Incompliment de les hipòtesis bàsiques del model de regressió amb R

Daniel Liviano Solís
Maria Pujol Jover

PID_00211044

UOC
Universitat Oberta de Catalunya

Índex

Introducció 5

Objectius 6

1. Propietats de l'estimació del model 7
1.1. Estimador MCO i la condició d'ortogonalitat 7
1.2. Biaix i consistència 7
1.3. Eficiència 9
1.4. Causes del biaix i de la inconsistència 9
1.4.1. Errors de mesura 9
1.4.2. Endogenitat 11

2. Heteroscedasticitat i autocorrelació 13
2.1. Definició teòrica 13
2.2. Exemple pràctic 14
2.3. Estimació eficient de l'MVC 17

3. Errors en la mostra 26
3.1. Multicol·linealitat 26
3.2. Observacions atípiques 32

4. Permanència estructural 37

Bibliografia 43

Introducció

En el primer mòdul hem estudiat com s'ha implementat l'estimador de mínims quadrats ordinaris (MCO) per a obtenir una estimació dels paràmetres d'un model de regressió. L'estimador MCO és la manera més simple i directa d'obtenir una estimació, però perquè sigui vàlida cal que es compleixin una sèrie de requisits (o restriccions) en les dades i en el model construït. Desafortunadament, molt sovint aquests requisits no es compleixen, de manera que cal acudir a altres tècniques per a obtenir una estimació fiable.

El primer capítol d'aquest mòdul és un repàs teòric de les propietats de l'estimació d'un model economètric: ortogonalitat, biaix, consistència i eficiència. El segon capítol s'encarrega del problema de l'eficiència d'una estimació, això és, la seva variança. D'aquesta manera, s'introdueixen les definicions d'heteroscedasticitat i autocorrelació, fenòmens que fan que la matriu de variàncies i covariàncies de l'estimació no sigui esfèrica. A més a més, amb un exemple s'estudia com es detecta i corregeix aquests fenòmens amb R i amb R-Commander. El tercer capítol analitza el fenomen d'errors en la mostra. La primera part estudia la multicol·linealitat, fenomen que apareix quan entre els regressors hi ha variables altament correlacionades entre si, cosa que dificulta l'estimació i mostra resultats erronis. La segona part analitza què succeeix quan hi ha observacions atípiques, això és, molt allunyades de la resta de les observacions. Finalment, el quart capítol es dedica a l'anàlisi de la permanència estructural, és a dir, si una mateixa estimació és vàlida per a totes les dades de la mostra o, al contrari, s'ha de dividir la mostra en diversos fragments, ja que entre aquests es detecta una relació funcional diferent.

Objectius

1. Comprendre totes les característiques i les propietats de l'estimació per mínims quadrats ordinaris (MCO) d'un model de regressió lineal.

2. Entendre quina és la condició d'ortogonalitat, i per què és fonamental per al resultat de l'estimació.

3. Saber diferenciar i explicar les propietats biaix, consistència i eficiència d'una estimació economètrica.

4. Estudiar les propietats de l'estimació de la variància d'un model, això és, l'esfericitat de la matriu de variàncies i covariàncies.

5. Saber relacionar la no-esfericitat de la matriu de variàncies i covariàncies amb els problemes d'heteroscedasticitat i autocorrelació.

6. Poder identificar la presència de multicol·linealitat entre els regressors d'un model de regressió, a més de dominar les tècniques pertinents per a solucionar-lo.

7. Ser capaç de detectar la presència d'observacions atípiques o *outliers*, i poder-lo tenir en compte a l'hora d'efectuar l'estimació economètrica.

8. Dominar les eines que permeten detectar un possible trencament de la permanència estructural, i també poder efectuar estimacions més adequades partint la mostra en diferents parts.

1. Propietats de l'estimació del model

1.1. Estimador MCO i la condició d'ortogonalitat

Hi ha un aspecte molt important de l'estimador MCO que s'ha de tenir en compte. Per construcció, l'estimador MCO garanteix la condició d'ortogonalitat. Dit d'una altra manera, una vegada obtenim els residus de l'estimació del model de regressió

Errors i residus

És molt important tenir present la diferència entre els errors del model de regressió e_i i els residus resultants de l'estimació del model $\hat{e}_i$.

$$\hat{e}_i = y_i - x_i'\hat{\beta},$$

essent-ne l'expressió matricial

$$\hat{e} = Y - X\hat{\beta},$$

és impossible verificar si es compleix la condició $E(X'e) = 0$, ja que l'estimador dels paràmetres fa que es compleixi el següent:

$$X'\hat{e} = X'(Y - X\hat{\beta}) = X'Y - X'X(X'X)^{-1}X'Y = X'Y - X'Y = 0.$$

Amb la qual cosa, l'investigador haurà de determinar si es compleix la condició d'ortogonalitat considerant altres criteris, tema que s'abordarà més endavant.

1.2. Biaix i consistència

L'estimador $\hat{\beta}$ és un estadístic, i com a tal té una distribució. En general, aquesta distribució és desconeguda. Si assumim que els errors segueixen una distribució normal, podem establir que l'estimador també segueix aquella distribució.

Abans de definir el biaix i la consistència d'un estimador, és útil fer la descomposició següent de l'estimador MCO:

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + e) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'e \\ &= \beta + (X'X)^{-1}X'e\end{aligned}$$

Aquesta descomposició mostra com la distribució de $\hat{\beta}$ està determinada únicament per la distribució conjunta de (x_i, e_i).

El **biaix** de l'estimador serà l'esperança matemàtica de la diferència entre el valor esperat de l'estimador i el paràmetre del model, és a dir, $E(\hat{\beta} - \beta)$. En el moment en què es compleix $E(\hat{\beta} - \beta) = 0$, o bé $E(\hat{\beta}) = \beta$, l'estimador $\hat{\beta}$ és **centrat**. Si prenem l'expressió de l'estimador que està determinada per (1,29), veiem que si es compleix la condició $E(X'e) = 0$, és a dir, si es compleix la condició d'ortogonalitat, l'estimador serà centrat:

$$E(\hat{\beta}) = \beta + E((X'X)^{-1})E(X'e) = \beta$$

El concepte de **consistència** fa referència a la convergència en probabilitat de l'estimador amb els vertaders paràmetres del model de regressió, a mesura que la mida mostral n tendeix a infinit. Seguint aquesta definició, direm que l'estimador $\hat{\beta}$ és consistent si es compleix que $plim_{n\to\infty}(\hat{\beta}) = \beta$, és a dir, si l'estimador convergeix en probabilitat amb el vertader paràmetre del model.

Segons la teoria asimptòtica, podem entendre el concepte de convergència en probabilitat com el límit que assoleix una determinada seqüència de valors a mesura que incrementa el conjunt d'informació.

