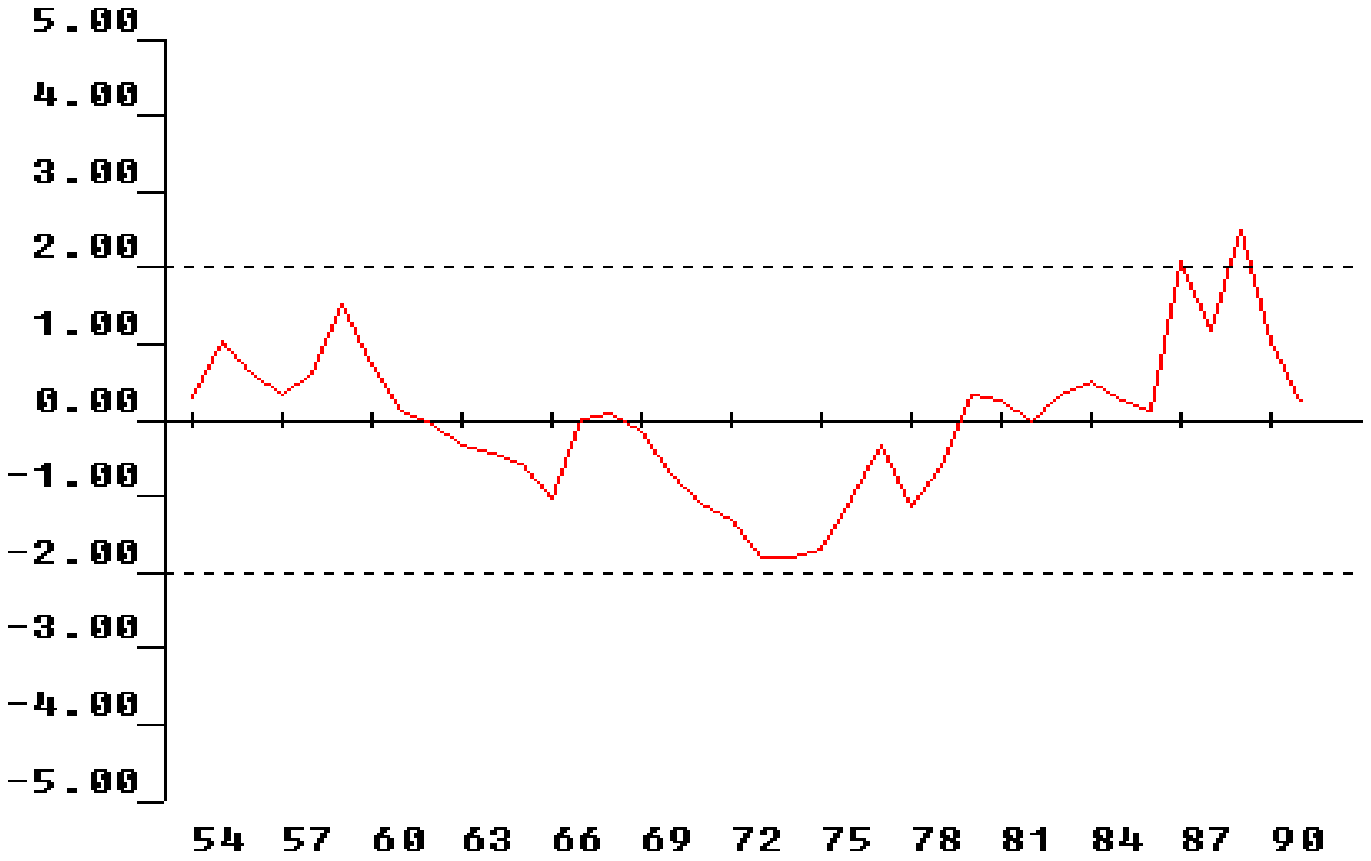
2. Els contrastos d'hipòtesis deixen de ser vàlids si s'utilitza una estimació de la variància dels estimadors que no és correcta. Cal corregir les variàncies dels estimadors. Cal estimar-les o calcular-les de forma robusta a la presència d'autocorrelació.

Com detectar si hi ha autocorrelació?

Com una primera aproximació pot analitzar-se el gràfic dels residus mínims quadràtics en funció del temps. Si els residus no estan correlacionats no presentaran cap comportament sistemàtic.

