
Introducció a
l'econometria
Ezequiel Uriel
Universidad de Valencia 2019

Disseny portada: Jordi Uriel

Introducció a l'Econometria

Ezequiel Uriel

Traducció al valencià Ezequiel Uriel

Revisió de la traducció Mireia Moltó

2019

Universitat de València

Desitjo agrair als professors Luisa Moltó, Amado Peiró, Paz Rico, Pilar Beneito i Javier Ferri els seus suggeriments per les errades que s'han detectat en versions prèvies, i per haver-me facilitat dades per a formular exercicis. També en la detecció d'errades han col·laborat alguns alumnes. En la versió al valencià desitge agrair a Núria Espí la seua col·laboració i, especialment, a Mireia Moltó, per la revisió de la traducció realitzada per l'autor. En qualsevol cas, sóc l'únic responsable de les errades que no han estat detectades.

Taula de contingut

1 ECONOMETRIA I DADES ECONÒMIQUES

1.1 ¿Què és l'econometria?

En primer lloc, vegem alguna cosa sobre l'origen de l'econometria com a disciplina. El terme econometria es creu que va ser encunyat per Ragnar Frisch co-guanyador del primer Premi Nobel en Ciències Econòmiques en 1969, juntament amb el també econòmetra Jan Tinbergen. Tots dos van ser fundadors de l'Econometric Society el 1933. A la secció I de la constitució d'aquesta societat, s'afirma que

"L'Econometric Society és una societat internacional per a l'avanç de la teoria econòmica en la seva relació amb l'estadística i les matemàtiques. El seu principal objectiu serà promoure que tinguin com a objectiu la unificació dels enfocaments quantitatiu-teòric i quantitatiu-empíric dels problemes econòmics i que són abordats de forma constructiva i rigorosa similar al que és l'enfocament predominant en les ciències naturals".

En el primer número d'Econometrica (1933), revista de l'Econometric Society, Ragnar Frisch ens dóna una explicació sobre el significat de l'econometria:

> "Però hi ha diversos aspectes de l'enfocament quantitatiu de l'economia, tot i que cap d'ells, pres aïlladament, hauria de confondre amb l'econometria. Així, l'econometria no és el mateix que l'estadística econòmica. Tampoc és idèntica al que anomenem teoria econòmica general, tot i que una part considerable d'aquesta teoria tinga un caràcter definitivament quantitatiu. Tampoc s'ha de prendre l'econometria com a sinònim de l'aplicació de les matemàtiques a l'economia. L'experiència ha demostrat que cadascun d'aquests tres punts de vista, el de l'estadística, la teoria econòmica i les matemàtiques, és una condició necessària, però no per si mateixa una condició suficient, per a una veritable comprensió de les relacions quantitatives en la vida econòmica moderna. Es tracta de la unificació dels tres aspectes el que li dóna gran abast. I és aquesta unificació, el que constitueix l'econometria".

Avui dia, també es diu que l'econometria és l'estudi combinat dels models econòmics, estadística matemàtica i dades econòmiques. Dins el camp de l'econometria es pot distingir la teoria economètrica de l'econometria aplicada.

La *teoria economètrica* es refereix al desenvolupament de les eines i mètodes, i l'estudi de les propietats dels mètodes economètrics. La teoria economètrica pertany a l'àmbit de l'estadística.

L'*econometria aplicada* és un terme que descriu l'elaboració de models econòmics quantitatius i l'aplicació de mètodes economètrics a aquests models utilitzant dades econòmiques. L'econometria aplicada es troba, bàsicament, dins el camp de l'economia aplicada.

¿Quins són els objectius de l'econometria? Anem a considerar tres objectius de l'econometria:

1) El coneixement de l'economia real. Els mètodes economètrics ens permeten estimar les magnituds econòmiques, com la propensió marginal al consum o l'elasticitat de la mà d'obra pel que fa a l'output. Aquestes estimacions se situen en un determinat temps i espai: per exemple, a Espanya en l'últim quart del segle XX. A més de l'estimació, en què s'obtenen els valors numèrics, els mètodes economètrics ens permeten realitzar contrastos d'hipòtesis, per exemple, ¿en una funció de producció és admissible la hipòtesi de rendiments a escala constants?
2) *Simulació de polítiques econòmiques*. Els mètodes d'econometria poden ser utilitzats per simular els efectes de polítiques alternatives. Per exemple, amb un model economètric apropiat, es podria determinar, en termes quantitatius, l'efecte de diferents tipus de l'impost del tabac sobre el consum de tabac.
3) Predicció. Molt sovint els mètodes economètrics s'utilitzen per a predir valors de variables econòmiques. Quan fem prediccions tractem de reduir la nostra incertesa sobre el futur de l'economia. Això no és una tasca fàcil, ja que, en general, les prediccions només són satisfactòries quan no hi ha canvis dràstics en l'economia. Seria molt convenient també predir aquests canvis dràstics, però les prediccions amb mètodes economètrics en general no són molt bones en aquests casos, tot i que tampoc funcionen altres mètodes alternatius.

1.2 Etapes en l'elaboració d'un model economètric

En l'elaboració d'un model economètric es poden distingir les següents etapes: especificació, estimació i validació.

Si bé en una primera aproximació aquestes etapes segueixen un ordre seqüencial, en l'elaboració d'un model economètric cal, per regla general, retrocedir en més d'una ocasió dins d'aquest ordre seqüencial. És a dir, en l'anàlisi economètrica no se segueix un ordre establert per endavant, sinó que és necessari confrontar contínuament el model amb les dades i amb qualsevol altra font d'informació, amb la finalitat d'obtenir un model economètric compatible amb les dades, que permeta analitzar la realitat, oferisca millors prediccions o constituïsca una bona base per prendre decisions. Es procedeix a continuació a descriure les etapes enumerades anteriorment.

(a) Especificació

La primera etapa de l'elaboració d'un model economètric la constitueix l'especificació. En l'etapa d'especificació, considerarem quatre elements: el model econòmic, el model economètric, els supòsits estadístics del model i les dades. En aquest apartat ens referirem als tres primers elements, mentre que en l'epígraf 1.3 examinarem els diferents tipus de dades utilitzades en l'anàlisi economètrica.

El primer element que necessitem és disposar d'un model econòmic. En alguns casos, un model formal econòmic s'especifica completament mitjançant la utilització de la teoria econòmica. En altres casos, la teoria econòmica s'utilitza menys formalment en la construcció d'un model econòmic.

Després d'obtenir un model econòmic, hem de convertir-lo a un model economètric. Anem a veure amb dos exemples com es realitza aquest procés.

EXEMPLE 1.1 Funció de consum keynesiana

Keynes va formular la seva coneguda funció de consum a través de les següents proposicions:

Proposició 1: El consum és una funció de la renda, i les dues variables estan mesurades en termes reals. Si les variables es mesuren en termes reals, això vol dir que quan els consumidors decideixen la proporció de renda que dedicaran al consum, no es veuen afectats per il·lusió monetària. Analíticament, la proposició 1 es pot expressar de la següent manera:

$$cons = f(renda) \tag{1-1}$$

Proposició 2: El consum és una funció creixent de la renda, però un augment de la renda produeix sempre un augment de menor magnitud en el consum.

Aquesta proposició implica que la propensió marginal al consum és més gran que 0 (que és una funció creixent), però és menor que 1 (un augment de la renda sempre causa un augment de menor magnitud en el consum). Analíticament, la proposició 2 es pot expressar de la següent manera:

$$0 < \frac{\mathrm{d}\, cons}{\mathrm{d}\, renta} < 1 \tag{1-2}$$

Proposició 3: La proporció de la renda dedicada al consum és menor quan augmenta la renda. És a dir, la proporció de l'últim euroguanyat destinat al consum és més xicoteta que la proporció de la renda total destinada al consum.

Analíticament, la proposició 3 es pot expressar de la següent manera:

$$\frac{\mathrm{d}con}{\mathrm{d}\, renda} < \frac{cons}{renda} \tag{1-3}$$

En altres paraules, la propensió marginal al consum és menor que la propensió mitjana al consum.

Aquestes tres proposicions constitueixen un model econòmic: la funció de consum keynesiana. Per estimar i contrastar aquest model hem de convertir-lo en un model economètric. Per aquesta conversió s'han de complir dos requisits.

D'acord amb el primer requisit, cal especificar la forma matemàtica de la funció. En aquest cas s'ha utilitzat la funció lineal, pel fet que, a més de ser simple, és compatible amb la descripció feta per Keynes.

Per tal de complir amb la segona exigència, cal tenir en compte que el model formulat en la proposició 1 és determinista. És a dir, la renda és l'únic factor que es té en compte per a la determinació del consum. Però en la vida real hi ha molts altres factors, diferents de la renda, que tenen una influència en el consum. En un model economètric tots els factors diferents de les variables independents incloses es reuneixen en una variable anomenada pertorbació aleatòria o error (u). Per tant, el segon requisit és la introducció del terme d'error en l'equació.

En general, tots els factors rellevants han de ser introduïts de manera explícita en el model economètric, i la resta dels factors s'agrupen en una única variable: l'error o pertorbació aleatòria. En la funció de consum keynesiana l'únic factor considerat rellevant és la renda.

Tenint en compte aquests dos requisits la funció de consum keynesiana es pot expressar de la següent manera:

$$cons = \beta_1 + \beta_2 renda + u \tag{1-4}$$

Aquest és un model economètric que pot estimar-se si es disposa de dades sobre consum i renda. Vegem ara les altres dues proposicions. En aquest model lineal la propensió marginal al consum és la següent:

$$\frac{\mathrm{d}\, cons}{\mathrm{d}\, renda} = \beta_2 \tag{1-5}$$

En conseqüència, la proposició 2 en aquest model és la següent:

$$0 < \beta_2 < 1 \tag{1-6}$$

Una vegada que el model s'ha estimat, és possible comprovar si l'estimació de β_2 es troba entre 0 i l.

En el model lineal la propensió mitjana al consum, considerant que l'error és igual a 0, és la següent:

$$\frac{cons}{renta} = \frac{\beta_1 + \beta_2 renta}{renta} = \frac{\beta_1}{renta} + \beta_2 \tag{1-7}$$

Per tant, la proposició 3 implica que

$$\frac{\beta_1}{renta} + \beta_2 > \beta_2 \quad \text{or} \quad \frac{\beta_1}{renta} > 0 \tag{1-8}$$

És a dir,

$$\beta_1 > 0 \tag{1-9}$$

Una vegada que el model s'ha estimat, contrastar la proposició 3 és equivalent a contrastar si el terme independent és significativament més gran que 0.

EXEMPLE 1.2 Determinació dels salaris

Model econòmic:

La teoria econòmica formal - la teoria del capital humà- diu que l'educació (*educ*), l'experiència (*exper*) i l'aprenentatge (*aprend*) són factors que afecten la productivitat i per tant al **salari**. Aleshores, un model econòmic per a explicar el salari podria ser el següent:

$$salari = f(educ, exper, aprend) \tag{1-10}$$

Per cert, en la seua opinió, creu. que falta alguna variable en aquest model?,

Modelo economètric:

El model economètric, que correspon utilitzant una forma lineal matemàtica, és el següent:

$$salari = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 aprend + u \tag{1-11}$$

En resum, per a convertir un model econòmic en un model economètric:

a) S'ha especificat la forma de la funció f (.).

b) S'ha inclòs en el model una pertorbació aleatòria que recull l'efecte conjunt d'altres variables que també afecten els salaris, però que no figuren en el model.

Un element important en l'especificació del model és la formulació d'un conjunt de supòsits estadístics, que s'utilitzen en les etapes següents. Aquests supòsits estadístics juguen un paper clau en el contrast d'hipòtesis i, en general, en tot el procés d'inferència dut a terme amb el model.

(b) Estimació

En l'estimació s'obtenen els valors numèrics dels coeficients d'un model economètric. Per completar aquesta etapa s'ha de disposar d'un conjunt d'observacions de totes les variables observables que apareixen en el model economètric especificat, i,

d'altra banda, cal seleccionar el mètode d'estimació apropiat, tenint en compte les implicacions d'aquesta elecció en les propietats estadístiques dels estimadors dels coeficients. La distinció entre un estimador i una estimació ha de quedar clara. Un estimador és el resultat d'aplicar un mètode d'estimació a una especificació economètrica. D'altra banda, una estimació consisteix en l'obtenció d'un valor numèric d'un estimador per a una mostra donada. Per exemple, l'aplicació del mètode de mínims quadrats a l'especificació de la funció de consum (1-4) proporciona expressions que determinen els estimadors $\hat{\beta}_1$ i $\hat{\beta}_2$. Substituint les dades mostrals en aquestes expressions, s'obtenen dos valors: un valor per $\hat{\beta}_1$ i un altre per $\hat{\beta}_2$, que són les estimacions dels paràmetres β_1 i β_2.

En general, és possible obtenir expressions analítiques dels estimadors, particularment en el cas de l'estimació de relacions lineals. No obstant això, en els procediments d'estimació no lineal és sovint difícil establir la seva expressió analítica.

(c) Validació

En l'etapa de validació s'avaluen els resultats. En aquesta etapa s'avalua si les estimacions obtingudes en l'etapa anterior són acceptables, tant per la teoria econòmica com des del punt de vista estadístic. S'analitza, d'una banda, si les estimacions dels paràmetres del model tenen els signes i magnituds esperats, és a dir, si satisfan les limitacions establertes per la teoria econòmica.

Des del punt de vista estadístic, d'altra banda, es duen a terme contrastos estadístics sobre la significativitat dels paràmetres del model en què s'utilitzen els supòsits estadístics formulats en l'etapa d'especificació. Al seu torn, és important contrastar si els supòsits estadístics del model economètric es compleixen, encara que cal tenir en compte que no tots els supòsits són contrastables. La violació d'algun d'aquests supòsits implica, en general, l'aplicació d'altres mètodes d'estimació, que permeten obtenir estimadors amb les millors propietats estadístiques possibles.

Una manera d'establir si el model és adequat per fer prediccions és utilitzar el model fora del període mostral, i després comparar els valors predits de la variable endògena amb els valors realment observats.

1.3 Dades econòmiques

Com hem vist, l'anàlisi empírica utilitza dades per contrastar una teoria o per estimar una relació. És important destacar que en Econometria utilitzem dades no experimentals. Les dades no experimentals es recullen mitjançant l'observació del món real d'una manera passiva. En aquest cas les dades no són el resultat d'experiments controlats. Les dades experimentals es recullen sovint en entorns de laboratori, com passa en les ciències naturals.

Ara, anem a veure tres tipus de dades que es poden utilitzar en l'estimació d'un model economètric: sèries temporals, dades de tall transversal i dades panell.

Sèries temporals

En les sèries temporals, les dades són observacions d'una variable al llarg del temps. Per exemple: magnituds dels comptes nacionals, com el consum, les importacions, ingressos, etc. L'ordre cronològic de les observacions proporciona informació potencialment important. En conseqüència, en una sèrie temporal l'ordenació de les observacions és rellevant.

No es pot assumir que les dades de sèries temporals siguen independents a través del temps. La majoria de les sèries econòmiques es relacionen amb les seves històries recents. Exemples típics són els agregats macroeconòmics com els preus i els tipus d'interès. Aquest tipus de dades es caracteritza per la dependència serial, de manera que el supòsit de mostreig aleatori resulta inadequat en aquest cas.

La majoria de les dades econòmiques agregades només estan disponibles per a freqüències baixes (anual, trimestral o mensual en algunes ocasions) per la qual cosa la mida de la mostra sol ser molt menor que en els típics estudis de tall transversal. L'excepció són les dades financeres, on es disposa de dades per a freqüències més elevades (setmanal, diària, per hora, etc.) de manera que les mides mostrals poden ser molt grans.

Dades de tall transversal

En les dades de tall transversal es disposa d'una observació per individu i es refereixen a un punt determinat en el temps. En la majoria dels estudis, els individus enquestats són persones (per exemple, en l'Enquesta de Població Activa (EPA), més de 100000 persones són entrevistades cada trimestre), llars (per exemple, l'Enquesta de Pressupostos Familiars), empreses (per exemple, l'Enquesta d'Empreses Industrials) o altres agents econòmics. Les enquestes són una font típica per dades de tall transversal. En molts estudis economètrics contemporanis de tall transversal la mida mostral és força elevat.

En les dades de tall transversal, les observacions han de ser obtingudes mitjançant un mostreig aleatori, el que implica que les observacions siguen independents entre si. L'ordre de les observacions en les dades de tall transversal no importa per a l'anàlisi economètric. Si les dades no s'obtenen amb una mostra aleatòria, tindrem un problema de selecció mostral.

Fins ara ens hem referit a dades de tipus de micro, però també es poden tenir dades de tall transversal relatius a unitats agregades, com països, regions, etc. Per descomptat, les dades d'aquest tipus no s'obtenen mitjançant un mostreig aleatori.

Dades de panell

Les dades de panell (o dades longitudinals) consisteixen en observacions de tall transversal repetides al llarg del temps. Així doncs, les dades panell combinen elements de dades de tall transversal i de sèries temporals. Aquests conjunts de dades consisteixen en un conjunt d'individus (en general persones, llars o empreses) enquestats repetidament al llarg del temps. En la modelització s'adopta generalment el cas que els individus són independents entre si, però que, per a un individu donat, les observacions al llarg del temps són mútuament dependents. Per tant, l'ordre dins d'un tall transversal d'un conjunt de dades panell no importa, però l'ordre en la dimensió temporal és rellevant. Si no tenim en compte el temps en dades de panell, es diu que estem utilitzant dades de tall transversal agrupats (pooled).

2 EL MODEL DE REGRESSIÓ LINEAL SIMPLE: ESTIMACIÓ I PROPIETATS

2.1 Algunes definicions en el model de regressió simple

2.1.1 El model de regressió poblacional i la funció de regressió poblacional

En el model de regressió simple, el *model de regressió poblacional* o, simplement, el *model poblacional* és el següent:

$$y = \beta_1 + \beta_2 x + u \tag{2-1}$$

Anem a veure els diferents elements del model (2-1) i la terminologia utilitzada per designar-los. En primer lloc, en el model hi ha tres tipus de variables: *y*, *x* i *u*. En aquest model l'únic factor explícit per explicar *y* és *x*. La resta dels factors que afecten *y* estan recollits en *u*.

Anomenem a *y* variable endògena (del grec: *generada dins*) o variable dependent. S'utilitzen també altres denominacions per designar *y*: variable explicada o regressant. En aquest model totes aquestes denominacions són equivalents, però en altres models, com veurem més endavant, pot haver-hi algunes diferències.

En la regressió lineal simple de *y* sobre *x*, a la variable *x* se li denomina variable exògena (del grec: *generat fora de*) o variable independent. Altres denominacions utilitzades també per designar *x* són: variable explicativa, regressor, covariable o variable de control. Totes aquestes denominacions són equivalents, però en altres models, com veurem més endavant, pot haver-hi algunes diferències.

La variable *u* recull tots aquells factors diferents de *x* que afecten *y*. És anomenada error o pertorbació aleatòria. El terme de pertorbació pot captar també l'error de mesurament de la variable dependent. La pertorbació és una variable no observable.

Els paràmetres β_1 i β_2 són fixos i desconeguts.

En el segon membre de (2-1) es poden distingir dos components: un component sistemàtic $\beta_1 + \beta_2 x$ i la pertorbació aleatòria *u*. Anomenant μ_y al component sistemàtic, podem escriure:

$$\mu_y = \beta_1 + \beta_2 x \tag{2-2}$$

Aquesta equació és coneguda com la funció de regressió poblacional (FRP) o recta poblacional. Per tant, com es pot veure a la figura 2.1, μ_y és una funció lineal de *x* amb terme independent igual a β_1 i pendent igual a β_2.

La linealitat significa que un augment d'una unitat en x implica que el *valor esperat* de *y* - *y* - $\mu_y = E(y)$ - varie en β_2 unitats.

Ara, suposem que disposem d'una mostra aleatòria de grandària *n* $\{(y_i, x_i)$: $i = 1, ...,n\}$ extreta de la població estudiada. Al diagrama de dispersió de la figura 2.2, es mostren els hipotètics valors de la mostra.

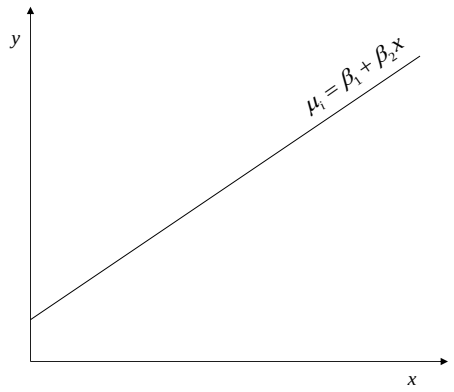


FIGURA 2.1. La funció de regressió poblacional. (*FRP*)

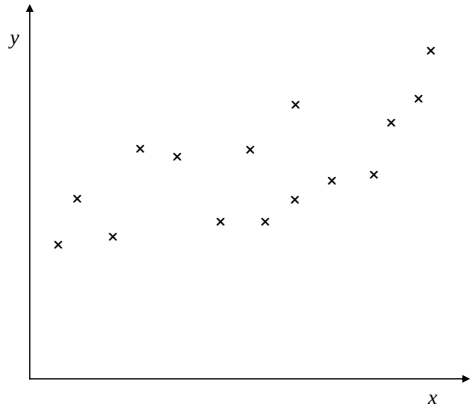


FIGURA 2.2. Diagrama de dispersió.

El model poblacional per a cada observació de la mostra es pot expressar de la següent manera:

$$y_i = \beta_1 + \beta_2 x_i + u_i \quad i = 1, 2, \cdots, n \tag{2-3}$$

A la figura 2.3 s'ha representat conjuntament la funció de regressió poblacional i el diagrama de dispersió, però és important no oblidar que β_1 i β_2 són fixos, però desconeguts. D'acord amb aquest model és possible, des d'un punt de vista teòric, fer la següent descomposició:

$$y_i = \mu_{yi} + u_i \quad i = 1, 2, \cdots, n \tag{2-4}$$

que ha estat representada a la figura 2.3 per a l'observació i-èsima. No obstant això, des d'un punt de vista empíric, no és possible fer-ho a causa que β_1 i β_2 són desconeguts i, conseqüentment, u_i és no observable.

2.1.2 La funció de regressió mostral

L'objectiu principal del *model de regressió* és la determinació o estimació de β_1 i β_2 a partir d'una mostra donada.

La *funció de regressió mostral* (*FRM*) és la contrapartida de la funció de regressió poblacional (*FRP*). Atès que la *FRM* s'obté per a una mostra donada, una nova mostra generarà una altra estimació diferent.

La *FRM*, que és una estimació de la *FRP*, ve donada per

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i \tag{2-5}$$

i permet calcular el valor ajustat ($\hat{y}_i$) per y tant $x = x_i$. A la *FRM* $\hat{\beta}_1$ i $\hat{\beta}_2$ són els estimadors dels paràmetres β_1 i β_2. Per a cada x_i tenim un valor observat (y_i) i un valor ajustat ($\hat{y}_i$).

A la diferència entre y_i i $\hat{y}_i$ se li denomina residu $\hat{u}_i$:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \tag{2-6}$$

En altres paraules, el residu $\hat{u}_i$ és la diferència entre el valor mostral y_i i el valor ajustat de $\hat{y}_i$, segons es pot veure a la figura 2.4. En aquest cas sí que és possible calcular empíricament la descomposició per a una mostra donada:

$$y_i = \hat{y}_i + \hat{u}_i$$

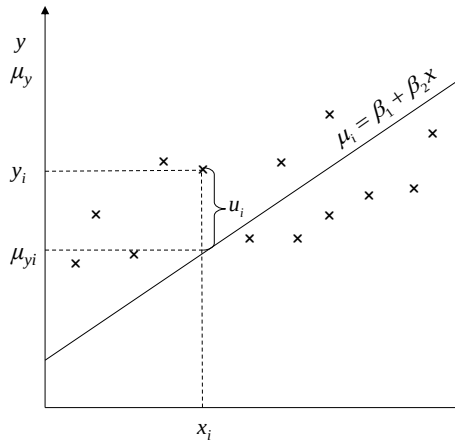


FIGURA 2.3. La funció de regressió poblacional i el diagrama de dispersió.

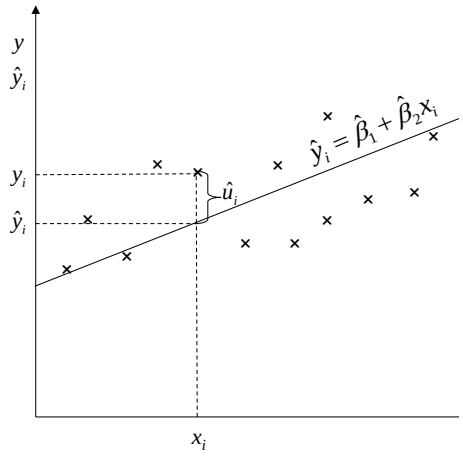


FIGURA 2.4. La funció de regressió mostral i el diagrama de dispersió.

Resumint $\hat{\beta}_1, \hat{\beta}_2, \hat{y}_i$ i $\hat{u}_i$ són la contrapartida mostral de β_1, β_2, μ_{yi} i u_i respectivament. És possible calcular $\hat{\beta}_1$ i $\hat{\beta}_2$, per a una mostra donada, però per a cada mostra les estimacions seran diferents. Per contra, β_1 i β_2 són fixos però desconeguts.

2.2 Obtenció de les estimacions per Mínims Quadrats Ordinaris (*MQO*)

2.2.1 Diferents criteris d'estimació

Abans d'obtenir les estimacions per mínims quadrats, anem a examinar tres mètodes alternatius per il·lustrar el problema que tenim plantejat. Aquests tres mètodes tenen en comú que tracten de minimitzar, d'alguna manera, el valor dels residus en el seu conjunt.

Criteri 1

Un primer criteri consistiria a prendre com estimadors $\hat{\beta}_1$ i $\hat{\beta}_2$ a aquells valors que facin la suma de tots els residus tan propera a zero com siga possible. Amb aquest criteri l'expressió a minimitzar seria la següent:

$$\text{Min}\left|\sum_{i=1}^{n}\hat{u}_i\right| \tag{2-7}$$

El problema principal d'aquest mètode d'estimació és que els residus de diferent signe poden compensar-se. Aquesta situació es pot observar gràficament a la figura 2.5, en la qual es representen tres observacions alineades (x_i,y_i), (x_2, y_2) i (x_3, y_3). En aquest cas, passa el següent

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y_3 - y_1}{x_3 - x_1}$$

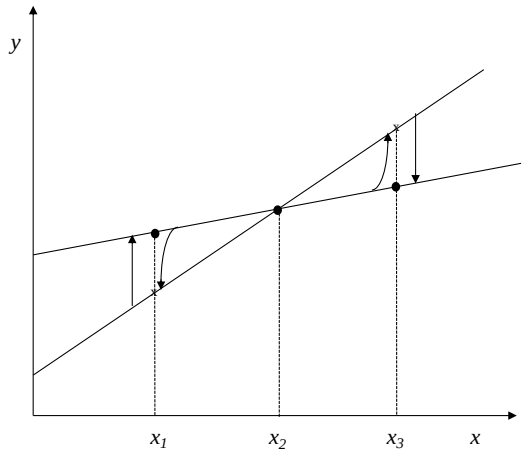


FIGURA 2.5. Els problemes del criteri 1.

Si una línia recta s'ajusta de manera que passe a través dels tres punts, cadascun dels residus prendrà el valor zero, de manera que

$$\left|\sum_{i=1}^{3}\hat{u}_i = 0\right|$$

Aquest ajust podria ser considerat òptim. Però també és possible obtenir $\left|\sum_{i=1}^{3}\hat{u}_i = 0\right|$, mitjançant la rotació de la línia recta - des del punt x_2, y_2 - en qualsevol direcció, com mostra la figura 2.5, perquè $\hat{u}_3 = -\hat{u}_1$. En altres paraules, fent girar d'aquesta manera la recta, s'obté sempre el resultat que $\left|\sum_{i=1}^{3}\hat{u}_i = 0\right|$. Aquest simple exemple mostra que aquest criteri no és adequat per a l'estimació dels paràmetres, ja que, per a qualsevol conjunt d'observacions, hi ha un nombre infinit de línies rectes que satisfan aquest criteri.

Criteri 2

Per tal d'evitar la compensació dels residus positius amb els negatius, d'acord amb aquest criteri es prenen els valors absoluts dels residus. En aquest cas es minimitzaria la següent expressió:

$$\text{Min} \sum_{i=1}^{n} |\hat{u}_i| \tag{2-8}$$

Malauradament, encara que els estimadors així obtinguts tenen algunes propietats interessants, el seu càlcul és complicat, requerint la resolució d'un problema de programació lineal o l'aplicació d'un procediment de càlcul iteratiu.

Criteri 3

Un tercer mètode consisteix a minimitzar la suma dels quadrats dels residus, és a dir,

$$\text{Min } S = \text{Min} \sum_{i=1}^{n} \hat{u}_i^2 \tag{2-9}$$

Els estimadors obtinguts s'anomenen estimadors de mínims quadrats (*MQ*), i gaudeixen de certes propietats estadístiques desitjables, que s'estudiaran més endavant. D'altra banda, davant del primer dels criteris examinats, en prendre els quadrats dels residus s'evita que es compensin, mentre que, a diferència del segon dels criteris, els estimadors de mínims quadrats són senzills d'obtenir. És important assenyalar que, des del moment en què prenem els quadrats dels residus, estem penalitzant més que proporcionalment als residus grans enfront dels xicotets (si un residu és el doble d'un altre, el seu quadrat serà quatre vegades més gran). Això caracteritza l'estimació de mínims quadrats respecte a altres procediments possibles.

2.2.2 Aplicació del criteri de mínim quadrats

A continuació, s'exposa el procés d'obtenció dels estimadors de *MQ*. L'objectiu és minimitzar la suma dels quadrats dels residus (*S*). Per això, en primer lloc expressem *S* com una funció dels estimadors, utilitzant (2-6):

Per tant

$$\underset{\hat{\beta}_1,\hat{\beta}_2}{\text{Min}}\, S = \underset{\hat{\beta}_1,\hat{\beta}_2}{\text{Min}} \sum_{t=1}^{T} \hat{u}_t^2 = \underset{\hat{\beta}_1,\hat{\beta}_2}{\text{Min}} \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \tag{2-10}$$

Per minimitzar *S*, derivem parcialment pel que fa a $\hat{\beta}_1$ i $\hat{\beta}_2$:

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) \tag{2-11}$$

$$\frac{\partial S}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) x_i \tag{2-12}$$

Els estimadors de *MQ* s'obtenen igualant les anteriors derivades a zero:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

Les equacions (2-11) i (2-12) es denominen *equacions normals* o *condicions de primer ordre* de *MQ*.

En les operacions amb sumatoris s'han de tenir en compte les següents regles: $\sum_{i=1}^{n} a = na$

$$\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Operant amb les equacions normals, s'obté que

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} x_i \qquad (2\text{-}13)$$

$$\sum_{i=1}^{n} y_i x_i = \hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 \qquad (2\text{-}14)$$

Dividint els dos membres de (2-13) per *n*, s'obté que

$$\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x} \qquad (2\text{-}15)$$

Per tant,

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \qquad (2\text{-}16)$$

Substituint aquest valor de $\hat{\beta}_1$ a la segona equació normal (2-14), s'obté que

$$\sum_{i=1}^{n} y_i x_i = (\bar{y} - \hat{\beta}_2 \bar{x})\sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} y_i x_i = \bar{y}\sum_{i=1}^{n} x_i - \hat{\beta}_2 \bar{x}\sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2$$

Resolent pe r $\hat{\beta}_2$ s'obté que:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i} \qquad (2\text{-}17)$$

O, com es pot veure en l'apèndix 2.1,

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2\text{-}18)$$

Si dividim numerador i denominador de (2-18) per *n*, es pot veure que $\hat{\beta}_2$ és el quocient entre la covariança de les dues variables i la variança de x. Per tant, el signe de $\hat{\beta}_2$ és el mateix que el signe de la covariança.

Un cop calculat $\hat{\beta}_2$, es pot obtenir $\hat{\beta}_1$ utilitzant l'equació (2-16).

Aquests són els estimadors de *MQ*. Atès que hi ha mètodes més complexos, que també es denominen de *MQ*, al mètode que acabem de desenvolupar li anomenarem mètode de mínims quadrats ordinaris (*MQO*), per la seua simplicitat.

En els epígrafs precedents, $\hat{\beta}_1$ i $\hat{\beta}_2$ s'han utilitzat per designar estimadors genèrics. A partir d'ara amb aquesta notació només designarem als estimadors *MQO*.

EXEMPLE 2.1 L'estimació de la funció de consum

Donada la funció de consum keynesiana,

$$cons = \beta_1 + \beta_2 renta + u_i$$

anem a estimar-la utilitzant les dades de 6 llars que apareixen en el quadre 2.1.

QUADRE 2.1. Dades i càlculs per estimar la funció de consum.

Observ	$cons_i$	$renda_i$	$cons_i \times renda_i$	$renda_i^2$	$cons_i - \overline{cons}$	$renda_i - \overline{renda}$	$(cons_i - \overline{cons}) \times (renda_i - \overline{renda})$	$(renda_i - \overline{renda})^2$
1	5	6	30	36	-4	-5	20	25
2	7	9	63	81	-2	-2	4	4
3	8	10	80	100	-1	-1	1	1
4	10	12	120	144	1	1	1	1
5	11	13	143	169	2	2	4	4
6	13	16	208	256	4	5	20	25
Suma	**54**	**66**	**644**	**786**	**0**	**0**	**50**	**60**

Calculant $\overline{cons}$ i $\overline{renda}$, i aplicant la fórmula (2-17), o alternativament (2-18), a les dades de la taula 2.1, obtenim:

$$\overline{cons} = \frac{54}{6} = 9\,;\ \overline{renda} = \frac{66}{6} = 11\,;\quad (2\text{-}17)\colon \hat{\beta}_2 = \frac{644 - 9 \times 66}{786 - 11 \times 66} = 0.8\hat{3}\,;\quad (2\text{-}18)\colon \hat{\beta}_2 = \frac{50}{60} = 0.8\hat{3}$$

Aplicant després (2-16), obtenim que $\hat{\beta}_1 = 9 - 0.8\hat{3} \times 11 = -0.1\hat{6}$

2.3 Algunes característiques dels estimadors de *MQO*

2.3.1 Algunes característiques dels estimadors de *MQO*

Les implicacions algebraiques de l'estimació són derivades exclusivament de l'aplicació del procediment de *MQO* al model de regressió lineal simple:

1. *La suma dels residus de MQO és igual a 0*:

$$\sum_{i=1}^{n} \hat{u}_i = 0 \qquad (2\text{-}19)$$

De la definició dels residus:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \qquad i = 1, 2, \cdots, n \qquad (2\text{-}20)$$

Si sumem per a les *n* observacions, s'obté:

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \qquad (2\text{-}21)$$

que és precisament la primera equació (2-11) del sistema d'equacions normals.

Cal observar que, si (2-21) es compleix, això implica que

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \qquad (2\text{-}22)$$

i, dividint (2-19) i (2-22) per *n*, s'obté

$$\bar{\hat{u}} = 0 \qquad \bar{y} = \bar{\hat{y}} \qquad (2\text{-}23)$$

2. La recta de regressió de MQO passa necessàriament pel punt ($\bar{x}, \bar{y}$).

Efectivament, dividint l'equació (2-13) per *n*, s'obté:

$$\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x} \qquad (2\text{-}24)$$

3. El producte creuat mostral entre cada un dels regressors i els residus de MQO és zero.

És a dir,

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0 \qquad (2\text{-}25)$$

Es pot veure que (2-25) és igual a la segona equació normal donada en (2-14):

$$\sum_{i=1}^{n} x_i \hat{u}_i = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

4. El producte creuat mostral entre els valors ajustats ($\hat{y}$) i els residus de MQO és igual a zero.

És a dir,

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0 \qquad (2\text{-}26)$$

Demostració

En efecte, tenint en compte les implicacions algebraiques 1 - (2-19) - i 3 - (2-25) -, s'obté que

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = \sum_{i=1}^{n} (\hat{\beta}_1 + \hat{\beta}_2 x_i) \hat{u}_i = \hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i \hat{u}_i = \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 0 = 0$$

2.3.2 Descomposició de la variança de *y*

Per definició

$$y_i = \hat{y}_i + \hat{u}_i \tag{2-27}$$

El producte creuat mostral entre cada un dels regressors i els residus de *MQO* és zero

$$y_i - \bar{y} = \hat{y}_i - \bar{\hat{y}} + \hat{u}_i$$

Elevant al quadrat els dos membres:

$$\left[y_i - \bar{y}\right]^2 = \left[(\hat{y}_i - \bar{\hat{y}}) + \hat{u}_i\right]^2 = (\hat{y}_i - \bar{\hat{y}})^2 + \hat{u}_i^2 + 2\hat{u}_i(\hat{y}_i - \bar{\hat{y}})$$

Sumant per a tot *i*:

$$\sum\left[y_i - \bar{y}\right]^2 = \sum(\hat{y}_i - \bar{\hat{y}})^2 + \sum\hat{u}_i^2 + 2\sum\hat{u}_i(\hat{y}_i - \bar{\hat{y}})$$

Tenint en compte les propietats algebraiques 1 i 4, el tercer terme del segon membre és igual a 0. Analíticament,

$$\sum\hat{u}_i(\hat{y}_i - \bar{\hat{y}}) = \sum\hat{u}_i\hat{y}_i - \bar{\hat{y}}\sum\hat{u}_i = 0 \tag{2-28}$$

Per tant, obtenim

$$\sum\left[y_i - \bar{y}\right]^2 = \sum(\hat{y}_i - \bar{\hat{y}})^2 + \sum\hat{u}_i^2 \tag{2-29}$$

En paraules,

Suma de quadrats totals (*SQT*) =

Suma de quadrats explicats (*SQE*) + Suma dels quadrats dels residus (*SQR*)

S'ha de complir la relació (2-15) per assegurar que (2-28) és igual a 0. Cal recordar que (2-15) està associada a la primera equació normal, és a dir, a l'equació corresponent al terme independent. Si en el model ajustat no hi ha terme independent, aleshores, en general, no es complirà la descomposició obtinguda en (2-29).

Aquesta descomposició es pot aplicar a les variances, dividint els dos membres de (2-29) per *n*:

$$\frac{\sum\left(y_i - \bar{y}\right)^2}{n} = \frac{\sum(\hat{y}_i - \bar{\hat{y}})^2}{n} + \frac{\sum\hat{u}_i^2}{n} \tag{2-30}$$

En paraules,

Variança total =variança explicada+variança residual

2.3.3 Bondat de l'ajust: Coeficient de determinació (R^2)

A priori, s'han obtingut uns estimadors que minimitzen la suma dels quadrats dels residus.

Ara, un cop feta l'estimació, podrem veure en quina mesura la recta de regressió mostral s'ajusta a les dades.

Una mesura que indique el grau d'ajust de la recta de regressió mostral amb les dades s'anomena mesura de *bondat de l'ajust*. Estudiarem ara la mesura més coneguda: el *coeficient de determinació* o R quadrat (R^2). Aquesta mesura es defineix de la següent manera:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2-31}$$

Per tant, R^2 és la proporció de la suma de quadrats totals (*SQT*), que s'explica per la regressió (*SQE*), és a dir, que s'explica pel model. També podem dir que 100 R^2 és el percentatge de variació mostral de *y* explicada per *x*.

Alternativament, tenint en compte (2-29), obtenim:

$$\sum (\hat{y}_i - \bar{\hat{y}})^2 = \sum (y_i - \bar{y})^2 - \sum \hat{u}_i^2$$

Substituint en (2-31), tenim

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\sum (y_i - \bar{y})^2 - \sum \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{SCR}{SCT} \tag{2-32}$$

Per tant, R^2 és igual a 1 menys la proporció de la suma de quadrats totals (*SQT*), que no és explicada per la regressió (*SQR*).

D'acord amb la definició de R^2, s'ha de complir que

$$0 \le R^2 \le 1$$

Casos extrems:

a) Si l'ajust és perfecte, aleshores es verificarà $\hat{u}_i = 0 \quad \forall i$. Això implica que

$$\hat{y}_i = y_i \quad \forall i \quad \Rightarrow \sum (\hat{y}_i - \bar{\hat{y}})^2 = \sum (y_i - \bar{y})^2 \quad \Rightarrow R^2 = 1$$

b) Si $\hat{y}_i = c \quad \forall i$, això implica que

$$\bar{\hat{y}} = c \quad \Rightarrow \hat{y}_i - \bar{\hat{y}} = c - c = 0 \quad \forall i \quad \Rightarrow \sum (\hat{y}_i - \bar{\hat{y}})^2 = 0 \quad \Rightarrow R^2 = 0$$

Si R^2 està proper a zero, això implica que l'ajust no és bo. En altres paraules, hi ha molt poca variació de *y* que siga explicada per *x*.

En molts casos, s'obté un R^2 elevat quan s'ajusta un model utilitzant dades de sèries temporals, a causa de l'efecte d'una tendència comuna. Per contra, quan fem servir dades de tall transversal és freqüent obtenir valors baixos, però això no vol dir que el model ajustat siga incorrecte.

Quina és la relació entre el coeficient de determinació i el coeficient de correlació estudiats en estadística descriptiva? El coeficient de determinació és igual al coeficient de correlació al quadrat, com es pot veure en l'apèndix 2.2:

$$r_{xy}^2 = R^2 \tag{2-33}$$

(Aquesta igualtat és vàlida en el model de regressió lineal simple, però no en el model de regressió lineal múltiple)

EXEMPLE 2.2 Compliment de les propietats algebraiques i R^2 en la funció de consum

A la columna 2 del quadre 2.1, es calcula $cons_i$ en les columnes 3, 4 i 5, es pot veure el compliment de les implicacions algebraiques 1, 3 i 4, respectivament. A la resta de les columnes es realitzen càlculs per tal d'obtenir

$$SCT = 42 \quad SCE = 41.67 \quad SCR = 42 - 41.67 = 0.33 \quad R^2 = \frac{41.67}{42} = 0.992$$

o, alternativament, $R^2 = 1 - \frac{0.33}{42} = 0.992$

QUADRE 2.2. Dades i càlculs per estimar la funció de consum.

Observ.	$cons_i$	$\hat{u}_i$	$\hat{u}_i \times renda_i$	$cons_i \times \hat{u}_i$	$cons_i^2$	$(cons_i - \overline{cons})^2$	$cons_i^2$	$(cons_i - \overline{cons})^2$
1	4.83	0.17	1.00	0.81	25	16	23.36	17.36
2	7.33	-0.33	-3.00	-2.44	49	4	53.78	2.78
3	8.17	-0.17	-1.67	-1.36	64	1	66.69	0.69
4	9.83	0.17	2.00	1.64	100	1	96.69	0.69
5	10.67	0.33	4.33	3.56	121	4	113.78	2.78
6	13.17	-0.17	-2.67	-2.19	169	16	173.36	17.36
	54.00	**0.00**	**0.00**	**0.00**	**528**	**42**	**527.67**	**41.67**

2.3.4 Regressió a través de l'origen

Si forcem que la línia de regressió passe pel punt (0,0) estem imposant la restricció que el terme independent siga zero, com es pot veure a la figura 2.6. A aquesta regressió se li denomina regressió a través de l'origen.

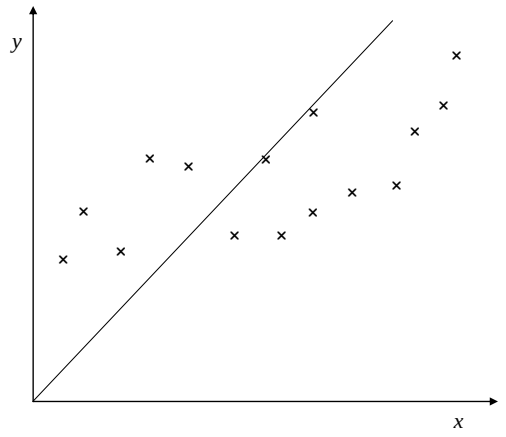


FIGURA 2.6. Una regressió a través de l'origen.

Ara, anem a estimar una recta de regressió a través de l'origen. El model ajustat és el següent:

$$\tilde{y}_i = \tilde{\beta}_2 x_i \tag{2-34}$$

Per tant, hem de minimitzar

$$\underset{\tilde{\beta}_2}{\text{Min}}\, S = \underset{\tilde{\beta}_2}{\text{Min}} \sum_{i=1}^{n} (y_i - \tilde{\beta}_2 x_i)^2 \tag{2-35}$$

Per minimitzar *S*, derivem respecte a $\tilde{\beta}_2$ e igualarem a 0:

$$\frac{dS}{d\tilde{\beta}_2} = -2\sum_{i=1}^{n} (y_i - \tilde{\beta}_2 x_i) x_i = 0 \tag{2-36}$$

Resolent per $\tilde{\beta}_2$

$$\tilde{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \tag{2-37}$$

Un altre problema que es planteja en ajustar una recta de regressió perquè passe per l'origen és que succeeix en general que:

$$\sum (y_i - \bar{y})^2 \neq \sum (\hat{y}_i - \bar{\hat{y}})^2 + \sum \hat{u}_i^2$$

Si no és possible la descomposició de la variança de *y* en dos components (explicada i residual), aleshores R^2 no té sentit. Aquest coeficient pot prendre valors negatius o superiors a 1 en el model sense terme independent.

Resumint, s'ha d'incloure sempre un terme independent en les regressions, llevat que hi hagi fortes raons en contra sustentades per la teoria econòmica.

2.4 Les unitats de mesura i la forma funcional

2.4.1 Unitats de mesura

Canvi d'unitats de mesura (canvi d'escala) en x

Si *x* és multiplicada/dividida per una constant $c \neq 0$, aleshores el pendent de *MQO* queda dividit/multiplicat per la mateixa constant, *c*. Així

$$\hat{y}_i = \hat{\beta}_1 + \left[\frac{\hat{\beta}_2}{c}\right](x_i \times c) \tag{2-38}$$

EXEMPLE 2.3 Suposem la següent funció del consum estimat, en què les dues variables es mesuren en milers d'euros:

$$cons_i = 0.2 + 0.85 \times renda_i \tag{2-39}$$

Si ara s'expressen la renda en euros (multiplicant per 1000) i es designa per *rendae*, el model ajustat a les noves unitats de mesura de la renda serà el següent:

$$cons_i = 0.2 + 0.00085 \times rendae_i$$

Com es pot veure, el canvi de les unitats de mesura de la variable explicativa no afecta el terme independent.

Canvi d'unitats de mesura (canvi d'escala) a **y**

Si *y* és multiplicada/dividida per una constant $c \neq 0$, aleshores el pendent i el terme independent calculats per *MQO* es multipliquen/divideixen per la mateixa constant, *c*. Així,

$$(\hat{y}_i \times c) = (\hat{\beta}_1 \times c) + (\hat{\beta}_2 \times c) x_i \tag{2-40}$$

EXEMPLE 2.4 Si expressem, en el model (2-39), el consum en euros (multiplicant per 1000) i en diem *conse*, el model ajustat a les noves unitats de mesura del consum serà el següent:

$$conse_i = 200 + 850 \times renda_i$$

Canvi de l'origen

Si se suma/resta una constant *d* a *x* i/o *y*, aleshores el pendent *MQO* no es veu afectada. No obstant això, si es canvia l'origen de *x* i/o *y* el terme independent de la regressió si es veu afectat.

Si es resta una constant *d* a *x*, el terme independent canvia de la següent manera:

$$\hat{y}_i = (\hat{\beta}_1 + \hat{\beta}_2 \times d) + \hat{\beta}_2 (x_i - d) \tag{2-41}$$

Si es resta una constant *d* a *y*, el terme independent canvia de la següent manera:

$$\hat{y}_i - d = (\hat{\beta}_1 - d) + \hat{\beta}_2 x_i \tag{2-42}$$

EXEMPLE 2.5 Suposem que la renda mitjana és de 20 mil euros. Si definim la variable $rendad_i = renda_i - \overline{renda}$ i les dues variables es mesuren en milers d'euros, el model ajustat amb aquest canvi en l'origen serà el següent:

$$cons_i = (0.2 + 0.85 \times 20) + 0.85 \times (renda_i - 20) = 17.2 + 0.85 \times rendad_i$$

EXEMPLE 2.6

Suposem que el consum mitjà és de 15 mil euros. Si definim la variable $consd_i = cons_i - \overline{cons}$ i mesurem les dues variables en euros, el model ajustat amb el canvi en l'origen serà el següent:

$$cons_i - 15 = 0.2 - 15 + 0.85 \times renda_i$$

És a dir,

$$consd_i = -14.8 + 0.85 \times renda_i$$

Cal observar que R^2 no varia en realitzar canvis d'unitats de *x* i/o *y*, i tampoc varia quan es canvia l'origen de les variables.

2.4.2 Forma funcional

En molts casos les relacions lineals no són adequades en les aplicacions econòmiques. No obstant això, en el model de regressió simple podem incorporar no

linealitats (en les variables) redefinint de manera apropiada la variable dependent i la variable independent.

Algunes definicions

Estudiarem ara algunes definicions de les mesures de variació que seran útils en la interpretació dels coeficients de diferents formes funcionals. En concret, estudiarem les següents mesures: canvi proporcional i canvi en logaritmes.

El *canvi proporcional* (o taxa de variació relativa) entre x_1 i x_0 ve donat per:

$$\frac{\Delta x_1}{x_0} = \frac{x_1 - x_0}{x_0} \tag{2-43}$$

Multiplicant un canvi proporcional per 100 s'obté un *canvi proporcional* en %. És a dir:

$$100\frac{\Delta x_1}{x_0}\% \tag{2-44}$$

El *canvi en logaritmes* i el *canvi en logaritmes* en *%* entre x_1 i x_0, vénen donats per

$$\begin{gathered}\Delta \ln(x) = \ln(x_1) - \ln(x_0) \\ 100\Delta \ln(x)\%\end{gathered} \tag{2-45}$$

El canvi en logaritmes és una aproximació del canvi proporcional, com es pot veure a l'apèndix 2.3. Aquesta aproximació és bona quan la variació és xicoteta, però les diferències poden ser importants quan el canvi proporcional és gran, com es pot observar al quadre 2.3.

QUADRE 2.3. Exemples de canvis proporcionals i canvis en logaritmes.

x_1	202	210	220	240	300
x_0	200	200	200	200	200
Canvi proporcional en %	1%	5.0%	10.0%	20.0%	50.0%
Canvi en logaritmes en %	1%	4.9%	9.5%	18.2%	40.5%

L'*elasticitat* és la raó entre els canvis relatius de dues variables. Si s'utilitzen canvis proporcionals, l'elasticitat de la variable *y* pel que fa a la variable *x* ve donada per

$$\varepsilon_{y/x} = \frac{\Delta y / y_0}{\Delta x / x_0} \tag{2-46}$$

Si es prenen logaritmes s'obtenen canvis infinitesimals, aleshores, l'elasticitat de la variable *y* amb respecte a una variable *x* ve donada per

$$\varepsilon_{y/x} = \frac{dy / y}{dx / x} = \frac{d\ln(y)}{d\ln(x)} \tag{2-47}$$

En general, en els models economètrics, l'elasticitat es defineix segons (2-47).

Formes funcionals alternatives

El mètode *MQO* també es pot aplicar a models en què s'hagin transformat la variable endògena i/o l'exògena. El model (2-1) ens mostra que la variable exògena i el regressor són termes equivalents. Però a partir d'ara, anomenarem regressor a la forma específica en què una variable exògena apareix en l'equació. Per exemple, en el model

$$y = \beta_1 + \beta_2 \ln(x) + u$$

la variable exògena és *x*, però el regressor és ln(*x*).

El model de (2-1) també ens indica que la variable endògena i el regressant són equivalents. Però d'ara en endavant, anomenarem regressant a la forma específica en què una variable endògena apareix en l'equació. Per exemple, en el model

$$\ln(y) = \beta_1 + \beta_2 x + u$$

la variable endògena és *y*, però el regressant és ln (*y*).

Tots dos models són lineals en els paràmetres, encara que no són lineals en la variable *x* (el primer) o en la variable *y* (el segon). En qualsevol cas, si un model és lineal en els paràmetres, es pot estimar aplicant el mètode de *MQO*. Per contra, si un model no és lineal en els paràmetres, l'estimació s'ha de fer per mètodes iteratius.

No obstant això, hi ha certs models no lineals que, per mitjà de transformacions adequades, poden convertir-se en lineals. Aquests models són denominats linealitzables.

Així, en algunes ocasions es postulen models potencials en la teoria econòmica, com és el cas de la coneguda funció de producció de Cobb-Douglas. Un model potencial amb una única variable explicativa ve donat per

$$y = e^{\beta_1} x^{\beta_2}$$

Si s'introdueix el terme de pertorbació de forma multiplicativa s'obté

$$y = e^{\beta_1} x^{\beta_2} e^{u} \tag{2-48}$$

Prenent logaritmes en tots dos membres de (2-48), s'obté un model lineal en els paràmetres:

$$\ln(y) = \beta_1 + \beta_2 \ln(x) + u \tag{2-49}$$

Per contra, si s'introdueix el terme de pertorbació de forma additiva, s'obté

$$y = e^{\beta_1} x^{\beta_2} + u$$

En aquest cas no hi ha una transformació que permeti convertir-lo en un model lineal. Serà, per tant, un model no linealitzable.

Ara, considerarem alguns models amb formes funcionals alternatives, però tots ells són lineals en els paràmetres. Estudiarem en cada cas la interpretació del coeficient $\hat{\beta}_2$:

a) Model lineal

El coeficient $\hat{\beta}_2$ mesura l'efecte del regressor *x* sobre *y*. Vegem això amb detall. observació i de la funció de regressió mostral s'expressa d'acord amb (2-24) per

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i \tag{2-50}$$

Considerem ara l'observació h del model ajustat en la qual el valor del regressor i, en conseqüència, del regressant han canviat pel que fa a (2-50):

$$\hat{y}_h = \hat{\beta}_1 + \hat{\beta}_2 x_h \tag{2-51}$$

Si restem (2-51) de (2-50), veiem que *x* té un efecte lineal sobre $\hat{y}$:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x \tag{2-52}$$

on $\Delta\hat{y} = \hat{y}_i - \hat{y}_h$ y $\Delta x = x_i - x_h$

Per tant, $\hat{\beta}_2$ és el canvi produït en *y* (a les unitats en les que estiga mesurada i) en canviar x en una unitat (a les unitats en les que estia mesurada x).

Per exemple, en la funció ajustada (2-39), si la renda augmenta en una unitat, el consum s'incrementarà en 0.85 unitats.

La linealitat d'aquest model implica que un canvi d'una unitat en *x* té sempre el mateix efecte en *y*, amb independència del valor de *x* considerat.

EXEMPLE 2.7 Quantitat de cafè venut com una funció del seu preu. Model lineal

En un experiment de màrqueting[1] es va formular el següent model per explicar la quantitat de cafè venut per setmana (*coffqty*) en funció del preu del cafè (*coffpric*)

$$coffqty = \beta_1 + \beta_2 coffpric + u$$

La variable *coffpric* pren el valor 1, el preu habitual, i també els valors 0.95 i 0.85 en dues accions els efectes estan sota investigació. L'experiment va durar 12 setmanes, *coffqty* està expressat en milers d'unitats i c*offpric* en francs francesos. Les dades apareixen en el quadre 2.4 i en el fitxer *coffee1*.

El model ajustat és el següent:

[1] Les dades d'aquest exercici s'han obtingut d'un experiment controlat de màrqueting, sobre la despesa en cafè en botigues de París. La referència és A. C. Bemmaor and D. Mouchoux, "Measuring the Short-Term Effect of In-Store Promotion and Retail Advertising on Brand Sales: A Factorial Experiment". *Journal of Marketing Research*, 28 (1991), 202–14.

$$coffqty = 774.9 - 693.33 coffpric \quad R^2 = 0.95 \quad n = 12$$

Interpretació del coeficient $\hat{\beta}_2$: si el preu del cafè s'incrementa en 1 franc francès, la quantitat venuda de cafè es reduirà en 693.33 milers d'unitats. En la mesura que el preu del cafè és una magnitud xicoteta, és preferible donar la següent interpretació: si augmenta el preu del cafè en 1 cèntim de franc francès, la quantitat venuda de cafè es reduirà en 6.93 milers d'unitats.

QUADRE 2.4. Dades sobre quantitats i preus del cafè.

setmana	*coffpric*	*coffqty*
1	1.00	89
2	1.00	86
3	1.00	74
4	1.00	79
5	1.00	68
6	1.00	84
7	0.95	139
8	0.95	122
9	0.95	102
10	0.85	186
11	0.85	179
12	0.85	187

EXEMPLE 2.8 Explicant el valor de mercat dels bancs espanyols. model lineal

Utilitzant dades de la Borsa de Madrid (Borsa de Madrid) del 18 d'agost de 1995 (fitxer *bolmad95*, 20 primeres observacions), s'ha estimat el següent model per explicar el valor de mercat de bancs i institucions financeres:

$$marktval = 29.42 + 1.219 bookval$$

$$R^2=0.836 \qquad n=20$$

on

- *marktval* és el valor en mercat d'una empresa. Es calcula multiplicant el preu de l'acció pel nombre d'accions emeses.
- *bookval* és el valor comptable o el valor net de la companyia. El valor comptable es calcula com la diferència entre els actius d'una empresa i els seus passius.
- Les dades de *marktva*l i *bookval* estan expressats en milions de pessetes.

Interpretació del coeficient β_2: si el valor comptable d'un banc s'incrementa en 1 milió de pessetes, la capitalització de mercat d'aquest banc s'incrementarà en 1.219 milions de pessetes.

b) Model lineal logarítmic

Un model lineal logarítmic s'expressa com

$$y = \beta_1 + \beta_2 \ln(x) + u \tag{2-53}$$

La funció ajustada corresponent és la següent:

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \ln(x) \tag{2-54}$$

Prenent primeres diferències en (2-54), i multiplicant i dividint el segon membre per 100, s'obté

$$\Delta\hat{y} = \frac{\hat{\beta}_2}{100} 100 \times \Delta \ln(x)\%$$

Per tant, si *x* augmenta un 1%, $\hat{y}$ s'incrementarà en $(\hat{\beta}_2 / 100)$ unitats.

c) Model logarítmic lineal

Un model logarítmic lineal s'expressa com

$$\ln(y) = \beta_1 + \beta_2 x + u \tag{2-55}$$

El model anterior s'obté prenent logaritmes naturals en els dos membres del següent model:

$$y = \exp(\beta_1 + \beta_2 x + u)$$

Per aquesta raó, el model (2-55) també es diu exponencial.

La funció de regressió mostral corresponent a (2-55) és la següent

$$\ln(y) = \hat{\beta}_1 + \hat{\beta}_2 x \tag{2-56}$$

Prenent les primeres diferències en (2-56), i multiplicant ambdós membres per 100, s'obté

$$100 \times \Delta \ln(y)\% \ = \ 100 \times \hat{\beta}_2 \Delta x$$

Per tant, si *x* augmenta en una unitat, aleshores $\hat{y}$ s'incrementarà un $100 \times \hat{\beta}_2$%.

d) Model doblement logarítmic

El model que figura a (2-49) és un model doblement logarítmic o, abans de la transformació, un model potencial (2-48). A aquest model també se l'anomena model d'elasticitat constant.

El model ajustat corresponent a (2-49) és el següent:

$$\ln(y) = \hat{\beta}_1 + \hat{\beta}_2 \ln(x) \tag{2-57}$$

Prenent primeres diferències en (2-57), s'obté

$$\Delta \ln(y) = \hat{\beta}_2 \Delta \ln(x)$$

Per tant, si *x* augmenta en 1%, aleshores $\hat{y}$ s'incrementarà un $\hat{\beta}_2$ %. Cal ressaltar que, en aquest model, $\hat{\beta}_2$ és l'elasticitat estimada de *y* pel que fa a *x*, per a qualsevol valor de *x* i *y*. En conseqüència, en aquest model l'elasticitat és constant.

A l'annex 1 en un cas d'estudi de la corba d'Engel per a la demanda de productes lactis s'analitzen sis formes funcionals alternatives.

EXEMPLE 2.9 Quantitat de cafè venut en funció del seu preu. Model doblement logarítmic (continuació de l'exemple 2.7)

Como una alternativa al model lineal s'ha estimat el model doblement logarítmic:

$$\ln(coffqty) = 4.415 - 5.132\ln(coffpric) \quad R^2 = 0.90 \quad n = 12$$

Interpretació del coeficient $\hat{\beta}_2$: si el preu del cafè augmenta en un 1%, la quantitat venuda de cafè es reduirà en un 5,13%. En aquest cas, $\hat{\beta}_2$ és l'estimador de l'elasticitat de la demanda/preu.

EXEMPLE 2.10 Explicant el valor de mercat dels bancs espanyols. Model doblement logarítmic (continuació de l'exemple 2.8)

Utilitzant dades de l'exemple 2.8, s'ha estimat el següent model doblement logarítmic:

$$\ln(marktval) = 0.6756 + 0.938\ln(bookval)$$

R^2=0.928 n=20

Interpretació del coeficient $\hat{\beta}_2$: si el valor comptable d'un banc s'incrementa en 1%, el valor de mercat d'aquest banc s'incrementarà en un 0,938%. En aquest cas $\hat{\beta}_2$ és l'estimador de l'elasticitat del valor de mercat/valor comptable.

En el quadre 2.5 es mostra, per al model ajustat, la interpretació dels quatre models estudiats. Si haguéssem considerat el model poblacional en lloc del mostral, la interpretació de β_2 és la mateixa però tenint en compte que Δu hauria de ser igual a 0.

QUADRE 2.5. Interpretació de $\hat{\beta}_2$ en els diferents models.

Model	Si x augmenta en	aleshores y s'incrementarà en
lineal	1 unitat	$\hat{\beta}_2$ unitats
lineal logarítmic	1%	$(\hat{\beta}_2/100)$ unitats
logarítmic lineal	1 unitat	$(100\hat{\beta}_2)\%$
doblement logarítmic	1%	$\hat{\beta}_2\%$

2.5 Supòsits i propietats estadístiques dels *MQO*

Anem ara a estudiar les propietats estadístiques dels estimadors de *MQO*, $\hat{\beta}_1$ i $\hat{\beta}_2$, del model de regressió lineal simple. Prèviament, cal formular un conjunt de supòsits estadístics. Específicament, al conjunt de supòsits que anem a formular se'ls denomina supòsits del model lineal clàssic (*MLC*). És important senyalar que els supòsits del *MLC* són senzills, i que els estimadors *MQO* tenen, sota aquests supòsits, molt bones propietats.

2.5.1 Supòsits estadístics del *MLC* en regressió lineal simple

a) Supòsit sobre la forma funcional

1) La relació entre el regressant, regressor i pertorbació aleatòria és lineal en els paràmetres:

$$y = \beta_1 + \beta_2 x + u \tag{2-58}$$

El regressant i els regressors poden ser qualsevol funció de la variable endògena i de les variables explicatives, respectivament, a condició que entre els regressors i el regressant existisca una relació lineal. És a dir, el model és lineal en els paràmetres. L'additivitat de la pertorbació garanteix la relació lineal amb la resta dels elements.

b) Supòsits sobre el regressor x

2) Els valors que pren x són fixos en repetides mostres:

D'acord amb aquest supòsit, cada observació del regressor pren el mateix valor per a diferents mostres del regressant. Aquest és un supòsit fort en el cas de les ciències socials, on, en general, no és possible l'experimentació. Les dades s'obtenen mitjançant observació, no mitjançant experimentació. És important destacar que els resultats obtinguts basats en aquest supòsit romanen virtualment idèntics als que s'obtenen quan assumim que els regressors són estocàstics, sempre, que postulem el supòsit addicional d'independència entre els regressors i la pertorbació aleatòria. Aquest supòsit alternatiu es pot formular així:

2) El regressor x es distribueix de forma independent de la pertorbació aleatòria.*

En qualsevol cas, al llarg d'aquest capítol i els següents anem a adoptar el supòsit 2.

3) El regressor x no conté errors de mesurament

Es tracta d'un supòsit que sovint no es compleix en la pràctica, ja que els instruments de mesurament no són sempre fiables en l'economia. Pensem, per exemple, en la multitud d'errors que es poden cometre en la recopilació d'informació quan es fan enquestes a les famílies.

4) La variança mostral de x és diferent de 0 i té un límit finit quan n tendeix a infinit

Per tant, aquest supòsit implica que

$$S_X^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n} \neq 0 \qquad (2\text{-}59)$$

c) Supòsit sobre els paràmetres

5) Els paràmetres β_i i β_2 són fixos

Si no s'adopta aquest supòsit, el model de regressió seria molt difícil d'aplicar. En qualsevol cas, pot ser acceptable postular que els paràmetres del model són estables en el temps (si no és un període molt llarg) o en l'espai (si és relativament limitat).

d) Supòsits sobre les pertorbacions aleatòries

6) L'esperança de les pertorbacions és zero,

$$E(u_i) = 0, \quad i = 1,2,3,\ldots,n \qquad (2\text{-}60)$$

Aquest no és un supòsit restrictiu, ja que sempre es pot utilitzar β_1 per normalitzar E(u) a 0. Suposem, per exemple, que, aleshores podríem redefinir el model de la manera:

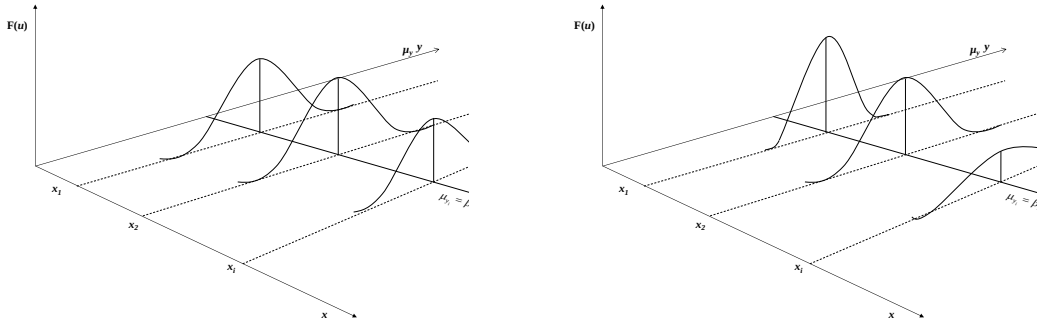
$$y = (\beta_1 + 4) + \beta_2 x + v$$

on $v = u - 4$. Per tant, l'esperança de la nova pertorbació, v, és 0 i l'esperança de u ha estat absorbida pel terme independent.

7) Les pertorbacions tenen una variança constant

$$var(u_i) = \sigma^2 \qquad i = 1, 2, \cdots n \tag{2-61}$$

A aquest supòsit se li denomina supòsit d'homoscedasticitat. Aquesta paraula ve del grec: *homo* (igual) i s*cedasticitat* (variabilitat). Això vol dir que la variabilitat al voltant de la línia de regressió és la mateixa en tota la mostra de *x*; és a dir, que no augmenta o disminueix quan *x* varia, com es pot veure a la figura 2.7, part a), on les pertorbacions són homoscedàstiques.



a) **b)**

FIGURA 2.7. Pertorbacions aleatòries: a) homoscedasticitat; b) heteroscedasticitat.

Si aquest supòsit no es compleix, com passa a la part b) de la figura 2.7, els estimadors de MQO no són eficients. Les pertorbacions en aquest cas es diu que són heteroscedàstiques (*hetero* vol dir diferent).

8) Les pertorbacions amb diferents subíndexs no estan correlacionades entre si (supòsit de no autocorrelació):

$$E(u_i u_j) = 0 \qquad i \neq j \tag{2-62}$$

És a dir, les pertorbacions corresponents a diferents individus o a diferents moments de temps, no estan correlacionades entre si. Aquest supòsit de no autocorrelació o no correlació serial, igual que en el cas d'homoscedasticitat, és contrastable *a posteriori*. La transgressió d'aquest supòsit es produeix amb força freqüència en els models que utilitzen dades de sèries temporals.

9) Les pertorbacions es distribueixen normalment

Tenint en compte els supòsits 6, 7 i 8 s'obté que

$$u_i \sim NID(0, \sigma^2) \qquad i = 1, 2, \cdots, n \tag{2-63}$$

on *NID* indica que les pertorbacions estan normal i independentment distribuïdes.

La raó d'aquest supòsit és que, si o es distribueix normalment, també ho faran *y* i els coeficients estimats de regressió, la qual cosa és útil en la realització de contrastos d'hipòtesis i en la construcció d'intervals de confiança per a β_1 i β_2 La justificació d'aquest supòsit es basa en el Teorema Central del Límit. En essència, aquest teorema indica que, si una variable aleatòria és el resultat agregat dels efectes d'un nombre indefinit de variables, tindrà una distribució aproximadament normal, fins i tot si els seus components no la tenen, a condició que cap d'ells siga dominant.

2.5.2 Propietats desitjables dels estimadors

Abans d'examinar les propietats dels estimadors mínim-quadràtics sota els supòsits estadístics de l'*MLC*, es pot plantejar la següent qüestió prèvia: ¿quines són les propietats desitjables per a un estimador?

Dues propietats desitjables per a un estimador és que siga no esbiaixat i que la seva variança siga el més xicoteta possible. Si això succeeix, el procés d'inferència es podrà dur a terme d'una manera satisfactòria.

Anem a il·lustrar aquestes propietats de forma gràfica. Considerem en primer lloc l'absència de biaix. A les figures 2.8 i 2.9 s'han representat les funcions de densitat de dos hipotètics estimadors obtinguts per dos procediments diferents:

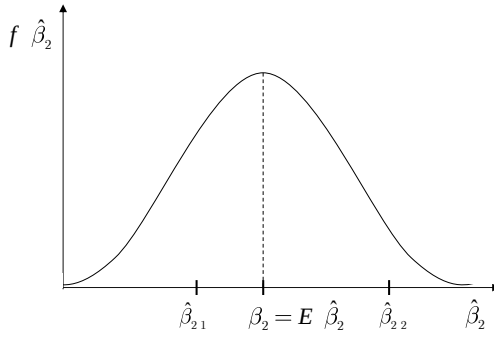


FIGURA 2.8. Estimador no esbiaixat.

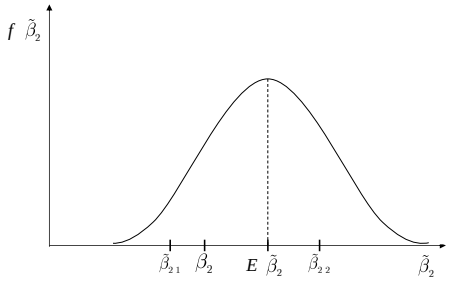


FIGURA 2.9. Estimador esbiaixat.

La $\hat{\beta}_2$ és no esbiaixada, és a dir, la seua esperança matemàtica és igual al paràmetre que tracta d'estimar, β_2. L'estimador $\hat{\beta}_2$ és una variable aleatòria que en cada mostra de *y*- les x són fixes en repetides mostra segons el supòsit 2- pren un valor diferent, però de mitjana, és a dir, tenint en compte els infinits valors que $\hat{\beta}_2$ pot prendre, és igual al paràmetre β_2. Amb cada mostra de *y* s'obté un valor específic de $\hat{\beta}_2$, és a dir, una estimació. A la figura 2.8 s'han representat dues estimacions β_2: $\hat{\beta}_{2(1)}$ i $\hat{\beta}_{2(2)}$. La primera estimació està relativament a prop de β_2, mentre que la segona està molt més allunyada. En tot cas, l'absència de biaix és una propietat desitjable, ja que ens assegura que l'estimador de mitjana està centrat sobre el paràmetre

L'estimador $\tilde{\beta}_2$, en la figura 2.9, és esbiaixat, ja que la seva esperança no és igual a β_2. El biaix és precisament $E\ \tilde{\beta}_2\ -\beta_2$. En aquest cas també s'han representat dues hipotètiques estimacions: $\tilde{\beta}_{2(1)}$ i $\tilde{\beta}_{2(2)}$. Com es pot veure $\tilde{\beta}_{2(1)}$ està més a prop de β_2 que l'estimador no esbiaixat $\hat{\beta}_{2(1)}$: és una qüestió d'atzar. En tot cas, per ser esbiaixat no està centrat en mitjana sobre el paràmetre. No hi ha dubte que sempre és preferible un estimador no esbiaixat ja que, amb independència del que passe en una mostra concreta, no té una desviació sistemàtica respecte al valor del paràmetre.

L'altra propietat desitjable és l'eficiència. Aquesta propietat fa referència a la variança dels estimadors. A les figures 2.10 i 2.11 s'han representat dos hipotètics estimadors no esbiaixats als que seguirem anomenant $\hat{\beta}_2$ i $\tilde{\beta}_2$. El primer d'ells té una variança més xicoteta que el segon.

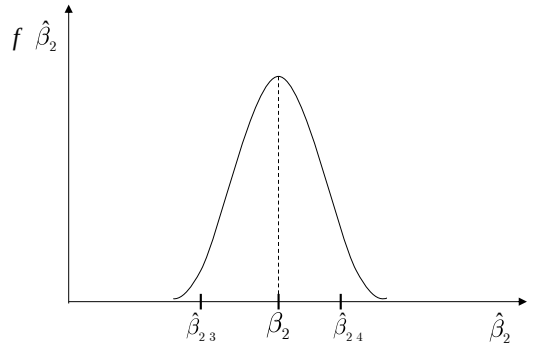


FIGURA 2.10. Estimador amb una variança xicoteta.

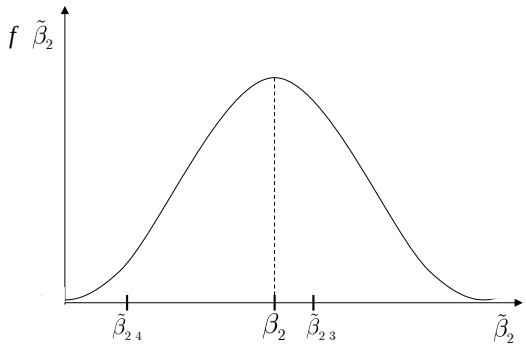


FIGURA 2.11. Estimador amb una variança gran.

En ambdues figures hem representat dues estimacions: $\hat{\beta}_{2(3)}$ i $\hat{\beta}_{2(4)}$ en l'estimador amb variança més xicoteta; $\tilde{\beta}_{2(3)}$ i $\tilde{\beta}_{2(4)}$ en l'estimador amb variança més gran. També aquí, per ressaltar el paper jugat per l'atzar, l'estimació que està més a prop de β_2 és precisament $\tilde{\beta}_{2(3)}$. En qualsevol cas, sempre és preferible que la variança de l'estimador siga el més xicoteta possible. Així per exemple, utilitzant l'estimador $\hat{\beta}_2$ és pràcticament impossible que una estimació estiga tan allunyada de β_2 com ho està $\tilde{\beta}_{2(4)}$, pel fet que el recorregut de $\hat{\beta}_2$ és molt més reduït que el que té $\tilde{\beta}_2$.

2.5.3 Propietats estadístiques dels estimadors *MQO*

Sota els supòsits anteriors, els estimadors *MQO* posseeixen algunes propietats ideals. Així, podem dir que els *MQO* són estimadors lineals no esbiaixats i òptims.

Els estimadors de MQO són lineals i no esbiaixats

L'estimador $\hat{\beta}_2$ de *MQO* és no esbiaixat. A l'apèndix 2.4 es demostra que és un estimador no esbiaixat utilitzant implícitament els supòsits 3, 4 i 5, i explícitament els supòsits 1, 2 i 6. En aquest annex també es pot veure que és un estimador lineal, utilitzant els supòsits 1 i 2. De la mateixa manera, es pot demostrar que l'estimador *MQO* $\hat{\beta}_1$ és no esbiaixat.

Recordem que l'absència de biaix és una propietat general de l'estimador, però que per a una mostra determinada l'estimació pot estar més "a prop" o més "lluny" del veritable paràmetre. En qualsevol cas, la distribució de l'estimador està centrada en el paràmetre poblacional.

Variances dels estimadors de MQO

Ara sabem que la distribució mostral del nostre estimador està centrada en el paràmetre poblacional, però ¿quina és la dispersió de la seva distribució? La variança, que és una mesura de dispersió, d'un estimador és un indicador de la precisió d'aquest estimador.

Per obtenir les variances de $\hat{\beta}_1$ i $\hat{\beta}_2$ es requereixen els supòsits 7 i 8, a més dels sis primers. Aquestes variances són les següents:

$$Var(\hat{\beta}_1) = \frac{\sigma^2 n^{-1}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2-64}$$

A l'apèndix 2.5 es mostra com s'obté la variança de $\hat{\beta}_2$.

Els estimadors de MQO són ELNEO

Els estimadors de *MQO* tenen la menor variança d'entre tots els estimadors lineals i no esbiaixats. Per aquesta raó es diu que els estimadors de *MQO* són estimadors *lineals no esbiaixats i òptims* (*ELNEO*), com es mostra a la figura 2.12. Aquesta propietat es coneix com el teorema de Gauss-Markov. Per a la demostració d'aquest teorema s'utilitzen els supòsits 1 a 8, com es pot veure a l'apèndix 2.6. Aquest conjunt de supòsits es coneix com els supòsits de Gauss-Markov.

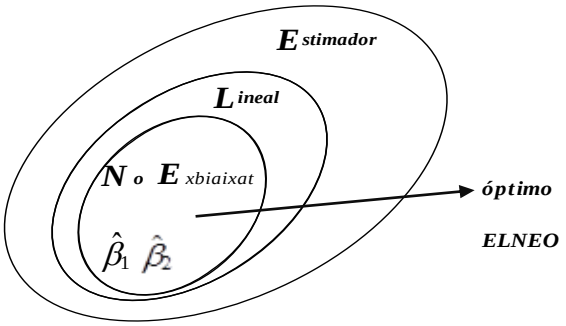


FIGURA 2.12. Els estimadors *MQO* són *ELNEO*.

L'estimació de la variança de les pertorbacions i de la variança dels estimadors

Atès que no coneixem el valor de la variança de la pertorbació, σ^2, hem d'estimar-lo. No obstant això, no podem estimar-utilitzant els valors mostrals de les pertorbacions u_i perquè no són observables En el seu lloc, hem d'utilitzar els residus de *MQO* ($\hat{u}_i$).

La relació entre les pertorbacions i els residus ve donada per

$$\begin{aligned} \hat{u}_i &= y_i - \hat{y}_i = \beta_1 + \beta_2 x_i + u_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \\ &= u_i - \left(\hat{\beta}_1 - \beta_1\right) - \left(\hat{\beta}_2 - \beta_2\right) x_i \end{aligned} \tag{2-65}$$

Per tant, $\hat{u}_i$ no és el mateix que u_i, encara que la diferència entre ells - $\left(\hat{\beta}_1 - \beta_1\right) - \left(\hat{\beta}_2 - \beta_2\right) x_i$ - té un valor esperat que és igual a zero. Per això, un primer estimador de σ^2 podria ser la variança residual:

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n} \tag{2-66}$$

No obstant això, aquest estimador és esbiaixat, essencialment perquè no té en compte les dues següents restriccions que han de ser satisfetes pels residus de *MQO* en el model de regressió lineal simple:

$$\begin{cases} \sum_{i=1}^{n} \hat{u}_i = 0 \\ \sum_{i=1}^{n} x_i \hat{u}_i = 0 \end{cases} \tag{2-67}$$

Una forma de veure aquestes restriccions és la següent: si coneixem *n*-2 dels residus, podem obtenir els altres dos residus mitjançant l'ús de les restriccions implícites en les equacions normals (2-67).

Per tant, només hi ha *n*-2 graus de llibertat en els residus de *MQO*, a diferència dels *n* graus de llibertat que tindrien les corresponents *n* pertorbacions. En l'estimador no esbiaixat de σ^2 mostrat a continuació es realitza un ajust en el qual es té en compte els graus de llibertat:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2} \tag{2-68}$$

Sota els supòsits 1-8 (supòsits Gauss-Markov), s'obté, com es pot veure a l'apèndix 2.7, que

$$E(\hat{\sigma}^2) = \sigma^2 \tag{2-69}$$

Si s'introdueix $\hat{\sigma}^2$ en les fórmules de la variança obtenim aleshores els estimadors no esbiaixats de var($\hat{\beta}_1$) i var($\hat{\beta}_2$)

L'estimador natural de σ és $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ i es diu *error estàndard de la regressió*. L'arrel quadrada de la variança es denomina *desviació estàndard* de $\hat{\beta}_2$, és a dir,

$$de(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{2-70}$$

Per tant, el seu estimador natural, al que es denomina *error estàndard* de $\hat{\beta}_2$, ve donat per

$$ee(\hat{\beta}_2) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{2-71}$$

Cal notar que $ee(\hat{\beta}_2)$, a causa de la presència de l'estimador $\hat{\sigma}$ en (2-71), és una variable aleatòria igual que $\hat{\beta}_2$. L'error estàndard d'una estimació ens ofereix una idea del necessari que és l'estimador.