Així doncs, afirmem que l'estimador serà consistent si l'error és asimptòticament ortogonal als regressors, és a dir:

plim significa límit en probabilitat. Una notació que també es fa servir i que indica convergència en probabilitat d'una variable amb una altra és $\hat{\beta} \longrightarrow_p \beta$.

$$plim_{n\to\infty}\left(\frac{X'e}{n}\right) = 0$$

En aquest cas, es complirà el següent:

$$\begin{aligned} plim_{n\to\infty}(\hat{\beta}) &= \beta + plim_{n\to\infty}\left[\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'e}{n}\right)\right] \\ &= \beta + plim_{n\to\infty}\left(\frac{X'X}{n}\right)^{-1} plim_{n\to\infty}\left(\frac{X'e}{n}\right) \\ &= \beta \end{aligned}$$

Convé recordar que, encara que un estimador sigui esbiaixat, és a dir, $E(x_i e_i) \neq 0$, és possible que asimptòticament l'error sigui ortogonal als regressors, de manera que $plim_{n\to\infty}\left(\frac{X'e}{n}\right) = 0$, i en aquest cas l'estimador és esbiaixat però consistent. Ara bé, un estimador inconsistent **sempre** serà esbiaixat.

1.3. Eficiència

L'eficiència d'un estimador és una propietat que fa referència a la seva variància. Un estimador serà **eficient** si assoleix una variància mínima entre altres possibles estimadors dels paràmetres del model. Si reprenem el model de regressió lineal:

$$y_i = x_i'\beta + e_i$$

$$E(e_i|x_i) = 0$$

veiem que estem imposant la condició que l'esperança condicional de l'error és nul·la i que aquesta variància condicional del model és:

$$E(e_i^2|x_i) = \sigma_i^2$$

En el capítol següent analitzem en detall els casos particulars en què un estimador no serà eficient, és a dir, en presència d'*heteroscedasticitat* i/o autocorrelació.

1.4. Causes del biaix i de la inconsistència

Com es demostra en la secció anterior, l'estimador MCO garanteix l'ortogonalitat dels regressors amb els residus, de manera que $E(X'\hat{e}) = 0$, per la qual cosa és impossible saber a partir d'aquesta estimació si l'error del model està correlacionat amb els regressors. Dit d'una altra manera, l'anàlisi dels residus de la regressió no conté informació sobre el biaix i la consistència de l'estimació. En aquesta secció es detallen les dues situacions en què no es compleixen les condicions d'ortogonalitat: errors de mesura i endogenitat.

1.4.1. Errors de mesura

Suposem que disposem del model següent de regressió lineal esfèric, en el qual només tenim un regressor (la variable x_i^*):

$$y_i = \alpha + \beta x_i^* + e_i$$

$$E(e_i|x_i^*) = 0$$

$$E(e_i^2|x_i^*) = \sigma^2$$

Si disposéssim de dades per a les variables (y_i, x_i^*), i suposant que es compleixen els dos supòsits del model, l'estimació MCO seria (1) centrada, (2) consistent i (3) efi-

cient. Desafortunadament, suposarem que mesurem el regressor amb error, de manera que no observem x_i^*, sinó x_i:

$$x_i = x_i^* + v_i$$

Suposem, a més a més, que l'error de mesura v_i és una variable aleatòria, amb mesura zero i variància constant, no correlacionada ni amb l'error de la regressió ni amb l'autèntica variable que no podem observar x_i^*:

$$E(v_i) = 0,$$

$$E(v_i^2) = \sigma_v^2,$$

$$E(v_i e_i) = 0,$$

$$E(v_i|x_i^*) = 0.$$

En aquest cas, com afecta aquest error de mesura en l'estimació? Bé, introduïm l'error de mesura en el model de regressió lineal:

$$\begin{aligned} y_i &= \alpha + \beta(x_i - v_i) + e_i \\ &= \alpha + \beta x_i - \beta v_i + e_i \\ &= \alpha + \beta x_i + u_i, \\ u_i &= e_i - \beta v_i \end{aligned}$$

Introduint l'error de mesura en el model, veiem que aquest error passa a ser $u_i = e_i - \beta v_i$. Amb aquest error, comprovem que la condició d'ortogonalitat no es compleix:

$$\begin{aligned} E(x_i u_i) = Cov(x_i, u_i) &= Cov(x_i^* + v_i, e_i - \beta v_i) \\ &= -\beta\, Cov(v_i, v_i) = -\beta\sigma_v^2 \end{aligned}$$

Això implica que l'estimació per MCO és esbiaixada i inconsistent. L'estimador MCO es pot expressar de la manera següent:

$$\hat{\beta}_{MCO} = \frac{(1/n)\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(1/n)\sum_{i=1}^n (x_i - \bar{x})^2} = \beta + \frac{\sum_{i=1}^n (x_i - \bar{x})(u_i - \bar{u})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

A continuació, analitzem la consistència de l'estimador:

$$plim_{n\to\infty}\hat{\beta}_{MCO} = \beta + \frac{plim(1/n)\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u})}{plim(1/n)(x_i - \bar{x})^2}$$

$$= \beta + \frac{Cov(x_i, u_i)}{Var(x_i)} = \beta + \frac{-\beta\sigma_v^2}{\sigma_{x^*}^2 + \sigma_v^2}$$

$$= \beta\left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_v^2}\right)$$

Com podem observar, en aquest cas l'error de mesura provoca un biaix en l'estimació del paràmetre cap a zero, és a dir, l'estimació del paràmetre mostra un valor inferior al del vertader valor. Aquest biaix creix a mesura que la variància de l'error σ_v^2 augmenta. A més a més, en cas que tinguéssim un model amb diversos paràmetres, les estimacions de tots ells es veurien afectades, encara que l'error de mesura es donés en una sola variable. Cal afegir que si hi ha més d'un regressor mesurant amb error, no se sabrà quina és la direcció del biaix.

1.4.2. Endogenitat

Un dels supòsits en què ens basem a l'hora de plantejar un model de regressió fa referència als regressors. Aquests han de ser exògens o predeterminats, és a dir, no hi ha d'haver cap element en el model que els determini. Un exemple d'endogenitat es dóna en els models d'equacions simultànies, en què els regressors d'una equació són generats en altres equacions amb una component estocàstica. Un altre exemple el trobem en els models que consideren dades temporals quan un dels regressors és la variable endògena retardada, això és:

Tècnicament, els conceptes d'exogenitat i predeterminació no són exactament equivalents, encara que nosaltres usem els dos termes indistintament.

$$y_t = \beta x_t + \gamma y_{t-1} + e_t$$

Aquest tipus de models **sempre** serà esbiaixat, és a dir, tindrem biaix per endogenitat. Ara bé, depenent de quina sigui l'estructura de l'error, les propietats asimptòtiques de l'error seran unes o unes altres. Suposem que el model és esfèric, de manera que l'error es caracteritza pel següent:

$$e_t \sim iid(0, \sigma^2 I_n)$$

En aquest cas, si analitzem la covariància entre regressor i error, obtenim:

$$Cov(y_{t-1}, e_t) = Cov(\beta x_{t-1} + \gamma y_{t-2} + e_{t-1}, e_t) = Cov(e_{t-1}, e_t) = 0$$

D'aquesta manera, obtenim consistència en l'estimador:

$$plim_{n\to\infty}(\hat{\beta}) = \beta + plim_{n\to\infty}\left(\frac{X'X}{n}\right)^{-1} plim_{n\to\infty}\left(\frac{X'e}{n}\right) = \beta$$