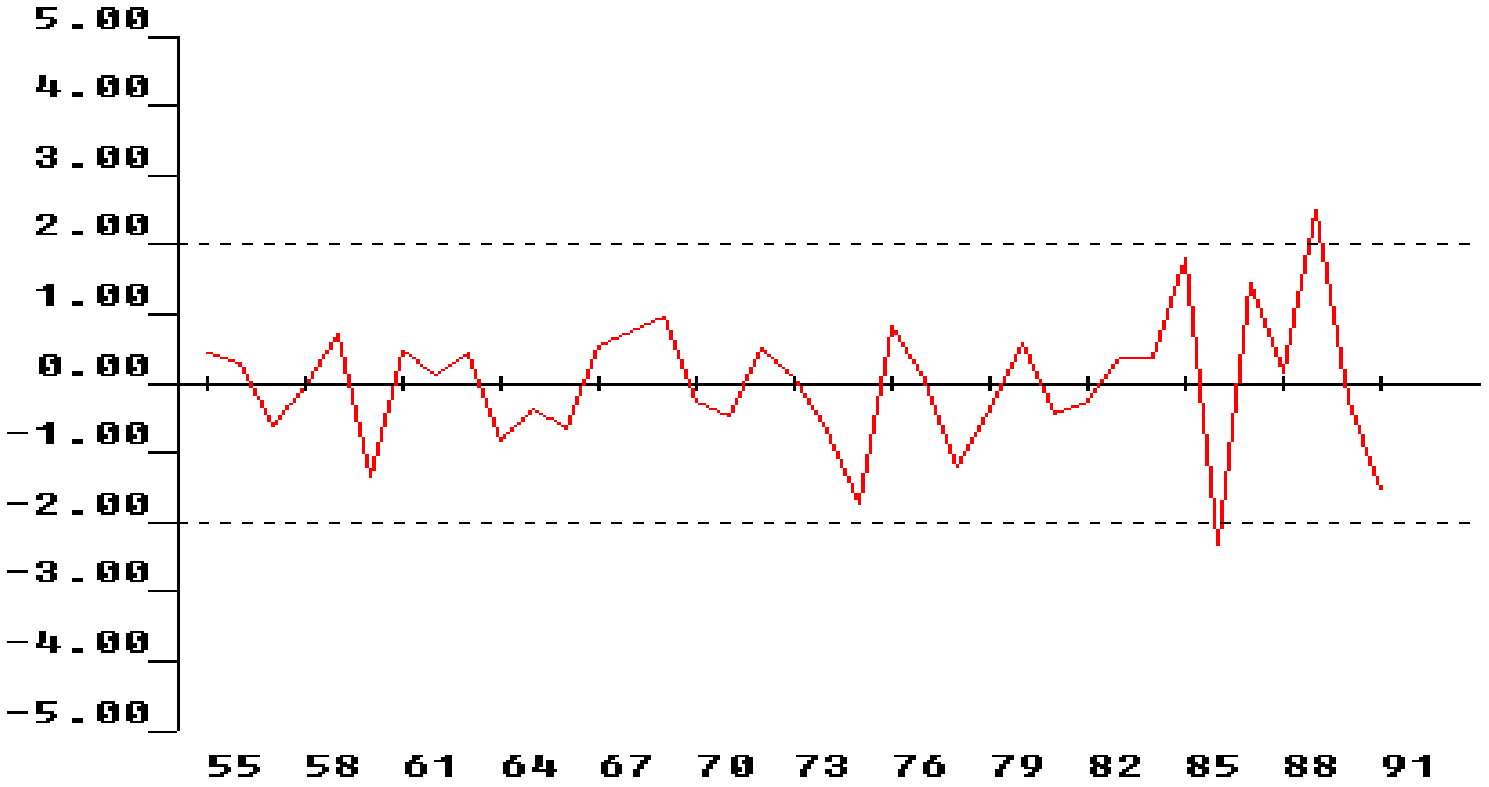
Un exemple

- El gràfic que es pot observar a continuació correspon als residus estandarditzats en l'estimació de l'equació de consum keynesiana (Uriel, 1997, 109). Com pot observar-se, no sembla que els residus siguen aleatoris ja que únicament creuen l'eix d'abscisses en poques ocasions; és a dir sembla que presenten un comportament sistemàtic:



Un altre exemple

- No obstant això, els residus del model de consum de Brown (Uriel, 1997, 142), que inclou el consum desfasat, no semblen presentar un comportament sistemàtic sinó que semblen distribuir-se de forma aleatòria:



Com detectar si hi ha autocorrelació? Contrast de Durbin-Watson

- Aquest contrast planteja que les pertorbacions segueixen un esquema autoregressiu d'ordre 1, un AR(1):

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad |\rho| < 1 \qquad \varepsilon_t \to N(0, \sigma_\varepsilon^2)$$