La consistència dels MQO i altres propietats asimptòtiques

A vegades no és possible obtenir un estimador no esbiaixat. Aleshores, es considera que la *consistència* és el requisit mínim que ha de complir l'estimador. Segons un enfocament intuïtiu, *consistència* vol dir que a mesura que $n \to \infty$, la funció de densitat de l'estimador convergeix al valor del paràmetre. Aquesta propietat pot expressar-se per l'estimador $\hat{\beta}_2$ com:

$$\operatorname{plim}_{n\to\infty} \hat{\beta}_2 = \beta_2 \tag{2-72}$$

on plim és el límit en probabilitat. En altres paraules, $\hat{\beta}_2$ convergeix en probabilitat a β_2.

És important tenir en ment que l'absència de biaix i consistència són conceptualment diferents. L'absència de biaix es manté per a qualsevol mida de la mostra, mentre que la consistència és una propietat estrictament de grans mostres o, de forma més precisa, és una *propietat asimptòtica*.

Sota els supòsits 1 a 6, els estimadors *MQO*, $\hat{\beta}_1$ i $\hat{\beta}_2$ són consistents. La demostració de la consistència de $\hat{\beta}_2$ es pot veure a l'apèndix 2.8.

Altres propietats asimptòtiques de $\hat{\beta}_1$ i $\hat{\beta}_2$: Sota els supòsits de Gauss-Markov 1 a 8, $\hat{\beta}_1$ i $\hat{\beta}_2$ tenen una *distribució asimptòticament normal* i és *asimptòticament eficient* dins de la classe d'estimadors consistents i asimptòticament normals.

Els estimadors MQO són estimadors de màxima versemblança (MV) i estimadors no esbiaixats de mínima variança (ENEMV)

Ara anem a introduir el supòsit 9 a la normalitat de les pertorbacions u. El conjunt de supòsits 1 a 9 es coneixen com els supòsits del *model lineal clàssic* (*MLC*)

Sota els supòsits de *MLC*, els estimadors de *MQO* són també *estimadors de màxima versemblança* (*MV*), com es pot veure a l'apèndix 2.8.

D'altra banda, sota els supòsits de *MLC*, els estimadors de *MQO* a més de ser *ELIO*, són *estimadors no esbiaixats de mínima variança* (*ENEMV*). Això significa que els estimadors de *MQO* tenen la variança més xicoteta entre tots els estimadors no esbiaixats, lineals o no lineals, segons es mostra a la figura 2.13. Per tant, ja no hem de restringir-als estimadors que són lineals en y_i.

També es compleix que qualsevol combinació lineal de $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, ..., \hat{\beta}_k$ es distribueix normalment, i qualsevol subconjunt de $\hat{\beta}_j$ les té una distribució normal conjunta.

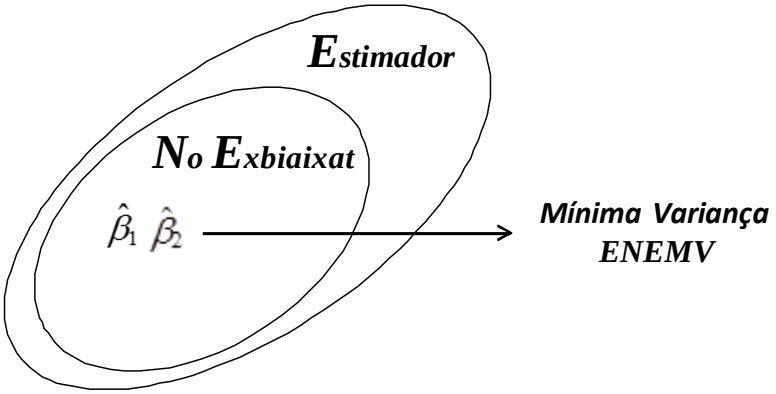


FIGURA 2.13. Els estimadors MQO són *ENEMV*.

En resum, hem vist que els estimadors de *MQO* tenen propietats molt desitjables quan es compleixen els supòsits estadístiques del *MLC*.

Exercicis

Exercici 2.1 El següent model ha estat formulat per explicar les vendes anuals (*vendes*) d'empreses fabricants de productes de neteja domèstica en funció d'un índex de preus relatiu (*ipr*):

$$vendes = \beta_1 + \beta_2 ipr + u$$

on la variable *vendes* està expressada en milions d'euros i *ipr* és un índex de preus relatius (preus de l'empresa/preus de l'empresa 1 de la mostra). Així, el valor 110 de l'empresa 2 indica que el seu preu és un 10% mal elevat que a l'empresa 1.

Per a això es disposa de les dades sobre deu empreses fabricants de productes de neteja domèstica:

empresa	*vendes*	*ipr*
1	10	100
2	8	110
3	7	130
4	6	100
5	13	80
6	6	80
7	12	90
8	7	120
9	9	120
10	15	90

a) Estimeu β_1 i β_2 per *MQO*.
b) Obtinga la suma dels quadrats dels residus.
c) Calculeu el coeficient de determinació.
d) Comproveu si es compleixen les implicacions algebraiques 1, 3 i 4 en l'estimació per *MQO*.

Exercici 2.2 Per estudiar la relació entre consum de combustible (*y*) i el temps de vol (*x*) en una companyia aèria s'ha formulat el següent model:

$$y = \beta_1 + \beta_2 x + u$$

on i està expressat en milers de lliures i x en hores, utilitzant-se com unitats d'ordre inferior fraccions decimals de l'hora.

De les estadístiques de «Temps de vol i consums de combustible» d'una companyia aèria s'han obtingut dades relatives a temps de vol i consums de combustible de 24 trajectes diferents realitzats per avions DC-9. A partir d'aquestes dades s'han elaborat els següents estadístiques:

$$\sum y_i = 219.719;\ \sum x_i = 31.470;\ \sum x_i^2 = 51.075;$$

$$\sum x_i y_i = 349.486;\ \sum y_i^2 = 2396.504$$

Es demana

a) L'estimació de β_1 i β_2.

b) La descomposició de la variança de *y* en variança explicada per la regressió i variança residual.

c) El coeficient de determinació.

d) Què consum total s'estimaria, en milers de lliures, per a un programa de vols compost per 100 vols de mitja hora, 200 d'una hora i 100 de dues hores?

Exercici 2.3 Un analista formula el següent model:

$$y = \beta_1 + \beta_2 x + u$$

Utilitzant una mostra donada, s'estima el model obtenint els següents resultats:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n} = 20 \qquad \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = 10 \qquad \begin{aligned} \overline{y} &= 8 \\ \overline{x} &= 4 \\ \hat{\beta}_2 &= 3 \end{aligned}$$

Li semblen coherents els resultats obtinguts per l'analista?

Exercici 2.4 Un econòmetra ha estimat el següent model amb una mostra de cinc observacions:

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

Un cop realitzada l'estimació l'econòmetra perd tota la informació excepte la que apareix en el quadre:

Obs.	x_i	$\hat{u}_t$
1	1	2
2	3	-3
3	4	0
4	5	¿?
5	6	¿?

Amb la informació anterior l'econòmetra ha de calcular la variança residual. Faci-ho en el seu lloc.

Exercici 2.5 Siga el següent model

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad 1 = 1, 2, \ldots, n$$

En estimar aquest model amb una mostra de mida 11 s'han obtingut els següents resultats:

$$\sum_{i=1}^{n} x_i = 0 \qquad \sum_{i=1}^{n} y_i = 0 \qquad \sum_{i=1}^{n} x_i^2 = B \qquad \sum_{i=1}^{n} y_i^2 = E \qquad \sum_{i=1}^{n} x_i y_i = F$$

a) Obtinga l'estimació de β_2 i β_1.
b) Obtinga la suma de quadrats dels residus.
c) Calculeu el coeficient de determinació.
d) Calculeu el coeficient de determinació sota el supòsit que $2F^2 = BE$.

Exercici 2.6 L'empresa A es dedica a muntar panells prefabricats per a naus industrials. Fins al moment ha realitzat 8 obres, en les quals el nombre de metres quadrats de panells i el d'hores de treball directament emprades en el muntatge han estat els següents:

Núm. de metres quadrats (milers)	**Núm. d'hores**
4	7400
6	9800
2	4600
8	12200
10	14000
5	8200
3	5800
12	17000

L'empresa A vol participar en un concurs per muntar 14000 m^2 de panell per a una nau industrial, per a això ha de presentar un pressupost.

Com a dades a tenir en compte en l'elaboració del pressupost, es coneixen els següents:

a) El pressupost s'ha de referir exclusivament als costos de muntatge, ja que el material el proporciona l'empresa que ha convocat el concurs.

b) El cost per hora treballada per a l'empresa A és de 1100 pessetes.

c) Per cobrir els restants costos, l'empresa A ha de carregar un 20% sobre l'import total del cost de mà d'obra emprada en el muntatge.

Per la situació en què es troba, a l'empresa A li interessa participar en el concurs amb un pressupost en el qual únicament es cobreixen els costos. En aquestes condicions, i sota el supòsit que el nombre d'hores treballades és funció lineal del nombre de metres quadrats de panells muntats, quin hauria de ser l'import del pressupost de l'empresa A?

Exercici 2.7 Considereu les següents igualtats:

1. $E[u] = 0$.

2. $E[\hat{u}] = 0$.

3. $\bar{u} = 0$.

4. $\bar{\hat{u}} = 0$.

En el context del model lineal, indiqueu si cadascuna de les anteriors igualtats es compleix o no, raonant la resposta.

Exercici 2.8 S'han estimat per mínims quadrats ordinaris els paràmetres β_1 i β_2 del model

$$y = \beta_1 + \beta_2 x + u$$

amb una mostra de mida 3.

Els valors de x_i són {1,2,3}. Se sap també que el residu corresponent a la primera observació és de 0.5.

A partir de l'anterior informació, és possible calcular la suma dels quadrats dels residus i obtenir una estimació de σ^2? En cas afirmatiu, realitzi els corresponents càlculs.

Exercici 2.9 Es tenen les següents dades, per estimar una relació entre *y* i *x*:

y	*x*
-2	-2
-1	0
0	1
1	0
2	1

a) Estimeu per *MQO* els paràmetres α i β del següent model:

$$y = \alpha + \beta x + \varepsilon$$

b) Estimeu var(ε_i).

c) D'altra banda, estime per *MQO* els paràmetres γ i δ del següent model:

$$x = \gamma + \delta y + \upsilon$$

d) Són les dues línies de regressió ajustades iguals? Expliqueu el resultat en termes de la metodologia mínim-quadràtica.

Exercici 2.10 Responeu a les següents preguntes:

a) Un investigador, després de realitzar l'estimació d'un model per *MQO*, calcula $\sum \hat{u}_i$ i comprova que no és 0. És això possible? Raoneu la resposta indicant en el seu cas les condicions en les quals pot haver-se produït aquest fet.

b) Obtinga un estimador no esbiaixat de σ^2, indicant els supòsits utilitzats. Raoneu la resposta.

Exercici 2.11 En el context del model de regressió lineal

$$y = \beta_1 + \beta_2 x + u$$

a) Indiqueu en què es basa el compliment, si s'escau, de les següents igualtats

$$\bar{u} = \frac{\sum_{i=1}^{n} u_i}{n} = 0; \qquad \bar{\hat{u}} = \frac{\sum_{i=1}^{n} \hat{u}_i}{n} = 0; \qquad E\left[x_i u_i\right] = 0; \qquad E\left[u_i\right] = 0;$$

b) Indiqueu la relació entre les dues expressions següents:

$$E\left[u_i^2\right]=\sigma^2; \qquad \hat{\sigma}^2=\frac{\sum \hat{u}_i^2}{n-k}$$

Exercici 2.12 Responeu a les següents preguntes:

a) Definiu les propietats probabilístiques dels estimadors de *MQO* sota els supòsits del *MLC*. Raoneu la resposta.
b) Què passa amb l'estimació del model de regressió lineal si la variança mostral de la variable explicativa és nul·la? Raoneu la resposta.

Exercici 2.13 Un investigador considera que la relació entre consum (*cons*) i renda disponible (*renda*) ha de ser estrictament proporcional. Per això, planteja el següent model

$$cons=\beta_2 renta+u$$

a) Dedueix la fórmula per estimar β_2.
b) Dedueix la fórmula per estimar σ^2.
c) En aquest model, a què és igual $\sum_{i=1}^{n}\hat{u}_i$?

Exercici 2.14 En el context del model de regressió lineal simple

$$y=\beta_1+\beta_2 x+u$$

a) Quins supòsits s'han de complir perquè els estimadors de mínims quadrats ordinaris siguen no esbiaixats?
b) Què supòsits es requereixen perquè la seva variança siga mínima dins del conjunt d'estimadors lineals i no esbiaixats?

Exercici 2.15 En llenguatge estadístic se solen fer en moltes ocasions afirmacions com la següent:

"Siga una mostra aleatòria de grandària n extreta d'una variable x amb distribució normal $N(\alpha,\sigma)$".

a) Expresse l'afirmació anterior amb llenguatge economètric, introduint un terme de pertorbació.
b) Deduïsca la fórmula per estimar α.
c) Deduïsca la fórmula per estimar σ^2.
d) En aquest model, a què seria igual? $\sum_{i=1}^{n}\hat{u}_i$?

Exercici 2.16 Siga el següent model que relaciona la despesa en educació (*geduc*) amb la renda disponible (*renda*):

$$geduc=\beta_1+\beta_2 renta+u$$

Utilitzant la informació obtinguda d'una mostra de 10 famílies s'han obtingut els següents resultats:

$\overline{geduc} = 7 \quad \overline{renda} = 50 \quad \sum_{i=1}^{10} renda_i^2 = 30.650 \quad \sum_{i=1}^{10} geduc_i^2 = 622 \quad \sum_{i=1}^{10} renda_i \times geduc_i = 4.34$

a) Estimeu β_1 i β_2 per *MQO*.
b) Estimeu l'elasticitat despesa en educació/renda per a la mitjana de les famílies de la mostra.
c) Descomponga la variança total de la despesa en educació de la mostra en variança explicada i variança residual.
d) Calculeu el coeficient de determinació.
e) Estimeu la variança de les pertorbacions

Exercici 2.17 Donat el model poblacional

$$y_i=3+2x_i+u_i \qquad i= 1, 2, 3$$

i sent els valors de $x_i=\{1,2,3\}$:

a) Genere 15 mostres de u_1, u_2 i u_3, i obtinga els corresponents valors de *y*, utilitzant els números aleatoris *N* (0,1).
b) Feu les corresponents estimacions de β_1 i β_2 en el model:

$$y = \beta_1 + \beta_2 x + u$$

c) Compare les mitjanes i variances mostrals de $\hat{\beta}_1$ i $\hat{\beta}_2$ amb les seues esperances i variances poblacionals.

Exercici 2.18 Basant-se en la informació subministrada en l'exercici 2.17, i amb les diferents estimacions de β_1 i β_2 obtingudes:

a) Calculeu els residus corresponents a cadascuna de les estimacions.
b) Expliqueu el motiu pel qual els residus adopten sempre la forma

$$\hat{u}_1 = -\hat{u}_2$$
$$\hat{u}_3 = 0$$

Exercici 2.19 El següent model es va formular per explicar el temps dedicat a dormir (*sleep*) en funció del temps dedicat al treball remunerat (*paidwork*):

$$sleep = \beta_1 + \beta_2 paidwork + u$$

on el *sleep* i la *paidwork* es mesuren en minuts per dia.

Usant una sub-mostra aleatòria, extreta de l'arxiu *timuse03*, van ser obtinguts els següents resultats:

$$sleep_i = 550.17 - 0.1783 paidwork$$

$$R^2= 0.2539 \qquad n=62$$

a) Interpreteu el coeficient de *paidwork*.
b) Quin és l'increment previst de son, de mitjana, si el temps dedicat al treball remunerat disminueix en una hora per dia?
c) Que part de la variació en el somni s'explica pel temps dedicat a treball remunerat?

Exercici 2.20 La quantificació de la felicitat no és una tasca fàcil. Els investigadors de l'Enquesta Mundial de Gallup van investigar sobre aquest tema mitjançant enquestes a milers de participants en 155 països, entre 2006 i 2009, per tal de mesurar dos tipus de benestar. Es va preguntar als enquestats sobre la satisfacció general en la seva vida, utilitzant una escala de puntuació d'1 a 10. Per explicar la satisfacció general (*stsfglo*) es va formular el següent model en què cada observació es refereix a les mitjanes obtingudes en els diferents països:

$$stsfglo = \beta_1 + \beta_2 lifexpec + u$$

on *lifexpec* és l'esperança de vida en néixer, és a dir, el nombre d'anys que s'espera que visca un nadó.

Utilitzant l'arxiu *HDR2010*, s'obté el següent model ajustat:

$$stsfglo = -1.499 + 0.1062 lifexpec$$

$$R^2 = 0.6135 \qquad n = 144$$

a) Interpreteu el coeficient de *lifexpec*.
b) Quina seria la mitjana de satisfacció global en un país amb una esperança de vida en néixer de 80 anys?
c) Quin ha de ser l'esperança de vida en néixer per obtenir una satisfacció global igual a 6?

Exercici 2.21 En economia es denomina intensitat en l'activitat en investigació i desenvolupament, o simplement R+D, a la relació entre la inversió d'una empresa en investigació i desenvolupament i les vendes d'aquesta empresa.

Per a l'estimació un model que explique la intensitat en R+D és necessari comptar amb una base de dades apropiada. A Espanya es pot utilitzar l'Enquesta sobre Estratègies Empresarials realitzada pel Ministeri d'Indústria. Aquesta enquesta, amb periodicitat anual, proporciona un profund coneixement de l'evolució del sector industrial a través del temps, ja que ofereix múltiples dades relatives al desenvolupament empresarial i a les decisions de l'empresa. Aquesta enquesta també està dissenyada per a generar informació microeconòmica que permet especificar i contrastar models economètrics. Quant a la seva cobertura, la població de referència d'aquesta enquesta són empreses amb deu o més treballadors de la indústria manufacturera. L'àrea geogràfica de referència és Espanya, i les dades són anuals. Una de les característiques més destacades d'aquesta enquesta és el seu alt grau de representativitat.

Utilitzant el fitxer *rdspain*, que és una base de dades de les empreses espanyoles des de 1983 a 2006, es va estimar la següent equació per explicar les despeses en recerca i desenvolupament (*rdintens*):

$$rdintens = -2.639 + 0.2123 \ln(sales)$$

$$R^2 = 0.0350 \qquad n = 1983$$

on *rdintens* s'expressa com un percentatge de les vendes, i les vendes es mesuren en milions d'euros.

a) Interprete el coeficient de ln (*sales*).

b) Si les vendes augmenten en un 50%, quin és el canvi estimat en punts percentuals de *rdintens*?
c) Quin percentatge de la variació de *rdintens* s'explica per les vendes? És elevat? Justifiqueu la resposta.

Exercici 2.22 El següent model es va formular per explicar el salari d'un graduat MBA (*salMBAgr*) en funció de les taxes de matrícula (*tuition*)

$$salMBAgr = \beta_1 + \beta_2 tuition + u$$

on *salMBApr* és el salari mitjà anual en dòlars per als estudiants matriculats a l'any 2010 de les 50 millors escoles de negocis americanes i *tuition* són els drets de matrícula, incloent totes les despeses necessàries per al programa complet (amb exclusió de les despeses de subsistència).

Utilitzant les dades de *MBAtui10*, es va obtenir el següent model ajustat:

$$\widehat{salMBAgr}_i = 54242 + 0.4313 tuition_i$$

$$R^2=0.4275 \qquad n=50$$

a) Quina és la interpretació del terme independent?
b) Quina és la interpretació del coeficient del pendent?
c) Quin és el valor predit de *salMBAgr* per a un estudiant de postgrau que va pagar 110000 dòlars pels drets de matrícula en un MBA de 2 anys?

Exercici 2.23 Usant una submostra de l'Enquesta Estructural de Salaris per a Espanya en 2006 (*wage06sp*), es va estimar el següent model per explicar els salaris:

$$\widehat{\ln(wage)} = 1.919 + 0.0527 educ$$

$$R^2=0.2445 \qquad n=50$$

on *educ* (educació) es mesura en anys i el salari (*wage*) en euros per hora.

a) Quina és la interpretació del coeficient *educ*?
b) Quants anys d'educació més es requereixen per obtenir un salari un 10% més elevat?
c) Sabent que $\overline{educ} = 10.2$, calculi l'elasticitat salari/educació.

Exercici 2.24 Utilitzant dades de l'economia espanyola per al període 1954-2010 (fitxer *consump*), es va estimar la funció de consum keynesiana:

$$\widehat{conspc}_t = -288 + 0.9416 incpc_t$$

$$R^2=0.994 \qquad n=57$$

on el consum (*conspc*) i la renda disponible (*incpc*) s'expressen en euros constants per càpita, prenent 2008 com a any de referència.

a) Quina és la interpretació del terme independent? Opini sobre el signe i magnitud del terme independent.

b) Interpreteu el coeficient d'*incpc*. Quin és el significat econòmic d'aquest coeficient?

c) Compari la propensió marginal a consumir amb la propensió mitjana al consum per al punt de la mitjana mostral $(\overline{conspc} = 8084,\ \overline{incpc} = 8896)$. Comenteu el resultat obtingut.

d) Calculeu l'elasticitat consum/renda per a la mitjana mostral.

Annex 2.1 Un cas d'estudi: corbes d'Engel per a la demanda de productes lactis

La corba d'Engel mostra la relació entre les diverses quantitats d'un bé que el consumidor està disposat a comprar per a diferents nivells de renda.

En una enquesta realitzada a 40 famílies s'han obtingut dades de despesa anual en productes lactis i de renda disponible que apareixen en el quadre 2.6. Per evitar distorsions degudes a la diferent grandària de les llars, tant el consum com la renda s'han expressat en termes per càpita. Les dades són expressades en milers d'euros al mes.

Abans de procedir a la seua estimació amb les dades del quadre 2.6, anem exposar diversos tipus de models que s'utilitzen en els estudis de demanda, analitzant les propietats de cada un d'ells. Els models que es van examinar són els següents: lineal, invers, semilogarítmic, potencial, exponencial i exponencial invers. En els tres primers models, el regressant de l'equació a estimar és directament la variable endògena, mentre que en els tres últims, després de realitzar les transformacions adequades, el regressant és el logaritme neperià de la variable endògena.

En tots els models es calcularà la propensió marginal, així com l'elasticitat de la demanda.

QUADRE 2.6 Despesa en productes lactis (*dairy*), renda disponible (*inc*) en termes *per càpita*. (Unitat: euros per mes)

família	*dairy*	*inc*	*família*	*dairy*	*inc*
1	8.87	1.250	21	16.20	2.100
2	6.59	985	22	10.39	1.470
3	11.46	2.175	23	13.50	1.225
4	15.07	1.025	24	8.50	1.380
5	15.60	1.690	25	19.77	2.450
6	6.71	670	26	9.69	910
7	10.02	1.600	27	7.90	690
8	7.41	940	28	10.15	1.450
9	11.52	1.730	29	13.82	2.275
10	7.47	640	30	13.74	1.620
11	6.73	860	31	4.91	740
12	8.05	960	32	20.99	1.125
13	11.03	1.575	33	20.06	1.335
14	10.11	1.230	34	18.93	2.875
15	18.65	2.190	35	13.19	1.680
16	10.30	1.580	36	5.86	870
17	15.30	2.300	37	7.43	1.620
18	13.75	1.720	38	7.15	960
19	11.49	850	39	9.10	1.125
20	6.69	780	40	15.31	1.875

Model lineal

El model lineal de la demanda de productes lactis és el següent:

$$dairy = \beta_1 + \beta_2 inc + u \tag{2-73}$$

Com sabem la propensió marginal de la despesa ens indica com canvia la despesa en variar la renda, i s'obté derivant la despesa pel que fa a la renda en l'equació de demanda. En el model lineal la propensió marginal de la despesa en productes lactis ve donada per

$$\frac{d\ dairy}{d\ inc} = \beta_2 \tag{2-74}$$

És a dir, en el model lineal la propensió marginal es manté constant i, per tant, és independent del valor que prenga la renda. El fet que siga constant és un avantatge, però al mateix temps té l'inconvenient que pot no ser adequada per descriure el comportament dels consumidors, especialment quan hi hagi diferències importants en la renda de les famílies analitzades. Així, no sembla plausible que una família amb uns ingressos mensuals de 700 euros dedique al consum de productes lactis de cada euro addicional de què dispose una proporció igual que la que dedicaria una família amb ingressos de 20000 euros. Ara bé, si la variació de la renda no és molt elevada un model lineal pot ser adequat per descriure la demanda de certs béns.

La propensió marginal mesura el canvi absolut que es produeix en la despesa en variar la renda. En moltes ocasions, però, l'investigador està més interessat a conèixer quina és la taxa de variació de la despesa davant d'una variació de la renda mesura en percentatge. Així, en aquest cas en concret l'investigador pot tenir un especial interès, per exemple, en conèixer el percentatge de variació de la despesa en productes lactis en incrementar la renda en un 1%. Aquest tipus d'aproximació requereix que es calcule l'elasticitat despesa/renda.

En termes matemàtics, l'elasticitat despesa/renda ve donada per

$$\varepsilon_{lacteos/rendis}^{linear} = \frac{d\ dairy}{d\ inc} inc = \beta_2 \frac{inc}{dairy} \tag{2-75}$$

Estimant el model (2-73) amb les dades del quadre 2.6, obtenim

$$dairy = 4.012 + 0.005288 \times inc \qquad R^2 = 0.4584 \tag{2-76}$$

Model invers

En el model invers s'estableix una relació lineal entre la despesa i la inversa de la renda. Per tant, aquest model és directament lineal en els paràmetres. La seva expressió és la següent:

$$dairy = \beta_1 + \beta_2 \frac{1}{inc} + u \tag{2-77}$$

El signe del coeficient β_2 serà negatiu en el cas normal que la renda estiga correlacionada positivament amb la despesa en el bé. Com es pot comprovar fàcilment, quan la renda tendeix cap infinit, la despesa tendeix a un límit que és igual a β_1. És a dir, β_1 representa el màxim consum que pot haver d'aquest bé.

A la figura 2.14 es pot veure la representació de la part sistemàtica d'aquest model. A la primera figura s'ha representat la relació entre la variable dependent i la variable

explicativa. A la segona s'ha representat la relació entre el regressant i el regressor. La segona funció és lineal com es pot veure a la figura.

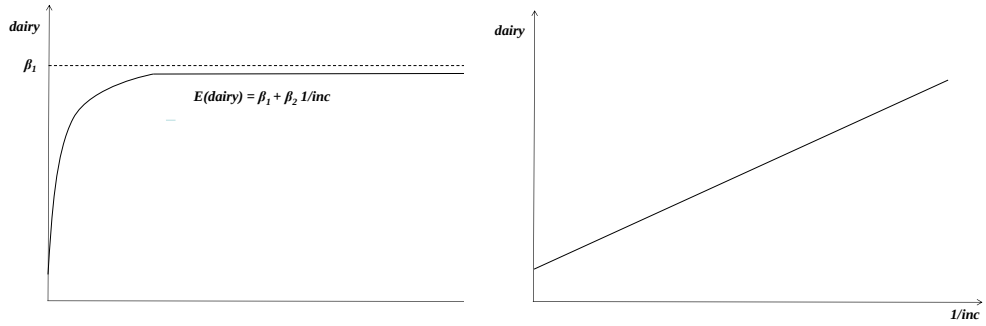


FIGURA 2.14. El model invers.

En el model invers la propensió marginal de la despesa ve donada per

$$\frac{d\ dairy}{d\ inc} = -\beta_2 \frac{1}{(inc)^2} \tag{2-78}$$

D'acord amb (2-78), la propensió marginal al consum va disminuint de forma inversament proporcional al quadrat del nivell de renda.

D'altra banda, l'elasticitat disminueix, segons es pot veure en la següent expressió, de forma inversament proporcional al producte de la despesa per la renda:

$$\varepsilon^{inv}_{dairy/inc} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = -\beta_2 \frac{1}{inc \times dairy} \tag{2-79}$$

Estimant el model (2-77) amb les dades del quadre 2.6, s'obté

$$dairy = 18.652 - 8702\frac{1}{inc} \qquad R^2 = 0.4281 \tag{2-80}$$

En aquest cas, el coeficient β_2 no té un significat econòmic.

Model lineal logarítmic

Aquest model rep la denominació de lineal logarítmic per ser la despesa una funció lineal del logaritme de la renda, és a dir,

$$dairy = \beta_1 + \beta_2 \ln(inc) + u \tag{2-81}$$

En aquest model, la propensió marginal a la despesa ve donada per

$$\frac{d\ dairy}{d\ inc} = \frac{d\ dairy}{d\ inc}\frac{inc}{inc} = \frac{d\ dairy}{d\ \ln(inc)}\frac{1}{inc} = \beta_2 \frac{1}{inc} \tag{2-82}$$

i l'elasticitat despesa/renda ve donada per

$$\varepsilon^{lin\text{-}\log}_{dairy/inc} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = \frac{d\ dairy}{d\ \ln(inc)}\frac{1}{dairy} = \beta_2 \frac{1}{dairy} \tag{2-83}$$

La propensió marginal és inversament proporcional al nivell de renda en el model lineal logarítmic, mentre que l'elasticitat és inversament proporcional al nivell de despesa en productes lactis.

A la figura 2.15, podem veure una doble representació de la funció poblacional corresponent a aquest model.

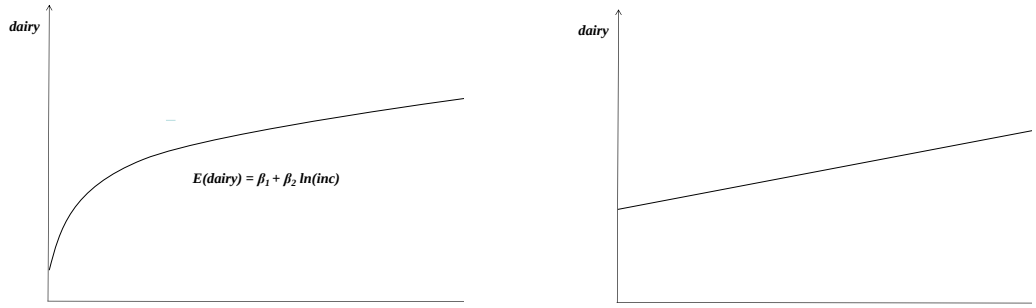


FIGURA 2.15. El model lineal logarítmic.

Estimant el model (2-81) amb les dades del quadre 2.6, s'obté

$$dairy = -41.623 + 7.399 \times \ln(inc) \quad R^2 = 0.4567 \tag{2-84}$$

La interpretació de $\hat{\beta}_2$ és la següent: si la renda augmenta en un 1%, la demanda de productes lactis s'incrementarà en 0,07399 euros.

Model potencial o doblement logarítmic

El model potencial es defineix de la manera:

$$dairy = e^{\beta_1} inc^{\beta_2} e^{u} \tag{2-85}$$

Aquest model no és lineal en els paràmetres, però és linealitzable, ja que en prendre logaritmes neperians s'obté el model:

$$\ln(dairy) = \beta_1 + \beta_2 \ln(inc) + u \tag{2-86}$$

A aquest model se l'anomena també doblement logarítmic, ja que aquesta és l'estructura de la versió linealitzada.

En el model potencial la propensió marginal de la demanda ve donada per

$$\frac{d\ dairy}{d\ inc} = \beta_2 \frac{dairy}{inc} \tag{2-87}$$

En el model potencial l'elasticitat és constant. Per tant, davant d'una variació donada de la renda, la despesa s'incrementarà en el mateix percentatge amb independència de quin siga el nivell de renda i despesa a què s'aplique. L'expressió de l'elasticitat és la següent:

$$\varepsilon_{dairy/inc}^{log\text{-}log} = \frac{d\ dairy}{d\ inc} \frac{inc}{dairy} = \frac{d\ \ln(dairy)}{d\ \ln(inc)} = \beta_2 \tag{2-88}$$

A la figura 2.16 es pot veure una doble representació de la funció poblacional corresponent a aquest model.

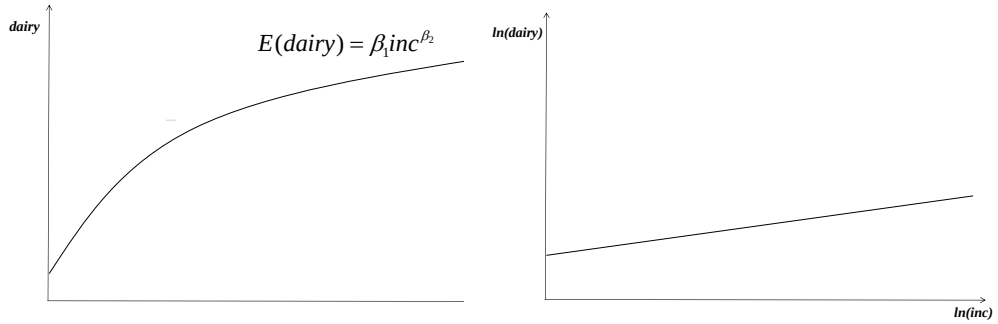


FIGURA 2.16. Model doblement logarítmic

Estimant el model (2-86) amb les dades del quadre 2.6, s'obté

$$\ln(dairy) = -2.556 + 0.6866 \times \ln(inc) \quad R^2 = 0.5190 \tag{2-89}$$

En aquest cas $\hat{\beta}_2$ és l'elasticitat de la despesa/renda. La seva interpretació és la següent: si l'ingrés augmenta en un 1%, la demanda de productes lactis s'incrementarà en un 0,68%.

Model exponencial

El model exponencial es defineix de la manera:

$$dairy = \exp(\beta_1 + \beta_2 inc + u) \tag{2-90}$$

Prenent logaritmes neperians en els dos membres de (2-90), s'obté el següent model que és lineal en els paràmetres:

$$\ln(dairy) = \beta_1 + \beta_2 inc + u \tag{2-91}$$

En el model exponencial la propensió marginal de la despesa ve donada per

$$\frac{d\ dairy}{d\ inc} = \beta_2 dairy \tag{2-92}$$

En el model exponencial, a diferència d'altres models vistos anteriorment, la propensió marginal augmenta quan el nivell de despesa ho fa. Per aquesta raó, aquest model és adequat per descriure la demanda de productes de luxe. D'altra banda, l'elasticitat és proporcional al nivell de renda:

$$\varepsilon^{exp}_{dairy/inc} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = \frac{d\ \ln(dairy)}{d\ inc} inc = \beta_2 inc \tag{2-93}$$

A la figura 2.17, podem veure una doble representació de la funció poblacional corresponent a aquest model.

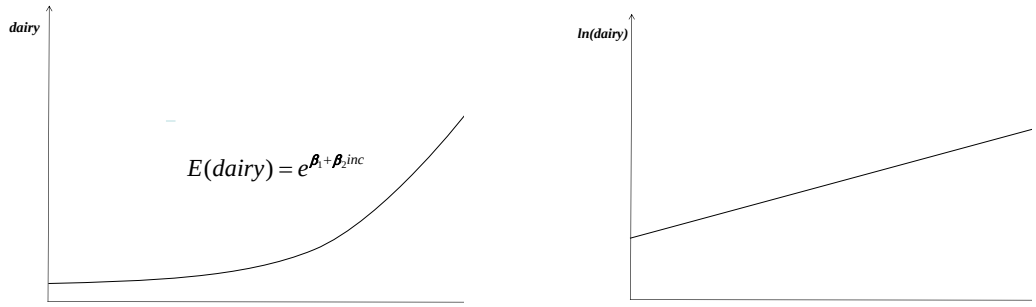


FIGURA 2.17. El model exponencial.

Estimant el model (2-91) amb les dades del quadre 2.6 s'obté

$$\ln(dairy) = 1.694 + 0.00048 \times inc \quad R^2 = 0.4978 \tag{2-94}$$

La interpretació de $\hat{\beta}_2$ és la següent: si la renda s'incrementa en 1 euro la demanda de productes lactis s'incrementarà en un 0,048%.

Model exponencial invers

El model exponencial invers és una barreja del model exponencial i del model invers, tenint propietats que el fan adequat per a determinar la demanda de productes en els quals hi ha un punt de saturació. La seva expressió és la següent:

$$dairy = \exp(\beta_1 + \beta_2 \frac{1}{inc} + u) \tag{2-95}$$

Prenent logaritmes neperians en els dos membres de (2-95) s'obté el següent model que és lineal en els paràmetres:

$$\ln(dairy) = \beta_1 + \beta_2 \frac{1}{inc} + u \tag{2-96}$$

En el model exponencial invers la propensió marginal de la despesa ve donada per

$$\frac{d\ dairy}{d\ inc} = -\beta_2 \frac{dairy}{(inc)^2} \tag{2-97}$$

y l'elasticitat per

$$\varepsilon_{dairy/inc}^{invexp} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = \frac{d\ \ln(dairy)}{d\ inc} inc = -\beta_2 \frac{1}{inc} \tag{2-98}$$

Estimant el model (2-96) amb les dades de la taula 2.6 s'obté

$$\ln(dairy) = 3.049 - 822.02 \frac{1}{inc} \quad R^2 = 0.5040 \tag{2-99}$$

En aquest cas, com en el model invers, el coeficient $\hat{\beta}_2$ no té un significat econòmic.

Al quadre 2.7, es mostren els resultats de la propensió marginal, l'elasticitat de la despesa/renda i el R^2 en els sis models ajustats.

QUADRE 2.7. Propensió marginal, elasticitat despesa/renda i R^2 en els models estimats per analitzar la demanda de productes lactis.

Model	*Propensió marginal*	*Elasticitat*	R^2
Lineal	$\hat{\beta}_2 = 0.0053$	$\hat{\beta}_2 \frac{\overline{inc}}{\overline{dairy}} = 0.6505$	0.4440
Invers	$-\hat{\beta}_2 \frac{1}{\left[\overline{inc}\right]^2} = 0.0044$	$-\hat{\beta}_2 \frac{1}{\overline{dairy} \times \overline{inc}} = 0.5361$	0.4279
Lineal logarítmic	$\hat{\beta}_2 \frac{1}{\overline{inc}} = 0.0052$	$\hat{\beta}_2 \frac{1}{\overline{dairy}} = 0.6441$	0.4566
Doblement logarítmic	$\hat{\beta}_2 \frac{\overline{dairy}}{\overline{inc}} = 0.0056$	$\hat{\beta}_2 = 0.6864$	0.5188
Logarítmic lineal	$\hat{\beta}_2 \times \overline{dairy} = 0.0055$	$\hat{\beta}_2 \times \overline{inc} = 0.6783$	0.4976
Logarítmic invers	$-\hat{\beta}_2 \frac{\overline{dairy}}{\left[\overline{inc}\right]^2} = 0.0047$	$-\hat{\beta}_2 \frac{1}{\overline{inc}} = 0.5815$	0.5038

El R^2 obtingut en els tres primers models no és comparable amb el R^2 obtingut en els tres últims perquè la forma funcional del regressant és diferent: *y* en els tres primers models i ln (*y*) en els tres últims.

Comparant els tres primers models entre si, el millor ajust s'obté amb el model lineal logarítmic si utilitzem R^2 com a mesura de bondat d'ajust. Comparant els tres últims models el millor ajust correspon al model doblement logarítmic. Si s'hagués utilitzat el Criteri d'Informació d'Akaike (*AIC*), que permet comparar els models amb diferents formes funcionals per al regressant, aleshores el model doblement logarítmic hauria estat el millor entre els sis models estimats. La mesura AIC serà estudiada en el capítol 3.

Apèndixs

Apèndix 2.1: Dues formes alternatives d'expressar $\hat{\beta}_2$

És fàcil veure que

$$\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x}) = \sum_{i=1}^{n}(y_i x_i - \bar{x}y_i - \bar{y}x_i + \overline{yx}) = \sum_{i=1}^{n} y_i x_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\overline{yx}$$

$$= \sum_{i=1}^{n} y_i x_i - n\overline{xy} - \bar{y}\sum_{i=1}^{n} x_i + n\overline{yx} = \sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i$$

D'altra banda, obtenim

$$\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}(x_i^2-2\overline{x}x_i+\overline{x}\overline{x})^2=\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{x}\overline{x}$$

$$=\sum_{i=1}^{n}x_i^2-2n\overline{x}^2+n\overline{x}^2=\sum_{i=1}^{n}x_i^2-\overline{x}\sum_{i=1}^{n}x_i$$

Per tant, (2-28) es pot expressar de la següent manera:

$$\hat{\beta}_2=\frac{\sum_{i=1}^{n}y_ix_i-\overline{y}\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2-\overline{x}\sum_{i=1}^{n}x_i}=\frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

Apèndix 2.2. Demostració que $r_{xy}^2=R^2$

En primer, lloc estudiarem una equivalència que es va a utilitzar en la demostració. Per definició,

$$\hat{y}_i=\hat{\beta}_1+\hat{\beta}_2x_i$$

De la primera equació normal, obtenim

$$\overline{y}=\hat{\beta}_1+\hat{\beta}_2\overline{x}$$

Restant la segona equació de la primera:

$$\hat{y}_i-\overline{y}=\hat{\beta}_2(x_i-\overline{x})$$

Elevant al quadrat els dos membres

$$(\hat{y}_i-\overline{y})^2=\hat{\beta}_2^2(x_i-\overline{x})^2$$

i sumant per a tot *i*, tenim

$$\sum(\hat{y}_i-\overline{y})^2=\hat{\beta}_2^2\sum(x_i-\overline{x})^2$$

Tenint en compte l'anterior equivalència, obtenim

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{\hat{y}})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=\frac{\hat{\beta}_2^2\sum_{i=1}^{n}(x_i-\overline{x})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=\frac{\left[\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})\right]^2}{\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^2}\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

$$=\frac{\left[\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})\right]^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\frac{1}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=r_{xy}^2$$

Apèndix 2.3. Canvi proporcional versus canvi en logaritmes

El canvi en logaritmes és una taxa de variació, que s'utilitza en la investigació econòmica. La relació entre el canvi proporcional i el canvi en logaritmes es pot veure si s'aplica un desenvolupament en sèrie de Taylor a (2-45):

$$\ln(x_1)-\ln(x_0)=\ln\left[\frac{x_1}{x_0}\right]$$

$$=\ln(1)+\left[\frac{x_1}{x_0}-1\right]\left[\frac{1}{\frac{x_1}{x_0}}\right]_{\frac{x_1}{x_0}=1}+\frac{1}{2}\left[\frac{x_1}{x_0}-1\right]^2\left[-\frac{1}{\left[\frac{x_1}{x_0}\right]}\right]_{\frac{x_1}{x_0}=1}$$

$$+\frac{1}{3\times 2}\left[\frac{x_1}{x_0}-1\right]^3\left[\frac{2}{\left[\frac{x_1}{x_0}\right]^3}\right]_{\frac{x_1}{x_0}=1}+\cdots\cdots \qquad (2\text{-}100)$$

$$=\left[\frac{x_1}{x_0}-1\right]-\frac{1}{2}\left[\frac{x_1}{x_0}-1\right]^2+\frac{1}{3}\left[\frac{x_1}{x_0}-1\right]^3+\cdots$$

$$=\frac{\Delta x_1}{x_0}-\frac{1}{2}\left[\frac{\Delta x_1}{x_0}\right]^2+\frac{1}{3}\left[\frac{\Delta x_1}{x_0}\right]^3+\cdots$$

Per tant, si prenem l'aproximació lineal en aquest desenvolupament, obtenim

$$\Delta\ln(x)=\ln(x_1)-\ln(x_0)=\ln\left[\frac{x_1}{x_0}\right]\approx\frac{\Delta x_1}{x_0} \qquad (2\text{-}101)$$

Apèndix 2.4. Demostració que els estimadors *MQO* són lineals i no esbiaixats

Només demostrarem l'absència de biaix de l'estimador $\hat{\beta}_2$ que és el més rellevant. Per a demostrar-ho, hem d'expressar el nostre estimador en termes del paràmetre poblacional. La fórmula (2-18) es pot expressar com

$$\hat{\beta}_2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (2\text{-}102)$$

ja que $\sum_{i=1}^{n}(x_i-\bar{x})\bar{y}=\bar{y}\sum_{i=1}^{n}(x_i-\bar{x})=\bar{y}\times 0=0$

Ara anem a expressar (2-102) de la següent manera:

$$\hat{\beta}_2=\sum_{i=1}^{n}c_i y_i \qquad (2\text{-}103)$$

on

$$c_i=\frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (2\text{-}104)$$

Els coeficients c_i tenen les següents propietats

$$\sum_{i=1}^{n} c_i = 0 \tag{2-105}$$

$$\sum_{i=1}^{n} c_i^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^2} = \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2-106}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n}(x_i - \overline{x})x_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 1 \tag{2-107}$$

Ara, si substituïm $y = \beta_1 + \beta_2 x + u$ (supòsit 1) a (2-102), obtenim

$$\begin{aligned} \hat{\beta}_2 &= \sum_{i=1}^{n} c_i y_i = \sum_{i=1}^{n} c_i(\beta_1 + \beta_2 x_i + u_i) \\ &= \beta_1 \sum_{i=1}^{n} c_i + \beta_2 \sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n} c_i u_i = \beta_2 + \sum_{i=1}^{n} c_i u_i \end{aligned} \tag{2-108}$$

Assumint que els regressors són no estocàstics (supòsit 2), c_i serà també no estocàstic. Per tant, $\hat{\beta}_2$ és un estimador que és funció lineal de *u*.

Prenent esperances en (2-108) i tenint en compte el supòsit 6, i implícitament els supòsits del 3 al 5, s'obté

$$E(\hat{\beta}_2) = \beta_2 + \sum_{i=1}^{n} c_i E(u_i) = \beta_2 \tag{2-109}$$

Per tant, $\hat{\beta}_2$ és un estimador no esbiaixat de β_2

Apèndix 2.5. Càlcul de la variança de $\hat{\beta}_2$:

$$\begin{aligned} E\left[\hat{\beta}_2 - \beta_2\right]^2 &= \left[\sum_{i=1}^{n} c_i u_i\right]^2 = \sum_{i=1}^{n} c_i^2 E(u_i^2) + \sum_{i \neq j}\sum_{i=1}^{n} c_i c_j E(u_i u_j) \\ &= \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sigma^2}{nS_X^2} \end{aligned} \tag{2-110}$$

En la demostració anterior, en passar de la segona a la tercera igualtat, s'han tingut en compte els supòsits 6 i 7.

Apèndix 2.6. Demostració del teorema de Gauss-Markov per a la pendent en la regressió simple

El procediment que seguirem per a la demostració és el següent. En primer lloc, definirem un estimador arbitrari, $\tilde{\beta}_2$, que és lineal en *y*. En segon lloc, anem a imposar les restriccions que es requereixen perquè siga no esbiaixat. En tercer lloc, es mostrarà que la variança d'aquest estimador arbitrari ha de ser més gran, o almenys igual, que la variança de $\hat{\beta}_2$.

Així doncs, definirem un estimador arbitrari, $\tilde{\beta}_2$, que és lineal en *y*:

$$\tilde{\beta}_2 = \sum_{i=1}^{n} h_i y_i \tag{2-111}$$

Ara, substituïm y_i pel seu valor en el model poblacional (supòsit 1):

$$\tilde{\beta}_2 = \sum_{i=1}^{n} h_i y_i = \sum_{i=1}^{n} h_i (\beta_1 + \beta_2 x_i + u_i) = \beta_1 \sum_{i=1}^{n} h_i + \beta_2 \sum_{i=1}^{n} h_i x_i + \sum_{i=1}^{n} h_i u_i \qquad (2\text{-}112)$$

Per a que l'estimador $\tilde{\beta}_2$ siga no esbiaixat cal que les restriccions següents es compleixen:

$$\sum_{i=1}^{n} h_i = 0 \qquad \sum_{i=1}^{n} h_i x_i = 1 \qquad (2\text{-}113$$

Per tant,

$$\tilde{\beta}_2 = \beta_2 + \sum_{i=1}^{n} h_i u_i \qquad (2\text{-}114)$$

La variança d'aquest estimador és la següent:

$$E\left[\tilde{\beta}_2 - \beta_2\right]^2 = \left[\sum_{i=1}^{n} h_i u_i\right]^2 = \sigma^2 \sum_{i=1}^{n} h_i^2 =$$

$$\sigma^2 \sum_{i=1}^{n} \left[h_i - \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^2 = \sigma^2 \sum_{i=1}^{n} \left[h_i - \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^2 \qquad (2\text{-}115)$$

$$+\sigma^2 \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^2 + 2\sigma^2 \sum_{i=1}^{n} \left[h_i - \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

El tercer terme de l'última igualtat és 0, com es mostra a continuació:

$$2\sigma^2 \sum_{i=1}^{n} \left[h_i - \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= 2\sigma^2 \sum_{i=1}^{n} \left[h_i \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] - 2\sigma^2 \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] = 2\sigma^2 \times 1 - 2\sigma^2 \times 1 = 0 \qquad (2\text{-}116)$$

Per tant, tenint en compte (2-116) i operant, obtenim

$$E\left[\tilde{\beta}_2 - \beta_2\right]^2 = \sigma^2 \sum_{i=1}^{n} \left[h_i - c_i\right]^2 + \sigma^2 \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (2\text{-}117)$$

on $c_i = \dfrac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

El segon terme de l'última igualtat és la variança de $\hat{\beta}_2$, mentre que el primer terme és sempre positiu, ja que és una suma de quadrats, excepte que es compleix que $h_i=c_i$, per a tot i, en aquest cas serà igual a 0, i aleshores $\tilde{\beta}_2 = \hat{\beta}_2$. Així doncs,

$$E\left[\tilde{\beta}_2-\beta_2\right]^2 \geq E\left[\hat{\beta}_2-\beta_2\right]^2 \tag{2-118}$$

Apèndix 2.7. Demostració que $\hat{\sigma}^2$ és un estimador no esbiaixat de la variança de les pertorbacions

El model poblacional és, per definició:

$$y_i=\beta_1+\beta_2 x_i+u_i \tag{2-119}$$

Si sumem els dos membres de (2-130) per a tot *i* i dividim per *n*, tenim

$$\bar{y}=\beta_1+\beta_2\bar{x}+\bar{u} \tag{2-120}$$

Restant (2-120) de (2-119), obtenim

$$y_i-\bar{y}=\beta_2\left(x_i-\bar{x}\right)+\left(u_i-\bar{u}\right) \tag{2-121}$$

D'altra banda, $\hat{u}_i$ és per definició:

$$\hat{u}_i=y_i-\hat{\beta}_1-\hat{\beta}_2 x_i \tag{2-122}$$

Si sumem els dos membres de (2-122) i dividim per *n*, tenim

$$\bar{\hat{u}}=\bar{y}-\hat{\beta}_1-\hat{\beta}_2\bar{x} \tag{2-123}$$

Restant (2-123) de (2-122), i tenint en compte que $\bar{\hat{u}}$ =0,

$$\hat{u}_i=\left(y_i-\bar{y}\right)-\hat{\beta}_2\left(x_1-\bar{x}\right) \tag{2-124}$$

Substituint (2-121) a (2-124), obtenim

$$\begin{aligned}\hat{u}_i&=\beta_2\left(x_i-\bar{x}\right)+\left(u_i-\bar{u}\right)-\hat{\beta}_2\left(x_1-\bar{x}\right)\\&=-\left(\hat{\beta}_2-\beta_2\right)\left(x_1-\bar{x}\right)+\left(u_i-\bar{u}\right)\end{aligned} \tag{2-125}$$

Elevant al quadrat i sumant en els dos membres de (2-125), s'obté que

$$\begin{aligned}\sum_{i=1}^n\hat{u}_i^2=\left[\tilde{\beta}_2-\beta_2\right]^2\sum_{i=1}^n(x_i-\bar{x})^2+\sum_{i=1}^n(u_i-\bar{u})^2\\-2\left[\tilde{\beta}_2-\beta_2\right]\sum_{i=1}^n(x_i-\bar{x})(u_i-\bar{u})\end{aligned} \tag{2-126}$$

Prenent les esperances en (2-126), s'obté que

$$\begin{aligned}E\left[\sum_{i=1}^n\hat{u}_i^2\right]&=\sum_{i=1}^n(x_i-\bar{x})^2E\left[\tilde{\beta}_2-\beta_2\right]^2+E\left[\sum_{i=1}^n(u_i-\bar{u})^2\right]\\&\quad-2E\left[\left(\tilde{\beta}_2-\beta_2\right)\sum_{i=1}^n(x_i-\bar{x})(u_i-\bar{u})\right]\\&=\sum_{i=1}^n(x_i-\bar{x})^2\frac{\sigma^2}{\sum_{i=1}^n(x_i-\bar{x})^2}+\left(n-1\right)\sigma^2-2\sigma^2=\left(n-2\right)\sigma^2\end{aligned} \tag{2-127}$$

Per obtenir el primer terme de l'última igualtat de (2-127), s'ha utilitzat (2-64). Per obtenir el segon i el tercer terme de l'última igualtat de (2-127) s'han utilitzat els

desenvolupaments que es fan a (2-128) i (2-129) respectivament. En tots dos casos s'han tingut en compte els supòsits 7 i 8.

$$E\left[\sum_{i=1}^{n}(u_i-\bar{u})^2\right]=E\left[\sum_{i=1}^{n}u_i^2-n\bar{u}^2\right]=E\left[\sum_{i=1}^{n}u_i^2-n\left(\frac{\sum_{i=1}^{n}u_i}{n}\right)^2\right] \tag{2-128}$$

$$=E\left[\sum_{i=1}^{n}u_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}u_i^2+\sum_{i\neq j}u_iu_j\right)\right]=n\sigma^2-\frac{n}{n}\sigma^2=(n-1)\sigma^2$$

$$E\left[\left(\tilde{\beta}_2-\beta_2\right)\sum_{i=1}^{n}(x_i-\bar{x})(u_i-\bar{u})\right]=E\left[\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}(x_i-\bar{x})u_i\sum_{i=1}^{n}(x_i-\bar{x})u_i\right]$$

$$=\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})E\left(u_i\right)\right]^2 \tag{2-}$$

$$=\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2E\left(u_i\right)^2+\sum_{i\neq j}\sum(x_i-\bar{x})(x_i-\bar{x})E\left(u_iu_j\right)\right]=\sigma^2$$

129)

D'acord amb (2-127), s'obté que

$$E\left[\sum_{i=1}^{n}\hat{u}_i^2\right]=(n-2)\sigma^2 \tag{2-130}$$

Per tant, un estimador no esbiaixat ve donat per

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\hat{u}_i^2}{n-2} \tag{2-131}$$

ja que

$$E\left(\hat{\sigma}^2\right)=\frac{1}{n-2}E\left(\sum_{i=1}^{n}\hat{u}_i^2\right)=\sigma^2 \tag{2-132}$$

Apèndix 2.8. Consistència dels estimadors de *MQO*

L'operador plim té la propietat d'invariança (propietat de Slutsky). És a dir, si $\hat{\theta}$ és un estimador consistent de θ i $g\left(\hat{\theta}\right)$ és qualsevol funció contínua de $\hat{\theta}$, aleshores

$$\operatorname*{plim}_{n\to\infty}g(\hat{\theta})=g(\theta) \tag{2-133}$$

Això vol dir que si $\hat{\theta}$ és un estimador consistent de θ, seguit de $1/\hat{\theta}$ i ln $(\hat{\theta})$ són també estimadors consistents d'$1/\theta$ i ln (θ), respectivament. Cal tenir en compte que aquestes propietats no són vàlides per a l'operador esperança *E*; per exemple, si és un

estimador no esbiaixat de θ [és a dir, E () = θ], no és cert que $1/\theta$ siga un estimador no esbiaixat d'una $1/\theta$, és a dir, $E(1/\hat{\theta}) \neq 1/E(\hat{\theta}) \neq 1/\theta$. Això és a causa del fet que l'operador esperança únicament pot ser aplicat a funcions lineals de variables aleatòries. D'altra banda, l'operador plim és aplicable a qualsevol funció contínua.

Sota els supòsits de l'1 al 6, els estimadors de *MQO*, $\hat{\beta}_1$ i $\hat{\beta}_2$ són consistents.

Ara demostrarem en particular que $\hat{\beta}_2$ és un estimador consistent. En primer lloc, $\hat{\beta}_2$ es pot expressar com:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(\beta_1+\beta_2 x_i+u_i)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$
$$= \frac{\beta_1\sum_{i=1}^{n}(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \beta_2\frac{\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \beta_2 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (2\text{-}134)$$

Per tal de comprovar la seva consistència necessitem prendre plim a (2-134) i aplicar la *Llei dels Grans Nombres*. Aquesta llei estableix que, en condicions generals, els moments mostrals convergeixen als seus corresponents moments poblacionals. Per tant, prenent plim a (2-134):

$$\operatorname*{plim}_{n\to\infty}\hat{\beta}_2 = \operatorname*{plim}_{n\to\infty}\left[\beta_2 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = \beta_2 + \frac{\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (2\text{-}135)$$

En aquesta última igualtat hem dividit el numerador i el denominador per *n*, perquè, si no ho fem, tots dos sumatoris tendiran a infinit quan *n* tendeix a infinit.

Si apliquem la llei de grans nombres al numerador i denominador de (2-134), convergiran en probabilitat a les quantitats poblacionals *cov*(*x*,*u*) i *var*(*x*) respectivament. Sempre que *var*(*x*)≠0 (supòsit 4), podem utilitzar les propietats del *límit de probabilitat* per obtenir

$$\operatorname{plim}\hat{\beta}_2 = \beta_2 + \frac{cov(x,u)}{var(x)} = \beta_2 \quad (2\text{-}136)$$

Per arribar a l'última igualtat, utilitzant els supòsits 2 i 6, obtenim que

$$cov(x,u) = E\left[(x-\bar{x})u\right] = (x-\bar{x})E\left[u\right] = (x-\bar{x})\times 0 = 0 \quad (2\text{-}137)$$

Per tant, $\hat{\beta}_2$ és un estimador consistent.

Apèndix 2.9 Estimació per màxima versemblança

Tenint en compte els supòsits de l'1 al 6 l'esperança de y_i és la següent:

$$E(y_i) = \beta_1 + \beta_2 x_i \quad (2\text{-}138)$$

Si tenim en compte el supòsit 7, la variança de y_i és igual a

$$\operatorname{var}(y_i) = E\left[y_i - E(y_i)\right]^2 = E\left[y_i - \beta_1 + \beta_2 x_i\right]^2 = E\left[u_i\right]^2 = \sigma^2 \quad \forall i \quad (2\text{-}139)$$

D'acord amb el supòsit 1 y_i és una funció lineal de u_i, i si u_i té una distribució normal (supòsit 9), aleshores y_i serà normal i independentment distribuïda (supòsit 8) amb mitjana $\beta_1 + \beta_2 x_i$ i variança σ^2.

Aleshores, la funció de densitat de probabilitat conjunta de $y_1, y_2, \cdots, y_n$ es pot expressar com un producte de *n* funcions de densitat individuals:

$$\begin{aligned} &f\left(y_1,\ y_2,\ \ldots, y_n \mid \beta_1 + \beta_2 x_i,\ \sigma^2\right) \\ &= f\left(y_1 \mid \beta_1 + \beta_2 x_i,\ \sigma^2\right) f\left(y_2 \mid \beta_1 + \beta_2 x_i,\ \sigma^2\right) \cdots f\left(y_n \mid \beta_1 + \beta_2 x_i,\ \sigma^2\right) \end{aligned} \tag{2-140}$$

on

$$f\left(y_i\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{\left[y_i - \beta_1 - \beta_2 x_i\right]^2}{\sigma^2}\right\} \tag{2-141}$$

que és la funció de densitat d'una variable distribuïda normalment amb la mitjana i la variança donada.

Substituint (2-141) a (2-140) per a cada y_i, s'obté

$$\begin{aligned} &f\left(y_1,\ y_2, \cdots,\ y_n\right) = f\left(y_1\right) f\left(y_2\right) \cdots f\left(y_n\right) \\ &= \frac{1}{\sigma^n \left(\sqrt{2\pi}\right)^n} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{\left[y_i - \beta_1 - \beta_2 x_i\right]^2}{\sigma^2}\right\} \end{aligned} \tag{2-142}$$

Si es coneixen $y_1, y_2, \cdots, y_n$, però β_2, β_3, i σ^2 són desconeguts, a la funció en (2-142) s'anomena *funció de versemblança*, i es denota per $L(\beta_2, \beta_3, \sigma^2)$ o simplement *L*. Si es prenen logaritmes a (2-142), s'obté

$$\begin{aligned} \ln L &= -n \ln \sigma - \frac{n}{2}\ln\left(\sqrt{2\pi}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left[y_i - \beta_1 - \beta_2 x_i\right]^2}{\sigma^2} \\ &= -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln\left(\sqrt{2\pi}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left[y_i - \beta_1 - \beta_2 x_i\right]^2}{\sigma^2} \end{aligned} \tag{2-143}$$

El mètode de màxima versemblança (*MV*), com el seu nom suggereix, consisteix en estimar els paràmetres desconeguts de tal manera que la probabilitat d'observar les y_i donades siga tan alta com siga possible. Per tant, tenim per trobar el màxim de la funció (2-143). Per maximitzar (2-143) cal derivar parcialment pel que fa a β_2, β_3, i σ^2 i igualar a 0. Denominant $\tilde{\beta}_1$, $\tilde{\beta}_2$ i $\tilde{\sigma}^2$ als estimadors de *MV*, obtenim que:

$$\begin{aligned} \frac{\partial \ln L}{\partial \tilde{\beta}_1} &= -\frac{1}{\tilde{\sigma}^2}\sum\left(y - \tilde{\beta}_1 - \tilde{\beta}_2 x_i\right)(-1) = 0 \\ \frac{\partial \ln L}{\partial \tilde{\beta}_2} &= -\frac{1}{\tilde{\sigma}^2}\sum\left(y - \tilde{\beta}_1 - \tilde{\beta}_2 x_i\right)(-x_i) = 0 \\ \frac{\partial \ln L}{\partial \tilde{\sigma}^2} &= -\frac{n}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4}\sum\left(y - \tilde{\beta}_1 - \tilde{\beta}_2 x_i\right)^2 = 0 \end{aligned} \tag{2-144}$$

Si prenem les dues primeres equacions de (2-144) i operem, obtenim

$$\sum y_i = n\tilde{\beta}_1 + \tilde{\beta}_2 \sum x_i \tag{2-145}$$

$$\sum y_i x_i = \tilde{\beta}_1 \sum x_i + \tilde{\beta}_2 \sum x_i^2 \tag{2-146}$$

Com es pot veure, (2-145) i (2-146) són iguals a (2-13) i (2-14), és a dir, els estimadors de *MV*, sota els supòsits del *MLC*, són iguals als estimadors de *MQO*.

Substituint $\tilde{\beta}_1$ i $\tilde{\beta}_2$, - obtinguts en resoldre (2-145) i (2-146) - en la tercera equació de (2-144)

$$\tilde{\sigma}^2 = \frac{1}{n}\sum \left(y_i - \tilde{\beta}_1 - \tilde{\beta}_2 x_i \right)^2 = \frac{1}{n}\sum \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \right)^2 = \frac{1}{n}\sum \hat{u}_i^2 \tag{2-147}$$

L'estimador de *MV* de $\tilde{\sigma}^2$ és esbiaixat, ja que, d'acord amb (2-131),

$$E(\tilde{\sigma}^2) = \frac{1}{n} E\left[\sum_{i=1}^{n} \hat{u}_i^2 \right] = \frac{n-2}{n}\sigma^2 \tag{2-148}$$

En qualsevol cas, $\tilde{\sigma}^2$ és un estimador consistent perquè

$$\lim_{n \to 0} \frac{n-2}{n} = 1 \tag{2-149}$$

3 EL MODEL DE REGRESSIÓ LINEAL MÚLTIPLE: ESTIMACIÓ I PROPIETATS

3.1 El model de regressió lineal múltiple

El model de regressió lineal simple no és adequat per a modelitzar molts fenòmens econòmics, ja que per a explicar una variable econòmica es requereix en general tenir en compte més d'un factor. Vegem alguns exemples.

En la funció keynesiana clàssica, el consum es fa dependre de la renda disponible com a única variable rellevant:

$$cons = \beta_1 + \beta_2 renda + u \tag{3-1}$$

No obstant això, hi ha altres factors que poden considerar-se rellevants en el comportament del consumidor. Un d'aquests factors podria ser la riquesa. Amb la inclusió d'aquest factor es tindrà un model amb dues variables explicatives:

$$cons = \beta_1 + \beta_2 renda + \beta_3 riquesa + u \tag{3-2}$$

En l'anàlisi de la producció s'utilitzen sovint les funcions potencials, que amb una especificació adequada poden ser transformades (prenent logaritmes naturals, en aquest cas) en models lineals en els paràmetres. Utilitzant un sol *input* (*treball*), un model per a explicar l'*output* s'especifica de la manera:

$$\ln(output) = \beta_1 + \beta_2 \ln(treball) + u \tag{3-3}$$

El model anterior és clarament insuficient per a l'anàlisi econòmic. Seria millor utilitzar el conegut model de Cobb-Douglas, en què es consideren dos inputs primaris (*treball* i *capital*):

$$\ln(output) = \beta_1 + \beta_2 \ln(treball) + \beta_3 \ln(capital) + u \tag{3-4}$$

D'acord amb la teoria microeconòmica, els costos totals (*costot*) s'expressen com una funció de la quantitat produïda (*cantprod*). Una primera aproximació per a explicar el cost total podria ser un model amb un únic regressor:

$$costot = \beta_1 + \beta_2 cantprod + u \tag{3-5}$$

No obstant això, és molt restrictiu considerar que, com seria el cas del model anterior, el cost marginal es manté constant, independentment de la quantitat produïda. En la teoria econòmica es proposa una funció cúbica, el que condueix al següent model economètric:

$$costot = \beta_1 + \beta_2 quantprod + \beta_3 quantprod^2 + \beta_4 quantprod^3 + u \qquad (3\text{-}6)$$

En aquest cas, a diferència dels anteriors, en el model només hi ha una variable explicativa, però que dóna lloc a tres regressors.

Els salaris es determinen per diferents factors. Un model relativament simple per a explicar els salaris en funció dels anys d'educació i dels anys d'experiència és el següent:

$$salaris = \beta_1 + \beta_2 educ + \beta_3 exper + u \qquad (3\text{-}7))$$

De totes maneres, altres factors importants per a explicar els salaris poden ser variables quantitatives com ara el temps de formació i l'edat, o variables qualitatives, com el sexe, el sector d'activitat, etc.

Finalment, per a explicar les despeses en consum de peix els factors rellevants poden ser el seu preu, el preu d'un producte substitutiu com la carn, i la renda disponible. És a dir:

$$depesapeix = \beta_1 + \beta_2 preupeix + \beta_3 preucarn + \beta_4 renda + u \qquad (3\text{-}8)$$

Per tant, els exemples anteriors han posat de relleu la necessitat d'utilitzar models de regressió múltiple. El tractament economètric del model de regressió lineal simple es va fer utilitzant àlgebra ordinària. El tractament d'un model economètric de dues variables explicatives mitjançant l'ús d'àlgebra ordinària és tediós i molest; d'altra banda, un model amb tres variables explicatives és pràcticament intractable amb aquesta eina. Per aquesta raó, el model de regressió es va a presentar utilitzant àlgebra matricial.

3.1.1 Model de regressió poblacional i funció de regressió poblacional

En el model de regressió lineal múltiple, el regressant-que pot ser la variable endògena o una transformació de les variables endògenes-, és una funció lineal de *k* regressors corresponents a les variables explicatives -o transformacions de les mateixes- i una pertorbació aleatòria o error. El model també inclou un terme independent. Si designem per *y* al regressant, per x_2, x_3,..., x_k als regressors i per *u* a l'error o pertorbació aleatòria, el model poblacional de regressió lineal múltiple vindrà donat per la següent expressió:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u \qquad (3\text{-}9)$$

Els paràmetres $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$ són fixos i desconeguts.

En el segon membre de (3-9) es poden distingir dos components: un component sistemàtic: μ_y

$$\mu_y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \qquad (3\text{-}10)$$

Aquesta equació és coneguda com a funció de regressió poblacional (*FRP*) o hiperplà poblacional. Quan *k*=2, la *FRP* és específicament una línia recta, quan *k*=3, la *FRP* és específicament un pla i, finalment, quan *k*>3, la *FRP* és denominada genèricament hiperplà, que no és susceptible de ser representat físicament.

D'acord amb (3-10), μ_y és una funció lineal en els paràmetres $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$. Ara, suposem que tenim una mostra aleatòria de grandària *n*, $\{(y_i, x_{2i}, x_{31}, \cdots, x_{ki}) : i = 1, 2, \cdots, n\}$, extreta de la població estudiada. Si expressem el model poblacional per a totes les observacions de la mostra, s'obté el següent sistema:

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 x_{21} + \beta_3 x_{31} + \cdots + \beta_k x_{k1} + u_1 \\ y_2 &= \beta_1 + \beta_2 x_{22} + \beta_3 x_{32} + \cdots + \beta_k x_{k2} + u_2 \\ \cdots & \qquad \cdots \qquad\qquad \cdots \quad \cdots \\ y_n &= \beta_1 + \beta_2 x_{2n} + \beta_3 x_{3n} + \cdots + \beta_k x_{kn} + u_n \end{aligned} \tag{3-11}$$

L'anterior sistema d'equacions pot expressar-se d'una forma més compacta usant la notació matricial. Així, anem a anomenar

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{21} & x_{31} & \dots & x_{k1} \\ 1 & x_{22} & x_{32} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & x_{3n} & \dots & x_{kn} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{bmatrix}$$

La matriu **X** és la matriu de regressors. Entre els regressors també s'inclou el regressor corresponent al terme independent. Aquest regressor, que sovint es denomina regressor *fictici*, pren el valor 1 per a totes les observacions.

El model de regressió lineal múltiple (3-11) expressat en notació matricial és el següent:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{21} & x_{31} & \dots & x_{k1} \\ 1 & x_{22} & x_{32} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & x_{3n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \tag{3-12}$$

L'anterior sistema d'equacions pot expressar-se d'una forma més compacta usant la notació matricial. Així, anem a anomenar:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{3-13}$$

on, d'acord amb la notació utilitzada, **y** és un vector $n \times 1$, **X** és una matriu $n \times k$, $\boldsymbol{\beta}$ és un vector $k \times 1$ i **u** és un vector $n \times 1$.

3.1.2 Funció de regressió mostral

La idea bàsica de la regressió consisteix en estimar els paràmetres poblacionals $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$, a partir d'una mostra donada.

La funció de regressió mostral (*FRM*) és la contrapartida de la funció de regressió poblacional (*FRP*). Atès que *FRM* s'obté d'una mostra donada, una nova mostra generarà diferents estimacions.

La *FRM*, que és una estimació de la *FRP*, que ve donada per

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_k x_{ki} \qquad i = 1, 2, \cdots, n \tag{3-14}$$

ens permet calcular el valor *ajustat* ($\hat{y}_i$) corresponent a cada y_i. A la *FRM*, $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \cdots, \hat{\beta}_k$ són els estimadors dels paràmetres $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$.

S'anomena residu a la diferència entre y_i. i $\hat{y}_i$. Això és

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \tag{3-15}$$

En altres paraules, el residu $\hat{u}_i$ és la diferència entre un valor mostral i el seu corresponent valor ajustat.

El sistema d'equacions (3-14) pot expressar-se d'una forma més compacta utilitzant notació matricial. Així, anem a denotar

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \cdots \\ \hat{y}_n \end{bmatrix} \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \qquad \hat{\mathbf{u}} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \cdots \\ \hat{u}_n \end{bmatrix}$$

El model ajustat corresponent, per a totes les observacions de la mostra, serà el següent:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{3-16}$$

El vector dels residus és igual a la diferència entre el vector de valors observats i el vector de valors ajustats, és a dir,

$$\hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{3-17}$$

3.2 Obtenció d'estimacions de mínims quadrats, interpretació dels coeficients, i altres característiques

3.2.1 Obtenció d'estimadors *MQO*

Anomenant *S* a la suma dels quadrats dels residus,

$$S = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right]^2 \tag{3-18}$$

per aplicar el criteri de mínims quadrats en el model de regressió lineal múltiple, calculem la primera derivada de S respecte a cada $\hat{\beta}_j$ en l'expressió (3-18).:

$$\begin{aligned}
&\frac{\partial S}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n} \left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right][-1] \\
&\frac{\partial S}{\partial \hat{\beta}_2} = 2\sum_{i=1}^{n} \left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right][-x_{2i}] \\
&\frac{\partial S}{\partial \hat{\beta}_3} = 2\sum_{i=1}^{n} \left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right][-x_{3i}] \\
&\cdots \qquad \cdots \qquad \cdots \qquad \cdots \\
&\frac{\partial S}{\partial \hat{\beta}_k} = 2\sum_{i=1}^{n} \left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right][-x_{ki}]
\end{aligned} \tag{3-19}$$

Els estimadors de mínims quadràtics s'obtenen en igualar a 0 les derivades anteriors:

$$\begin{aligned}
&\sum_{i=1}^{n} \left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right] = 0 \\
&\sum_{i=1}^{n} \left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right] x_{2i} = 0 \\
&\sum_{i=1}^{n} \left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right] x_{3i} = 0 \\
&\cdots \qquad \cdots \qquad \cdots \qquad \cdots \\
&\sum_{i=1}^{n} \left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right] x_{ki} = 0
\end{aligned} \tag{3-20}$$

o, amb notació matricial,

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{3-21}$$

Al sistema anterior se li denomina genèricament sistema d'equacions normals de l'*hiperplà*.

En notació matricial ampliada, el sistema d'equacions normals és el següent:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{2i} & \dots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{2i}^2 & \cdots & \sum_{i=1}^{n} x_{2i}x_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{2i}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix} \tag{3-22}$$

Cal observar que:

a) a) $\mathbf{X}'\mathbf{X}/n$ és la matriu de moments mostrals de segon ordre, pel que fa a l'origen, dels regressors, entre els quals s'inclou el regressor fictici (x_{1i}) associat al terme independent, que pren el valor x_{1i}=1 per a tot *i*.

b) $\mathbf{X}'\mathbf{y}/n$ és el vector de moments mostrals de segon ordre, pel que fa a l'origen, entre el regressant i els regressors.

En aquest sistema hi ha *k* equacions i *k* incògnites $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \cdots, \hat{\beta}_k)$. En aquest sistema hi ha *k* equacions i *k* incògnites. Aquest sistema es pot resoldre fàcilment utilitzant àlgebra matricial. Per tal de resoldre unívocament el sistema (3-21) pel que fa a $\hat{\boldsymbol{\beta}}$, cal que el rang de la matriu $\mathbf{X}'\mathbf{X}$ siga igual a *k*. Si això es compleix, tots dos membres de (3-21) poden ser premultiplicats per $\left[\mathbf{X}'\mathbf{X}\right]^{-1}$:

$$\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y}$$

obtenint-se l'expressió del vector d'estimadors de mínims quadrats, o més exactament, el vector d'estimadors de mínims quadrats ordinaris (*MQO*), perquè $\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$. Per tant, la solució és la següent:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y} \tag{3-23}$$

Com la matriu de segones derivades, $2\mathbf{X}'\mathbf{X}$, és una matriu definida positiva, la conclusió és que S presenta un mínim en $\hat{\boldsymbol{\beta}}$.

3.2.2 Interpretació dels coeficients

El coeficient $\hat{\beta}_j$ mesura l'efecte parcial del regressor x_i, mantenint els altres regressors fixos. Anem a veure el significat d'aquesta expressió.

El model estimat per a l'observació *i*-èsima ve donat per

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_j x_{ji} + \cdots + \hat{\beta}_k x_{ki} \tag{3-24}$$

Considerem ara el model estimat per a l'observació *h*-èsima, en què els valors de les variables explicatives i, en conseqüència, de *y* hauran canviat respecte a la (3-24):

$$\hat{y}_h = \hat{\beta}_1 + \hat{\beta}_2 x_{2h} + \hat{\beta}_3 x_{3h} + \cdots + \hat{\beta}_j x_{jh} + \cdots + \hat{\beta}_k x_{kh} \tag{3-25}$$

Restant (3-255) de (3-24), obtenim

$$\Delta\hat{y} = \hat{\beta}_2 \Delta x_2 + \hat{\beta}_3 \Delta x_3 + \cdots + \hat{\beta}_j \Delta x_j + \cdots + \hat{\beta}_k \Delta x_k \tag{3-26}$$

On $\Delta\hat{y} = \hat{y}_i - \hat{y}_h, \Delta x_2 = x_{2i} - x_{2h}, \Delta x_3 = x_{3i} - x_{3h}, \cdots \Delta x_k = x_{ki} - x_{kh}$.

L'expressió anterior capta la variació de $\hat{y}$ a canvis en tots els regressors. Si només canvia x_j, haurem de

$$\Delta\hat{y} = \hat{\beta}_j \Delta x_j \tag{3-27}$$

Si x_k s'incrementa en una unitat, tenim

$$\Delta\hat{y} = \hat{\beta}_j \quad \text{for } \Delta x_j = 1 \tag{3-28}$$

En conseqüència, el coeficient $\hat{\beta}_j$ mesura el canvi en *y* quan x_j augmenta en 1 unitat, *mantenint fixos els regressors* $x_2, x_3, \cdots, x_{j-1}, x_{j+1}, \cdots, x_k$. És molt important en la interpretació dels coeficients tenir en compte aquesta clàusula *ceteris paribus*.

Aquesta interpretació no és vàlida, per descomptat, per al terme independent.

EXEMPLE 3.1 Quantificant la influència de l'edat i del salari sobre l'absentisme a l'empresa Buenosaires

Buenosaires és una empresa dedicada a la fabricació de ventiladors, havent tingut resultats relativament acceptables en els últims anys. Els directius consideren, però, que els resultats haurien estat millors si l'absentisme a l'empresa no fos tan alt. Per a aquest propòsit, el model que es proposa és el següent:

$$absent = \beta_1 + \beta_2 age + \beta_3 tenure + \beta_4 wage + u$$

on l'absència, *absent*, es mesura en dies per any, el salari, *wage*, en milers d'euros a l'any; els anys treballats a l'empresa, *tenure*, i l'edat, *age*, s'expressen en anys.

Utilitzant una mostra de mida 48 (fitxer *absent*), s'ha estimat la següent equació:

$$\underset{}{absent} = \underset{(1.603)}{14.413} - \underset{(0.048)}{0.096}\, age - \underset{(0.067)}{0.078}\, tenure - \underset{(0.007)}{0.036}\, wage$$

$$R^2 = 0.694 \qquad n = 48$$

La interpretació de $\hat{\beta}_2$ és la següent: mantenint fix el salari i els anys treballats a l'empresa, si l'edat s'incrementa en un any, l'absentisme laboral es reduirà en 0.096 dies a l'any. La interpretació de $\hat{\beta}_3$ és com segueix: mantenint fix el salari i l'edat, si els anys treballats a l'empresa s'incrementen en un any, l'absentisme laboral es reduirà en 0.078 dies a l'any. Finalment, la interpretació de $\hat{\beta}_4$ és la següent: mantenint fixa l'edat i els anys treballats a l'empresa, si el salari s'incrementa en 1.000 euros a l'any, l'absentisme laboral es reduirà en 0.036 dies per any.

EXEMPLE 3.2 Demanda de serveis hotelers

Per a explicar la demanda de serveis hotelers es va formular el següent model:

$$\ln\ hostel\ = \beta_1 + \beta_2 \ln(inc) + \beta_3 hhsize + u \tag{3-29}$$

on *hostel* és la despesa en serveis hotelers i *inc* és la renda disponible; ambdues variables estan expressades en euros per mes. La variable *hhize* és el nombre de membres de la llar.

L'equació estimada amb una mostra de 40 llars, utilitzant el fitxer *hostel*, és la següent:

$$\ln(hostel_i) = -27.36 + 4.442 \ln(inc_i) - 0.523 hhsize_i$$

$$R^2=0.738 \qquad n=40$$

A la vista d'aquests resultats, podem dir que els serveis hotelers són un bé de luxe, ja que l'elasticitat de la demanda/renda per aquest bé és molt alta (4.44). Això vol dir que, si la renda s'incrementa en un 1%, la despesa en serveis hotelers s'incrementarà un 4.44%, mantenint fix la mida de la família. D'altra banda, si la mida de la llar augmenta en un membre, aleshores la despesa en serveis hotelers disminuirà en un 52%.

EXEMPLE 3.3 Una regressió hedònica per a cotxes

El model hedònic de mesurament de preus es basa en el supòsit que el valor d'un bé depèn del valor de les seues diferents característiques. Així, el preu d'un cotxe dependrà del valor que el comprador assigne als seus atributs: qualitatius (per exemple, canvi automàtic, potència, dièsel, direcció assistida, aire condicionat), i quantitatius (per exemple, consum de combustible, pes, etc .). La base de dades per a aquest exemple és el fitxer *hedcarsp* (preus hedònics dels cotxes a Espanya) i cobreix els anys 2004 i 2005. Un primer model basat només en atributs quantitatius és el següent:

$$\ln(price) = \beta_1 + \beta_2 volume + \beta_3 fueleff + u$$

on *volume* és longitud× amplada×alçada en m^3 i *fueleff* és la ràtio litres per 100 km/cavalls de vapor expressada en percentatge.

L'equació estimada amb una mostra de 214 observacions és la següent:

$$\ln(price)_i = \underset{(0.151)}{14.97} + \underset{(0.009)}{0.0956}\, volume_i - \underset{(0.010)}{0.1608}\, fueleff_i$$

$$R^2=0.765 \qquad n=214$$

La interpretació de $\hat{\beta}_2$ i $\hat{\beta}_3$ Mantenint *fueleff* fix, si augmenta *volume* en 1 m^3, el preu dels cotxes s'incrementarà en un 9.56%. Mantenint fix *volume*, si la ràtio litres per 100 km/cavalls de vapor augmenta en un punt percentual, el preu dels automòbils es reduirà en un 16,08%.