Ara bé, suposem que el terme d'error està correlacionat i segueix una estructura autoregressiva, amb la qual cosa l'error ja no és esfèric:

$$e_t = \rho\, e_{t-1} + u_t,$$

$$u_t \sim iid(0, \sigma_u^2 I_n)$$

Fixem-nos que en aquest cas el model incorpora el regressor estocàstic $\rho\, e_{t-1}$. A l'hora d'analitzar la covariància entre regressor i error, obtenim:

$$Cov(y_{t-1}, e_t) = Cov(\beta x_{t-1} + \gamma y_{t-2} + e_{t-1}, \rho\, e_t + u_t) = \rho\, Cov(e_{t-1}, e_{t-1}) = \rho\, \sigma^2$$

En aquest cas, l'estimació ja no és consistent, ja que

$$plim_{n\to\infty}(\hat{\beta}) \neq \beta$$

2. Heteroscedasticitat i autocorrelació

2.1. Definició teòrica

En primer lloc, definim la matriu de variàncies i covariàncies de l'error del model de regressió:

$$MVC(e) = E(ee') = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}_{n \times n}$$

Els elements de la diagonal són les variàncies dels errors, i fora de la diagonal estan situades les covariàncies. Aquí podem trobar diverses situacions:

- **Elements de la diagonal**. El model de regressió lineal és **homoscedàstic** si els elements de la diagonal són tots idèntics, això és, si es compleix que $\sigma_i^2 = \sigma^2$. En aquest cas, l'esperança del quadrat de l'error no varia a través dels elements mostrals. En canvi, serem davant un model de regressió lineal **heteroscedàstic** si es compleix que $\sigma_i^2 = \sigma^2(x_i)$, és a dir, si σ_i^2 varia per a cada element i.

- **Elements de fora de la diagonal**. Si aquests no són nuls, això és, $\sigma_{ij} \neq 0,\ \forall i \neq j$, el model de regressió està **autocorrelacionat**, i anàlogament si són nuls, el model no estarà autocorrelacionat.

Partint d'aquestes definicions, diem que som davant un **model de regressió lineal esfèric** (també se sol denominar un model de regressió lineal amb una matriu de variàncies i covariàncies esfèrica) si la matriu de variàncies i covariàncies és homoscedàstica i no correlacionada, de manera que podem expressar la matriu de variàncies i covariàncies de la manera següent:

$$MVC(e) = E(ee') = \sigma^2 I_n$$

Essent I_n la matriu identitat de dimensió $n \times n$. En aquest cas, l'estimació del model per MCO és **eficient**.

El fet de ser davant un model de regressió lineal homoscedàstic o heteroscedàstic té implicacions a l'hora de valorar tant els paràmetres del model com la matriu de variàncies i covariàncies. Això és, en presència d'heteroscedasticitat i/o autocorrelació,

tindrem un **model de regressió lineal no esfèric**. En aquest cas, l'estimació del model per MCO no serà eficient, ja que no estarem incorporant l'estructura de l'error en l'estimació dels paràmetres. En aquest cas, el *teorema de Gauss-Markov* estableix que el millor estimador lineal centrat i de mínima variància és el de **mínims quadrats generalitzats (MCG)**. Així, suposant que la matriu de variàncies i covariàncies adquireix la forma $MVC(e) = E(ee') = \Omega$, aquest estimador es defineix de la manera següent:

$$\hat{\beta}_{MCG} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

Amb freqüència no es coneix Ω, per la qual cosa s'ha de valorar (o bé directament o bé imposant una estructura). Una vegada obtenim l'estimació $\hat{\Omega}$, podem calcular l'estimador per mínims quadrats generalitzats factibles (MCGF):

$$\hat{\beta}_{MCGF} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}Y$$

2.2. Exemple pràctic

En aquesta secció farem un exercici pràctic d'anàlisi d'heteroscedasticitat i autocorrelació amb R-Commander. Per a això, analitzarem el següent model temporal de consum amb dades simulades:

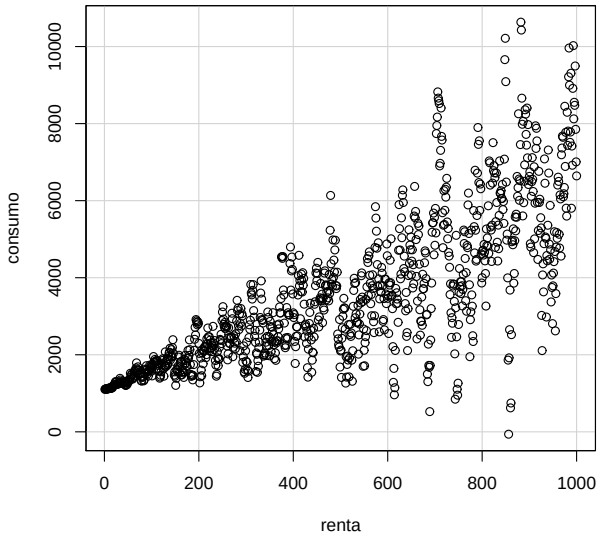
$$C_t = \beta_0 + \beta_1 R_t + e_t$$

On C_t correspon al consum i R_t és el nivell de renda. Les dades són temporals i corresponen a una economia, de manera que $t = 1, \ldots, T$.

Una vegada importades les dades, un bon inici és una representació gràfica de les dades, cosa que és immediata si només hi ha un regressor. Mitjançant la ruta següent, obtenim un diagrama de dispersió de les variables explicativa i explicada:

Gràfiques / Diagrama de dispersió

Cosa que resulta en el gràfic següent:

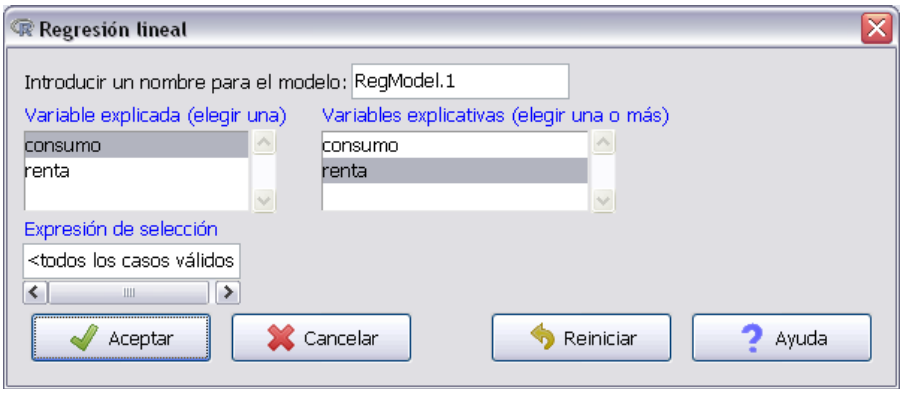


La interpretació d'aquest gràfic és molt intuïtiva. Per a nivells baixos de renda, els nivells de consum varien poc en l'eix d'ordenades (y). No obstant això, a mesura que augmenten els nivells de renda, s'observa una variabilitat superior de la variable explicativa. Això és un signe d'heteroscedasticitat, l'existència de la qual ha de ser validadada estadísticament mitjançant els contrastos corresponents.