- La hipòtesi nul·la del contrast de DW és la de no autocorrelació (d'ordre 1): H_0: $\rho = 0$

- L'estadístic de Durbin-Watson és:

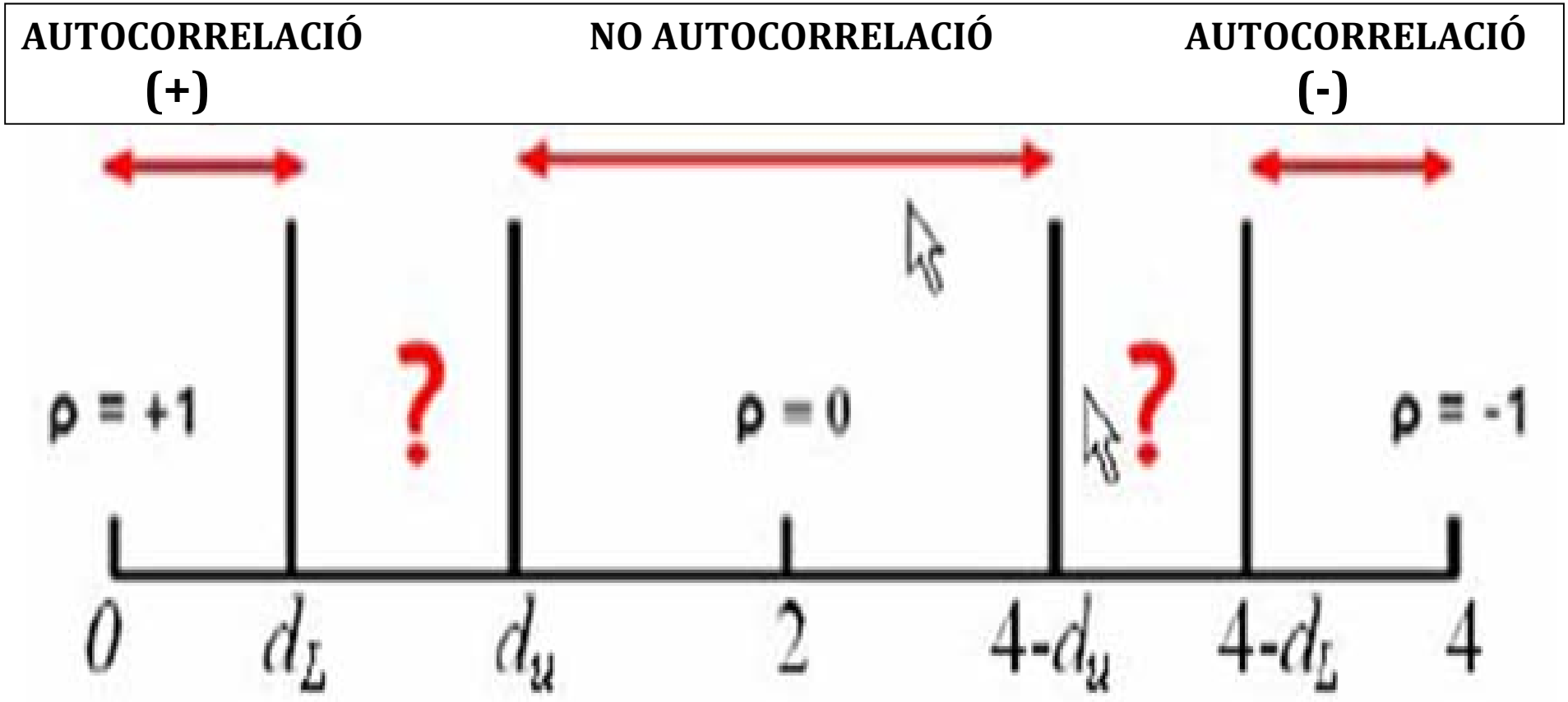
$$d = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}_t^2}$$

- Pot demostrar-se que $d \cong 2(1 - \hat{\rho})$. Per tant, com que $\hat{\rho}$ està fitat entre -1 i +1, l'estadístic de DW estarà fitat entre 0 i 4. L'estadístic d, **si la H_0 és certa, estarà entorn de 2** .

- No obstant això, no hi ha un valor crític únic que ens porti a rebutjar la hipòtesi nul·la (perquè la distribució de l'estadístic depèn de N, de les dades de les variables explicatives i del nombre de regressors. Per a k i N fixos, la distribució de l'estadístic depèn dels valors de les variables explicatives).