EXEMPLE 3.4. Vendes i publicitat: el cas de Lydia E. Pinkham

En aquest cas es va a estimar un model amb dades de sèries temporals a fi de mesurar l'efecte que puguen tenir les despeses de publicitat, realitzats al llarg de diferents períodes de temps, sobre les vendes del moment actual. Es designen per V_t i P_t a les vendes i a les despeses en publicitat, realitzats en el moment *t*, el model plantejat inicialment per a explicar les vendes, en funció de les despeses en publicitat presents i passats, és el següent:

$$V_t = \alpha + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 P_{t-2} + \cdots + u_t \tag{3-30}$$

En l'expressió anterior els punts suspensius indiquen que les despeses en publicitat realitzats en el passat segueixen exercint influència de manera indefinida, tot i que, se suposa, que amb un impacte decreixent sobre les vendes del moment actual. Naturalment, el model anterior no és operatiu, ja que té un nombre indefinit de coeficients. Per solucionar el problema es poden adoptar, en principi, dos enfocaments. El primer enfocament consisteix a fixar *a priori* el nombre màxim de períodes durant els quals la publicitat manté els seus efectes sobre les vendes. En el segon enfocament, es postula que els coeficients es comporten d'acord amb alguna llei que permet determinar el seu valor en funció d'un nombre reduït de paràmetres, possibilitant a més una ulterior simplificació.

En el primer enfocament el problema que sorgeix és que en general no existeixen criteris precisos i informació suficient que permetin la fixació a priori del nombre màxim de períodes. Per aquesta raó, anem a veure un cas particular del segon enfocament que té gran interès per la plausibilitat del supòsit i la seua fàcil aplicació. El cas que considerarem consisteix a establir que els coeficients β_i disminueixen de forma geomètrica a mesura que ens allunyem cap enrere en el temps segons l'esquema:

$$\beta_i = \beta_1 \lambda^i \quad \forall i \qquad 0 < \lambda < 1 \tag{3-31}$$

A l'anterior transformació se li denomina transformació de Koyck, ja que va ser aquest autor el que la va introduir en 1954 per a l'estudi de la inversió.

Substituint (3-31) a (3-30), s'obté que

$$V_t = \alpha + \beta_1 P_t + \beta_1 \lambda P_{t-1} + \beta_1 \lambda^2 P_{t-2} + \cdots + u_t \tag{3-32}$$

El model anterior continua tenint infinits termes, però només tres paràmetres i a més es pot simplificar. En efecte, si expressem l'equació (3-32) per al període *t*-1 i multipliquem els dos membres per λ s'obté que

$$\lambda V_{t-1} = \alpha\lambda + \beta_1 \lambda P_{t-1} + \beta_1 \lambda^2 P_{t-2} + \beta_1 \lambda^3 P_{t-3} + \cdots + \lambda u_{t-1} \tag{3-33}$$

Restant membre a membre (3-33) de (3-32), i tenint en compte que els factors λ^i tendeixen a 0 en tendir *i* a infinit, s'arriba al següent resultat:

$$V_t = \alpha(1-\lambda) + \beta_1 P_t + \lambda V_{t-1} + u_t - \lambda u_{t-1} \tag{3-34}$$

El model ha quedat simplificat de manera que només té tres regressors, tot i que, a canvi, s'ha passat a un terme de pertorbació compost. Abans de veure l'aplicació d'aquest model es va analitzar el significat del coeficient λ i el problema de la durada dels efectes de les despeses en publicitat sobre les vendes. El paràmetre λ és la taxa a que decauen els efectes de les despeses en publicitat presents sobre les vendes presents i futures. Els efectes acumulats de la despesa en publicitat d'una unitat monetària sobre les vendes després de transcorreguts m períodes vénen donats per

$$\beta_1(1 + \lambda + \lambda^2 + \lambda^3 + \cdots + \lambda^m) \tag{3-35}$$

Per calcular la suma acumulada dels efectes, donada en (3-35), tindrem en compte que aquesta expressió és la suma dels termes d'una progressió geomètrica[2], de manera que es pot expressar de la següent manera:

$$\frac{\beta_1(1-\lambda^m)}{1-\lambda} \tag{3-36}$$

Quan *m* tendeix a infinit, aleshores la suma dels efectes acumulats ve donada per

$$\frac{\beta_1}{1-\lambda} \tag{3-37}$$

[2] Designant per a_p, a_u i *r* al primer terme, a l'últim terme i a la raó respectivament, la suma dels termes d'una progressió geomètrica convergent ve donada per

$$\frac{a_p - a_u}{1-r}$$

Una qüestió interessant és determinar quants períodes de temps es requereixen perquè s'obtinga el *p*% (per exemple, el 50%) de l'efecte total. Designant per *h* el nombre de períodes requerits per obtenir aquest percentatge, es pot establir que

$$p = \frac{\text{Efecte en } h \text{ periodes}}{\text{Efecte total}} = \frac{\dfrac{\beta_1(1-\lambda^h)}{1-\lambda}}{\dfrac{\beta_1}{1-\lambda}} = 1-\lambda^h \tag{3-38}$$

Fixat *p* es pot calcular *h* d'acord amb (3-38). En efecte, aïllant *h* en aquesta expressió s'obté que

$$h = \frac{\ln(1-p)}{\ln \lambda} \tag{3-39}$$

Aquest model el va utilitzar Kristian S. Palda en la seva tesi doctoral publicada en 1964, titulada *The Measurement of Cumulative Advertising Effects*, per analitzar els efectes acumulats de les despeses en publicitat, en el cas de la companyia Lydia E. Pinkham. Aquest cas, així com l'estudi de Palda, han estat la base a partir de la qual s'ha desenvolupat la investigació dels efectes de les despeses en publicitat. Anem a veure a continuació algunes característiques d'aquest cas:

1) La Lydia E. Pinkham Medicine Company fabricava des de 1873 un extracte d'herbes diluït en una solució alcohòlica. Aquest producte s'anunciava inicialment com un analgèsic i també com un remei per a una enorme varietat de malalties.

2) En general, en els diferents tipus de productes sol haver competència entre diferents marques, com puga ser el cas paradigmàtic de la Coca-Cola i la Pepsi-Cola en el camp de les colas. Quan això passa, per analitzar els efectes de les despeses en publicitat hi ha tenir en compte el comportament dels principals competidors. Lydia E. Pinkham tenia l'avantatge de no tenir competidors, i que, en la seua línia de producte, actuava a la pràctica com monopolista.

3) Una altra característica del cas Lydia E. Pinkham era que la major part de les despeses de distribució s'assignaven a la publicitat, ja que la companyia no tenia agents comercials, sent molt elevada la relació despeses en publicitat/vendes.

4) El producte va passar al llarg del temps per diferents avatars. Així, el 1914 la Food and Drug Administration (organisme dels Estats Units que estableix els controls per als productes alimentaris i els medicaments) el va acusar de publicitat enganyosa pel que va haver de canviar els seus missatges publicitaris. També la Internal Revenue (oficina d'impostos) va amenaçar amb aplicar-li una taxa sobre begudes alcohòliques, ja que el contingut alcohòlic del producte era del 18%. Per tots aquests motius es van produir canvis en la presentació i contingut durant el període 1915-1925. El 1925 la Food and Drug Administration va prohibir que el producte s'anunciés com a medicina, passant a distribuir-se com a beguda tònica. En el període 1926-1940 es van incrementar considerablement les despeses en publicitat per després decaure.

L'estimació del model (3-34) amb dades des de 1907 a 1960, recollits en el fitxer *pinkham*, és la següent:

$$\underset{}{ventas_t} = \underset{(95.7)}{138.7} + \underset{(0.156)}{0.3288}\, gpub_t + \underset{(0.0915)}{0.7593}\, ventas_{t-1}$$

$$R^2=0.877 \qquad n=53$$

La suma dels efectes acumulats de les despeses en publicitat sobre les vendes s'obté aplicant la fórmula (3-37):

$$\frac{\hat{\beta}_1}{1-\hat{\lambda}} = \frac{0.3288}{1-0.7593} = 1.3660$$

D'acord amb aquest resultat, per cada unitat monetària addicional gastada en publicitat s'obté un efecte acumulat total en les vendes de 1.366 unitats monetàries. Atès que és important no només determinar l'efecte total, sinó també com es distribueixen aquests efectes al llarg del temps, anem a contestar ara a la

següent pregunta: Quants períodes de temps es requereixen per assolir la meitat dels efectes totals? Aplicant la fórmula (3-39) per al cas de *p*= 0,5, s'obté el següent resultat:

$$\hat{h}(0.5) = \frac{\ln(1-0.5)}{\ln(0.7593)} = 2.5172$$

3.2.3 Implicacions algebraiques de l'estimació

Les implicacions algebraiques de l'estimació es deriven exclusivament de l'aplicació del mètode de *MQO* al model de regressió lineal múltiple:

1. *La suma dels residus de MQO és igual a 0:*

$$\sum_{i=1}^{n} \hat{u}_i = 0 \tag{3-40}$$

De la definició de residus

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki} \qquad i = 1, 2, \cdots, n \tag{3-41}$$

Si sumem per a les n observacions, obtenim:

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} - \cdots - \hat{\beta}_k \sum_{i=1}^{n} x_{ki} \tag{3-42}$$

D'altra banda, la primera equació del sistema d'equacions normals (3-20) és igual a

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} - \cdots - \hat{\beta}_k \sum_{i=1}^{n} x_{ki} = 0 \tag{3-43}$$

Si comparem (3-42) i (3-43), arribem a la conclusió que (3-40) es compleix.

Recordeu que, si (3-40) es compleix, això implica que

$$\sum_{i=1}^{n} y = \sum_{i=1}^{n} \hat{y}_i \tag{3-44}$$

i, dividint (3-40) i (3-44) per *n*, obtenim

$$\bar{\hat{u}} = 0 \qquad \bar{y} = \bar{\hat{y}} \tag{3-45}$$

2. *L'hiperplà MQO passa sempre a través del punt de mitjanes mostrals* $\left(\bar{y}, \bar{x}_2, \cdots, \bar{x}_k\right)$.

En efecte, dividint l'equació (3-43) per *n* s'ha de:

$$\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x}_2 + \cdots + \hat{\beta}_k \bar{x}_k \tag{3-46}$$

3. *El producte creuat mostral entre cada un dels regressors i els residus MQO és zero.*

És a dir,

$$\sum_{i=1}^{n} x_{ji}\hat{u}_i = 0 \qquad j = 2,3,\cdots,k \tag{3-47}$$

Utilitzant les últimes *k* equacions normals (3-20) i tenint en compte que, per definició $\hat{u}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki}$, podem veure que

$$\begin{aligned} &\sum_{i=1}^{n} \hat{u}_i x_{2i} = 0 \\ &\sum_{i=1}^{n} \hat{u}_i x_{3i} = 0 \\ &\cdots \qquad\qquad \cdots \\ &\sum_{i=1}^{n} \hat{u}_i x_{ki} = 0 \end{aligned} \tag{3-48}$$

4. *El producte creuat mostral entre els valors ajustats ($\hat{y}$) i els residus MQO és zero.*

És a dir,

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0 \tag{3-49}$$

Tenint en compte (3-40) i (3-48), obtenim

$$\begin{aligned} \sum_{i=1}^{n} \hat{y}_i \hat{u}_i &= \sum_{i=1}^{n} (\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki})\hat{u}_i = \hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i}\hat{u}_i + \cdots \hat{\beta}_k \sum_{i=1}^{n} x_{ki}\hat{u}_i \\ &= \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 0 + \cdots \hat{\beta}_k \times 0 = 0 \end{aligned} \tag{3-50}$$

3.3 Supòsits i propietats estadístiques dels estimadors de *MQO*

Abans d'estudiar les propietats estadístiques dels estimadors de *MQO* en el model de regressió lineal múltiple, necessitem formular un conjunt de supòsits estadístics. Específicament, al conjunt de supòsits que anem a formular se'ls denomina supòsits del *model lineal clàssic* (*MLC*). És important destacar que els supòsits estadístiques del *MLC* són molt simples, i que els estimadors de *MQO* tenen, sota aquests supòsits, molt bones propietats.

3.3.1 Supòsits estadístiques del *MLC* en la regressió lineal múltiple

a) Supòsit sobre la forma funcional

1) La relació entre el regressant, els regressors i l'error és lineal en els paràmetres:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \tag{3-51}$$

o, alternativament, per a totes les observacions,

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u} \tag{3-52}$$

b) Supòsits sobre els regressors

2) Els valors de $x_2, x_3, \cdots, x_k$ *són fixos en repetides mostres, o la matriu* **X** *és fixa en repetides mostres:*

Aquest és un supòsit fort en el cas de les ciències socials, on, en general, no és possible experimentar. Una hipòtesi alternativa pot formular-se així:

2) Els regressors* $x_2, x_3, \cdots, x_k$ *es distribueixen independentment de la pertorbació aleatòria.* Formulat d'una altra manera, **X** es distribueix de forma independent del vector de pertorbacions aleatòries, el que implica que $E(\mathbf{X}'\mathbf{u}) = \mathbf{0}$

Com vam fer en el capítol 2, anem a adoptar també el supòsit 2).

3) La matriu de regressors, **X***, no conté errors de mesurament.*

4) La matriu de regressors, **X***, té rang igual a k:*

$$\rho(\mathbf{X}) = k \tag{3-53}$$

Recordem que la matriu de regressors conté *k* columnes, corresponents als *k* regressors del model, i *n* files, corresponents al nombre d'observacions. El supòsit 4 té dues implicacions:

1. El nombre d'observacions, *n*, ha de ser igual o major que el nombre de regressors, *k*. Intuïtivament, això té sentit: per estimar *k* paràmetres, necessitem almenys *k* observacions

2. Cada regressor ha de ser linealment independent, el que implica que no hi ha relacions lineals exactes entre els regressors. Si un regressor és una combinació lineal exacta d'altres regressors, aleshores es diu que hi ha *multicolinealitat perfecta*, i el model no pot estimar-se.

Si hi ha una relació lineal aproximada - és a dir, no una relació exacta -, aleshores es poden estimar els paràmetres, encara que la seva fiabilitat es veurà afectada. En aquest cas, es diu que hi ha *multicolinealitat no perfecta*.

c) Supòsit sobre els paràmetres

5) Els paràmetres $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$ *són constants, o* $\boldsymbol{\beta}$ *és un vector constant.*

d) Supòsits sobre la pertorbació aleatòria

6) La mitjana de les pertorbacions és zero,

$$E(u_i) = 0, \quad i = 1, 2, 3, \ldots, n \quad \text{o} \quad E(\mathbf{u}) = \mathbf{0} \tag{3-54}$$

7) Les pertorbacions tenen una variança constant (supòsit d'homoscedasticitat):

$$var(u_i) = \sigma^2 \qquad i = 1, 2, \cdots n \tag{3-55}$$

8) Les pertorbacions amb diferents subíndexs no estan correlacionades entre si (supòsit de no autocorrelació):

$$E(u_i u_j) = 0 \qquad i \neq j \tag{3-56}$$

La formulació dels supòsits d'homoscedasticitat i no autocorrelació permet especificar la matriu de covariances del vector de pertorbacions:

$$\begin{aligned} E\left[\left[\mathbf{u}-E(\mathbf{u})\right]\left[\mathbf{u}-E(\mathbf{u})\right]'\right] &= E\left[\left[\mathbf{u}-\mathbf{0}\right]\left[\mathbf{u}-\mathbf{0}\right]'\right] = E\left[\left[\mathbf{u}\right]\left[\mathbf{u}\right]'\right] \\ = E\left[\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}\begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}\right] &= E\begin{bmatrix} u_1^2 & u_1u_2 & \cdots & u_1u_n \\ u_2u_1 & u_2^2 & \cdots & u_2u_n \\ \vdots & \vdots & \ddots & \vdots \\ u_nu_1 & u_nu_2 & \cdots & u_n^2 \end{bmatrix} \\ = \begin{bmatrix} E(u_1^2) & E(u_1u_2) & \cdots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \cdots & E(u_2u_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(u_nu_1) & E(u_nu_2) & \cdots & E(u_n^2) \end{bmatrix} &= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \end{aligned} \tag{3-57}$$

Per aconseguir la igualtat s'ha tingut en compte que la variança de cada element és constant i iguala σ^2, d'acord amb (3-55), i que la covariança entre cada parell d'elements és 0, d'acord amb (3-56).

El resultat anterior pot expressar-se de forma compacta així:

$$E(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I} \tag{3-58}$$

A la matriu donada en (3-58) se li denomina matriu escalar, ja que és un escalar (σ^2, en aquest cas) multiplicat per la matriu identitat.

9) La pertorbació **u** *té una distribució normal.*

Tenint en compte els supòsits 6 a 9, obtenim

$$u_i \sim NID(0, \sigma^2) \quad i = 1, 2, \cdots, n \qquad \text{o} \qquad \mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \tag{3-59}$$

on *NID* vol dir que la pertorbació està *normal i independent distribuïda*.

3.3.2 Propietats estadístiques de l'estimador de *MQO*

Sota els supòsits del *MLC*, l'estimador de *MQO* posseeix bones propietats. En les demostracions d'aquest apartat implícitament es tindran en compte sempre els supòsits 3, 4 i 5.

Els estimadors per MQO són lineals i no esbiaixats

Ara, anem a demostrar que l'estimador de *MQO* és linealment no esbiaixat. En primer lloc expressarem $\hat{\boldsymbol{\beta}}$ com una funció del vector **u**, utilitzant el supòsit 1, d'acord amb (3-60):

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\left[\mathbf{X}\boldsymbol{\beta}+\mathbf{u}\right] = \boldsymbol{\beta}+\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{u} \tag{3-60}$$

L'estimador de *MQO* pot expressar de la manera per tal de veure de forma més clara la propietat de linealitat:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}+\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{u} = \boldsymbol{\beta}+\mathbf{A}\mathbf{u} \tag{3-61}$$

on $\mathbf{A} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'$ és fixa sota el supòsit 2. Així doncs, $\hat{\boldsymbol{\beta}}$ és una funció lineal de **u** i, conseqüentment, és un estimador *lineal*.

Prenent les esperances en (3-60) i aplicant el supòsit 6, s'obté

$$E\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}+\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'E\left[\mathbf{u}\right] = \boldsymbol{\beta} \tag{3-62}$$

Per tant, $\hat{\boldsymbol{\beta}}$ és un estimador *no esbiaixat*.

Variança de l'estimador de MQO

Per calcular la matriu de covariances de $\hat{\boldsymbol{\beta}}$ són necessaris els supòsits 7 i 8, a més dels 6 primers:

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}) &= E\left[\hat{\boldsymbol{\beta}}-E(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}}-E(\hat{\boldsymbol{\beta}})\right]' = E\left[\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\right]\left[\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\right]' \\ &= E\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\right] = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}')\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1} \\ &= \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'E(\sigma^2\mathbf{I})\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1} = \sigma^2\left[\mathbf{X}'\mathbf{X}\right]^{-1}\end{aligned} \tag{3-63}$$

En el tercer pas de la demostració anterior s'ha tingut en compte que, d'acord amb (3-60), $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{u}$. El supòsit 2 s'ha tingut en compte en el quart pas. Finalment, els supòsits 7 i 8 s'han utilitzat en l'últim pas.

Per tant, $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2\left[\mathbf{X}'\mathbf{X}\right]^{-1}$ és la matriu de covariances del vector $\hat{\boldsymbol{\beta}}$. En aquesta matriu de covariances, la variança de cada element $\hat{\beta}_j$ apareix en la diagonal principal, mentre que les covariances entre cada parell d'elements es troben fora de la diagonal principal. Específicament, la variança de $\hat{\beta}_j$ (per j=2,3,…,k) és igual a σ^2 multiplicada per l'element corresponent de la diagonal principal de $\left[\mathbf{X}'\mathbf{X}\right]^{-1}$. Després d'operar, la variança de $\hat{\beta}_j$ pot expressar-se com

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{nS_j^2(1-R_j^2)} \tag{3-64}$$

on R_j^2 és el R quadrat de la regressió de cada x_j sobre la resta de regressors, n és la mida de la mostra i S_j^2 és la variança mostral del regressor x_j.

La fórmula (3-64) és vàlida per a tots els coeficients de pendent, però no per al terme independent.

A l'arrel quadrada de (3-64) se li denomina desviació estàndard (*de*) de $\hat{\beta}_j$:

$$de(\hat{\beta}_j) = \frac{\sigma}{\sqrt{nS_j^2(1-R_j^2)}} \tag{3-65}$$

Els estimadors de MQO són ELNEO

Sota els supòsits 1 a 8 del *MLC*, que són anomenats supòsits de Gauss-Markov, els estimadors de *MQO* són *estimadors lineals no esbiaixats* i *òptims* (*ELNEO*).

El teorema de Gauss Markov estableix que els estimadors *MQO* són estimadors òptims dins la classe dels estimadors lineals no esbiaixats. En aquest context òptim, vol dir que és un estimador amb la variança més xicoteta per a un determinat mida de mostra. Anem ara a comparar la variança d'un element de $\hat{\beta}$ ($\hat{\beta}_j$),amb qualsevol altre estimador $\tilde{\beta}_j$ que siga lineal (tal que $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$) i no esbiaixat (de manera que els pesos, w_j, han de complir algunes restriccions). La propietat que $\hat{\beta}_j$ és un estimador *ELNEO* té les següents implicacions en comparar la seua variança amb la variança de $\tilde{\beta}_j$

1) La variança d'un coeficient $\tilde{\beta}_j$ o és més gran que o igual a, la variança de $\hat{\beta}_j$ obtingut per *MQO*:

$$\text{var}(\tilde{\beta}_j) \geq \text{var}(\hat{\beta}_j) \qquad \text{j} = 1, 2, \ldots, k \tag{3-66}$$

2) La variança de qualsevol combinació lineal $\tilde{\beta}_j$ és més gran que, o igual a, la variança de la corresponent combinació lineal de $\hat{\beta}_j$.

A l'apèndix 3.1 es pot veure la demostració del teorema de Gauss-Markov.

Estimador de la variança de la pertorbació

Tenint en compte el sistema d'equacions normals (3-20), si coneixem *n-k* residus, podem obtenir els altres *k* residus utilitzant les restriccions que imposa aquest sistema als residus.

Per exemple, la primera equació normal ens permet obtenir el valor de $\hat{u}_n$ en funció dels residus restants:

$$\hat{u}_n = -\hat{u}_1 - \hat{u}_2 - \cdots - \hat{u}_{n-1}$$

Per tant, només hi ha *n-k* graus de llibertat en els residus de *MQO*, a diferència dels *n* graus de llibertat en les pertorbacions. Recordeu que els graus de llibertat es defineixen com la diferència entre el nombre d'observacions i el nombre de paràmetres estimats.

L'estimador no esbiaixat de σ^2 s'ajusta tenint en compte els graus de llibertat:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k} \tag{3-67}$$

Sota els supòsits 1 a 8, s'obté que

$$E(\hat{\sigma}^2) = \sigma^2 \tag{3-68}$$

Vegeu l'apèndix 3.2 per a la demostració.

A l'arrel quadrada de (3-67), $\hat{\sigma}$ se li denomina error estàndard de la regressió i és un estimador de σ.

Estimadors de les variances de $\hat{\boldsymbol{\beta}}$ ***i dels coeficients de pendent*** $\hat{\beta}_j$

L'estimador de la matriu de covariances de $\hat{\boldsymbol{\beta}}$ ve donat per

$$Var(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \left[\mathbf{X}'\mathbf{X}\right]^{-1} = \begin{bmatrix} \text{var}(\hat{\beta}_1) & Cov(\hat{\beta}_1,\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_1,\hat{\beta}_j) & \cdots & Cov(\hat{\beta}_1,\hat{\beta}_k) \\ Cov(\hat{\beta}_2,\hat{\beta}_1) & \text{var}(\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_2,\hat{\beta}_j) & \cdots & Cov(\hat{\beta}_2,\hat{\beta}_k) \\ \cdots & \cdots & \ddots & \cdots & \cdots & \cdots \\ Cov(\hat{\beta}_j,\hat{\beta}_1) & Cov(\hat{\beta}_j,\hat{\beta}_2) & \cdots & \text{var}(\hat{\beta}_j) & \cdots & Cov(\hat{\beta}_j,\hat{\beta}_k) \\ \cdots & \cdots & \cdots & \cdots & \ddots & \cdots \\ Cov(\hat{\beta}_k,\hat{\beta}_1) & Cov(\hat{\beta}_k,\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_k,\hat{\beta}_j) & \cdots & \text{var}(\hat{\beta}_k) \end{bmatrix} \tag{3-69}$$

La variança del coeficient del pendent $\hat{\beta}_j$, donada en (3-64), és una funció del paràmetre desconegut σ^2. Quan σ^2 se substitueix per la seva estimador $\hat{\sigma}^2$, s'obté un estimador de la variança de $\hat{\beta}_j$:

$$\text{var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{nS_j^2(1-R_j^2)} \tag{3-70}$$

D'acord amb l'expressió anterior, l'estimador de la variança de ve afectat pels següents factors:

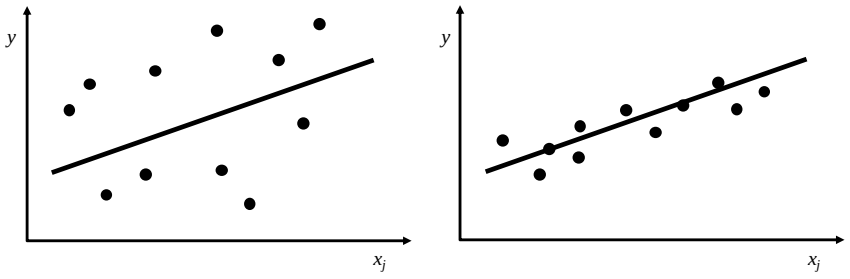
a) Com més gran és $\hat{\sigma}^2$, més gran és la variança de l'estimador. Això no és sorprenent en absolut: com més "soroll" hi hagi en l'equació, i, en conseqüència, més gran serà, de manera que serà $\hat{\sigma}^2$ més difícil estimar amb precisió l'efecte parcial de qualsevol regressor sobre i. (Vegeu figura 3.1).

b) A mesura que s'incrementa la grandària de la mostra, la variança de l'estimador es redueix.

c) Com més xicoteta siga la variança mostral d'un regressor, més gran és la variació del coeficient corresponent. Mantenint els altres factors igual, per a estimar β_j és preferible que la variació mostral de *xj* siga el més gran possible, tal com s'il·lustra a la figura 3.2. Hi ha moltes línies hipotètiques

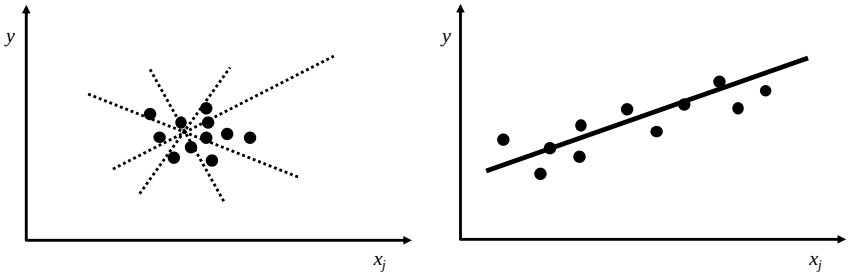
que podrien ajustar-se a les dades quan la variança mostral de x_j, (S_j^2) és xicoteta, com es pot veure en la part a) de la figura. En qualsevol cas, no està permès pel supòsit 4 que S_j^2=0.

d) Com més gran és R_j^2, (és a dir, com més gran siga la correlació del regressor j-èsim amb la resta dels regressors), major serà la variança de $\hat{\beta}_j$.



a) $\hat{\sigma}^2$ gran b) $\hat{\sigma}^2$ xicoteta

FIGURA 3.1. Influència de $\hat{\sigma}^2$ sobre l'estimador de la variança.



a) S_j^2 xicotet b) S_j^2 gran

FIGURA 3.2. Influència de S_j^2 sobre l'estimador de la variança.

A l'arrel quadrada de (3-70) se li denomina *error estàndard* (*ee*) de $\hat{\beta}_j$:

$$ee(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{nS_j^2(1-R_j^2)}} \qquad (3\text{-}71)$$

Altres propietats dels estimadors MQO

Sota els supòsits 1 a 6 del MLC, l'estimador de *MQO*, és consistent, com es pot veure en l'apèndix 3.3, *asimptòtica i normalment distribuït*, i també *asimptòticament eficient* dins de la classe dels estimadors consistents i asimptòticament normals.

Sota els supòsits 1 a 9 del MLC, l'estimador *MQO* és també l'estimador de *màxima versemblança* (*MV*), com es prova en l'apèndix 3.4, i és *l'estimador no esbiaixat de mínima variança* (*ENEMV*). Això últim vol dir que l'estimador de *MQO* té la menor variança entre tots els estimadors no esbiaixats, siguen lineals o no.

3.4 Més sobre formes funcionals

En aquest apartat examinarem dos temes sobre formes funcionals: l'ús dels logaritmes en models economètrics i les funcions polinomiales.

3.4.1 Utilització de logaritmes en els models economètrics

Algunes variables s'utilitzen sovint en forma logarítmica. Així és en el cas de les variables monetàries que, en general, són positives o d'altres variables amb valors elevats. La utilització de models amb transformacions logarítmiques té a més els seus avantatges. Una d'elles és que els coeficients tenen interpretacions atractives (elasticitats o semi-elasticitats). Una altra és la invariança dels coeficients de pendent quan hi ha canvis d'escala en les variables. Prendre logaritmes pot ser convenient pel fet que redueix el rang de les variables, el que fa que les estimacions siguen menys sensibles als valors extrems de les variables. Els supòsits de l'*MLC* se satisfan més sovint en models que apliquen logaritmes a la variable endògena, que en els models que no apliquen cap transformació. Així, succeeix que la distribució condicional de y és freqüentment heteroscedàstica, mentre que ln(y) pot ser homoscedàstica.

Una limitació de la transformació logarítmica és que no es pot utilitzar quan la variable original pren valors zero o negatius. D'altra banda, algunes variables que es mesuren en anys i en altres variables que són una proporció o un percentatge, s'utilitza la variable original sense cap transformació.

3.4.2 Funcions polinomiales

Les funcions polinomiales s'han utilitzat àmpliament en la investigació Econometrica. Quan en el model només hi ha regressors corresponents a una funció polinomial diem que és un *model polinomial*. La forma general del model polinomial de grau k pot expressar-se com

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \cdots + \beta_k x^k + u \tag{3-72}$$

Funcions quadràtiques

Un cas interessant de funcions polinomiales és la *funció quadràtica*, que és una *funció polinomial de segon grau*. Quan hi ha només regressors corresponents a la funció quadràtica, tenim un *model quadràtic*:

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + u \tag{3-73}$$

Les funcions quadràtiques s'utilitzen molt sovint en economia aplicada per captar la disminució o l'augment dels efectes marginals. És important observar que, en aquest cas, β_2 no mesura el canvi en y pel que fa a x, perquè no té sentit mantenir x^2 fix, mentre canvia x. L'efecte marginal de x sobre y, que depèn linealment del valor de x, és el següent:

$$em = \frac{dy}{dx} = \beta_2 + 2\beta_3 x \tag{3-74}$$

En una aplicació específica, aquest efecte marginal s'avaluarà per valors específics de x. Si β_2 i β_3 tenen signes oposats del punt de canvi està situat en

$$x^* = -\frac{\beta_2}{2\beta_3} \tag{3-75}$$

Si $\beta_2>0$ i $\beta_3<0$, l'efecte marginal de *x* sobre *y* és positiu al principi, però serà negatiu quan *x* siga més gran que x^*. Si $\beta_2<0$ i $\beta_3>0$, l'efecte marginal de *x* sobre *y* és negatiu al principi, però serà positiu per *x* més gran que x^*.

Exemple 3.5 Salaris i anys d'antiguitat a l'empresa

Utilitzant les dades de *ceosal2* per estudiar el tipus de relació entre el salari (*salary*) dels consellers delegats (CEO) a Estats Units i els anys de permanència en l'empresa com CEO de la companyia (*ceoten*), es va estimar el següent model:

$$\ln(salary) = \underset{(0.086)}{6.246} + \underset{(0.0001)}{0.0006}\, profits + \underset{(0.0156)}{0.0440}\, ceoten - \underset{(0.00052)}{0.0012}\, ceoten^2$$

$$R^2=0.1976 \qquad n=177$$

on els beneficis de les companyies (profits) estan expressats en milions de dòlars i el salari és la remuneració anual expressada en milers de dòlars.

L'efecte marginal de ceoten sobre salary expressat en percentatge és el següent:

$$em_{salario/ceoten}\% = 4.40 - 2\times 0.12 ceoten$$

Així, per a un conseller delegat amb 10 anys en la seua companyia, si està un any més a l'empresa, el seu salari s'incrementarà en un 2%. Igualant a zero l'expressió anterior i aïllant *ceoten*, ens trobem amb que l'efecte màxim de permanència com a conseller delegat sobre el salari s'aconsegueix als 18 anys. És a dir, fins als 18 anys com CEO l'efecte marginal del salari pel que fa als anys de permanència en la companyia és positiu. Per contra, des dels 18 anys en endavant, aquest efecte marginal és negatiu.

Funcions cúbiques

Un altre cas interessant és la funció cúbica o *funció polinomial de tercer grau*. Si en el model hi ha només regressors corresponents a la funció cúbica, tenim un *model cúbic*:

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3 + u \tag{3-76}$$

Els models cúbics s'utilitzen molt sovint en economia aplicada per captar variacions en els efectes marginals, particularment en les funcions de costos. L'*efecte marginal* (*em*) de *x* sobre *y*, que depèn, segons una forma quadràtica, del valor de *x*, serà el següent:

$$em = \frac{dy}{dx} = \beta_2 + 2\beta_3 x + 3\beta_4 x^2 \tag{3-77}$$

El mínim d'*em* es produirà quan

$$\frac{dem}{dx} = 2\beta_3 + 6\beta_4 x = 0 \tag{3-78}$$

Per tant,

$$em_{\min} = \frac{-\beta_3}{3\beta_4} \tag{3-79}$$

En un model cúbic d'una funció de costos, s'ha de complir la restricció $\beta_3^2 < 3\beta_4\beta_2$ per garantir que em_{mi} siga positiu. Altres restriccions que la funció de costos ha de complir són les següents: β_1, β_2, i $\beta_4>0$; i $\beta_3<0$.

Exemple 3.6 Efecte marginal en una funció de costos

Utilitzant les dades de 11 empreses de plantes de cel·lulosa (fitxer *costfunc*) per estudiar la funció de costos, es va estimar el següent model:

$$\underset{}{cost} = \underset{(1.602)}{29.16} + \underset{(0.2167)}{2.316}\,output - \underset{(0.0081)}{0.0914}\,output^2 + \underset{(0.000086)}{0.0013}\,output^3$$

$$R^2=0.9984 \qquad n=11$$

on *output* és la producció de pasta de paper en milers de tones i cost és el cost total en milions d'euros.

El *cost marginal* és el següent:

$$marcost = 2.316 - 2\times 0.0914 output + 3\times 0.0013 output^2$$

Per tant, en una empresa amb una producció de 30 mil tones de pasta de paper, si l'empresa augmenta la producció de cel·lulosa en mil de tones, el cost s'incrementarà en 0.754.000 d'euros. Calculant el mínim de l'expressió anterior i resolent per l'output, ens trobem que el cost marginal mínim és igual a una producció de 23222 tones de pasta de paper.

3.5 Bondat de l'ajust i selecció de regressors

Una vegada que s'han aplicat els mínims quadrats, és convenient tenir alguna mesura de la bondat de l'ajust del model a les dades. En el cas que s'hagin estimat diversos models alternatius, les mesures de la bondat de l'ajust podrien ser utilitzades per a seleccionar el model més apropiat.

En la literatura Econometrica ha nombroses mesures de bondat de l'ajust. La més popular és el coeficient de determinació, que es designa per R^2 o *R*-quadrat, i el coeficient de determinació ajustat, que es designa per $\bar{R}^2$ o *R*-quadrat ajustat. Atès que aquestes mesures tenen algunes limitacions, ens referirem també al criteri d'informació d'Akaike (*AIC*) i al criteri de Schwarz (*SC*).

3.5.1 Coeficient de determinació

Com hem vist al capítol 2, el coeficient de determinació es basa en la següent descomposició:

$$SQT = SQE + SQR \tag{3-80}$$

on *SQT* és la *suma de quadrats totals*, *SQE* és la *suma de quadrats explicats* i *SQR* és la *suma de quadrats residual*.

Basant-se en aquesta equació, el coeficient de determinació es defineix com:

$$R^2 = \frac{SQE}{SQT} \tag{3-81}$$

Alternativament, i d'una forma equivalent, el coeficient de determinació es pot definir com

$$R^2 = 1 - \frac{SQR}{SQT} \tag{3-82}$$

Els valors extrems del coeficient de determinació són: 0, quan la variança explicada és zero, i 1, quan la variança residual és zero, és a dir, quan l'ajust és perfecte. Per tant,

$$0 \leq R^2 \leq 1 \tag{3-83}$$

Un R^2 xicotet implica que la variança de la pertorbació (σ^2) és gran en relació a la variació de *y*, el que significa que β_j no pot ser estimada amb precisió. Però cal recordar, que una variança de la pertorbació gran pot compensar-se amb una mida mostral elevat, de manera que si *n* és prou gran, podem ser capaços d'estimar els coeficients amb precisió tot i que no s'hagin controlat molts dels factors no observats.

Per interpretar el coeficient de determinació adequadament, s'han de tenir en compte les següents cauteles:

a) Quan s'afegeixen noves variables explicatives, el coeficient de determinació augmenta el seu valor o, almenys, manté el mateix valor. Això succeeix tot i que la variable o variables afegides no tinguen relació amb la variable endògena. Així doncs, sempre es verifica que

$$R_j^2 \geq R_{j-1}^2 \tag{3-84}$$

on R_{j-1}^2 és el *R* quadrat en un model amb *j*-1 regressors, i R_j^2 és el *R* quadrat en un model amb un regressor addicional. És a dir, si s'afegeixen variables a un model determinat, R^2 mai disminuirà, encara que aquestes variables no tenen una influència significativa.

b) Si el model no té terme independent, el coeficient de determinació no té una interpretació clara, perquè la descomposició donada (3-80) no es compleix. A més, les dues formes de càlcul esmentades - (3-81) i (3-82) - en general condueixen a resultats diferents, que en alguns casos poden quedar fora de l'interval [0,1].

c) El coeficient de determinació no es pot utilitzar per comparar models en què la forma funcional de la variable endògena és diferent. Per exemple, R^2 no es pot aplicar per comparar dos models en què el regressant és la variable original en un, *y*, i ln(*y*) en l'altre.

3.5.2 R quadrat ajustat

Per a superar una de les limitacions del R^2, aquest coeficient es pot "ajustar" de manera que tinga en compte el nombre de variables incloses en un model donat. Per veure com R^2 usual podria ajustar-se, és útil expressar com

$$R^2 = 1 - \frac{SCR / n}{SCT / n} \tag{3-85}$$

on, en el segon terme del segon membre, apareix la variança residual dividida per la variança del regressant.

El R^2, tal com està definit en (3-85), és una mesura mostral. Ara bé, si volem una mesura poblacional (R_{POB}^2), aquesta es podria definir com

$$R^2_{POB} = 1 - \frac{\sigma^2_u}{\sigma^2_y} \tag{3-86}$$

No obstant això, cal tenir en compte que es disposa d'una millor estimació de les variances, σ^2_u i σ^2_y, que les utilitzades en (3-85). En el seu lloc, utilitzarem estimacions no esbiaixats d'aquestes variancess:

$$\bar{R}^2 = 1 - \frac{SCR/(n-k)}{SCT/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k} \tag{3-87}$$

Aquesta mesura es denomina *R* quadrat ajustat, o $\bar{R}^2$. El principal atractiu del $\bar{R}^2$ és que imposa una penalització a l'afegir altres regressors a un model. Si s'afegeix un regressor al model la *SQR* decreix $\bar{R}^2$, en el pitjor dels casos queda, igual. D'altra banda, els graus de llibertat de la regressió (*n-k*) sempre disminueixen. Per això, el $\bar{R}^2$ pot créixer o decréixer quan s'afegeix un nou regressor al model. És a dir:

$$\bar{R}^2_j \geq \bar{R}^2_{j-1} \quad \text{o} \quad \bar{R}^2_j \leq \bar{R}^2_{j-1} \tag{3-88}$$

Un resultat algebraic interessant és el fet que si afegim un nou regressor a un model, el $\bar{R}^2$ s'incrementa si, i només si, l'estadístic *t* del nou regressor és més gran que un en valor absolut. Així, veiem immediatament que $\bar{R}^2$ podria ser utilitzat per decidir si un determinat regressor addicional ha de ser inclòs en el model. El $\bar{R}^2$ té una cota superior que és igual a 1, però estrictament no té una cota inferior, ja que pot prendre un valor negatiu, encara que molt a prop de 0.

Les observacions b) i c) fetes per el *R* quadrat segueixen sent vàlides per al *R* quadrat ajustat.

3.5.3 Criteri d'informació d'Akaike (*AIC*) i criteri de Schwarz (*SC*)

Aquests dos criteris -criteri d'informació d'Akaike (*AIC*) i criteri de Schwarz (*SC*)- tenen una estructura molt similar. Per aquesta raó, s'examinaran conjuntament.

L'estadístic *AIC*, proposat per Akaike (1974) i basat en la teoria de la informació, té la següent expressió:

$$AIC = -\frac{2l}{n} + \frac{2k}{n} \tag{3-89}$$

on *l* és el logaritme de la funció de versemblança (suposant que les pertorbacions tinguen una distribució normal) avaluada per als valors estimats dels coeficients.

L'estadístic *SC* proposat per Schwarz (1978), té la següent expressió:

$$SC = -\frac{2l}{n} + \frac{k\ln(n)}{n} \tag{3-90}$$

Els estadístics *AIC* i *SC*, a diferència dels coeficients de determinació (R^2 i $\bar{R}^2$), indiquen millors ajustos com més baixos siguen els seus valors. És important destacar que els estadístics *AIC* i *SC* no tenen cotes, a diferència del R^2.

a) Els estadístics *AIC* i *SC* penalitzen la introducció de nous regressors. En el cas d'*AIC*, com es pot veure en el segon terme del segon membre de (3-89), el nombre de regressors *k* apareix en el numerador. Per tant, el creixement de *k* s'incrementarà el valor de l'*AIC* i per tant empitjorarà la bondat de l'ajust, si no es veu compensat per un creixement suficient de *l*. En el cas del *SC*, com es pot veure en el segon terme del segon membre de (3-90), el numerador és $k\ln(n)$. Per $n>7$, passa el següent: $k\ln(n)> 2k$. Per tant, el *SC* imposa una penalització major a la introducció de regressors que l'*AIC* quan la mida de la mostra és major de 7.

b) Els estadístics *AIC* i *SC* es pot aplicar a models estadístics sense terme independent.

c) Els estadístics *AIC* i *SC* no són mesures relatives com ho són els coeficients de determinació. Per tant, la seua magnitud, en si mateixa, no ofereix cap informació.

d) Els estadístics *AIC* i *SC* es pot aplicar per comparar models en els quals les variables endògenes tenen diferents formes funcionals. En particular, anem a comparar dos models en els quals els regressades són *y* i ln (*y*). Quan el regressant és *y*, s'aplica la fórmula (3-89) en el cas de l'*AIC*, o (3-90) en el cas del *SC*. Quan el regressant és ln(*y*), i a més volem comparar amb un altre model en què el regressant és *y*, cal corregir aquests estadístiques de la següent manera

$$AIC_C = AIC + 2\overline{\ln(Y)} \tag{3-91}$$

$$SC_C = SC + 2\overline{\ln(Y)} \tag{3-92}$$

on AIC_C i SC_C són els estadístics corregits, i *AIC* i *SC* són els estadístics que subministra qualsevol paquet economètric com, per exemple, l'E-views.

Exemple 3.7 Selecció del millor model

Per analitzar els determinants de la despesa en productes lactis, s'han considerats els següents models alternatius:

1) $dairy = \beta_1 + \beta_2 inc + u$
2) $dairy = \beta_1 + \beta_2 \ln(inc) + u$
3) $dairy = \beta_1 + \beta_2 inc + \beta_3 punder5 + u$
4) $dairy = \beta_2 inc + \beta_3 punder5 + u$
5) $dairy = \beta_1 + \beta_2 inc + \beta_3 hhsize + u$
6) $\ln(dairy) = \beta_1 + \beta_2 inc + u$
7) $\ln(dairy) = \beta_1 + \beta_2 inc + \beta_3 punder5 + u$
8) $\ln(dairy) = \beta_2 inc + \beta_3 punder5 + u$

on *inc* és la renda disponible de les llars, *hhsize* és el nombre de membres de la llar i *punder5* és la proporció de nens menors de cinc anys a la llar.

Utilitzant una mostra de 40 llars (fitxer *demand*), i tenint en compte que $\overline{\ln(dairy)} = 2.3719$, els estadístics de bondat de l'ajust obtinguts pels 8 models es mostren en el quadre 3.1. En particular, l'estadístic *AIC* corregit pel model 6) s'ha calculat com segueix:

$$AIC_C = AIC + 2\overline{\ln(Y)} = 0.2794 + 2 \times 2.3719 = 5.0232$$

Conclusions

a) El *R*-quadrat pot ser utilitzat per comparar els següents parells de models: 1) amb 2), i 3) amb 5).

b) El *R*-quadrat ajustat només es pot utilitzar per comparar els models 1) amb 2), 3) i 5); i 6) amb 7.

c) El millor dels vuit models és el model de 7) d'acord amb els criteris *AIC* i *SC*.

QUADRE 3.1. Mesures de bondat d'ajust de vuit models.

Nombre de model	1	2	3	4	5	6	7	8
Regressant	*dairy*	*dairy*	*dairy*	*dairy*	*dairy*	ln(*dairy*)	ln(*dairy*)	ln(*dairy*)
Regressors	*intercept inc*	*intercept* ln(*inc*)	*intercept inc punder5*	*inc punder5*	*intercept Inc househsize*	*intercept inc*	*intercept inc punder5*	*inc punder5*
R-quadrat	0.4584	0.4567	0.5599	0.5531	0.4598	0.4978	0.5986	-0.6813
R-quadrat ajustat	0.4441	0.4424	0.5361	0.5413	0.4306	0.4846	0.5769	-0.7255
Criteri d'informació d'Akaike	5.2374	5.2404	5.0798	5.0452	5.2847	0.2794	0.1052	1.4877
Criteri de Schwarz	5.3219	5.3249	5.2065	5.1296	5.4113	0.3638	0.2319	1.5721
Criteri d'informació d'Akaike corregit						5.0232	4.8490	6.2314
Criteri de Schwarz corregit						5.1076	4.9756	6.3159

Exercicis

Exercici 3.1 Considereu el model de regressió lineal $\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$, on $\mathbf{X}$ és una matriu 50×5.

Contesteu de manera raonada a les següents qüestions:

a) Quines són les dimensions dels vectors $\mathbf{y}$, $\boldsymbol{\beta}$, $\mathbf{u}$?

b) Quantes equacions hi ha al sistema d'equacions normals $\mathbf{X'X\hat{\beta}} = \mathbf{X'y}$?

c) Quina condició ha de complir per poder obtenir $\hat{\boldsymbol{\beta}}$?

Exercici 3.2 Donat el model

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

i les següents dades:

y	x_2	x_3
10	1	0
25	3	-1
32	4	0
43	5	1
58	7	-1
62	8	0
67	10	-1
71	10	2

a) Estimeu β_1, β_2 i β_3 per *MQO*.

b) Calculeu la suma dels quadrats dels residus.

c) Obtinga la variança residual.

d) Obtinga la variança explicada per la regressió.

e) Obtinga la variança de la variable endògena.

f) Calculeu el coeficient de determinació.

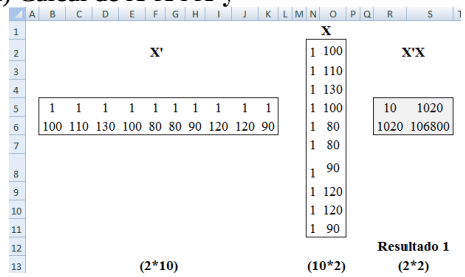
g) Obtinga una estimació no esbiaixada de σ^2.

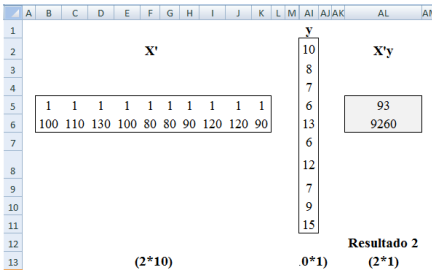
h) Estimeu la variança de $\hat{\beta}_2$.

Per respondre a aquestes preguntes pot utilitzar Excel. Vegeu el requadre 3.1.

Requadre 3.1

a) Càlcul de **X'X** i **X'y**





Explicació per **X'X**

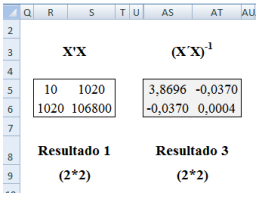
a) Introduïu les matrius **X'** i **X**: B5:K6 i N2:O11 en Excel

b) El producte **X'X** es calcula seleccionant prèviament les cel·les on voleu afegir la matriu resultant (R5:S6).

c) Un cop seleccionades les cel·les per a la matriu resultant, i mentre encara està ressaltada, escriviu la fórmula següent: = MMULT (B5:K6;N2:O11)

d) Quan la fórmula s'hagi introduït, premeu la *tecla Ctrl* i la tecla de majúscules simultàniament, aleshores, tenint pressionades aquestes dues tecles, premeu la *tecla Enter* també.

2) Càlcul de $\mathbf{(X'X)^{-1}}$



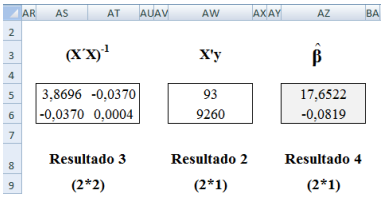
a) Introduïu la matriu **X'X** en Excel: R5:S6

b) Hi trobem la inversa de la matriu **X'X**, seleccionant prèviament les cel·les on volem col·locar la matriu resultant (AS5:AT6)

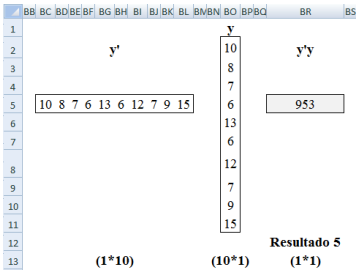
c) Un cop seleccionades les cel·les per a la matriu resultant, i mentre encara està ressaltada, escriviu la fórmula següent: = MINVERSA (AO5:AP6).

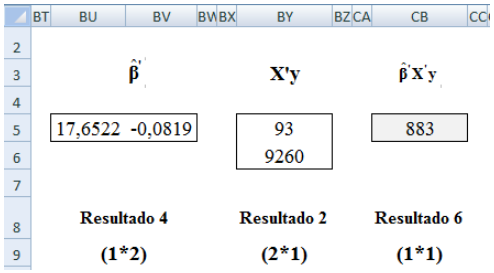
d) Quan la fórmula s'haja introduït, premeu la *tecla Ctrl* i la tecla de majúscules simultàniament, llavors, tenint pressionades aquestes dues tecles, premeu la *tecla Enter* també.

3) Càlcul de vector $\hat{\boldsymbol{\beta}}$



4) Càlcul de $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ i σ^2





$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{y}}'\hat{\mathbf{y}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = R.5 - R.6 = 953 - 883 = 70$$

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-2} = \frac{70}{8} = 8.6993$$

5) Càlcul de la matriu de covariances de $\hat{\boldsymbol{\beta}}$

$$\text{var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2\left[\mathbf{X}'\mathbf{X}\right]^{-1} = 8.6993\begin{pmatrix} 3.8696 & -0.0370 \\ -0.0370 & 0.0004 \end{pmatrix} = \begin{pmatrix} 33.6624 & -0.3215 \\ -0.3215 & 0.0032 \end{pmatrix}$$

Exercici 3.3 El següent model ha estat estimat per explicar les vendes anuals d'empreses fabricants de productes de neteja domèstica en funció d'un índex de preus relatiu (*ipr*) i de despeses de publicitat (*dpub*):

$$vendes = \beta_1 + \beta_2 ipr + \beta_3 dpub + u$$

on les *vendes* estan expressades en milions d'euros, *ipr* és un índex de preus relatius (preus de l'empresa/preus de l'empresa 1 de la mostra) i *dpub* són les despeses anuals realitzats en publicitat i campanyes de promoció i difusió, expressats també en milions d'euros.

Per a això es disposa de les dades sobre deu empreses fabricants de productes de neteja domèstica:

empresa	*vendes*	*ipr*	*dpub*
1	10	100	300
2	8	110	400
3	7	130	600
4	6	100	100
5	13	80	300
6	6	80	100
7	12	90	600
8	7	120	200
9	9	120	400
10	15	90	700

Utilitzant un full excel:

a) Estimeu els paràmetres del model proposat.
b) Estimeu la matriu de covariances.
c) Calculeu el coeficient de determinació.

Nota: En el requadre 3.1 s'estima el model $ventas = \beta_1 + \beta_2 rpi + u$ utilitzant Excel. Allà també es poden veure les instruccions per a fer-ho.

Exercici 3.4 Un investigador, que està elaborant un model economètric amb el qual desitja explicar el comportament de la renda, formula la següent especificació:

$$renda = \alpha + \beta cons + \gamma estalvi + u \qquad [1]$$

on *renda* és la renda disponible de les famílies, *cons* és el consum total i *estalvi* és l'estalvi total de les famílies.

L'investigador no va tenir en compte que les tres magnituds anteriors estan lligades per la identitat

$$\text{renda} = \text{cons} + \text{estavi} \qquad [2]$$

L'equivalència entre els models [1] i [2] exigeix que, a més de desaparèixer el terme de pertorbació, els paràmetres del model [1] agafen els següents valors:

$$\alpha = 0,\ \beta = 1,\ \gamma = 1$$

Si es fan servir les dades d'un país per ajustar l'equació [1] per MQO, es pot esperar, en general, que les estimacions obtingudes agafen els valors

$$\hat{\alpha} = 0, \hat{\beta} = 1, \hat{\gamma} = 0\,?$$

Justifiqueu la resposta, utilitzant notació matemàtica.

Exercici 3.5 Un investigador planteja el següent model economètric per a explicar els ingressos totals per turisme en un país determinat (*ingtotur*):

$$ingtotur = \beta_1 + \beta_2 ingmitur + \beta_3 nomtur + u$$

on *ingmitur* és l'ingrés mitjà per turista i *nomtur* és el nombre total de turistes.

a) És obvi que *ingtotur*, *ingmitur* i *nomtur* estan lligats també per la relació *ingtotur*=*ingmitur*×*nomtur*; afectarà aquest fet d'alguna manera a les estimacions dels paràmetres del model proposat?

b) Hi ha una altra manera funcional del model que implique restriccions més forts sobre els paràmetres? Si l'hagués, indiqueu.

c) Li sembla raonable utilitzar el model indicat per a explicar el comportament dels ingressos per turisme?

Exercici 3.6 Suposem que s'ha d'estimar el model

$$\ln(y) = \beta_1 + \beta_2 \ln(x_2) + \beta_3 \ln(x_3) + \beta_4 \ln(x_4) + u$$

utilitzant els següents observacions:

x_2	x_3	x_4
3	12	4
2	10	5
4	4	1
3	9	3
2	6	3
5	5	1

Quins problemes pot plantejar l'estimació d'aquest model?

Exercici 3.7 Contesteu a les següents preguntes:

a) Expliqueu que mesuren els coeficients de determinació (R^2) i de determinació corregit ($\overline{R}^2$). Per a què es poden utilitzar? Raoneu la resposta.

b) Donats els models

$$\ln(y)=\beta_1+\beta_2\ln(x)+u \quad (1)$$
$$\ln(y)=\beta_1+\beta_2\ln(x)+\beta_3\ln(z)+u \quad (2)$$
$$\ln(y)=\beta_1+\beta_2\ln(z)+u \quad (3)$$
$$y=\beta_1+\beta_2 z+u \quad (4)$$

indique quina mesura de bondat de l'ajust és adequada per comparar els següents parells de models: (1)-(2); (1)-(3); i (1)-(4). Raoneu la resposta.

Exercici 3.8 S'estima per *MQO* el següent model:

$$\ln(y) = \beta_1 + \beta_2 \ln(x) + \beta_3 \ln(z) + u$$

a) Poden ser tots els residus mínim-quadràtics positius? Raoneu la resposta.

b) Sota la hipòtesi bàsica de no autocorrelació de les pertorbacions, són independents dels residus mínim-quadràtics? Raoneu la resposta.
c) Suposant que les pertorbacions no tinguen distribució normal, els estimadors mínim-quadràtics són no esbiaixats? Raoneu la resposta.

Exercici 3.9 Penseu el model de regressió

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$$

on **y** i **u** són vectors 8×1, **X** és una matriu 8×3 i $\boldsymbol{\beta}$ és un vector 3×1 de paràmetres desconeguts. A més, es disposa de la següent informació:

$$\mathbf{X}'\mathbf{X}=\begin{bmatrix}2 & 0 & 0\\ 0 & 3 & 0\\ 0 & 0 & 3\end{bmatrix} \qquad \hat{\mathbf{u}}'\hat{\mathbf{u}}=22$$

Responeu a les següents preguntes, justificant la resposta:

a) Indiqueu la mida de la mostra, el nombre de regressors, el nombre de paràmetres i els graus de llibertat de la suma dels quadrats dels residus.
b) Deduïu la matriu de covariances del vector $\hat{\boldsymbol{\beta}}$, explicitant els supòsits utilitzats. Estimeu les variances dels estimadors dels paràmetres del model.
c) Conté el model de regressió terme constant? Quines implicacions té la contestació a aquesta pregunta en el significat del R^2 en aquest model?

Exercici 3.10 Argumenteu la veracitat o falsedat de les següents afirmacions:

a) En un model de regressió lineal, la suma dels residus és zero.

b) El coeficient de determinació (R^2) és sempre una bona mesura de la qualitat del model.

c) L'estimador per mínims quadrats és un estimador esbiaixat.

Exercici 3.11 El següent model es formula per a explicar el temps emprat en dormir:

$$sleep = \beta_1 + \beta_2\ totalwrk + \beta_3 leisure + u$$

on el temps dedicat a dormir (*sleep*), al treball -remunerat i no remunerat- (*totalwrk*), i l'oci (*leisure*) (temps no dedicat a dormir o treballar) estan mesurats en minuts per dia.

L'equació estimada amb una mostra de 200 observacions, utilitzant el fitxer *timuse03*, és la següent:

$$\widehat{sleep} = 1440 - 1 \times\ total\_work - 1 \times leisure$$

$$R^2=1.000 \qquad n=1000$$

a) Quina és la seua opinió sobre aquests resultats?
b) Quin és el significat del terme independent estimat?

Exercici 3.12 Utilitzant una submostra de l'Enquesta d'Estructura Salarial per a Espanya en 2006 (arxiu *wage06sp*) es va estimar el següent model per a explicar el salari (*wage*):

$$\ln(wage) = 1.565 + 0.0448educ + 0.0177tenure + 0.0065age$$

$$R^2=0.337 \quad n=800$$

on educació (*educ*), permanència en l'empresa (*tenure*) i edat (*age*) estan mesurats en anys i el salari en euros per hora.

a) Quina és la interpretació dels coeficients *educ*, *tenure* i *age*?
b) Quants anys ha d'augmentar l'edat perquè tinga un efecte similar a l'increment d'1 any en l'educació, mantenint fixos en cada cas, els altres dos regressors?
c) Sabent que $\overline{educ}=10.2$, $\overline{tenure}=7.2$ i $\overline{age}=42.0$, calculi les elasticitats dels salaris pel que fa a l'educació, permanència en l'empresa i edat, mantenint fixos els altres regressors. Considera vostè que aquestes elasticitats són altes o baixes?

Exercici 3.13 La següent equació descriu el preu de l'habitatge en termes del nombre de dormitoris de la casa (*bedrooms*), del nombre de banys complets (*bathrms*) i de la mida de la parcel·la en peus quadrats (*lotsize*):

$$price = \beta_1 + \beta_2 bedrooms + \beta_3 bathrms + \beta_4 lotsize + u$$

on el preu (*price*) de l'habitatge es mesura en dòlars.

Utilitzant les dades de la ciutat de Windsor continguts en el fitxer *housecan*, s'estima el següent model:

$$price = -2418 + 5827bedrooms + 19750bathrms + 5.411lotsize$$

$$R^2=0.486 \quad n=546$$

a) Quin és l'augment estimat en el preu d'una casa amb un dormitori i un bany addicionals, mantenint *lotsize* constant?
b) Quin percentatge de variació en el preu s'explica pel nombre de dormitoris, el nombre de banys complets i la mida de l'habitatge en conjunt?
c) Determineu el preu de venda predit per a una casa de la mostra amb *bedrooms* = 3, *bathrms* = 2 i *lotsize* = 3880.
d) El preu de venda real de la casa al c) va ser de 66000 $. Trobeu el valor del residu per a aquesta casa. A la vista d'aquest resultat, el comprador va pagar de més o de menys per la casa?

Exercici 3.14 Per examinar els efectes dels rendiments de les empreses sobre els salaris dels seus consellers delegats es va formular el següent model:

$$\ln(salary) = \beta_1 + \beta_2 roa + \beta_3 \ln(sales) + \beta_4 profits + \beta_5 tenure + u$$

on *roa*, és la ràtio beneficis/actius expressada en percentatge i *tenure* és el nombre d'anys a l'empresa com a conseller delegat (= 0 si és menor de 6 mesos). El salari (*salary*) està expressat en milers de dòlars, mentre que les vendes (*sales*) i els beneficis (*profits*) estan en milions de dòlars.

S'ha utilitzat el fitxer *ceoforbes* per a l'estimació del model. Aquest fitxer conté dades sobre 447 executius de les 500 empreses més grans dels EUA (52 de les 500 empreses van ser excloses per manca de dades sobre una o més variables. Apple Computer també va ser exclòs perquè Steve Jobs, conseller delegat d'Apple el 1999, no va rebre cap compensació durant aquest període.) Les dades de les empreses provenen de la revista Fortune i es refereixen a 1999, les dades dels consellers delegats provenen de la revista Forbes i es refereixen també a 1999. Els resultats obtinguts van ser els següents:

$$\ln(salary) = 4.641 + 0.0054 roa + 0.2893 \ln(sales) + 0.0000564\, profits + 0.0122 tenure$$

$$R^2 = 0.232 \qquad n = 447$$

a) Interpreteu el coeficient del regressor *roa*.
b) Interpreteu el coeficient del regressor ln*(sales*). Quina és la seva opinió sobre la magnitud de l'elasticitat del *salary*/*sales*?
c) Interpreteu el coeficient del regressor *profits*.
d) Quina és l'elasticitat de *salary*/*profits* per al punt de les mitjanes mostrals ($\overline{salary}$ =2028 i $\overline{profits}$ =700).

Exercici 3.15 (Continuació de l'Exercici 2.21) Utilitzant una base de dades de 1983 empreses enquestada a l'any 2006 (fitxer *rdspain*), es va estimar la següent equació:

$$rdintens = -1.8168 + 0.1482 \ln(sales) + 0.0110 exponsal$$

$$R^2 = 0.048 \qquad n = 1983$$

on *rdintens* és la despesa en recerca i desenvolupament (R+D) expressat com a percentatge de les vendes, les vendes (sales) es mesuren en milions d'euros, i *exponsal* són les exportacions preses com a percentatge de les vendes.

a) Interpreteu el coeficient de ln(*sales*). En particular, si les vendes augmenten en un 100%, quin és el percentatge de canvi estimat de rdintens? És aquest un efecte econòmic gran?
b) Interpreteu el coeficient d'*exponsal*. És gran aquest coeficient des d'un punt de vista econòmic?
c) Quin percentatge de la variació en *rdintens* s'explica per les vendes i per les exportacions preses com a percentatge de les vendes?
d) Quina és l'elasticitat *rdintens*/*sales* per a la mitjana mostral ($\overline{rdintens}$ = 0.732 i $\overline{sales}$ =63544960). Comenteu el resultat.
Quina és l'elasticitat *rdintens*/*exponsal* per a la mitjana mostral.
($\overline{rdintens}$ = 0.732 i $\overline{exponsal}$ 17.657)? Comenteu el resultat.

Exercici 3.16 La següent regressió hedònica es va formular per a explicar els preus dels cotxes (vegeu exemple 3.3):

$$\ln(price) = \beta_1 + \beta_2 cid + \beta_3 hpweight + \beta_4\, fueleff + u$$

on *cid*, és el desplaçament en polzades cúbiques, *hpweight* és la ràtio potència/pes en kg, expressada en percentatge i *fueleff* és la ràtio litres per 100 km/cavalls de vapor expressada en percentatge.

a) Quins són els signes probables de β_2, β_3 i β_4? Explique.

b) Estimeu el model utilitzant el fitxer *hedcarsp* i expresse els resultats en forma d'equació.
c) Interpreteu el coeficient del regressor *cid*.
d) Interpreteu el coeficient del regressor *hpweight*.
e) Expandiu el model, introduint un regressor relatiu a la mida de cotxe, com el volum o el pes. Què passa si s'introdueixen els dos a la regressió? Quina creu que és la relació entre el pes i el volum?

Exercici 3.17 El concepte de treball cobreix un ampli espectre d'activitats possibles en l'economia productiva. Una part important del treball és no remunerat, no passa pel mercat i, per tant, no té preu. El treball no remunerat més important és el treball realitzat a la llar (*houswork*) dut a terme principalment per dones. Per tal d'analitzar els factors que influeixen en el treball de la llar, s'ha formulat el següent model:

$$houswork = \beta_1 + \beta_2 educ + \beta_3 hhinc + \beta_4 age + \beta_5 paidwork + u$$

on *educ* són els anys d'educació assolits, *hhinc* són els ingressos de les llars en euros per mes, *age* és l'edat de la persona entrevistada i *paidwork* és el treball remunerat. Les variables *houswork* i *paidwork* estan mesurades en minuts per dia.

Utilitzeu les dades contingudes en el fitxer *timuse03* per estimar el model. Aquest fitxer conté 1000 observacions corresponents a una submostra aleatòria extreta de l'enquesta d'Ús del Temps a Espanya que es va dur a terme en el període 2002-2003

a) Quins signes esperaria per β_2, β_3, β_4 i β_5? Explique.
b) Expresseu els resultats en forma d'equació
c) Creu vostè que hi ha factors rellevants omesos en l'equació anterior? Explique.
d) Interpretació dels coeficients dels regressors *educ*, *hhinc*, *age* i *paidwork*.

Exercici 3.18 (Continuació de l'Exercici 2.20) Per a explicar la satisfacció general de les persones (*stsfglo*) es formula el següent model:

$$stsfglo = \beta_1 + \beta_2 gnipc + \beta_3 lifexpec + u$$

on *gnipc* és la renda nacional bruta per càpita expressada en dòlars (USA) PPA (paritat del poder adquisitiu) a preus de 2008 i *lifexpec* és l'esperança de vida en néixer, és a dir, el nombre d'anys que un nadó pot esperar viure. Quan una magnitud s'expressa en dòlars nord-americans PPA, això vol dir que s'ha convertit a dòlars internacionals usant taxes PPA. (Un dòlar internacional té el mateix poder adquisitiu que un dòlar dels EUA als Estats Units.)

Utilitzeu el fitxer *HDR2010* per a l'estimació del model.

a) Quins signes esperaria pe β_2 i β_3? Explique.
b) Quina seria la de satisfacció global mitjana d'un país els habitants tenen una esperança de vida en néixer de 80 anys i que tenen una renda nacional bruta per càpita de 30.000 dòlars PPA expressats en dòlars de 2008 d'Estats Units?
c) Interpretació dels coeficients de *gnipc* i *lifexpe*.
d) Tenint en compte un país l'esperança de vida en néixer és igual a 50 anys, quina hauria de ser la renda nacional bruta per càpita per obtenir una satisfacció global igual a 5?

Exercici 3.19 (Continuació de l'Exercici 2.24) A causa dels problemes sorgits en el model keynesià, Brown va introduir en la funció de consum, a més de la renda, el consum retardat per reflectir la persistència d'hàbits del consumidor:

$$conspc = \beta_1 + \beta_2 incpc + \beta_3 conspc(-1) + u$$

Com en aquest model s'inclou el consum retardat, cal distingir entre propensió marginal al consum a curt termini i a llarg termini. La propensió marginal a curt termini es calcula de la mateixa manera que en la funció de consum de Keynes. Per calcular la propensió marginal a llarg termini s'ha de considerar una situació d'equilibri, en la qual no hi ha variacions en les variables. Es designen $conspc^e$ i $incpc^e$ al consum i a la renda d'equilibri i prescindint de la pertorbació aleatòria, el model anterior en situació d'equilibri ve donat per

$$conspc^e = \beta_1 + \beta_2 incpc^e + \beta_3 conspc^e$$

La funció de consum de Brown es va estimar amb dades de l'economia espanyola per al període 1954-2010 (fitxer *consumsp*), obtenint-se els següents resultats:

$$conspc_t = -7.156 + 0.3965 incpc_t + 0.5771 conspc_{t-1}$$

$$R^2 = 0.997 \qquad n = 56$$

a) Interpreteu el coeficient d'*incpc*. En la seua interpretació, inclouria la clàusula "mantingut fix l'altre regressor? Justifiqueu la resposta.

b) Calculeu l'elasticitat a curt termini per a les mitjanes mostrals ($\overline{conspc}$ =8084, $\overline{incpc}$ =8896).

c) Calculeu l'elasticitat a llarg termini per a les mitjanes mostrals.

d) Comenteu la diferència entre els valors obtinguts per als dos tipus d'elasticitat.

Exercici 3.20 Per a explicar la influència dels incentius i les despeses de publicitat en les vendes, s'han formulat els models alternatius següents:

$$sales = \beta_1 + \beta_2 advert + \beta_3 incent + u \quad (1)$$

$$\ln(sales) = \beta_1 + \beta_2 \ln(advert) + \beta_3 \ln(incent) + u \quad (2)$$

$$\ln(sales) = \beta_1 + \beta_2 advert + \beta_3 incent + u \quad (3)$$

$$sales = \beta_2 advert + \beta_3 incent + u \quad (4)$$

$$\ln(sales) = \beta_1 + \beta_2 \ln(incent) + u \quad (5)$$

$$sales = \beta_1 + \beta_2 incent + u \quad (6)$$

a) Utilitzant una mostra de 18 àrees de venda (fitxer *advincen*), estime els models anteriors:

b) Dins de cada un dels següents grups seleccione el millor model, indicant quins han estat els criteris que s'han utilitzat. Justifiqueu la resposta.

b1) (1) i (6)
b2) (2) i (3)
b3) (1) i (4)
b4) (2), (3) i (5)

*b*5) (1), (4) i (6)
*b*6) (1), (2), (3), (4), (5) i (6)

Apèndixs

Apèndix 3.1 Demostració del Teorema de Gauss-Markov

Per demostrar aquest teorema, s'utilitzen els supòsits 1 a 8 del *MLC*.

Considerem un altre estimador $\tilde{\boldsymbol{\beta}}$ que és una funció de **y** (recordeu que $\hat{\boldsymbol{\beta}}$ és també una funció de **y**), donat per

$$\tilde{\boldsymbol{\beta}} = \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\mathbf{y} \tag{3-93}$$

on **A** és una matriu arbitrària, $k \times n$, que és funció de **X** i/o altres variables no estocàstiques, però no és funció de **y**. Perquè $\tilde{\boldsymbol{\beta}}$ siga no esbiaixat, s'han de complir certes condicions.

Tenint en compte (3-52), obtenim que

$$\tilde{\boldsymbol{\beta}} = \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right] = \boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\mathbf{u} \tag{3-94}$$

Prenent les esperances en els dos membres de (3-94), obtenim que

$$E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]E(\mathbf{u}) = \boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta} \tag{3-95}$$

Perquè $\tilde{\boldsymbol{\beta}}$ siga no esbiaixat, és a dir, $E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, s'ha de complir el següent:

$$\mathbf{A}\mathbf{X} = \mathbf{I} \tag{3-96}$$

Conseqüentment,

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\mathbf{u} \tag{3-97}$$

Tenint en compte els supòsits 7 i 8, i (3-96), la $Var(\tilde{\boldsymbol{\beta}})$ és igual a

$$\begin{aligned} Var(\tilde{\boldsymbol{\beta}}) &= E((\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})') = E\left[\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\mathbf{u}\mathbf{u}'\left[\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1} + \mathbf{A}'\right]\right] \\ &= E\left[\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{u}\mathbf{u}'\left[\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\right] + \mathbf{A}\mathbf{A}'\right] = \sigma^2\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1} + \mathbf{A}\mathbf{A}'\right] \end{aligned} \tag{3-98}$$

La diferència entre les dues variances és la següent:

$$Var(\tilde{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}) = \sigma^2\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1} + \mathbf{A}\mathbf{A}' - \left[\mathbf{X}'\mathbf{X}\right]^{-1}\right] = \sigma^2\mathbf{A}\mathbf{A}' \tag{3-99}$$

El producte d'una matriu per la seua transposada és sempre una matriu semidefinida positiva. Per tant,

$$Var(\tilde{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{A}\mathbf{A}' \geq 0 \tag{3-100}$$

La diferència entre la variança d'un estimador $\tilde{\boldsymbol{\beta}}$- -arbitrari, però lineal i no esbiaixat - i la variança de l'estimador $\hat{\boldsymbol{\beta}}$ és una matriu semidefinida positiva. En conseqüència $\hat{\boldsymbol{\beta}}$ és un *Estimador Lineal No Esbiaixat Òptim*, és a dir, és un estimador *ELNEO*.

Apèndix 3.2 Demostració: $\hat{\sigma}^2$ és un estimador no esbiaixat de la variança de la pertorbació

Per tal de veure quin és l'estimador més convenient de σ^2, es van a analitzar primer les propietats de la suma dels quadrats dels residus. Aquest és precisament el numerador de la variança residual.

Tenint en compte (3-17) i (3-23), anem a expressar el vector de residus com una funció del regressant

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y} = \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{y} = \mathbf{M}\mathbf{y} \tag{3-101}$$

on **M** és la matriu idempotent.

Alternativament, el vector de residus es pot expressar com una funció del vector de les pertorbacions:

$$\begin{aligned}\hat{\mathbf{u}} &= \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{y} = \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right] \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{u} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\boldsymbol{\beta}\mathbf{u} \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} + \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{u} = \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{u} \\ &= \mathbf{M}\mathbf{u}\end{aligned} \tag{3-102}$$

Tenint en compte (3-102), la suma dels quadrats dels residus (*SQR*) es pot expressar en la forma següent:

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u} = \mathbf{u}'\mathbf{M}\mathbf{u} \tag{3-103}$$

Ara bé, tenint en compte que estem buscant un estimador no esbiaixat de σ^2, calcularem l'esperança de l'expressió anterior:

$$\begin{aligned}E\left[\hat{\mathbf{u}}'\hat{\mathbf{u}}\right] &= E\left[\mathbf{u}'\mathbf{M}\mathbf{u}\right] = trE\left[\mathbf{u}'\mathbf{M}\mathbf{u}\right] = E\left[tr\mathbf{u}'\mathbf{M}\mathbf{u}\right] \\ &= E\left[tr\mathbf{M}\mathbf{u}\mathbf{u}'\right] = tr\mathbf{M}E\left[\mathbf{u}\mathbf{u}'\right] = tr\mathbf{M}\sigma^2\mathbf{I} \\ &= \sigma^2 tr\mathbf{M} = \sigma^2(n-k)\end{aligned} \tag{3-104}$$

En la deducció de (3-104), hem utilitzat la propietat de la traça que $tr(\mathbf{AB}) = tr(\mathbf{BA})$). Tenint en compte aquesta propietat de la traça, s'obté el valor de: $tr\mathbf{M}$:

$$\begin{aligned}tr\mathbf{M} &= tr\left[\mathbf{I}_{n\times n} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right] = tr\mathbf{I}_{n\times n} - tr\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' \\ &= tr\mathbf{I}_{n\times n} - tr\mathbf{I}_{k\times k} = n-k\end{aligned}$$

D'acord amb (3-104), es compleix que

$$\sigma^2 = \frac{E\left[\hat{\mathbf{u}}'\hat{\mathbf{u}}\right]}{n-k} \tag{3-105}$$

A la vista de (3-105), un estimador de la variança no esbiaixat vindrà donat per:

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k} \tag{3-106}$$

ja que, d'acord amb (3-104),

$$E(\hat{\sigma}^2) = E\left[\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}\right] = \frac{E(\hat{\mathbf{u}}'\hat{\mathbf{u}})}{n-k} = \frac{\sigma^2(n-k)}{n-k} = \sigma^2 \tag{3-107}$$

El denominador de (3-106) són els graus de llibertat que corresponen a la *SQCR* que apareix en el numerador. Aquest resultat es justifica pel fet que les equacions normals de l'hiperplà imposen *k* restriccions sobre els residus. Per tant, el nombre de graus de llibertat de la *SQR* és igual al nombre d'observacions (*n*) menys el nombre de restriccions *k*.

Apèndix 3.3 La consistència de l'estimador de *MQO*

A l'apèndix 2.8 hem provat en el model de regressió simple la consistència de l'estimador *MQO*. Ara anem a provar la consistència del vector $\hat{\boldsymbol{\beta}}$ obtingut per *MQO*.

En primer lloc, l'estimador de mínims quadrats $\hat{\boldsymbol{\beta}}$, donat en (3-23) pot expressar-se així

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{n}\mathbf{X'X}\right)^{-1}\left(\frac{1}{n}\mathbf{X'u}\right) \tag{3-108}$$

Ara, prenem límits a l'últim factor de (3-108) i anomenem **Q** al resultat:

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{X'X} = \mathbf{Q} \tag{3-109}$$

Si **X** és fixa en mostres repetides, d'acord amb el supòsit 2, llavors (3-109) implica que **Q**=(l/*n*)**X'X**. D'acord amb el supòsit 3, i pel fet que la matriu inversa és una funció contínua de la matriu original, hi ha $\mathbf{Q}^{-1}$. Per tant, podem escriure que

$$\text{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{Q}^{-1}\text{plim}\left[\frac{1}{n}\mathbf{X'u}\right]$$

L'últim terme de (3-108) es pot expressar com

$$\frac{1}{n}\mathbf{X'u}=\frac{1}{n}\begin{bmatrix}1 & 1 & \cdots & 1 & \cdots & 1\\ x_{21} & x_{22} & \cdots & x_{2i} & \cdots & x_{2n}\\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\ x_{j1} & x_{j2} & \cdots & x_{ji} & \cdots & x_{jn}\\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\ x_{k1} & x_{k2} & \cdots & x_{ki} & \cdots & x_{kn}\end{bmatrix}\begin{bmatrix}u_1\\ u_2\\ \vdots\\ u_i\\ \vdots\\ u_n\end{bmatrix}$$

$$=\frac{1}{n}\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_i & \cdots & \mathbf{x}_n\end{bmatrix}\begin{bmatrix}u_1\\ u_2\\ \vdots\\ u_i\\ \vdots\\ u_n\end{bmatrix}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i=\overline{\mathbf{x}_i u_i} \qquad (3\text{-}110)$$

on $\mathbf{x}_i$ és el vector de la columna corresponent a l'observació *i-èsima*.

Ara, anem a calcular l'esperança i la variança de (3-110),

$$E\left[\overline{\mathbf{x}_i u_i}\right]=\frac{1}{n}\sum_{i=1}^{n}E\ \mathbf{x}_i u_i\ =\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i E\ u_i\ =\frac{1}{n}\mathbf{X'}E\ \mathbf{u}\ =\mathbf{0} \qquad (3\text{-}111)$$

$$var\left[\overline{\mathbf{x}_i u_i}\right]=E\left[\overline{\mathbf{x}_i u_i}(\overline{\mathbf{x}_i u_i})'\right]=\frac{1}{n}\mathbf{X'}E\ \mathbf{uu}'\ \mathbf{X}=\frac{1}{n}\frac{\sigma^2}{n}\frac{\mathbf{X'X}}{n}=\frac{\sigma^2}{n^2}\mathbf{Q} \qquad (3\text{-}112)$$

ja que $E\ \mathbf{uu}'=\sigma^2\mathbf{I}$ d'acord amb els supòsits 7 i 8.