Per estimar el model amb R-Commander, anirem a la ruta següent:

Estadístics / Ajust de models / Regressió lineal

Apareixerà el quadre de diàleg següent, en el qual introduïm la variable explicativa i l'explicada:



El resultat de l'estimació MCO del model és el següent:

```
> summary(RegModel.1)

Call:
lm(formula = consumo ~ renta, data = Datos)

Residuals:
    Min      1Q  Median      3Q     Max
-5475.7  -560.9    96.0   513.0  5082.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  920.372     76.855   11.97   <2e-16 ***
renta          5.250      0.133   39.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1214 on 998 degrees of freedom
Multiple R-squared:  0.6095, Adjusted R-squared:  0.6091
F-statistic:  1558 on 1 and 998 DF,  p-value: < 2.2e-16
```

Un test adequat per a detectar la possible heteroscedasticitat és el de Breusch-Pagan. Aquest test, vàlid quan es disposa de mostres prou grans, pressuposa que és possible expressar la variància del terme de pertorbació com una combinació lineal d'un nombre determinat (p) de variables explicatives. El contrast es planteja de la manera següent:

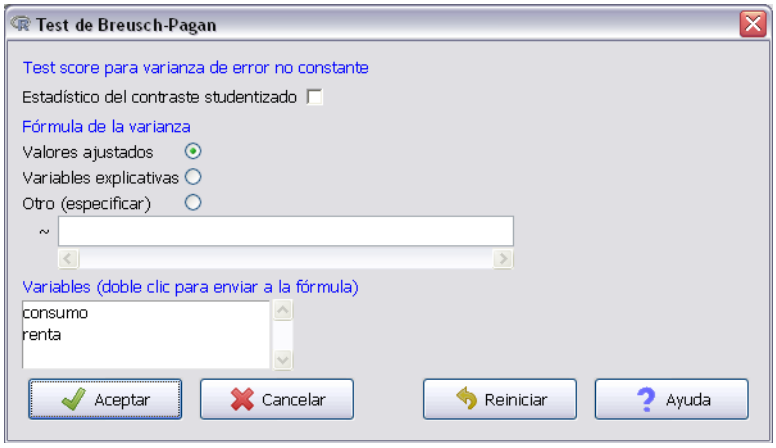
$$H_0 : \sigma_t^2 = \sigma^2$$

$$H_1 : \sigma_t^2 \neq \sigma^2$$

És a dir, segons la hipòtesi alternativa, la variància no és constant, sinó que depèn d'alguna variable. Amb R-Commander, aquest test es fa accedint a la ruta següent:

Diagnòstics numèrics / Test de Breusch-Pagan per a heteroscedasticitat

Apareixerà el quadre de diàleg següent, en què haurem d'introduir els valors del contrast. És a dir, tenim la possibilitat d'introduir la forma funcional de la variància, en cas de conèixer-la. En el nostre cas, acceptarem l'opció per defecte, que adquireix els valors ajustats de la regressió com a fórmula per a la variància:



El resultat del test ens indica que caiem en la regió de rebuig de la hipòtesi nul·la, de manera que determinem que hi ha heteroscedasticitat en el nostre model.

```
> bptest(consumo ~ renta, varformula = ~ fitted.values(RegModel
    .1), studentize=FALSE, data=Datos)

Breusch-Pagan test

data:  consumo ~ renta
BP = 351.9272, df = 1, p-value < 2.2e-16
```

El segon problema que s'ha d'analitzar és la possible existència d'autocorrelació en el model. Per a això farem el contrast de Durbin-Watson. Aquest test permet contrastar

si el terme de pertorbació està autocorrelacionat segons un esquema AR(1), és a dir, la hipòtesi nul·la indica que si el terme de pertorbació és de la forma $e_t = \rho e_{t-1} + \varepsilon_t$. Específicament, el contrast es defineix de la manera següent:

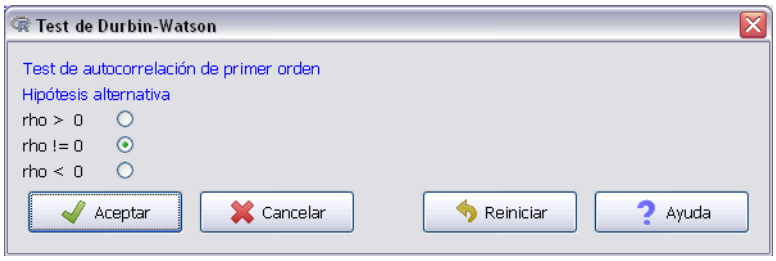
$$H_0 : e_t \sim AR(1) \qquad \text{amb} \qquad \rho = 0$$

$$H_1 : e_t \sim AR(1) \qquad \text{amb} \qquad \rho \gtrless 0$$

Amb R-Commander, aquest test es fa accedint a la ruta següent:

Diagnòstics numèrics / Test de Durbin-Watson per a autocorrelació

Apareixerà el quadre de diàleg següent, on hem d'indicar la hipòtesi alternativa. Si tenim informació prèvia que el vertader valor del paràmetre ρ és positiu, seleccionarem $H_1 : \rho > 0$, i el que correspongui per a un valor negatiu de ρ. Si no tenim informació prèvia sobre aquest paràmetre, seleccionarem $H_1 : \rho \neq 0$:



El resultat del test ens indica clarament, per a qualsevol nivell de confiança, que rebutgem la hipòtesi nul·la, és a dir, hi ha autocorrelació en el model.

```
> dwtest(consumo ~ renta, alternative="two.sided", data=Datos)

Durbin-Watson test

data:  consumo ~ renta
DW = 0.4037, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is not 0
```

2.3. Estimació eficient de l'MVC

En aquesta secció ens encarreguem de com es fa una estimació eficient en presència d'autocorrelació i/o heteroscedasticitat. White (1980) va argumentar que no sempre és possible conèixer l'estructura dels errors i valorar el model mitjançant MCG. Quan això succeeix, en el cas de ser davant un model heteroscedàstic, la millor opció és

valorar els paràmetres del model mitjançant MCO i intentar obtenir una estimació robusta de la matriu de variàncies i covariàncies dels paràmetres mitjançant la fórmula:

$$MVC(\hat{\beta}_{MCO}) = n(X'X)^{-1}n^{-1}\sum_{i=1}^{n}\hat{e}_i^2 x_i x_i'(X'X)^{-1}$$

Aquest procediment es coneix amb diversos noms en la literatura: fórmula de White, fórmula d'Eicker-White, fórmula de Huber, fórmula de Huber-White o matriu de covariàncies GMM, entre d'altres. D'altra banda, és problemàtic en mostres petites.

En aquest capítol veurem com s'estima un model davant **heteroscedasticitat** i/o **autocorrelació**. Com veurem, hi ha dues grans aproximacions sobre això:

1) Estimar el model mitjançant mínims quadrats generalitzats (MCG).

2) Estimar el model mitjançant MCO i a continuació valorar eficientment la matriu de variàncies i covariàncies.