- Malgrat no haver-hi un valor únic, DW van trobar un límit inferior (d_L) i un límit superior (d_O), que ja no depenen de les dades de les X's, tals que si el valor de l'estadístic DW cau fora d'aquests límits pot prendre's una decisió sobre si rebutjar o no la H_0.

- Regles d'aplicació del contrastament de DW:



- Aspectes a tenir en compte a l'hora d'aplicar el contrast de DW:

  a) El model ha d'incloure constant.

  b) Les X´s són no estocàstiques.

  c) u_t segueix un procés AR(1) estacionari: $u_t = \rho u_{t-1} + \varepsilon_t \qquad |\rho| < 1 \qquad \varepsilon_t \rightarrow N(0, \sigma_\varepsilon^2)$

  d) El model no ha d'incloure com a regressor la variable depenent desfasada.

Contrastament de Breusch-Godfrey

- Aquest contrast és aplicable a esquemes autoregressius d'ordre superior al del DW i també pot utilitzar-se quan hi ha regressors desfasats.

- L'esquema que considera aquest contrast és:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t \qquad |\rho| < 1 \qquad \varepsilon_t \to N(0, \sigma_\varepsilon^2)$$

- La hipòtesi nul·la a contrastar és:

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_p = 0$$

- Els passos que es requereixen en aquest contrast són els següents:

  1) S'estima el model original i s'obtenen els residus mínims quadràtics.

  2) S'efectua la següent regressió auxiliar:

$$\hat{u}_t = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \cdots + \alpha_k X_{kt} + \gamma_1 \hat{u}_{t-1} + \cdots \gamma_p \hat{u}_{t-p} + \varepsilon_t$$

  En aquesta regressió auxiliar el regressand seran els residus del model original i els regressors són els regressors del model original i els residus retardats p períodes. La regressió auxiliar sempre ha de tenir terme independent.

  3) Designant R^2_{RA} el coeficient de determinació de la regressió auxiliar, es calcula el següent estadístic:

$$NR^2_{RA}$$

  Sota la hipòtesi nul·la d'aquest estadístic es distribueix asimptòticament com una khi-quadrat amb k+p graus de llibertat.

Un exemple

(Examen 6/2/2007)

S'ha estimat la següent funció de consum de l'economia espanyola amb dades trimestrals del període 1966-2005:

$$\hat{C}_t = 1,36 + 0,77 R_t + 0,10 W_t$$

$R^2 = 0,932$ $\quad \overline{R}^2 = 0,927$ $\quad SCR = 5,525$ $\quad$ d=0,952

on *C* és el consum, *R* és la renda disponible, *W* és riquesa. Es demana:

a) Expliqueu breument en què consisteix el problema de l'autocorrelació en les pertorbacions.
b) Contrasteu detalladament si existeix autocorrelació d'ordre un.