Prenent límits en (3-112), se segueix aleshores que

$$\lim_{n\to\infty}\text{var}\left[\overline{\mathbf{x}_i\mathbf{u}_i}\right]=\lim_{n\to\infty}\frac{\sigma^2}{n^2}\mathbf{Q}=0(\mathbf{Q})=\mathbf{0} \qquad (3\text{-}113)$$

Donat que l'esperança de $\overline{\mathbf{x}_i\mathbf{u}_i}$ és idèntica a zero i la seva variança convergeix a zero, $\overline{\mathbf{x}_i\mathbf{u}_i}$ convergeix en mitjana quadràtica a zero. La convergència en mitjana quadràtica implica la convergència en probabilitat, de manera que plim($\overline{\mathbf{x}_i\mathbf{u}_i}$)=0. Per tant,

$$\text{plim}(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta}+\mathbf{Q}^{-1}\text{plim}(\mathbf{x}_i\mathbf{u}_i)=\boldsymbol{\beta}+\mathbf{Q}^{-1}\text{plim}\left[\frac{1}{n}\mathbf{X'u}\right]=\boldsymbol{\beta}+\mathbf{Q}^{-1}\times 0=\boldsymbol{\beta} \qquad (3\text{-}114)$$

En conseqüència, $\hat{\beta}$ és un estimador consistent.

Apèndix 3.4 Estimador de màxima versemblança

El mètode de màxima versemblança és un mètode àmpliament utilitzat en econometria. Aquest mètode proposa com estimadors dels paràmetres aquells valors per als quals la probabilitat d'obtenir les observacions donades és màxima. En l'estimació de mínims quadrats no es va adoptar *a priori* cap supòsit; per contra, l'estimació per màxima

versemblança requereix establir *a priori* supòsits estadístiques sobre els diversos elements del model. Així, en l'estimació per màxima versemblança anem a adoptar tots els supòsits del model lineal clàssic (*MLC*).

Per tant, en l'estimació per màxima versemblança de $\boldsymbol{\beta}$ i σ^2 en el model (3-52), es prenen com estimadors a aquells valors que maximitzen la probabilitat d'obtenir les observacions d'una mostra donada.

Anem a veure el procediment per obtenir els estimadors de màxima versemblança $\boldsymbol{\beta}$ i σ^2. D'acord amb els supòsits del *MLC*

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \tag{3-115}$$

L'esperança i la variança de la distribució de $\mathbf{y}$ estan donades per

$$E(\mathbf{y}) = E\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right] = \mathbf{X}\boldsymbol{\beta} + E(\mathbf{u}) = \mathbf{X}\boldsymbol{\beta} \tag{3-116}$$

$$\text{var}(\mathbf{y}) = E\left[\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\right] = E\left[\mathbf{u}\mathbf{u}'\right] = \sigma^2\mathbf{I} \tag{3-117}$$

Per tant,

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \tag{3-118}$$

La funció de densitat de $\mathbf{y}$ (o funció de versemblança), considerant $\mathbf{X}$ i $\mathbf{y}$ fixes i $\boldsymbol{\beta}$ i σ^2 variables, serà d'acord amb (3-118) igual a

$$L = f(y \mid \boldsymbol{\beta}, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left(-\left(1/2\sigma^2\right)\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right) \tag{3-119}$$

Atès que el màxim per a L s'aconsegueix en el mateix punt que $\ln(L)$, per ser la funció logaritme monòtona, podem, a efectes de maximització, treballar amb $\ln(L)$ en lloc de L. Aleshores,

$$\ln(L) = -\frac{n\ln(2\pi)}{2} - \frac{n\ln(\sigma^2)}{2} - \frac{1}{2\sigma^2} - \mathbf{(y - X\beta)'(y - X\beta)} \tag{3-120}$$

Per a maximitzar $\ln(L)$, cal derivar respecte a $\boldsymbol{\beta}$ i σ^2:

$$\frac{\delta \ln(L)}{\delta \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(-2\mathbf{X'y} + 2\mathbf{X'X\beta}) \tag{3-121}$$

$$\frac{\delta \ln(L)}{\delta \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\mathbf{(y - X\beta)'(y - X\beta)}}{2\sigma^4} \tag{3-122}$$

Igualant (3-121) a zero, veiem que l'estimador de màxima versemblança de $\boldsymbol{\beta}$, denotat per $\tilde{\boldsymbol{\beta}}$, satisfà que

$$\mathbf{X'X}\tilde{\beta} = \mathbf{X'y} \tag{3-123}$$

Com se suposa que $\mathbf{X'X}$ és invertible, obtenim

$$\tilde{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1} = \mathbf{X}'\mathbf{y} \tag{3-124}$$

En conseqüència, l'estimador de màxima versemblança de **β**, sota els supòsits del *MLC*, coincideix amb l'estimador *MQO*, és a dir,

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} \tag{3-125}$$

Per tant,

$$(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\mathbf{u}}'\hat{\mathbf{u}} \tag{3-126}$$

Igualant (3-122) a zero i substituint **β** per $\tilde{\boldsymbol{\beta}}$, obtenim

$$-\frac{n}{2\tilde{\sigma}^2} + \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{2\tilde{\sigma}^4} = 0 \tag{3-127}$$

on hem designat per $\tilde{\sigma}^2$ l'estimador de màxima versemblança de la variança de les pertorbacions aleatòries. De (3-127) es dedueix que

$$\tilde{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n} \tag{3-128}$$

Com es pot veure l'estimador de màxima versemblança no és igual a l'estimador no esbiaixat que s'ha obtingut en (3-106). De fet, si prenem esperances en (3-128),

$$E\left[\tilde{\sigma}^2\right] = \frac{1}{n}E\left[\hat{\mathbf{u}}'\hat{\mathbf{u}}\right] = \frac{n-k}{n}\sigma^2$$

És a dir, l'estimador de màxima versemblança, $\tilde{\sigma}^2$, és un estimador esbiaixat, encara que el seu biaix tendeix a zero quan *n* tendeix a infinit, ja que

$$\lim_{n\to\infty}\frac{n-k}{n} = 1 \tag{3-130}$$

4 CONTRAST D'HIPÒTESIS EN EL MODEL DE REGRESSIÓ MÚLTIPLE

4.1 El contrast d'hipòtesis: una panoràmica

Abans de contrastar les hipòtesis en el model de regressió múltiple, oferirem una panoràmica sobre el contrast d'hipòtesis.

El contrast d'hipòtesis permet realitzar inferències sobre paràmetres poblacionals utilitzant dades provinents d'una mostra. Per realitzar el contrast d'hipòtesis estadístic, en general, cal realitzar els següents passos:

1) Establir una hipòtesi nul·la i una hipòtesi alternativa relatives als paràmetres de la població.

2) Construir un estadístic per contrastar les hipòtesis formulades.

3) Definir una regla de decisió per determinar si la hipòtesi nul·la ha de ser, o no, rebutjada en funció del valor que prenga l'estadístic construït.

Anem a examinar a continuació cada un d'aquests passos.

4.1.1 Formulació de la hipòtesi nul·la i de la hipòtesi alternativa

Abans de referir-nos a la manera de formular la hipòtesi nul·la i alternativa, distingirem entre *hipòtesis simples* i *hipòtesis compostes*. Les hipòtesis que es formulen mitjançant una o més igualtats s'anomenen hipòtesis simples. Quan per a formular una hipòtesi s'utilitzen els operadors "desigualtat", "més gran que" i "menor que", aleshores a aquesta hipòtesi se li denomina composta.

És important assenyalar que el contrast d'hipòtesis es refereix sempre als paràmetres poblacionals. El contrast d'hipòtesis implica prendre la decisió, sobre la base de les dades mostrals, de rebutjar o no que certes restriccions siguen satisfetes pel model bàsic assumit. Les restriccions que es volen contrastar es coneixen com la *hipòtesi nul·la*, a la qual designa H_0. Així doncs, una hipòtesi nul·la és una declaració sobre els paràmetres poblacionals.

Encara que és possible formular hipòtesis nul·les compostes en el context del model de regressió, sempre considerarem que la hipòtesi nul·la és una hipòtesi simple. És a dir, per a formular una hipòtesi nul·la, utilitzarem sempre l'operador "igualtat". Vegem a continuació alguns exemples d'hipòtesis nul·les referides al model de regressió:

a) $H_0 : \beta_1=0$

b) $H_0 : \beta_1+ \beta_2=0$

c) $H_0 : \beta_1=\beta_2=0$

d) $H_0 : \beta_2+\beta_3=1$

També definirem una hipòtesi alternativa, designada per H_1, que representa la nostra conclusió en el cas que el contrast concloga que H_0 és falsa.

Tot i que les hipòtesis alternatives poden ser també simples o compostes, en el model de regressió, prendrem sempre, com hipòtesi alternativa, una hipòtesi composta. La hipòtesi alternativa, a la qual es designa H_1, es formula mitjançant l'operador "desigualtat" en la major part dels casos. Així, per exemple, donada la H_0:

$$H_0 : \beta_j = 1 \tag{4-1}$$

podem formular la següent H_1:

$$H_1 : \beta_j \neq 1 \tag{4-2}$$

que és una hipòtesi "alternativa de dues cues".

Les següents hipòtesis es diuen "alternatives d'una cua":

$$H_1 : \beta_j < 1 \tag{4-3}$$

$$H_1 : \beta_j > 1 \tag{4-4}$$

4.1.2 Estadístic de contrast

Un estadístic de contrast és una funció d'una mostra aleatòria, de manera que també és una variable aleatòria. Quan es calcula l'estadístic de contrast per a una mostra donada, s'obté un resultat, és a dir, un nombre. Al realitzar un contrast estadístic seria convenient conèixer la distribució de l'estadístic de contrast sota la hipòtesi nul·la. Aquesta distribució depèn en gran mesura de les hipòtesis formulades en el model. Si en l'especificació del model s'assumeix el supòsit de normalitat, aleshores la distribució estadística apropiada serà la distribució normal o alguna de les distribucions associades a aquesta, com són la *Chi-quadrat*, la *t de Student*, o la *F de Snedecor*.

En el quadre 4.1 es mostren algunes distribucions, que són apropiades en diferents situacions, sota el supòsit de normalitat de les pertorbacions del model.

QUADRE 4.1. Algunes distribucions utilitzades en el contrast d'hipòtesis.

	1 restricció	*1 o més restriccions*
σ^2 *coneguda*	*N*	*Chi-quadrat*
σ^2 *desconeguda*	*t* de Student	*F* de Snedecor

L'estadístic utilitzat per al contrast es construeix tenint en compte la H_0 i les dades mostrals. A la pràctica, com σ^2 és sempre desconeguda, s'utilitzaran sempre les distribucions *t* i *F*.

4.1.3 Regla de decisió

Per al contrast d'hipòtesis considerarem dos enfocaments: l'enfocament clàssic i un enfocament alternatiu basat en els valors-*p*. Però abans de veure la manera d'aplicar la regla de decisió, examinarem els tipus d'error que es poden cometre en el contrast d'hipòtesis.

Tipus d'errors en el contrast d'hipòtesis

En el contrast d'hipòtesis, podem cometre dos tipus d'errors: *error de tipus* I i *error de tipus* II.

Error de tipus I

Nosaltres podem rebutjar la H_0 quan en realitat és certa. A aquest error se li denomina *error de tipus* I. Generalment, es defineix el nivell de significació (α) d'un contrast com la probabilitat de cometre un error de tipus I. Simbòlicament,

$$\alpha = \Pr(\textit{rebutjar } H_0 \mid H_0) \tag{4-5}$$

Dit d'una altra manera, el nivell de significació és la probabilitat de rebutjar la H_0 quan la H_0 és certa. Les regles per al contrast d'hipòtesis es construeixen fent que la probabilitat d'un *error de tipus I* siga prou xicoteta. Els nivells de significació usuals per a són 0.10, 0.05 i 0.01. Algunes vegades també s'utilitza 0.001.

Després d'haver pres la decisió de rebutjar o no la H_0, la decisió pot haver estat la correcta o pot ser que s'haja comès un error. Mai sabrem amb certesa si es va cometre un error. No obstant això, podem calcular la probabilitat d'haver comès un *error de tipus I* o un *error de tipus II*.

Error de tipus II

Podem fracassar a rebutjar la H_0 quan en realitat és falsa. Si això succeeix s'ha comès un *error de tipus* II.

$$\beta = \Pr(\textit{No rechazar } H_0 \mid H_1) \tag{4-6}$$

En altres paraules, β és la probabilitat de no rebutjar H_0 quan H_1 és certa.

És important assenyalar que no és possible minimitzar els dos tipus d'error de forma simultània. A la pràctica el que fem és seleccionar.

Enfocament clàssic: Aplicació de la regla de decisió

El mètode clàssic implica els següents passos:

a) *Elecció de* α. El contrast d'hipòtesis clàssic requereix que inicialment s'especifique un *nivell de significació*. Quan s'especifica un valor per *a*, essencialment el

que estem quantificant és la nostra tolerància per a un *error de tipus* I. Si α= 0.05, aleshores l'investigador està disposat a rebutjar H_0 falsament en un 5% dels casos.

b) *Obtenció de c*, *valor crític*, utilitzant taules estadístiques. El valor *c* es determina pel valor de α.

El valor crític en un contrast d'hipòtesis és un llindar amb el qual es compara l'estadístic de contrast per determinar si la hipòtesi nul·la es rebutja o no.

c) *Comparant el resultat de l'estadístic de contrast, s,* amb *c*, la H_0 es rebutja o no per a un valor donat de α.

La regió de rebuig (*RR*), delimitada pel valor crític *(c*), és un conjunt de valors de l'estadístic de contrast per als quals es rebutja la hipòtesi nul·la. (Vegeu figura 4.1). És a dir, l'espai mostral de l'estadístic de contrast es divideix en dues regions: una regió (la regió de rebuig) ens porta a rebutjar la hipòtesi nul·la H_0, mentre que l'altra no ens deixa rebutjar la hipòtesi nul·la. Per tant, si el valor observat de l'estadístic de contrast *s* es troba a la regió crítica, rebutgem la H_0; en el cas que no es troba a la regió de rebuig arribem a la conclusió, de *no rebutjar* la H_0 o de *fracassar a rebutjar* la H_0.

Simbòlicament,

$$\begin{array}{llll} \text{Si} & s \geq c & \text{es rebutja} & H_0 \\ \text{Si} & s < c & \text{no es rebutja} & H_0 \end{array} \qquad (4\text{-}7)$$

Si la hipòtesi nul·la es rebutja amb l'evidència de la mostra, aquesta és una conclusió *forta*. No obstant això, l'acceptació de la hipòtesi nul·la és una conclusió *feble* perquè no coneixem quina és la probabilitat de no rebutjar la hipòtesi nul·la quan ha de ser rebutjada. És a dir, no coneixem la probabilitat de cometre un *error de tipus* II. Per tant, en lloc d'utilitzar l'expressió de l'acceptació de la hipòtesi nul·la, és més correcte dir *fracassar a rebutjar* la hipòtesi nul·la, o *no rebutjar*, ja que el que realment passa és que no tenim prou evidència empírica per rebutjar la hipòtesi nul·la.

En el procés de contrastació, la part més subjectiva és la determinació *a priori* del nivell de significació. Quins criteris es poden utilitzar per determinar α? En general, es tracta d'una decisió arbitrària, encara que com ja hem dit, els nivells d'1%, 5% i 10% per α són els més utilitzats a la pràctica. De vegades s'efectua el contrast condicionat a diferents nivells de significació.

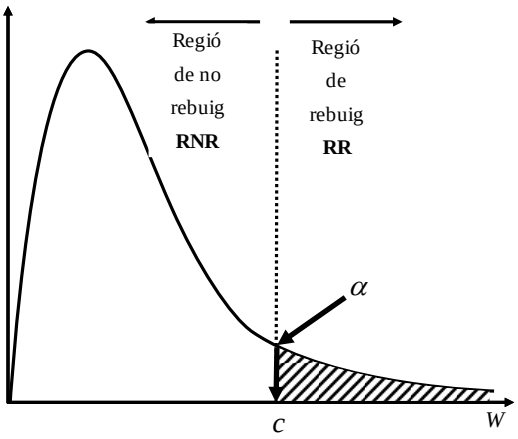


FIGURA 4.1. Contrast d'hipòtesis: enfocament clàssic.

Un enfocament alternatiu: el valor-p

Amb la utilització d'ordinadors, el contrast d'hipòtesis pot contemplar-se des d'una altra perspectiva molt més racional. Així, els programes d'ordinador solen oferir, al costat de l'estadístic de contrast una probabilitat. Aquesta probabilitat, a la qual se li denomina *valor-p* (*p*-value) -és a dir, valor de probabilitat-, també és coneguda com nivell de significació crític o exacte, o probabilitat exacta de cometre un *error de tipus* I. Més tècnicament, el *valor-p* es defineix com el més baix nivell de significació al que pot ser rebutjada una hipòtesi nul·la.

Una vegada que el *valor-p* ha estat determinat, sabem que la hipòtesi nul·la es rebutja per a qualsevol nivell de significació $\alpha \geq$ *valor-p*; per contra, la hipòtesi nul·la no es rebutja quan $\alpha <$ *valor-p*. Per tant, el *valor-p* és un indicador del nivell d'admissibilitat de la hipòtesi nul·la: com més gran siga el *valor-p*, més confiança podem tenir en la hipòtesi nul·la. L'ús de *valor-p* canvia per complet l'enfocament en el contrast d'hipòtesis. Així, en lloc de fixar *a priori* el nivell de significació, es calcula el *valor-p*, que ens permet determinar els nivells de significació per als que es rebutja la hipòtesi nul·la.

A les seccions següents veurem a la pràctica l'ús de *valor-p* en el contrast d'hipòtesis.

4.2 Contrast d'hipòtesis utilitzant l'estadístic *t*

4.2.1 Contrast d'un sol paràmetre

L'estadístic t

Sota els supòsits de l'*MCL* de l'1 al 9,

$$\hat{\beta}_j \sim N\left[\beta_j, \text{var}(\hat{\beta}_j)\right] \qquad j = 1, 2, 3, \cdots, k \tag{4-8}$$

Si tipifiquem

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N[0,1] \qquad j = 1, 2, 3, \cdots, k \tag{4-9}$$

El supòsit de normalitat es recolza en el Teorema del Límit Central (*TCL*), però aquest teorema és restrictiu en alguns casos. És a dir, la normalitat no sempre es pot assumir. En qualsevol aplicació, assumir o no el supòsit de normalitat de *u* és realment una qüestió empírica. Sovint, mitjançant una transformació, per exemple, prenent logaritmes, s'obté una distribució que està més propera a la normalitat i que és més fàcil de gestionar des d'un punt de vista matemàtic. Les mostres grans ens permeten prescindir del supòsit de normalitat sense afectar massa els resultats.

Sota els supòsits del *MLC* de l'1 al 9, s'obté una distribució *t* de Student

$$\frac{\hat{\beta}_j - \beta_j}{ee(\hat{\beta}_j)} \sim t_{n-k} \tag{4-10}$$

on *k* és el nombre de paràmetres desconeguts en el model poblacional (*k*-1 paràmetres del pendent i el terme independent, β_1). L'expressió (4-10) és important perquè ens permet contrastar la hipòtesi sobre β_j.

Si comparem (4-10) amb (4-9), veiem que la distribució de *t* de Student deriva del fet que el paràmetre σ de $ee(\hat{\beta}_j)$ ha estat reemplaçat pel seu estimador $\hat{\sigma}$, que és una variable aleatòria. Així doncs, els graus de llibertat de la t són *n-k*, corresponents als graus de llibertat utilitzats en l'estimació de $\hat{\sigma}^2$.

Quan una distribució *t* té molts graus de llibertat (*gl*) s'aproxima a una distribució normal estàndard. A la figura 4.2 s'ha representat la funció de densitat normal i la funció de densitat de la *t* per a diferents graus de llibertat. Com es pot veure, les funcions de densitat de la *t* són més aplanades (platicúrtiques) i amb les cues més amples que la funció de densitat normal, però a mesura que augmenten els *gl*, la funció de densitat de la *t* està més pròxima de la funció de densitat normal. De fet, el que passa és que la distribució *t* té en compte que s'ha estimat per ser desconeguda. Donada aquesta incertesa, la distribució de la *t* s'estén més que la de la normal. No obstant això, quan els *gl* creixen, la distribució *t* està més a prop de la distribució normal perquè la incertesa de no conèixer σ^2 disminueix.

Per tant, s'hauria de tenir en ment la següent convergència en distribució:

$$t_n \xrightarrow[n\to\infty]{} N(0,1) \tag{4-11}$$

Així doncs, quan el nombre de graus de llibertat d'una *t* de Student tendeix cap a infinit convergeix cap a una distribució *N*(0,1). En el context del contrast d'hipòtesis, si creix la mida de la mostra, també ho faran els graus de llibertat. Això implica que per mides grans es pot utilitzar, de forma pràcticament equivalent, la distribució normal per contrastar hipòtesis amb una sola restricció, encara que no es conega la variança poblacional. Com a regla pràctica, quan els *gl* són més grans que 120, es poden agafar valors crítics de la distribució normal.

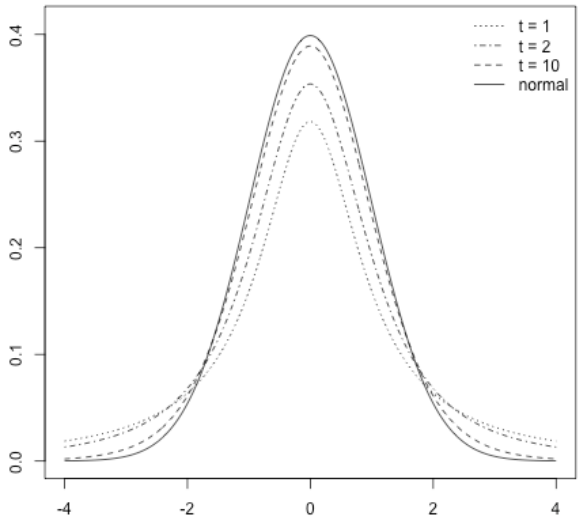


FIGURA 4.2. Funcions de densitat: normal i *t* per a diferents graus de llibertat

Considere hipòtesi nul·la,

$$H_0 : \beta_j = 0$$

Ja que β_j mesura l'efecte parcial de x_j sobre y, després de controlar per a totes les altres variables independents, $H_0 : \beta_j = 0$ vol dir que, una vegada que x_2, x_3, …,x_{j-1}, x_{j+1},…, x_k han estat tinguts en compte, x no té efecte sobre y. Aquesta H_0 correspon al denominat *contrast de significativitat*. L'estadístic que s'utilitza per contrastar $H_0 : \beta_j = 0$ contra qualsevol altra alternativa, s'anomena l'*estadístic t*, o la *ràtio t*, de $\hat{\beta}_j$ i s'expressa com

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{ee(\hat{\beta}_j)}$$

Quan anem a contrastar $H_0 : \beta_j = 0$ és natural tenir present a $\hat{\beta}_j$, el nostre estimador no esbiaixat de β_j. En una mostra donada $\hat{\beta}_j$ mai serà zero exactament, però un valor xicotet indicarà una hipòtesi nul·la veritable, mentre que un valor gran indicarà una hipòtesi nul·la falsa. La pregunta és: ¿fins a quin punt $\hat{\beta}_j$ està allunyada de zero?

Hem de recordar que hi ha un error de mostreig en l'estimació de $\hat{\beta}_j$, de manera que la mida de $\hat{\beta}_j$ s'ha de comparar amb el seu error de mostreig. Això és precisament el que fem quan utilitzem $t_{\hat{\beta}_j}$, ja que aquest estadístic mesura quants errors estàndard està $\hat{\beta}_j$ allunyada de zero. A fi de determinar un regle per rebutjar H_0, hem de decidir sobre la hipòtesi alternativa rellevant. Hi ha tres possibilitats: hipòtesis alternatives unilaterals (cua dreta i esquerra) i hipòtesi alternativa de dues cues.

Hipòtesi alternativa d'una cua: dreta

En primer lloc, anem a considerar la hipòtesi nul·la

$$H_0 : \beta_j = 0$$

contra la hipòtesi alternativa

$$H_1 : \beta_j > 0$$

Aquest és un contrast de *significació positiva*. La regla de decisió és en aquest cas la següent:

Regla de decisió

Si	$t_{\hat{\beta}_j} \geq t^{\alpha}_{n-k}$	es rebutja	H_0	(4-12)
Si	$t_{\hat{\beta}_j} < t^{\alpha}_{n-k}$	no es rebutja	H_0	

Per tant, rebutgem $H_0 : \beta_j = 0$ en favor de $H_1 : \beta_j > 0$ quan $t_{\hat{\beta}_j} \geq t^{\alpha}_{n-k}$ com es pot veure a la figura 4.3. Està molt clar que per rebutjar H_0 contra $H_1 : \beta_j > 0$ el valor de $t_{\hat{\beta}_j}$ ha de ser positiu. Un resultat negatiu de $t_{\hat{\beta}_j}$, no importa el gran que siga, no proporciona cap evidència a favor de $H_1 : \beta_j > 0$. D'altra banda, per obtenir t^{α}_{n-k} a la taula estadística de la *t*, només ens cal conèixer el nivell de significació α i els graus de llibertat. En tot cas, és important destacar que quan α disminueix, t^{α}_{n-k} augmenta.

Fins a cert punt, l'enfocament clàssic és, en algun sentit arbitrari, ja que s'ha de triar un α per endavant, i, depenent d'aquesta elecció, la H_0 es rebutja o no.

A la figura 4.4 es representa l'enfocament alternatiu. Com es desprèn de l'examen de la figura, la determinació del *valor-p* és l'operació inversa de trobar el valor en les taules estadístiques per a un determinat nivell de significació. Una vegada que el *valor-p* ha estat determinat, sabem que es rebutja la H_0 per a qualsevol nivell de significació en què α>*valor-p*, per contra, la hipòtesi nul·la no es rebutja quanα<*valor-p*.

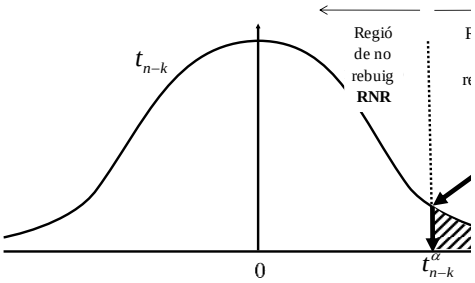


FIGURA 4.3. Regió de rebuig utilitzant la *t*: hipòtesi alternativa de cua a la dreta

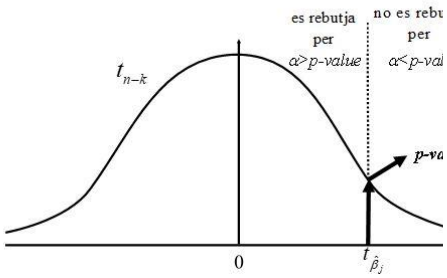


FIGURA 4.4. El *valor-p* utilitzant la *t*: hipòtesi alternativa de cua a la dreta.

EXEMPLE 4.1 És la propensió marginal a consumir menor que la propensió mitjana al consum?

Com es veu en l'exemple 1.1, contrastar la proposició 3 de la funció de consum keynesiana, en un model lineal, és equivalent a contrastar si el terme independent és significativament més gran que 0. És a dir, en el model

$$cons = \beta_1 + \beta_2 inc + u$$

Hem de contrastar si

$$\beta_1 > 0$$

Amb una mostra aleatòria de 42 observacions, s'han obtingut els següents resultats

$$\underset{}{cons_i} = \underset{(0.350)}{0.41} + \underset{(0.062)}{0.843} inc_i$$

Els nombres entre parèntesis, sota dels coeficients, són els errors estàndard (*ee*) dels estimadors.

La pregunta que ens plantegem és la següent: ¿és la tercera proposició de la teoria keynesiana admissible? A continuació, responem a aquesta pregunta.

1) En aquest cas, les hipòtesis nul·la i alternativa són les següents:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 > 0$$

2) El contrast estadístic és:

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{ee(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{ee(\hat{\beta}_1)} = \frac{0.41}{0.35} = 1.171$$

3) Regla de decisió

És convenient utilitzar diversos nivells de significació. Comencem amb un nivell de significació de 0.10 perquè el valor de t és relativament xicotet (menor que 1.5). En aquest cas, els graus de llibertat són 40 (42 observacions menys 2 paràmetres estimats). Si ens fixem en la taula estadística de la *t* (fila 40 i columna de 0.10 o 0.20, a les taules d'una cua, o de dues cues, respectivament), trobem que. $t_{40}^{0.10} = 1.303$.

Com $t<1.303$, no es rebutja la H_0. Si no rebutgem la H_0 per α=0.10, tampoc es rebutjarà per α=0.05 ($t_{40}^{0.05} = 1.684$) o α=0.01 ($t_{40}^{0.01} = 2.423$), com es pot veure a la figura 4.5. En aquesta figura la regió de rebuig correspon a $\alpha = 0.10$. Per tant, no es pot rebutjar la H_0 en favor de la H_1. En altres paraules, les dades mostrals no són consistents amb la proposició 3 de Keynes.

En l'enfocament alternatiu, com es pot veure a la figura 4.6, el valor-p corresponent $t_{\hat{\beta}_1}$ =1.171 per una *t* amb 40 *gl* és igual a 0.124. Per α<0.124 - per exemple, 0.10, 0.05 i 0.01-, no es rebutja la H_0.

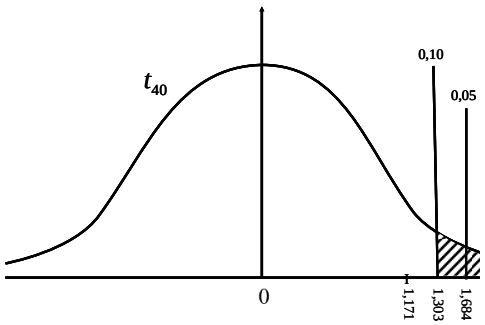


FIGURA 4.5. Exemple 4.1: Regió de rebuig utilitzant la *t*: hipòtesi alternativa de cua a la dreta.

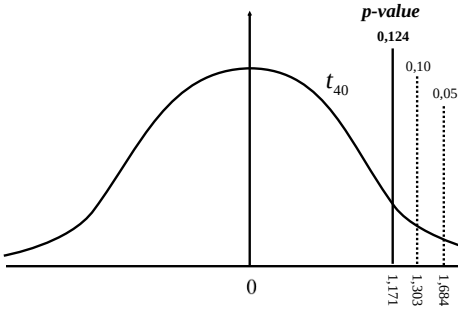


FIGURA 4.6. Exemple 4.1: El *valor-p* utilitzant la *t*: hipòtesi alternativa de cua a la dreta.

Hipòtesi alternativa d'una cua: esquerra

Considerem ara la hipòtesi nul·la

$$H_0 : \beta_j = 0$$

contra la hipòtesi alternativa

$$H_1 : \beta_j < 0$$

Aquest és un contrast de significació negativa.

La regla de decisió és en aquest cas és la següent:

Regla de decisió

Si	$t_{\hat{\beta}_j} \leq -t^{\alpha}_{n-k}$	se rebutja	H_0	(4-13)
Si	$t_{\hat{\beta}_j} > -t^{\alpha}_{n-k}$	no se rebutja	H_0	

Per tant, rebutgem $H_0 : \beta_j = 0$ en favor de $H_1 : \beta_j < 0$ per un α donat quan $t_{\hat{\beta}_j} \leq -t^{\alpha}_{n}$, es pot veure en la figura 4.7. Està molt clar que per a rebutjar H_0 en contra de $H_1 : \beta_j < 0$, el valor de $t_{\hat{\beta}_j}$ ha de ser negatiu. Un valor positiu de $t_{\hat{\beta}_j}$, no importa lo gran que siga, no proporciona cap evidència a favor de $H_1 : \beta_j < 0$.

A la figura 4.8 es representa l'enfocament alternatiu. Una vegada que el valor-p ha estat determinat, se sap que es rebutja H_0 per a qualsevol nivell de significació tal que α>valor-p; per contra, la hipòtesi nul·la no es rebutja quan α<*valor-p*.

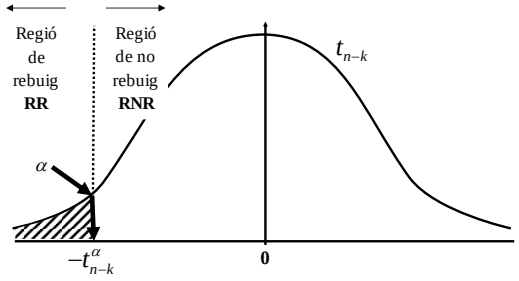


FIGURA 4.7. Regio de rebuig utilitzant la *t*: hipòtesi alternativa de cua a l'esquerra.

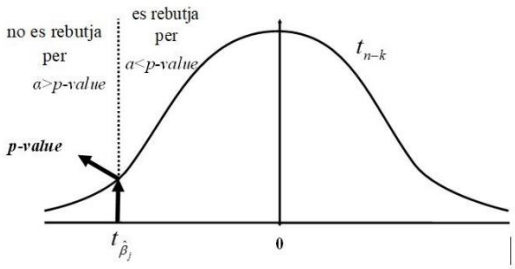


FIGURA 4.8. El *valor-p* utilitzant la *t*: hipòtesi alternativa de cua a l'esquerra.

EXEMPLE 4.2 Té la renda una influència negativa sobre la mortalitat infantil?

El següent model ha estat plantejat per explicar les morts de nens menors de 5 anys per 1000 nascuts vius (fitxer *deathun5*).

$$deathun5 = \beta_1 + \beta_2 gnipc + \beta_3 ilitrate + u$$

on *gnipc* és la renda nacional bruta per càpita i *ilitrate* és la taxa d'analfabetisme d'adults (15 anys o més) en percentatge.

Amb una mostra de 130 països (fitxer *hdr2010*) s'ha realitzat la següent estimació:

$$\underset{}{deathun5_i} = \underset{(5.93)}{27.91} - \underset{(0.00028)}{0.000826}\, gnipc_i + \underset{(0.183)}{2.043}\, ilitrate_i$$

Els nombres entre parèntesis, sota dels coeficients, són els errors estàndard (*ee*) dels estimadors.

Una de les preguntes formulades pels investigadors és si la renda té una influència negativa sobre la mortalitat infantil. Per respondre a aquesta pregunta es porta a terme un contrast en el qual les hipòtesis nul·la i alternativa i l'estadístic de contrast són els següents:

$$H_0 : \beta_2 = 0 \qquad t = \frac{\hat{\beta}_2}{ee(\hat{\beta}_2)} = \frac{-0.000826}{0.00028} = -2.966$$
$$H_1 : \beta_2 < 0$$

Atès que el valor de *t* és relativament alt, anem a començar a contrastar amb un nivell de l'1%. Per α=0.01, $t^{0.01}_{130-2} \approx t^{0.01}_{60} = 2.390$. Tenint en compte que *t*<-2.390, com es mostra a la figura 4.9, es rebutja a favor de la H_1. Per tant, la renda nacional bruta per càpita té una influència que és significativament negativa en la mortalitat de nens menors de 5 anys; és a dir, com més gran siga la renda nacional bruta per càpita més baix serà el percentatge de mortalitat de nens menors de 5 anys. Com la H_0 s'ha rebutjat per α = 0.01, també serà rebutjada pels nivells de 5% i 10%

En l'enfocament alternatiu, com es pot veure a la figura 4.10, el *valor-p* corresponent a un $t_{\hat{\beta}_1}$ =-2.966 per *t* amb menys de 61 *gl* és igual a 0.0000. Per a tots els α> 0.0000 com 0.01, 0.05 i 0.10, es rebutja la H_0.

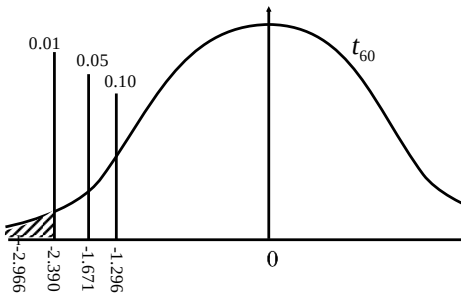


FIGURA 4.9. Exemple 4.2: Regió de rebuig utilitzant la *t*: hipòtesi alternativa de cua a l'esquerra.

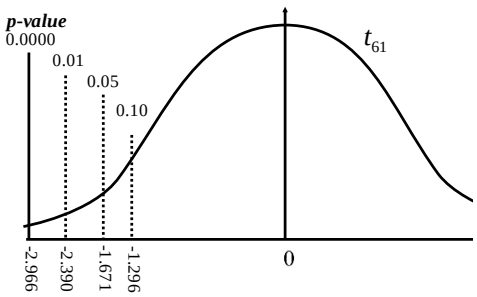


FIGURA 4.10. Exemple 4.2: El *valor-p* utilitzant la *t*: hipòtesi alternativa de cua a l'esquerra.

Hipòtesi alternativa amb dues cues

Considerem ara la hipòtesi nul·la

$$H_0 : \beta_j = 0$$

contra la hipòtesi alternativa

$$H_1 : \beta_j \neq 0$$

Aquesta és l'alternativa rellevant quan el signe de β_j no està ben determinat per la teoria o el sentit comú. Quan l'alternativa és de dues cues, estem interessats en el *valor absolut* de l'estadístic t. Aquest és un *contrast de significació*.

La regla de decisió en aquest cas és la següent:

Regla de decisió

$$\begin{array}{lll} \text{Si} & \left| t_{\hat{\beta}_j} \right| \geq t_{n-k}^{\alpha/2} & \text{se rebutja} \quad H_0 \\ \text{Si} & \left| t_{\hat{\beta}_j} \right| < t_{n-k}^{\alpha/2} & \text{no se rebutja} \quad H_0 \end{array} \qquad (4\text{-}14)$$

Per tant, rebutgem $H_0 : \beta_j = 0$ en favor de $H_1 : \beta_j < 0$ per un α donat quan $\left| t_{\hat{\beta}_j} \right| \geq t_{n-k}^{\alpha/2}$, com es pot veure a la figura 4.11. En aquest cas, per rebutjar la H_0 a favor de $H_1 : \beta_j \neq 0$, $t_{\hat{\beta}_j}$ ha de ser prou gran, bé siga positiu o negatiu.

És important fer notar que quan α decreix, $t_{n-k}^{\alpha/2}$ augmenta en valor absolut.

En l'enfocament alternatiu, una vegada que el valor-p ha estat determinat, se sap que es rebutja H_0 per a qualsevol nivell de significació si α>*valor-p*; per contra, la hipòtesi nul·la no es rebutja quan α<*valor-p*. En aquest cas *valor-p* es distribueix entre les dues cues de manera simètrica, com es mostra a la figura 4.12.

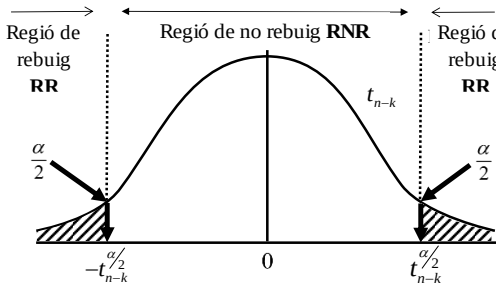


FIGURA 4.11. Regió de rebuig usant t: hipòtesi alternativa de dues cues.

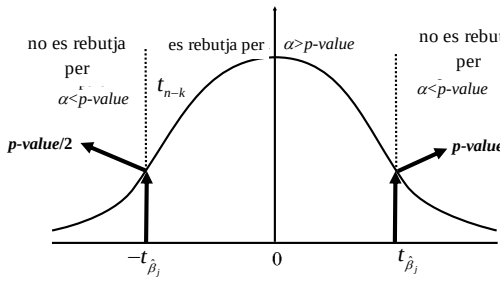


FIGURA 4.12. El *valor-p* usant t: hipòtesi alternativa de dues cues.

Quan no s'especifica una hipòtesi alternativa, en general, es considera que el contrast d'hipòtesis és de dues cues. Si es rebutja la H_0 a favor de la H_1 per a un α donat, se sol dir que "x_j *és estadísticament significativa* per el nivell α".

EXEMPLE 4.3 La taxa de delinqüència, ¿juga un paper en el preu de l'habitatge d'una àrea?

Per explicar el preu de l'habitatge en una ciutat nord-americà s'ha estimat el següent model:

$$price = \beta_1 + \beta_2 rooms + \beta_3 lowstat + \beta_4 crime + u$$

on *rooms* és el nombre d'habitacions de la casa, *lowstat* és el percentatge de persones de "classe marginal" a la zona i *crim* són els delictes comesos per càpita a la zona.

Els resultats del model ajustat realitzat amb *E-views*, utilitzant el fitxer *hprice2* (primeres 55 observacions), apareix en el quadre 4.2. El significat de les tres primeres columnes és clar: "Estadístic t" és la dada requerida per a fer un contrast de significació, és a dir, és la relació entre el "Coeficient" i l'"Error estàndard", i "Prob" és el *valor -p* per a un contrast de dues cues.

En relació amb aquest model la pregunta que es fan els investigadors és si la taxa de criminalitat juga un paper important en el preu de les cases de la zona. Per respondre a aquesta pregunta, s'ha dut a terme el següent procediment.

En aquest cas, la hipòtesi nul·la i alternativa, i l'estadístic de contrast, són els següents:

$$H_0 : \beta_4 = 0$$
$$H_1 : \beta_4 \neq 0$$
$$t = \frac{\hat{\beta}_4}{ee(\hat{\beta}_4)} = \frac{-3854}{960} = -4.016$$

QUADRE 4.2. Sortida estàndard en una regressió per a explicar el preu d'una casa. *n*=55.

Variable	*Coeficient*	*Error estàndard*	*Estadístic t*	*Prob.*
C	-15693.61	8021.989	-1.956324	0.0559
rooms	6788.401	1210.720	5.606910	0.0000
lowstat	-268.1636	80.70678	-3.322690	0.0017
crime	-3853.564	959.5618	-4.015962	0.0002

Atès que el valor de t és relativament alt, anem a començar a contrastar per a un nivell de l'1%. Per α=0.01, $t_{51}^{0.01/2} \approx t_{50}^{0.01/2} = 2.69$. (A les taules estadístiques habituals per a la distribució t, no hi ha informació per a cada *gl*, un a un, més enllà de 20). Tenint en compte que $|t| > 2.69$, rebutgem la H_0 a favor de la H_1. Per tant, la delinqüència té una influència significativa en el preu de l'habitatge amb un nivell de significació de l'1% i, per tant, d'un 5% i 10%.

En l'enfocament alternatiu podem realitzar el contrast amb major precisió. En el quadre 4.2, veiem que el valor de *valor-p* per al coeficient *crime* és de 0.0002. Això vol dir que la probabilitat que l'estadístic *t* siga més gran de 4.016 és 0.0001 i la probabilitat que t siga menor de 4.016 és de 0.0001. És a dir, el valor de *valor-p*, com es mostra a la figura 4.13 es distribueix en els dos costats. Com es pot veure en aquesta

figura, es rebutja H_0 per a tots els nivells de significació superiors a 0.0002, com ara 0.01, 0.05 i 0.10. Si es tractés d'un contrast d'una cua, en l'enfocament alternatiu el *valor-p* seria igual a 0.0002/1 = 0.0001.

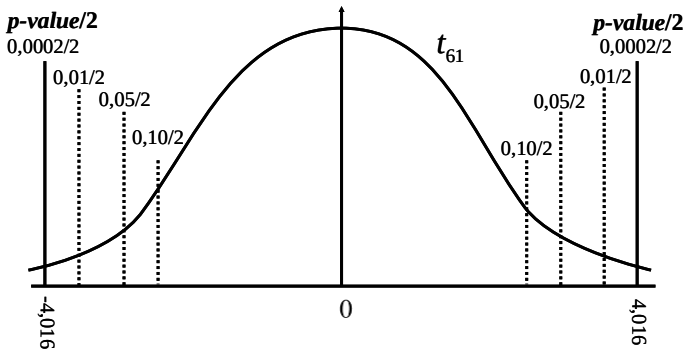


FIGURA 4.13. Exemple 4.3: El *valor-p* usant t: hipòtesi alternativa de dues cues.

Fins ara hem vist el contrast significatiu d'una cua i de dues cues en el qual un paràmetre pren el valor 0 a la H_0. Ara anem a veure un cas més general en què el paràmetre en la H_0 pren un valor específic qualsevol:

$$H_0 : \beta_j = \beta_j^0$$

En aquest cas, l'estadístic *t* adequat és

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j^0}{ee(\hat{\beta}_j)}$$

Igual que abans, $t_{\hat{\beta}_j}$ mesura la quantitat de desviacions estàndard està $\hat{\beta}_j$ distanciada de β_j^0, valor que pren el paràmetre en la hipòtesi nul·la.

EXEMPLE 4.4 És l'elasticitat de la despesa en fruites/renda igual a 1? És la fruita un bé de luxe?

Per respondre a aquestes dues preguntes utilitzarem el següent model per explicar la despesa en fruita (*fruit*):

$$\ln(fruit) = \beta_1 + \beta_2 \ln(inc) + \beta_3 househsize + \beta_4 punders + u$$

on *inc* és la renda disponible de les llars, *househsize* és el nombre de membres de la família i *punder5* és la proporció de nens menors de cinc anys a la llar.

Atès que les variables *fruit* i *inc* apareixen expressades en logaritmes naturals, aleshores β_2 és l'elasticitat de la despesa/renda. Utilitzant una mostra de 40 llars (fitxer *demand*), s'han obtingut els resultats del quadre 4.3.

QUADRE 4.3. Sortida estàndard d'una regressió que explica les despeses en fruita. *n*=40.

Variable	*Coeficient*	*Error estàndard*	*Estadístic t*	*Prob.*
C	-9.767654	3.701469	-2.638859	0.0122
ln(inc)	2.004539	0.512370	3.912286	0.0004
househsize	-1.205348	0.178646	-6.747147	0.0000
punder5	-0.017946	0.013022	-1.378128	0.1767

És l'elasticitat de la despesa en fruites/renda igual a 1?

Per respondre a aquestes dues preguntes utilitzarem el següent model per explicar la despesa en fruita (fruit):

$$H_0 : \beta_2 = 1 \qquad H_1 : \beta_2 \neq 1 \qquad t = \frac{\hat{\beta}_2 - \beta_2^0}{ee(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - 1}{ee(\hat{\beta}_2)} = \frac{2.005 - 1}{0.512} = 1.961$$

Per α=0.10, ens trobem que $t_{36}^{0.10/2} \approx t_{35}^{0.10/2} = 1.69$. Com $|t|$>1.69 es rebutja H_0. Per a α=0.05, $t_{36}^{0.05/2} \approx t_{35}^{0.05/2} = 2.03$. Com $|t|$<2.03 no rebutgem H_0 per α=0.05, ni per α=0.01. Per tant, rebutgem que l'elasticitat de la despesa en fruita/renda siga igual a 1 per a α= 0.10, però no rebutgem per α= 0.05, ni paraα= 0.01.

És la fruita un bé de luxe?

Segons la teoria econòmica, una mercaderia és un bé de luxe quan l'elasticitat de la despesa pel que fa a la renda és més gran que 1. Per tant, per respondre a aquesta segona qüestió, i tenint en compte que l'estadístic *t* és el mateix, s'ha dut a terme el següent procediment:

$$H_0 : \beta_2 = 1 \qquad H_1 : \beta_2 > 1.$$

Per α=0.10, ens trobem que $t_{36}^{0.10} \approx t_{35}^{0.10} = 1.31$. Quan *t*>1.31, rebutgem H_0 en favor de H_1. Per a α=0.05, $t_{36}^{0.05} \approx t_{35}^{0.05} = 1.69$ .Com *t*>1.69, rebutgem la H_0 en favor de la H_1. Per α=0.01, $t_{36}^{0.01} \approx t_{35}^{0.01} = 2.44$. Com *t*<2.44, no rebutgem H_0. Per tant, la fruita és un bé de luxe per α=0.10 i α=0.05, però no es pot rebutjar la H_0 en favor de la H_1 per α=0.01.

EXEMPLE 4.5 És la Borsa de Madrid un mercat eficient?

Abans de respondre a la qüestió plantejada, examinarem alguns conceptes previs. La taxa de *rendiment d'un actiu* en un període de temps es defineix com la variació percentual que experimenta el valor invertit en aquest actiu durant aquest període de temps. Anem a considerar ara com a actiu específic, a una acció d'una companyia industrial adquirida en una borsa espanyola al final d'un any i que es manté fins al final de l'any següent. A aquests dos moments de temps els designarem *t*-1 i *t* respectivament. La taxa de rendiment d'aquesta acció al cap d'aquest any pot expressar-se mitjançant la següent relació:

$$RA_t = \frac{\Delta P_t + D_t + A_t}{P_{t-1}} \tag{4-15}$$

on

P_t: és la cotització de l'acció al final del període *t*

D_t: són els dividends percebuts per l'acció durant el període *t*

A_t: és el valor dels drets que eventualment han pogut correspondre a l'acció durant el període *t*

Així doncs, en el numerador de (4-15) es recullen els tres tipus de guanys de capital que s'han pogut percebre pel manteniment d'una acció durant l'any *t*: increment - o pèrdua en el seu cas - en la cotització, dividends i drets d'ampliació. En dividir per P_{t-1} s'obté la taxa de guany sobre el valor de l'acció a la fi del període anterior. Dels tres components el més important és l'increment en la cotització. Tenint en compte només a aquesta component, la taxa de rendiment de l'acció pot expressar-se per

$$RA1_t = \frac{\Delta P_t}{P_{t-1}} \tag{4-16}$$

o, alternativament si utilitzem una taxa de variació natural, per

$$RA2_t = \Delta \ln P_t \tag{4-17}$$

De la mateixa manera que Ra_t, en qualsevol de les dues expressions, representa la taxa de rendiment d'una acció concreta, es pot estimar també la taxa de rendiment del conjunt d'accions cotitzades en borsa. A aquesta última taxa de rendiment, a la qual designarem per RM_t, se li denomina taxa de rendiment de mercat.

Fins ara hem considerat la taxa de rendiment en un any, però igualment es pot aplicar expressions del tipus (4-16), o (4-17), per obtenir taxes de rendiment diari. Analitzant el comportament d'aquestes taxes cal preguntar-se si les taxes de rendiment en el passat són d'utilitat per predir les taxes de rendiment en el futur. Aquesta pregunta està relacionada amb el concepte d'eficiència d'un mercat. Un mercat és *eficient* si els preus incorporen tota la informació disponible, de manera que hi ha possibilitat d'obtenir guanys extraordinaris en utilitzar aquesta informació.

Per contrastar l'eficiència d'un mercat definirem el següent model, utilitzant taxes de rendiment diàries definides segons (4-16):

$$rmad92_t = \beta_1 + \beta_2 rmad92_{t-1} + u_t \tag{4-18}$$

Si un mercat és eficient, aleshores el paràmetre β_2 de l'anterior model ha de ser 0. Anem a contrastar ara si la Borsa de Madrid, pel que fa a la renda variable, és o no eficient en el seu conjunt.

El model de (4-18) s'ha estimat amb dades diàries de la Borsa de Madrid per a l'any 1992, utilitzant el fitxer *bolmadef*. Els resultats obtinguts han estat els següents:

$$rmad92_t = \underset{(0.0007)}{-0.0004} + \underset{(0.0629)}{0.1267}\, rmad92_{t-1}$$

$$R^2=0.0163 \qquad n=247$$

Els resultats obtinguts poden resultar paradoxals. D'una banda, el valor del coeficient de determinació és molt baix (0,0163), el que significa que només el 1.63% de la variança total de la taxa de rendiment s'explica per la taxa de rendiment del dia anterior. D'altra banda, però, el coeficient corresponent a la taxa de significació del dia anterior és estadísticament significatiu per a un nivell del 5% però no per a un nivell d'1%, perquè l'estadístic *t* és igual a 0,1267/0,0629 = 2.02 que és lleugerament més gran en valor absolut que $t_{245}^{0.01} \simeq t_{60}^{0.01}$=2.00. El motiu d'aquesta aparent paradoxa es deu al fet que la grandària de la mostra és molt elevat. Així, tot i que la incidència de la variable explicativa sobre la variable endògena és relativament reduïda (com indica el coeficient de determinació), però, aquesta incidència és significativa (com ho confirma l'estadístic *t*) a causa que s'ha disposat d'una mostra de dades prou gran.

Contestant a la pregunta de si la Borsa de Madrid és o no un mercat eficient, en una primera anàlisi la resposta és que no és totalment eficient. No obstant això, aquesta resposta ha de matisar-se. En economia financera és coneguda l'existència d'una relació de dependència entre la taxa de rendiment d'un dia i la del dia precedent. Aquesta relació no és molt forta, encara que sí és estadísticament significativa en moltes borses mundials, i es deu a les friccions del mercat. En qualsevol cas, aquest fenomen no es pot explotar de forma lucrativa pels agents del mercat, de manera que no es pot qualificar a aquests mercats d'ineficients, d'acord amb la definició donada anteriorment sobre el concepte d'eficiència.

EXEMPLE 4.6 **La rendibilitat de la Borsa de Madrid, es veu afectada per la rendibilitat de la Borsa de Tòquio?**

L'estudi de la relació entre diferents mercats d'accions (Borsa de Nova York, Borsa de Tòquio, Borsa de Madrid, Borsa de Londres, etc.) ha rebut una gran atenció en els últims anys, a causa d'una major llibertat en la circulació de capitals i a la conveniència d'utilitzar mercats estrangers per reduir el risc en la gestió de carteres, ja que l'absència d'una perfecta integració dels mercats permet la diversificació del risc. De tota manera, cada vegada es camina cap a una major integració mundial dels mercats financers, en general, i dels mercats d'accions, en particular.

Si els mercats són eficients, i hem vist en l'Exemple 4.5 que es pot admetre que ho siguen, les notícies que es van produint, a les que s'anomenen *innovacions*, durant un període de 24 hores s'aniran veient reflectides en els diferents mercats.

Convé distingir entre dos tipus d'innovacions: a) *innovacions globals*, que són notícies que es generen al voltant del món i que es capten en els preus de les accions en tots els mercats; b) *innovacions específiques*, que és la informació generada durant un període de 24 hores que només afecta els preus d'un mercat particular. Així, la informació sobre evolució dels preus del petroli es pot considerar com una innovació global, mentre que una nova regulació del sector financer en un país seria considerada possiblement com una innovació específica.

D'acord amb l'exposició anterior, en una sessió de Borsa d'un determinat mercat, els preus de les accions que en ell es cotitzen vindran afectats per les innovacions globals recollides en un altre mercat que

haja tancat abans. Així, les innovacions globals recollides al mercat de Tòquio influiran en les cotitzacions del mercat de Madrid d'aquest mateix dia. El següent model recull la transmissió d'efectes entre la Borsa de Tòquio i la Borsa de Madrid:

$$rmad92_t = \beta_1 + \beta_2 rtok92_t + u_t \qquad (4\text{-}19)$$

on $rmad92_t$ és la taxa de rendibilitat de la Borsa Madrid en el període *t*, i $rtok92_t$ és la taxa de rendibilitat de la Borsa de Tòquio en el període t. Les rendibilitats dels mercats s'han calculat d'acord amb (4-16).

Al fitxer *madtok* es poden trobar els índexs generals de la Borsa de Madrid i la Borsa de Valors de Tòquio durant els dies en què les dues bosses estaven obertes de forma simultània. És a dir, s'han eliminat les observacions dels dies en què qualsevol de les borses estigués tancada. En total, el nombre d'observacions és de 234, en comparació amb els 247 i 246 dies en què les Borsa de Madrid i de Tòquio van estar obertes respectivament.

L'estimació del model (4-19) és com segueix:

$$\widehat{rmad92}_t = \underset{(0.0007)}{-0.0005} + \underset{(0.0375)}{0.1244}\, rtok92_t$$

$$R^2 = 0.0452 \qquad n = 235$$

Cal observar que el coeficient de determinació és relativament baix. No obstant això, per contrastar H_0: $\beta_2=0$, s'obté que l'estadístic $t = (0{,}1244/0{,}0375) = 3.32$ qual cosa implica que es rebutja la hipòtesi nul·la de que la taxa de rendiment de la Borsa de Valors de Tòquio no tinga cap efecte sobre la taxa de rendibilitat de la Borsa de Madrid, per a un nivell de significació de 0.01.

Un cop més ens trobem amb la mateixa paradoxa aparent que va aparèixer quan es va analitzar l'eficiència de la Borsa de Madrid en l'Exemple 4.5 a excepció d'una diferència. En aquest últim cas, la taxa de rendiment del dia anterior apareixia com a significativa, a causa de problemes derivats de l'elaboració de l'índex general de la Borsa de Madrid.

En conseqüència, el fet que la hipòtesi nul·la siga rebutjada implica que hi ha evidència empírica que dóna suport a la teoria que les innovacions globals de la Borsa de Tòquio es transmeten a les cotitzacions de la Borsa de Madrid aquest mateix dia.

4.2.2 Els intervals de confiança

Sota els supòsits del *MLC*, és fàcil construir un interval de confiança (*IC*) per al paràmetre de la població, β_j. Als *IC* també se'ls anomena estimacions per interval, ja que proporcionen un rang de valors versemblants per β_j, i no només una estimació puntual.

L'*IC* està construït de tal manera que el paràmetre desconegut està contingut dins del recorregut de l'*IC* amb una probabilitat prèviament especificada.

Utilitzant el fet que

$$\frac{\hat{\beta}_j - \beta_j}{ee(\hat{\beta}_j)} \sim t_{n-k}$$

$$\Pr\left[-t_{n-k}^{\alpha/2} \le \frac{\hat{\beta}_j - \beta_j}{ee(\hat{\beta}_j)} \le t_{n-k}^{\alpha/2} \right] = 1-\alpha$$

Operant per situar el paràmetre β_j sol al centre de l'interval, obtenim que

$$\Pr\left[\hat{\beta}_j - ee(\hat{\beta}_j) \times t_{n-k}^{\alpha/2} \le \beta_j \le \hat{\beta}_j + ee(\hat{\beta}_j) \times t_{n-k}^{\alpha/2} \right] = 1-\alpha$$

Per tant, els límits inferior i superior, respectivament, d'un IC de probabilitat (1-α) estan donats per

$$\underline{\beta}_j = \hat{\beta}_j - ee(\hat{\beta}_j) \times t_{n-k}^{\alpha/2}$$

$$\overline{\beta}_j = \hat{\beta}_j + ee(\hat{\beta}_j) \times t_{n-k}^{\alpha/2}$$

Si les mostres es van obtenir a l'atzar de forma repetida calculant $\underline{\beta}_j$ i $\overline{\beta}_j$ cada vegada, el paràmetre poblacional (desconegut) cauria en l'interval ($\underline{\beta}_j$, $\overline{\beta}_j$) en un (1−α)% de les mostres. Desafortunadament, per a la mostra individual que s'utilitza en la construcció de l'*IC*, no sabem si β_j està o no realment continguda dins de l'interval.

Una vegada que l'*IC* s'ha construït, és fàcil dur a terme contrastos d'hipòtesis de dues cues. Si la hipòtesi nul·la és $H_0 : \beta_j = a_j$, aleshores la H_0 es rebutja contra $H_1 : \beta_j \neq a_j$ per al nivell de significació del 5%, si i només si, a_j no està en l'IC del 95%.

Per a il·lustrar tot l'anterior, a la figura 4.14 s'han construït intervals de confiança del 90%, 95% i 99%, per la propensió marginal al consum -b2- corresponent a l'Exemple 4.1.

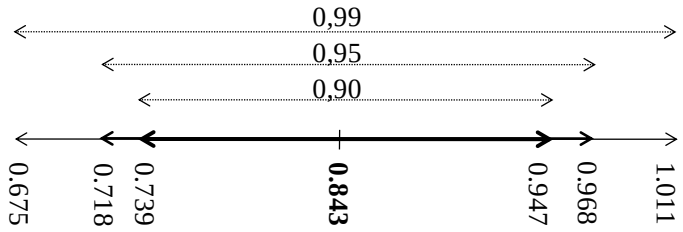


FIGURA 4.14. Intervals de confiança per a la propensió marginal al consum en e l'Exemple 4.1.

4.2.3 Contrast d'hipòtesis sobre una combinació lineal de paràmetres

En moltes aplicacions, estem interessats en contrastar hipòtesis en què estan implicats més d'un paràmetre poblacional. L'estadístic *t* també es pot utilitzar per contrastar una combinació de paràmetres, en què dos o més paràmetres estan implicats.

El contrast sobre una combinació lineal de paràmetres es pot fer per dos procediments diferents. En el primer procediment, l'error estàndard de la combinació lineal dels paràmetres corresponents a la hipòtesi nul·la es calcula utilitzant informació sobre la matriu de covariança dels estimadors. En el segon procediment, el model es reparametritza mitjançant la introducció d'un nou paràmetre deduït de la hipòtesi nul·la, després s'estima el model reparametritzat i el contrast del nou paràmetre ens indica si es rebutja, o no, la hipòtesi nul·la. El següent exemple il·lustra tots dos procediments.

EXEMPLE 4.7 Hi ha rendiments constants a escala en el sector de metalls primari?

Per examinar si hi ha rendiments constants a escala en el sector de metalls primari es va a utilitzar la funció de producció Cobb-Douglas, donada per

$$\ln(output) = \beta_1 + \beta_2 \ln(labor) + \beta_3 \ln(capital) + u \qquad (4\text{-}20)$$

En l'anterior model els paràmetres β_2 i β_3 són elasticitats (producció/treball i producció/capital)

Abans de fer inferències hem de recordar que els *rendiments a escala* es refereixen a una característica tècnica de la funció de producció que analitza els canvis en la producció degudes a un canvi en la mateixa proporció de tots els inputs, que en aquest cas són el treball i el capital. Si la producció canvia en la mateixa proporció que els inputs aleshores es diu que hi ha *rendiments constants a escala*. Els rendiments constants a escala impliquen que si els factors *treball* i *capital* augmenten a una certa taxa (diguem el 10%), la producció augmentarà en la mateixa proporció (això és en un 10%). Si la producció augmenta en una major proporció, aleshores hi ha *rendiments creixents a escala*. Si augmenta la producció en una menor proporció, hi ha *rendiments decreixents a escala*. En el model anterior, succeeix que:

- si $\beta_2+\beta_3=1$, hi ha *rendiments constants a escala*.

- si $\beta_2+\beta_3>1$, hi ha *rendiments creixents a escala*.

- si $\beta_2+\beta_3<1$, hi ha *rendiments decreixents a escala*

Les dades utilitzades en aquest exemple són una mostra de 27 empreses del sector de metalls primari (fitxer *prodmet*), on *output* és el valor afegit brut, *labor* és una mesura de la mà d'obra i *capital* és el valor brut de la planta i equip. Detalls addicionals sobre la construcció d'aquestes dades s'ofereixen en Aigne *et al*. (1977) i en Hildebrand i Liu (1957). Els resultats obtinguts en l'estimació del model (4-20) apareixen en el quadre 4.4.

QUADRE 4.4. Sortida estàndard de l'estimació de la funció de producció: model (4-20).

Variable	*Coeficient*	*Error estàndard*	*Estadístic t*	*Prob.*
constant	1.170644	0.326782	3.582339	0.0015
ln(*labor*)	0.602999	0.125954	4.787457	0.0001
ln(*capital*)	0.375710	0.085346	4.402204	0.0002

Per respondre a la pregunta plantejada en aquest Exemple, hem de contrastar:

$$H_0 : \beta_2 + \beta_3 = 1$$

contra la hipòtesi alternativa següent:

$$H_1 : \beta_2 + \beta_3 \neq 1$$

D'acord amb la H_0, es dedueix que $\beta_2 + \beta_3 - 1 = 0$. Per tant, l'estadístic *t* es basa ara en si la suma estimada $\hat{\beta}_2 + \hat{\beta}_3 - 1$ és prou diferent de 0 com per rebutjar la H_0 a favor de la H_1.

Per contrastar aquesta hipòtesi es van a utilitzar dos procediments. En el primer procediment s'utilitza la matriu de covariances dels estimadors. En el segon, el model es reparametritza introduint un nou paràmetre.

Procediment que utilitza la matriu de covariances dels estimadors

D'acord amb H_0, s'estima que $\beta_2 + \beta_3 - 1 = 0$. Per tant, l'estadístic t està basat ara en si la suma estimada $\hat{\beta}_2 + \hat{\beta}_3 - 1$ és prou diferent de 0 per rebutjar la H_0 a favor de la H_1. Per tenir en compte l'error de mostreig en els nostres estimadors, s'estandarditza aquesta suma dividint pel seu error estàndard:

$$t_{\hat{\beta}_2+\hat{\beta}_3} = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{ee(\hat{\beta}_2 + \hat{\beta}_3)}$$

Així que, si $t_{\hat{\beta}_2+\hat{\beta}_3}$ és prou gran, anem a concloure, en un contrast de dues cues, que no hi ha *rendiments constants a escala*. D'altra banda, si $t_{\hat{\beta}_2+\hat{\beta}_3}$ ¡ és positiu i prou gran, anem a rebutjar, en un contrast alternatiu d'una cua (la dreta), a H_0 en favor de $H_1 : \beta_2 + \beta_3 > 1$, concloent que sí que hi ha rendiments creixents a escala.

D'altra banda, obtenim que

$$ee(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{\text{var}(\hat{\beta}_2 + \hat{\beta}_3)}$$

on

$$\text{var}(\hat{\beta}_2 + \hat{\beta}_3) = \text{var}(\hat{\beta}_2) + \text{var}(\hat{\beta}_3) + 2 \times \text{covar}(\hat{\beta}_2, \hat{\beta}_3)$$

Per tant, per estimar $ee(\hat{\beta}_2 + \hat{\beta}_3)$ es necessita informació sobre la covariança estimada dels estimadors. Molts paquets de programari economètric, com l'E-Views, té una opció per mostrar les estimacions de la matriu de covariances del vector dels estimadors. En aquest cas, la matriu de covariances obtinguda apareix en el quadre 4.5. Amb aquesta informació s'obté que

$$ee(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{0.015864 + 0.007284 - 2 \times 0.009616} = 0.0626$$

$$t_{\hat{\beta}_2 + \hat{\beta}_3} = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{ee(\hat{\beta}_2 + \hat{\beta}_3)} = \frac{-0.02129}{0.0626} = -0.3402$$

QUADRE 4.5. Matriu de covarianços de la funció de producció.

	constant	ln(*labor*)	ln(*capital*)
constant	0.106786	-0.019835	0.001189
ln(*labor*)	-0.019835	0.015864	-0.009616
ln(*capital*)	0.001189	-0.009616	0.007284

Tenint en compte que *t*= 0,3402, és evident que no podem rebutjar l'existència de rendiments constants d'escala, per als nivells de significació habituals. Atès que l'estadístic *t* pren un valor negatiu, no té sentit contrastar si hi ha rendiments creixents a escala.

Procediment en el qual es reparametritza el model mitjançant la introducció d'un nou paràmetre

L'aplicació d'aquest segon procediment és una forma més fàcil de realitzar aquest contrast. En aquest procediment s'estima un model diferent que proporciona directament l'error estàndard en què estem interessats. Així, en l'exemple anterior definirem:

$$\theta = \beta_2 + \beta_3 - 1$$

de manera que la hipòtesi nul·la que hi ha *rendiments constants a escala* és equivalent a postular que. $H_0 : \theta = 0$.

De la definició de θ s'obté que $\beta_2 = \theta - \beta_3 + 1$. Substituint β_2 en l'equació original:

$$\ln(output) = \beta_1 + (\theta - \beta_3 + 1)\ln(labor) + \beta_3 \ln(capital) + u$$

Per tant,

$$\ln(output / labor) = \beta_1 + \theta \ln(labor) + \beta_3 \ln(capital / labor) + u$$

Així doncs, contrastar si hi ha rendiments constants a escala és equivalent a realitzar un contrast de significació del coeficient ln(*labor*) en el model anterior. L'estratègia de reescriure el model, de manera que continga el paràmetre d'interès, funciona en tots els casos i normalment és fàcil d'implementar. Si apliquem aquesta transformació en aquest exemple, s'obtenen els resultats del quadre 4.6.

Com es pot veure s'obté el mateix resultat:

$$t_{\hat{\theta}} = \frac{\hat{\theta}}{ee(\hat{\theta})} = -0.3402$$

QUADRE 4.6. Sortida de l'estimació de la funció de producció: model reparametritzat.

Variable	*Coeficient*	*Error estàndard*	*Estadístic t*	*Prob.*
constant	1.170644	0.326782	3.582339	0.0015
ln(*labor*)	-0.021290	0.062577	-0.340227	0.7366
ln(*capital/labor*)	0.375710	0.085346	4.402204	0.0002

EXEMPLE 4.8 Publicitat o incentius?

La Companyia Bush es dedica a la venda i distribució de regals importats d'Orient Pròxim. L'article més popular al catàleg és la polsera de Guantánamo. Té algunes propietats relaxants. Els agents de vendes reben una comissió del 30% de l'import total de les vendes. Per tal d'augmentar les vendes sense necessitat d'expandir la xarxa comercial, la companyia va establir incentius especials per a aquells agents que sobrepassen l'objectiu de vendes durant l'últim any.

D'altra banda, s'emeten anuncis publicitaris en ràdio en diferents regions per enfortir la promoció de les vendes. En aquests llocs es va fer especial èmfasi a destacar el benestar de portar un braçalet de Guantánamo.

El gerent de l'Empresa Bush es pregunta si un dòlar gastat en incentius especials té, o no, una major incidència en les vendes que un dòlar gastat en publicitat. Per respondre a aquesta pregunta l'econòmetra de la companyia suggereix el següent model per explicar les vendes (*sales*):

$$sales = \beta_1 + \beta_2 advert + \beta_3 incent + u$$

on *incent* són els incentius als venedors i *advert* són les despeses en publicitat. Les variables *sales*, *incent* i *advert* estan expressades en milers de dòlars.

Utilitzant una mostra de 18 àrees de venda (fitxer *advincen*), s'han obtingut els resultats de la regressió i la matriu de covariances dels coeficients que apareixen en els quadres 4.7 i 4.8, respectivament.

QUADRE 4.7. Sortida estàndard de la regressió per l exemple 4.8.

Variable	*Coeficient*	*Error estàndard*	*Estadístic t*	*Prob.*
constant	396.5945	3548.111	0.111776	0.9125
advert	18.63673	8.924339	2.088304	0.0542
incent	30.69686	3.604420	8.516448	0.0000

QUADRE 4.8 Matriu de covarianços de l'exemple 4.8.

	C	*advert*	*incent*
constant	12589095	-26674	-7101
advert	-26674	79.644	2.941
incent	-7101	2.941	12.992

En aquest model, el coeficient β_2 indica l'augment de les vendes produïdes per un dòlar d'increment en la despesa en publicitat, mentre β_3 indica l'augment que es produeix en les vendes per un dòlar d'increment en els incentius especials, mantenint fix en tots dos casos l'altre regressor.

Per respondre a la pregunta plantejada en aquest exemple, la hipòtesi nul·la i la hipòtesi alternativa són les següents

$$H_0 : \beta_3 - \beta_2 = 0$$
$$H_1 : \beta_3 - \beta_2 > 0$$

L'estadístic *t* es construeix utilitzant la informació sobre la matriu de covariances dels estimadors:

$$t_{\hat{\beta}_3 - \hat{\beta}_2} = \frac{\hat{\beta}_3 - \hat{\beta}_2}{ee(\hat{\beta}_3 - \hat{\beta}_2)}$$

$$ee(\hat{\beta}_3 - \hat{\beta}_2) = \sqrt{79.644 + 12.992 - 2 \times 2.941} = 9.3142$$

$$t_{\hat{\beta}_3 - \hat{\beta}_2} = \frac{\hat{\beta}_3 - \hat{\beta}_2}{ee(\hat{\beta}_3 - \hat{\beta}_2)} = \frac{30.697 - 18.637}{9.3142} = 1.295$$

Per α=0.10, ens trobem que $t_{15}^{0.10} = 1.341$. Com t<1.341, no rebutgem H_0 per α=0.10, ni per α=0.05 o α=0.01. Per tant, no hi ha evidència empírica que un dòlar gastat en incentius especials tinga una major incidència en les vendes que un dòlar gastat en publicitat.

EXEMPLE 4.9 Contrast de la hipòtesi d'homogeneïtat en la demanda de peix

En el cas d'estudi del capítol 2 van ser estimats diversos models, utilitzant dades de tall transversal, per explicar la demanda de productes lactis en què la renda disponible era l'única variable explicativa. No obstant això, el preu del producte estudiat i, en major o menor mesura, els preus d'altres productes són determinants en la demanda. L'anàlisi de la demanda sobre la base de dades de tall transversal, té precisament la limitació que no és possible examinar l'efecte dels preus en la demanda perquè els preus es mantenen constants, ja que les dades es refereixen totes al mateix punt en el temps. Per analitzar l'efecte dels preus és necessari l'ús de dades de sèries temporals o, alternativament, de dades de panell. A continuació, examinarem, breument, alguns aspectes de la teoria de la demanda d'un bé per després passar a l'estimació d'una funció de demanda amb dades de sèries temporals. Com a colofó a aquest cas, anem a contrastar una de les hipòtesis que, en determinades circumstàncies, ha de satisfer un model teòric.

La demanda d'un bé -com el bé *j*- es pot expressar, d'acord amb un procés d'optimització dut a terme pel consumidor, en termes de renda disponible, del preu de la mercaderia i dels preus de la resta de béns. Analíticament:

$$q_j = f_j(p_1, p_2, \cdots, p_j, \cdots, p_m, R) \tag{4-21}$$

on

- R és la renda disponible dels consumidors.
- $p_1, p_2, \cdots, p_j, \cdots p_m$ són els preus dels béns que es tenen en compte pels consumidors quan adquireixen el bé *j*.

En els estudis de demanda, els models logarítmics són atractius, ja que els coeficients són directament elasticitats. El model logarítmic s'expressa de la següent manera:

$$\ln(q_j) = \beta_1 + \beta_2 \ln(p_1) + \beta_3 \ln(p_2) + \cdots + \beta_j \ln(p_j) + \cdots + \beta_{m+1} \ln(p_m) + \beta_{m+2} \ln(R) + u \tag{4-22}$$

Com es pot veure de forma immediata, tots els coeficients β, excloent el terme constant, són elasticitats de diferents tipus i, per tant, són independents de les unitats de mesura de les variables. Quan no hi ha il·lusió monetària, si tots els preus i la renda creixen a la mateixa taxa, la demanda d'un bé no es veu afectada per aquests canvis. Per tant, suposant que els preus i la renda es multipliquen per λ, si el consumidor no té il·lusió monetària, s'ha de satisfer que

$$f_j(\lambda p_1, \lambda p_2, \cdots, \lambda p_j, \cdots, \lambda p_m, \lambda R) = f_j(p_1, p_2, \cdots, p_j, \cdots p_m, R) \tag{4-23}$$

Des d'un punt de vista matemàtic, la condició anterior implica que la funció de demanda ha de ser homogènia de grau 0. Aquesta condició es diu la *restricció d'homogeneïtat*. Aplicant el teorema d'Euler, la restricció d'homogeneïtat, al seu torn implica que la suma de l'elasticitat de demanda/renda i de totes les elasticitats de demanda/preu és zero, és a dir:

$$\sum_{h=1}^{m} \varepsilon_{q_j/p_h} + \varepsilon_{q_j/R} = 0 \tag{4-24}$$

Aquesta restricció aplicada al model logarítmic (4-22) implica que

$$\beta_2 + \beta_3 + \cdots + \beta_j + \cdots + \beta_{m+1} + \beta_{m+2} = 0 \tag{4-25}$$

A la pràctica, quan s'estima una funció de demanda, els preus de molts béns no estan inclosos, sinó només aquells que estan estretament relacionats, ja siga per ser complementaris o per ser substitutius del bé estudiat. També és ben sabut que l'assignació pressupostària de la despesa se sol realitzar en diverses etapes.

A continuació, estudiarem la demanda de peix a Espanya, utilitzant un model similar a (4-22). Tinguem en compte que en una primera assignació, el consumidor distribueix la seva renda entre el consum total i l'estalvi. En una segona etapa, la despesa de consum per funció es porta a terme tenint en compte el consum total i els preus rellevants en cada funció. En concret, a la demanda de peix s'ha suposat que només és rellevant el preu del peix i el preu de la carn que és el substitutiu més important.

Tenint en compte les consideracions anteriors, s'ha formulat el següent model:

$$\ln(fish) = \beta_1 + \beta_2 \ln(fishpr) + \beta_3 \ln(meatpr) + \beta_4 \ln(cons) + u \tag{4-26}$$

on *fish* és la despesa de peix a preus constants, *fishpr* és el preu del peix, *meatpr* és el preu de la carn i *cons* és el consum total a preus constants.

El fitxer *fishdem* conté informació sobre aquesta sèrie per al període 1964-1991. Els preus són nombres índexs amb base 1986, i *fish* i *cons* són magnituds a preus constants també amb base a 1986. Els resultats de l'estimació de model (4-26) són els següents:

$$\ln(\widehat{fish}_i) = \underset{(2.30)}{7.788} - \underset{(0.133)}{0.460}\ln(fishpr_i) + \underset{(0.112)}{0.554}\ln(meatpr_i) + \underset{(0.137)}{0.322}\ln(cons_i)$$

Com es pot observar, els signes de les elasticitats són correctes: l'elasticitat de la demanda és negatiu pel que fa al preu del propi bé, mentre que les elasticitats respecte al preu del bé substitutiu i pel que fa al consum total són positius.

En el model de (4-26) la restricció d'homogeneïtat implica la següent hipòtesi nul·la:

$$\beta_2 + \beta_3 + \beta_4 = 0 \tag{4-27}$$

Per realitzar aquest contrast utilitzarem un procediment similar al de l'Exemple 4.6. Ara, el paràmetre θ es defineix de la següent manera

$$\theta = \beta_2 + \beta_3 + \beta_4 \tag{4-28}$$

Fent $\beta_2 = \theta - \beta_3 - \beta_4$, s' ha estimat el següent model:

$$\ln(fish) = \beta_1 + \theta \ln(fishpr) + \beta_3 \ln(meatpr / fishpr) + \beta_4 \ln(cons / fishpr) + u \tag{4-29}$$

Els resultats obtinguts han estat els següents:

$$\ln(\widehat{fish}_i) = \underset{(2.30)}{7.788} - \underset{(0.1334)}{0.4596}\ln(fishpr_i) + \underset{(0.112)}{0.554}\ln(meatpr_i) + \underset{(0.137)}{0.322}\ln(cons_i)$$

Usant (4-28), contrastar la hipòtesi nul·la (4-27) és equivalent a contrastar que el coeficient de ln(*fishpr*) a (4-29) és igual a 0. Atès que l'estadístic *t* per aquest coeficient és igual a -3.44 i $t_{24}^{0.01/2}$ =2.8, rebutgem la hipòtesi d'homogeneïtat de la demanda de peix.

4.2.4 Importància econòmica versus significació estadística

Fins ara hem posat èmfasi en la significació estadística. No obstant això, és important recordar que hem de prestar atenció a la magnitud i el signe dels coeficients estimats, a més dels estadístics *t*.

La significació estadística de la variable x_j es determina completament per la mida de $t_{\hat{\beta}_j}$, mentre que la importància econòmica d'una variable es relaciona amb la mida (i signes) de $\hat{\beta}_j$. Si es posa massa èmfasi en la significació estadística pot conduir-nos a la falsa conclusió que una variable és "important" per explicar i, encara que el seu efecte estimat siga modest.

Així que, fins i tot si una variable és estadísticament significativa, cal analitzar la magnitud del coeficient estimat per tenir una idea de la seua importància pràctica o econòmica.

4.3 Contrast de restriccions lineals múltiples utilitzant l'estadístic *F*.

A les restriccions lineals múltiples distingirem tres tipus: les restriccions d'exclusió, la significativitat del model i altres restriccions lineals $\beta_1, \beta_2, \beta_3, ..., \beta_k$.

A les restriccions lineals múltiples distingirem tres tipus: les *restriccions d'exclusió*, la *significativitat del model* i *altres restriccions lineals*.

4.3.1 Restriccions d'exclusió

Hipòtesi nul·la i alternativa; models no restringit i restringit

Hipòtesi nul·la i alternativa; models no restringit i restringit

Comencem contrastant si un conjunt de variables independents té o no un efecte parcial sobre la variable dependent, *y*. A aquest contrast se li denomina de *restriccions d'exclusió*. Així, considerant el model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u \qquad (4\text{-}30)$$

la hipòtesi nul·la en un típic exemple de restriccions d'exclusió podria ser la següent:

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

Aquest és un exemple de restriccions múltiples, ja que s'ha imposat més d'una restricció en els paràmetres del model. Un contrast de restriccions múltiples és denominat també contrast conjunt d'hipòtesis.

La hipòtesi alternativa es pot expressar de la següent manera:

H_1: H_0 no és certa

És important destacar que s'estima H_0 conjuntament i no individualment. Ara, anem a distingir entre el model no restringit (*NR*) i el restringit (*R*). El model no restringit és el model de referència o model inicial. En aquest exemple el model no restringit és el model que figura a (4-30). El model restringit s'obté mitjançant la imposició de H_0 en el model original. En l'anterior exemple el model restringit és

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

Per definició, el model restringit sempre té un menor nombre de paràmetres que el model no restringit. A més, sempre es verifica que

$$SQR_R \geq SQR_{NR}$$

on SQR_R és la *SQR* del model restringit, i SQR_{NR} és la *SQR* del model no restringit. Recordeu que, pel fet que les estimacions dels coeficients per *MQO* es trien de manera que minimitzin la suma dels quadrats dels residus, la *SQR* no disminueix (i en general augmenta) quan algunes restriccions (com l'eliminació de variables) s'introdueixen en el model.