Per il·lustrar-ho amb un exemple, amb R crearem unes dades fictícies que generin un model heteroscedàstic i autocorrelacionat. Abans de res, carregarem tres biblioteques que ens hi ajudaran:

```
> library(sandwich)
> library(lmtest)
> library(nlme)
```

Suposem el model de regressió lineal següent:

$$y_t = \alpha + \beta x_t + u_t, \qquad t = 1, \ldots, T.$$

Simularem les dades, de manera que els paràmetres poblacionals són $\alpha = 100$ i $\beta = 5$. A més a més, fixem la grandària mostral com a $T = 1000$. El model es construeix de manera que el terme d'error no és esfèric, ja que estarà autocorrelacionat i serà heteroscedàstic:

$$u_t = \rho u_{t-1} + \varepsilon_t$$

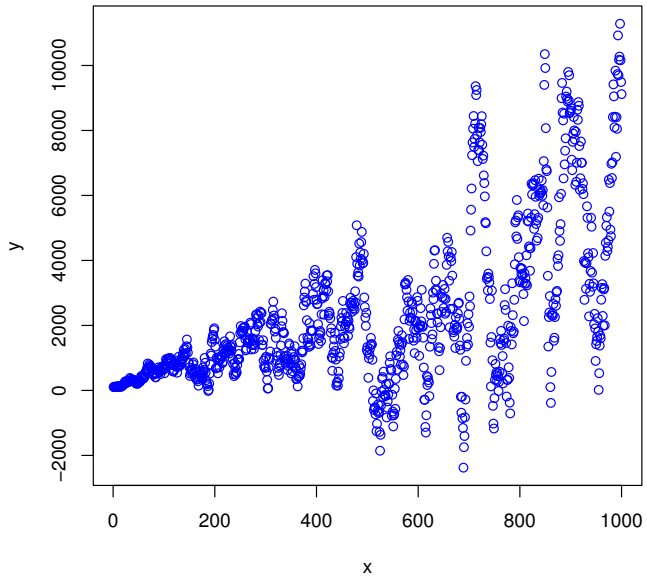
$$\varepsilon_t \sim N(0, \gamma t)$$

Fixem els valors $\rho = 0,95$ i $\gamma = 1,1$. Amb el model definit, l'introduïm en R i el representem gràficament:

```
> T <- 1000
> alpha <- 100
> beta <- 5
> rho <- 0.95
> gamma <- 1.1
> x <- 1:T
> y0 <- alpha + beta * x
> err <- rep(0, T)
> set.seed(12)
> err[1] <- rnorm(1, 0, 1)
> set.seed(12)
> for (i in 2:T) {
+     err[i] <- err[i - 1] * rho + rnorm(1, 0, i * gamma)
+ }
> y <- y0 + err
```

Visualitzarem les variables creades per veure com es relacionen entre elles. Aquest gràfic ja ens ha de donar la impressió que la variància no es comporta aleatòriament.

```
> plot(x, y, col = "blue")
```



Com veiem, el model per construcció no té un terme de pertorbació esfèric. Quin és el problema d'aplicar l'estimador de mínims quadrats ordinaris (MCO)? Bé, perquè l'estimador MCO sigui eficient (mínima variància de l'estimació), la matriu de variàncies i covariàncies de u ha de ser esfèrica, és a dir:

1) Homoscedàstica: la variància de u no varia entre els elements de la mostra, de manera que $\sigma_t^2 = \sigma^2$ i els elements de la diagonal de $MVC(u)$ són idèntics.

2) No autocorrelacionada: si els elements de fora de la diagonal no són nuls ($\sigma_{ij} \neq 0,\ \forall i \neq j$), el model de regressió està autocorrelacionat, i viceversa.

Si *a)* i *b)* es compleixen, la matriu $MVC(u)$ serà:

$$MVC(u) = E(uu') = \sigma^2 I_T$$

Essent I_T la matriu identitat de dimensió $T \times T$.

En el nostre cas, veiem que això no es compleix. Valorarem primer l'estimador MCO i veurem com es comporta:

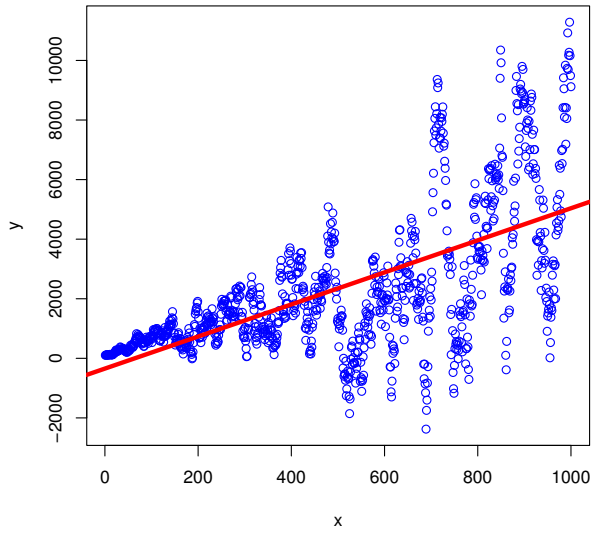
```
> m_mco <- lm(y ~ x)
> summary(m_mco)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-5740.7  -964.1   157.1   724.6  6262.2

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -345.0854   119.8442  -2.879  0.00407 **
x              5.3828     0.2074  25.951  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1893 on 998 degrees of freedom
Multiple R-squared: 0.4029, Adjusted R-squared: 0.4023
F-statistic: 673.5 on 1 and 998 DF,  p-value: < 2.2e-16
```

Vegem un interval de confiança al 95% per als paràmetres estimats:

```
> confint(m_mco)

                  2.5 %      97.5 %
(Intercept) -580.260897 -109.909888
x              4.975785    5.789846
```

Representarem visualment la recta estimada ($\hat{\alpha}$ i $\hat{\beta}$) sobre el diagrama de dispersió dels punts:

```
> plot(x, y, col = "blue")
> abline(lsfit(x, y), lty = 1, lwd = 4, col = "red")
```



Aquest estimador es construeix mitjançant la fórmula següent:

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$$

I calcula la variància i covariàncies de $\hat{\beta}$ així:

$$MVC(\hat{\beta}) = \hat{\sigma}_u^2(X'X)^{-1}$$

No obstant això, hem vist que l'MVC del terme de pertorbació és realment:

$$MVC(u) = E(uu') = \Omega = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1T} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{T1} & \sigma_{T2} & \cdots & \sigma_T^2 \end{pmatrix}_{T\times T}$$

Amb la qual cosa, en realitat, la variància dels paràmetres és:

$$MVC(\beta) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

La qüestió és, com estimem el model? Hi ha dues opcions. Teòricament, si coneixem exactament la forma de Ω, la podem introduir directament en l'estimador per mínims quadrats generalitzats (MCG):

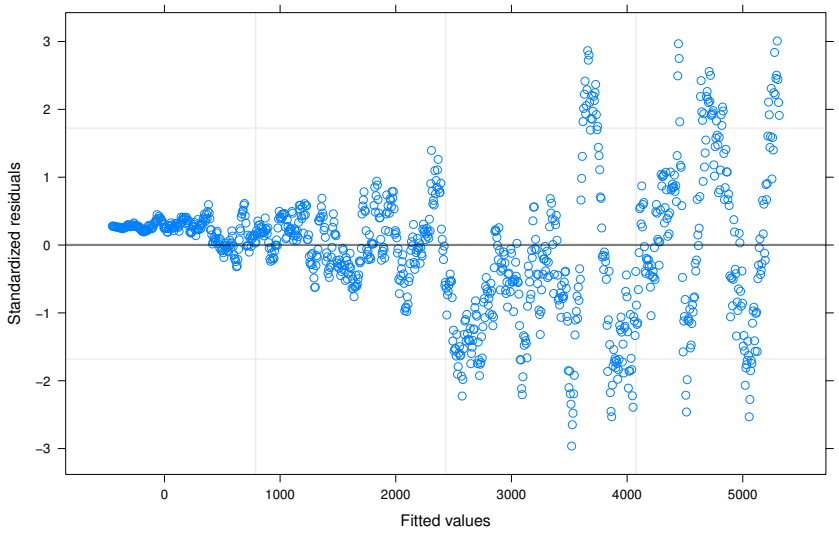
$$\hat{\beta}_{MCG} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