Un altre exemple

Archivo Herramientas Datos Ver Añadir Muestra Variable Modelo

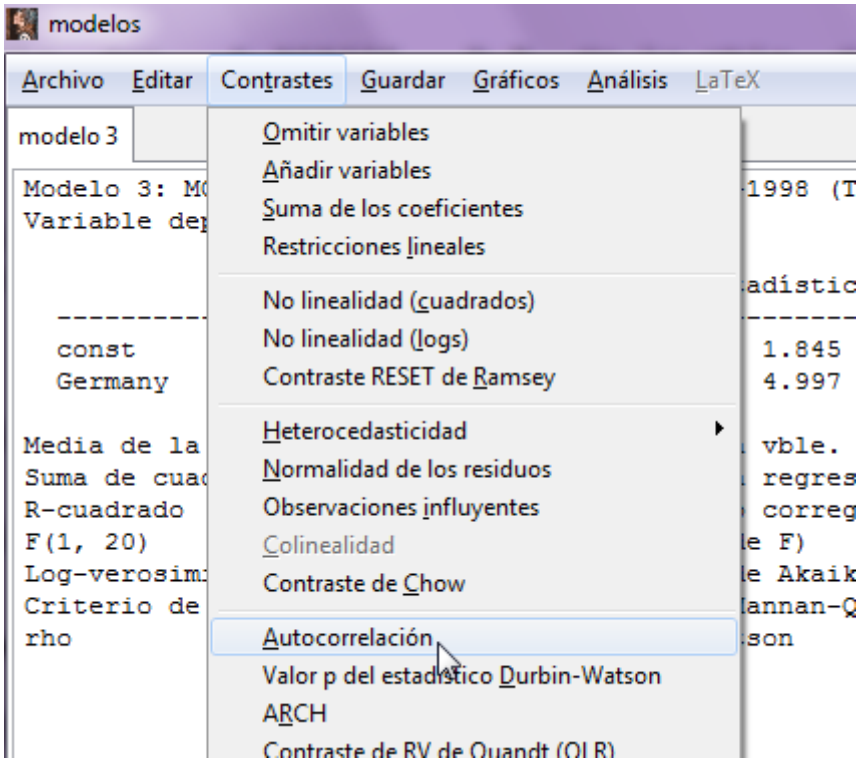
Table_1.3.gdt *

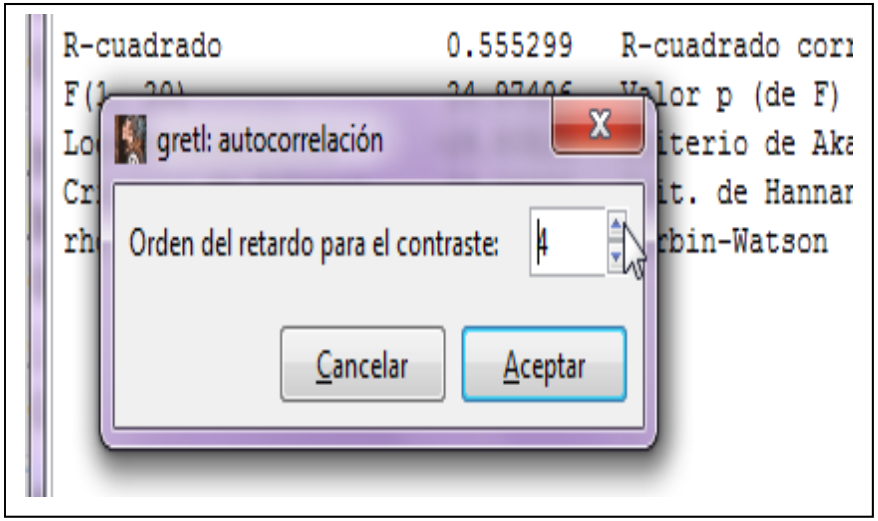
ID #	Nombre de variable	Etiqueta descriptiva
0	const	
1	Canada	currency units per U.S. dollar, Canada
2	France	currency units per U.S. dollar, France
3	Germany	currency units per U.S. dollar, Germany
4	UK	U.S. dollars per pound sterling

```
Modelo 3: MCO, usando las observaciones 1977-1998 (T = 22)
Variable dependiente: France

            Coeficiente   Desv. Típica   Estadístico t   Valor p
  ------------------------------------------------------------
  const        1.61889       0.877602          1.845     0.0799   *
  Germany      2.18583       0.437393          4.997     6.91e-05 ***

Media de la vble. dep.  5.908205   D.T. de la vble. dep.   1.256155
Suma de cuad. residuos  14.73579   D.T. de la regresión    0.858365
R-cuadrado              0.555299   R-cuadrado corregido    0.533064
F(1, 20)                24.97406   Valor p (de F)          0.000069
Log-verosimilitud      -26.80826   Criterio de Akaike      57.61651
Criterio de Schwarz     59.79860   Crit. de Hannan-Quinn   58.13054
rho                     0.828485   Durbin-Watson           0.135990
```

a) Contrasteu si existeix autocorrelació d'ordre un.

b) Contrasteu si existeix autocorrelació fins a quart ordre.





Contrastament de Breusch-Godfrey que ofereix Gretl

```
Contraste Breusch-Godfrey de autocorrelación hasta el orden 4
MCO, usando las observaciones 1977-1998 (T = 22)
Variable dependiente: uhat

             Coeficiente   Desv. Típica   Estadístico t   Valor p
  ------------------------------------------------------------------
  const       0.795064       0.972152         0.8178      0.4255
  Germany    -0.407170       0.498773        -0.8163      0.4263
  uhat_1      0.988328       0.246484         4.010       0.0010   ***
  uhat_2     -0.0360242      0.339919        -0.1060      0.9169
  uhat_3     -0.0583468      0.340075        -0.1716      0.8659
  uhat_4     -0.323000       0.336933        -0.9586      0.3520

  R-cuadrado = 0.719478

Estadístico de contraste: LMF = 10.259137,
con valor p  = P(F(4,16) > 10.2591) = 0.000259

Estadístico alternativo: TR^2 = 15.828518,
con valor p  = P(Chi-cuadrado(4) > 15.8285) = 0.00326

Ljung-Box Q' = 30.4676,
con valor p  = P(Chi-cuadrado(4) > 30.4676) = 3.93e-006
```