L'augment en la *SQR* quan s'imposen restriccions pot indicar-nos alguna cosa sobre si és versemblant la H_0. Si l'increment és gran, això és una evidència en contra de H_0, i aquesta hipòtesi serà rebutjada. Si l'increment és xicotet, això no serà una evidència en contra de H_0, i aquesta hipòtesi no serà rebutjada. La pregunta és aleshores si l'augment observat en la *SQR* quan s'imposen les restriccions, és prou gran, en relació amb la *SQR* en el model no restringit, per justificar el rebuig de H_0.

La resposta depèn és clar de α, però no podem dur a terme el contrast sobre el α seleccionat fins que tinguem un estadístic la distribució del qual siga coneguda i estiga tabulat sota la H_0. Per tant, necessitem una manera de combinar la informació de la SQR_R i de la SQR_{NR} per obtenir un estadístic de contrast amb una distribució coneguda sota H_0.

Ara, vegem el cas general, on el model no restringit és el següent:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + u \tag{4-31}$$

Suposem que hi ha *q* restriccions d'exclusió a contrastar. Aleshores, H_0 postula que q variables tenen coeficients zero. Si s'assumeix que són les últimes *q* variables, la H_0 s'expressa com

$$H_0 : \beta_{k-q+1} = \beta_{k-q+2} = ... = \beta_k = 0 \tag{4-32}$$

El model restringit s'obté mitjançant la imposició de *q* restriccions de la H_0 en el model no restringit:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_{k-q} x_{k-q} + u \tag{4-33}$$

La H_1 s'expressa com

$$H_1: H_0 \text{ no és certa} \tag{4-34}$$

Estadístic de contrast: la ràtio F

L'estadístic *F*, o la *ràtio F*, està definit per

$$F = \frac{(SCR_R - SCR_{SR}) / q}{SCR_{SR} / (n-k)} \tag{4-35}$$

on SQR_R és el *SQR* del model restringit, SQR_{NR} és el *SQR* del model no restringit i *q* és el nombre de restriccions, és a dir, el nombre d'igualtats en la hipòtesi nul·la.

Per poder utilitzar l'estadístic *F* per al contrast d'hipòtesis, hem de conèixer la seva distribució mostral sota H_0 per tal d'escollir el valor de *c* per a un α donat, i determinar la regla de rebuig. Es pot demostrar que, sota la H_0, i assumint que els supòsits del *MLC*

es mantenen, l'estadístic *F* es distribueix com una variable aleatòria *F* de Snedecor amb *q* i *n-k* graus de llibertat. Vam escriure aquest resultat de la següent manera

$$F \mid H_0 \sim F_{q,n-k} \tag{4-36}$$

Una *F* de Snedecor amb *q graus de llibertat en el numerador* i *n-k graus de llibertat en el denominador* és igual a

$$F_{q,n-k} = \frac{\chi_q^2 / q}{\chi_{n-k}^2 / n-k} \tag{4-37}$$

on x_q^2 i x_{n-k}^2 són distribucions Chi-quadrat independents l'una de l'altra.

En (4-35) s'observa que els graus de llibertat que corresponen a la SQR_{NR} (gl_{NR}) són *n-k.* Recordeu que

$$\hat{\sigma}_{NR}^2 = \frac{SQR_{NR}}{n-k} \tag{4-38}$$

D'altra banda, els graus de llibertat que corresponen a la SQR_R (gl_R) són *n-k+q*, perquè en el model restringit s'estimen *k-q* paràmetres. Els graus de llibertat que correspon a SQR_R-SQR_{SR} són

$$(n\text{-}k+q)\text{-}(n\text{-}k)=q = \textit{nombre de graus de llibertat}=gl_R\text{-}gl_{NR}$$

Així, en el numerador de la *F*, la diferència entre les *SQR* es divideix per *q*, que és el nombre de restriccions imposades en passar del model no restringit al restringit. En el denominador de la *F*, SQR_{NR} De fet, el denominador de la F és justament l'estimador de σ^2 en el model no restringit gl_{NR}. De fet, el denominador de la *F* és justament l'estimador de σ^2 en el model no restringit.

La *ràtio F* ha de ser més gran que o igual a 0, ja que $SQR_R - SQR_{NR} \geq 0$.

Sovint és convenient tenir una forma de l'estadístic *F* tal que puga ser calculada a partir de R^2 dels models restringit i no restringit.

Utilitzant el fet de que $SQR_R = SQT(1 - R_R^2)$ i $SQR_{NR} = SQT(1 - R_{NR}^2)$, podem expressar (4-35) de la següent manera:

$$F = \frac{(R_{SR}^2 - R_R^2) / q}{(1 - R_{SR}^2) / (n-k)} \tag{4-39}$$

ja que el terme *SQT* es cancel·la.

Això expressió s'anomena la forma *R-quadrat de l'estadístic F*.

Atès que, encara que la forma *R-quadrat de l'estadístic F* és molt convenient per a contrastar restriccions d'exclusió, no pot aplicar-se per contrastar tota mena de restriccions lineals. Per exemple, la *ràtio F* (4-39) no es pot utilitzar quan el model no té terme independent ni quan la forma funcional de la variable endògena en el model restringit no és la mateixa que en el model no restringit.

Regla de decisió

La distribució $F_{q,n-k}$ està tabulada i disponible en taules estadístiques, on es busca el valor crític ($F^{\alpha}_{q,n-k}$), que depèn de α (nivell de significació), q (*gl* del numerador), i $n-k$, (*gl* del denominador). Tenint en compte l'anterior, la regla de decisió és molt simple.

Regla de decisió

$$\begin{array}{llll} \text{Si} & F \geq F^{\alpha}_{q,n-k} & \text{es rebutja} & H_0 \\ \text{Si} & F < F^{\alpha}_{q,n-k} & \text{no es rebutja} & H_0 \end{array} \qquad (4\text{-}40)$$

Per tant, rebutgem la H_0 en favor de la H_1 en α quan $F \geq F^{\alpha}_{q,n-k}$, com es pot veure a la figura 4.15. És important destacar que quan α disminueix, augmenta $F^{\alpha}_{q,n-k}$. Si es rebutja la H_0, aleshores diem que $x_{k-q+1}, x_{k-q+2}, \cdots, x_k$ són *estadísticament significatius conjuntament*, o, més breu, *significatius conjuntament*, per al nivell de significació seleccionat.

Aquest contrast per si sol no ens permet dir quines de les variables tenen un efecte parcial sobre *y*, ja que totes elles poden afectar *y*, o potser només una afecta *y*. Si no es rebutja H_0, aleshores diem que no són estadísticament significatives conjuntament, o simplement que no són significatives conjuntament, el que sovint justifica la seua eliminació del model. L'estadístic *F* és sovint útil per contrastar l'exclusió d'un grup de variables quan les variables del grup estan altament correlacionades entre si

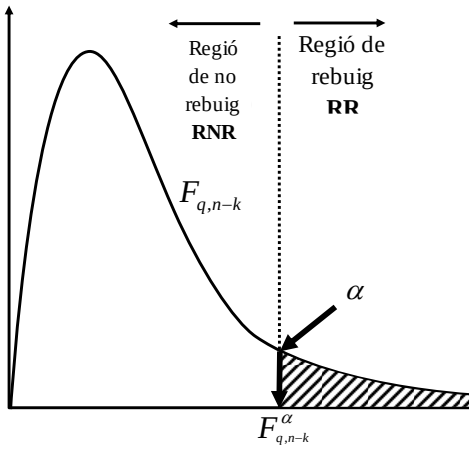


FIGURA 4.15. Regió de rebuig i regió de no rebuig utilitzant la distribució *F*.

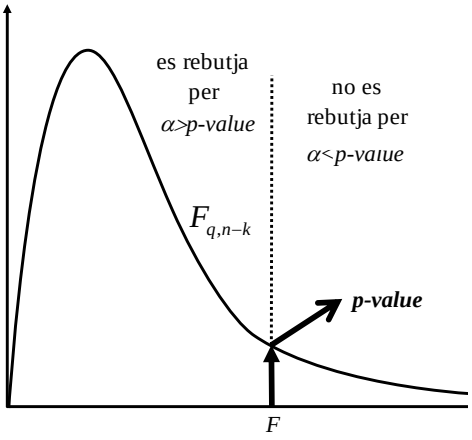


FIGURA 4.16. *Valor-p* utilitzant la distribució *F*.

En el context de l'estadístic *F*, el *valor-p* es defineix como

$$valor\text{-}p = \Pr(F > F' \mid H_0)$$

on *F* és el valor real de l'estadístic de contrast i F' designa una variable aleatòria *F* de Snedecor amb *q* i *n-k graus de llibertat*.

El *valor-p* té la mateixa interpretació que en l'estadístic *t*. Un *valor-p* xicotet és una evidència en contra de la H_0. Per contra, un *valor-p* elevat no constitueix una evidència en contra de la H_0. Una vegada que *valor-p* ha estat calculat, el contrast *F* es

pot dur a terme per a qualsevol nivell de significació. A la figura 4.16 es representa aquest enfocament alternatiu. Com es desprèn de l'observació de la figura, la determinació del valor-p és l'operació inversa a la de trobar el valor en les taules estadístiques per a un determinat nivell de significació. Una vegada que el *valor-p* ha estat determinat, se sap que es rebutja H_0 per a qualsevol nivell de significació tal que α>*valor-p*; per contra, la hipòtesi nul·la no es rebutja quan α<*valor-p*.

EXEMPLE 4.10 Salaris, experiència, antiguitat i edat

El següent model ha estat formulat per analitzar els factors determinants dels salaris:

$$\ln(wage) = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 tenure + \beta_5 age + u$$

on *wage* són els salaris mensuals, *educ* són els anys d'educació, *expert* són els anys d'experiència laboral, *tenure* són els anys treballant amb l'empresa actual, i *age* és l'edat en anys.

L'investigador té la intenció d'excloure *tenure* del model, ja que en molts casos és igual a l'experiència, i també l'edat, ja que està altament correlacionada amb l'experiència. És acceptable l'exclusió de les dues variables?

Les hipòtesis nul·la i alternativa són les següents:

$$H_0 : \beta_4 = \beta_5 = 0$$
$$H_1 : H_0 \text{ no es certa}$$

El model restringit corresponent a aquesta H_0 és

$$\ln(wage) = \beta_1 + \beta_2 educ + \beta_3 exper + u$$

Utilitzant una mostra de 53 observacions del fitxer *wage2*, s'han obtingut les següents estimacions per als models no restringit i restringit:

$$\ln(wage_i) = 6.476 + 0.0658 educ_i + 0.0267 exper_i - 0.0094 tenure_i - 0.0209 age_i \quad SCR = 5.954$$
$$\ln(wage_i) = 6.157 + 0.0457 educ_i + 0.0121 exper_i \quad SCR = 6.250$$

La ràtio *F* obtinguda és la següent:

$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(n-k)} = \frac{(6.250 - 5.954)/2}{5.954/48} = 1.193$$

Tenint en compte que l'estadístic *F* és baix, anem a veure què passa amb un nivell de significació del 0,10. En aquest cas, els graus de llibertat per al denominador són 48 (53 observacions menys 5 paràmetres estimats). Si busquem a les taules l'estadístic *F* per a 2 graus de llibertat al numerador i 45 graus de llibertat en el denominador, trobem $F^{0.10}_{2,48} \simeq F^{0.10}_{2,45} = 2.42$. Com *F*<2.42 no es rebutja H_0. Si no es rebutja la H_0 per 0.10, no s'ha de rebutjar tampoc per 0.05 o 0.01, com es pot a la figura 4.17. Per tant, no podem rebutjar H_0 en favor de H_1. En altres paraules, l'antiguitat en l'empresa i l'edat no són conjuntament significatives.

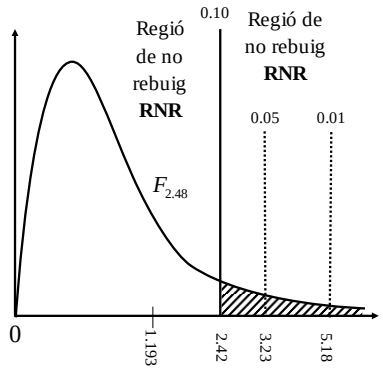


FIGURA 4.17. Exemple 4.10: Regió de rebuig en la distribució F (els valors α són per $F_{2.40}$).

4.3.2 Significació global del model

Contrastar la significació del model, o significació global del model, és un cas particular dels contrastos de restriccions d'exclusió. Es podria pensar que aquest contrast la H_0 hauria de ser la següent:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_k = 0 \tag{4-41}$$

No obstant això, això no és la H0 adequada per contrastar la significativitat global del model. Si $\beta_2 = \beta_3 = ... = \beta_k = 0$, aleshores el model restringit seria el següent:

$$y = \beta_1 + u \tag{4-42}$$

Si prenem esperances en (4-42), obtenim que

$$E(y) = \beta_1 \tag{4-43}$$

Així, la H_0 en (4-41) implica que no només que les variables explicatives no tenen cap influència sobre la variable endògena, sinó també que la mitjana de la variable endògena -per exemple, el consum mitjà- és igual a 0.

Per tant, si volem conèixer si el model és globalment significatiu, la H_0 ha de ser la següent:

$$H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0 \tag{4-44}$$

El corresponent model restringit donat a (4-42) no explica res i, per tant, R_R^2 es igual a 0. Aleshores, contrastar la H_0 donada en (4-44) és molt fàcil utilitzant la forma *R-quadrat* de l'estadístic *F*:

$$F = \frac{R^2 / k}{(1 - R^2) / (n - k)} \tag{4-45}$$

on $R^2 = R_{NR}^2$, ja que només cal estimar el model no restringit, a causa que el R^2 del model (4-42) -model restringit es igual a 0.

EXEMPLE 4.11 Salaris de directors executius

Considere la següent equació per a explicar els salaris (*salary*) dels directors executius en funció de les vendes anuals de l'empresa (*sales*), la rendibilitat sobre recursos propis (*roe*) (en forma de percentatge), i el rendiment de les accions de l'empresa (*ros*) (en forma de percentatge):

$$\ln(salary) = \beta_1+\beta_2\ln(sales)+\beta_3 roe+\beta_4 ros+ \text{ u}.$$

La pregunta que es planteja és si el rendiment de l'empresa mesurat per les variables *sales*, *roe* i *ros* és crucial per establir els salaris dels directors executius. Per respondre a aquesta pregunta farem un contrast de significació global. Les hipòtesis nul·la i alternativa són les següents:

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$

H_1: H_0 no és certa

El quadre 4.9 mostra una sortida d'E-views completa de *mínims quadrats* (*ls*), utilitzant el fitxer *ceosal1*. A la part inferior es pot veure el "estadístic F" per al contrast de significació global, així com "Prob.", que és el *valor-p* corresponent a aquest estadístic. En aquest cas *valor-p* és igual a 0, és a dir, es rebutja *H0* per a tots els nivells de significació (Vegeu la figura 4.18). Per tant, podem rebutjar que el rendiment de l'empresa no tinga cap influència sobre els salaris dels directors executius.

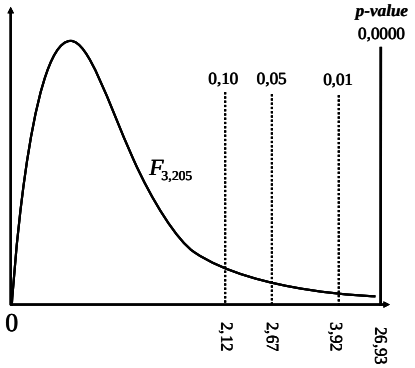


FIGURA 4.18. Exemple 4.11: *Valor-p* utilitzant la distribució *F* (els valors α per a una $F_{3,140}$).

QUADRE 4.9. Sortida completa d'E-views en l'Exemple 4.11.

Dependent Variable: LOG(SALARY)
Method: Least Squares
Date: 04/12/12 Time: 19:39
Sample: 1 209
Included observations: 209

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.311712	0.315433	13.66919	0.0000
LOG(SALES)	0.280315	0.03532	7.936426	0.0000
ROE	0.017417	0.004092	4.255977	0.0000
ROS	0.000242	0.000542	0.446022	0.6561

R-squared	0.282685	Mean dependent var	6.950386
Adjusted R-squared	0.272188	S.D. dependent var	0.566374
S.E. of regression	0.483185	Akaike info criterion	1.402118
Sum squared resid	47.86082	Schwarz criterion	1.466086
Log likelihood	-142.5213	**F-statistic**	**26.9293**
Durbin-Watson stat	2.033496	**Prob(F-statistic)**	**0.0000**

4.3.3 Estimant altres restriccions lineals

Fins al moment, hem contrastat hipòtesi amb restriccions d'exclusió mitjançant la utilització estadístic *F*. Però també podem contrastar hipòtesis amb restriccions lineals de qualsevol tipus. Per tant, podem combinar diversos tipus de restriccions: restriccions d'exclusió, restriccions que imposen determinats valors als paràmetres i restriccions relatives a una combinació lineal dels paràmetres.

Així, considerem el següent model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u$$

amb la hipòtesi nul·la:

$$H_0 : \begin{cases} \beta_2 + \beta_3 = 1 \\ \beta_4 = 3 \\ \beta_5 = 0 \end{cases}$$

El model restringit corresponent a aquesta hipòtesi nul·la és

$$(y - x_2 - 3x_4) = \beta_1 + \beta_3 (x_3 - x_2) + u$$

En l'Exemple 4.12, que es veurà a continuació, la hipòtesi nul·la consisteix en dues restriccions: una combinació lineal de paràmetres i una restricció d'exclusió.

EXEMPLE 4.12 Una restricció addicional en la funció de producció. (Continuació de l'Exemple 4.7)

En la funció de producció de Cobb-Douglas, anem a contrastar la següent H_0, que té 2 restriccions:

$$H_0 : \begin{cases} \beta_2 + \beta_3 = 1 \\ \beta_1 = 0 \end{cases}$$
$$H_1 : H_0 \text{ no es certa}$$

A la primera restricció s'imposa que hi ha rendiments constants a escala. A la segona restricció β_1, el paràmetre relacionat amb la productivitat total dels factors, és igual a 0.

Substituint la restricció de la H_0 en el model original (model sense restriccions), tenim

$$\ln(output) = (1 - \beta_3)\ln(labor) + \beta_3 \ln(capital) + u$$

Operant, s'obté el model restringit:

$$\ln(output / labor) = \beta_3 \ln(capital / labor) + u$$

En l'estimació dels models no restringit i restringit, obtenim que SQR_R=3.1101 i SQR_{NR}=0.8516. Per tant, la *ràtio F* és igual a

$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(n-k)} = \frac{(3.1101 - 0.8516)/2}{0.8516/(27-3)} = 13.551$$

Hi ha dues raons per no utilitzar R^2 en aquest cas. En primer lloc, el model restringit no té terme independent. En segon lloc, el regressant del model restringit és diferent del regressant del model no restringit.

Atès que el valor de F és relativament alt, anem a començar el contrastar amb un nivell de significació de l'1%. Per α=0.01, $F_{2,24}^{0.01} = 5.61$. Tenint en compte que F>5.61, rebutgem H_0 en favor de H_1.

Per tant, es rebutja la hipòtesi conjunta que hi ha rendiments constants a escala i que el paràmetre β_1 és igual a 0. Si es rebutja H_0 per α=0.01, també serà rebutjada per als nivells de 5% i 10%.

4.3.4 Relació entre els estadístics *F* i *t*

Fins ara, hem vist com s'utilitza l'estadístic *F* per contrastar diverses restriccions en el model, però també pot utilitzar-se per contrastar una sola restricció. En aquest cas, podem triar entre l'estadístic *F* o l'estadístic *t* per fer un contrast de dues cues. En qualsevol cas, les conclusions seran exactament les mateixes.

Ara bé, ¿quina és la relació entre una *F* amb un grau de llibertat en el numerador (per contrastar una sola restricció) i una *t*? Es pot demostrar que

$$t_{n-k}^2 = F_{1,n-k} \tag{4-46}$$

Aquest fet s'il·lustra a la figura 4,19. S'observa que la cua de la *F* s'ha desdoblat en les dues cues de la *t*. Per tant, els dos enfocaments condueixen exactament al mateix resultat, sempre que la hipòtesi alternativa siga de dues cues. No obstant això, l'estadístic *t* és més flexible per contrastar una hipòtesi amb una sola restricció, ja que pot utilitzar-se per al contrastar la H_0 contra una alternativa de sola cua.

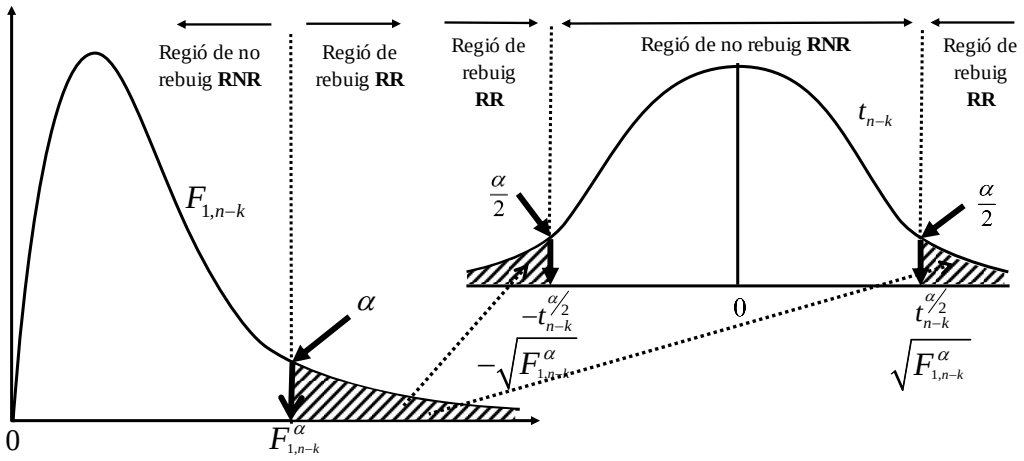


FIGURA 4.19. Relació entre $F_{1,n-k}$ i t_{n-k}.

A més, atès que els estadístics *t* són també més fàcils d'obtenir que els estadístics *F*, en realitat no hi ha una bona raó per utilitzar un estadístic *F* per contrastar una hipòtesi amb una sola restricció.

4.4 Contrastos sense normalitat

La normalitat dels estimadors de *MQO* depèn crucialment del supòsit de normalitat de les pertorbacions. Què passa si les pertorbacions no tenen una distribució normal? Hem vist que les pertorbacions sota els supòsits de Gauss-Markov, i en conseqüència, els estimadors de *MQO* tenen una distribució asimptòtica normal, és a dir, tenen aproximadament una distribució normal.

Si les pertorbacions no són normals, l'estadístic *t* no té una distribució *t* exacta, sinó només aproximada. Com es pot veure a la taula de la *t* de Student, per una mida mostral de 60 observacions dels punts crítics són pràcticament igual als de la distribució normal estàndard.

De la mateixa manera, si les pertorbacions no són normals, l'estadístic *F* no té una distribució *F exacta*, sinó només *aproximada*, quan la grandària mostral és prou gran i els supòsits de Gauss-Markov es compleixen. Aleshores, podem utilitzar l'estadístic *F* per contrastar les restriccions lineals en models lineals com un contrast aproximat.

Hi ha altres contrastos asimptòtics (la raó de versemblança, el multiplicador de Lagrange i el contrast de Wald) basats en les funcions de versemblança que poden utilitzar-se en contrastar restriccions lineals quan les pertorbacions no es distribueixen normalment. Aquests tres estadístiques també poden aplicar-se quan a) les restriccions són no lineals, i b) el model és no lineal en els paràmetres. Per a les restriccions no lineals, en els models lineals i no lineals, el contrast més utilitzat és el contrast de Wald.

Per contrastar els supòsits del model (per exemple, els d'homoscedasticitat i no autocorrelació) s'aplica generalment el multiplicador de Lagrange (*ML*). En l'aplicació del contrast de *ML* s'estima sovint una *regressió auxiliar*. El nom de regressió auxiliar significa que els coeficients no són d'interès directe: d'aquesta regressió auxiliar només es conserva el R^2. En una regressió auxiliar el regressant és, en general, els residus o funcions dels residus, obtinguts en l'estimació per *MQO* del model original, mentre que les variables independents són, sovint, els regressors (i/o funcions dels mateixos) del model original.

4.5 Predicció

En aquest epígraf s'examinaran dos tipus de predicció: predicció puntual i predicció per intervals.

4.5.1 Predicció puntual

L'obtenció d'una predicció puntual no planteja cap problema especial, ja que és una operació simple d'extrapolació en el context de mètodes descriptius.

Designem per $x_2^0, x_3^0, \cdots, x_k^0$ als valors específics de cada un dels *k* regressors en la predicció; aquests valors poden o no correspondre a un punt real de dades a la nostra mostra. Si substituïm aquests valors en el model de regressió múltiple, tenim

$$y^0 = \beta_1 + \beta_2 x_2^0 + \beta_3 x_3^0 + ... + \beta_k x_k^0 + u^0 = \theta^0 + u^0 \qquad (4\text{-}47)$$

Per tant, l'esperança, o mitjana, del valor de *y* ve donada per

$$E(y^0) = \beta_1 + \beta_2 x_2^0 + \beta_3 x_3^0 + ... + \beta_k x_k^0 = \theta^0 \qquad (4\text{-}48)$$

La predicció puntual s'obté de forma immediata mitjançant la substitució dels paràmetres de (4-48) pels estimadors de *MQO* corresponents:

$$\hat{\theta}^0 = \hat{\beta}_1 + \hat{\beta}_2 x_2^0 + \hat{\beta}_3 x_3^0 + \ldots + \hat{\beta}_k x_k^0 \qquad (4\text{-}49)$$

Per obtenir (4-49) no es necessita postular cap supòsit estadístic. Però, si adoptem els supòsits 1 a 6 del *MLC*, és fàcil concloure que $\hat{\theta}^0$ és un predictor no esbiaixat de θ^0

$$E\left[\hat{\theta}^0\right] = E\left[\hat{\beta}_1 + \hat{\beta}_2 x_2^0 + \hat{\beta}_3 x_3^0 + \ldots + \hat{\beta}_k x_k^0\right] = \beta_1 + \beta_2 x_2^0 + \beta_3 x_3^0 + ... + \beta_k x_k^0 = \theta^0 \quad (4\text{-}50)$$

D'altra banda, adoptant els supòsits de Gauss Markov (1 a 8), es pot demostrar que aquest predictor puntual és l'estimador lineal no esbiaixat òptim (*ELNEO*).

Tenim ara una predicció puntual per θ^0, però, ¿quina és la predicció puntual per y^0? Per respondre a aquesta pregunta hem de predir u_0. Com l'error no és observable, el millor predictor de u_0 és el seu valor esperat, que és 0. Per tant,

$$\hat{y}^0 = \hat{\theta}^0 \tag{4-51}$$

4.5.2 Predicció per intervals

Les prediccions puntuals fetes amb un model economètric no coincidiran, en general, amb els valors observats, a causa de la incertesa que envolta els fenòmens econòmics.

La primera font d'incertesa és que no podem utilitzar la funció de regressió poblacional, ja que no coneixem els paràmetres β. En el seu lloc hem d'utilitzar la funció de regressió mostral. L'*interval de confiança per al valor esperat*, és a dir, pe θ^0, que examinarem a continuació, inclou només aquest tipus d'incertesa.

La segona font d'incertesa és que en un model economètric, a més de la part sistemàtica, hi ha una pertorbació que no és observable. La predicció per *interval per un valor individual*, és a dir per a y^0, que es discutirà més endavant inclou tant la incertesa derivada de l'estimació com del terme de pertorbació.

Una tercera font d'incertesa pot provenir del fet de no saber exactament els valors que les variables explicatives prendran en el moment de predicció. Aquesta tercera font d'incertesa, que no s'aborda aquí, complica els càlculs per a la construcció d'intervals.

Interval de confiança per al valor esperat

Si volem predir el valor esperat de *y*, això és θ^0, aleshores, l'error de predicció $\hat{e}_1^0$ serà $\hat{e}_1^0 = \theta^0 - \hat{\theta}^0$. D'acord amb (4-50), l'error de predicció esperat és zero. Sota els supòsits del *MLC*,

$$\frac{\hat{e}_1^0}{ee(\hat{\theta}^0)} = \frac{\theta^0 - \hat{\theta}^0}{ee(\hat{\theta}^0)} \sim t_{n-k}$$

Per tant, podem escriure que

$$\Pr\left[-t_{n-k}^{\alpha/2} \le \frac{\theta^0 - \hat{\theta}^0}{ee(\hat{\theta}^0)} \le t_{n-k}^{\alpha/2} \right] = 1-\alpha$$

Operant, podem construir un interval de confiança (*IC*) del (1-α)% θ^0 amb l'estructura següent:

$$\Pr\left[\hat{\theta}^0 - ee(\hat{\theta}^0) \times t_{n-k}^{\alpha/2} \le \theta^0 \le \hat{\theta}^0 + ee(\hat{\theta}^0) \times t_{n-k}^{\alpha/2} \right] = 1-\alpha \tag{4-52}$$

Per obtenir un *IC* per θ^0, necessitem conèixer l'error estàndard ($ee(\hat{\theta}_0)$) per $\hat{\theta}^0$. En qualsevol cas, hi ha una manera fàcil d'estimar-lo. Així, resolent (4-48) per β_1 veiem que $\beta_1 = \theta^0 - \beta_2 x_2^0 - \beta_3 x_3^0 - ... - \beta_k x_k^0$. Posant aquest resultat en l'equació (4-47), obtenim

$$y = \theta^0 + \beta_2(x_2 - x_2^0) + \beta_3(x_3 - x_3^0) + \ldots + \beta_k(x_k - x_k^0) + u \qquad (4\text{-}53)$$

Aplicant *MQO* a (4-53), a més de la predicció puntual, s'obté $ee(\hat{\theta}^0)$, l'error estàndard corresponent al terme independent d'aquesta regressió. El mètode anterior ens permet obtenir un *IC* d'E(*y*), de l'estimació per *MQO*, per a diferents valors dels regressors.

Predicció per intervals per a un valor individual

Ara construirem un interval per y^0, al que es denomina *predicció per intervals per un valor individual*, o d'una manera més breu, *predicció per intervals*. D'acord amb (4-47), y^0 té dos components:

$$y^0 = \theta^0 + u^0 \qquad (4\text{-}54)$$

L'*interval per al valor esperat* construït prèviament és un interval de confiança al voltant de θ^0, que és una combinació dels paràmetres. En contrast, l'interval per y^0 és aleatori, perquè un dels seus components, u^0, és aleatori. Per tant, l'interval per y^0 és un interval probabilístic i no un interval de confiança. El mecanisme per a la seva obtenció és el mateix, però tenint en compte que ara considerarem que el conjunt $x_2^0, x_3^0, \cdots, x_k^0$ es troba fora de la mostra utilitzada per estimar la regressió.

L'error de predicció de $\hat{y}^0$ a predir y^0 és

$$\hat{e}_2^0 = y^0 - \hat{y}^0 = \theta^0 + u^0 - \hat{y}^0 \qquad (4\text{-}55)$$

Tenint en compte (4-51) i (4-50), i que $E(u^0)=0$, l'error de predicció esperat és zero. En la recerca de la variança de $\hat{e}_2^0$, cal tenir en compte que u^0 no està correlacionat amb $\hat{y}^0$, perquè $x_2^0, x_3^0, \cdots, x_k^0$ no pertanyen a la mostra.

Per tant, la diferència d'error de predicció (condicionada a *x*) és la suma de les variances:

$$Var(\hat{e}_2^0) = Var(\hat{y}^0) + Var(u^0) = Var(\hat{y}^0) + \sigma^2 \qquad (4\text{-}56)$$

Així doncs, hi ha dos fonts de variació en $\hat{e}_2^0$:

1. L'error mostral de $\hat{y}^0$ sorgeix perquè s'han estimat les β_j.
2. La ignorància dels factors no observats que afecten *y*, els quals es reflecteixen en σ^2.

Sota els supòsits del *MLC*, $\hat{e}_2^0$ també es distribueix normalment. Utilitzant l'estimador no esbiaixat de σ^2 i tenint en compte que $var(\hat{y}^0) = var(\hat{\theta}^0)$, podem definir l'error estàndard (*ee*) de $\hat{e}_2^0$ com

$$ee(\hat{e}_2^0) = \left\{ \left[ee(\hat{\theta}^0) \right]^2 + \hat{\sigma}^2 \right\}^{\frac{1}{2}} \qquad (4\text{-}57)$$

En general, $\hat{\sigma}^2$ és més gran que $\left[ee(\hat{\theta}^0) \right]^2$. Sota els supòsits del *MLC*,

$$\frac{\hat{e}_2^0}{ee(\hat{e}_2^0)} \sim t_{n-k} \tag{4-58}$$

Per tant, podem escriure que

$$\Pr\left[-t_{n-k}^{\alpha/2} \le \frac{\hat{e}_2^0}{ee(\hat{e}_2^0)} \le t_{n-k}^{\alpha/2}\right] = 1-\alpha \tag{4-59}$$

En substituir $\hat{e}_2^0 = y^0 - \hat{y}^0$ en (4-59) i reordenant s'obté un *interval de predicció* del (1-α)% per y^0:

$$\Pr\left[\hat{y}^0 - ee(\hat{e}_2^0)\times t_{n-k}^{\alpha/2} \le y^0 \le \hat{y}^0 + ee(\hat{e}_2^0)\times t_{n-k}^{\alpha/2}\right] = 1-\alpha \tag{4-60}$$

EXEMPLE 4. 13 Quina és la puntuació esperada en l'examen final si s'han obtingut 7 punts en la primera avaluació?

S'ha estimat el següent model per comparar les puntuacions en l'examen final (*finalmrk*) i en la primera avaluació (*primeval*) d'Econometria:

$$finalmrk_i = \underset{(0.715)}{4.155} + \underset{(0.123)}{0.491}\, primeval_i$$

$$\hat{\sigma} = 1.649 \quad R^2 = 0.533 \quad n = 16$$

Per estimar la nota final esperada per a un estudiant amb $primeval^0 = 7$ a la primera avaluació, es va estimar el següent model, d'acord amb (4-53):

$$finalmrk_i = \underset{(0.497)}{7.593} + \underset{(0.123)}{0.491}\ (primeval_i - 7)$$

$$\hat{\sigma} = 1.649 \quad R^2 = 0.533 \quad n = 16$$

La predicció puntual de partida per $primeval^0 = 7$ és $\hat{\theta}_0 = 7.593$ i els límits inferior i superior d'un *IC* del 95%, respectivament, vénen donats per

$$\underline{\theta}^0 = \hat{\theta}^0 - ee(\hat{\theta}^0)\times t_{14}^{0.05/2} = 7.593 - 0.497\times 2.14 = 6.5$$

$$\overline{\theta}^0 = \hat{\theta}^0 + ee(\hat{\theta}^0)\times t_{14}^{0.05/2} = 7.593 + 0.497\times 2.14 = 8.7$$

Per tant, l'estudiant tindrà una confiança del 95% d'obtenir, de mitjana, una qualificació final situada entre 6.5 i 8.7.

La predicció puntual pot obtenir-se també a partir de la primera equació estimada:

$$finalmrk = 4.155 + 0.491\times 7 = 7.593$$

Ara, anem a estimar un interval de probabilitat del 95% per al valor individual. L'error estàndard de $\hat{e}_2^0$ és igual a

$$ee(\hat{e}_2^0) = \left\{\left[ee(\hat{y}^0)\right]^2 + \hat{\sigma}^2\right\}^{\frac{1}{2}} = \sqrt{0.497^2 + 1.649^2} = 1.722$$

on 1.649 és l'error estàndard de la regressió (E.E.), s'ha obtingut de la sortida d'E-views directament.

Els límits inferior i superior d'un interval de probabilitat del 95%, respectivament, vénen donats per

$$\underline{y}^0 = \hat{y}^0 - ee(\hat{e}_2^0) \times t_{14}^{0.025} = 7.593 - 1.722 \times 2.14 = 3.7$$

$$\overline{y}^0 = \hat{y}^0 + ee(\hat{e}_2^0) \times t_{14}^{0.025} = 7.593 + 1.722 \times 2.14 = 11.3$$

Cal tenir en compte que aquest interval de probabilitat és bastant gran a causa de que la mida de la mostra és molt xicotet.

EXEMPLE 4.14 Predient el salari dels directors executius

Utilitzant les dades de les empreses més importants dels EUA agafades de Forbes (fitxer *ceoforbes*), s'ha estimat la següent equació per explicar els salaris - *salary* -(incloent bonificacions) anuals obtinguts (en milers de dòlars) en 1999 pels directors executius d'aquestes empreses:

$$\underset{}{salary_i} = \underset{(104)}{1381} + \underset{(0.0013)}{0.008377}\, assets_i + \underset{(8.671)}{32.508}\, tenure_i + \underset{(0.0538)}{0.2352}\, profits_i$$

$$\hat{\sigma} = 1506 \quad R^2 = 0.2404 \quad n = 447$$

on *assets* són els actius totals de la firma a milions de dòlars, *tenure* és el nombre d'anys com a director executiu de la companyia, i *profits* són els beneficis en milions de dòlars.

En el quadre 4.10 apareixen les mesures descriptives de les variables explicatives del model dels salaris dels directors generals.

QUADRE 4.10. Mesures descriptives de les variables del model sobre el salari dels executius.

	assets	*tenure*	*profits*
Mitjana	27054	7.8	700
Mediana	7811	5.0	333
Màxim	668641	60.0	22071
Mínim	718	0.0	-2669
N. d'observacions	447	447	447

Els salaris predits i els corresponents $ee(\hat{\theta}_0)$ per a valors seleccionats (màxim, mitja, mitjana i mínim), utilitzant un model com el (4-53), es presenten en el quadre 4.11.

QUADRE 4.11. Prediccions per als valors seleccionats.

	Predicció $\hat{\theta}_0$	*E. estàndard* $ee(\hat{\theta}_0)$
Valor mitjà	2026	71
Valor de la mediana	1688	78
Valor del màxim	14124	1110
Valor del mínim	760	195

4.5.3 Predicció de *y* en un model logarítmic

Considere el model en logaritmes:

$$\ln(y) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u \tag{4-61}$$

Un cop obtingudes les estimacions per *MQO*, podem predir ln(*y*) de la següent manera

$$\ln(y) = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \tag{4-62}$$

Prenent antilogaritmes a (4-62), obtenim el valor de predicció per *y*:

$$\tilde{y} = \exp(\ln(y)) = \exp(\hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k) \tag{4-63}$$

No obstant això, aquesta predicció és esbiaixada i inconsistent, ja que sistemàticament subestima el valor esperat de *y*. Vegem el per què. Si apliquem antilogaritmes a (4-61), obtenim que

$$y = \exp(\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k) \times \exp(u) \tag{4-64}$$

Abans de prendre les esperances en (4-64), cal tenir en compte que si $u \sim N(0,\sigma^2)$, aleshores $E(\exp(u)) = \exp\left(\frac{\sigma^2}{2}\right)$. Per tant, sota els supòsits 1 al 9 del *MLC* obtenim que

$$E(y) = \exp(\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k) \times \exp(\sigma^2 / 2) \tag{4-65}$$

Prenent com a referència (4-65) el predictor adequat de *y* és

$$\hat{y} = \exp(\hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k) \times \exp(\hat{\sigma}^2 / 2) = \tilde{y} \times \exp(\hat{\sigma}^2 / 2) \tag{4-66}$$

on $\hat{\sigma}^2$ és l'estimador no esbiaixat de σ^2.

És important destacar que $\hat{y}$ és un predictor esbiaixat però consistent, mentre que $\tilde{y}$ és esbiaixat i inconsistent.

EXEMPLE 4.15 Predient el salari dels directors executius amb un model logarítmic (continuació de 4.14)

Utilitzant les mateixes dades que a l'Exemple 4.14, es va estimar el següent model:

$$\ln(salary_i) = \underset{(0.210)}{5.5168} + \underset{(0.0232)}{0.1885}\ln(assets_i) + \underset{(0.0032)}{0.0125}\,tenure_i + \underset{(0.0000195)}{0.00007}\,profits_i$$

$$\hat{\sigma} = 0.5499 \quad R^2 = 0.2608 \quad n = 447$$

En el salari, *salary*, i en els actius, *assets*, s'han pres logaritmes naturals, però els guanys, *profits*, estan en nivells - és a dir, sense cap transformació, a causa que algunes observacions són negatives, pel que no és possible prendre logaritmes.

En primer lloc, anem a estimar la predicció d'acord amb (4-63) per a un director executiu que treballa en una empresa amb *assets*=10000, *tenure*=10 anys i *profits*=1000:

$$\begin{aligned} salary_i &= \exp(\ln(salary_i)) \\ &= \exp(5.5168 + 0.1885\ln(10000) + 0.0125 \times 10 + 0.00007 \times 1000) = 1716 \end{aligned}$$

Ara, utilitzant (4-66), s'obté la següent predicció que és consistent:

$$salary = \exp(0.5499^2 / 2) \times 1716 = 1996$$

4.5.4 Avaluació de les prediccions i predicció dinàmica

En aquesta secció anem a comparar les prediccions realitzades amb un model economètric i els valors realment observats per tal d'avaluar la capacitat predictiva del model. També examinarem la predicció dinàmica en models en què hi ha variables endògenes retardades incloses com regressors.

Estadístics per avaluació de prediccions

Suposem que es realitzen prediccions per a i=n+1, n+2,…, n+h, i que es designa el valor real i el previst en el període i com y_i i $\hat{y}_i$, respectivament. Ara, presentarem alguns dels estadístics més usuals utilitzats per a l'avaluació de les prediccions.

Error absolut mitjà (EAM)

El *EAM* es defineix com la mitjana dels valors absoluts dels errors:

$$EAM = \frac{\sum_{i=n+1}^{n+h} |\hat{y}_i - y_i|}{h} \tag{4-67}$$

En prendre valors absoluts dels errors s'evita que els errors positius es compensin amb els negatius.

Error absolut mitjà en percentatge (EAMP),

$$EAMP = \frac{\sum_{i=n+1}^{n+h} \frac{|\hat{y}_i - y_i|}{y_i}}{h} \times 100 \tag{4-68}$$

Arrel de l'error quadràtic mitjà (AECM)

Aquest estadístic es defineix com l'arrel quadrada de l'error quadràtic mitjà:

$$AECM = \sqrt{\frac{\sum_{i=n+1}^{n+h} \hat{y}_i - y_i^{\,2}}{h}} \tag{4-69}$$

Com es prenen els quadrats dels errors, s'evita la compensació entre errors positius i negatius. És important remarcar que l'*AECM* imposa una penalització major als errors de predicció que l'*EAM*.

Coeficient de desigualtat de Theil (U)

Aquest coeficient es defineix com segueix:

$$U = \frac{\sqrt{\frac{\sum_{i=n+1}^{n+h} \hat{y}_i - y_i^{\,2}}{h}}}{\sqrt{\frac{\sum_{i=n+1}^{n+h} \hat{y}_i^2}{h}} + \sqrt{\frac{\sum_{i=n+1}^{n+h} y_i^2}{h}}} \tag{4-70}$$

Com més xicotet és U més precises són les prediccions. L'escala d'U és tal que sempre està entre 0 i 1. Si U=0, y_i= $\hat{y}_i$ per a totes les prediccions; si U=1, la predicció és tan dolenta com puga ser. L'estadístic U de Theil pot ser reescalat i es pot descompondre en 3 proporcions: el biaix, la variança i la covariança. Per descomptat, la suma d'aquestes tres proporcions és 1. La interpretació d'aquestes tres proporcions és la següent:

1) El *biaix* reflecteix errors sistemàtics. Qualsevol que siga el valor d'*U*, esperaríem que el biaix estiga proper a 0. Un biaix gran suggereix que les prediccions estan sistemàticament per sobre, o per sota, dels valors reals.
2) La *variança* reflecteix també errors sistemàtics. La mida d'aquesta proporció és una indicació de la incapacitat de les prediccions per replicar la variabilitat de la variable a predir.
3) La *covariança* mesura errors de caràcter no sistemàtic. Idealment, aquesta hauria de ser la major proporció de la desigualtat de Theil.

A més del coeficient definit en (4-70), Theil va proposar altres coeficients per a l'avaluació de prediccions.

Predicció dinàmica

Siga el següent model:

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + u_t \tag{4-71}$$

Suposem que es volen fer prediccions per als períodes $i=n+1,\ldots,i=n+h$, i que es designa el valor real i previst en el període i com y_i i $\hat{y}_i$ respectivament. La predicció per al període $n+1$ és

$$\hat{y}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1} + \hat{\beta}_3 y_n \tag{4-72}$$

Com es pot observar, per a la predicció s'utilitza el valor observat de y (y_n) perquè està dins de la mostra utilitzada en l'estimació. Per a la predicció de la resta dels períodes utilitzem de forma recursiva la predicció del valor retardat de la variable dependent (predicció dinàmica), és a dir,

$$\hat{y}_{n+i} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+i} + \hat{\beta}_3 \hat{y}_{n-1+i} \qquad i = 2,3,\cdots,h \tag{4-73}$$

Així, des del període $n+2$ al $n+h$ la predicció realitzada per a un període s'utilitza per pronosticar la variable endògena en el període següent.

Exercicis

Exercici 4.1 Per explicar el preu de l'habitatge en una ciutat americana es formula el següent model:

$$price = \beta_1 + \beta_2 rooms + \beta_3 lowstat + \beta_4 crime + u$$

on *rooms* és el nombre d'habitacions de la casa, *lowstat* és el percentatge de persones de "classe marginal" a la zona i *crime* és el nombre de delictes per càpita a la zona. Els preus de les cases estan mesurats en dòlars.

Utilitzant les dades del fitxer *hprice2*, s'ha estimat el model anterior

$$price = \underset{(8022)}{-15694} + \underset{(1211)}{6788}\, rooms - \underset{(81)}{268}\, lowstat - \underset{(960)}{3854}\, crime$$

$$R^2=0.771 \qquad n=55$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Interpreteu el significat dels coeficients $\hat{\beta}_2$, $\hat{\beta}_3$ i $\hat{\beta}_4$.
b) Té el percentatge de persones de "classe marginal" una influència negativa sobre el preu de les cases en aquesta àrea?
c) Té el nombre d'habitacions una influència positiva sobre el preu de l'habitatge?

Exercici 4.2 Considere el següent model:

$$\ln(\mathit{fruit}) = \beta_1 + \beta_2 \ln(\mathit{inc}) + \beta_3 \mathit{hhsize} + \beta_4 \mathit{punder5} + u$$

on *fruit* és la despesa en fruites, *inc* és la renda disponible de la llar, *hhsize* és el nombre de membres de la llar i *punder5* és la proporció de nens menors de cinc anys a la llar.

Usant les dades del fitxer *demand*, s'ha estimat el model anterior:

$$\ln(\mathit{fruit}) = \underset{(3.701)}{-9.768} + \underset{(0.512)}{2.005}\ln(\mathit{inc}) - \underset{(0.179)}{1.205}\,\mathit{hhsize} - \underset{(1.302)}{1.795}\,\mathit{punder5}$$

$$R^2=0.728 \qquad n=40$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Interpreteu el significat dels coeficients $\hat{\beta}_2$, $\hat{\beta}_3$ i $\hat{\beta}_4$.
b) Té el nombre de membres de la llar un efecte estadísticament significatiu sobre la despesa en fruita?
c) És la proporció de nens menors de cinc anys a la llar un factor que té influència negativa en la despesa fruita?
d) És la fruita un bé de luxe?

Exercici 4.3 (Continuació de l'Exercici 2.5). Tenint en compte el model

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad \mathrm{i} = 1, 2, \ldots, n$$

s'han obtingut els següents resultats amb una mida mostral d'11 observacions:

$$\sum_{i=1}^{n} x_i = 0 \qquad \sum_{i=1}^{n} y_i = 0 \qquad \sum_{i=1}^{n} x_i^2 = B \qquad \sum_{i=1}^{n} y_i^2 = E \qquad \sum_{i=1}^{n} x_i y_i = F$$

(Recordeu que $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}$)

a) Construïu un estadístic per contrastar $H_0 : \beta_2 = 0$ contra $H_1 : \beta_2 \neq 0$
b) Contraste les hipòtesis de la qüestió a) tant $EB = 2F^2$.
c) Contraste les hipòtesis de la qüestió a) tant $EB = F^2$.

Exercici 4.4 S'ha formulat el següent model per explicar la despesa d'aliments (*alim*):

$$\mathit{alim} = \beta_1 + \beta_2 \mathit{renda} + \beta_3 \mathit{pralim} + u$$

on *renda* és la renda disponible i *pralim* és l'índex de preus relatius dels aliments pel que fa als altres productes de consum.

Prenent una mostra d'observacions corresponents a 20 anys successius s'obtenen els següents resultats:

$$\underset{}{alim_i} = \underset{(4.92)}{1.40} + \underset{(0.01)}{0.126}\, renta_i - \underset{(0.07)}{0.036}\, pralim_i$$

$$R^2=0.996; \qquad \sum \hat{u}_t^2 = 0.196$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Contraste la hipòtesi nul·la de que el coeficient de *pralim* és menor que 0.
b) Obtinga un interval de confiança del 95% per a la propensió marginal al consum d'aliments pel que fa a la renda.
c) Contraste la significativitat conjunta del model.

Exercici 4.5 S'han utilitzat mínims quadrats ordinaris per estimar la següent funció de demanda de lloguer d'habitatges:

$$\ln(dllog_i)=\beta_1+\beta_2\ln(pllog_i)+ \beta_3\ln(renda_i)+\varepsilon_i$$

on *dllog* és la despesa en lloguer d'habitatges, *pllog* és el preu de lloguer, i *renda* és la renda disponible.

Utilitzant una mostra de 403 observacions, s'obtenen els següents resultats:

$$\ln(dllog_i) = 10 - 0.7\ln\left(pllog_i\right) + 0.9\ln\left(renda_i\right)$$

amb R^2= 0.39 i la matriu estimada de covariances

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 1.0 & 0 & 0 \\ 0 & 0.09 & 0.085 \\ 0 & 0.085 & 0.09 \end{bmatrix}$$

a) Interpretació dels coeficients de ln(*dllog*) i ln(*pllog*).
b) Utilitzant un nivell de significació de 0,01, contraste la hipòtesi nul·la que $\beta_2=\beta_3=0$.
c) Contraste la hipòtesi nul·la que $\beta_2=0$, enfront de l'alternativa que $\beta_2<0$.
d) Contraste la hipòtesi nul·la que $\beta_3=1$, enfront de l'alternativa que $\beta_3 \neq 11$.
e) Contraste la hipòtesi nul·la que un augment en el preu de l'habitatge i un augment en la renda, en la mateixa proporció, no tenen cap efecte sobre la demanda d'habitatges.

Exercici 4.6 Utilitzant una mostra de 30 empreses s'han estimat els següents models corresponents a les funcions del cost mitjà (*cm*):

$$\underset{}{cm_i} = \underset{(11.97)}{172.46} + \underset{(3.70)}{35.72}\, quant_i$$

$$R^2 = 0.838 \qquad SCR = 8090 \qquad (1)$$

$$cm_i = \underset{(29.44)}{310.07} - \underset{(33.81)}{85.39}\, quant_i + \underset{(11.61)}{26.73}\, quant_i^2 - \underset{(1.22)}{1.40}\, quant_i^3$$
$$R^2 = 0.978 \quad SCR = 1097 \qquad (2)$$

on *cm* és el cost mitjà i *quant* és la quantitat produïda.

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Contraste si els termes quadràtic i cúbic de la quantitat produïda són significatius en la determinació el cost mitjà.
b) Contraste la significació global del model 2.

Exercici 4.7 Utilitzant una mostra de 35 observacions, s'han estimat els següents models per explicar la despesa en cafè

$$\ln(coffee) = 21.32 + \underset{(0.01)}{0.11}\ln(inc) - \underset{(0.23)}{1.33}\ln(cprice) + 1.35\ln(tprice) \qquad (1)$$
$$R^2 = 0.905 \qquad SCR = 254$$

$$\ln(coffee) = 19.9 + \underset{(0.02)}{0.14}\ln(inc) - \underset{(0.21)}{1.42}\ln(cprice) \qquad (2)$$
$$SCR = 529$$

on *inc* és la renda disponible, *cprice* és el preu del cafè i *tprice* és el preu del te.

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Contraste la significativitat global del model (1)
b) L'error estàndard de ln(*tprice*) no apareix en el model (1), ho pot calcular?
c) Contraste si el preu del te és estadísticament significatiu.
d) Com es contrastaria la hipòtesi que l'elasticitat del preu del cafè és igual però amb signe oposat al de l'elasticitat del preu del te? Detall el procediment.

Exercici 4.8 Ha estat formulat el següent model per analitzar els determinants de la qualitat de l'aire (*airqual*) en 30 àrees SMSA (Standard Metropolitan Statistical Areas) de Califòrnia:

$$airqual = \beta_1 + \beta_2\, popln + \beta_3 medincm + \beta_4\, poverty + \beta_5\, fueoil + \beta_6 valadd + u$$

on *airqual* és el pes en $\mu g/m^3$ de partícules en suspensió, *popln* és la població en milers, *medincm* és la renda mitjana per càpita en dòlars, *poverty* és el percentatge de famílies amb una renda inferior al nivell de pobresa, *fueloil* són milers de barrils de petroli consumit en indústries manufactureres, i *valadd* és el valor afegit de les indústries manufactureres en l'any 1972 en milers de dòlars.

Utilitzant les dades del fitxer *airqualy*, s'ha estimat el model anterior:

$$airqual_i = \underset{(10.19)}{97.35} + \underset{(0.0311)}{0.0956}\, popln_i - \underset{(0.0055)}{0.0170}\, medincm_i - \underset{(0.0089)}{0.0254}\, poverty_i$$
$$- \underset{(0.0017)}{0.0031}\, fueoil_i - \underset{(0.0025)}{0.0011} valadd_i$$

$$R^2=0.415 \quad n=30$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Interpretació dels coeficients de *medincm*, *poverty* i *valadd*

b) Són els coeficients de pendent significatius al 10% de forma individual?

c) Contrast de significació conjunta de *fuel* i *valadd*, sabent que

$$airqual_i = \underset{(10.41)}{97.67} + \underset{(0.020)}{0.0566}\, popln_i - \underset{(0.0039)}{0.0102}\, medincm_i - \underset{(0.0078)}{0.0174}\, poverty_i$$

$$R^2 = 0.339 \quad n = 30$$

d) Si s'omet en el primer model la variable de la pobresa s'obtenen els següents resultats:

$$airqual_i = \underset{(10.02)}{82.98_i} + \underset{(0.031)}{0.0523}\, popln_i - \underset{(0.0055)}{0.0097}\, medincm_i$$

$$- \underset{(0.0017)}{0.00063}\, fueoil_i - \underset{(0.0028)}{0.00037}\, valadd_i$$

$$R^2 = 0.218 \quad n = 30$$

Són els coeficients de pendent significatius al 10% de forma individual en el nou model? Considera que aquests resultats són raonables en comparació amb els obtinguts en l'apartat b)? Comparant R^2 en els dos models estimats, quin és el paper exercit per la pobresa en la determinació de la qualitat de l'aire?

e) Si es fa una regressió per explicar *airqual* usant com regressors únicament el terme independent i *poverty*, s'obté que R^2=0.037. Considera raonable aquest valor tenint en compte els resultats obtinguts en l'apartat d)?

Exercici 4.9 S'han estimat pel mètode de *MQO* amb una mostra de 39 observacions les següents funcions de producció:

$$output_t = \hat{\alpha} labor_t^{1.30} capital_t^{0.32} \exp(0.0055 trend_t) \qquad R^2 = 0.9945$$

$$output_t = \hat{\beta} labor_t^{1.41} capital_t^{0.47} \qquad R^2 = 0.9937$$

$$output_i = \hat{\gamma} \exp(0.0055 trend_t) \qquad R^2 = 0.9549$$

on *trend* és una variable de tendència.

a) Contraste la significativitat conjunta de *labor* i *capital*.

b) Contraste la significativitat del coeficient de la variable *trend*.

c) Indiqueu els supòsits estadístics sota les quals els contrastos realitzats en els dos apartats anteriors són correctes. Una qüestió addicional: Escriu el model poblacional corresponent a la primera de les tres especificacions anteriors.

Exercici 4.10 En una investigació s'ha formulat el següent model:

$$y = \beta_1 + \beta_2\ x_2 + \beta_3 x_3 + u$$

Amb una mostra de 43 observacions s'han obtingut els següents resultats:

$$\hat{y}_i = \text{-}0.06 + 1.44\ x_{2i} - 0.48\ x_{3i}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.1011 & -0.0007 & -0.0005 \\ & 0.0231 & -0.0162 \\ & & 0.0122 \end{bmatrix} \quad \sum y_i^2 = 444$$

$$\sum \hat{y}_i^2 = 424.92$$

a) Contraste que el terme independent és menor que 0.
b) Contraste que β_2=2.
c) Contraste de la hipòtesi nul·la $\beta_2+3\beta_3=0$.

Exercici 4.11 Donada la funció de producció

$$q = ak^{\alpha}l^{\beta}\exp(u)$$

s'ha procedit a la seva estimació amb dades de l'economia espanyola dels últims 20 anys, obtenint-se els següents resultats:

$$\ln(q_i) = 0.15 + 0.73\ln(k_i) + 0.47\ln(l_i)$$

$$\left[\mathbf{X}'\mathbf{X}\right]^{-1} = \begin{bmatrix} 4129 & -95 & -266 \\ -95 & 3 & 5 \\ -266 & 5 & 19 \end{bmatrix} \qquad SCR = 0.017$$

a) Contraste la significativitat individual dels coeficients de *k* i *l*.
b) Contraste si el paràmetre α és significativament diferent d'1.
c) Contraste si hi ha rendiments creixents a escala.

Exercici 4.12 Siga el següent model de regressió múltiple:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + u$$

Amb una mostra de 33 observacions s'ha estimat aquest model per MQO, obtenint-se els següents resultats:

$$\hat{y}_i = 12.7 + 14.2x_{1i} + 2.1x_{2i}$$

$$\hat{\sigma}^2\left[\mathbf{X}'\mathbf{X}\right]^{-1} = \begin{bmatrix} 4.1 & -0.95 & -0.266 \\ -0.95 & 3.8 & 0.5 \\ -0.266 & 0.5 & 1.9 \end{bmatrix}$$

a) Contraste la hipòtesi nul·la $\alpha_0 = \alpha_1$.
b) Contraste si $\alpha_1 / \alpha_2 = 7$.

Són significatius individualment els coeficients α_0, α_1, i α_2?

Exercici 4.13 Amb una mostra de 30 empreses s'han estimat les següents funcions de cost:

$$a)\ cost_i = \underset{(11.97)}{172.46} + \underset{(3.70)}{35.72}\, x_i \qquad R^2 = 0.838 \quad \bar{R}^2 = 0.829 \quad SCR = 8090$$

$$b)\ cost_i = \underset{(29.44)}{310.07} - \underset{(33.81)}{85.39}\, x_i + \underset{(11.61)}{26.73}\, x_i^2 - \underset{(1.22)}{1.40}\, x_i^3 \quad R^2 = 0.978 \quad \bar{R}^2 = 0.974 \quad SCR = 1097$$

on *cost* és el cost mitjà i *x* és la quantitat produïda

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Quin dels dos models triaria? En base a quin criteri?

b) Contraste si els termes quadràtic i cúbic de la quantitat produïda influeixen significativament en la determinació del cost mitjà.

c) Contraste la significativitat del model b

Exercici 4.14 Un investigador formula el següent model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

Utilitzant una mostra de 13 observacions obté els següents resultats:

$$\begin{aligned} &\hat{y}_i = 1.00 - 1.82 x_{2i} + 0.36 x_{3i} \\ &R^2 = 0.50 \quad n = 13 \end{aligned} \qquad (1)$$

$$\text{var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 0.25 & -0.01 & 0.04 \\ -0.01 & 0.16 & -0.15 \\ 0.04 & -0.15 & 0.81 \end{bmatrix}$$

Utilitzant la informació disponible:

a) Contraste la hipòtesi nul·la de que $\beta_2 = 0$ enfront de la hipòtesi alternativa que. $\beta_2 < 0$.

b) Contraste la hipòtesi nul·la de que $\beta_2 + \beta_3 = -1$ enfront de la hipòtesi alternativa que. $\beta_2 + \beta_3 \neq -1$, amb un nivell de significació del 5%.

c) El model en el seu conjunt és significatiu?

d) Suposant que les variables a (1) estan mesurades en logaritmes naturals, ¿quina és la interpretació del coeficient corresponent a x_3?

Exercici 4.15 Amb una mostra de 50 empreses del sector de l'automòbil s'han estimat les següents funcions de producció, utilitzant com a variable endògena el valor afegit brut de la producció d'automòbils (*vab*) i com a variables explicatives el factor treball (*labor*) i el factor capital (*capital*).

$$1)\ \begin{aligned} &\ln(vab_i) = 3.87 + \underset{(0.11)}{0.80} \ln(labor_i) + \underset{(0.24)}{1.24} \ln(capital_i) \\ &SCR = 254 \quad R^2 = 0.75 \quad \bar{R}^2 = 0.72 \end{aligned},$$

2) $$\ln(vab_i) = 19.9 + 1.04\ln(capital_i)$$
$$SCR = 529 \quad R^2 = 0.84, \bar{R}^2 = 0.81$$

3) $$\ln(vab/labor_i) = 15.2 + 0.87\ln(capital_i/labor_i)$$
$$SCR = 380$$

(Entre parèntesi apareixen els errors estàndard dels estimadors).

a) Contraste la significativitat conjunta dels dos factors de la funció de producció.
b) Contraste si el factor treball té una influència significativament positiva sobre el valor afegit de la producció d'automòbils.
c) Contraste la hipòtesi de rendiments constants a escala. Raoneu la resposta.

Exercici 4.16 Amb una mostra de 35 observacions anuals s'han estimat dues funcions de demanda de vi de Rioja, utilitzant com a variable endògena la despesa en vi de reserva (*vi*) i com a variables explicatives la renda disponible (*renda*), el preu mitjà d'una ampolla de vi de Rioja de reserva (*pvirioj*) i el preu mitjà d'una ampolla de vi de Ribera de Duero de reserva (*pvinduer*). Els resultats són els següents:

$$\ln(vi_i) = 21.32 + \underset{(0.01)}{0.11}\ln(renda_i) - \underset{(0.23)}{1.33}\ln(pvirioj_i) + \underset{(0.233)}{1.35}\ln(pviduer_i)$$
$$R^2 = 0.905 \quad SCR = 254$$

$$\ln(vi_i) = 19.9 + \underset{(0.02)}{0.14}\ln(renda_i) - \underset{(0.21)}{1.42}\ln(pvirioj_i)$$
$$SCR = 529$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Contraste la significativitat global del primer model.
b) Contraste si el preu del vi de Ribera de Duero té una influència significativa, aplicant dues estadístiques que no utilitzen la mateixa informació. Mostri que tots dos procediments són equivalents
c) Com contrastaria la hipòtesi que l'elasticitat preu del vi de Rioja és igual però amb signe contrari a l'elasticitat preu del vi de Ribera de Duero? Detall el procediment que seguiria.

Exercici 4.17 Per analitzar la demanda de te de Ceilan (*teceil*) s'ha formulat el següent model economètric:

$$\ln(teceil) = \beta_1 + \beta_2\ln(renta) + \beta_3\ln(pteceil) + \beta_4\ln(pteind) + \beta_5\ln(pcabras) + u$$

on *renda* és la renda disponible, *pteceil* el preu del te de Ceilan, *pteind* és el preu del te de l'Índia i *pcabras* és el preu del cafè del Brasil.

Amb una mostra de 22 observacions s'han realitzat les següents estimacions:

$$\ln(teceil_i) = 2.83 + \underset{(0.17)}{0.25}\ln(renda_i) - \underset{(0.98)}{1.48}\ln(pteceil_i) + \underset{(0.69)}{1.18}\ln(pteind_i) + \underset{(0.15)}{0.20}\ln(pcafbras_i)$$

SQR=0.4277

$$\ln(teceil_i \times pteceil) = 0.74 + \underset{(0.16)}{0.26}\ln(renda_i) + \underset{(0.15)}{0.20}\ln(pcafbras_i)$$

SQR=0.6788

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Contraste la significativitat de la renda disponible.
b) Contraste la hipòtesi $\beta_3 = -1$ i $\beta_4 = 0$, explicant el procediment aplicat.
c) Si en lloc de disposar d'informació sobre la *SQR*, només conegués el R^2 de cada model, com procediria per realitzar el contrast de la part b)?

Exercici 4.18 Amb una mostra de 64 països s'han obtingut les següents estimacions per explicar les defuncions de menors de 5 anys per 1000 habitants nascuts vius (*defmen5*)

1) $defmen5_i = 263.64 - \underset{(0.0019)}{0.0056}renpc_i + \underset{(0.21)}{2.23}tanaldon_i$; $\quad R^2 = 0.7077$

2) $defmen5_i = 168.31 - \underset{(0.0018)}{0.0055}renpc_i + \underset{(0.25)}{1.76}tanaldon_i + 12.87tfec_i$, $\quad R^2 = 0.7474$

on *renpc* és la renda per càpita, *tanaldon* és la taxa d'analfabetisme de les dones i *tfec* és la taxa de fecunditat (TFI).

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Contraste la significativitat conjunta de renda, taxa d'analfabetisme i taxa de fecunditat.
b) Contraste la significativitat de la taxa de fecunditat.
c) Quin dels dos models triaria? Raoneu la resposta.

Exercici 4.19 S'ha estimat la següent funció de vendes d'automòbils d'una determinada marca utilitzant una mostra de 32 observacions anuals:

$$\hat{v}_i = \underset{(6.48)}{104.8} - \underset{(3.19)}{6.64}\,p_i + \underset{(0.16)}{2.98}\,d_i$$

$$\sum \hat{u}_i^2 = 1805.2; \qquad \sum (v_i - \bar{v})^2 = 13581.4$$

on *v* són vendes, *p* és el preu dels automòbils i *d* són despeses en publicitat.

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Són significatius conjuntament el preu i les despeses en publicitat? Raoneu la resposta.
b) És admissible la hipòtesi que els preus tinguen una influència negativa sobre les vendes? Raoneu la resposta.
c) Descriviu detalladament com contrastaria la hipòtesi que l'impacte de les despeses en publicitat sobre les vendes és més gran que menys 0.4 vegades l'impacte del preu.

Exercici 4.20 En un estudi sobre els costos de producció (cp_i) de 62 mines de carbó s'elabora el següent model:

$$\widehat{cp_i} = \underset{(3.4)}{2.20} - \underset{(0.005)}{0.104} gm_i + \underset{(2.2)}{3.48} dg_i + \underset{(0.15)}{0.104} pa_i$$

on gm_i és el grau de mecanització, dg_i és una mesura de dificultats geològiques i pa_i és el percentatge d'absentisme.

(Els números entre parèntesis són els errors estàndard dels estimadors.)

Es disposa a més de la següent informació:

$$\sum \left[cp_i - \overline{cp} \right]^2 = 109.6 \qquad \sum \hat{u}_i^2 = 18.48$$

a) Contraste la significativitat de cada un dels coeficients del model.
b) Contraste la significativitat global del model.

Exercici 4.21 Amb quinze observacions s'ha obtingut la següent estimació:

$$\hat{y}_i = 8.04 - \underset{(1.00)}{2.46} x_{i2} + \underset{(0.60)}{0.23} x_{i3}$$
$$\bar{R}^2 = 0.30$$

on els valors entre parèntesis són els errors estàndard dels estimadors i el coeficient de determinació és el corregit.

a) És significatiu el coeficient de la variable x_3?
b) Es rebutja la hipòtesi que un augment en x_2 de dues unitats ocasiona una disminució en y de vuit unitats?
c) Analitzeu la significativitat conjunta del model.

Exercici 4.22 Considere la següent especificació economètrica:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

Amb una mostra de 26 observacions s'han obtingut les dues següents estimacions:

1) $\hat{y}_i = 2 + \underset{(1.9)}{3.5} x_{1i} - \underset{(2.2)}{0.7} x_{2i} - \underset{(1.5)}{2} x_{3i} + u_i$ $\qquad R^2$=0.982

2) $\hat{y}_i = 1.5 + \underset{(2.7)}{3} (x_{1i} + x_{2i}) - \underset{(2.4)}{0.6} x_{3i} + u_i$ $\qquad R^2$= 0.876

(Entre parèntesi figuren els estadístics t)

a) Demostreu que les següents expressions de l'estadístic F són equivalents:

$$F = \frac{\left(SCR_R - SCR_{NR} \right) / r}{SCR_{NR} / (n-k)} \qquad F = \frac{\left(R_{NR}^2 - R_R^2 \right) / q}{(1 - R_{NR}^2) / (n-k)}$$

b) Contraste la hipòtesi nul·la $\beta_2 = \beta_3$.

Exercici 4.23 En l'estimació del model de Brown, realitzada en l'Exercici 3.19 utilitzant el fitxer *consumsp*, es va obtenir el següent resultat:

$$\underset{}{conspc_t} = \underset{(84.88)}{-7.156} + \underset{(0.0857)}{0.3965}\, incpc_t + \underset{(0.0903)}{0.5771}\, conspc_{t-1}$$

$$R^2=0.997; \qquad SQR=1891320; \qquad n=56$$

També es realitza l'estimació dels següents models:

$$conspc_t - conspc_{t-1} = \underset{(84.43)}{-98.13} + \underset{(0.0803)}{0.2757}(incpc_t - conspc_{t-1})$$

$$R^2=0.1792; \qquad SQR=2199474; \qquad n=56$$

$$conspc_t - incpc_{t-1} = \underset{(84.88)}{-7.156} - \underset{(0.0090)}{0.0264}\, incpc + \underset{(0.0903)}{0.5771}(conspc_{t-1} - incpc_t)$$

$$R^2=0.6570; \qquad SQR=1891320; \qquad n=56$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Contraste la significativitat de cada un dels coeficients del primer model.

b) Contraste que el coeficient d'*incpc* en el primer model és menor que 0.5.

c) Contraste la significativitat global del primer model.

d) És admissible la hipòtesi $\beta_2 + \beta_3 = 1$?

e) Demostre que operant en el tercer model pot arribar als mateixos coeficients que en el primer model.

Exercici 4.24 El següent model va ser formulat per analitzar els determinants del salari mitjà en dòlars que s'obté en graduar-se en la classe del 2010 en les millors escoles de negocis dels EUA (*SalMBAgr*):

$$salMBAgr = \beta_1 + \beta_2 tuition + \beta_3 salMBApr + u$$

on *tuition* són els drets de matrícula, incloent a més tots els altres honoraris per al programa complet, però amb exclusió de les despeses de manutenció, i *salMBApr* és el salari mitjà anual en dòlars obtingut prèviament per la classe de 2010.

Utilitzant les dades del fitxer *MBAtui10*, s'ha estimat el model anterior:

$$\underset{}{salMBAgr_i} = \underset{(5415)}{42489} + \underset{(0.0628)}{0.1881}\, tuition_i + \underset{(0.1015)}{0.5992}\, salMBApr_i$$

$$R^2=0.703 \qquad n=39$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Quins dels regressors inclosos en el model anterior són individualment significatius a l'1% i el 5%?

b) Contraste la significació global del model.

c) Quin és el valor predit de *salMBAgr* per a un estudiant de postgrau que va pagar 100000 dòlars de drets de matrícula en un màster MBA de dos anys i que anteriorment va tenir un salari de 70000 dòlars anuals? Quants anys

de treball necessita l'estudiant per compensar les despeses de matrícula? Per respondre a aquestes preguntes suposi que la taxa de descompte és igual a la taxa esperada d'augment salarial i que l'estudiant no va rebre cap ingrés salarial durant els 2 cursos de durada del màster.

d) Si hi afegim el regressor *rank2010* (el rang de cada escola de negocis el 2010), s'obtenen els següents resultats:

$$\underset{}{salMBAgr_i} = \underset{(8520)}{61320} + \underset{(0.0626)}{0.1229}\, tuition_i + \underset{(0.1055)}{0.4662}\, salMBApr_i - \underset{(85.13)}{232.06}\, rank2010_i$$

$$R^2 = 0.755 \quad n = 39$$

Quin dels regressors inclosos en aquest model són individualment significatius al 5%?

Quina és la interpretació del coeficient de *rank2010*?

e) La variable *rank2010* s'ha construït a partir de tres components: *gradpoll* és una classificació basada en enquestes als graduats en MBA i contribueix amb un 45 per cent a la classificació final; *corppoll* és una classificació basada en enquestes realitzades a reclutadors dels MBA i contribueix amb el 45 per cent a la classificació final; i *intellec* és una classificació basada en la revisió de la recerca universitària publicada durant un període de cinc anys a les 20 principals revistes acadèmiques, i en els llibres universitaris revisats per *The New York Times*, *The Wall Street Journal* i *Bloomberg Businessweek* en el mateix període; aquesta última classificació aporta el 10 per cent a la classificació final. En el següent model estimat *rank2010* ha estat substituït pels seus tres components:

$$\underset{}{salMBAgr_i} = \underset{(10700)}{79904} + \underset{(0.0696)}{0.0305}\, tuition_i + \underset{(0.107)}{0.3751}\, salMBApr_i - \underset{(94.54)}{303.82}\, gradpoll_i - \underset{(61.26)}{33.829}\, corppoll_i - \underset{(64.09)}{113.36}\, intellec_i$$

$$R^2 = 0.797 \quad n = 39$$

Quin és el pes en percentatge de cada un d'aquestes tres components en la determinació de *salMBAgr*? Compare els resultats amb la contribució de cadascuna en la definició de *rank2010*.

f) Són *gradpoll*, *corppoll* i *intellec* conjuntament significatius al 5%? Són individualment significatius al 5%?

Exercici 4.25 (Continuació de l'Exercici 3.12). El model poblacional que correspon a aquest exercici és el següent:

$$\ln(wage) = \beta_1 + \beta_2 educ + \beta_3 tenure + \beta_4 age + u$$

Utilitzant el fitxer *wage06s*, s'ha estimat el model anterior:

$$\ln(wage)_i = \underset{(0.073)}{1.565} + \underset{(0.0035)}{0.0448}\, educ_i + \underset{(0.0019)}{0.0177}\, tenure_i + \underset{(0.0016)}{0.0065}\, age_i$$

$$R^2 = 0.337 \quad n = 800$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Contraste la significativitat global del model.
b) És *tenure* estadísticament significativa al 10%? És *age* positivament significativa al 10%?
c) És admissible que el coeficient d'*educ* siga igual al de *tenure*? És admissible que el coeficient d'*educ* siga el triple del de tenure? Per respondre a aquestes preguntes disposa de la següent informació addicional:

$$\ln(wage)_i = \underset{(0.073)}{1.565} + \underset{(0.0042)}{0.0271}educ_i + \underset{(0.0019)}{0.0177}(educ + tenure)_i + \underset{(0.0016)}{0.0065}age_i$$

$$\ln(wage)_i = \underset{(0.073)}{1.565} - \underset{(0.0071)}{0.0082}educ_i + \underset{(0.0019)}{0.0177}(3\times educ + tenure)_i + \underset{(0.0016)}{0.0065}age_i$$

d) Es pot calcular el R2 en les dues equacions de l'apartat c)? Si us plau, si la resposta és positiva, feu-ho.

Exercici 4.26 (Continuació de l'Exercici 3.13). Prenguem el model poblacional d'aquest Exercici com a model de referència. En el model estimat, amb el fitxer *housecan*, els errors estàndard dels coeficients apareixen entre parèntesis:

$$price_i = \underset{(3379)}{-2418} + \underset{(1207)}{5827}bedrooms_i + \underset{(1785)}{19750}bathrms_i + \underset{(0.388)}{5.411}lotsize_i$$

$$R^2=0.486 \quad n=546$$

a) Contraste la significativitat global del model.
b) Contraste la hipòtesi nul·la que un bany addicional té la mateixa influència sobre el preu de l'habitatge que 4 dormitoris addicionals. D'altra banda, contrast si una cambra de bany addicional té més influència sobre el preu de l'habitatge que 4 dormitoris addicionals. (Informació addicional: $\text{var}(\hat{\beta}_2)$ =1455813, $\text{var}(\hat{\beta}_3)$ =3186523 i $\text{var}(\hat{\beta}_2, \hat{\beta}_3)$ =-764846).
c) Si afegim al model el regressor *stories* (nombre de plantes, exclòs el soterrani), s'obtenen els següents resultats:

$$price_i = \underset{(3603)}{-4010} + \underset{(1215)}{2825}bedrooms_i + \underset{(1734)}{17105}bathrms_i + \underset{(0.369)}{5.429}lotsize_i + \underset{(1008)}{7635}stories_i$$

$$R^2=0.536 \quad n=546$$

Quina és la seva opinió sobre el signe i magnitud del coeficient de *stories*? És un resultat sorprenent? Quina és la interpretació d'aquest coeficient? Estimeu si el nombre de *stories* té una influència significativa sobre el preu de l'habitatge.
d) Repetiu els contrastos de l'apartat b) amb el model estimat en l'apartat (Informació addicional: $\text{var}(\hat{\beta}_2)$ =1475758, $\text{var}(\hat{\beta}_3)$ =3008262 i $\text{var}(\hat{\beta}_2, \hat{\beta}_3)$ =-554381).

Exercici 4.27 (Continuació de l'Exercici 3.14). Prenguem el model poblacional d'aquest Exercici com a model de referència. Usant el fitxer *ceoforbes*, el model estimat és el següent:

$$\ln(salary)_i = \underset{(0.377)}{4.641} + \underset{(0.0033)}{0.0054}roa_i + \underset{(0.0425)}{0.2893}\ln(sales_i) + \underset{(0.0000220)}{0.0000564}profits_i + \underset{(0.0032)}{0.0122}tenure_i$$

R^2=0.232 n=447

(Els números entre parèntesi són els errors estàndard dels estimadors.)

a) Té *roa* un efecte significatiu sobre el salari? Té *roa* un efecte positivament significatiu sobre el salari? Feu tots dos contrastos per al 10% i el 5% de nivell de significació.

b) En el cas que *roa* s'incrementés en 20 punts, en quin percentatge s'incrementaria el salari?

c) Contraste la hipòtesi nul·la de que l'elasticitat salari/vendes és igual a 0.4.

d) Si a aquest model afegim el regressor *age*, s'obtenen els següents resultats:

$$\ln(salary)_i = \underset{(0.442)}{4.159} + \underset{(0.0033)}{0.0055}\, roa_i + \underset{(0.0423)}{0.2903}\ln(sales_i) + \underset{(0.0000220)}{0.0000539}\, profits + \underset{(0.0035)}{0.00924}\, tenure_i + \underset{(0.0043)}{0.00880}\, age_i \qquad R^2=0.240$$

n=447

Són els coeficients estimats molt diferents dels obtinguts en l'estimació del model de referència? Què passa amb el coeficient de *tenure*? Explique-.

e) Té l'edat (*age*) del director executiu un efecte significatiu sobre el salari?

f) És admissible que el coeficient d'*age* siga igual al coeficient de *tenure*?

(Informació addicional: $\text{var}(\hat{\beta}_5)$ =1.24E-05; $\text{var}(\hat{\beta}_6)$ =1.82E-05 i $\text{var}(\hat{\beta}_5, \hat{\beta}_6)$ =-6.09E-06).

Exercici 4.28 (Continuació de l'Exercici 3.15). Prenguem el model poblacional d'aquest Exercici com a model de referència. Utilitzant el fitxer *rdspain*, el model estimat va ser el següent:

$$rdintens_i = \underset{(0.428)}{-1.8168} + \underset{(0.0278)}{0.1482}\ln(sales_i) + \underset{(0.0021)}{0.0110}\, exponsal_i$$

R^2=0.048 n=1983

a) És la variable de vendes (*sales*) individualment significativa a l'1%?

b) Contraste la hipòtesi nul·la de que el coeficient de les vendes és igual a 0.2?

c) Contraste la significativitat global del model de referència.

d) Si hi afegim el regressor ln(*workers*), - on *workers* és el nombre de treballadors-, s'obtenen els següents resultats:

$$rdintens = \underset{(0.750)}{0.480} - \underset{(0.0687)}{0.08585}\ln(sales) + \underset{(0.0021)}{0.01049}\, exponsal + \underset{(0.09198)}{0.3422}\ln(workers)$$

R2=0.055 n=1983

És la variable vendes individualment significativa a l'1% en el nou model estimat?

e) Contraste la hipòtesi nul·la que el coeficient de ln(*workers*) és més gran que 0.5.

Exercici 4.29 (Continuació de l'Exercici 3.16). Prenguem el model poblacional d'aquest Exercici com a model de referència. Utilitzant el fitxer *hedcarsp*, el model estimat va ser el següent:

$$\ln(price)_i = \underset{(0.154)}{14.42} + \underset{(0.0000438)}{0.000581}\, cid_i + \underset{(0.0079)}{0.003823}\, hpweight_i - \underset{(0.0122)}{0.07854}\, fueleff_i$$

$$R^2=0.830 \quad n=214$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Quina de les variables explicatives incloses en el model de referència són individualment significatives a l'1%?
b) Afegiu la variable *volume* (volum) al model de referència. Té *volume* un efecte estadísticament significatiu sobre ln(*price*)? Té el volum un efecte positiu estadísticament significatiu sobre el ln(*price*)?
c) És admissible que el coeficient estimat de *volume* en l'apartat b) siga igual però amb signe oposat que el coeficient de *fueloff*?
d) Afegiu les variables *length*, *width* i *height* (longitud, amplada i alçada) al model estimat en l'apartat b). Tenint en compte que *volume* = *length*× *width*×*height*, hi ha multicolinealitat perfecta al nou model? Per què? Per què no? Estimeu el nou model si és possible.
e) Afegiu la variable ln(*volume*) al model de referència. Contraste la hipòtesi nul·la per la qual l'elasticitat de preu/volum és igual a 1?
f) Què passaria si s'afegeix al model estimat en l'apartat e) els regressors ln(*length*), ln(*width*) i ln(*height*)?