En R, estimarem MCG sabent que $\rho = 0,95$ i $\gamma = 1,1$. Primer assumint solament autocorrelació:

```
> gls_1 <- gls(y ~ x, correlation = corAR1(rho))
> summary(gls_1)

Generalized least squares fit by REML
  Model: y ~ x
  Data: NULL
       AIC      BIC    logLik
  15686.43 15706.05 -7839.214

Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
     Phi
0.950404

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) -455.9404  753.9217 -0.604758  0.5455
x              5.7711    1.2923  4.465694  0.0000

 Correlation:
  (Intr)
x -0.858

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.96297027 -0.53634973  0.07080453  0.37691878  3.00744997

Residual standard error: 1990.35
Degrees of freedom: 1000 total; 998 residual
```

La funció `plot` aplicada al model estimat per MCG ens mostra el gràfic dels residus:

```
> plot(gls_1)
```

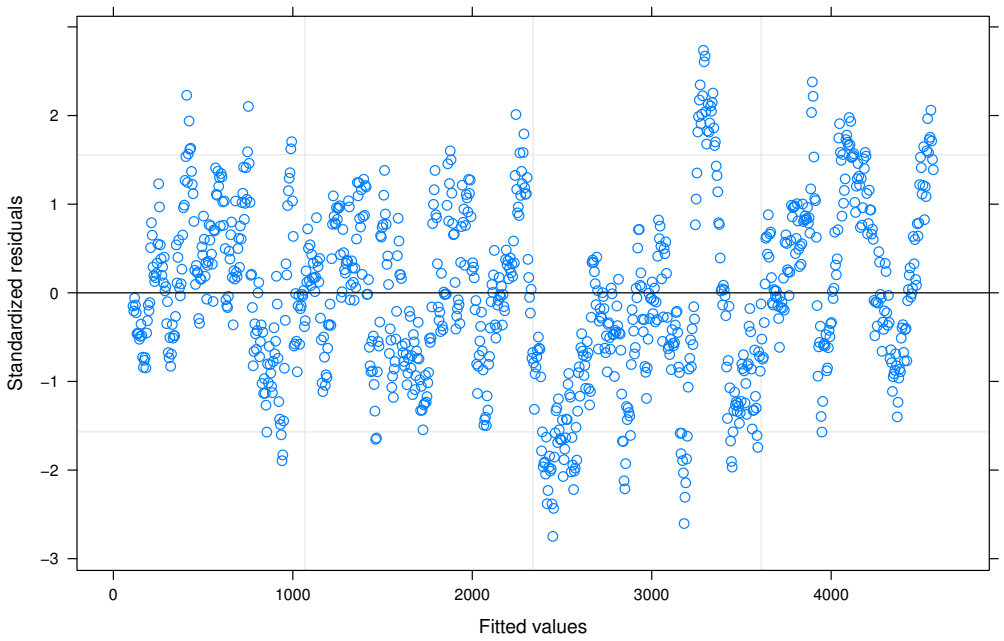


I ara estimem novament el model mitjançant MCG, assumint aquesta vegada tant autocorrelació com heteroscedasticitat:

```
> gls_2 <- gls(y ~ x, correlation = corAR1(rho), weights =
    varPower(gamma))
> summary(gls_2)

Generalized least squares fit by REML
  Model: y ~ x
  Data: NULL
       AIC     BIC    logLik
  14797.58 14822.1 -7393.788

Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.9353005
Variance function:
 Structure: Power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
   power
1.183458

Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 104.94599  37.42416 2.804231  0.0051
x             4.46403   0.53408 8.358306  0.0000

 Correlation:
  (Intr)
x -0.293

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.74838335 -0.69033419 -0.05426096  0.70224709  2.73602972

Residual standard error: 0.1528423
Degrees of freedom: 1000 total; 998 residual
```

Novament, la funció `plot` aplicada al model estimat per MCG ens mostra el gràfic dels residus:

```
> plot(gls_2)
```



És rellevant recordar que White (1980) va argumentar que no sempre és possible conèixer l'estructura dels errors i valorar el model mitjançant MCG. Quan això succeeix, en el cas de ser davant un model heteroscedàstic, la millor opció és estimar els paràmetres del model mitjançant MCO i intentar obtenir una estimació robusta de la matriu de variàncies i covariàncies dels paràmetres mitjançant la fórmula:

$$MVC(\hat{\beta}_{MCO}) = n(X'X)^{-1}n^{-1}\sum_{i=1}^{n}\hat{u}_i^2 x_i x_i'(X'X)^{-1}$$

En aquest sentit, hi ha moltes maneres de calcular eficientment $\hat{\Omega}$. El programa R ens n'ofereix dues:

1) *HC* : *Heteroskedasticity Consistent matrix.*

2) *HAC* : *Heteroskedasticity and Autocorrelation Consistent matrix.*

Llavors, a partir d'MCO, calculem $\hat{\Omega}$ de les dues maneres i així recalculem les variàncies (i els contrastos de significació associats) dels coeficients:

```
> coeftest(m_mco)

t test of coefficients:

             Estimate Std. Error t value  Pr(>t)
(Intercept) -345.08539  119.84419 -2.8795  0.004069 **
x              5.38282    0.20742 25.9512 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> coeftest(m_mco, vcovHC(m_mco))



t test of coefficients:

             Estimate Std. Error t value  Pr(>t)
(Intercept) -345.08539   72.74869 -4.7435 2.406e-06 ***
x              5.38282    0.22307 24.1310 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> coeftest(m_mco, vcovHAC(m_mco))

t test of coefficients:

             Estimate Std. Error t value  Pr(>t)
(Intercept) -345.08539  302.01809 -1.1426    0.2535
x              5.38282    0.75061  7.1712 1.446e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Com veiem, el fet d'estimar la vertadera matriu MVC revela que les variàncies reals són en realitat més grans que les estimades per MCO i, per tant, els intervals de confiança per a $\hat{\beta}$ són també més grans.

3. Errors en la mostra

3.1. Multicol·linealitat

La multicol·linealitat apareix quan dues o més variables explicatives en un model de regressió múltiple estan altament correlacionades. De manera alternativa, es pot afirmar que, en presència de multicol·linealitat, una variable explicativa es pot predir linealment a partir d'altres variables explicatives.

La multicol·linealitat implica que les estimacions dels coeficients de la regressió múltiple poden canviar de manera erràtica davant petits canvis en l'especificació del model o canvis en les dades. A més a més, un alt grau de multicol·linealitat pot causar problemes a l'hora de calcular la matriu inversa de $X'X$, necessària per a calcular els coeficients de regressió.

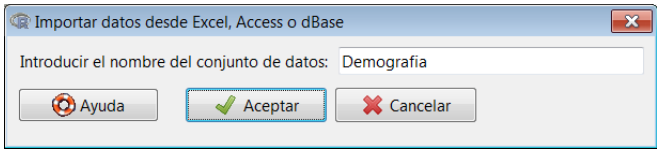
Recordem que hi ha tres graus de multicol·linealitat:

1) Absència total de multicol·linealitat. Passa quan no hi ha correlació entre les variables explicatives del model.

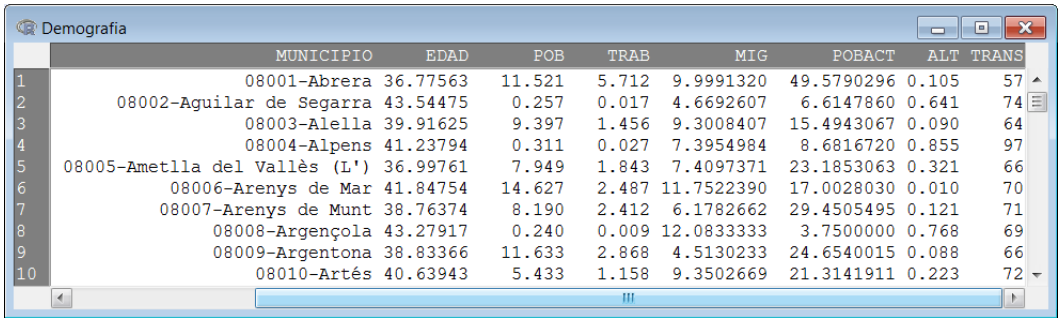
2) Presència d'un cert grau de multicol·linealitat. Hi ha un alt grau de correlació lineal entre algunes variables explicatives. Com més elevat sigui aquest grau de correlació (és a dir, el coeficient de correlació de Pearson s'acosti a 1), més gran serà el grau de multicol·linealitat.

3) Presència de multicol·linealitat perfecta. Hi ha alguna variable explicativa que es pot obtenir a partir de la combinació lineal d'altres variables explicatives, cosa que implica que algunes variables explicatives són linealment dependents entre si. En aquest cas, l'estimació del model és impossible degut a la impossibilitat d'invertir la matriu $X'X$.

Vegem un exemple pràctic, amb R-Commander, de com s'analitza el problema de la multicol·linealitat. Per a això considerarem un estudi demogràfic per als municipis de Catalunya l'any 2009. El primer pas serà importar les dades d'un arxiu d'Excel i crear un conjunt de dades que anomenarem de *Demografia*:



Si visualitzem les dades importades, observem que estan incloses les variables següents:



	MUNICIPIO	EDAD	POB	TRAB	MIG	POBACT	ALT	TRANS
1	08001-Abrera	36.77563	11.521	5.712	9.9991320	49.5790296	0.105	57
2	08002-Aguilar de Segarra	43.54475	0.257	0.017	4.6692607	6.6147860	0.641	74
3	08003-Alella	39.91625	9.397	1.456	9.3008407	15.4943067	0.090	64
4	08004-Alpens	41.23794	0.311	0.027	7.3954984	8.6816720	0.855	97
5	08005-Ametlla del Vallès (L')	36.99761	7.949	1.843	7.4097371	23.1853063	0.321	66
6	08006-Arenys de Mar	41.84754	14.627	2.487	11.7522390	17.0028030	0.010	70
7	08007-Arenys de Munt	38.76374	8.190	2.412	6.1782662	29.4505495	0.121	71
8	08008-Argençola	43.27917	0.240	0.009	12.0833333	3.7500000	0.768	69
9	08009-Argentona	38.83366	11.633	2.868	4.5130233	24.6540015	0.088	66
10	08010-Artés	40.63943	5.433	1.158	9.3502669	21.3141911	0.223	72

La descripció de les variables és la següent:

MUNICIPIO: codi postal i nom del municipi.

EDAD: mitjana d'edat de la població.

POB: població total (en milers de persones).

TRAB: nombre de treballadors (en milers de persones).

MIG: percentatge de població immigrant.

POBACT: percentatge de població activa.

ALT: altitud del municipi (en quilòmetres).

TRANS: temps de transport fins a la capital més propera.

El primer model de regressió considera la variable *EDAD* com a variable explicada, i la resta de les variables com a variables explicatives. Per valorar un model de regressió lineal, com sabem, tenim la ruta següent en el menú desplegable:

Estadístics / Ajust de models / Regressió lineal

Seleccionem el nom del model estimat i les variables que s'han d'incloure en l'estimació en el quadre de diàleg següent:

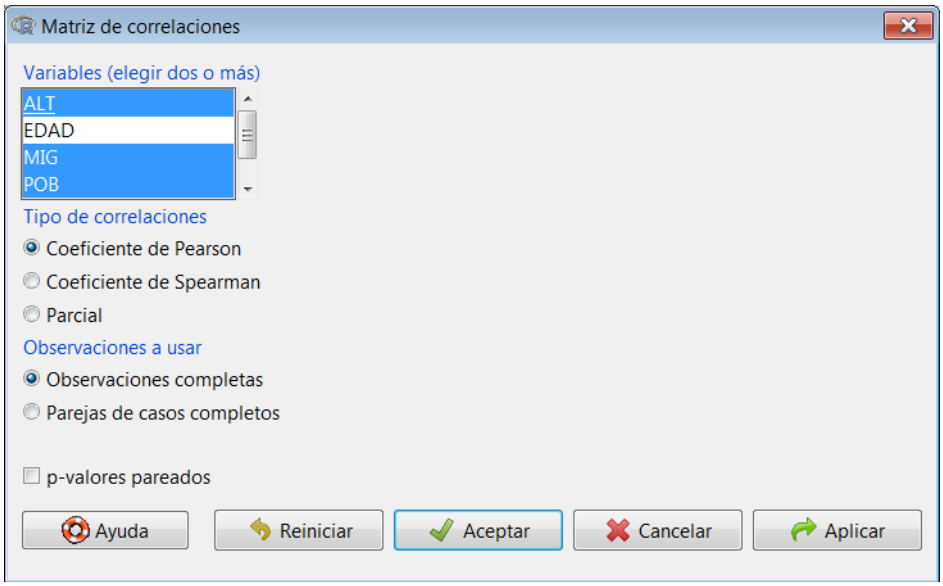
El resultat de l'estimació es mostra a continuació. A simple vista, encara que l'ajust del model sigui més aviat pobre ($R^2 = 0,3$), tots els coeficients estimats són significatius amb un nivell de significació més petit que 1%, i l'estimació és significativa en conjunt, ja que el resultat del test *F*.