Exercici 4.30 (Continuació de l'Exercici 3.17). Prenguem el model poblacional d'aquest Exercici com a model de referència. Utilitzant el fitxer *timuse03*, el corresponent model ajustat va ser el següent:

$$housework_i = \underset{(23.27)}{141.9} + \underset{(1.621)}{3.850}\, educ_i - \underset{(0.00539)}{0.00917}\, hhinc_i + \underset{(0.311)}{1.767}\, age_i - \underset{(0.0229)}{0.2289}\, paidwork_i$$

$$R^2=0.1440 \quad n=1000$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Quina de les variables explicatives incloses en el model de referència són individualment significatives al 5% i l'1%?
b) Estimeu un model en el qual es puga contrastar directament si un any més d'educació té el mateix efecte sobre el temps dedicat al treball domèstic que 2 anys d'edat addicionals. Quina és la seva conclusió?
c) Contraste la significació conjunta d'*educ* i *hhnc*.
d) Estimeu una regressió en la qual s'afegeix la variable *childup3* (nombre de nens de fins a 3 anys) al model de referència. En el nou model, ¿quins dels regressors són individualment significatius al 5% i l'1%?
e) En el model formulat en l'apartat d), quina és la variable més influent? Per què?

Exercici 4.31 (Continuació de l'Exercici 3.18). Prenguem el model poblacional d'aquest Exercici com a model de referència. Utilitzant el fitxer *hdr2010*, el corresponent model ajustat va ser el següent:

$$stsfglo_i = \underset{(0.584)}{-0.375} + \underset{(0.00000617)}{0.0000207}\, gnipc_i + \underset{(0.009)}{0.0858}\, lifexpec_i$$

$$R^2=0.642 \quad n=144$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Quina de les variables explicatives incloses en el model de referència són individualment significatives a l'1%?
b) Estimeu un model en el qual s'han afegit les variables *popnosan* (percentatge de població sense accés a serveis de sanejament millorat) i *gnirank* (rang de la renda nacional bruta) al model de referència. Quin dels regressors inclosos en el nou model són individualment significatius a l'1%? Intèrpret els coeficients de *popnosan* i *gnirank*.
c) Són *popnosan* i *gnirank* conjuntament significatius?
d) Contraste la significativitat global del model formulat en l'apartat b).

Exercici 4.32 Amb una mostra de 42 observacions s'ha estimat el següent model

$$\hat{y}_t = -670.591 + 1.008x_t$$

Per a l'observació 43 se sap que el valor de x és 1571.9.

a) Calculeu el predictor puntual per a l'observació *43.*
b) Sabent que la variança de l'error de predicció $\hat{e}_2^{43} = y^{43} - \hat{y}^{43}$ és igual a $(24.9048)^2$, calculi un interval de probabilitat del 90% del valor individual.

Exercici 4.33 Sobre la funció de consum Brown, a més de l'estimació presentada l'Exercici 4.23, es disposa de la següent estimació:

$$\underset{}{conspc_t} = \underset{(64.35)}{12729} + \underset{(0.0857)}{0.3965}(incpc_t - 13500) + \underset{(0.0903)}{0.5771}(conspc_{t-1} - 12793.6)$$

$$R^2 = 0.997; \qquad SQR = 1891320; \qquad n = 56$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Obtinga el predictor puntual del consum per càpita el 2011, sabent que *conspc2010*=12793.6 i *incpc2011*=13500.
b) Obtinga un interval de confiança del 95% per al valor esperat del consum per càpita el 2011.
c) Obtinga un interval de predicció del 95% per al valor individual del consum per càpita el 2011.

Exercici 4.34 (Continuació de l'Exercici 4.30) Contesteu les següents preguntes:

a) Utilitzant la primera estimació de l'Exercici 4.30, obtenir una predicció per *houswork* (minuts dedicats a les tasques domèstiques per dia), quan en l'equació es fa *educ*=10 (anys), *hhinc*=1200 (euros al mes), *age*=50 (anys) i *paidwork*= 400 (minuts per dia).
b) Feu una regressió, utilitzant el fitxer *timuse03*, que li permeta calcular un *IC* del 95% amb les característiques tingudes en compte en l'apartat a).
c) Obtinga un interval de predicció del 95% per al valor individual de *houswork* amb les característiques tingudes en compte en l'apartat a).

Exercici 4.35 (Continuació de l'Exercici 4.29) Contesteu les següents preguntes:

a) En la primera equació de l'Exercici 4.29 considere que *cid*=2000 (polzades cúbiques de desplaçament), *hpweight*=10 (relació potència/pes en kg, expressats en percentatge) i *fueleff*=6 (minuts per dia). Obtinga el predictor

puntual del consum per càpita el 2011, sabent que *incpc2011*=12793.6 i *conspc2010*=13500.

b) Obtinga una estimació consistent de *price* per les característiques utilitzades en l'apartat a).

c) Feu una regressió que li permeta calcular un *IC* del 95% per a les característiques utilitzades en l'apartat a).

d) Obtinga un interval de predicció del 95% per al valor individual del preu.

5 ANÀLISI DE REGRESSIÓ MÚLTIPLE AMB INFORMACIÓ QUALITATIVA

5.1 Introducció d'informació qualitativa en els models economètrics

Fins ara les variables que hem utilitzat per explicar la variable endògena tenien un caràcter quantitatiu. No obstant això, hi ha altres variables de caràcter qualitatiu que poden ser importants per explicar el comportament de la variable endògena, com el sexe, raça, religió, nacionalitat, regió geogràfica, etc. Per exemple, mantenint tots els altres factors constants, s'ha constatat que les dones treballadores tenen uns salaris inferiors que els seus homòlegs masculins. Aquest resultat pot ser conseqüència de la discriminació per gènere, però qualsevol que siga la raó, les variables qualitatives com el gènere sembla que influeixen en la variable endògena i s'haurien d'incloure en molts casos entre les variables explicatives. Els factors qualitatius, sovint, però no sempre, es presenten en forma d'informació binària, és a dir, una persona és home o dona, està casada o no, etc. Quan els factors qualitatius es presenten en forma dicotòmica la informació rellevant pot mostrar-se com una variable binària o una variable de zero-un. En econometria, les variables binàries que s'utilitzen com regressors són comunament anomenades variables fictícies. En la definició d'una variable dicotòmica, hem de decidir a quin cas se li assigna el valor 1 i a qual se li assigna el valor 0.

En el cas del gènere podem definir

$$mujer = \begin{cases} 1 & \text{si la persona es una dona} \\ 0 & \text{si la persona es un home} \end{cases}$$

Però, per descomptat, també podem definir

$$hombre = \begin{cases} 1 & \text{si la persona es un home} \\ 0 & \text{si la persona es una dona} \end{cases}$$

És important assenyalar que les dues variables, dona i home, contenen la mateixa informació. Utilitzar les variables zero-un per captar informació qualitativa és una decisió arbitrària, però amb aquesta elecció els paràmetres tenen una interpretació natural.

5.2 Una sola variable fictícia independent.

Analitzarem com es pot incorporar la informació dicotòmica en els models de regressió. Considere el model per a la determinació del salari per hora, en funció dels anys d'educació (*educ*):

$$salari = \beta_1 + \beta_2 educ + u \tag{5-1}$$

Per mesurar la discriminació salarial deguda al gènere s'introdueix una variable fictícia (*dona*) com a variable independent en el model definit anteriorment,

$$salari = \beta_1 + \delta_1 dona + \beta_2 educ + u \tag{5-2}$$

L'*atribut* gènere té dues *categories*: *dona* i *home*. La categoria *dona* ha estat inclosa en el model; mentre que la categoria *home*, que ha estat omesa, és la categoria de referència. El model (5-2) es mostra a la figura 5.1, prenent $\delta_1<0$. La interpretació de δ_1 és la següent: δ_1 és la diferència en el salari per hora entre dones i homes, donat el mateix nivell d'educació (i el mateix terme de pertorbació, *u*). Així, el coeficient δ_1 determina si hi ha una discriminació contra les dones o no. Si $\delta_1<0$, aleshores, per al mateix nivell d'altres factors (educació, en aquest cas), les dones guanyen menys que els homes de mitjana. Suposant que l'esperança de la pertorbació és zero, si es prenen esperances en les dues categories s'obté:

$$\begin{aligned} \mu_{salari|dona} &= E(salari \mid dona = 1, educ) = \beta_1 + \delta_1 + \beta_2 educ \\ \mu_{salari|home} &= E(salari \mid dona = 0, educ) = \beta_1 + \beta_2 educ \end{aligned} \tag{5-3}$$

Com es pot veure a (5-3), el terme independent per als homes és β_1, i $\beta_1+\delta_1$ per a les dones. Gràficament, com es pot veure a la figura 5.1, hi ha un desplaçament del terme independent, però les línies per a homes i dones són paral·leles.

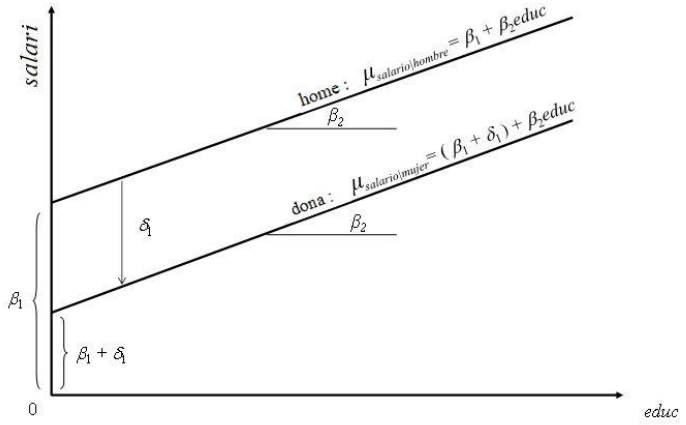


FIGURA 5.1. Mateix pendent, terme independent diferent.

En (5-2) hem inclòs una variable fictícia per a les dones, però no per als homes perquè incloure les dues variables fictícies hauria estat redundant. De fet, tot el que necessitem és dos termes independents, un per a dones i un altre per als homes. Com hem vist, amb la introducció de la variable fictícia dona, ens permet obtenir un terme independent per a cada gènere. La introducció de dues variables fictícies causaria multicolinealitat perfecta, ja que *dona*+*home*=1, el que significa que *home* és una funció lineal exacta de *dona* i del terme independent. La inclusió de variables fictícies per a tots

dos sexes, a més el terme independent, és l'exemple més senzill de l'anomenada trampa de les variables fictícies, com veurem més endavant:

Si fem servir *home* en lloc de *dona*, l'equació de salaris seria la següent:

$$salari = \alpha_1 + \gamma_1 home + \beta_2 educ + u \tag{5-4}$$

No ha canviat amb la nova equació, amb excepció de la interpretació de α_1 i γ_1: α_1 és el terme independent per a les dones, que ara és la categoria de *referència*, i $\alpha_1+\gamma_1$ és el terme independent per als homes. Això implica la següent relació entre els coeficients:

$$\alpha_1=\beta_1+\delta_1 \text{ y } \alpha_1+\gamma_1=\beta_1 \Rightarrow \gamma_1=-\delta_1$$

En qualsevol aplicació, no importa com triem la categoria de referència, ja que només afecta la interpretació dels coeficients associats a les variables fictícies, però és important tenir present quina categoria és la categoria de referència. L'elecció d'una categoria de referència és, generalment, una qüestió de conveniència. També és possible eliminar el terme independent i incloure una variable fictícia per a cada categoria. L'equació serà aleshores:

$$salari = \mu_1 home + \nu_1 dona + \beta_2 educ + u \tag{5-5}$$

on el terme independent és μ_1 per als homes i μ_1 per a les dones.

El contrast d'hipòtesis es realitza com de costum. En el model (5-2) la hipòtesi nul·la de no discriminació entre homes i dones és $H_0 : \delta_1 = 0$, mentre que la hipòtesi alternativa que hi ha discriminació contra la dona és $H_1 : \delta_1 < 0$. Per tant, en aquest cas, hem d'aplicar un contrast *t* d'una sola cua (l'esquerra).

En les especificacions formulades en el treball aplicat és usual transformar la variable dependent prenent logaritmes, ln(*y*), en models d'aquest tipus. Per exemple:

$$\ln(salari) = \beta_1 + \delta_1 dona + \beta_2 educ + u \tag{5-6}$$

Vegem la interpretació del coeficient de la variable fictícia en un model amb logaritmes. En el model (5-6), prenent *u*=0, el salari per a una dona i per a un home són els següents:

$$\ln(salari_M) = \beta_1 + \delta_1 + \beta_2 educ \tag{5-7}$$

$$\ln(salari_H) = \beta_1 + \beta_2 educ \tag{5-8}$$

Donat el mateix nivell d'educació, si restem (5-7) de (5-8), tenim

$$\ln(salari_M) - \ln(salari_H) = \delta_1 \tag{5-9}$$

Prenent antilogaritmes a (5-9) i restant 1 de tots dos membres de (5-9), obtenim

$$\frac{salari_M}{salari_H} - 1 = e^{\delta_1} - 1 \tag{5-10}$$

és a dir

$$\frac{salari_M - salari_H}{salari_H} = e^{\delta_1} - 1 \tag{5-11}$$

D'acord amb (5-11), la taxa de variació entre el salari femení i el salari dels homes, per a un mateix nivell d'educació, és igual a $e^{\delta_1}-1$. Per tant, la taxa exacta de variació percentual entre el salari per hora d'homes i dones és de $100\times(e^{\delta_1}-1)$. Com una aproximació a aquest canvi es pot utilitzar $100\times(e^{\delta_1}-1)$, però si la magnitud del percentatge és alta aquesta aproximació no és tan bona.

EXEMPLE 5.1 Existeix discriminació salarial per a la dona a Espanya?

Utilitzant dades de l'*Enquesta d'Estructura Salarial d'Espanya* per a 2002 (fitxer *wage02sp*), s'ha estimat el model (5-6) i s'han obtingut els següents resultats:

$$\ln(wage) = \underset{(0.026)}{1.731} - \underset{(0.022)}{0.307}\ female + \underset{(0.0025)}{0.055}\,educ$$

$$SQR=393 \quad R^2=0.243 \quad n=2000$$

on *wage* és el salari hora en euros, *femal*e és una variable fictícia que pren el valor 1 si és dona, i *educ* són els anys d'educació. (Els números entre parèntesis són els errors estàndard dels estimadors.)

Per respondre a la pregunta plantejada més amunt, hem de contrastar $H_0:\delta_1=0$ contra de $H_1:\delta_1<0$. Atès que l'estadístic *t* és igual a -14.27 es rebutja la hipòtesi nul·la per α=0.01. És a dir, hi ha evidències d'una discriminació a Espanya contra la dona en l'any 2002. De fet, la diferència percentual en el salari per hora entre homes i dones és $100\times(e^{0.307}-1)=35.9\%$, donats uns mateixos anys en educació.

EXEMPLE 5.2 Anàlisi de la relació entre la capitalització de mercat i el valor comptable: el paper de l'IBEX-35

Un investigador desitja estudiar la relació entre la capitalització de mercat i el valor comptable de les accions cotitzades en el mercat continu de la Borsa de Madrid. En aquest mercat algunes empreses estan incloses en l'IBEX-35, que és un índex selectiu. L'investigador també vol saber si accions incloses en l'Ibex 35 tenen una major capitalització de mitjana. Amb aquest propòsit en ment, l'investigador formula el següent model:

$$\ln(marktval) = \beta_1 + \delta_1 ibex35 + \beta_2 \ln(bookval) + u \qquad (5\text{-}12)$$

- *marktval* és el valor de mercat d'una companyia. Es calcula multiplicant el preu de l'acció pel nombre d'accions emeses.
- *bookval* és el valor comptable d'una companyia. També es coneix com a valor net de la companyia. El valor comptable es calcula com la diferència entre els actius d'una companyia i els seus passius..
- *ibex35* és una variable fictícia que pren el valor 1 si la companyia està inclosa en el selectiu Ibex 35.

Utilitzant les 92 companyies que van cotitzar el 15 de novembre 2011, i que van subministrar informació sobre el valor comptable (fitxer *bolmad11*), es van obtenir els següents resultats:

$$\ln(marktval) = \underset{(0.243)}{1.784} + \underset{(0.179)}{0.690}\,ibex35 + \underset{(0.037)}{0.675}\ln(bookval)$$

$$SQR=35.672 \quad R^2=0.893 \quad n=92$$

L'elasticitat *marktval*/*bookval* és igual a 0.690, és a dir, si el valor comptable s'incrementa en 1%, la capitalització borsària de les accions que cotitzen augmentarà en un 0,675%.

Contrastar si les accions incloses en l'IBEX-35 tenen de mitjà una major capitalització implica contrastar $H_0:\delta_1=0$ contra $H_1:\delta_1>0$. Atès que l'estadístic *t* és (0.690/0.179) = 3.85, aleshores rebutgem la hipòtesi nul·la per als nivells habituals de significació. D'altra banda, veiem que les accions incloses en l'IBEX-35 cotitzen el 99,4% més elevat que les accions no incloses. El percentatge s'obté com segueix: $100\times(e^{0.690}-1)=99.4\%$

EXEMPLE 5.3 Gasten més en peix les persones que viuen en zones urbanes que les que viuen en zones rurals?

Per veure si les persones que viuen en zones urbanes gasten més en peix que les persones que viuen en zones rurals, s'ha proposat el model següent:

$$\ln(fish) = \beta_1 + \delta_1 urban + \beta_2 \ln(inc) + u \tag{5-13}$$

on *fish* és despesa en peix, *urban* és una variable fictícia que pren el valor 1 si la persona viu en una zona urbana i *inc* és la renda disponible.

Utilitzant una mostra de mida 40 (fitxer *demand*), es va estimar el model (5-13):

$$\ln(fish) = \underset{(0.511)}{-6.375} + \underset{(0.055)}{0.140}\, urban + \underset{(0.070)}{1.313}\ln(inc)$$

$$SQR=1.131 \quad R^2=0.904 \quad n=40$$

D'acord amb aquests resultats, les persones que viuen en zones urbanes gasten en peix aproximadament un 14% més que les persones que viuen en zones rurals. Si es contrasta $H_0 : \delta_1 = 0$ contra $H_1 : \delta_1 > 0$, constatem que l'estadístic t és (0.140 / 0.055) = 2.55. Tenint en compte que $t_{37}^{0.01} \approx t_{35}^{0.01}$ =2.44, es rebutja la hipòtesi nul·la en favor de l'alternativa per als nivells habituals de significació. És a dir, hi ha evidència empírica que les persones que viuen a les zones urbanes gasten més en peix que les persones que viuen a les zones rurals.

5.3 Categories múltiples per a un atribut

En l'epígraf anterior vam considerar un atribut (gènere) que té dues categories (dona i home). Ara considerarem atributs amb més de dues categories. En concret, examinarem un atribut amb 3 categories

Per mesurar l'impacte de la mida de l'empresa sobre el salari, podem utilitzar variables dicotòmiques. Suposem que les empreses es classifiquen en tres grups segons la seva grandària: xicotetes (fins a 49 treballadors), mitjanes (de 50 a 199 treballadors) i grans (més de 199 treballadors). Amb aquesta informació podem construir 3 variables fictícies:

$$chicoteta = \begin{cases} 1 & \text{fins 49 treballadors} \\ 0 & \text{en altres casos} \end{cases}$$

$$mitjana = \begin{cases} 1 & \text{de 50 a 199 treballadors} \\ 0 & \text{en altres casos} \end{cases}$$

$$gran = \begin{cases} 1 & \text{mes de 199 treballadors} \\ 0 & \text{en altres casos} \end{cases}$$

Si volem explicar el salari per hora introduint en el model la mida de l'empresa, cal ometre una de les categories. En el següent model la categoria omesa són les empreses xicotetes:

$$salari = \beta_1 + \theta_1 mitjana + \theta_2 gran + \beta_2 educ + u \tag{5-14}$$

L'interpretació de los coeficients θ_j és la següent: θ_1 (θ_2) és la diferència en el salari per hora entre les empreses mitjanes (grans) i les xicotetes, donat un mateix nivell d'educació (i un mateix terme de pertorbació, *u*).

Anem a veure què passa si també incloem en (5-14) la categoria petita. En aquest cas, tindríem el següent model:

$$salari = \beta_1 + \theta_0 chicoteta + \theta_1 mitjana + \theta_2 gran + \beta_2 educ + u \qquad (5\text{-}15)$$

Ara, considerem que tenim una mostra de sis observacions: les observacions 1 i 2 corresponen a empreses xicotetes, la 3 i la 4 a mitjanes, i la 5 i 6 a grans. En aquest cas, la matriu **X** de regressors tindria la següent configuració:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & educ_1 \\ 1 & 1 & 0 & 0 & educ_2 \\ 1 & 0 & 1 & 0 & educ_3 \\ 1 & 0 & 1 & 0 & educ_4 \\ 1 & 0 & 0 & 1 & educ_5 \\ 1 & 0 & 0 & 1 & educ_6 \end{bmatrix}$$

Com es pot veure en la matriu **X**, la columna 1 d'aquesta matriu és igual a la suma de les columnes 2, 3 i 4. Per tant, hi ha *multicolinealitat perfecta*, a causa de l'anomenada *parany de les variables fictícies*. Generalitzant, si un atribut té *g* categories, en el model únicament hem d'incloure g-1 variables fictícies, juntament amb el terme independent. El terme independent per a la categoria de referència és el terme independent general del model, i el coeficient de la variable fictícia d'un grup particular representa la diferència estimada entre els termes independents entre aquesta categoria i la categoria de referència. Si incloem *g* variables fictícies, juntament amb un terme independent es caurà en el parany de les variables fictícies. Una alternativa és incloure *g* variables fictícies, i excloure el terme independent general. En el cas que ens ocupa, el model seria el següent:

$$salari = \theta_0 chicoteta + \theta_1 mitjana + \theta_2 gran + \beta_1 educ + u \qquad (5\text{-}16)$$

Aquesta solució no és aconsellable per dues raons. Amb aquesta configuració del model és més difícil contrastar les diferències respecte a una categoria de referència. En segon lloc, aquesta solució només funciona en el cas d'un model amb només un atribut.

EXEMPLE 5.4 Influeix la mida de l'empresa en la determinació dels salaris?

Utilitzant la mostra del Exemple 5.1 (fitxer *wage02sp*), es va estimar el model (5-14), prenent log en *wage*:

$$\ln(wage) = \underset{(0.027)}{1.566} + \underset{(0.025)}{0.281}\, medium + \underset{(0.024)}{0.162}\, large + \underset{(0.003)}{0.048}\, educ$$

$$SQR=406 \quad R^2=0.218 \quad n=2000$$

on *medium* i *large* són dues variables dicotòmiques per a designar a empreses de mida mitjana i gran respectivament.

Per respondre a la pregunta inicial no farem un contrast *individual* de θ_1 o θ_2. En comptes d'això, contrastarem conjuntament si la mida de les empreses té una influència significativa sobre el salari. És a dir, hem de contrastar si les mitjanes i grans empreses preses conjuntament tenen una influència significativa en la determinació del salari. En aquest cas, les hipòtesis nul·la i alternativa, prenent a (5-14) com el model no restringit, seran les següents:

$$H_0 : \theta_1 = \theta_2 = 0$$
$$H_1 : H_0 \text{ no es certa}$$

El model restringit en aquest cas és el següent:

$$\ln(wage) = \beta_1 + \beta_2 educ + u \qquad (5\text{-}17)$$

L'estimació d'aquest model és la següent:

$$\ln(wage) = \underset{(0.026)}{1.657} + \underset{(0.003)}{0.053}educ$$

$$SQR=433 \quad R^2=0.166 \quad n=2000$$

Per tant, l'estadístic F és

$$F = \frac{[SQR_R - SCR_{NR}]/q}{SQR_{NR}/(n-k)} = \frac{[433-406]/2}{406/(2000-4)} = 66.4$$

Així, d'acord amb el valor de l'estadístic F, es pot concloure que la mida de la firma té una influència significativa en la determinació dels salaris per als nivells usuals de significació.

EXEMPLE 5.5 En el cas de Lydia E. Pinkham, són significatives les variables temporals fictícies de forma individual i conjunta?

Al Exemple 3.4 vam veure el cas de Lydia E. Pinkham en què les vendes, *sales*, d'un extracte d'herbes d'aquesta empresa (en milers de dòlars) s'explicava en termes de la despesa en publicitat en milers de dòlars (*advexp*) així com les vendes de l'any anterior ($salest_{-1}$). No obstant això, el seu autor, a més d'aquestes dues variables, va incloure tres variables temporals fictícies: $d1$, $d2$ i $d3$. Aquestes variables fictícies abasten les diferents situacions per les quals va passar la companyia. Així, $d1$ pren el valor 1 en el període de 1907-1914 i 0 en els períodes restants, $d2$ pren el valor 1 en el període de 1915-1925 i 0 en altres períodes, i finalment, $d3$ pren el valor 1 en el període de 1926 a 1940 i 0 en els restants períodes. Per tant, la categoria de referència és el període 1941-1960. En conseqüència, la formulació final del model va ser la següent:

$$sales_t=\beta_1+\beta_2 advexp_t+\beta_3\, sales_{t-1}+\beta_4 d1_t+\beta_5 d2_t+\beta_6 d3_t+u_t \qquad (5\text{-}18)$$

Els resultats obtinguts en la regressió, utilitzant el fitxer *pinkham*, van ser els següents:

$$sales_t = \underset{(96.3)}{254.6} + \underset{(0.136)}{0.5345}\, advexp_t + \underset{(0.0814)}{0.6073}\, sales_{t-1} - \underset{(89)}{133.35}\, d1_t + \underset{(67)}{216.84}\, d2_t - \underset{(67)}{202.50}\, d3_t$$

$$R^2=0.929 \quad n=53$$

Per contrastar si les variables fictícies de forma individual tenen un efecte significatiu en les vendes, les hipòtesis nul·la i alternativa són:

$$\begin{cases} H_0 : \theta_i = 0 \\ H_1 : \theta_i \neq 0 \end{cases} \qquad i = 1,2,3$$

Els corresponents estadístiques t són els següents:

$$t_{\hat{\theta}_1} = \frac{-133.35}{89} = -1.50 \qquad t_{\hat{\theta}_2} = \frac{216.84}{67} = 3.22 \qquad t_{\hat{\theta}_3} = \frac{-202.50}{67} = -3.02$$

Com es pot veure, el regressor $d1$ no és significatiu per als nivells habituals de significació, mentre que per contra els regressors $d2$ i $d3$ són significatius per a qualsevol dels nivells habituals.

La interpretació del coeficient del regressor $d2$, per exemple, és la següent: mantenint fix la despesa en publicitat i donades les vendes de l'any anterior, les vendes per a un any del període 1915-1920 són 21.684 dòlars grans, de mitjana, que les d'un any qualsevol del període 1941-1960.

Per estimar l'efecte conjunt de les variables temporals fictícies, les hipòtesis nul·la i alternativa són

$$\begin{cases} H_0 : \theta_1 = \theta_2 = \theta_3 = 0 \\ H_1 : H_0 \text{ no es certa} \end{cases}$$

i el contrast estadístic corresponent és

$$F = \frac{(R^2_{NR} - R^2_R)/q}{(1-R^2_{NR})/(n-k)} = \frac{(0.9290-0.8770)/3}{(1-0.9290)/(53-6)} = 11.47$$

Per a qualsevol dels nivells habituals de significació la hipòtesi nul·la és rebutjada. Per tant, les variables temporals fictícies tenen un efecte significatiu sobre les vendes.

5.4 Diversos atributs

Ara considerarem la possibilitat de tenir en compte dos atributs per explicar la determinació del salari: el gènere i durada de la jornada de treball (a temps parcial i a temps complet). La variable fictícia tempar, va ser una variable binària que pren el valor 1 quan el tipus de contracte és a temps parcial i 0 si és a temps complet. En el següent model s'introdueixen les dues variables fictícies: *dona* i *tempar*:

$$salari = \beta_1 + \delta_1 dona + \phi_1 tempar + \beta_2 educ + u \tag{5-19}$$

En aquest model, ϕ_1 és la diferència en el salari per hora entre les persones que treballen a temps parcial, per a un gènere donat i per al mateix nivell d'educació (i també el mateix terme de pertorbació, *u*).

Cada un d'aquests dos atributs té una categoria de referència, que és la categoria omesa. En aquest cas, home és la categoria de referència per al gènere i temps complet per al tipus de contracte. Si prenem les esperances per a les quatre categories implicades, s'obté:

$$\begin{aligned}
\mu_{wage|dona,tempar} &= E[wage \mid dona, tempar, educ] = \beta_1 + \delta_1 + \phi_1 + \beta_2 educ \\
\mu_{wage|dona,temcom} &= E[wage \mid dona, temcom, educ] = \beta_1 + \delta_1 + \beta_2 educ \\
\mu_{wage|home,tempar} &= E[wage \mid home, tempar, educ] = \beta_1 + \phi_1 + \beta_2 educ \\
\mu_{wage|home,tiemcom} &= E[wage \mid home, temcom, educ] = \beta_1 + \beta_2 educ
\end{aligned} \tag{5-12}$$

El terme independent general en l'equació reflecteix l'efecte de les dues categories de referència, home i temps complet; és a dir, la categoria de referència és home amb jornada a temps complet. En (5-20) es pot veure el terme independent per a cada combinació de categories.

EXEMPLE 5.6 La influència de gènere i durada de la jornada de treball en la determinació dels salaris

El model (5-19), prenent log en wage, es va estimar utilitzant dades de l'Enquesta d'Estructura Salarial d'Espanya per a l'any 2006 (fitxer *wage06sp*):

$$\ln(wage) = \underset{(0.026)}{2.005} - \underset{(0.021)}{0.233}\, female - \underset{(0.027)}{0.087}\, partime + \underset{(0.002)}{0.053} educ$$

$$SQR=365 \quad R^2=0.235 \quad n=2000$$

on *partime* és un contracte a temps parcial.

D'acord amb els valors dels coeficients i els corresponents errors estàndard, és evident que cadascuna de les dues variables fictícies, female i partime, són estadísticament significatives per als nivells habituals de significació.

EXEMPLE 5.7 Anàlisi de l'absentisme laboral a l'empresa Buenosaires

Buenosaires és una empresa dedicada a la fabricació de ventiladors, havent tingut resultats relativament acceptables en els últims anys. Els directius consideren que aquests haurien estat millors si l'absentisme a l'empresa no fos tan alt. Amb el propòsit d'analitzar els factors que determinen l'absentisme, es proposa el següent model:

$$absent = \beta_1 + \delta_1 bluecoll + \phi_1 male + \beta_2 age + \beta_3 tenure + \beta_4 wage + u \qquad (5\text{-}21)$$

on *bluecoll* és una variable fictícia que indica que la persona és un treballador manual (la categoria de referència és coll blanc), *male* és una variable dicotòmica que pren el valor 1 si el treballador és home. Les variables *tenure* i *age* són contínues que reflecteixen els anys treballant a l'empresa i l'edat respectivament.

Utilitzant una mostra de mida 48 (fitxer *absent*), s'ha estimat la següent equació.

$$\underset{}{absent} = \underset{(1.640)}{12.444} + \underset{(0.669)}{0.968}\, bluecoll + \underset{(0.712)}{2.049}\, male - \underset{(0.047)}{0.037}\, age - \underset{(0.065)}{0.151}\, tenure - \underset{(0.007)}{0.044}\, wage$$

$$SQR=161.95 \quad R^2=0.760 \quad n=48$$

Ara veurem si bluecoll és significativa. Contrastant $H_0 : \delta_1 = 0$ contra $H_1 : \delta_1 \neq 0$, l'estadístic *t* és (0968/0669) = 1.45. Com $t_{40}^{0.10/2}$ =1.68, fracassem en rebutjar la hipòtesi nul·la para α=0.10. Aleshores no hi ha evidència empírica per afirmar que l'absentisme dels treballadors manuals (coll blau) és diferent del dels treballadors d'oficina (coll blanc). Però si es contrasta $H_0 : \delta_1 = 0$ contra $H_1 : \delta_1 > 0$, com $t_{40}^{0.10}$ =1.30 per a α=0.10, no es pot rebutjar que l'absentisme dels treballadors de coll blau siga més gran que el dels treballadors de coll blanc.

Per contra, en el cas de la variable fictícia *male*, contrastant $H_0 : \varphi_1 = 0$ contra $H_1 : \varphi_1 \neq 0$, atès que l'estadístic *t* és (2.049/0.712)=2.88 i $t_{40}^{0.01/2}$ =2.70, rebutgem que l'absentisme siga igual en homes i dones per als nivells habituals de significació.

EXEMPLE 5.8 Mida de l'empresa i gènere en la determinació del salari

Per conèixer si la mida de l'empresa i el gènere, de forma conjunta, són dos factors rellevants en la determinació del salari, es formula el següent model:

$$\ln(wage) = \beta_1 + \delta_1 female + \theta_1 medium + \theta_2 large + \beta_2 educ + u \qquad (5\text{-}22)$$

En aquest cas hem de fer un contrast conjunt, on les hipòtesis nul·la i alternativa són,

$$H_0 : \delta_1 = \theta_1 = \theta_2 = 0$$
$$H_1 : H_0 \text{ no es cierta}$$

El model restringit en aquest cas és el model de (5-17), que es va estimar en l'exemple 5.4 (fitxer *wage02sp*). L'estimació del model no restringit és la següent:

$$\underset{}{\ln(wage)} = \underset{(0.026)}{1.639} - \underset{(0.021)}{0.327}\, female + \underset{(0.023)}{0.308}\, medium + \underset{(0.023)}{0.168}\, large + \underset{(0.0024)}{0.050}\, educ$$

$$SQR=361 \quad R^2=0.305 \quad n=2000$$

L'estadístic *F* es

$$F = \frac{\left[SQR_R - SCR_{NR}\right]/q}{SQR_{NR}/(n-k)} = \frac{\left[433-361\right]/3}{361/(2000-5)} = 133$$

Per tant, d'acord amb el valor de *F*, es pot concloure que la mida de la signatura i el gènere tenen conjuntament una influència significativa en la determinació del salari.

5.5 Les interaccions que impliquen variables fictícies

5.5.1 Interaccions entre dues variables fictícies

Per permetre la possibilitat que hi hagi una interacció entre el gènere i durada de la jornada de treball en la determinació salarial podem afegir al model (5-19) un terme d'interacció entre *dona* i *tempar*, de manera que el model a estimar serà el següent:

$$salari = \beta_1 + \delta_1 dona + \phi_1 tempar + \varphi_1 dona \times tempar + \beta_2 educ + u \qquad (5\text{-}23)$$

Això permet determinar si l'efecte de la durada de la jornada de treball en el salari depèn, o no, del gènere. Anàlogament, també permet si la influència del gènere en el salari depèn, o no, de la durada de la jornada de treball.

EXEMPLE 5.9 És la interacció entre les dones i el treball a temps parcial significativa?

El model (5-23) es va estimar utilitzant les dades de l'Enquesta d'Estructura Salarial d'Espanya per a 2006 (fitxer *wage06sp*):

$$\ln(wage) = \underset{(0.026)}{2.007} - \underset{(0.022)}{0.259}\, female - \underset{(0.047)}{0.198}\, partime + \underset{(0.058)}{0.167}\, female \times partime + \underset{(0.002)}{0.054}\, educ$$

$$SQR=363 \quad R^2=0.238 \quad n=2000$$

Per respondre a la pregunta plantejada, hem de contrastar $H_0 : \varphi_1 = 0$ contra $H_0 : \varphi_1 \neq 0$. Atès que l'estadístic t és (0.167 / 0.058) = 2.89, i tenint en compte que $t_{60}^{0.01/2} = 2.66$ es rebutja la hipòtesi nul·la en favor de la hipòtesi alternativa. Per tant, hi ha evidència empírica que la interacció entre female i partime és estadísticament significativa.

EXEMPLE 5.10 Discriminen les empreses petites a les dones més, o menys, que les empreses grans?

Per respondre a aquesta pregunta es formula el següent model:

$$\begin{aligned} \ln(wage) &= \beta_1 + \delta_1\, female + \theta_1 medium + \theta_2 large \\ &+ \varphi_1\, female \times medium + \varphi_2\, female \times large + \beta_2 educ + u \end{aligned} \qquad (5\text{-}24)$$

Utilitzant la mostra de l'Exemple 5.1 (fitxer *wage02sp*), va ser estimat el model (5-24):

$$\begin{aligned} \ln(wage) = \underset{(0.027)}{1.624} - \underset{(0.034)}{0.262}\, female + \underset{(0.028)}{0.361}\, medium + \underset{(0.027)}{0.179}\, large \\ - \underset{(0.050)}{0.159}\, female \times medium - \underset{(0.051)}{0.043}\, female \times large + \underset{(0.0024)}{0.050}\, educ \end{aligned}$$

$$SCR=359 \quad R^2=0.308 \quad n=2000$$

Si en (5-24) els paràmetres φ_1 i φ_2 són igual a 0, això implica que, en l'equació per a la determinació del salari, no hi ha interacció entre gènere i mida de l'empresa. Així per respondre a la pregunta plantejada prenem (5-24) com el model no restringit. Les hipòtesis nul·la i alternativa seran les següents:

$$H_0 : \varphi_1 = \varphi_2 = 0$$
$$H_1 : H_0 \text{ no es certa}$$

Per tant, el model restringit és, en aquest cas, el model (5-22), que es va estimar en l'Exemple 5.7. L'estadístic *F* pren el valor

$$F = \frac{[SCR_R - SCR_{NR}]/q}{SCR_{NR}/(n-k)} = \frac{[361-359]/2}{359/(2000-7)} = 5.55$$

Per α=0.01, resulta que $F_{2,1993}^{0.01} \simeq F_{2,60}^{0.01} = 4.98$. Com *F*>5.61, rebutgem H_0 en favor de H_1. Si se rebutja H_0 per α=0.01, també serà rebutjada per als nivells de 5% i 10%. Per tant, per als nivells usuals de significació, la interacció entre gènere i mida d'empresa és rellevant en la determinació del salari.

5.5.2 Interaccions entre una variable fictícia i una variable quantitativa

Fins ara, en els exemples sobre determinació del salari s'ha utilitzat una variable fictícia per desplaçar el terme independent o per estudiar la seva interacció amb una altra variable fictícia, però mantenint la pendent d'*educ* constant. Ara bé, també es poden utilitzar les variables fictícies per desplaçar pendents si interactuen amb qualsevol

variable explicativa contínua. Per exemple, en el següent model la variable fictícia *dona* interactua amb la variable contínua *educ*:

$$salario = \beta_1 + \beta_2 educ + \delta_1 mujer \times educ + u \tag{5-25}$$

En aquest model, com es pot veure a la figura 5.2, el terme independent és el mateix per a homes i per a dones, però el pendent és més gran en homes que en dones, perquè δ_1 és negativa.

En el model de (5-25), els rendiments d'un any addicional en educació depenen del gènere de l'individu. De fet,

$$\frac{\partial salari}{\partial educ} = \begin{cases} \beta_2 + \delta_1 & \text{per a dones} \\ \beta_2 & \text{per a homes} \end{cases} \tag{5-26}$$

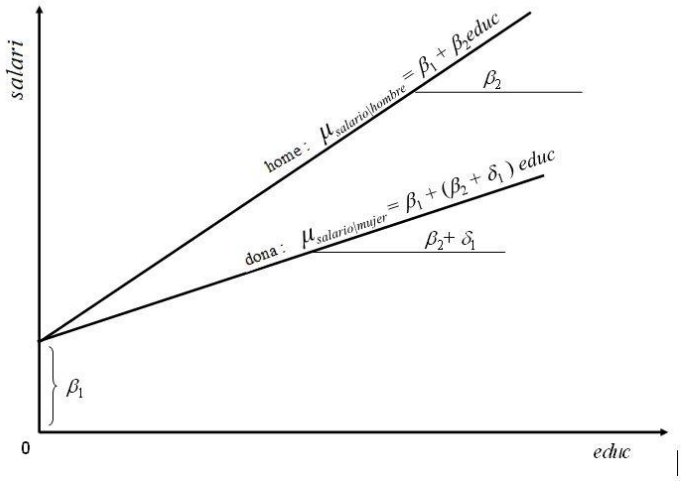


FIGURA 5.2. Diferent pendent, mateix terme independent.

EXEMPLE 5.11 És el rendiment de l'educació per als homes més que per a les dones?

Utilitzant la mostra del Exemple 5.1 (fitxer *wage 02sp*), prenent logaritmes en wage, s'ha estimat el model (5-25):

$$\ln(wage) = \underset{(0.025)}{1.640} + \underset{(0.0026)}{0.063} educ - \underset{(0.0021)}{0.027} educ \times female$$

$$SCR=400 \quad R^2=0.229 \quad n=2000$$

En aquest cas necessitem contrastar $H_0 : \delta_1 = 0$ contra $H_1 : \delta_1 < 0$. Atès que l'estadístic *t* és (-0.028/0.0002)=-12.81, es rebutja la hipòtesi nul·la en favor de la hipòtesi alternativa per a qualsevol nivell de significació. És a dir, existeix evidència empírica que el rendiment d'un any addicional d'educació és més gran per a homes que per a dones

5.6 Contrast de canvi estructural

Fins ara hem contrastat les hipòtesis que un paràmetre, o un subconjunt de paràmetres del model, són diferents per a dos grups (dones i homes, per exemple). Però de vegades, volem contrastar la hipòtesi nul·la que dos grups tenen la mateixa funció de regressió poblacional, enfront de l'alternativa que no és la mateixa. En altres paraules, volem contrastar si la mateixa equació és vàlida per als dos grups. Hi ha dos procediments

per realitzar aquest contrast, denominat contrast de canvi estructural: utilitzant variables fictícies i realitzant regressions separades mitjançant el contrast de Chow.

5.6.1 Utilitzant variables fictícies

En aquest procediment, contrastar si hi ha diferències entre grups consisteix a realitzar un contrast de significació conjunt de la variable fictícia que diferència entre els dos grups i de les seves interaccions amb totes els altres regressors. Per tant, estimem el model amb (model no restringit) i sense (model restringit) la variable fictícia i totes les seves interaccions.

De l'estimació de les dues equacions s'obté l'estadístic F, ja siga a través de la *SQR* o del R^2. En el següent model, per a la determinació del salari, tant el terme independent com el pendent són diferents per a homes i dones:

$$salari = \beta_1 + \delta_1 dona + \beta_2 educ + \delta_2 dona \times educ + u \qquad (5\text{-}27)$$

A la figura 5.3, ha estat representada la funció de regressió poblacional d'aquest model. Com es pot veure, si *dona*=1, s'obté que

$$salari = (\beta_1 + \delta_1) + (\beta_2 + \delta_2) educ + u \qquad (5\text{-}28)$$

Aleshores, per a les dones el terme independent és $\beta_1 + \delta_1$ i el pendent $\beta_2 + \delta_2$. Per *dona*=0, obtenim l'equació (5-1). Per *dona*=0, obtenim l'equació (5-1). En aquest cas, per als homes el terme independent és Per tant β_1 i el pendent β_2. Per tant, δ_1 mesura la diferència entre els termes independents per a dones i homes i δ_2 mesura al seu torn la diferència en el rendiment l'educació entre dones i homes. La figura 5.3 mostra un terme independent i un pendent menors per a dones que per a homes. Això vol dir que les dones guanyen menys que els homes en tots els nivells de l'educació, i que la bretxa augmenta a mesura que educ es fa més gran, és a dir, un any addicional d'educació té un rendiment inferior per a dones que per a homes.

L'estimació (5-27) és equivalent a l'estimació de dues equacions de salaris, un per a homes i un altre per a les dones, per separat. L'única diferència és que (5-27) imposa la mateixa variança als dos grups, mentre que les regressions per separat no ho fan. Aquesta especificació del model és ideal, com veurem més endavant, per contrastar la igualtat de pendents, la igualtat de termes independents, o la igualtat tant de termes independents com de pendents en els dos grups.

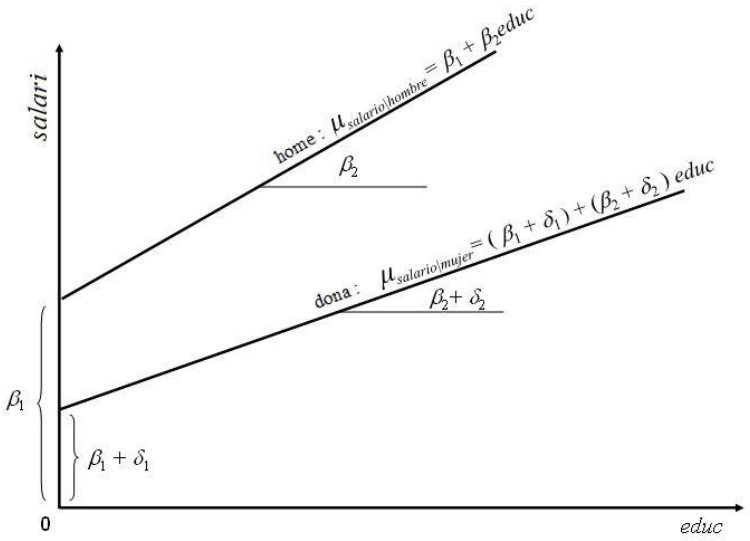


FIGURA 5.3. Pendent diferent, diferent terme independent.

EXEMPLE 5.12 És l'equació de salaris vàlida tant per a homes com per a dones?

Si els paràmetres δ_1 i δ_2 són iguals a 0 en el model de (5-27), implica que l'equació per a la determinació dels salaris és la mateixa per a homes i dones. Aleshores per respondre a la qüestió plantejada, prenem (5-27), però expressant el salari en logaritmes, com el model el model no restringit. Les hipòtesis nul·la i alternativa seran les següents:

$$H_0 : \delta_1 = \delta_2 = 0$$
$$H_1 : H_0 \text{ no es cierto}$$

Per tant, el model restringit s'obté aplicant la hipòtesi nul·la al model (5-27). Utilitzant la mateixa mostra que en l'Exemple 5.1 (fitxer *wage02sp*), hem obtingut la següent estimació dels models (5-27) i (5-17):

$$\ln(wage) = \underset{(0.030)}{1.739} - \underset{(0.055)}{0.332}\, female + \underset{(0.0030)}{0.054}\, educ - \underset{(0.0054)}{0.003}\, educ \times female$$

SQR=393 R^2=0.243 n=2000

$$\ln(wage) = \underset{(0.026)}{1.657} + \underset{(0.0026)}{0.0525}\, educ$$

SQR=433 R^2=0.166 n=2000

L'estadístic F pren el valor

$$F = \frac{[SQR_R - SQR_{NR}]/q}{SQR_{NR}/(n-k)} = \frac{[433-393]/2}{393/(2000-4)} = 102$$

Quan contrastem a l'Exemple 5.1 si hi havia discriminació contra la dona a Espanya ($H_0 : \delta_1 = 0$ contra $H_1 : \delta_1 < 0$), es va assumir que el pendent d'*educ* (model (5-6)) era la mateixa per a homes i dones. Ara també és possible utilitzar el model (5- 27) per contrastar la mateixa hipòtesi nul·la, però assumint que el pendent és diferent. Atès que l'estadístic *t* és (-0.332/0,0546)= - 6.06, aleshores es rebutja la hipòtesi nul·la mitjançant l'ús d'aquest model més general que el de l' Exemple 5.1.

A l'Exemple 5.11 es va contrastar si el coeficient de δ_2 en el model de (5-25), prenent log en *wage*, era 0, suposant que el terme independent és el mateix per a homes i dones. Ara bé, si prenem (5-27), prenent log en *wage*, com a model no restringit, podem contrastar la mateixa hipòtesi nul·la, però assumint

que el terme independent és diferent per a homes i dones. Atès que l'estadístic *t* és (0.0027/0,0054)=0.493, aleshores no es pot rebutjar la hipòtesi nul·la que no hi ha interacció entre gènere i educació.

EXEMPLE 5.13 Tenen els consumidors urbans el mateix patró de comportament que els rurals pel que fa a la despesa en peix?

Per respondre a aquesta pregunta es formula el següent model, que es prendrà com a model no *restringit*:

$$\ln(fish) = \beta_1 + \delta_1 urban + \beta_2 \ln(inc) + \delta_2 \ln(inc) \times urban + u \tag{5-29}$$

Les hipòtesis nul·la i alternativa seran les següents:

$$H_0 : \delta_1 = \delta_2 = 0$$
$$H_1 : H_0 \text{ no es certa}$$

El model restringit corresponent a aquesta H_0 és

$$\ln(fish) = \beta_1 + \beta_2 \ln(inc) + u \tag{5-30}$$

Utilitzant la mostra de l'Exemple 5.3 (fitxer *demand*), s'han estimat els models (5-29) i (5-30):

$$\ln(fish) = \underset{(0.627)}{-6.551} + \underset{(1.095)}{0.678} urban + \underset{(0.087)}{1.337} \ln(inc) - \underset{(0.152)}{0.075} \ln(inc) \times urban$$

SQR=1.123 R^2=0.904 n=40

$$\ln(fish) = \underset{(0.542)}{-6.224} + \underset{(0.075)}{1.302} \ln(inc)$$

SQR=1.325 R^2=0.887 n=40

L'estadístic F pren el valor

$$F = \frac{[SCR_R - SCR_{NR}]/q}{SCR_{NR}/(n-k)} = \frac{[1.325 - 1.123]/2}{1.123/(40-4)} = 3.24$$

Si mirem a la taula estadística de la F per a 2 *graus de llibertat* al numerador i 35 *gl* al denominador per α=0.10, veiem que $F^{0.10}_{2,36} \simeq F^{0.10}_{2,35} = 2.46$. Com F>2.46 rebutgem la H_0. No obstant això, com $F^{0.05}_{2,36} \simeq F^{0.05}_{2,35} = 3.27$, vam fracassar a rebutjar H_0 a favor de H_1 per α=0.05 i, per tant, per α=0.01. Conclusió: no hi ha una evidència forta que les famílies que viuen a les zones rurals tinguin un patró de consum diferent de peix pel que fa a les famílies que viuen en zones urbanes.

Exemple 5.14 Ha canviat l'estructura productiva de les regions espanyoles?

La pregunta que s'ha de respondre és específicament la següent: ¿ha canviat l'estructura productiva de les regions espanyoles entre 1995 i 2008? El problema que es planteja és un problema d'estabilitat estructural. Per especificar el model que es pren com a referència en l'estimació, definirem la variable fictícia *y*2008, que pren el valor 1 si l'any és 2008 i 0 si l'any és 1995.

El model de referència és un model de Cobb-Douglas, que introdueix paràmetres addicionals per recollir els canvis estructurals que pugan haver passat. La seua expressió és la següent:

$$\ln(q) = \gamma_1 + \alpha_1 \ln(k) + \beta_1 \ln(l) + \gamma_2 y2008 + \alpha_2 y2008 \times \ln(k) + \beta_2 y2008 \times \ln(l) + u \tag{5-31}$$

És fàcil veure, d'acord amb la definició de la variable fictícia *y*2008 que les elasticitats de producció/capital en 1995 i 2008 són diferents. En concret, prenen els següents valors:

$$\varepsilon_{Q/K(1995)} = \frac{\partial \ln(Q)}{\partial \ln(K)} = \alpha_1 \qquad \varepsilon_{Q/K(2008)} = \frac{\partial \ln(Q)}{\partial \ln(K)} = \alpha_1 + \alpha_2$$

En el cas que α_2 siga igual a 0, llavors l'elasticitat de la producció / capital és la mateixa en els dos períodes.

De la mateixa manera, les elasticitats de producció/treball per als dos períodes vénen donades per

$$\varepsilon_{Q/K(1995)} = \frac{\partial \ln(L)}{\partial \ln(K)} = \beta_1 \qquad \varepsilon_{Q/K(2008)} = \frac{\partial \ln(L)}{\partial \ln(K)} = \beta_1 + \beta_2$$

El terme independent de la funció Cobb-Douglas és un paràmetre que mesura l'eficiència. En el model de (5-31) es considera la possibilitat que el paràmetre d'eficiència (*PEF*) siga diferent en els dos períodes. així,

$$PEF(1995) = \gamma_1 \qquad PEF(2008) = \gamma_1 + \gamma_2$$

Si els paràmetres α_2, β_2 i γ_2 són zero en el model (5-31), la funció de producció és la mateixa en els dos períodes. Per tant, en l'estimació d'estabilitat estructural de la funció de producció, les hipòtesis nul·la i alternativa són:

$$\begin{aligned} &H_0 : \gamma_2 = \ \alpha_2 = \beta_2 \\ &H_1 : H_0 \text{ no es certa} \end{aligned} \tag{5-32}$$

Sota la hipòtesi nul·la, les restriccions donades en (5-32) condueixen al model restringit següent:

$$\ln(q) = \gamma_1 + \alpha_1 \ln(k) + \beta_1 \ln(l) + u \tag{5-33}$$

El fitxer *prodsp* conté informació per a cadascuna de les regions espanyoles el 1995 i 2008 sobre el valor afegit brut en milions d'euros (*gdp*), l'ocupació en milers de llocs de treball (*labor*), i el capital productiu en milions d'euros (*captot*). També en aquest arxiu es pot trobar la variable fictícia 2008.

A continuació es mostren els resultats del model de regressió no restringit (5-31). És evident que no podem rebutjar la hipòtesi nul·la que cada un dels coeficients α_2, β_2 i γ_2, considerats individualment, siga 0, ja que cap dels estadístics t arriba a 0.1 en valor absolut.

$$\begin{aligned} \ln(gva) = {} & \underset{(0.916)}{0.0559} + \underset{(0.185)}{0.6743}\ln(captot) + \underset{(0.185)}{0.3291}\ln(labor) \\ & - \underset{(2.32)}{0.1088}\, y2008 + \underset{(0.419)}{0.0154}\, y2008 \times \ln(captot) - \underset{(0.418)}{0.0094}\, y2008 \times \ln(labor) \end{aligned}$$

$$R^2 = 0.99394 \qquad n = 34$$

Els resultats del model restringit (5-33) són els següents:

$$\ln(gva) = -\underset{(0.200)}{0.0690} + \underset{(0.036)}{0.6959}\ln(captot) + \underset{(0.042)}{0.311}\ln(labor)$$

$$R^2 = 0.99392 \qquad n = 34$$

Com es pot veure, les R^2 dels dos models són pràcticament idèntiques ja que difereixen només a partir del cinquè decimal. No és estrany, per tant, que l'estadístic *F* per al contrast de la hipòtesi nul·la (5-32) tinga un valor proper a 0:

$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1 - R^2_{UR})/(n-k)} = \frac{(0.99394 - 0.99392)/3}{(1 - 0.99394)/(34-6)} = 0.0308$$

Així doncs, la hipòtesi alternativa que existisca canvi estructural en l'economia productiva de les regions espanyoles entre 1995 i 2008 es rebutja per a qualsevol nivell de significació.

5.6.2 Utilitzant regressions separades: el contrast de Chow

Aquest contrast va ser introduït pel económetra Chow (1960). Aquest autor va considerar el problema de contrastar la igualtat de dos conjunts de coeficients de

regressió. En el contrast de Chow el model restringit és el mateix que en el cas d'ús de variables fictícies per diferenciar entre grups. Ara, però, el model no restringit, en lloc de distingir el comportament de dos grups mitjançant variables fictícies, consisteix simplement en regressions separades. Així, en l'exemple determinació dels salaris, el model no restringit consta de dues equacions:

$$\begin{aligned} &dona: \quad salari = \beta_{11} + \beta_{21} educ + u \\ &home: \quad salari = \beta_{12} + \beta_{22} educ + u \end{aligned} \tag{5-34}$$

Si estimem les dues equacions per *MQO*, es pot demostrar que la *SQR* del model no restringit, SQR_{NR}, és igual a la suma de la *SQR* obtinguda de l'estimació per a les dones, SQR_1, i per als homes, SQR_2, és a dir,

$$SQR_{NR}=SQR_1+SQR_2$$

La hipòtesi nul·la estableix que els paràmetres de les dues equacions en (5-34) són iguals. Aleshores,

$$\begin{aligned} &H_0: \begin{cases} \beta_{11} = \beta_{12} \\ \beta_{21} = \beta_{22} \end{cases} \\ &H_1: \text{No } H_0 \end{aligned}$$

Aplicant la hipòtesi nul·la al model (5-34), s'obté el model (5-17), que és el model restringit. L'estimació d'aquest model per a tota la mostra se sol denominar regressió agrupada o *pooled* regressió (*P*). Per tant, considerarem que SQR_R i SQR_P són expressions equivalents.

Per tant, l'estadístic *F* serà la següent:

$$F = \frac{\left[SQR_P - \left(SQR_1 + SQR_2\right)\right]/k}{\left[SQR_1 + SQR_2\right]/\left[n-2k\right]} \tag{5-35}$$

És important assenyalar que, sota la hipòtesi nul·la, han de ser iguals les variances de la pertorbació per als grups. Observe que tenim *k* restriccions: els *k*-1 coeficients de pendent (interaccions), més el coeficient del terme independent. Cal notar també que en el model no restringit estimem 2 termes independents diferents i 2 coeficients de pendent diferents, de manera que els *gl* del model són *n*-2*k*.

Una limitació important del contrast de Chow és que sota la hipòtesi nul·la no hi ha diferències en absolut entre els grups. En la majoria dels casos, és més interessant permetre diferències parcials entre els dos grups, com hem fet mitjançant la utilització de variables fictícies.

El contrast de Chow es pot generalitzar a més de dos grups d'una manera natural. Des del punt de vista pràctic, és probablement més fàcil estimar regressions separades per a cada grup que utilitzar el procediment basat en la introducció de variables fictícies en el model.

En el cas de tres grups l'estadístic *F* en el contrast de Chow té la següent configuració:

$$F = \frac{\left[SQR_P - (SQR_1 + SQR_2 + SQR_3)\right] / 2 \times k}{(SQR_1 + SQR_2 + SQR_3) / (n - 3k)} \tag{5-36}$$

Observe que, com a regla general, el nombre de *gl* del numerador és igual al (nombre de grups-1) × k, mentre que el nombre de *gl* del denominador és igual a n menys (nombre de grups) × *k*.

EXEMPLE 5.15 Una altra forma d'abordar la qüestió de la determinació dels salaris per criteri de gènere

Utilitzant la mateixa mostra que en l'Exemple 5.1 (fitxer *wage02sp*), hem obtingut l'estimació de les equacions en (5-34), prenent log en wage, per a homes i dones, les quals preses conjuntament donen lloc a l'estimació del model *no restringit*:

Equació per a la dona $\ln(wage) = \underset{(0.042)}{1.407} + \underset{(0.0041)}{0.057}\, educ$

SQR=104 R^2=0.236 n=617

Equació per a l'home $\ln(wage) = \underset{(0.031)}{1.739} + \underset{(0.0032)}{0.054}\, educ$

SQR=289 R^2=0.175 n=1383

El model restringit, que s'estima en l'Exemple 5.4, té la mateixa configuració que les equacions (5-34), però referit en aquest cas per a tota la mostra. Per tant, és la regressió agrupada (*P*) corresponent al model restringit. L'estadístic *F* pren el valor

$$F = \frac{\left[SQR_P - (SQR_F + SQR_M)\right] / k}{SQR_F + SQR_M) / (n - 2k)} = \frac{\left[433 - (104 + 289)\right] / 2}{(104 + 289) / (2000 - 2 \times 2)} = 102$$

L'estadístic *F* ha de ser, i ho és, igual al de l'Exemple 5.12. En conseqüència, les conclusions són les mateixes.

EXEMPLE 5.16 El model de determinació dels salaris és el mateix per a diferents mides d'empresa?

En altres exemples bé el terme independent o bé el pendent corresponent a la variable educació, va ser diferent per a tres diferents mides d'empresa (xicoteta, mitjana i gran). Ara considerem una equació completament diferent per a cada mida de l'empresa. Per tant, el model no restringit estarà compost per tres equacions:

$$\begin{aligned} &xicoteta: \ln(wage) = \beta_{11} + \delta_{11}\, female + \beta_{21} edu + u \\ &mitjana: \ln(wage) = \beta_{12} + \delta_{12}\, female + \beta_{22} edu + u \\ &gran \;\; : \ln(wage) = \beta_{13} + \delta_{13}\, female + \beta_{23} edu + u \end{aligned} \tag{5-37}$$

Les hipòtesis nul·la i alternativa seran les següents:

$$H_0 : \begin{cases} \beta_{11} = \beta_{12} = \beta_{13} \\ \delta_{11} = \delta_{12} = \delta_{13} \\ \beta_{21} = \beta_{22} = \beta_{23} \end{cases}$$
$$H_1 : \text{No } H_0$$

Donada aquesta hipòtesi nul·la, el model restringit és el model de (5-2).

Les estimacions de les tres equacions de (5-37), utilitzant el fitxer *wage02sp*, són les següents:

xicoteta $\ln(wage) = \underset{(0.0346)}{1.706} - \underset{(0.0312)}{0.249}\, female + \underset{(0.0038)}{0.040}\, educ$

SQR=121 R^2=0.160 n=801

$$\textit{mitjana} \qquad \ln(wage) = \underset{(0.0514)}{1.934} - \underset{(0.0390)}{0.422}\, female + \underset{(0.0046)}{0.055}\, educ$$

$$SQR = 123 \qquad R^2 = 0.302 \qquad n = 590$$

$$\textit{gran} \qquad \ln(wage) = \underset{(0.0462)}{1.749} - \underset{(0.0385)}{0.303}\, female + \underset{(0.0044)}{0.055}\, educ$$

$$SQR = 114 \qquad R^2 = 0.273 \qquad n = 609$$

La regressió agrupada (*P*) ja ha estat estimada en l'Exemple 5.1. L'estadístic *F* pren el valor:

$$F = \frac{\left[SQR_P - (SQR_S + SQR_M + SQR_L)\right] / 2 \times k}{(SQR_S + SQR_M + SQR_L) / (n - 3k)}$$
$$= \frac{\left[393 - (121 + 123 + 114)\right] / 6}{(121 + 123 + 114) / (2000 - 3 \times 3)} = 32.4$$

Per a qualsevol nivell de significació, rebutgem que les equacions per a la determinació dels salaris siguen les mateixes per als tres mides d'empresa considerats.

EXEMPLE 5.17 És el model Pinkham vàlid per als quatre períodes?

En l'Exemple 5.5 es van introduir variables fictícies temporals i es va contrastar si el terme independent era diferent per a cada període. Ara, anem a contrastar si el model en el seu conjunt és vàlid per als quatre períodes considerats. Per tant, el model restringit estarà compost per quatre equacions:

$$\begin{aligned}
&\text{1907-1914} \qquad sales_t = \beta_{11} + \beta_{21} advexp_t + \beta_{31}\, sales_{t-1} + u_t \\
&\text{1915-1925} \qquad sales_t = \beta_{12} + \beta_{22} advexp_t + \beta_{32}\, sales_{t-1} + u_t \\
&\text{1926-1940} \qquad sales_t = \beta_{13} + \beta_{23} advexp_t + \beta_{33}\, sales_{t-1} + u_t \\
&\text{1941-1960} \qquad sales_t = \beta_{14} + \beta_{24} advexp_t + \beta_{34}\, sales_{t-1} + u_t
\end{aligned} \tag{5-38}$$

Les hipòtesis nul·la i alternativa seran les següents:

$$H_0 : \begin{cases} \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} \\ \beta_{21} = \beta_{22} = \beta_{23} = \beta_{24} \\ \beta_{31} = \beta_{32} = \beta_{33} = \beta_{34} \end{cases}$$
$$H_1 : \text{No } H_0$$

Donada aquesta hipòtesi nul·la, el model restringit és el següent

$$sales_t = \beta_1 + \beta_2 advexp_t + \beta_3\, sales_{t-1} + u_t \tag{5-39}$$

Les estimacions de les quatre equacions (5-38) són les següents:

$$\text{1907-1914} \qquad sales_t = \underset{(603)}{64.84} + \underset{(1.025)}{0.9149}\, advexp + \underset{(0.425)}{0.4630}\, sales_{t-1} \qquad SQR = 36017 \quad n = 7$$

$$\text{1915-1925} \qquad sales_t = \underset{(190)}{221.5} + \underset{(0.557)}{0.1279}\, advexp + \underset{(0.425)}{0.9319}\, sales_{t-1} \qquad SQR = 400605 \quad n = 11$$

$$\text{1926-1940} \qquad sales_t = \underset{(112)}{446.8} + \underset{(0.115)}{0.4638}\, advexp + \underset{(0.0827)}{0.4445}\, sales_{t-1} \qquad SQR = 201614 \quad n = 15$$

$$\text{1941-1960} \qquad sales_t = -\underset{(134)}{182.4} + \underset{(0.241)}{1.6753}\, advexp + \underset{(0.111)}{0.3042}\, sales_{t-1} \qquad SQR = 187332 \quad n = 20$$

La regressió agrupada, estimada en l'Exemple 3.4, és la següent:

$$sales_t = \underset{(95.7)}{138.7} + \underset{(0.156)}{0.3288}\, advexp + \underset{(0.0915)}{0.7593}\, sales_{t-1} \qquad SQR = 2527215 \quad n = 53$$

L'estadístic *F* pren el valor

$$F = \frac{\left[SCR_P - (SCR_1 + SCR_2 + SCR_3 + SCR_4)\right]/3\times k}{(SCR_1 + SCR_2 + SCR_3 + SCR_4)/(n-4k)}$$
$$= \frac{\left[2527215 - (36017 + 400605 + 201614 + 187332)\right]/9}{(36017 + 400605 + 201614 + 187332)/(53 - 4\times 3)} = 9.16$$

Per a qualsevol nivell de significació, rebutgem que el model (5-39) siga el mateix per als quatre períodes considerats.

Exercicis

Exercici 5.1 Responeu a les dues següents qüestions relatives a un model amb variables explicatives fictícies:

a) Quina és la interpretació dels coeficients de les variables fictícies?

b) Per què no s'han d'incloure el mateix nombre de variables fictícies que categories?

Exercici 5.2 S'han obtingut les següents estimacions de demanda d'habitatges per a lloguer amb una mostra de 560 famílies.

$$\hat{q}_i = \underset{(0.11)}{4.17} - \underset{(0.017)}{0.247}\, p_i + \underset{(0.026)}{0.960}\, y_i$$

$$R^2=0.371 \qquad n=560$$

$$\hat{q}_i = \underset{(0.13)}{5.27} - \underset{(0.030)}{0.221}\, p_i + \underset{(0.031)}{0.920}\, y_i + \underset{(0.120)}{0.341} d_i y_i$$

$$R^2=0.380$$

on q_i és el logaritme de la despesa en lloguer d'habitatge de la família i-èsima, p_i és el logaritme del preu de lloguer per m^2 a l'àrea que viu la família i-èsima, y_i és el logaritme de la renda familiar disponible *i*-èsima i d_i és una variable fictícia que pren el valor un si la família resideix en un municipi urbà i zero en un rural.

(Els números entre parèntesis són els errors estàndard dels estimadors.).)

a) Contrast, en el primer model ajustat, la hipòtesi que l'elasticitat de la despesa en lloguer d'habitatge pel que fa a la renda es 1.

b) Contrast si la interacció entre la variable fictícia i la renda és significativa. Existeix una diferència significativa de l'elasticitat despesa de lloguer renda entre les àrees rurals i urbanes?

Exercici 5.3 En un model de regressió lineal amb variables fictícies contesti a les següents preguntes:

a) Significat i interpretació dels coeficients de les variables fictícies en models amb diferents formes funcionals de la variable endògena.

b) Per què no és convenient incloure el mateix nombre de fictícies que de categories existents a la variable qualitativa?

c) Expresse com es veu afectat un model en el qual s'han introduït variables fictícies en forma additiva i un altre en què només

s'introdueixen en forma multiplicativa respecte a una variable quantitativa.

Exercici 5.4 En el context del model de regressió lineal múltiple,

a) Què és una variable fictícia? Posi un exemple d'especificació d'un model economètric amb variables fictícies. Interpretació dels coeficients, raonant la resposta.
b) Quina relació pot existir entre el problema de multicolinealitat i les variables fictícies?

Exercici 5.5 Amb dades corresponents als treballadors d'un departament d'una certa empresa s'ha obtingut la següent estimació:

$$salari_i = 500 + 50 antigüedat_i + 200 nivelldeestudis_i + 100 home_i$$

on *salari* és el salari en euros mensuals, *antiguitat* és l'antiguitat laboral mesura en anys, *nivellestudis* és una variable fictícia que pren valor 1 si el treballador té estudis superiors i 0 en cas contrari i *home* és una variable fictícia que pren el valor 1 si el treballador és home i 0 en cas contrari.

a) Quin salari prediria per a una treballadora amb 6 anys d'antiguitat laboral i amb estudis superiors?
b) Suposant que totes les dones treballadores tenen estudis superiors i cap dels homes treballadors tenen estudis superiors, escriviu una hipotètica matriu de regressors (**X**) per a sis observacions. En aquest cas, ¿es plantejaria algun problema en l'estimació del model? Expliqueu la seua resposta.
c) Plantege un nou model economètric que permeta dilucidar si hi ha diferències salarials entre els treballadors amb estudis primaris, amb estudis mitjans i amb estudis superiors.

Exercici 5.6 Considere el model de regressió lineal:

$$y_i = \alpha + \beta x_i + \gamma_1 d_{1i} + \gamma_2 d_{2i} + u_i \qquad (1)$$

on *y* és el salari mensual d'un professor, x és el nombre d'anys d'experiència docent i d_1 i d_2 són dues variables fictícies que prenen els següents valors

$$d_{1i} = \begin{cases} 1 \text{ si el professor és home} \\ 0 \text{ en tots els altres casos} \end{cases} \qquad d_{2i} = \begin{cases} 1 \text{ si el professor és de raça blanca} \\ 0 \text{ en tots els altres casos} \end{cases}$$

a) Quina és la categoria de referència en el model?
b) Interprete el significat de γ_1 i γ_2. Quin és el salari esperat per a totes les categories possibles?

Per millorar la capacitat explicativa del model es va considerar la següent especificació alternativa

$$y_i = \alpha + \beta x_i + \gamma_1 d_{1i} + \gamma_2 d_{2i} + \gamma_3 (d_{1i} d_{2i}) + u_i \qquad (2)$$

c) Quin és el significat del terme $(d_{1i} d_{2i})$? Interpreteu el significat de γ_3.

d) Quin és el salari esperat per a totes les categories possibles en el model (2)?

Exercici 5.7 S'ha obtingut la següent equació estimada per mínims quadrats ordinaris amb una mostra de 36 observacions:

$$\hat{y}_t = \underset{(0.12)}{1.10} - \underset{(0.34)}{0.96}\, x_{t1} - \underset{(3.35)}{4.56}\, x_{t2} + \underset{(0.07)}{0.34}\, x_{t3}$$

$$\sum_{t=1}^{n}\left(\hat{y}_t - \bar{y}\right)^2 = 109.24 \quad \sum_{t=1}^{n}\hat{u}_t^2 = 20.22$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Contrast la significativitat individual del coeficient associat a x_2.
b) Calcule el coeficient de determinació, R^2, i done una interpretació del mateix.
c) Contrast la significativitat conjunta del model.
d) Dos regressions addicionals, amb la mateixa especificació, van ser realitzades per als dos grups, A i B, inclosos en la mostra (n_1=21 i n_2=15). En aquestes estimacions es van obtenir les següents *SCR*, 11.09 i 2.17, respectivament. Contrast si els grups A i B tenen un diferent comportament.

Exercici 5.8 Per explicar el temps dedicat a activitats esportives (*esport*) s'ha formulat el següent model:

$$esport = \beta_1 + \delta_1 dona + \varphi_1\, fumador + \beta_2 edat + u \qquad (1)$$

on *esport* són els minuts dedicats al dia, de mitjana, a activitats esportives en minuts; *dona* i *fumador* són variables fictícies que prenen el valor 1 si la persona és una dona o si fuma almenys 5 cigarrets diaris, respectivament. La variable edat està expressada en anys.

a) Interprete el significat de δ_1, φ_1 i β_2.
b) Quin és el temps esperat dedicat a activitats esportives per a totes les categories possibles?
c) Per millorar la capacitat explicativa del model es va considerar la següent especificació alternativa:

$$\begin{aligned} depor = {} & \beta_1 + \delta_1 mujer + \varphi_1\, fumador + \gamma_1 mujer \times fumador \\ & + \delta_2 mujer \times edad + \varphi_2\, fumador \times edad + \beta_2 edad + u \end{aligned} \qquad (2)$$

En el model (2), quin és el significat de γ_1? Quin és el significat de δ_2 i φ_2?

d) Quins són els possibles efectes marginals d'*esport* pel que fa a l'*edat* en el model (2)? Detalle'ls.

Exercici 5.9 Utilitzant informació de les regions espanyoles en els anys 1995 i 2000 s'han estimat diverses funcions de producció.

Per al conjunt dels dos períodes es van obtenir els següents resultats

$$\ln(q) = 5.72 + 0.26\ln(k) + 0.75\ln(l) - 1.14f + 0.11f \times \ln(k) - 0.05f \times \ln(l) \qquad (1)$$

$$R^2 = 0.9594 \quad \bar{R}^2 = 0.9510 \quad SCR = 0.9380 \quad n = 34$$

$$\ln(q) = 3.91 + 0.45\ln(k) + 0.60(l) \qquad (2)$$

$$R^2 = 0.9567 \quad \bar{R}^2 = 0.9525 \quad SQR = 1.0007$$

D'altra banda, per a cada un dels anys es van estimar separadament els següents models:

1995 $$\ln(q) = 5.72 + 0.26\ln(k) + 0.75l \qquad (3)$$

$$R^2 = 0.9527 \quad \bar{R}^2 = 0.9459 \quad SCR = 0.6052$$

2000 $$\ln(q) = 4.58 + 0.37\ln(k) + 0.70l \qquad (4)$$

$$R^2 = 0.9629 \quad \bar{R}^2 = 0.9555 \quad SQR = 0.3331$$

on q és producció, k és capital, l és ocupació i f és una variable fictícia que pren el valor 1 per a les dades de 1995 i 0 per als de l'any 2000.

a) Contrast si es produeix un canvi estructural entre 1995 i 2000.
b) Compare els resultats de les estimacions (3) i (4) amb l'estimació (1).
c) Contrast la significativitat global del model (1).

Exercici 5.10 Amb una mostra de 300 empreses del sector de serveis, es va estimar la següent funció de *cost*:

$$cost_i = 0.847 + \underset{(0.025)}{0.899}\, qty_i \quad SCR = 901.074 \quad n = 300$$

on qty_i és la quantitat produïda.

Les 300 empreses estan distribuïdes en tres grans àrees (100 en cadascuna). Els resultats obtinguts van ser els següents:

Àrea 1: $cost_i = 1.053 + \underset{(0.038)}{0.876}\, qty_i \qquad \hat{\sigma}^2 = 0.457$

Àrea 2: $cost_i = 3.279 + \underset{(0.096)}{0.835}\, qty_i \qquad \hat{\sigma}^2 = 3.154$

Àrea 3: $cost_i = 5.279 + \underset{(0.10)}{0.984}\, qty_i \qquad \hat{\sigma}^2 = 4.255$

a) Calculeu una estimació no esbiaixada σ^2 de la funció de costos per al conjunt de les 300 empreses.
b) És la mateixa funció de cost vàlida per a les tres àrees?

Exercici 5.11 Per a l'estudi de la despesa en revistes (*rev*) s'han formulat els següents models:

$$\ln(rev) = \beta_1 + \beta_2 \ln(renda) + \beta_3 edat + \beta_4 home + u \qquad (1)$$

$$\ln(rev) = \beta_1 + \beta_2 \ln(renda) + \beta_3 edat + \beta_4 home + \beta_5 prim + \beta_6 sec + u \qquad (2)$$

on *renda* és la renda disponible, *edat* és l'edat en anys, *home* és una variable dicotòmica que pren el valor 1 si és home, *prim* i són variables fictícies que prenen el valor 1 quan l'individu ha assolit, com a molt, els nivells primaris i secundaris d'estudis, respectivament.

Amb una mostra de 100 observacions, s'han obtingut els següents resultats:

$$\ln(rev)_i = \underset{(0.124)}{1.27} + \underset{(0040)}{0.756} \ln(renda_i) + \underset{(0.001)}{0.031} edat_i - \underset{(0.022)}{0.017} home_i$$

$$SQR=1.1575 \qquad R^2=0.9286$$

$$\ln(rev_i) = \underset{(0.020)}{1.26} + \underset{(0.007)}{0.811} \ln(renda_i) + \underset{(0.0002)}{0.030} edat_i + \underset{(0.003)}{0.003} home_i - \underset{(0.004)}{0.250} prim_i + \underset{(0.005)}{0.108} sec_i$$

$$SQR=0.0306 \qquad R^2=0.9981$$

a) És l'educació un factor rellevant per explicar la despesa en revistes? Quina és la categoria de referència per a l'educació?
b) En el primer model, ¿és més gran la despesa en revistes per a homes que per a dones? Justifique la resposta.
c) Interprete el coeficient de la variable home en el segon model. És major la despesa en revistes per a homes que per a dones? Compare amb el resultat obtingut en la part *a)*.

Exercici 5.12 Considerem que fruit és la despesa en fruites en un any, expressat en euros, realitzat per una llar i r_1, r_2, r_3, i r_4 són variables dicotòmiques que reflecteixen les quatre regions d'un país.

a) Si es realitza una regressió de fruit sobre r_1, r_2, r_3, i r_4 sense terme independent, quina és la interpretació dels coeficients?
b) Si es realitza una regressió de fruit sobre r_1, r_2, r_3, i r_4 i amb un terme independent, què passaria? Per què?
c) Si es realitza una regressió de fruit sobre r_2, r_3, i r_4 sense terme independent, quina és la interpretació dels coeficients?
d) Si es realitza una regressió de fruit sobre r_1-r_2, r_2, r_4-r_3, l r_4 sense terme independent, quina és la interpretació dels coeficients?

Exercici 5.13 Considere el següent model

$$salario = \beta_1 + \delta_1 dona + \beta_2 educ + u$$

Ara, considerarem tres possibilitats de definir la variable fictícia *dona*:

$$1)\ dona = \begin{cases} 1 & \text{per a dona} \\ 0 & \text{per a home} \end{cases} \quad 2)\ dona = \begin{cases} 2 & \text{per a dona} \\ 1 & \text{per a home} \end{cases} \quad 3)\ dona = \begin{cases} 2 & \text{per a dona} \\ 0 & \text{per a home} \end{cases}$$

a) Interprete el coeficient de la variable fictícia per a cada definició.
b) És alguna definició preferible a les altres? Justifique la resposta.

Exercici 5.14 Es considera el següent model de regressió:

$$salari = \beta_1 + \delta_1 dona + u$$

on *dona* és una variable dicotòmica que pren el valor 1 per a les dones i el valor 0 per als homes.

Demostre que aplicant les fórmules de *MQO* per a la regressió simple s'obté que

$$\hat{\beta}_1 = \overline{salari_H}$$

$$\hat{\delta}_1 = \overline{salari_D} - \overline{salari_H}$$

on *D* indica dona i *H* home.

Per tal de facilitar l'obtenció de la solució, considere que en la mostra hi ha n_1 dones i n_2 homes: la mostra total és $n= n_1+n_2$.

Exercici 5.15 Les dades d'aquest exercici es van obtenir d'un experiment de màrqueting controlat a les botigues a París sobre la despesa en cafè, publicat per CA Bemmaor i Mouchoux D., "Measuring the Short-Term Effect of In-Store Promotion and Retail Advertising on Brand Sales: A Factorial Experiment ", *Journal of Marketing Research*, 28 (1991), 202-14. En aquest experiment es va formular el següent model per explicar la quantitat venuda de cafè per setmana:

$$\ln(coffqty) = \beta_1 + \delta_1 advert + \beta_2 \ln(coffpric) + \delta_2 advert \times \ln(coffpric) + u$$

on *coffpric* pren tres valors: 1, que és el preu habitual, 0.95 i 0.85; *advert* és una variable dicotòmica que pren valor 1 si es fa publicitat en aquesta setmana, i 0 si no es fa. L'experiment va durar 18 setmanes. El model original i tres models més van ser estimats, utilitzant el fitxer *coffee2*:

1)
$$\ln(coffqty_i) = \underset{(0.04)}{5.85} + \underset{(0.099)}{0.2565}\, advert_i - \underset{(0.450)}{3.9760} \ln(coffpric_i) - \underset{(0.883)}{1.069}\, advert_i \times \ln(coffpric_i)$$
$$R^2 = 0.9468 \qquad n = 18$$

2)
$$\ln(coffqty_i) = \underset{(0.04)}{5.83} + \underset{(0.057)}{0.3559}\, advert_i - \underset{(0.393)}{4.2539} \ln(coffpric_i)$$
$$R^2 = 0.9412 \quad n = 18$$

3)
$$\ln(coffqty_i) = \underset{(0.04)}{5.88} - \underset{(0.513)}{3.6939}\ln(coffpric_i) - \underset{(0.582)}{2.9575}\,advert_i \times \ln(coffpric_i)$$
$$R^2 = 0.9214 \quad n = 18$$

4)
$$\ln(coffqty_i) = \underset{(0.07)}{5.89} - \underset{(0.674)}{5.1727}\ln(coffpric)$$
$$R^2 = 0.7863 \quad n = 18$$

a) En el model (2), ¿quina és la interpretació del coeficient d'*advert*?
b) En el model (3), quina és la interpretació del coeficient d'*advert*×ln *(coffpric)*?
c) En el model (2), té el coeficient d'*advert* un efecte positiu significatiu al 5% i l'1%?
d) És el model (4) vàlid per setmanes amb publicitat i per setmanes sense publicitat?
e) En el model (1), és el terme independent el mateix per setmanes amb publicitat i per setmanes sense publicitat?
f) En el model (3), és l'elasticitat de demanda de cafè/preu diferent en setmanes amb publicitat i en setmanes sense publicitat?
g) En el model (4), és l'elasticitat de demanda de cafè/preu inferior a -4?

Exercici 5.16 (Continuació del Exercici 4.39). Utilitzant el fitxer *timuse03*, s'han estimat els següents models:

$$housework_i = \underset{(23)}{132} + \underset{(1.497)}{2.787}\,educ_i + \underset{(0.308)}{1.847}\,age_i - \underset{(0.023)}{0.2337}\,paidwork_i \tag{1}$$
$$R^2 = 0.142 \quad n = 1000$$

$$housework_i = -\underset{(22.29)}{3.02} + \underset{(1.356)}{3.641}\,educ_i + \underset{(0.279)}{1.775}\,age_i - \underset{(0.021)}{0.1568}\,paidwork_i + \underset{(2.16)}{32.11}\,female_i$$
$$R^2 = 0.298 \quad n = 1000 \tag{2}$$

$$housework_i = -\underset{(35.18)}{8.04} + \underset{(2.352)}{4.847}\,educ_i + \underset{(0.502)}{1.333}\,age_i - \underset{(0.032)}{0.0871}\,paidwork_i + \underset{(8.15)}{32.75}\,female_i$$
$$-\underset{(0.546)}{0.1650}\,educ_i \times female_i + \underset{(0.112)}{0.1019}\,age_i \times female_i - \underset{(0.009)}{0.02625}\,paidwork_i \times female_i$$
$$R^2 = 0.306 \quad n = 1000 \tag{3}$$

a) En el model (1), hi ha una compensació estadísticament significativa entre el temps dedicat a treball remunerat i el temps dedicat a treball domèstic?
b) Mantenint igual tots els altres factors i prenent com a model de referència a (2), hi ha evidència que les dones dediquen més temps al treball domèstic que els homes?
c) Compare el R^2 dels models (1) i (2). Quina és la seva conclusió?
d) En el model (3), quin és l'efecte marginal del temps dedicat al treball domèstic que fa al temps dedicat al treball remunerat?
e) És significativa la interacció entre *paidwork* i *gènere*?
f) Són les interaccions entre gènere i les variables quantitatives del model conjuntament significatives?

Exercici 5.17 Utilitzant dades de la Borsa de Madrid del 19 de novembre de 2011 (fitxer *bolmad11*), s'han estimat els següents models:

$$\ln(marktval_i) = \underset{(0.243)}{1.784} + \underset{(0.179)}{0.6998}\, ibex35_i + \underset{(0.0369)}{0.6749}\ln(bookval_i) \qquad (1)$$

$$SQR=35.69 \quad R^2=0.8931 \quad n=92$$

$$\begin{aligned}\ln(marktval_i) = \underset{(0.275)}{1.828} + \underset{(0.778)}{0.4236}\, ibex35_i + \underset{(0.0423)}{0.6678}\ln(bookval_i) \\ + \underset{(0.088)}{0.0310}\, ibex35_i \times \ln(bookval_i)\end{aligned} \qquad (2)$$

$$SQR=35.622 \quad R^2=0.8933 \quad n=92$$

$$\begin{aligned}\ln(marktval_i) = \underset{(0.310)}{2.323} + \underset{(0.785)}{0.1987}\, ibex35_i + \underset{(0.0405)}{0.6688}\ln(bookval_i) \\ + \underset{(0.089)}{0.0369}\, ibex35_i \times \ln(bookval_i) - \underset{(0.236)}{0.6613}\, services_i - \underset{(0.221)}{0.6698}\, consump_i \\ - \underset{(0.263)}{0.1931}\, energy_i - \underset{(0.207)}{0.3895}\, industry_i - \underset{(0.324)}{0.7020}\, itt_i\end{aligned} \qquad (3)$$

$$SQR=30.781 \quad R^2=0.9078 \quad n=92$$

$$\ln(marktval_i) = \underset{(0.234)}{1.366} + \underset{(0.0305)}{0.7658}\ln(bookval_i) \qquad (4)$$

$$SQR=41.625 \quad R^2=0.8753 \quad n=92$$

Per *finance*=1 $$\ln(markval_i) = \underset{(0.560)}{0.558} + \underset{(0.0702)}{0.9346}\ln(bookval_i) \qquad (5)$$

$$SQR=2.7241 \quad R^2=0.9415 \quad n=13$$

on

- *marktval* és el valor de mercat d'una companyia.
- *bookval* és el valor de comptable d'una companyia.
- *ibex35* és una variable fictícia que pren el valor 1 si la companyia està inclosa en el selectiu Ibex 35.
- *services*, *consump* (*consum*), *energy*, *industry* i *itc* (tecnologies de la informació i la comunicació) són variables fictícies. Cada un d'elles pren el valor 1 si la companyia està classificada en aquest sector a la Borsa de Madrid. La categoria de referència és el sector financer (*finance*).

a) En el model (1), ¿quina és la interpretació del coeficient d'*ibex-35?*
b) En el model (1), és l'elasticitat *marktval/bookval* igual a 1?
c) En el model (2), és l'elasticitat *marktval/bookval* la mateixa per a totes les companyies incloses en la mostra?
d) És el model (4) vàlid tant per a les companyies incloses en l'*ibex 35* i per a les companyies excloses?
e) En el model (3), quina és la interpretació del coeficient de *consump*?
f) És el coeficient de *consump* significativament negatiu?
g) Està justificada estadísticament la introducció de variables fictícies per als diferents sectors?

h) És l'elasticitat *marktval/bookval* per al sector financer igual a 1?

Exercici 5.18 (Continuació del Exercici 4.37). Utilitzant el fitxer *rdspain*, s'han estimat les equacions que apareixen en el quadre adjunt

Les variables que apareixen en el quadre són les següents:

- *rdintens* és la despesa en recerca i desenvolupament (R+D) mesurat com a percentatge de les vendes,
- *sales*, vendes mesures en milions d'euros,
- *exponsal* són les exportacions mesures com a percentatge de les vendes,
- *medtech* i *hightech* són dues variables fictícies que reflecteixen si l'empresa pertany a un sector de mitjana o d'alta tecnologia. La categoria de referència correspon a les empreses de baixa tecnologia.
- *workers* és el nombre de treballadors de l'empresa.

	(1) *rdintens*	(2) *rdintens*	(3) *rdintens*	(4) *rdintens* per a *hightech*=1	(5) *rdintens* per a *medtech*=1	(6) *rdintens* per a *lowtech*=1
exponsal	0.0136 (0.00195)	0.0101 (0.00193)	0.00968 (0.00189)	0.00584 (0.00792)	0.0116 (0.00300)	0.00977 (0.00169)
workers	0.000433 (0.0000740)	0.000392 (0.0000725)	0.000394 (0.000208)	0.00196 (0.000338)	0.0000563 (0.0000815)	0.000393 (0.000121)
hightech		1.448 (0.141)	0.976 (0.151)			
medtech		0.361 (0.109)	0.472 (0.112)			
hightech× *workers*			0.00153 (0.000271)			
medtech× *workers*			-0.000326 (0.000222)			
terme independent	0.394 (0.0598)	0.137 (0.0691)	0.143 (0.0722)	1.211 (0.313)	0.577 (0.103)	0.142 (0.0443)
n	1983	1983	1983	296	616	1071
R^2	0.0507	0.0986	0.138	0.113	0.0278	0.0459
SQR	9282.7	8815.0	8425.3	4409.0	2483.6	1527.5
F	52.90	54.06	52.90	18.71	8.776	25.72
df_n	2	4	6	2	2	2
df_d	1980	1978	1976	293	613	1068

(Errors estàndard entre parèntesis)

a) En el model (2), mantenint-se igual tots els altres factors, hi ha evidència que la despesa en recerca i desenvolupament (expressat com un percentatge de les vendes) en empreses d'alta tecnologia siga més gran que en empreses de baixa tecnologia? És fort l'evidència empírica?

b) En el model (2), mantenint-se igual tots els altres factors, hi ha evidència que la despesa en R+D, *rdintens*, en les empreses de tecnologia mitjana siga igual al d'empreses de baixa tecnologia? És fort l'evidència empírica?

c) Prenent com a model de referència (2), si hagués de contrastar la hipòtesi que *rdintens* a les empreses d'alta tecnologia és igual a les empreses de tecnologia mitjana, formule un model que li permeta contrastar aquesta hipòtesi sense necessitat d'utilitzar la informació sobre la matriu de covariances dels estimadors.

d) Hi ha influència dels treballadors associats a *rdintens* amb el nivell de tecnologia a les empreses?

e) És el model (1) vàlid per a totes les empreses independentment del seu nivell tecnològic?

Exercici 5.19 Per explicar la satisfacció general de les persones (*stsfglo*) es van estimar els següents models utilitzant dades del fitxer *hdr2010*:

$$\begin{gathered} stsfglo_i = \underset{(0.584)}{-0.375} + \underset{(0.00000617)}{0.0000207}\, gnipc_i + \underset{(0.009)}{0.0858}\, lifexpec_i \\ R^2 = 0.642 \quad n = 144 \end{gathered} \tag{1}$$

$$\begin{gathered} stsfglo_i = \underset{(0.897)}{2.911} + \underset{(0.00000572)}{0.0000381}\, gnipc_i + \underset{(0.18)}{1.215}\, lifexpec_i \\ + \underset{(0.179)}{1.215}\, dlatam_i - \underset{(0.259)}{0.7901}\, dafrica_i \\ R^2 = 0.748 \quad n = 144 \end{gathered} \tag{2}$$

$$\begin{gathered} stsfglo_i = \underset{(1.146)}{0.6984} + \underset{(0.000006)}{0.0000198}\, gnipc_i + \underset{(0.0164)}{0.0724}\, lifexpec_i + \underset{(1.950)}{4.099}\, dafrica_i \\ + \underset{(0.000052)}{0.0000801}\, gnipc_i \times dafrica_i - \underset{(0.0336)}{0.0896}\, lifexpec_i \times dafrica_i \\ R^2 = 0.6840 \quad n = 144 \end{gathered} \tag{3}$$

on

- *gnipc* és el producte nacional brut per càpita expressat en *PPA* (paritat de poder adquisitiu) en dòlars americans de 2008,
- *lifexpec* és l'esperança de vida en néixer, és a dir, el nombre d'anys que un nadó pot esperar viure,
- *dafrica* és una variable dicotòmica que pren el valor 1 si el país es troba a l'Àfrica
- *dlatam* és una variable dicotòmica que pren el valor 1 si el país està a Amèrica Llatina.

a) En el model (2), quina és la interpretació dels coeficients de *dlatam* i *dafrica*?

b) En el model (2), *dlatam* i *dafrica*, individualment, tenen una influència significativament positiva sobre la satisfacció global?
c) En el model (2), *dlatam* i *dafrica* ¿tenen una influència conjunta sobre la satisfacció global?
d) És la influència de l'esperança de vida sobre la satisfacció global menor a Àfrica que en altres regions del món?
e) És la influència de la variable de *gnipc* major a Àfrica que en altres regions del món en un 10%?
f) Són les interaccions de les persones que viuen a l'Àfrica i les variables *gnipc* i *lifexpec* conjuntament significatives?

Exercici 5.20 Les equacions que apareixen en el quadre adjunt s'han estimat utilitzant les dades del fitxer timuse03. Aquest fitxer conté 1000 observacions corresponents a una submostra aleatòria extreta de l'enquesta d'ús del temps a Espanya que es va dur a terme en el període 2002-2003.

Les variables que apareixen en el quadre són:

- *educ* són els anys d'educació assolits,
- *sleep* (dormir), *paidwork* (treball remunerat) and *unpaidwrk* (treball no remunerat es mesuren en minuts per dia,
- *female* (dona), *workday* (dilluns a divendres), spaniard (espanyol) i *houswife* (mestressa de casa) són variables fictícies.

a) En el model (1), hi ha una compensació estadísticament significativa entre el temps dedicat al treball remunerat i el temps dedicat a dormir?
b) En el model (1), és el coeficient d'*unpaidwk* estadísticament significatiu?
c) Hi ha evidència que les dones dormen més que els homes?
d) En el model (2), són *workday* i *spaniard* individualment significatives? Són conjuntament significatives?
e) És el coeficient de *housewife* estadísticament significatiu?
f) Són les interaccions de *female* amb *educ*, *paidwork* i *unpaidwk* conjuntament significatives?

	(1) *Sleep*	*(2)* *Sleep*	*(3)* *Sleep*	*(4)* *Sleep*	*(5)* *Sleep*	*(6)* *sleep*
educ	-4.669	-4.787	-4.805	-4.754	-4.782	-4.792
	(0.916)	(0.912)	(0.912)	(0.913)	(0.917)	(0.917)
persinc	0.0238	0.0207	0.0195	0.0210	0.0208	0.0208
	(0.00587)	(0.00600)	(0.00607)	(0.00601)	(0.00601)	(0.00601)
age	0.854	0.879	0.895	0.884	0.879	0.891
	(0.174)	(0.174)	(0.174)	(0.174)	(0.174)	(0.302)
paidwork	-0.258	-0.247	-0.246	-0.248	-0.246	-0.247
	(0.0150)	(0.0159)	(0.0159)	(0.0160)	(0.0210)	(0.0159)
unpaidwk	-0.205	-0.198	-0.188	-0.224	-0.198	-0.198
	(0.0184)	(0.0184)	(0.0196)	(0.0365)	(0.0185)	(0.0184)
female	4.161	3.588	3.981	2.485	3.638	3.727
	(1.465)	(1.467)	(1.493)	(1.975)	(1.691)	(3.287)
workday		-19.31	-19.46	-19.47	-19.30	-19.30
		(7.168)	(7.165)	(7.171)	(7.173)	(7.172)
spaniard		-47.50	-46.88	-47.90	-47.63	-47.51
		(19.99)	(19.98)	(20.00)	(20.10)	(20.00)
houswife			-14.71			
			(10.42)			
unpaidwk ×*female*				0.00607		
				(0.00726)		
paidwork ×***female***					-0.000324	
					(0.00540)	
age× *female*						-0.00308
						(0.0652)
terme independent	588.9	648.3	646.6	651.9	648.2	647.8
	(13.62)	(24.34)	(24.36)	(24.73)	(24.39)	(26.40)
N	1000	1000	1000	1000	1000	1000
R^2	0.316	0.325	0.326	0.325	0.325	0.325
SQR	9913901.3	9789312.3	9769648.2	9782424.0	9789276.9	9789290.3
F	76.58	59.62	53.27	53.06	52.95	52.95
df_n	6	8	9	9	9	9
df_d	993	991	990	990	990	990

Errors estàndar entre paréntesis

Exercici 5.21 Per a l'estudi de la mortalitat infantil al món s'han estimat els següents models a partir de les dades del fitxer *hdr2010*:

$$\underset{}{deathinf_i} = \underset{(4.58)}{93.02} - \underset{(0.0002)}{0.00037}\, gnipc_i - \underset{(0.1866)}{0.6046}\, physicn_i - \underset{(0.003)}{0.003}\, contrcep_i \qquad (1)$$

$$SQR=40285 \quad R^2=0.6598 \quad n=108$$

$$\begin{gathered} deathinf_i = \underset{(5.96)}{78.55} - \underset{(0.0002)}{0.00042}\, gnipc - \underset{(0.1879)}{0.3809}\, physicn_i - \underset{(0.1042)}{0.6989}\, contrcep_i \\ + \underset{(5.05)}{17.92}\, dafrica \end{gathered} \tag{2}$$

$$SQR=35893 \quad R^2=0.6851 \quad n=108$$

$$\begin{gathered} deathinf_i = \underset{(6.76)}{72.58} - \underset{(0.0002)}{0.00044}\, gnipc - \underset{(0.1879)}{0.3994}\, physicn_i - \underset{(0.1234)}{0.5857}\, contrcep_i \\ + \underset{(5.05)}{17.92}\, dafrica - \underset{(0.000826)}{0.0000914}\, gnipc \times dafrica - \underset{(2.2351)}{2.0013}\, physicn \times dafrica \\ - \underset{(0.2716)}{0.2172}\, contrcep_i \times dafrica \end{gathered} \tag{3}$$

$$SQR=34309 \quad R^2=0.7109 \quad n=108$$

on

- *deatinf* és el nombre de morts infantils (d'un any o menys) per cada 1000 nascuts vius el 2008,
- *gnipc* és el producte nacional brut per càpita expressat en PPA en dòlars americans de 2008,
- *physicn* són els metges per cada 10000 habitants en el període 2000-2009,
- *contrcep* és la taxa d'ús d'anticonceptius de qualsevol tipus, expressada com % de les dones casades de 15-49 anys per al període 1990-2008,
- *dafrica* és una variable dicotòmica que pren el valor 1 si el país es troba a l'Àfrica.

a) En el model (1), quin és la interpretació dels coeficients de *gnipc*, *physicn* i *contrcep*?
b) En el model (2), ¿quina és la interpretació del coeficient de *dafrica*?
c) En el model (2), mantenint igual tots els altres factors, ¿tenen els països d'Àfrica una mortalitat infantil més gran que els països d'altres regions del món?
d) Quin és l'efecte marginal de la variable *gnipc* sobre la mortalitat infantil al model (3)?
e) És el pendent corresponent a l'regressor *contrcep* significativament més gran per als països de l'Àfrica?
f) Són els pendents corresponents als regressors *gnipc*, *physicn* i *contrcep* conjuntament diferents per als països de l'Àfrica?
g) És el model (1) vàlid per a tots els països del món?

Exercici 5.22 Utilitzant una submostra aleatòria de 2000 observacions extretes de les enquestes d'ús del temps per a Espanya dutes a terme en els períodes 2002-2003 i 2009-2010 (fitxer timus309), s'han estimat els següents models per explicar el temps que es passa veient la televisió:

$$\begin{gathered} watchtv = \underset{(9.46)}{114} - \underset{(0.620)}{3.523}\, educ + \underset{(0.129)}{1.330}\, age - \underset{(0.010)}{0.1111}\, paidwork \\ R^2 = 0.169 \quad n = 2000 \end{gathered} \tag{1}$$

$$watchtv = \underset{(9.92)}{127} - \underset{(0.615)}{3.653}\,educ + \underset{(0.129)}{1.291}\,age - \underset{(0.010)}{0.120}\,paidwork - \underset{(4.903)}{25.15}\,female + \underset{(5.25)}{17.14}\,y2009 \qquad R^2 = 0.184 \quad n = 2000 \tag{2}$$

$$watchtv = \underset{(10.01)}{123} - \underset{(0.615)}{3.583}\,educ + \underset{(0.129)}{1.302}\,age - \underset{(0.012)}{0.1053}\,paidwork - \underset{(4.90)}{24.87}\,female + \underset{(6.115)}{24.54}\,y2009 - \underset{(0.021)}{0.0501}\,y2009 \times paidwork \qquad R^2 = 0.186 \quad n = 2000 \tag{3}$$

on

- *educ* són els anys d'educació assolits,
- *watchtv* (veure televisió) i *paidwork* (treball remunerat) es mesuren en minuts per dia,
- *female* (dona) és una variable fictícia que pren valor 1 si l'entrevistat és una dona,
- *y2009* és una variable fictícia que pren valor 1 si l'enquesta es va dur a terme en el bienni 2008-2009.

a) En el model (1), ¿quina és la interpretació del coeficient d'*educ*?
b) En el model (1), hi ha una compensació estadísticament significativa entre el temps dedicat a la feina i el temps dedicat a veure la televisió?
c) Mantenint igual tots els altres factors i prenent com a model (2) com a referència, hi ha evidència que els homes veuen la televisió més que les dones? És fort aquesta possible evidència?
d) En el model (2), ¿quina és la diferència estimada del temps dedicat a veure la televisió entre les dones enquestades en 2008-2009 i els homes enquestats al període 2002-2003? És aquesta diferència estadísticament significativa?
e) En el model (3), quin és l'efecte marginal del temps dedicat al treball remunerat sobre el temps dedicat a veure televisió?
f) Hi ha una interacció significativa entre l'any de l'enquesta i el temps dedicat al treball remunerat?

Exercici 5.23 Utilitzant el fitxer *consumsp*, es van estimar els següents models per analitzar si l'entrada d'Espanya a la Comunitat Europea el 1986 va tenir algun impacte en el comportament dels consumidors espanyols:

$$conspc_t = -\underset{(84.88)}{7.156} + \underset{(0.0857)}{0.3965}\,incpc_t + \underset{(0.0903)}{0.5771}\,conspc_{t-1} \tag{1}$$

$$R^2=0.9967 \qquad SQR=1891320 \qquad n=56$$

$$conspc_t = -\underset{(108)}{102.4} + \underset{(0.0879)}{0.3573}\,incpc_t + \underset{(0.0901)}{0.5992}\,conspc_{t-1} + \underset{(92.56)}{148.60}\,y1986_t \tag{2}$$

$$R^2=0.9968 \qquad SQR=1802007 \qquad n=56$$

$$conspc_t = \underset{(114)}{78.18} + \underset{(0.1100)}{0.5181}\,incpc_t + \underset{(0.1199)}{0.4186}\,conspc_{t-1} + \underset{(456.3)}{819.82}\,y1986_t - \underset{(0.2338)}{0.5403}\,incpc_t \times y1986_t + \underset{(0.2182)}{0.5424}\,conspc_{t-1} \times y1986_t \tag{3}$$

$$R^2=0.9972 \quad SQR=1600714 \quad n=56$$

$$\begin{aligned} conspc_t = &\underset{(118)}{117.03} + \underset{(0.0968)}{0.3697}\, incpc_t + \underset{(0.1051)}{0.5823}\, conspc_{t-1} + \underset{(348)}{41.62}\, y1986_t \\ &+ \underset{(0.0326)}{0.0104}\, incpc_t \times y1986_t \end{aligned} \tag{4}$$

$$R^2=0.9968 \quad SQR=1798423 \quad n=56$$

$$conspc_t = \underset{(114)}{120.1} + \underset{(0.0854)}{0.3750}\, incpc_t + \underset{(0.0890)}{0.5758}\, conspc_{t-1} + \underset{(0.0087)}{0.0141}\, incpc_t \times y1986_t \tag{5}$$

$$R^2=0.9968 \quad SQR=1798927 \quad n=56$$

on el consum (*conspc*) i la renda disponible (*incpc*) s'expressen en euros constants per càpita, prenent 2008 com a any de referència.

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Comprove en el model (6) si la propensió marginal al consum a curt termini es va reduir el 1986 i anys successius.
b) Són les interaccions de y1986 amb les variables quantitatives del model significatives en forma conjunta?
c) Estime si hi va haver un canvi estructural en la funció de consum el 1986 i següents anys.
d) Comprovar si el coeficient de $conspc_{t-1}$ va canviar el 1986.
e) Existeix una bretxa entre el consum que es realitzava abans de 1986 pel que fa al consum el 1986 i anys successius

6 RELAXACIÓ DELS SUPÒSITS EN EL MODEL LINEAL CLÀSSIC

6.1 Relaxació dels supòsits del *MLC*: una panoràmica

En els capítols 2 i 3 es va formular el model de regressió lineal, simple i múltiple, incloent el conjunt de supòsits estadístics anomenats supòsits del model lineal clàssic (*MLC*). Ara, anem a examinar els problemes que planteja l'incompliment de cadascun dels supòsits del *MLC*, així com els mètodes alternatius que es plantegen per a estimar el model lineal.

Supòsits sobre la forma funcional

En el supòsit 1 es postula quin és el model poblacional:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \tag{6-1}$$

Aquest supòsit especifica quina és la variable endògena i la forma funcional amb que apareix en l'equació, quines són les variables explicatives i les seues respectives formes funcionals. A més, s'estableix que el model és lineal en els paràmetres.

Quan s'estima un model poblacional diferent es comet un error d'especificació. Les conseqüències d'aquest tipus d'errors s'examinen en l'epígraf 6.2.

Supòsits sobre els regressors

Sobre els regressors es van formular els supòsits 2, 3, i 4. En el model de regressió lineal múltiple, en el supòsit 2 es postulava que els valors de $x_2, x_3, \cdots, x_k$ són fixos en repetides mostres, és a dir, els regressors són no estocàstics. Aquesta és un supòsit raonable quan els regressors s'obtenen a partir de variables controlades experimentalment. En canvi, és menys admissible en variables obtingudes mitjançant observació de caràcter passiu, com seria el cas de la renda en la funció del consum.

Quan els regressors són estocàstics, la relació estadística entre els regressors i la pertorbació aleatòria és un punt crucial en l'elaboració d'un model economètric. Per això es va formular el supòsit alternatiu 2*: els regressors $x_2, x_3, \cdots, x_k$ es distribueixen independentment de la pertorbació aleatòria. Quan assumim aquest supòsit alternatiu, la inferència, condicionada a la matriu dels regressors, porta a uns resultats que són pràcticament coincidents amb el cas en que la matriu **X** és fixa. En altres paraules, en el cas d'independència entre els regressors i la pertorbació aleatòria, el mètode de mínims

quadrats ordinaris segueix sent el mètode òptim per a l'estimació del vector de coeficients. ordinaris segueix sent el mètode òptim per a l'estimació del vector de coeficients.

En el supòsit 3 es postulava que la matriu de regressors **X** no conté errors de mesura. En el cas que els tingués es planteja un problema economètric molt greu, la solució és complexa.

El supòsit 4 estableix que no existeix relació lineal exacta entre els regressors, o, en altres paraules, estableix que no existeix multicolinealitat perfecta en el model. Aquest supòsit és necessari per al càlcul del vector d'estimadors mínims quadràtics. La multicolinealitat perfecta no se sol presentar en la pràctica. En canvi, sí que és freqüent que entre els regressors hi hagi una relació aproximadament lineal, en aquest cas els estimadors que s'obtinguin seran en general poc precisos, encara que segueixen conservant la propietat de ser estimadors *ELNEO*. En altres paraules, la relació entre regressors fa que siga difícil quantificar amb precisió l'efecte que cada regressor exerceix sobre el regressant, el que determina que les variances dels estimadors siguen elevades. Quan es presenta una relació aproximadament lineal entre els regressors, es diu que hi ha multicolinealitat no perfecta. L'epígraf 6.3 es dedica a examinar la detecció de la multicolinealitat (no perfecta), així com algunes de les possibles solucions.

Supòsit sobre els paràmetres

En el supòsit 5 es va assumir que els paràmetres $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$ són no aleatoris. L'anàlisi del món real pot suggerir que aquesta constància dels coeficients no siga raonable. Així, en els models que utilitzen dades de sèries temporals, pot quedar de manifest que al llarg del temps s'han produït canvis en els patrons de comportament, el que implicaria naturalment canvis en els coeficients de regressió. Sobre aquesta qüestió, en l'epígraf 5.6 s'ha examinat el contrast de canvi estructural que permet determinar si s'ha produït algun canvi en els paràmetres al llarg del temps.

Supòsits sobre la pertorbació aleatòria

En el supòsit 5 es va assumir que $E(\mathbf{u})=\mathbf{0}$. Aquest supòsit no és contrastable empíricament en el cas general de models amb terme independent.

Abans de passar a altres supòsits sobre la pertorbació aleatòria ui convé remarcar que aquesta és una variable no observable. La informació sobre u_i l'obtenim indirectament a través dels residus, que són els que haurem d'utilitzar per realitzar contrastos sobre el comportament de les pertorbacions. No obstant això, la utilització dels residus per a realitzar contrastos sobre les pertorbacions planteja el següent problema. Quan es compleixen els supòsits del *MLC*, les pertorbacions aleatòries són homoscedàstiques i no autocorrelacionades, però en canvi els residus són heteroscedàstics i estan autocorrelacionados, sota aquests supòsits. Aquesta circumstància s'ha de tenir en compte en el disseny dels contrastos estadístics sobre els supòsits d'homoscedasticitat i no autocorrelació.

Si no es compleixen els supòsits 7 d'homoscedasticitat i/o 8 de no autocorrelació dels estimadors obtinguts per mínims quadrats segueixen sent lineals, no esbiaixats, però no òptims.

Els supòsits d'homoscedasticitat i no autocorrelació formulades en el tema 3 es poden formular conjuntament indicant que la matriu de covariances de les pertorbacions aleatòries és una matriu escalar, és a dir,

$$E(\mathbf{uu}') = \sigma^2\mathbf{I} \tag{6-2}$$

Quan no es compleix un, o els dos, dels supòsits assenyalats, aleshores la matriu de covariances serà menys restrictiva. Així, considerarem la següent matriu de covariances de les pertorbacions:

$$E(\mathbf{uu}') = \sigma^2\mathbf{\Omega} \tag{6-3}$$

on l'única restricció que s'imposa a $\mathbf{\Omega}$ és que siga una matriu definida positiva.

Quan la matriu de covariances és una matriu no escalar, com (6-3), aleshores poden obtenir uns estimadors lineals, no esbiaixats i òptims mitjançant l'aplicació del mètode de mínims quadrats generalitzats (*MQG*). L'expressió d'aquests estimadors és la següent:

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y} \tag{6-4}$$

A la pràctica, no se sol aplicar directament la fórmula (6-4). En el seu lloc s'aplica un procediment en dues etapes, que condueix exactament als mateixos resultats.

En epígraf 6.5 s'examinaran els contrastos per determinar si existeix o no heteroscedasticitat, així com la particularització del mètode de *MQG* a aquest cas concret. En l'epígraf 6.6 s'exposaran procediments de contrast, així com el tractament de models amb pertorbacions autocorrelacionades.

El supòsit 9 de normalitat postulat en el *MLC* permet construir estadístiques per realitzar inferències amb distribucions conegudes. Si el supòsit de normalitat no és adequat, aleshores els contrastos només tindran una validesa aproximada. En l'epígraf 6.4 s'exposa un contrast de normalitat de les pertorbacions que s'utilitza per determinar si aquest supòsit és acceptable o no.

6.2 Errors d'especificació

Com hem indicat es produeix un error d'especificació quan s'estima un model diferent del model poblacional. El problema en les ciències socials, i en particular en economia, és que generalment no coneixem el model poblacional.

Tenint en compte aquesta observació, considerarem tres tipus d'errors d'especificació:

- Inclusió d'una variable irrellevant
- Exclusió d'una variable rellevant.
- Forma funcional incorrecta

6.2.1 Conseqüències de l'especificació errònia

A continuació, examinarem les conseqüències en els estimadors *MQO* de cada tipus d'especificació errònia.

Inclusió d'una variable irrellevant

Suposem que el model poblacional és el següent:

$$y = \beta_1 + \beta_2 x_2 + u \tag{6-5}$$

Per tant, la *funció de regressió poblacional* (*FRP*) - part sistemàtica d'aquest model- ve donada per

$$\mu_y = \beta_1 + \beta_2 x_2 \tag{6-6}$$

Ara suposem que la *funció de regressió mostral* (*FRP*) estimada és la següent:

$$\tilde{y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 x_{2i} + \tilde{\beta}_3 x_{3i} \tag{6-7}$$

Aquest és el cas d'inclusió d'una variable irrellevant: específicament en (6 7) hem introduït la variable irrellevant x_3. Quin són els efectes de la inclusió d'una variable irrellevant en els estimadors obtinguts per *MQO*?

Pot demostrar-se que els estimadors corresponents a (6-7) són no esbiaixats, és a dir,

$$E(\tilde{\beta}_1) = \beta_1 \qquad E(\tilde{\beta}_2) = \beta_2 \qquad E(\tilde{\beta}_3) = 0$$

No obstant això, les variances d'aquests estimadors seran més grans que les obtingudes en estimar (6-5) on s'ha omès (correctament) x_3.

Aquest resultat és generalitzable: si incloem una o més variables irrellevants, es que els estimadors *MQO* són no esbiaixats, però amb variances més grans que quan no s'inclouen variables irrellevants en el model estimat.

Exclusió d'una variable rellevant

Suposem que el model poblacional és el següent:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \tag{6-8}$$

Aleshores la *FRP* ve donada per

$$\mu_y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 \tag{6-9}$$

Ara suposem que la *FRM* estimada, causa de la nostra ignorància o a la no disponibilitat de dades, és la següent

$$\tilde{y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 x_{2i} \tag{6-10}$$

Aquest és un cas d'exclusió d'una variable rellevant: específicament en (6-10) hem omès la variable rellevant x_3. És $\tilde{\beta}_2$, obtingut mitjançant aplicació de *MQO* a (6-10), un estimador no esbiaixat de β_2?

Com es mostra en l'apèndix 6.1 l'estimador $\tilde{\beta}_2$ és esbiaixat. El biaix és

$$Bias(\tilde{\beta}_2) = \beta_3 \frac{\sum_{i=1}^{n} (x_{2i} - \bar{x}_2) x_{3i}}{\sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2} \tag{6-11}$$

Aquest biaix és nul si, d'acord amb (6-11), la covariança entre x_2 i x_3 és 0. És important advertir que la *rati*

$$\frac{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)x_{3i}}{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2}$$

és justament el pendent ($\hat{\delta}_2$) en la regressió de x_3 sobre x_2. És a dir,

$$\hat{x}_2=\hat{\delta}_1+\hat{\delta}_2\hat{x}_2=\hat{\delta}_1+\frac{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)x_{3i}}{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2}\hat{x}_2 \qquad (6\text{-}12)$$

Així doncs, d'acord amb (6-72) - en l'apèndix 6.1 i (6-12), podem dir que

$$E(\tilde{\beta}_2)=\beta_2+\beta_3\hat{\delta}_2 \qquad (6\text{-}13)$$

En conseqüència, el biaix és igual a $\beta_3\hat{\delta}_2$. En el quadre 6.1 es pot veure un resum del signe del biaix en $\tilde{\beta}_2$ quan s'omet x_2 en l'equació estimada. Per a la millor comprensió del contingut d'aquest quadre s'ha de tenir en compte que el signe de $\hat{\delta}_2$ té el mateix signe que la correlació mostral entre x_2 i x_3.

QUADRE 6.1. Resum del biaix en $\tilde{\beta}_2$ quan s'omet x_2 en li equació estimada.

	Corr(x_2,x_3)>0	*Corr(x_2,x_3)<0*
β_3>0	Biaix positiu	Biaix negatiu
β_3<0	Biaix negatiu	Biaix positiu

Forma funcional incorrecta

Si utilitzem una forma funcional diferent del model poblacional veritable, aleshores els estimadors *MQO* estaran esbiaixats.

En resum, si hi ha exclusió de variables rellevants i/o s'ha mutilat d'una manera funcional incorrecta, es que els estimadors *MQO* estaran esbiaixats i a més seran també inconsistents. En conseqüència els procediments convencionals d'inferència quedaran invalidats en aquests dos casos.

6.2.2 Contrastos d'especificació: el contrast RESET

Per contractar si s'han inclòs en el model variables irrellevants, es poden aplicar els contrastos d'exclusió examinats en el capítol 4.

Per contrastar l'exclusió de variables rellevants o la utilització d'una forma funcional incorrecta, pot aplicar-se el contrast RESET (Regression Equation Specification Error Test). Aquest contrast és un contrast general per errors d'especificació proposat per Ramsey (1969). Per explicar-ho, considerarem que el model inicial és el següent:

$$y=\beta_1+\beta_2x_2+\beta_3x_3+u \qquad (6\text{-}14)$$

Ara, anem a introduir un model augmentat en el qual apareixen dues noves variables (z_1 i z_2):

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha_1 z_1 + \alpha_2 z_2 + u \tag{6-15}$$

Tenint en compte l'especificació dels dos models, les hipòtesis nul·la i alternativa seran les següents:

$$\begin{aligned} &H_0 : \alpha_1 = \alpha_2 = 0 \\ &H_1 : H_0 \text{ no es cierta} \end{aligned} \tag{6-16}$$

La qüestió clau per construir aquest contrast és determinar les variables o regressors *z* que s'han d'introduir. En el cas d'exclusió de variables rellevants, les variables *z* seran els regressors omesos o també quadrats o potències de nous regressors. El contrast a aplicar seria similar als contrastos d'exclusió, però amb els papers invertits: el model restringit és ara el model *inicial*, mentre que el model no restringit es correspon amb el model *augmentat*.

En el contrast per a formes funcionals incorrectes, considerem, per exemple, que s'ha especificat (6-14) en lloc de la veritable relació:

$$\ln(y) = \beta_1 + \beta_2 \ln(x_2) + \beta_3 \ln(x_3) + u \tag{6-17}$$

En el model (6-17) hi ha una relació multiplicativa entre els regressors. Ramsey, va tenir en compte que una aproximació per sèries de Taylor d'una relació multiplicativa donaria lloc a una expressió que inclouria potències i productes creuats de les variables explicatives. Per aquesta raó, aquest autor suggereix la inclusió, en el model augmentat, de potències dels valors predits de la variable independent (que són, per descomptat, combinacions de potències i productes creuats de les variables explicatives):

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha_1 \hat{y}^2 + \alpha_2 \hat{y}^3 + u \tag{6-18}$$

on les $\hat{y}$ són els valors ajustats per *MQO* corresponents al model (6-14). Els superíndexs indiquen les potències a les que aquests valors predits estan elevats. No s'inclou la primera potència perquè seria perfectament colineal amb la resta dels regressors del model inicial l.

Els passos implicats en el contrast RESET són els següents:

Pas 1. S'estima el model inicial i es calculen els valors ajustats, $\hat{y}_i$.

Pas 2. S'estima el model augmentat (6-18), el qual pot incloure una o més potències de $\hat{y}_i$.

Pas 3. Prenent el R^2_{inic} corresponent al model inicial i el R^2_{aum} corresponent al model augmentat, es calcula l'estadístic *F*:

$$F = \frac{(R^2_{aum} - R^2_{inic}) / r}{(1 - R^2_{aum}) / (n - h)} \tag{6-19}$$

on *r* és el nombre de nous paràmetres que s'han afegit al model inicial, i *h* és el nombre de paràmetres del model augmentat, inclòs el terme independent.

Sota la hipòtesi nul·la, aquest estadístic es distribueix com segueix:

$$F \mid H_0 \sim F_{r,n-h} \tag{6-20}$$

Pas 4. Per a un nivell de significació α, i designant per $F^{\alpha}_{r,n-h}$ el corresponent valor a la taula de la F, la decisió a prendre és la següent:

Si $F \geq F^{\alpha}_{r,n-h}$ es rebutja H_0

Si $F < F^{\alpha}_{r,n-h}$ no es rebutja H_0

En conseqüència, valors elevats d'aquest estadístic conduiran a rebutjar el model inicial.

En el contrast RESET es contrasta una hipòtesi nul·la contra una hipòtesi alternativa que no indica quin hauria de ser l'especificació correcta del model. Així doncs aquest contrast és un contrast d'especificació que pot indicar que hi ha algun tipus d'especificació errònia però sense donar cap pista de quina és l'especificació correcta.

EXEMPLE 6.1 Especificació errònia en un model de determinació dels salaris

Utilitzant una de l'Enquesta d'Estructura Salarial per a Espanya en 2006 (arxiu *wage06sp*) es va estimar el següent model per explicar els salaris:

$$\underset{}{wage_i} = \underset{(1.55)}{4.679} + \underset{(0.146)}{0.681}educ_i + \underset{(0.071)}{0.293}tenure_i$$

R^2=0.249 n=150

on educació (*educ*) i antiguitat a l'empresa (*tenure*) estan mesurats en anys i el salari (*wage*) en euros per hora.

Considerant que podia haver-hi un problema de manera funcional incorrecta, es va estimar un model augmentat. En aquest model augmentat - a més d'*educ*, *tenure*, i el terme independent - $wage_i^2$ i $wage_i^3$, obtinguts a partir de l'estimació del model inicial, van ser inclosos com regressors. L'estadístic F calculat utilitzant R^2_{inic} i R^2_{augm}, d'acord a (6-18), és igual a 4.18. Atès que $F^{0.05}_{2,145} \simeq F^{0.05}_{2,60} = 3.15$, es rebutja, per als nivells α=0.05 i α= 0.10, que la forma lineal siga l'adequada per explicar la determinació dels salaris. Per contra, atès que $F^{0.01}_{2,145} \simeq F^{0.01}_{2,60} = 4.98$, la H_0 no se rebutja per α=0.01.

6.3 Multicolinealitat

6.3.1 Plantejament

La multicolinealitat perfecta no se sol presentar en la pràctica, llevat que es dissenyi malament el model com veurem en l'epígraf següent. En canvi, sí que és freqüent que entre els regressors hi hagi una relació aproximadament lineal, en aquest cas els estimadors que s'obtinguin seran en general poc precisos, encara que segueixen conservant la propietat de ser estimadors *ELNEO*. En altres paraules, la relació entre regressors fa que siga difícil quantificar amb precisió l'efecte que cada regressor exerceix

sobre el regressant, el que determina que les variances dels estimadors siguen elevades. Quan es presenta una relació aproximadament lineal entre els regressors, es diu que hi ha multicolinealitat no perfecta. És important assenyalar que el problema de multicolinealitat, sorgeix perquè no hi ha informació suficient per obtenir una estimació precisa dels paràmetres del model.

Per analitzar aquest problema, examinarem la variança d'un estimador. En el model de regressió lineal múltiple, l'estimador de la variança d'un coeficient de pendent qualsevol - per exemple, de $\hat{\beta}_j$ - es pot formular de la següent manera:

$$\text{var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{nS_j^2(1-R_j^2)} \tag{6-21}$$

on $\hat{\sigma}^2$ és l'estimador no esbiaixat de σ^2, n és la mida de la mostra, S_j^2 és la variança mostral del regressor X_j i R_j^2 es el coeficient de determinació obtingut en efectuar la regressió de X_j sobre la resta dels regressors del model.

L'últim d'aquests quatre factors que determinen el valor de la variança de $\hat{\beta}_j$ és el que es refereix a la multicolinealitat. Diem que la multicolinealitat sorgeix en estimar β_j quan R_j^2 està "pròxim" a 1 però no hi ha una cota que es puga fixar per concloure que la multicolinealitat és realment un problema per a la precisió dels estimadors. Encara que el problema de la multicolinealitat no pot definir-se clarament, és cert que, a l'estimar β_j, és millor que la variable x_j tinga menys correlació amb les altres variables independents. Si un R_j^2 és igual a 1, tindríem multicolinealitat perfecta i cap possibilitat d'obtenir estimacions dels coeficients. Quan un o més dels R_j^2 s'aproximen a 1 la multicolinealitat té una certa gravetat. En aquest cas, es presenten els següents problemes al realitzar inferències amb el model:

a) Les variances dels estimadors són molt grans.
b) Els coeficients estimats seran molt sensibles davant xicotets canvis en les dades.

6.3.2 Detecció

Com la multicolinealitat és un problema *mostral*, ja que va associada a la configuració concreta de la matriu dels regressors, no existeixen contrastos estadístics, pròpiament dits, que siguen aplicables per a la seva detecció. (Recordeu que els contrastos estadístics van referits a paràmetres poblacionals). En canvi, s'han desenvolupat nombroses regles pràctiques que tracten de determinar en quina mesura la multicolinealitat afecta greument les inferències realitzades amb un model. Aquestes regles no són sempre fiables, i en alguns casos molt discutibles. En qualsevol cas, es van a exposar algunes mesures que són útils per detectar el grau de multicolinealitat: el *factor d'engrandiment de la variança (FEV)* i la *tolerància*, i el *nombre de condició* i el *coeficient de descomposició de la variança*

Factor d'engrandiment de la variança (FEV) i tolerància

A fi d'explicar el significat d'aquestes mesures, suposem que no hi ha cap tipus de relació lineal entre el regressor x_j i la resta de regressors del model, és a dir, el regressor

xj és ortogonal amb la resta dels regressors. Aleshores, R_j^2 serà 0 i la variança de $\hat{\beta}_j$ serà igual a

$$\text{var}(\beta_j^*) = \frac{\hat{\sigma}^2}{nS_j^2} \tag{6-22}$$

El quocient entre (6-20) i (6-21) és precisament el factor d'engrandiment de la variança (*FEV*), l'expressió serà

$$FEV(\hat{\beta}_j) = \frac{1}{1-R_j^2} \tag{6-23}$$

A l'estadístic *FEV* calculat d'acord a (6-23) se li denomina de vegades "*FEV centrat*" per distingir-lo del "*FEV no centrat*" el qual té interès en els models sense terme independent. El programa E-views ofereix ambdós estadístics.

La tolerància, que és la inversa de *FEV*, es defineix com,

$$Tolerancia(\hat{\beta}_j) = \frac{1}{FEV} = 1-R_j^2 \tag{6-24}.$$

Així, doncs, el $FEV(\hat{\beta}_j)$ és la ràtio entre la variança observada i la que hauria estat en cas que x_j estigués incorrelacionat amb la resta de regressors del model. Dit d'una altra manera, el *FEV* mostra de quina manera «s'engrandeix» la variança de l'estimador com a conseqüència de la no ortogonalitat dels regressors. Es pot veure fàcilment que com més elevat siga el *FEV* (o com més baixa siga la tolerància), més elevada serà la variança de $\hat{\beta}_j$.

El procediment consisteix a triar a cada regressor com a variable dependent, i calcular la regressió sobre la resta dels regressors. D'aquesta manera s'obtindrien k valors del *FEV*. Si algun d'això és elevat, és un indici de multicolinealitat. Desafortunadament, però, no hi ha cap indicador teòric per determinar si el *FEV* és "alt". Tampoc, hi ha cap teoria que ens diga que fer en cas que hi hagi multicolinealitat.

El *FEV* i la tolerància són mesures utilitzades àmpliament. Alguns autors consideren que hi ha un problema greu de multicolinealitat quan el *FEV* d'algun coeficient és major de 10, és a dir, quan el *FEV*>10, o anàlogament quan la *tolerància*<0.10, però aquesta regla no té una justificació científica.

El problema que té el *FEV* (o la tolerància) és que no subministra cap informació que puga utilitzar-se per a tractar el problema.

EXEMPLE 6.2 Analitzant la multicolinealitat en el cas de l'absentisme laboral

Al Exemple 3.1 es va formular i va estimar, utilitzant el fitxer absent, un model per explicar l'absentisme laboral en funció de les variables edat, antiguitat i salari.

En quadre 6.2 s'ofereix informació de la tolerància i del *FEV* de cada variable. Segons aquests estadístiques la multicolinealitat no sembla afectar el salari però si té un cert grau d'importància en les variables edat i antiguitat. En tot cas el problema de multicolinealitat d'aquest model no sembla ser seriós ja que tots els FAV estan per sota de 5.

QUADRE 6.2. Tolerància i *FEV*.

	Estadístiques de colinealitat	
	Tolerància	*FEV*
edat	0.2346	4.2634
antiguitat	0.2104	4.7532
salari	0.7891	1.2673

Número de condició i el coeficient de descomposició de la variança

Aquest mètode, desenvolupat per Belsey et al. (1982), està basat en la descomposició de la variança de cada coeficient de regressió en funció dels arrels característiques λ_h de la matriu **X'X** i dels corresponents vectors característiques associats. No es discutirà aquí sobre els arrels i vectors característics, ja que van més enllà de l'objecte d'aquest llibre, però en tot cas veurem la seua aplicació.

El *número de condició* és una mesura estàndard del mal condicionament d'una matriu, i indica la sensibilitat potencial d'una matriu inversa calculada pel que fa a xicotets canvis en la matriu de partida (**X'X** en el cas de la regressió). Com més a prop està la matriu de ser singular, més xicotets són els valors característics. El número de condició (κ) es defineix com l'arrel quadrada de la major arrel característica ($\lambda_{\max}$) dividida per la més xicoteta ($\lambda_{\min}$):

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \tag{6-25}$$

Quan no hi ha multicolinealitat en absolut, totes les arrels característiques i el nombre de condició serà igual a 1. En créixer la multicolinealitat, les arrels característiques seran més grans i més xicotetes que 1 (les arrels característiques pròximes a 0 indiquen que hi ha un problema de multicolinealitat), i el número de condició creixerà. Una regla pràctica de caràcter informal estableix que si el número de condició és més gran que 15, aleshores la multicolinealitat és un problema, i si és més gran que 30 la multicolinealitat és un problema molt seriós.

La variança $\hat{\beta}_j$ segons les contribucions que aporta cadascuna de les arrels característiques pot expressar-se de la manera:

$$\text{var}(\hat{\beta}_j) = \sigma^2 \sum_h \frac{u_{jh}^2}{\lambda_h} \tag{6-26}$$

Així, la proporció de la contribució de λ_h a la variança de $\hat{\beta}_j$ es igual a

$$\phi_{jh} = \frac{\dfrac{u_{jh}^2}{\lambda_h}}{\displaystyle\sum_{h=0}^{k} \frac{u_{jh}^2}{\lambda_h}} \tag{6-27}$$

Valors elevats de ϕ_{jh} indiquen que, com a conseqüència de la multicolinealitat, hi ha una inflació de la variança. Atès que les arrels característiques pròximes a 0 indiquen un problema de multicolinealitat, és important prestar una especial atenció a les arrels característiques més xicotetes. Les contribucions corresponents a l'arrel característica més

xicoteta poden donar una clau de quins són els regressors que estan implicats en el problema de multicolinealitat.

EXEMPLE 6.3 Analitzant la multicolinealitat dels factors que determinen el temps dedicat al treball domèstic

A fi d'analitzar els factors que influeixen sobre el temps dedicat al treball domèstic (*houswork*), es va formular el següent model en exercici 3.17, utilitzant l'arxiu *timuse03*:

$$housework = \beta_1 + \beta_2 educ + \beta_3 hhinc + \beta_4 age + \beta_5 paidwork + u$$

on *educ* són els anys d'educació assolida, *hhinc* és la renda de la família en euros per mes. Les variables *houswork* i *paidwork* estan mesurades en minuts per dia.

El quadre 6.3 proporciona informació sobre les arrels característiques, ordenades de la més xicoteta a la més gran, i les proporcions de descomposició de la variança per a cada arrel característica estan calculades segons (6-26). El número de condició és igual a

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = \sqrt{\frac{542.14}{7.06E-06}} = 8782$$

Com es pot veure, el númtero de condició és molt elevat, el que indicaria que el problema de multicolinealitat és molt important.

Com es pot veure en el quadre 6.3[3] les proporcions més elevades associades a l'arrel característica més xicoteta, que és la responsable de la multicolinealitat en aquest model, corresponen als regressors *educ* i *age*. Aquests dos regressors estan inversament correlacionats. Les proporcions més elevades associades a la segona arrel característica més xicoteta corresponen als regressors educació assolida i renda de la llar, que estan positivament correlacionades.

QUADRE 6.3. Arrels característiques i proporcions de descomposició de la variança.

Arrels característiques	7.03E-06	0.000498	0.025701	1.861396	542.1400

Proporcions de descomposició de la variança

	Associated Eigenvalue				
Variable	1	2	3	4	5
C	0.999995	4.72E-06	8.36E-09	1.23E-13	1.90E-15
EDUC	0.295742	0.704216	4.22E-05	2.32E-09	3.72E-11
HHINC	0.064857	0.385022	0.209016	0.100193	0.240913
AGE	0.651909	0.084285	0.263805	5.85E-07	1.86E-08
PAIDWORK	0.015405	0.031823	0.007178	0.945516	7.80E-05

6.3.3 Solucions

En principi, el problema de la multicolinealitat està relacionat amb deficiències en la informació mostral. El disseny no experimental de la mostra és, sovint, el responsable

[3] En el quadre 6.3 les arrels característiques estan ordenades de menys a més el mateix que les arrels característiques associades (*associated eigenvalue*) les proporcions de descomposició de la variança. Convé advertir que en l'E-views les arrels característiques estan ordenades de major a menor. D'altra banda, el número de condició està definit de forma diferent a l'usual dels manuals d'econometria que hem seguit.

d'aquestes deficiències. Vegem a continuació algunes de les solucions proposades per resoldre el problema de la multicolinealitat.

Eliminació de variables

La multicolinealitat pot atenuar si s'eliminen els regressors que són més afectats per la multicolinealitat. El problema que planteja aquesta solució és que els estimadors del nou model seran esbiaixats en el cas que el model original fos el correcte. Sobre aquesta qüestió convé fer la següent reflexió. L'investigador està interessat en què un estimador siga no esbiaixat (o, si no pot ser, que tinga un biaix xicotet) i tinga una variança reduïda. L'error quadràtic mitjà (*EQM*) recull tots dos factors. Així, per l'estimador $\hat{\beta}_j$, l'*EQM* es defineix de la següent manera:

$$EQM(\hat{\beta}_j) = \left[biaix(\hat{\beta}_j) \right]^2 + var(\hat{\beta}_j) \tag{6-28}$$

Si un regressor és eliminat del model, l'estimador d'un regressor que es manté (per exemple, $\hat{\beta}_j$) serà esbiaixat, però, no obstant això, el seu *EQM* pot ser menor que el corresponent al model original, a causa de que l'omissió d'una variable pot fer disminuir suficientment la variança de l'estimador. En resum, tot i que l'eliminació d'una variable no és una pràctica que en principi siga aconsellable, en certes circumstàncies pot tenir la seua justificació quan contribueix a disminuir l'*EQM*.

Augment de la mida de la mostra

Tenint en compte que un cert grau de multicolinealitat comporta problemes quan augmenta ostensiblement les variances mostrals dels estimadors, les solucions han d'anar encaminades a reduir aquestes variances. Aquesta solució no sempre és viable, ja que les dades utilitzades en les contrastacions empíriques procedeixen generalment de fonts estadístiques diverses, intervenint en comptades ocasions l'investigador en la recollida d'informació.

D'altra banda, quan es tracta de dissenys experimentals, es pot incrementar directament la variabilitat dels regressors sense necessitat d'incrementar la mida de la mostra.

Utilització d'informació extramostral

Una altra possibilitat és la utilització d'informació extramostral, bé establint restriccions sobre els paràmetres del model, bé aprofitant estimadors procedents d'altres estudis.

L'establiment de restriccions sobre els paràmetres del model redueix el nombre de paràmetres a estimar i, per tant, pal·lia les possibles deficiències de la informació mostral. En qualsevol cas, perquè aquestes restriccions siguen útils han d'estar inspirades en el propi model teòric o, almenys, tenir un significat econòmic.

En general, un inconvenient d'aquesta forma de procedir és que el significat atribuïble a l'estimador obtingut amb dades de tall transversal és molt diferent de l'obtingut amb dades temporals, en el cas que es combinin ambdós tipus d'informació. De vegades, aquests estimadors poden resultar realment «estranys» o aliens a l'objecte d'estudi.

Utilització de ràtios

Si en lloc del regressant i dels regressors del model original s'utilitzen *ràtios* respecte al regressor que tinga major colinealitat, pot fer que la correlació entre els regressors del model disminueixi. Una solució d'aquest tipus resulta molt atractiva, per la seua senzillesa d'aplicació. No obstant això, les transformacions de les variables originals del model que s'estima utilitzant ràtios poden provocar un altre tipus de problemes. Suposant admissibles els supòsits del *MLC* pel que fa a les pertorbacions originals del model, aquesta transformació modificaria implícitament les propietats del model, de tal manera que les pertorbacions del model transformat utilitzant ràtios ja no serien pertorbacions homoscedàstiques, sinó heteroscedàstiques.

6.4 Contrast de normalitat

Els contrastos de significativitat *F* i *t* construïts en el capítol 4 estan basats en el supòsit de normalitat de les pertorbacions. No obstant això, no és usual realitzar contrastos de normalitat, potser pel fet que sovint no es disposa d'una mostra prou gran -per exemple, 50 o més observacions- que és necessària per a realitzar contrastos sobre aquest supòsit. De tota manera, recentment els contrastos sobre normalitat estan rebent un interès creixent tant en els estudis teòrics com aplicats.

Anem a examinar a continuació un contrast per verificar el supòsit de normalitat de les pertorbacions en un model economètric. Aquest contrast va ser proposat per Bera i Jarque, i està basat en els estadístics d'asimetria i curtosi dels residus.

L'estadístic d'asimetria és un moment de tercer ordre estandarditzat, aplicat als residus, i la seua expressió és la següent:

$$\gamma_{1(\hat{u})} = \frac{\sum \hat{u}_i^3 / n}{\left[\sum \hat{u}_i^2 / n\right]^{3/2}} \tag{6-29}$$

En una distribució simètrica, com és el cas de la distribució normal, el coeficient d'asimetria és 0.

L'estadístic de curtosi, que és un moment de quart ordre estandarditzat, té la següent expressió quan s'aplica als residus:

$$\gamma_{2(\hat{u})} = \frac{\sum \hat{u}_i^4 / n}{\left[\sum \hat{u}_i^2 / n\right]^{2}} \tag{6-30}$$

En una distribució normal estàndard, és a dir, en una distribució $N(0,1)$, el coeficient de curtosi és igual a 3.

L'estadístic de Bera i Jarque (*BJ*) ve donat per

$$BJ = \left[\frac{n}{6}\left(\gamma_{1(\hat{u})}\right)^2 + \frac{n}{24}\left(\gamma_{2(\hat{u})} - 3\right)^2\right] \tag{6-31}$$

En una distribució normal teòrica, l'anterior expressió prendrà un valor nul, ja que els coeficients d'asimetria i curtosi prenen respectivament els valors de 0 i 3. L'estadístic *BJ* prendrà valors elevats en la mesura que el coeficient d'asimetria s'allunyi de 0 i que el coeficient de curtosi s'allunyi de 3. Sota la hipòtesi nul·la de normalitat, l'estadístic *BJ* té la següent distribució:

$$BJ \xrightarrow[n\to\infty]{} \chi_2^2 \qquad (6\text{-}32)$$

Amb l'indicarien de $n \to \infty$, es vol assenyalar que és un contrast asimptòtic, és a dir, que té validesa quan la mostra siga prou gran.

EXEMPLE 6.4 És acceptable la hipòtesi de normalitat en el model per analitzar l'eficiència de la Borsa de Madrid?

En l'Exemple 4.5, utilitzant el fitxer *bolmadef*, es va analitzar l'eficiència del mercat de la Borsa de Madrid el 1992, mitjançant un model que relaciona la taxa de rendiment d'un dia sobre la taxa de rendiment del dia precedent. Ara, realitzarem contrastos de normalitat sobre les pertorbacions d'aquest model. Donada la poca proporció de variança explicada amb aquest model (vegeu Exemple 4.5), el contrast de normalitat de les pertorbacions és pràcticament equivalent a contrastar la normalitat de la variable endògena.

En el quadre 6.4 es mostren els coeficients d'asimetria, curtosi i l'estadístic de Bera i Jarque, aplicat als residus del model estimat. El coeficient d'asimetria (-0.04) no està molt allunyat del valor 0 corresponent a una distribució $N(0,1)$. D'altra banda, el coeficient de curtosi (4.43) és una mica diferent del valor 3, que pren en la distribució normal. En aquest cas, es rebutja el supòsit de normalitat per als nivells usuals de significació, ja que l'estadístic de Bera i Jarque pren el valor de 21.02, que és més gran que $\chi_2^{2(0.01)}$ =9.21.

QUADRE 6.4. Contrast de normalitat en el model de la Borsa de Madrid.

coeficient d'asimetria	*coeficient de curtosi*	*estadístic Bera i Jarque*
-0.0421	4.4268	21.0232

El fet que es rebutgi amb tanta contundència el supòsit de normalitat pot semblar paradoxal, ja que els valors de curtosi i, especialment, d'asimetria no difereixen de forma substancial dels valors que prenen aquests coeficients en una distribució normal. No obstant això, les discrepàncies són prou significatives perquè estan avalades per una mida de mostra elevat (247 observacions). Si *n* (la mida de la mostra) hagués estat de 60 en lloc de 247, l'estadístic *BJ*, calculat segons (6-30) i utilitzant els mateixos coeficients d'asimetria i curtosi, pren el valor de 5.1068, que és més xicotet que $\chi_2^{2(0.01)}$ =9.21. Dit d'una altra manera, amb els mateixos coeficients, però amb una mostra menor, no proporcionen prou evidències empíriques per rebutjar la hipòtesi nul·la de normalitat. Cal observar que això es deu al fet que l'estadístic *BJ* creix proporcionalment amb la mida de la mostra, però els graus de llibertat (2) romanen inalterables.

6.5 Heteroscedasticitat

El supòsit d'homoscedasticitat (supòsit 7 del MLC) postula que les pertorbacions tenen una variança constant, és a dir,

$$var(u_i) = \sigma^2 \qquad i = 1, 2, \cdots n \qquad (6\text{-}33)$$

Suposant que només hi ha una variable independent, el supòsit d'homoscedasticitat vol dir que la variabilitat al voltant de la línia de regressió és la mateixa al llarg de tota la mostra de les *x*; és a dir, que no augmenta o disminueix quan *x* varia, com es pot veure a la figura 2.7, part a) del capítol 2. A la figura 6.1 s'ha representat el diagrama de dispersió corresponent a un model en què les pertorbacions són homoscedàstiques. Si el supòsit d'homoscedasticitat no es compleix es diu que existeix heteroscedasticitat, o que les pertorbacions són heteroscedàtiques. A la figura 2.7, part b) es va representar un model amb pertorbacions heteroscedàstiques en què la dispersió augmentava en incrementar el valor de *x*. A la figura 6.2 s'ha representat el diagrama de dispersió corresponent a un model en què la dispersió de les pertorbacions creix en créixer *x*.

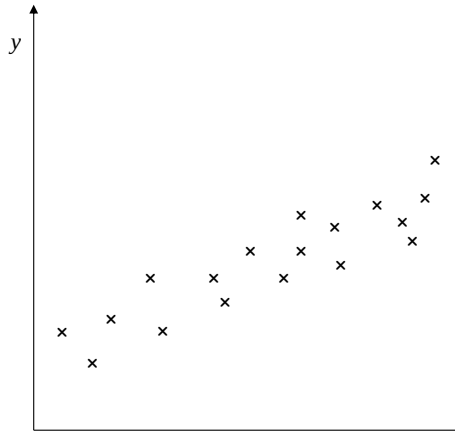


FIGURA 6.1. Diagrama de dispersió corresponent a un model amb pertorbacions homoscedàstiques.

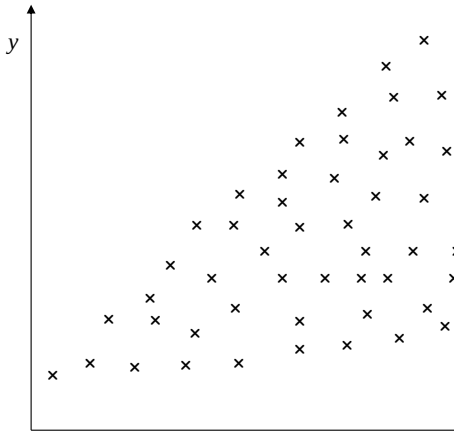


FIGURA 6.2. Diagrama de dispersió corresponent a un model amb pertorbacions heteroscedàstiques.

6.5.1 Causes de l'heteroscedasticitat

En els models estimats amb dades de tall transversal, com per exemple en els estudis de demanda basats en enquestes de pressupostos familiars, és freqüent que es presentin problemes d'heteroscedasticitat. De tota manera, l'heteroscedasticitat també es pot presentar en models estimats amb sèries temporals.

Anem a considerar ara alguns factors que poden causar que les pertorbacions d'un model siguen heteroscedàstiques:

a) Influència de la mida d'una variable explicativa en la mida de la pertorbació. Examinem aquest factor servir un exemple. Suposem un model en què la despesa en hotels és una funció lineal de la renda disponible. Si es disposa d'una mostra representativa de la població d'un país es pot comprovar la gran variabilitat de la renda percebuda per les diferents famílies. Lògicament, les famílies amb rendes baixes tenen poques possibilitats d'efectuar una despesa elevada en hotels, podent-se esperar en aquest cas que les oscil·lacions en la despesa d'unes famílies a altres no siguen importants. En canvi, en les famílies amb rendes altes es pot esperar una major variabilitat en aquest tipus de despesa. En efecte, les famílies amb rendes elevades poden optar entre gastar en hotels una part substancial de la seua renda o no gastar pràcticament res. El diagrama de la figura 6.2 pot ser adequat per representar el que succeeix en un model per explicar la demanda d'un bé de luxe com és el cas de la despesa en hotels.

b) La presència de valors anòmals (outliers) pot causar heteroscedasticitat. Un *outlier* és una observació generada aparentment per una població diferent de la que ha generat la resta de les observacions mostrals. Quan la mida de mostra és xicoteta la inclusió o exclusió d'una observació d'aquest tipus pot alterar substancialment els resultats de l'anàlisi de regressió i causar heteroscedasticitat.

c) Transformació de les dades. Com hem vist en un epígraf previ una de les solucions per resoldre el problema de la multicolinealitat consistia a transformar el model pres ràtios respecte a una variable (diguem, X_{ji}), és a dir, dividint ambdós membres del model per *Xji*. En conseqüència, la pertorbació serà ara u_i/X_{ji}, en lloc de u_i. Suposant que

ui compleix el supòsit d'homoscedasticitat, les pertorbacions del model transformat (u_i/X_{ji}) ja no seran homoscedàstiques sinó heteroscedàstiques.

6.5.2 Conseqüències de l'heteroscedasticitat

Quan hi ha heteroscedasticitat el mètode de mínims quadrats ordinaris (*MQO*), ja no és el més adequat, ja que en aquest cas els estimadors obtinguts no són òptims, és a dir, els estimadors de *MQO* no són *ELNEO*.

D'altra banda, els estimadors obtinguts per *MQO* en el cas que hi hagi heteroscedasticitat, a més de no ser *ELNEO*, presenten el següent problema. L'estimació de la matriu de covariances dels estimadors obtinguda aplicant la fórmula usual no és vàlida quan existeix heteroscedasticitat. Conseqüentment, els estadístics *t* i *F* basats en aquesta estimació de la matriu de covariances donaran lloc a inferències errònies.

6.5.3 Contrastos d'heteroscedasticitat

Anem a examinar dos contrastos d'heteroscedasticitat: Breusch-Paguen-Godfrey i White. Tots dos contrastos són asimptòtics i tenen la forma d'un contrast de multiplicadors de Lagrange (*ML*).

***Contrast de Breusch-Pagan-Godfrey* (*BPG*)**

El contrast *BPG* és un contrast asimptòtic, és a dir, vàlid només per a mostres grans. Les hipòtesis nul·la i alternativa d'aquest contrast poden formular-se de la següent manera:

$$\begin{aligned} &H_0 : E\left(u_i^2\right) = \sigma^2 \quad \forall i \\ &H_1 : \sigma_i^2 = \alpha_1 + \alpha_2 z_{2i} + \alpha_3 z_{3i} + \cdots + \alpha_m z_{mi} \end{aligned} \tag{6-34}$$

on les z_i poden ser totes o algunes de les x_i del model.

Prenent com a referència l'anterior H_1, aleshores H_0 pot expressar-se com

$$H_0 : \alpha_2 = \alpha_3 = \cdots \alpha_m = 0 \tag{6-35}$$

Els passos que necessiten és contrari als següents:

Pas 1. S'estima el model original i es calculen els residus mínims-quadrats.

Pas 2. Es realitza la següent regressió auxiliar, prenent com es recupera el quadrat dels residus obtinguts en l'estimació del model original ($\hat{u}_i^2$), ja que no se sap ni σ_i^2 ni u_i^2:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 z_{2i} + \alpha_3 z_{3i} + \cdots + \alpha_m z_{mi} + \varepsilon_i \tag{6-36}$$

A la regressió auxiliar ha d'aparèixer un terme independent, encara que el model original s'ha estimat sense ell. D'acord amb l'expressió (6-36), a la regressió auxiliar hi ha *m* regressors.

Pas 3. Designant per R_{ra}^2 al coeficient de determinació de la regió auxiliar, es calcula l'estadístic nR_{ra}^2.

Sota la hipòtesi nul·la, aquest estadístic (*BPG*) té la següent distribució:

$$BPG = nR_{ra}^2 \xrightarrow[n\to\infty]{} \chi_m^2 \tag{6-37}$$

Paso 4. Per a un nivell de significació α, i designant per $\chi_m^{2(\alpha)}$ al valor en la taula de la χ^2, la decisió que cal prendre és el següent:

Si $BPG > \chi_m^{2(\alpha)}$ es rebutja la H_0

Si $BP \leq \chi_m^{2(\alpha)}$ no se rebutja la H_0

En aquest contrast valors elevats de l'estadístic corresponen a una situació d'heteroscedasticitat, és a dir, al rebuig de la hipòtesi nul·la.

EXEMPLE 6.5 Aplicació del contrast de Breusch-Pagan-Godfrey

Apliquem a continuació aquest contrast a una submostra de 10 observacions, que s'han utilitzat per estimar les despeses en hostaleria (*hostel*) en funció de la renda disponible (*renda*). Les dades apareixen en el quadre 6.5.

QUADRE 6.5. Dades de *hostel* i *renda*.

i	*hostel*	*renda*
1	17	500
2	24	700
3	7	250
4	17	430
5	31	810
6	3	200
7	8	300
8	42	760
9	30	650
10	9	320

Pas 1. S'apliquen *MQO* al model

$$hostel = \beta_1 + \beta_2 renda + u$$

i, utilitzant les dades del quadre 6.5, s'obté el següent model estimat:

$$\underset{}{hostel_i} = \underset{(3.48)}{-7.427} + \underset{(0.0065)}{0.0533}\, renda_i$$

Els residus corresponents a aquest model ajustat apareixen en el quadre 6.6.

QUADRE 6.6. Residus de la regressió de *hostel* sobre *renda*.

i	1	2	3	4	5	6	7	8	9	10
$\hat{u}_i$	-2.226	-5.888	1.100	1.505	-4.751	-0.234	-0.565	8.913	2.777	-0.631

Pas 2. La regressió auxiliar a estimar serà la següent:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 renda_i + \eta_i$$

Aplicant MQO a l'anterior model s'obté la següent estimació:

$$\hat{u}_i^2 = -23.93 + 0.0799 renda_i\,; \qquad R^2 = 0.5045$$

Pas 3. A partir del valor de R^2 s'obté el següent valor de l'estadístic BPG:

$$BPG=nR^2=10(0.56)=5.05.$$

Pas 4. Atès que $\chi_1^{2(0.01)}=3.84$, es rebutja la hipòtesi nul·la d'homoscedasticitat per a un nivell del 5%, ja que $BPG>3.84$, però no per al nivell de significació de l'1%.

Recordeu que la validesa d'aquest contrast és asimptòtica. No obstant això, la mostra utilitzada en aquest exemple és molt xicoteta.

Contrast de White

En el contrast de White no s'especifiquen les variables que determinen l'heteroscedasticitat. Aquest és un contrast no constructiu ja que no dóna cap tipus d'indicació de l'esquema d'heteroscedasticitat quan la hipòtesi nul·la és rebutjada

El contrast de White està basat en el fet que els errors estàndard són vàlids asimptòticament si se substitueix el supòsit d'homoscedasticitat pel supòsit més feble que la pertorbació al quadrat, u^2, està incorrelacionada amb tots els regressors, els seus quadrats i els productes mixtes entre ells. Tenint en compte aquest fet, White va proposar fer la regressió auxiliar de $\hat{u}_i^2$, ja que u_i^2 és desconegut, pel que fa a tots els factors que s'acaben d'esmentar. Si els coeficients de la regressió auxiliar són conjuntament no significatius, aleshores podem admetre que les pertorbacions són homoscedàstiques. D'acord amb el supòsit adoptat, el contrast de White és asimptòtic.

L'aplicació del contrast de White pot plantejar problemes en models amb molts regressors. Per exemple, si el model original té 5 variables independents, la regressió auxiliar de White té 16 regressors (a menys que alguns siguen redundants), el que implica que la regressió es realitza amb una pèrdua de 16 graus de llibertat. Per aquesta raó, quan el model té molts regressors s'aplica sovint una versió simplificada del contrast de White. En aquesta versió simplificada s'ometen els productes creuats de la regressió auxiliar.

Els passos que es requereixen en aquest contrast són els següents:

Pas 1. S'estima el model original i es calculen els residus mínim-quadràtics.

Pas 2. Es realitza la següent regressió auxiliar, prenent com regressant al quadrat dels residus obtinguts en l'estimació del model original:

$$\hat{u}_i^2=\alpha_1+\alpha_2\psi_{2i}+\alpha_3\psi_{3i}+\cdots+\alpha_m\psi_{mi}+\varepsilon_i \qquad (6\text{-}38)$$

Els regressors de la regressió auxiliar anterior ψ_{ji} són els regressors del model original, els quadrats dels regressors i els productes creuats dels regressors.

En qualsevol cas, cal eliminar les possibles redundàncies que es produeixen (és a dir, regressors que apareguen repetits). Per exemple, no poden aparèixer simultàniament com regressors el terme independent (que és un 1 per a totes les observacions) i el quadrat d'aquest regressor, ja que són idèntics. La introducció simultània d'aquests dos regressors donaria lloc a una situació de multicolinealitat perfecta.

En la regressió auxiliar ha d'aparèixer un terme independent, tot i que el model original s'hagi estimat sense ell. D'acord amb l'expressió (6-38), s'ha considerat que en la regressió auxiliar hi ha *m* regressors.

Pas 3 Designant per R^2_{ra} al coeficient de determinació de la regressió auxiliar, es calcula l'estadístic nR^2_{ra}.

Sota la hipòtesi nul·la, aquest estadístic (*W*) té la següent distribució:

$$W = nR^2_{ra} \xrightarrow[n\to\infty]{} \chi^2_m \qquad (6\text{-}39)$$

Amb l'estadístic nR^2_{ra} es contrasta la significativitat global del model (6-38).

Pas 4. És similar al pas 4 en el contrast de Breusch-Pagan-Godfrey.

EXEMPLE 6.6 Aplicació del contrast de White

Aquest contrast es va aplicar a les dades del quadre 6.5.

Pas 1. Aquest pas és igual que en el contrast de Breusch-Paguen-Godfrey.

Pas 2. Com hi ha dos regressors en el model original (terme independent i *renda*), els regressors de la regressió auxiliar són

$$\psi_{1i} = 1 \qquad \forall i$$
$$\psi_{2i} = 1 \times renda_i$$
$$\psi_{3i} = renda_i^2$$

En conseqüència, el model a estimar serà

Aplicant *MQO* a l'anterior model, utilitzant dades del quadre 6.5, s'obté la següent estimació:

$$\hat{u}_i^2 = 14.29 - 0.10 renda_i + 0.00018 renda_i^2 \qquad R^2=0.56$$

Pas 3. A partir del valor de R^2 s'obté l'estadístic *W*:

$$W=nR^2 = 10(0.56) = 5.60.$$

El nombre de graus de llibertat és 2.

Pas 4. Atès que $\chi_2^{2(0.10)}$ =4.61, es rebutja la hipòtesi nul·la d'homoscedasticitat per a un nivell del 10% ja que $W=nR^2>4.61$, però no per nivells de significació del 5% i de l'1%.

Cal tenir en compte que la validesa d'aquest contrast també és asimptòtica.

EXEMPLE 6.7 Contrastos d'heteroscedasticitat en la determinació del valor de les accions dels bancs espanyols

Per explicar el valor de mercat (*marktval*) dels bancs espanyols en funció del seu valor comptable (*bookval*) s'han formulat dos models, un lineal (Exemple 2.8) i l'altre doblement logarítmic (Exemple 2.10).

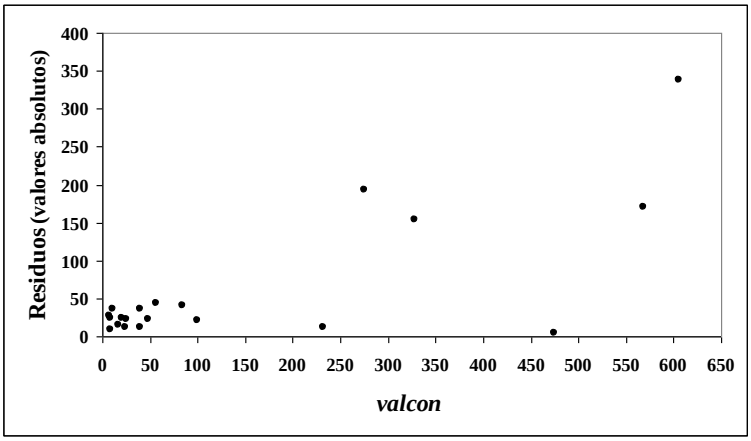
Heteroscedasticitat en el model lineal

El model lineal ve donat per

$$marktval=\beta_1+ \beta_2 bookval+u$$

Utilitzant dades de 20 bancs i entitats d'assegurances (fitxer *bolmad95*) s'han obtingut els següents resultats:

En el gràfic 6.1 s'ha representat el diagrama de dispersió entre els residus en valor absolut (en ordenades) i la variable *bookval* (en abscisses). De l'examen d'aquest gràfic es desprèn que els valors absoluts dels residus, que són indicatius de la dispersió d'aquesta sèrie, creixen en incrementar els valors de la variable *bookval*. En altres paraules, aquest gràfic constitueix un indici, però no una prova formal, de l'existència d'heteroscedasticitat de les pertorbacions associada a la variable *bookval*.



GRÀFIC 6.1. Diagrama de dispersió entre els residus en valor absolut i la variable *bookval* en el model lineal.

L'estadístic de Breusch-Pagan-Godfrey pren el següent valor:

$$BPG = nR_{ra}^2 = 20\times 0.5220 = 10.44$$

Com $\chi_1^{2(0.01)}$ =6.64<10.44, es rebutja la hipòtesi nul·la d'homoscedasticitat per a un nivell de significació de l'1%, i, en conseqüència per α=0.05 i per α=0.10.

Anem a aplicar a continuació el contrast de White. En aquest cas, en la regressió auxiliar s'inclouen com regressors el terme independent, la variable *bookval*, i el quadrat d'aquesta variable. L'estadístic de White pren el següent valor,

$$W = nR_{ra}^2 = 20\times 0.6017 = 12.03$$

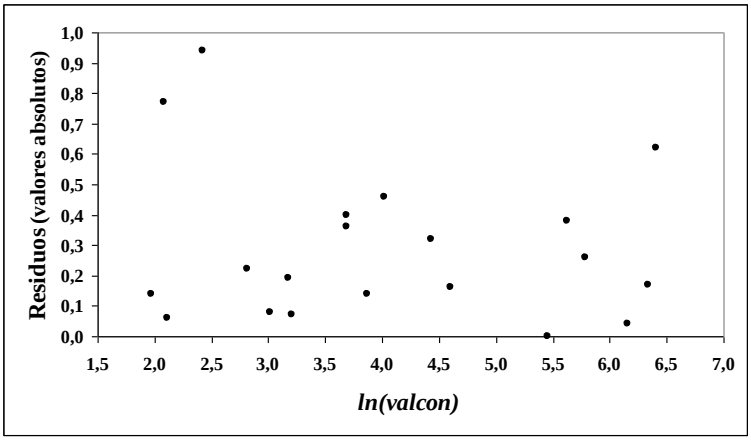
Com $\chi_2^{2(0.01)}$ =9.21<12.03, es rebutja la hipòtesi nul·la d'homoscedasticitat per a un nivell de significació de l'1%.

Heteroscedasticitat en el model doblement logarítmic

L'estimació del model doblement logarítmic amb la mateixa mostra ha estat la següent:

$$\ln(marktval) = \underset{(0.265)}{0.676} + \underset{(0.062)}{0.9384}\ln(bookval)$$

En el gràfic 6.2 s'ha representat el diagrama de dispersió entre els residus en valor absolut (en ordenades), obtinguts en estimar el model el model anterior, i la variable ln(*bookval*) (en abscisses). Com es pot veure, els dos residus més grans corresponen a dos bancs amb valor comptable xicotet. Encara no tenint en compte aquests dos casos, no sembla que hi hagi una relació entre els residus i la variable explicativa del model.



GRÀFIC 6.2. Diagrama de dispersi entre els residus en valor absolut i la variable ln(*bookval*) en el model doblement logarítmic

Els resultats dels dos contrastos d'heteroscedasticitat aplicats es presenten en el quadre 6.7.

QUADRE 6. 7. Contrastos d'heteroscedasticitat en el model doblement logarítmic per explicar el valor de mercat dels bancs espanyols.

Contrast	*Estadístic*	*Valors taules*
Breusch-Pagan	$BP = nR_{ra}^2 = 1.05$	$\chi_2^{2(0.10)} = 4.61$
White	$W = nR_{ra}^2 = 2.64$	$\chi_2^{2(0.10)} = 4.61$

Com es pot veure, tots dos contrastos són concloents en què no es pot rebutjar la hipòtesi nul·la d'homoscedasticitat enfront de la hipòtesi alternativa que la variança de les pertorbacions està associada a la variable explicativa del model.

Una conclusió important d'aquest cas és la següent. Quan en l'estimació d'un model economètric amb dades de tall transversal hi ha unitats de molt diferent mida, els problemes d'escala poden provocar heteroscedasticitat en les pertorbacions. Aquests problemes es poden resoldre en moltes ocasions utilitzant models logarítmics.

Exemple 6.8 Existeix heteroscedasticitat en la demanda de serveis d'hostaleria?

En general, en la demanda de béns alimentaris no sol aparèixer heteroscedasticitat en les pertorbacions. En canvi, en la demanda de béns de luxe l'heteroscedasticitat sol ser molt més freqüent, pel fet que en la demanda d'aquests béns pot haver una disparitat molt gran en el comportament de les llars amb rendes elevades, enfront de les llars amb rendes baixes en què és molt improbable que existisca tal disparitat donat el reduït de la renda.

A la vista de les consideracions anteriors, anem a estimar un mo del en el qual s'explica el logaritme de la despesa en serveis d'hostaleria -ln (*hostel*) - en funció del logaritme de la renda disponible -ln (*inc*) - i d'altres variables demogràfiques i socials.

L'especificació utilitzada per a l'estimació de la demanda dels serveis d'hostaleria és la següent:

$$\ln\ hostel = \beta_1 + \beta_2 \ln(inc) + \beta_3 secstud + \beta_4 terstud + \beta_5 hhsize + u \qquad (6\text{-}40)$$

on *inc* és la renda disponible de la llar, *hhsize* és el nombre de membres de la llar, i *secstud* i *terstud* són dues variables fictícies que el valor 1 si han completat estudis secundaris i terciaris respectivament.

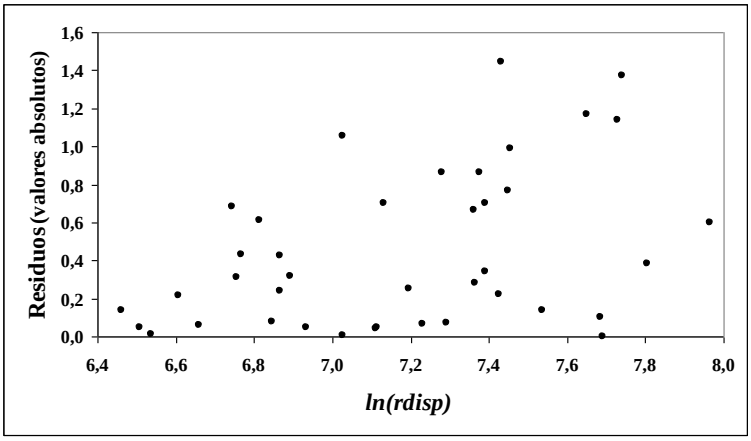
Els resultats de la regressió obtinguts són els següents (fitxer *hostel*):

$$\ln(hostel)_i = \underset{(2.26)}{-16.37} + \underset{(0.324)}{2.732}\ln(inc)_i + \underset{(0.258)}{1.398}\, secstud_i + \underset{(0.333)}{2.972}\, terstud_i - \underset{(0.088)}{0.444}\, hhsize_i$$

A la vista d'aquests resultats, es pot afirmar que els serveis d'hostaleria són un bé de luxe, ja que l'elasticitat demanda/renda per aquest bé és molt elevada (2.73). Això vol dir que, si la renda s'incrementa en un 1%, la despesa en serveis d'hostaleria augmentarà, de mitjana, en un 2.73%. Com es pot veure les famílies en què el sustentador principal té estudis mitjans (*secstud*) o, en major mesura, estudis superiors (*terstud*), realitzen una major despesa en serveis d'hostaleria que quan el sustentador principal només té estudis primaris. Per contra, aquesta despesa disminueix en augmentar la mida de la llar (*hhsize*).

En el gràfic 6.3 s'ha representat el gràfic de dispersió entre els residus en valor absolut i la variable ln (inc), ja que, en els models de demanda, en els quals apareix la renda (o una transformació de la mateixa) com a variable explicativa, és aquesta variable la principal candidata, per no dir l'única, per explicar la hipotètica heteroscedasticitat en les pertorbacions. Com es pot veure en el gràfic, la dispersió dels residus és més reduïda per a les rendes baixes, que en les rendes mitjanes o altes.

Anem a aplicar a continuació els dos contrastos d'heteroscedasticitat que s'han exposat en aquest apartat.



GRÀFIC 6.3. Diagrama de dispersió entre els residus en valor absolut i la variable ln(*inc*) en l'estimació del model d'hostaleria.

Els resultats dels dos contrastos d'heteroscedasticitat examinats es presenten en el quadre 6.8.

QUADRE 6. 8. Contrastos d'heteroscedasticitat en el model de demanda de serveis d'hostaleria.

Contrast	*Estadístic*	*Valors taules*
Breusch-Pagan	$BP = nR^2_{ra}$ =7.83	$\chi^{2(0.05)}_2$ =5.99
White	$W = nR^2_{ra}$ =12.24	$\chi^{2(0.01)}_2$ =9.21

En el contrast de *BPG* es rebutja la hipòtesi nul·la d'homoscedasticitat per a un nivell de significació de α=0.05 però no per a un nivell de α=0.01.

En l'aplicació del contrast de White, atès que hi ha moltes variables dicotòmiques en el model, la inclusió dels productes creuats en la regressió auxiliar pot donar lloc a seriosos problemes de multicolinealitat. Per aquesta raó, en la regressió auxiliar no s'han inclòs els productes creuats. Com és lògic, entre els regressors de la regressió auxiliar no figuren els quadrats de *secstud* i *terstud*, ja que els quadrats d'aquests regressors són ells mateixos per tractar-se de variables dicotòmiques. Donat el valor obtingut en l'estadístic de White, es rebutja la hipòtesi nul·la d'homoscedasticitat per a un nivell de significació de α=0.01. En conseqüència, el contrast de White és més concloent en el rebuig del supòsit d'homoscedasticitat.

6.5.4 Estimació de la matriu de covariances consistent sota heteroscedasticitat

Quan existeix heteroscedasticitat i apliquem *MQO*, no podem fer inferències correctes si utilitzem la matriu de covariances associada a les estimacions per *MQO*, ja que aquesta matriu no és un estimador consistent de la matriu de covariances dels

coeficients. En conseqüència, els estadístics *t* i *F* basats en la matriu de covariances estimada condueixen a inferències errònies.

Per tant, si hi ha heteroscedasticitat i ha estat aplicat el mètode de *MQO*, per a realitzar inferències hauria de buscar-un estimador de la matriu de covariances que siga consistent sota el supòsit heteroscedasticitat. White va proposar un estimador que és consistent sota aquest supòsit. Tanmateix, és important tenir en compte que aquest estimador no treballa bé en xicotetes mostres, ja que és una aproximació asimptòtica.

La majoria dels paquets economètrics permeten calcular desviacions estàndard dels estimadors pel procediment de White. Utilitzant aquests errors estàndard consistents es poden fer contrastos correctes sota el supòsit heteroscedasticitat.

EXEMPLE 6.9 Errors estàndard consistents en la determinació del valor de les accions dels bancs espanyols (continuació Exemple 6.7)

En la següent equació estimada del model lineal les desviacions típiques dels estimadors són calculades pel procediment de White i, per tant, són consistents sota el supòsit heteroscedasticitat:

$$\underset{}{marktval} = \underset{(18.67)}{29.42} + \underset{(0.249)}{1.219}\, bookval$$

Com es pot comprovar, l'error estàndard del coeficient de *bookval* passa de 0.127 aplicant el procediment usual a 0.249 en el procediment de White. De tota manera, el nivell de significació crític segueix sent molt baix, ja que el seu valor se situa en 0.0001. En conseqüència, se segueix mantenint la significativitat de la variable *bookval* per a tots els nivells usuals. Per contra, el terme independent, que no té especial rellevància en el model, té ara un error estàndard (18.67), que és inferior a l'obtingut amb el procediment usual (30.85).

Si apliquem el procediment de White al model doblement logarítmic s'obtenen els següents resultats:

$$\ln(marktval) = \underset{(0.3218)}{0.676} + \underset{(0.0698)}{0.9384}\ln(bookval)$$

En aquest cas, l'error estàndard del coeficient ln(*bookval*) és pràcticament el mateix en els dos procediments.

Dels anteriors resultats es poden obtenir les següents conclusions. En la determinació del valor de les accions dels bancs espanyols, les pertorbacions del model lineal són fortament heteroscedàstiques. Per això, en realitzar una estimació consistent, la desviació típica gairebé es duplica respecte al procediment usual. Per contra, en el model doblement logarítmic, que no està afectat per l'heteroscedasticitat, tot just hi ha diferències entre els errors estàndard que s'obtenen per ambdós procediments.

6.5.5 Tractament de l'heteroscedasticitat

Per realitzar l'estimació d'un model amb pertorbacions heteroscedàstiques cal conèixer o, en cas que no es conega, estimar l'esquema d'heteroscedasticitat. Així, suposem que la desviació típica de les pertorbacions segueix el següent esquema:

$$\sigma_i = f\left(x_{ji}\right) \tag{6-41}$$

Com s'ha indicat en l'epígraf 6.1, l'aplicació del mètode de *MQG* permet obtenir estimadors *ELNEO* quan les pertorbacions són heteroscedàstiques. Conegut l'esquema (6-41), l'aplicació de *MQG* es realitza en dues fases. En la primera etapa es transforma el model original dividint ambdós membres per la desviació estàndard. Per tant, d'acord amb (6-40), el model transformat vindrà donat per

$$\frac{y_i}{f\left(x_{ji}\right)}=\beta_1\frac{1}{f\left(x_{ji}\right)}+\beta_2\frac{x_{1i}}{f\left(x_{ji}\right)}+\beta_3\frac{x_{2i}}{f\left(x_{ji}\right)}+\cdots+\beta_k\frac{x_{ki}}{f\left(x_{ki}\right)}+\frac{u_i}{f\left(x_{ji}\right)} \qquad (6\text{-}42)$$

Es pot veure fàcilment que les pertorbacions del model anterior, ($u_i/f(x_{ji})$),són homoscedàstiques. Per això, en la segona etapa s'apliquen *MQO* al model transformat, ja que s'obtindran estimadors *ELNEO*. Atès que, en dividir per $f(x_{ji})$, s'està ponderant cada observació per l'invers del valor que pren aquesta funció, al procediment anterior se li denomina freqüentment mínims quadrats ponderats (*MQP*). En aquest cas, el factor de ponderació és $1/f(x_{ji})$.

Si no es coneix la funció $f(x_{ji})$, cal procedir a la seva estimació. En aquest cas, el mètode d'estimació no serà exactament *MQG*, ja que l'aplicació d'aquest mètode implica el coneixement de la matriu de covariances, o almenys el coneixement d'una matriu que siga proporcional a aquesta. Quan s'estima la matriu de covariances, a més dels paràmetres, es diu que s'apliquen *MQG* factibles. En el cas de pertorbacions heteroscedàstiques, a la particularització del mètode de *MQG* factibles, se li denomina *MQP* en dues etapes. En la primera etapa s'estima la funció $f(x_{ji})$, mentre que en la segona etapa s'aplica *MQO* al model transformat utilitzant les estimacions de $f(x_{ji})$.

Per veure com es pot aplicar el mètode de *MQP* en dues etapes, anem a partir de la següent relació, que simplement defineix la variança de les pertorbacions, en el cas d'heteroscedasticitat,

$$E\left(u_i^2\right)=\sigma_i^2 \qquad (6\text{-}43)$$

Per tant, la pertorbació al quadrat es pot fer igual, com en el model de regressió, al seu esperança més una variable aleatòria, és a dir,

$$u_i^2=\sigma_i^2+\varepsilon_i \qquad (6\text{-}44)$$

Com les pertorbacions no són observables, es pot establir una relació anàloga a l'anterior utilitzant els residus en lloc de les pertorbacions. Per tant, s'obté que

$$\hat{u}_i^2=\sigma_i^2+\eta_{2i} \qquad (6\text{-}45)$$

Cal tenir en compte que la relació anterior no té exactament les mateixes propietats que (6-44), a causa que els residus estan correlacionats i són heteroscedàstics, encara que les pertorbacions complisquen amb tots els supòsits de l'MLC. No obstant això, en grans mostres les propietats són les mateixes.

Si utilitzem els residus com regressant, en lloc dels residus al quadrat, caldrà prendre valors absoluts, ja que la desviació estàndard només pren valors positius. Si es té en compte (6-45), es pot establir la següent relació:

$$\left|\hat{u}_i\right|=\sigma_i^2+\eta_{2i}=f\left(x_{ji}\right)+\eta_{2i} \qquad (6\text{-}46)$$

Atès que la funció $f(x_{ji})$ serà en general desconeguda, se solen assajar diferents funcions. A continuació, presentem algunes de les funcions més usuals:

$$\begin{aligned}
|\hat{u}_i| &= \alpha_1 + \alpha_2 x_{ji} + \eta_{2i} \\
|\hat{u}_i| &= \alpha_1 + \alpha_2 \sqrt{x_{ji}} + \eta_{2i} \\
|\hat{u}_i| &= \alpha_1 + \alpha_2 \frac{1}{x_{ji}} + \eta_{2i} \\
|\hat{u}_i| &= \alpha_1 + \alpha_2 \ln(x_{ji}) + \eta_{2i}
\end{aligned} \tag{6-47}$$

A la vista dels resultats, se selecciona aquella forma funcional amb la que s'obtinga un millor ajust (un coeficient de determinació més elevat o un estadístic *AIC* més xicotet). Per a la transformació del model es contemplen dues circumstàncies, segons quina siga la significativitat del terme independent. Si aquest coeficient és estadísticament significatiu, es transforma el model dividint pels valors ajustats de l'equació seleccionada. Si no és estadísticament significatiu, es transforma el model dividend pel regressor corresponent a l'equació seleccionada. Així, si l'equació seleccionada fos la segona de (6-47), i no és significatiu el terme independent, el model transformat seria el següent:

$$\frac{y_i}{\sqrt{x_{ji}}} = \beta_1 \frac{1}{\sqrt{x_{ji}}} + \beta_2 \frac{x_{2i}}{\sqrt{x_{ji}}} + \beta_3 \frac{x_{3i}}{\sqrt{x_{ji}}} + \cdots + \beta_k \frac{x_{ki}}{\sqrt{x_{ji}}} + \frac{u_i}{\sqrt{x_{ji}}} \tag{6-48}$$

Cal observar que en el cas de que el terme independent no siga significatiu, en la transformació del model no intervenen paràmetres estimats, però si ho faran en el cas que siga significatiu dit terme independent. Com els estimadors dels models (6-47) no són no esbiaixats, encara que sí consistents, no és convenient realitzar transformacions amb valors ajustats -en el càlcul intervenen $\hat{\alpha}_1$ i $\hat{\alpha}_2$ - llevat que siga molt fort (per exemple, superior a l'1%) la significativitat del terme independent.

EXEMPLE 6.10 Aplicació de mínims quadrats ponderats en la demanda de serveis d'hostaleria (continuació 6.8)

Atès que els dos contrastos aplicats al model per explicar la despesa dels serveis d'hostaleria indiquen que les pertorbacions són heteroscedàstiques, anem a aplicar el mètode de mínims quadrats ponderats per estimar el model (6-40).

En primer lloc, s'estimen els quatre models (6-47), utilitzant com `regressant als residus en valor absolut $|\hat{u}_i|$ obtinguts en l'estimació del model (6-40) per *MQO*. Els resultats d'aquestes estimacions es presenten a continuació:

$$|\hat{u}_i| = \underset{(0.143)}{0.0239} + \underset{(2.73)}{0.0003}\, inc \qquad R^2 = 0.1638$$

$$|\hat{u}_i| = \underset{(-1.34)}{-0.4198} + \underset{(2.82)}{0.0235}\sqrt{inc} \qquad R^2 = 0.1733$$

$$|\hat{u}_i| = \underset{(5.39)}{0.8857} - \underset{(-2.87)}{532.1}\frac{1}{inc} \qquad R^2 = 0.1780$$

$$|\hat{u}_i| = \underset{(-2.46)}{-2.7033} + \underset{(2.88)}{0.4389}\ln(inc) \qquad R^2 = 0.1788$$

En els resultats anteriors sota de cada coeficient apareix l'estadístic *t*.

La forma funcional seleccionada és la que utilitza ln(*inc*) com regressor, ja que per a ella s'obté el R^2 més elevat. Atès que el coeficient del terme independent no és estadísticament significatiu a l'1% i seguint la recomanació feta, es van a aplicar *MQP*, prenent com a ponderació 1/ln(*inc*). En l'estimació per *MQP* s'han obtingut els següents resultats:

$$\ln(hostel)_i = \underset{(2.15)}{-16.21} + \underset{(0.309)}{2.709}\ln(inc)_i + \underset{(0.247)}{1.401}\,secstud_i + \underset{(0.326)}{2.982}\,terstud_i - \underset{(0.085)}{0.445}\,hhsize_i$$

$$R^2=0.914 \qquad n=40$$

Comparant amb l'estimació per *MQO*, feta en l'Exemple 6.5, es pot veure que les diferències són molt xicotetes, el que és indicatiu de la robustesa del model.

6.6 Autocorrelació

El supòsit de *no autocorrelació*, o de *no correlació serial*, (supòsit 8 del MLC) postula que les pertorbacions amb diferents subíndexs no estan correlacionades entre si:

$$E(u_i u_j) = 0 \qquad i \neq j \tag{6-49}$$

És a dir, les pertorbacions corresponents a diferents períodes de temps, o a individus diferents, no estan correlacionades entre si. A la figura 6.3 es mostra un gràfic que correspon a pertorbacions que no estan autocorrelacionades. L'eix *x* és el temps. Com es pot observar, les pertorbacions es distribueixen aleatòriament per sobre *y* per sota de la línia 0 (mitjana teòrica de *u*). A la figura, cada pertorbació està unida per una línia a la pertorbació del període següent: en total aquesta línia creua la línia 0 en 13 ocasions.

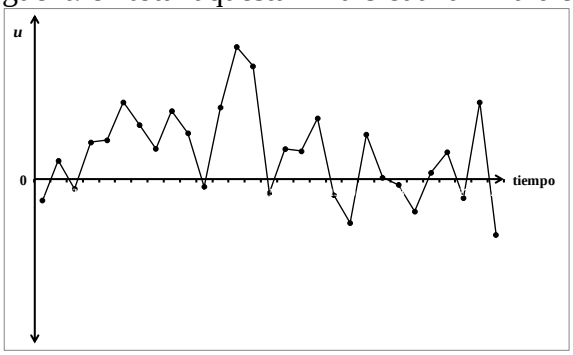


FIGURA 6.3. Gràfic de pertorbacions no autocorrelacionades.

La transgressió del supòsit de no autocorrelació es produeix amb força freqüència en els models que utilitzen dades de sèries temporals. Cal assenyalar també que l'autocorrelació pot ser tant positiva com negativa. L'autocorrelació positiva es caracteritza per deixar una estela al llarg del temps, pel fet que el valor de cada pertorbació es troba pròxim al valor de la pertorbació que el precedeix. L'autocorrelació positiva es produeix molt més freqüentment en la pràctica que la negativa. A la figura 6.4 es mostra un gràfic que correspon a les pertorbacions que estan positivament autocorrelacionades. Com es pot veure, la línia que uneix les pertorbacions successives creua la línia 0 en només 4 vegades.

Per contra, les pertorbacions afectades per autocorrelació negativa presenten una configuració de dents de serra, i sovint cada pertorbació té el signe oposat de la pertorbació que el precedeix. A la figura 6.5 el gràfic correspon a pertorbacions que estan negativament autocorrelacionades. Ara la línia 0 és creuada en 21 ocasions per la línia que uneix les pertorbacions successives.

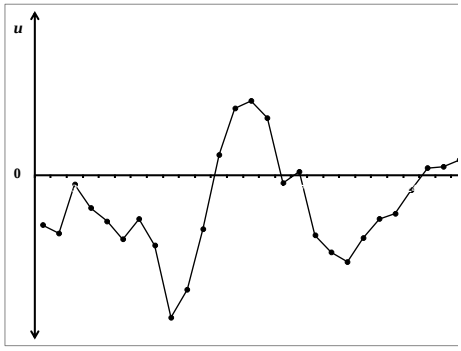


FIGURA 6.4. Gràfic de pertorbacions autocorrelacionades positivament.

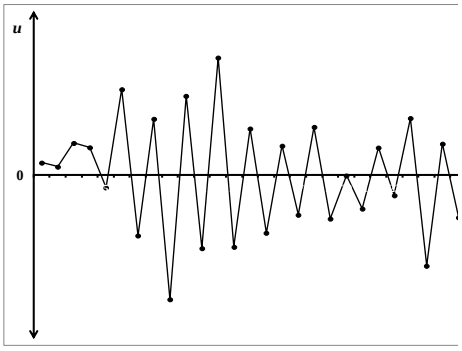


FIGURA 6.5. Gràfic de pertorbacions autocorrelacionades negativament.

6.6.1 Causes d'autocorrelació

Hi ha diverses causes per la presència d'autocorrelació en un model. Vegem a continuació algunes d'elles.

a) *Biaix d'especificació*. Pot ser degut a l'ús d'una forma funcional incorrecta o a l'omissió d'una variable rellevant.

Suposem que la forma funcional correcta per determinar el *salari* en funció dels anys d'experiència (*exp*) és la següent:

$$salari = \beta_1 + \beta_2 exp + \beta_3 exp^2 + u$$

En volta d'aquest model s'ajusta el següent:

$$salari = \beta_1 + \beta_2 exp + v$$

En el segon model de la pertorbació té un component sistemàtic ($v = \beta_3 exp^2 + u$). A la figura 6.5 s'ha representat un diagrama de dispersió (generat pel primer model) i la funció ajustada del segon model. Com es pot veure, per a valors d'*exp* baixos se sobreestimen els salaris; per a valors intermedis d'*exp* es subestimen els salaris; finalment, per a valors elevats d'*exp* el model ajustat sobreestima de nou als salaris. Aquest exemple il·lustra un cas en què l'ús d'una forma funcional incorrecta provoca autocorrelació positiva.

D'altra banda, l'omissió d'una variable rellevant en el model podria induir autocorrelació positiva si aquesta variable té, per exemple, un comportament cíclic..

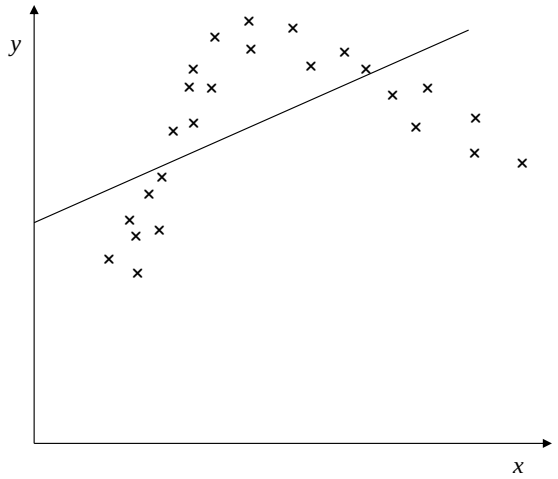


FIGURA 6.6. Pertorbacions autocorrelacionades degudes a un biaix d'especificació.

b) Inèrcia. El terme de pertorbació en una equació de regressió reflecteix la influència de les variables que afecten la variable dependent no incloses en l'equació de regressió. Precisament, la inèrcia o la persistència dels efectes de les variables excloses del model -i incloses en *u*- és probablement la causa més freqüent que hi haja autocorrelació positiva. Com és ben sabut, les sèries temporals macroeconòmiques tals com el *PIB*, la producció, l'ocupació i els índexs de preus tendeixen a moure conjuntament: en períodes d'expansió aquestes sèries tendeixen a augmentar de forma més o menys paral·lela, mentre que en els temps de contracció del cicle tendeixen a disminuir també en una forma paral·lela. Per aquesta raó, en les regressions amb dades de sèries temporals, és molt probable que les observacions successives de la pertorbació depenguen dels valors previs. Amb això, aquest comportament cíclic pot produir autocorrelació en les pertorbacions.

c) Transformació de dades. A modo d'exemple considerem el següent model que explica el consum en funció de la renda:

$$cons_t = \beta_1 + \beta_2 inc_t + u_t \tag{6-50}$$

Per a l'observació *t*-1, obtenim que

$$cons_{t-1} = \beta_1 + \beta_2 inc_{t-1} + u_{t-1} \tag{6-51}$$

Si restem (6-51) de (6-50), obtenim

$$\Delta cons_t = \beta_2 \Delta inc_t + \Delta u_t \tag{6-52}$$

on $\Delta cons_t = cons_t - cons_{t-1}$, $\Delta inc_t = inc_t - inc_{t-1}$ i $v_t = \Delta u_t = u_t - u_{t-1}$.

A l'equació (6-50) se li coneix com equació en forma de *nivells*, mentre que a l'equació (6-52) se li coneix com equació en forma de *primeres diferències*. En l'anàlisi empírica s'utilitzen les dues especificacions. Si la pertorbació no està autocorrelacionada en (6-50), la pertorbació en (6-52), que és igual a $v_t = u_t - u_{t-1}$, sí que ho estarà, ja que v_t i v_{t-1} tenen un element en comú (u_{t-1}). En qualsevol cas, convé advertir que el model (6-52) tal com està especificat pot plantejar altres problemes economètrics que no seran examinats aquí.

6.6.2 Conseqüències de l'autocorrelació

Les conseqüències de l'autocorrelació per *MQO* són similars a les de l'heteroscedasticitat. Per tant, si les pertorbacions estan autocorrelacionades, l'estimador per *MQO* no és *ELNEO*, ja que es pot trobar un altre estimador no esbiaixat alternatiu que tinga menor variança. A més de no ser *ELNEO*, l'estimador obtingut per *MQO* sota el supòsit d'autocorrelació presenta el problema que l'estimació de la matriu de covariances dels estimadors calculada per les fórmules usuals de *MQO* està esbiaixada i, per tant, els estadístics *t* i *F* basats en aquesta matriu de covariances pot portar a inferències errònies.

6.6.3 Contrastos d'autocorrelació

Per realitzar contrastos d'autocorrelació cal especificar la hipòtesi alternativa que definisca un esquema d'autocorrelació de les pertorbacions. A continuació, es van a examinar tres dels més coneguts contrastos. En dos d'ells (el contrast de Durbin i Watson i el contrast *h* de Durbin) la hipòtesi alternativa és un esquema autoregressiu de primer ordre, mentre que en el tercer, denominat contrast de Breusch-Godfrey, és un contrast general d'autocorrelació aplicable a esquemes autoregresius d'ordre més elevat.

Contrast de Durbin i Watson

El contrast *d* de Durbin i Watson va ser proposat per aquests econòmetres en l'any 1950. Per referir-se a aquest estadístic és també usual la denominació de *DW*.

Durbin i Watson proposen el següent esquema sobre les pertorbacions aleatòries u_t:

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad |\rho| < 1 \qquad \varepsilon_t \to NID(0, \sigma^2) \tag{6-53}$$

L'esquema proposat per a les u_t és un esquema autoregressiu de primer ordre, ja que les pertorbacions apareixen com regressant i també com regressor amb un període de desfasament. En la terminologia usual de l'anàlisi de sèries temporals, a l'esquema (6-53) se li denomina *AR*(1), és a dir, un procés autoregressiu d'ordre 1. El coeficient d'aquest esquema és ρ al qual s'exigeix que siga menor que 1 en valor absolut a fi de que les pertorbacions no tinguen un caràcter explosiu, en créixer indefinidament *n*. La variable ε_t és una variable aleatòria per la qual es postula una distribució normal i independent (això és el que vol dir *NID*) amb mitjana 0 i variança σ^2. En conseqüència, sobre la variable u_t es postulen els mateixos supòsits que es van postular per u_t en els supòsits del *MLC*. A la variable que gaudeix d'aquestes propietats se li sol denominar variable soroll blanc.

Segons que el valor de ρ siga positiu o negatiu l'autocorrelació serà positiva o negativa. L'autocorrelació positiva és, amb molta diferència, la qual es presenta amb molta més freqüència en la pràctica. D'altra banda, gairebé sempre es realitzen contrastos d'una sola cua, és a dir, es pren com a hipòtesi alternativa o l'autocorrelació positiva o l'autocorrelació negativa.

La construcció d'un contrast d'autocorrelació de les pertorbacions presenta el problema que aquestes no són observables, de manera que el contrast s'ha de basar en els residus obtinguts per *MQO*. Aquesta circumstància planteja problemes, ja que, sota la hipòtesi nul·la de que les pertorbacions no estan autocorrelacionades, els residus en canvi sí que ho estan. Durbin i Watson, en la construcció del seu contrast, sí van tenir en compte aquesta circumstància.

Vegem ara com s'aplica aquest contrast. Prenent com a referència l'esquema definit en (6-53), Durbin i Watson formulen les següents hipòtesis nul·la i alternativa d'autocorrelació positiva

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho > 0 \end{aligned} \tag{6-54}$$

Així doncs, sota la hipòtesi nul·la es verifica que $u_t=\varepsilon_t$, és a dir, el model compleix els supòsits del *MLC*.

L'estadístic que utilitzen Durbin i Watson per al contrast de les hipòtesis (6-54) és l'estadístic *d*, o *DW*, definit de la següent manera:

$$d = DW = \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})}{\sum_{t=1}^{n}\hat{u}_t^2} \tag{6-55}$$

La distribució de l'estadístic *d*, que és simètrica amb una mitjana igual a 2, és molt complicada, ja que depèn de la forma concreta de la matriu de regressors **X**, de la mida de la mostra (*n*) i del nombre de regressors (*k*) exclòs el terme independent

De tota manera, Durbin i Watson, per a diferents nivells de significació, tabularen dos valors (d_L i d_U) per a cada valor de *n* i de *k*. Les regles per contrastar autocorrelació positiva són les següents:

$$\begin{aligned} &\text{Si } d < d_L \text{ , existeix autocorrelació positiva.} \\ &\text{Si } d_L \le d \le d_U \text{ , no es concloent el contrast.} \\ &\text{Si } d > d_U \text{ , no existeix autocorrelació positiva.} \end{aligned} \tag{6-56}$$

Com es pot veure, hi ha uns valors en els quals el contrast no és concloent. Això es deu a l'efecte que la configuració concreta de la matriu **X** té en la distribució de *d*.

Si es vol fer el contrast d'autocorrelació negativa, la hipòtesi alternativa és la següent:

$$H_1 : \rho < 0 \tag{6-57}$$

Per aplicar el contrast d'autocorrelació negativa es té en compte que l'estadístic *d* té una distribució simètrica amb un recorregut entre 0 i 4. Les regles, per tant, són les següents:

$$\begin{aligned} &\text{Si } d > 4 - d_L \text{ , existeix autocorrelació negativa.} \\ &\text{Si } 4 - d_U \le d \le 4 - d_L \text{ , no es concloent el contrast.} \\ &\text{Si } d < 4 - d_U \text{ , no existeix autocorrelació negativa.} \end{aligned} \tag{6-58}$$

El contrast de Durbin i Watson no és aplicable quan entre els regressors haja variables endògenes desfasades.

Per a la seua aplicació a dades trimestrals, Wallis va considerar el següent esquema autoregressiu de quart ordre:

$$u_t = \rho_4 u_{t-4} + \varepsilon_i \qquad |\rho_4| < 1 \qquad \varepsilon_t \to NID(0, \sigma^2) \tag{6-59}$$

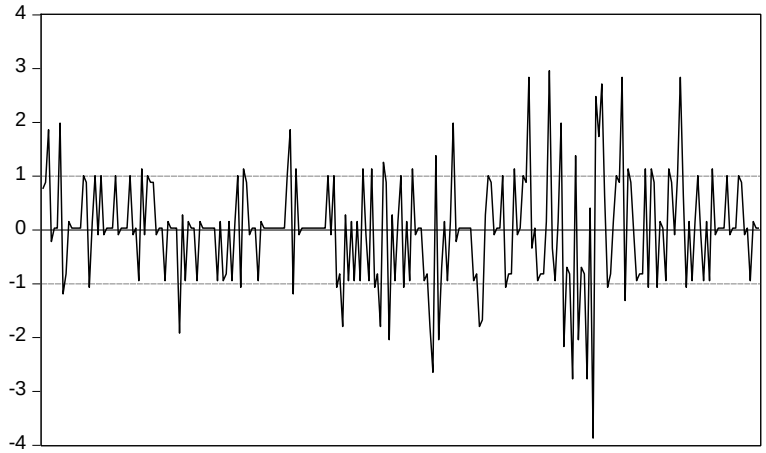
L'anterior esquema és similar a (6-53), amb la diferència que la pertorbació del segon membre està retardada 4 períodes. L'estadístic de contrast de Wallis és similar a (6-55), però tenint en compte que ara el retard és de 4 períodes. Aquest autor va dissenyar unes taules *ad hoc* per a contrastar el model (6-59)-

EXEMPLE 6.11 Autocorrelació en el model per determinar l'eficiència de la Borsa de Madrid

En l'Exemple 4.5 es va formular un model per a determinar l'eficiència de la borsa de Madrid. Per tenir una primera impressió, el gràfic 6.4 mostra els residus estandarditzats[4] corresponents a l'estimació d'aquest model, utilitzant el fitxer *bolmadef*. L'estadístic *DW* és igual a 2.04. (L'estadístic *DW* apareix a la sortida de qualsevol paquet economètric). Com les taules publicades no recullen els valors significatius per a una mida de mostra de 247, utilitzarem els corresponents a n=200 i k'=1. (A la nomenclatura d'aquest contrast s'utilitza **k**' per referir-se al nombre total de regressors exclòs el terme independent). Com la mida de la mostra és molt elevat utilitzarem un nivell de significació α=0.01, és a dir de l'1%. A la tabulació realitzada per Durbin i Watson els valors inferior i superior, que corresponen a les anteriors especificacions, són els següents:

$$d_L = 1.664 \quad ; \quad d_U = 1.684$$

Ja que DW=2.04>d_U, s'accepta la hipòtesi nul·la de que les pertorbacions no estan autocorrelacionades, per a un nivell de significació de l'1%, enfront de la hipòtesi alternativa d'autocorrelació positiva segons l'esquema (6-52).



GRÀFIC 6.4. Residus estandarditzats en l'estimació del model per a determinar l'eficiència de la Borsa de Madrid

EXEMPLE 6.12 Autocorrelació en el model sobre la demanda de peix

En l'Exemple 4.9 es va estimar el model (4-44), utilitzant el fitxer *fishdem*, per explicar la demanda de peix a Espanya. En el gràfic 6.5 es mostren els residus estandarditzats corresponents a l'estimació d'aquest model. De l'examen del gràfic no es desprèn que hi haja un esquema d'autocorrelació apreciable. En aquest sentit, convé assenyalar que, sobre un total de 28 observacions, la línia que uneix els punts dels residus creua l'eix 0 en 11 ocasions, el que és indici d'una certa aleatorietat de la distribució dels residus.

El valor de l'estadístic *DW*, per al contrast de l'esquema (6-52), és 1.202. Per n=28 i k'=3, i per a un nivell de significació de l'1%, s'obtenen els següents valors en la taula tabulada per Durbin i Watson:

$$d_L = 0.969 \quad ; \quad d_U = 1.415$$

Atès que d_L<1.202<d_U, no hi ha evidències suficients ni per acceptar la hipòtesi nul·la, ni per rebutjar-la.

[4] Els residus estandarditzats són igual als residus dividits per $\hat{\sigma}$.

GRÀFIC 6.5. Residus estandarditzats en l'estimació del model de demanda de peix

Contrast h de Durbin

Durbin va proposar el 1970 un estadístic, al que va denominar h, per contrastar les hipòtesis (6-54) en el cas que hi hagi una o més variables endògenes desfasades, que apareguin com a variables explicatives del model. L'expressió de l'estadístic *h* és la següent:

$$h = \hat{\rho}\sqrt{\frac{n}{1 - n\text{var}\ \hat{\beta}_j}} \tag{6-60}$$

on $\hat{\rho}$ és el coeficient de correlació entre $\hat{u}_i$ i $\hat{u}_{i-1}$, n és la mida de la mostra, i var $\hat{\beta}_j$ és la variança corresponent al coeficient de la variable endògena desfasada.

L'estadístic $\hat{\rho}$ pot estimar utilitzant la següent aproximació: $DW = d \simeq 2(1-\hat{\rho})$. En el cas que apareguen com regressors la variable endògena amb diferents desfasaments se seleccionarà la variança corresponent al coeficient de la variable endògena amb menor desfasament.

Sota els supòsits (6-54), l'estadístic *h* té la distribució:

$$h \xrightarrow[n\to\infty]{} N(0,1) \tag{6-61}$$

La regió crítica es troba, doncs, en les cues de la distribució normal: cua de la dreta per a l'autocorrelació positiva i cua de l'esquerra per a l'autocorrelació negativa.

El contrast (6-60) no es pot calcular quan $n\text{var}\ \hat{\beta}_j \geq 1$. En aquest cas Durbin proposa com a alternativa estimar una regressió auxiliar, en la qual es pren com regressant els residus mínim quadràtics i com regressors els mateixos del model original i, a més, els residus desfasats un període. Si el coeficient corresponent als residus desfasats no fos significatiu, es rebutja la hipòtesi alternativa.

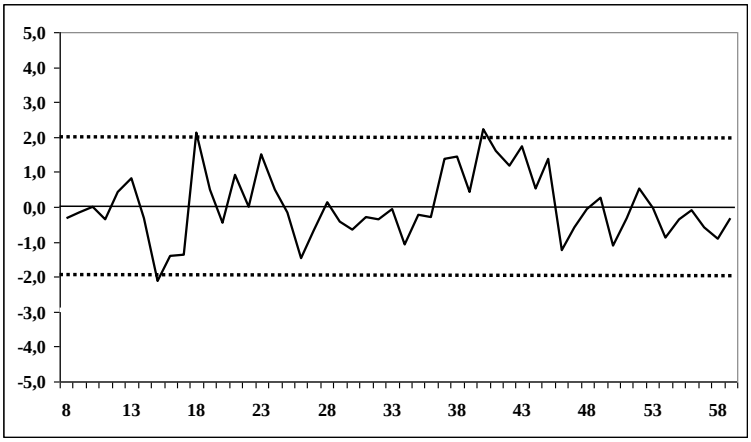
EXEMPLE 6.13 Autocorrelació en el cas de Lydia E. Pinkham

En l'Exemple 5.5 es va examinar el cas Lydia E. Pinkham en el qual es va estimar un model per explicar les vendes d'un extracte herbal, utilitzant el fitxer *pinkham*. A fi de tenir una primera impressió, al gràfic 6.6 es mostra el gràfic dels residus estandarditzats d'aquest model. Com es pot observar, no sembla que els residus es distribuïsquen de forma totalment aleatòria, ja que, per exemple, a partir de 1936 els residus prenen valors positius durant 8 anys consecutius.

El contrast d'autocorrelació idoni per a aquest model és l'estadístic *h* de Durbin, a causa de la presència de la variable endògena desfasada $sales_{t-1}$. L'estadístic *h* és:

$$h = \hat{\rho}\sqrt{\frac{n}{1-n\text{var }\hat{\beta}_j}} = \left[1-\frac{d}{2}\right]\sqrt{\frac{n}{1-n\text{var }\hat{\beta}_j}} = \left[1-\frac{1.2012}{2}\right]\sqrt{\frac{53}{1-53\times 0.0814^2}} = 3.61$$

Donat aquest valor de *h*, es rebutja la hipòtesi nul·la de no autocorrelació, ja que la hipòtesi nul·la es rebutja per α=0.01 i, fins i tot, per α=0.001, d'acord amb la taula de la normal.



GRÀFIC 6.6. Residus estandarditzats en l'estimació del model del cas Lydia E. Pinkham

Contrast de Breusch–Godfrey (BG)

El contrast de Breusch-Godfrey (1978) és un contrast general d'autocorrelació aplicable a esquemes autoregresius d'un ordre superior, i pot utilitzar-se quan hi ha regressors estocàstics com el regressant retardat. Aquest és un contrast asimptòtic al qual també es coneix com el contrast general de *ML* (multiplicadors de Lagrange) per autocorrelació.

En el contrast *BG* s'assumeix que les pertorbacions u_t segueixen un procés autoregressiu d'ordre *p*, *AR*(*p*):

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t \qquad |\rho| < 1 \qquad \varepsilon_t \to NID(0,\sigma^2) \qquad (6\text{-}62)$$

Aquest és simplement una extensió de l'esquema AR (1) del contrast de Durbin i Watson.

Les hipòtesis nul·la i alternativa a contrastar són

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_p = 0$$
$$H_1 : H_0 \text{ no es cierto}$$

El contrast *BG* implica els següents passos:

Pas 1. S'estima el model original i es calculen els residus per *MQO* ($\hat{u}_i$).

Pas 2. S'estima la regressió auxiliar en la qual es pren com regressant als residus ($\hat{u}_i$) i com regressors als regressors del model original i els residus retardats 1, 2, .. i *p* períodes:

$$\hat{u}_t = \alpha_1 + \alpha_2 x_{2t} + \cdots + \alpha_k x_{kt} + \gamma_1 \hat{u}_{t-1} + \cdots + \gamma_1 \hat{u}_{t-p} + \varepsilon_i \qquad (6\text{-}63)$$

La regressió auxiliar hauria de tenir un terme independent, tot i que el model original no ho tingués. D'acord amb (6-63), en la regressió auxiliar hi ha $k+p$ regressors a més del terme independent.

Pas 3. Designant per R^2_{ar} al coeficient de determinació de la regressió auxiliar, es calcula l'estadístic nR^2_{ar}.

Sota la hipòtesi nul·la, l'estadístic *BG* es distribueix de la manera:

$$BG = nR^2_{ar} \xrightarrow[n\to\infty]{} \chi^2_{k+p} \qquad (6\text{-}64)$$

L'estadístic *BG* s'utilitza per realitzar un contrast global del model (6-63). Per a aquest propòsit es pot utilitzar també l'estadístic *F*, encara que en aquest cas només té validesa asimptòtica, com passa amb l'estadístic *BG*.

Pas 4. Per a un nivell de significació α, i designant per $\chi^{2(\alpha)}_{k+p}$ al corresponent valor en la taula χ^2, la decisió a prendre és la següent:

Si $BG > \chi^{2(\alpha)}_{k+p}$ Es rebutja H_0

Si $BG \leq \chi^{2(\alpha)}_{k+p}$ No es rebutja H_0

Com un cas particular el contrast *BG* pot aplicar-se a dades trimestrals utilitzant un esquema $AR(4)$.

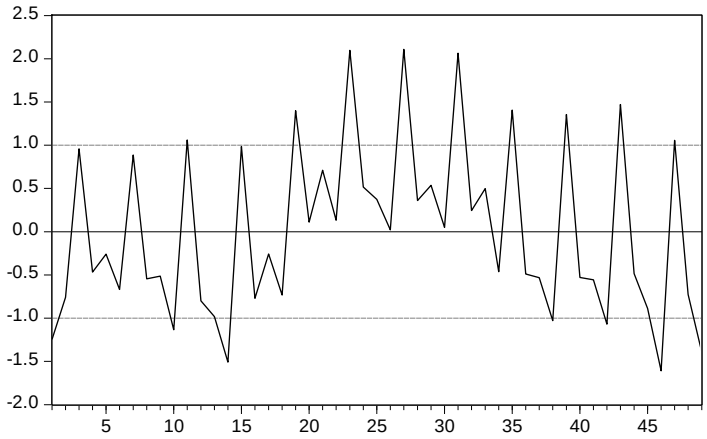
EXEMPLE 6.14 Autocorrelació en un model per explicar les despeses dels residents a l'estranger

Per explicar les despeses dels residents a l'estranger (*turimp*), es va estimar el següent model utilitzant dades trimestrals de l'economia espanyola (arxiu *qnatacsp*):

$$\ln(turimp_t) = \underset{(3.43)}{-17.31} + \underset{(0.276)}{2.0155}\ln(gdp_t)$$

R^2=0.531 *DW*=2.055 *n*=49

on *gdp* és el producte interior brut.



GRÀFIC 6.7. Residus estandarditzats en el model per explicar les despeses dels residents a l'estranger.

El gràfic 6.7 mostra els residus estandarditzats corresponents a aquest model. Com es pot veure, sembla que els residus no estan distribuïts de forma aleatòria, perquè per exemple, s'observen pics cada 4 trimestres, el que és un indicatiu que l'autocorrelació segueix un esquema *AR*(4).

L'estadístic *BG*, calculat per a un esquema *AR*(4), és igual a nR_{ar}^2 =36.35. Donat aquest valor de *BG*, es rebutja la hipòtesi de no autocorrelació per α=0.01, ja que $\chi_5^{2(\alpha)}$ =15.09. En la regressió auxiliar, en la qual s'han utilitzat com a regressors $\hat{u}_{t-1}, \hat{u}_{t-2}, \hat{u}_{t-3}$ i $\hat{u}_{t-4}$, l'únic que ha resultat significatiu ha estat $\hat{u}_{t-4}$.

6.6.4 Errors estàndard *HAC*

Com una extensió dels errors estàndard consistents per heteroscedasticitat de White, examinats en la secció 6.5.2, Newey i West van proposar un mètode conegut com a errors estàndard *HAC* (*heteroskedasticity and autocorrelation consistent*) que permeten corregir els errors estàndard de *MQO* no solament en situacions d'autocorrelació sinó també en cas d'heteroscedasticitat. Recordeu que el procediment de White va ser dissenyat específicament per a heteroscedasticitat. És important ressaltar que el procediment Newey i West és vàlid, estrictament parlant, per a grans mostres i pot no ser apropiat per a xicotetes mostres. Es pot considerar que una mida de 50 observacions és raonablement gran.

EXEMPLE 6.15 Errors estàndard HAC en el cas de Lydia E. Pinkham (Continuació de l'Exemple 6.13)

Donada l'existència d'autocorrelació en el model del cas Lydia E. Pinkham, s'han calculat els errors estàndard d'acord amb el procediment de Newey i West, el que permetrà realitzar correctament contrastos d'hipòtesis sobre els paràmetres. En el quadre 6.9 apareixen els estadístics *t* obtinguts pel procediment convencional i pel procediment *HAC*, així com la ràtio entre tots dos. Com es pot veure les *t* obtingudes pel procediment *HAC* són lleugerament inferiors a les obtingudes pel mètode convencional, amb l'excepció del coeficient d'*advexp*, la *t* sorprenentment és molt més gran quan s'aplica el procediment *HAC*. En qualsevol cas, en realitzar contrastos de significativitat de cada un dels paràmetres s'obtenen exactament les mateixes conclusions per ambdós procediments per als nivells de significació de 0.1, 0.05 i 0.01.

QUADRE 6.9. Estadístiques *t*, convencional i *HAC*, en el cas de Lydia E. Pinkham.

regressor	*t* convencional	*t HAC*	*ràtio*
intercept	2.644007	1.779151	1.49
advexp	3.928965	5.723763	0.69
sales(-1)	7.45915	6.9457	1.07
*d*1	-1.499025	-1.502571	1.00
*d*2	3.225871	2.274312	1.42
*d*3	-3.019932	-2.658912	1.14

6.6.5 Tractament de l'autocorrelació

Per realitzar l'estimació d'un model economètric, on les pertorbacions segueixen l'esquema *AR*(1) considerarem en primer lloc el cas en què ρ és conegut. Aquest és més aviat un supòsit acadèmic que no es presenta a la realitat, però que és convenient adoptar com a supòsit inicial a l'efecte d'exposició. Siga el següent model de regressió lineal múltiple:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \cdots + \beta_k x_{kt} + u_t \qquad (6\text{-}65)$$

Si en (6-65) es considera un desfasament i es multipliquen els dos membres per ρ, s'obté que

$$\rho y_{t-1} = \rho\beta_1 + \rho\beta_2 x_{2,t-1} + \rho\beta_3 x_{3,t-1} + \cdots + \rho\beta_k x_{k,t-1} + \rho u_{t-1} \quad (6\text{-}66)$$

Restant (6-66) de (6-65) s'obté el següent:

$$y_t - \rho y_{t-1} = \beta_{1,1}(1-\rho) + \beta_2\left(x_{2,t} - \rho x_{2,t-1}\right) + \cdots + \beta_k\left(x_{k,t} - \rho x_{k,t-1}\right) + \left(u_t - \rho u_{t-1}\right) \quad (6\text{-}67)$$

Com es pot veure, d'acord amb l'esquema donat a (6-53), el terme de pertorbació de (6-67) compleix amb els supòsits del *MLC*.

El model (6-67) es pot estimar directament per mínims quadrats si es coneix el valor de ρ. Els estimadors obtinguts s'aproximen al mètode de *MQG* si la mostra és prou gran. Estrictament parlant el mètode de *MQG*, consisteix a transformar les observacions 2 a *n* segons l'esquema (6-67) i, a més, en transformar la primera observació de la següent manera:

$$y_t\sqrt{1-\rho^2} = \beta_1\sqrt{1-\rho^2} + \beta_2\sqrt{1-\rho^2}x_{2t} + \cdots + \beta_k\sqrt{1-\rho^2}x_{kt} + \varepsilon_t \quad (6\text{-}68)$$

Quan s'estima ρ conjuntament amb la resta dels paràmetres del model, aleshores al mètode de *MQG* se li denomina *MQG* factibles.

En general, en els diferents mètodes per aplicar *MQG* factibles es fa cas omís de la transformació de la primera observació realitzada a (6-68). Els mètodes de *MQG* factibles per a l'estimació d'un model en què les pertorbacions segueixen un esquema *AR*(1) es poden agrupar en tres blocs: a) mètodes en dues etapes; b) mètodes iteratius; i c) mètodes de rastreig.

A continuació, exposarem dos mètodes corresponents al bloc a), denominats mètode directe i mètode de Durbin en dues etapes.

A la primera etapa del mètode directe i en el mètode proposat per Durbin es procedeix a estimar ρ. En el mètode directe, ρ. s'estima fàcilment a partir de l'estadístic *DW*, utilitzant l'aproximació $DW \simeq 2(1-\hat{\rho})$. En el mètode de Durbin en dues etapes s'estima el següent model de regressió en el qual les variables explicatives són els regressors del model original, els regressors desfasats un període i la variable endògena desfasada un període:

$$y_t = \alpha_1 + \alpha_{2,0}x_{2t} + \alpha_{2,1}x_{2,t-1} + \cdots + \alpha_{k0}x_{kt} + \alpha_{k1}x_{k,t-1} + \rho y_{t-1} + \upsilon_t \quad (6\text{-}69)$$

El coeficient de la variable endògena desfasada és precisament el paràmetre ρ. En la primera etapa, s'estima el model (6-69) per *MQO*, prenent del mateix l'estimació de ρ. En la segona etapa, aplicable als dos mètodes, es transforma el model amb l'estimació de ρ calculada en la primera etapa de la següent manera:

$$y_t - \hat{\rho}y_{t-1} = \beta_1(1-\hat{\rho}) + \beta_2\left(x_{2t} - \hat{\rho}x_{2,t-1}\right) + \cdots + \beta_k\left(x_{kt} - \hat{\rho}x_{k,t-1}\right) + \xi_t \quad (6\text{-}70)$$

Aplicant MQO al model transformat s'obtenen les estimacions dels paràmetres. Una exposició dels mètodes iteratius i de rastreig es pot veure a Uriel, E .; Contreras, D; Moltó, M. L. i Peiró, A. (1990): *Econometria. El model lineal*. Editorial AC. Madrid.

Exercicis

Exercici 6.1 Considerem el següent model poblacional**:**

$$y_i = \beta_1 + \beta_2 x_i + u_i \quad (1)$$

En el seu lloc, es va estimar el següent model estimat:

$$\tilde{y}_i = \tilde{\beta}_2 x_{2i} \tag{2}$$

¿És $\tilde{\beta}_2$, obtingut en aplicar *MQO* a (2), un estimador no esbiaixat de β_2?

Exercici 6.2 Considerem el següent model poblacional:

$$y_i = \beta_2 x_i + u_i \tag{1}$$

En el seu lloc, es va estimar el següent model estimat:

$$\tilde{y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 x_{2i} \tag{2}$$

¿És $\tilde{\beta}_2$, obtingut en aplicar *MQO* a (2), un estimador no esbiaixat de β_2?

Exercici 6.3 Siguen els següents models:

$$imp = \beta_1 + \beta_2 gdp + \beta_3 rpimp + u \tag{1}$$

$$\ln(imp) = \beta_1 + \beta_2 \ln(gdp) + \beta_3 \ln(rpimp) + u \tag{2}$$

on *imp* és la importació de béns, *gdp* és el producte interior brut a preus de mercat, i *rpimp* són els preus relatius importacions/pib. Les magnituds *imp* i *gdp* estan expressades en milions de pessetes.

a) Utilitzant una mostra del període 1971-1997 per a Espanya (arxiu *importsp*), estime els models (1) i (2).
b) Interpretació dels coeficients β_2 i β_3 en tots dos models.
c) Aplique el procediment RESET al model (1).
d) Aplique el procediment RESET al model (2).
e) Utilitze l'especificació més adequada usant els valors *p* obtinguts en les seccions b) i c).

Exercici 6.4 Considere el següent mode de demanda d'aliments

$$alim = \beta_1 + \beta_2 pr + \beta_3 renda + u$$

on *alim* és la despesa en aliments, *pr* són els preus relatius i *renda* és la renda disponible.

L'investigador A omet per oblit la variable renda, obtenint la següent estimació del model:

$$\underset{}{alim_i} = \underset{(11.85)}{89.97} + \underset{(0.118)}{0.107}\ pr_i$$

L'investigador B, que és més acurat, obté la següent estimació del model:

$$\underset{}{alim_i} = \underset{(5.84)}{92.05} - \underset{(0.067)}{0.142}\ pr_i + \underset{(0.031)}{0.236}\ renda_i$$

(Entre parèntesi figuren desviacions típiques)

Al llarg de la discussió entre tots dos investigadors sobre quin dels dos models estimats és el més adequat, l'investigador A tracta de justificar el seu oblit, atribuint l'omissió de la variable renda al problema de la multicolinealitat.

a) En favor de quin dels investigadors s'inclinaria vostè, a la vista dels resultats obtinguts. Argumente raonadament el seu punt de vista.

b) Obtinga analíticament l'expressió del biaix d'estimació de l'estimador del paràmetre β_2 en el model amb error d'especificació per omissió de variable rellevant.

Exercici 6.5 Per estimar una funció de producció s'ha formulat el següent model

$$\ln(output) = \beta_1 + \beta_2 \ln(labor) + \beta_3 \ln(capital) + u$$

on *output* és la quantitat d'output produït, *labor* és la quantitat de mà d'obra, i *capital* és la quantitat de capital

Es disposa de les següents observacions corresponents a 9 empreses:

$output_i$	230	140	180	270	300	240	230	350	120
$labor_i$	30	10	20	40	50	20	30	60	40
$capital_i$	160	50	100	200	240	190	160	300	150

Un investigador estima el model prenent equivocadament només 8 observacions, i obté els següents resultats:

$$output_i = \underset{(1.956)}{97.259} + \underset{(0.124)}{0.970}\, labor_i + \underset{(0.027)}{0.650}\, capital_i$$

$$R^2 = 0.999; \quad F=3422$$

Els valors entre parèntesis són els errors estàndard dels estimadors i l'estadístic *F* correspon al contrast global del model.

Quan s'adona de l'error comès, estima el model amb totes les observacions (n=9), obtenint en aquest cas els següents resultats:

$$output_i = \underset{(32.046)}{75.479} - \underset{(1.742)}{1.970}\, labor_i + \underset{(0.376)}{1.272}\, capital_i$$

$$R^2 = 0.824 \qquad F= 14.056$$

El seu desconcert és gran en comparar les dues estimacions, i no pot comprendre com, per utilitzar una sola observació més, els resultats obtinguts arriben a ser tan diferents. Pot trobar alguna explicació que puga justificar aquestes diferències?

Exercici 6.6 Suposem que en el model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

el *R* quadrat obtingut en la regressió de x_1 sobre x_2, al qual anomarem $R^2_{1/2}$, és zero.

D'altra banda, si estima els següents models:

$$y = \lambda_0 + \lambda_1 x_1 + u$$

$$y = \gamma_0 + \gamma_1 x_2 + u$$

a) Serà $\hat{\lambda}_1$ igual a $\hat{\beta}_1$ i $\hat{\gamma}_1$ igual a $\hat{\beta}_2$?

b) Serà $\hat{\beta}_0$ igual a $\hat{\lambda}_0$ o $\hat{\beta}_0$ igual a $\hat{\gamma}_0$?

c) Serà var($\hat{\lambda}_1$) igual a var($\hat{\beta}_1$) i var($\hat{\gamma}_1$) igual a var($\hat{\beta}_2$)?

Exercici 6.7 Un analista desitja estimar el següent model utilitzant les observacions del quadre adjunt:

$$y_i = e^{\beta_1} x_{2i}^{\beta_2} x_{3i}^{\beta_3} x_{4i}^{\beta_4} e^{u_i}$$

x_2	x_3	x_4
3	12	4
2	10	5
4	4	1
3	9	3
2	6	3
5	5	1

Quins problemes es poden presentar a l'estimació d'aquest model amb aquestes dades?

Exercici 6.8 En l'Exercici 4.8, utilitzant el fitxer *airqualy*, es va estimar el següent model:

$$\begin{aligned} airqual_i = \underset{(10.19)}{97.35} + \underset{(0.0311)}{0.0956}\, popln_i - \underset{(0.0055)}{0.0170}\, medincm_i - \underset{(0.0089)}{0.0254}\, poverty_i \\ - \underset{(0.0017)}{0.0031}\, fueoil_i - \underset{(0.0025)}{0.0011} valadd_i \end{aligned}$$

$$R^2 = 0.415 \qquad n = 30$$

a) Calculeu l'estadístic *FEV* per a cada coeficient.
b) Quina és la seua conclusió?

Exercici 6.9 Per examinar els efectes dels rendiments de l'empresa sobre els salaris dels directors executius s'ha formulat el següent model:

$$\ln(salary) = \beta_1 + \beta_2 roa + \beta_3 \ln(sales) + \beta_4 profits + \beta_5 tenure + \beta_6 age + u$$

on *roa* és la ràtio beneficis/actius expressats en percentatge, *tenure* és el nombre d'anys com a conseller delegat a l'empresa (= 0 si és menys de 6 mesos), i *age* és l'edat en anys. Els salaris estan expressats en milers de dòlars, i *sales* (vendes) i *profits* (beneficis) en milions de dòlars.

a) Utilitzant una mostra de 447 observacions del fitxer *ceoforbes*, estime el model per *MQO*.
b) Aplique el contrast de normalitat als residus.
c) Utilitzant les 60 primeres observacions, estime el model per *MQO*. Compare els coeficients i el R^2 d'aquesta estimació amb els obtinguts en l'apartat *a)*. Quina és la seva conclusió?
d) Aplique el contrast de normalitat als residus obtinguts en l'apartat *c)*. Quina és la seua conclusió al comparar aquest resultat amb l'obtingut en l'apartat

b)? Quina és la seva conclusió al comparar aquest resultat amb l'obtingut en l'apartat *b)*?

Exercici 6.10 Siga el model

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad [1]$$

sent

$$\sigma_i^2 = \sigma^2 x_i, \quad x_i > 0, \quad \forall i$$

a) Aplíque *MQG* al model [1] per estimar β_i.
b) Calcule la variança de l'estimador per *MQG*:

Exercici 6.11 Siga el model

$$y_i = \beta x_i + u_i \qquad [1]$$

on

$$\sigma_i^2 = \sigma^2 x_i, \quad x_i > 0, \quad \forall i$$

a) Estime β del model [1] per mínims quadrats generalitzats.

b) Calcule la variança del estimador obtingut.

Exercici 6.12 Siga el model

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad [1]$$

on la variança de las pertorbacions és igual a

$$\sigma_i^2 = \sigma^2 x_i, \quad x_i > 0, \quad \forall i$$

1) Aplicant *MQO* al model [1] i tenint en compte els supòsits Gauss-Markov, la variança de l'estimador, d'acord amb (2-16) és

$$\frac{\sigma^2}{\sum (x_i - \bar{x})^2} \qquad [2]$$

2) Aplicant *MQO* al model [1] i tenint en compte que i la resta de supòsits Gauss-Markov, la variança de l'estimador és aleshores igual a

$$\frac{\sigma^2 \sum (x_i - \bar{x})^2 x_i}{(\sum (x_i - \bar{x})^2)^2} \qquad [3]$$

3) Aplicant *MQG* al model [1] i tenint en compte que $\sigma_i^2 = \sigma^2 x_i$, i la resta de supòsits Gauss-Markov, la variança de l'estimador és

$$\frac{\sigma^2}{\sum \frac{(x_i - \bar{x})^2}{x_i}} \qquad [4]$$

a) ¿Són correctes les variances [2] i [3]?

b) Demostre que [4] és menor o igual que [3]. (Consell: Aplique la desigualtat Cauchy-Schwarz que diu que $\left[\sum w_i z_i\right]^2 \leq \left[\sum w_i^2\right]\left[\sum z_i^2\right]$ és veritat)

Exercici 6.13 Siga el model

$$hostel = \alpha_1 + \alpha_2 renda + u$$

on *hostel* és la despesa en hostaleria i *renda* és la renda anual disponible

Es disposa de la següent informació sobre 9:

família	*hostel*	*renda*
1	13	300
2	3	200
3	38	700
4	47	900
5	14	400
6	18	500
7	25	800
8	1	100
9	21	600

Les variables *hostel* i *renda* estan expressades en milers de pessetes.

a) Estime el model per *MQO*.

b) Aplique el contrast d'heteroscedasticitat de White.

c) Aplique el contrast d'heteroscedasticitat de Breusch-Pagan-Godfrey.

d) Li apareix adequat utilitzar els anteriors contrastos d'heteroscedasticitat en aquest cas?

Exercici 6.14 Amb referència al model de l'Exercici 4.5, se suposa ara que

$$\text{var}(\varepsilon_i) = \sigma^2 \ln(y_i)$$

a) Són, en aquest cas, no esbiaixats dels estimadors obtinguts per *MQO*?

b) Són eficients els estimadors *MQO*?

c) Podria suggerir un estimador millor que *MQO*?

Exercici 6.15 Indique quines de les següents afirmacions són veritat, justificant les respostes, quan existeix heteroscedasticitat:

a) Els estimadors *MQO* deixen de ser estimadors *ELNE*O.

b) Els estimadors *MQO* $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \cdots, \hat{\beta}_k$ son inconsistents.

c) Els contrastos convencionals *t* i *F* són no vàlids.

Exercici 6.16 En l'Exercici 3.19, utilitzant l'arxiu *consumsp*, es va estimar el model de Brown per a l'economia espanyola en el període 1954-2010. Els resultats obtinguts van ser els següents:

$$\underset{}{conspc_t} = \underset{(84.88)}{-7.156} + \underset{(0.0857)}{0.3965}\, incpc_t + \underset{(0.0903)}{0.5771}\, conspc_{t-1}$$

Utilitzant els residus de l'anterior model ajustat, es va obtenir la següent regressió:

$$(\hat{u}_t^2) = 141568 + 89.71 incpc_t - 149.2 conspc_{t-1}$$
$$-0.183 incpc_t^2 - 0.221 conspc_{t-1}^2 + 0.406 incpc_t \times conspc_{t-1}$$

$$R^2 = 0.285$$

a) Existeix heteroscedasticitat en aquesta funció de consum?

b) Es va obtenir la següent estimació, amb errors estàndard consistents per heteroscedasticitat de White*:*

$$\underset{}{conspc_t} = \underset{(66.92)}{?} + \underset{(0.0669)}{?}\, incpc_t + \underset{(0.0741)}{?}\, conspc_{t-1}$$

Pot omplir els espais amb interrogant? Si us plau, feu-ho.

Explique la diferència entre els errors estàndard consistents per a l'heteroscedasticitat de White i els errors estàndard usuals.

c) Contraste si el coeficient d'*incpc* és igual a 5. Quins errors estàndard utilitzaria en el procés d'inferència? Per què?

Exercici 6.17 Suponga la següent especificació:

$$c_i = \gamma_1 + \gamma_2 h_i + \gamma_3 m_i + u_i$$
$$\sigma_i^2 = \sigma^2 h_i^2$$

Seria adequat per eliminar l'heteroscedasticitat realitzar la següent transformació del model?

$$\frac{c_i}{h_i} = \gamma_1 + \gamma_2 h_i + \gamma_3 m_i + u_i \; ?$$

Raone la resposta.

Exercici 6.18 Siga el model

$$y = \beta_1 + \beta_2 x + u$$

i es disposa de la següent informació:

y_i	x_i	$\hat{u}_i$
2	-3	1.37
3	-2	-0.42
7	-1	0.79
6	0	-3.00
15	1	3.21
8	2	-6.58
22	3	4.63

a) Aplique el contrast d'heteroscedasticitat de White.
b) Aplique el contrast d'heteroscedasticitat de Breusch-Paguen-Godfrey.
c) Per què la significació obtinguda en els dos contrastos és tan diferent?

Exercici 6.19 Responga a las següents preguntes

a) Explique detalladament en què consisteix el problema de l'heteroscedasticitat en el model de regressió lineal.
b) Il·lustre breument el problema de l'heteroscedasticitat amb un exemple.
c) Propose solucions al problema de l'heteroscedasticitat.

Exercici 6.20 Utilitzant una mostra corresponent a 17 regions s'han obtingut les següents estimacions:

$$\hat{y}_i = -309.8 + 0.76z_i + 3.05h_i \qquad R^2 = 0.989$$

$$\hat{u}_i^2 = -1737.2 - 17.8z_i + 0.09z_i^2 + 0.65z_ih_i + 10.6h_i - 0.31h_i^2 \qquad R^2 = 0.705$$

on *y* és la despesa en educació, *z* és el *PIB* i *h* es el nombre d'habitants.

a) Existeix un problema d'heteroscedasticitat? Detall el procediment de contrast.
b) Suposant que es detectés la presència d'heteroscedasticitat en el model de regressió, quina solució adoptaria per analitzar la significativitat de les variables explicatives del model? Raone la resposta.

Exercici 6.21 Utilitzant dades de l'economia espanyola per al període 1971-1997 (arxiu *importsp*), es va estimar el següent model per explicar les importacions (*imp*):

$$\ln(imp_t) = \underset{(3.65)}{-26.58} + \underset{(0.210)}{2.4336}\ln(gdp_t) - \underset{(0.0232)}{0.4494}\ln(rpimp_t)$$

$$R^2=0.997 \quad n=27$$

on *gdp* és el producte interior brut a preus de mercat, i *rpimp* són els preus relatius importacions/pib. Les variables *imp* i *gdp* estan expressades en milions de pessetes.

a) Formule i estimi la regressió auxiliar per realitzar el contrast d'heteroscedasticitat de Breusch-Paguen-Godfrey.
b) Aplique el contrast d'heteroscedasticitat de Breusch-Paguen-Godfrey utilitzant la regressió formulada en la secció *a).*
c) Formule i estime la regressió auxiliar per a realitzar el contrast complet de White d'heteroscedasticitat.
d) Aplique el contrast d'heteroscedasticitat complet de White utilitzant la regressió formulada en la secció *c).*
e) Formule i estime la regressió auxiliar per a realitzar el contrast simplificat d'heteroscedasticitat de White.
f) Aplique el contrast d'heteroscedasticitat simplificat de White utilitzant la regressió formulada en la secció *e).*
g) Compare els resultats dels contrastos realitzats en les seccions *b*), *d*) i *f*).

Exercici 6.22 Utilitzant dades de l'arxiu *tradocde*, es va estimar el següent model per explicar les importacions (*impor*) en els països de l'OCDE:

$$\ln(impor_i) = \underset{(6.67)}{18.01} + \underset{(0.658)}{1.6425}\ln(gdp_i) - \underset{(0.636)}{0.5151}\ln(popul_i)$$

$$R^2=0.614 \qquad n=34$$

on *gdp* és el producte interior brut a preus de mercat, i *popul* és la població de cada país.

a) Quina és la interpretació del coeficient de *gdp?*
b) Formule i estime la regressió auxiliar per a realitzar el contrast de White d'heteroscedasticitat.
c) Aplique el contrast d'heteroscedasticitat de White utilitzant la regressió formulada en la secció *b).*
d) Contrast si l'elasticitat *import/gdp* és més gran que 1. Per realitzar aquest contrast, necessita utilitzar els errors estàndard consistents per a l'heteroscedasticitat de White?

Exercici 6.23 Expliqueu detalladament quin seria el contrast d'autocorrelació apropiat en cada situació:

a) Quan el model no té variables endògenes retardades i les observacions són anuals.
b) Quan el model té variables endògenes retardades i les observacions són anuals.
c) Quan el model no té variables endògenes retardades i les observacions són trimestrals

Exercici 6.24 S'han estimat dos models alternatius del cost mitjà de producció anual d'automòbils d'una determinada marca en el període 1980-1999.

$$c = \alpha + \beta p + u \qquad R^2 = 0.848; \quad \bar{R}^2 = 0.812; \quad d = DW = 0.51$$
$$c = \alpha + \beta p + \gamma p^2 + u \qquad R^2 = 0.852; \quad \bar{R}^2 = 0.811; \quad d = DW = 2.11$$

a) En comparar les dues estimacions, indique si s'observa algun problema economètric. Explique.
b) En funció de la seua resposta a l'apartat anterior, Quin dels dos models triaria?

Exercici 6.25 En el període 1950-1980 s'ha estimat la següent funció de producció:

$$\ln(o_t) = \underset{(0.24)}{-3.94} + \underset{(0.083)}{1.45}\ln(l_t) + \underset{(0.048)}{0.38}\ln(k_t)$$

$$R^2 = 0.994 \qquad DW = 0.858 \qquad \hat{\rho} = 0.559$$

on *o* és la producció, *l* es el treball, i *k* és el capital.

(Els números entre parèntesis són les desviacions estàndard dels estimadors)

a) Contrast detalladament l'existència d'autocorrelació.
b) Si el model tingués una variable endògena retardada com a variable explicativa indique de quina manera contrastaria l'autocorrelació.

Exercici 6.26 Utilitzant una mostra de 38 observacions de periodicitat anual s'ha estimat la següent funció de demanda d'un producte

$$d_i = 2.47 + \underset{(0.39)}{0.35}\, p_i + \underset{(0.06)}{0.9}\, d_{i-1} \qquad R^2 = 0.98 \qquad DW = 1.82$$

on *d* és la quantitat demandada i *p* és el preu.

(Els números entre parèntesis són les desviacions estàndard dels estimadors).

a) Existeix un problema d'autocorrelació? Raone la resposta.

b) Enumere les condicions sota quals seria adequat utilitzar el contrast de Durbin Watson.

Exercici 6.27 S'ha estimat el següent model de demanda d'habitatge amb observacions anuals corresponents al període 1960-1994:

$$\ln(v_t) = \underset{(0.15)}{-0.39} + \underset{(0.05)}{0.3}\ln(r_t) - \underset{(0.02)}{0.67}\ln(p_t) + \underset{(0.04)}{0.70}\ln(v_{t-1})$$

$$R^2 = 0.999 \qquad DW = 0.52$$

on *v* és la despesa en habitatge, *r* és la renda disponible, *p* és el preu de l'habitatge.

(Els números entre parèntesis són les desviacions estàndard dels estimadors).

a) Existeix un problema d'autocorrelació? Raone la resposta.
b) Enumere les condicions sota quals seria adequat utilitzar el contrast de Durbin Watson.

Exercici 6.28 Conteste a les següents preguntes:

a) En un model per explicar les vendes es realitza l'estimació utilitzant dades trimestrals. Explique com pot contrastar si existeix autocorrelació.
b) Descriga detalladament, introduint els supòsits que considere oportuns, com estimaria el model quan es rebutja la hipòtesi nul·la de no autocorrelació.

Exercici 6.29 En l'estimació de la funció de consum keynesiana de l'economia francesa s'han obtingut els següents resultats:

$$consum_t = \underset{(-0.73)}{-485.22} + \underset{(79.39)}{0.913}\, renda_t$$

$$R^2 = 0.9936 \quad \text{DW=0.4205} \qquad n\text{=}30$$

(Els números entre parèntesis són els estadístics t dels estimadors).

Un investigador considera que s'ha de centrar l'atenció en la funció d'estalvi, en lloc de fer-ho en la funció de consum. En conseqüència, proposa el següent model:

$$estalvi_t = \alpha_1 + \alpha_2 renda + v_t \qquad [1]$$

on

$$estalvi_t = renda_t - consum_t$$

Utilitzant la informació donada en el present Exercici, si això és possible:

a) Obtinga les estimacions de α_1 i α_2.

b) Estime les variances de $\hat{\alpha}_1$ i $\hat{\alpha}_2$.

c) Calculeu l'estadístic DW (Durbin-Watson) del model d'estalvi.

d) Calcule el coeficient R^2 per al model d'estalvi.

Exercici 6.30 Siga el model

$$\begin{aligned} y_t &= \beta x_t + u_t \\ u_t &= \rho u_{t-1} + \varepsilon_t; \quad E\left[\varepsilon_t^2\right] = \sigma^2 \qquad \forall i \end{aligned} \tag{1}$$

a) Si el model [1] es transforma prenent primeres diferències, sota quins supòsits resulta avantatjosa l'estimació per *MQO* del model transformat pel que fa a l'estimació per *MQO* del model [1]?

b) És adequat utilitzar el R^2 per comparar el model [1] i el model transformat? Explique la seva resposta.

Exercici 6.31 Siga el modelo

$$y_t = \beta_1 + \beta_2 x_t + u_t \tag{1}$$

S'obté la següent mostra d'observacions per a les variables *x* i *y*:

y_i	6	3	1	1	1	4	6	16	25	36	49	64
x_i	-4	-3	-2	-1	1	2	3	4	5	6	7	8

a) Estime el model [1] per *MQO* i calcule el corresponent coeficient de determinació corregit.

b) Calcule l'estadístic de Durbin-Watson corresponent a l'estimació realitzada en *a).*

c) A la vista de contrast de Durbin i Watson i de la representació de la recta ajustada i dels residus, és convenient reformular el model [1]? Justifique la resposta i, en cas que aquesta siga afirmativa, estime el model alternatiu que es considere més adequat a les dades.

Exercici 6.32 En el següent model:

$$y_t = \beta_1 + \beta_2 x_t + u_t$$

$$u_t = \rho u_{t-1} + \varepsilon_t; \qquad \varepsilon_t \sim NI\ \ 0, \sigma^2$$

La següent informació addicional està també disponible:

$$\rho = 0.5$$

y_i	22	26	32	31	40	46	46	50
x_i	4	6	10	12	13	16	20	22

a) Estime el model per *MCO*.

b) Estime el model per *MQG* sense la transformació de la primera observació.

c) Quin dels dos estimadors de β_2 és més eficient?

Exercici 6.33 En un estudi sobre la demanda d'un producte s'han obtingut els següents resultats:

$$\hat{y}_t = \underset{(7.17)}{2.30} + \underset{(0.05)}{0.86}\, x_t$$

$$R^2 = 0.9687 \quad \text{DW=3.4} \quad n = 15$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

A més, es disposa de la següent informació addicional sobre les regressions dels errors, en valor absolut:

1. $|\hat{u}_t| = \underset{(0.210)}{0.167} + \underset{(0.180)}{0.127}\, x_t$
2. $|\hat{u}_t| = \underset{(0.098)}{0.231} + \underset{(0.095)}{0.218}\, x_t^{1/2}$

a) Detecte si existeix autocorrelació.
b) Detecte si existeix heteroscedasticitat.
c) Quin seria el procediment més adequat per evitar el possible problema d'heteroscedasticitat?

Exercici 6.34 Utilitzant una mostra per al període 1971-1997 (arxiu *importsp*), es va estimar el següent model, utilitzant errors estàndard *HAC*, per explicar les importacions de béns a Espanya (*imp*):

$$\ln(imp_t) = \underset{(3.65)}{-26.58} + \underset{(0.210)}{2.434}\ln(gdp_t) - \underset{(0.023)}{0.4494}\ln(rpimp_{t-1})$$

$$R^2 = 0.997 \quad \text{DW=0.73} \quad n = 27$$

(Els números entre parèntesis són els errors estàndard dels estimadors.)

A més, es disposa de la següent informació addicional sobre les regressions dels errors, en valor absolut:

a) Interprete el coeficient de *rpimp*.
b) Hi ha autocorrelació en aquest model?
c) Contraste si l'elasticitat *imp/gdp* més quatre vegades l'elasticitat *imp/rpimp* és igual a 0. (Informació addicional: $\text{var}(\hat{\beta}_2)$ =0.044247; $\text{var}(\hat{\beta}_3)$ =0.000540; i $\text{var}(\hat{\beta}_2, \hat{\beta}_3)$ =0.004464).
d) Contraste la significació global.

Exercici 6.35 Utilitzant una mostra per al període 1954-2009 (arxiu *electsp*), es va estimar el següent model per explicar el consum d'electricitat a Espanya (*conselec*):

$$\ln(conselec_t) = \underset{(0.46)}{-9.98} + \underset{(0.035)}{1.469}\ln(gdp_t)$$

$$R^2 = 0.9805 \quad \text{DW=0.18} \quad n = 37 \qquad (1)$$

on *gdp* és el producte interior brut a preus de mercat. La variable *conselec* està expressada en milers de tones equivalents de petroli *(ktep*) i *gdp* està expressat en milions de pessetes.

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Contraste si hi ha autocorrelació mitjançant l'aplicació de l'estadístic Durbin-Watson.
b) Contraste si hi ha autocorrelació mitjançant l'aplicació de l'estadístic Breusch-Godfrey per a un esquema *AR*(2).
c) També va ser estimat el següent model:

$$\ln(conselec_t) = \underset{(0.75)}{-0.917} + \underset{(0.107)}{0.164}\ln(gdp_t) + \underset{(0.072)}{0.871}\ln(conselec_{t-1})$$

$$R^2 = 0.997 \quad \text{DW=0.93} \quad n = 36 \qquad (2)$$

Contraste si hi ha autocorrelació mitjançant l'aplicació del procediment que estime oportú.
d) Contraste si l'elasticitat *conselec/gdp* en una situació d'equilibri $(\ln(conselec^e) = \beta_1 + \beta_2 \ln(gdp^e) + \beta_3 \ln(conselec^e))$ és més gran que 1 utilitzant un procediment adequat.

Exercici 6.36 La corba de Phillips representa la relació entre la taxa d'inflació (*inf*) i la taxa d'atur (*unemp*). Mentre que a curt termini s'ha observat un *tradeoff* estable entre atur i inflació, aquest fenomen no s'ha constatat a llarg termini.

El següent model reflecteix la corba de Phillips estàtica:

$$inf = \beta_1 + \beta_2 unempl + u$$

Utilitzant una mostra de l'economia espanyola per al període 1970-2010 (arxiu *phillipsp*), es van obtenir els següents resultats:

$$inf_t = \underset{(1.79)}{12.59} - \underset{(0.120)}{0.3712}\, unempl_t$$

R^2=0.198; DW=0.219; n=41

(Els números entre parèntesis són els errors estàndard dels estimadors.)

a) Interprete el coeficient d'*unempl*.
b) Contraste si hi ha autocorrelació de primer ordre mitjançant l'aplicació de l'estadístic Durbin-Watson.
c) Utilitzant la informació que té disponible fins ara, pot contrastar de forma adequada el coeficient d'*unempl*?
d) Utilitzant els errors estàndard HAC, contrast la significació del coeficient d'*unempl*.

Exercici 6.37 És important remarcar que la corba de Phillips té un caràcter relatiu. La inflació és considerada alta o baixa en relació a la taxa d'inflació esperada i la desocupació és considerada alta o baixa en relació amb la denominada taxa natural d'atur. A la corba *augmentada* de Phillips tot això es té en compte:

$$inf_t - inf^e_{t/t-1} = \beta_2(unempl_t - \lambda_0) + u_t$$

on λ_0 és la taxa natural d'atur i $inf^e_{t/t-1}$ és la taxa d'inflació esperada en *t* i formada en *t*-1. Si considerem que taxa esperada per *t* és igual a la inflació en *t*-1 ($inf^e_{t/t-1} = inf_{t-1}$) i fent $\beta_1 = -\beta_2\lambda_0$, la corba augmentada de Phillips pot expressar-se així:

$$inf_t - inf_{t-1} = \beta_1 + \beta_2 unempl_t + u_t$$

a) Utilitzant l'arxiu *phillipsp*, estime el model anterior.
b) Interprete el coeficient d'*unempl.*
c) Contraste si hi ha correlació de segon ordre.
d) Contraste si la taxa natural d'atur és més gran que 10.

Apèndix 6.1

En primer lloc anem a expressar l'estimador tenint en compte que i ha estat generada pel model (6-8):

$$\begin{aligned}
\tilde{\beta}_2 &= \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)(y_i-\bar{y})}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} = \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)y_i}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \\
&= \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)(\beta_1+\beta_2 x_{1i}+\beta_3 x_{2i}+u_i)}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \\
&= \beta_2 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{1i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)u_i}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \\
&= \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)u_i}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2}
\end{aligned} \tag{6-71}$$

Si prenem esperança en els dos membres de (6-71), obtenim que

$$\begin{aligned}
E(\tilde{\beta}_2) &= \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)E(u_i \mid x_2, x_3)}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \\
&= \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2}
\end{aligned} \tag{6-72}$$