```
> RegModel.1 <- lm(EDAD~ALT+MIG+POB+POBACT+TRAB+TRANS, data=
    Demografia)

> summary(RegModel.1)

Call:
lm(formula = EDAD ~ ALT + MIG + POB + POBACT + TRAB + TRANS,
    data = Demografia)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4061  -2.4548  -0.3131   2.3616  16.2820

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) 39.634134   0.513051  77.252  < 2e-16 ***
ALT          1.407579   0.450937   3.121  0.00185 **
MIG         -0.154346   0.018279  -8.444  < 2e-16 ***
POB         -0.037418   0.012169  -3.075  0.00217 **
POBACT      -0.040815   0.005446  -7.495 1.54e-13 ***
TRAB         0.074415   0.023209   3.206  0.00139 **
TRANS        0.059654   0.006064   9.838  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.62 on 934 degrees of freedom
Multiple R-squared:  0.3021,  Adjusted R-squared:  0.2976
F-statistic: 67.37 on 6 and 934 DF,  p-value: < 2.2e-16
```

Significa això que el resultat de l'estimació és satisfactori, i que podem donar aquest resultat com a vàlid? La veritat és que no necessàriament. Abans de fer una estimació, és útil visualitzar la matriu de correlacions simple entre totes les variables. Encara que hi ha tècniques més avançades i eficients per a detectar la multicol·linealitat, aquesta matriu sempre mostrarà informació útil:

Resums / Matriu de correlacions

En el quadre d'opcions resultant seleccionem totes les variables explicatives, així com el coeficient de correlació de Pearson.



Aquesta ruta ens mostra la informació següent:

```
> cor(Demografia[,c("ALT","MIG","POB","POBACT","TRAB","TRANS")
    ], use="complete")

          ALT    MIG    POB POBACT   TRAB  TRANS
ALT      1,00  -0,32  -0,11  -0,11  -0,07   0,43
MIG     -0,32   1,00   0,10   0,00   0,06   0,12
POB     -0,11   0,10   1,00   0,07   0,98  -0,12
POBACT  -0,11   0,00   0,07   1,00   0,08  -0,11
TRAB    -0,07   0,06   0,98   0,08   1,00  -0,08
TRANS    0,43   0,12  -0,12  -0,11  -0,08   1,00
```

Per a facilitar la interpretació del resultat, s'ha limitat a dos decimals cada valor d'aquesta matriu. En realitat, el resultat mostra més decimals.

Què podem destacar d'aquesta matriu de correlacions? La correlació lineal entre les variables *POB* (població) i *TRAB* (treballadors) és de *0,98*, és a dir, és una correlació lineal positiva gairebé perfecta. Realment, és necessari incorporar al model que s'ha d'estimar dues variables que aporten gairebé la mateixa informació? Això no solament té conseqüències negatives quant al procés d'estimació, sinó que també pot comportar estimacions errònies dels coeficients.

Un procediment més refinat per a avaluar la possible existència de multicol·linealitat entre les variables explicatives (o regressors) és el **factor d'increment de la variància (FIV)** de cada una de les variables explicatives. El FIV és un estadístic que permet determinar si la variància d'un estimador està inflada per la presència de multicol·linealitat en el model respecte al cas d'ortogonalitat entre regressors. Això és, si la correlació entre tots els regressors fos igual a zero (ortogonalitat perfecta), la variància de l'estimació seria òptima i el FIV de cada regressor seria igual a zero. En la

En el càlcul del FIV no afecta quina sigui la variable explicada, ja que només hi intervenen les variables explicatives o regressores.

pràctica, cada regressor tindrà un FIV més elevat com més gran sigui la seva correlació amb la resta dels regressors. En la pràctica, no hi ha un valor llindar dels FIV a partir del qual s'hagi d'afirmar que hi ha problemes greus de multicol·linealitat, però habitualment es considera que, per a cada regressor, un $FIV > 5$ indica un grau de multicol·linealitat elevat que s'ha de corregir.

A partir del model estimat anteriorment, amb R-Commander calcularem el FIV accedint a la ruta següent:

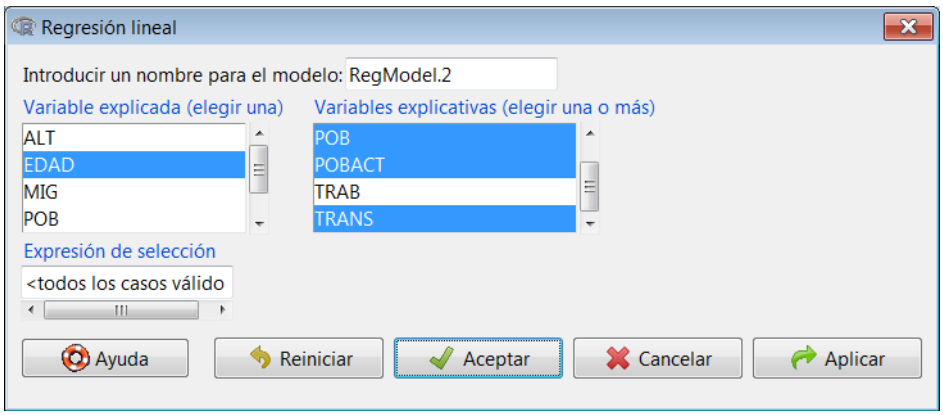
Models / Diagnòstics numèrics / Factors d'inflació de variància

El resultat mostra clarament com totes les variables tenen un FIV baix menys dos: *POB* i *TRAB*. Per a aquestes dues variables el valor del FIV és altíssim, amb la qual cosa una de les dues ha de ser eliminada de l'especificació del model.

```
> vif(RegModel.1)
      ALT        MIG        POB     POBACT       TRAB      TRANS
 1.514863   1.306153  33.629262   1.029765  33.160244   1.432465
```

Ara optarem per retirar la variable *TRAB* de l'especificació, i estimar un segon model de manera anàloga al cas anterior:

Estadístics / Ajust de models / Regressió lineal



El resultat del segon model estimat ens mostra una contradicció respecte a la primera estimació. El coeficient associat a la variable *POB* ara no és significatiu, mentre que en el model estimat anteriorment sí que ho era. Què ens indica això? Doncs que **no s'ha de confiar en les estimacions de paràmetres en presència de multicol·linealitat**.

```
> RegModel.2 <- lm(EDAD~ALT+MIG+POB+POBACT+TRANS, data=
    Demografia)

> summary(RegModel.2)

Call:
lm(formula = EDAD ~ ALT + MIG + POB + POBACT + TRANS, data =
    Demografia)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6718  -2.4594  -0.3481   2.4163  16.5107

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.2802715  0.5035203  78.011  < 2e-16 ***
ALT          1.4575008  0.4528992   3.218  0.00133 **
MIG         -0.1654489  0.0180364  -9.173  < 2e-16 ***
POB          0.0009936  0.0021441   0.463  0.64318
POBACT      -0.0392892  0.0054517  -7.207 1.18e-12 ***
TRANS        0.0631734  0.0059932  10.541  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.638 on 935 degrees of freedom
Multiple R-squared:  0.2944,  Adjusted R-squared:  0.2906
F-statistic: 78.01 on 5 and 935 DF,  p-value: < 2.2e-16
```

Per assegurar-nos que el problema de multicol·linealitat està resolt, obtindrem els VIF dels coeficients d'aquesta segona estimació.

Models / Diagnòstics numèrics / Factors d'inflació de variància

```
> vif(RegModel.2)
     ALT      MIG      POB   POBACT    TRANS
1.513057 1.259275 1.033843 1.021903 1.385527
```

Clarament, tots els valors són més petits que 5, amb la qual cosa hem resolt el problema de multicol·linealitat.

3.2. Observacions atípiques

Aquest problema sorgeix quan en la mostra algunes observacions manifesten un valor molt diferent de la resta. Visualment, això es correspon amb un núvol de punts de la variable en què un punt és molt allunyat de la resta de les observacions. Dues explicacions poden respondre a aquest fet:

1) Hi ha errors en la recollida de la mostra, de manera que hi ha valors erronis que no es corresponguin amb la realitat.

2) El valor recollit en la mostra d'aquestes observacions *outliers* es deu a particularitats de l'observació, de manera que no hi ha cap error en la mostra.

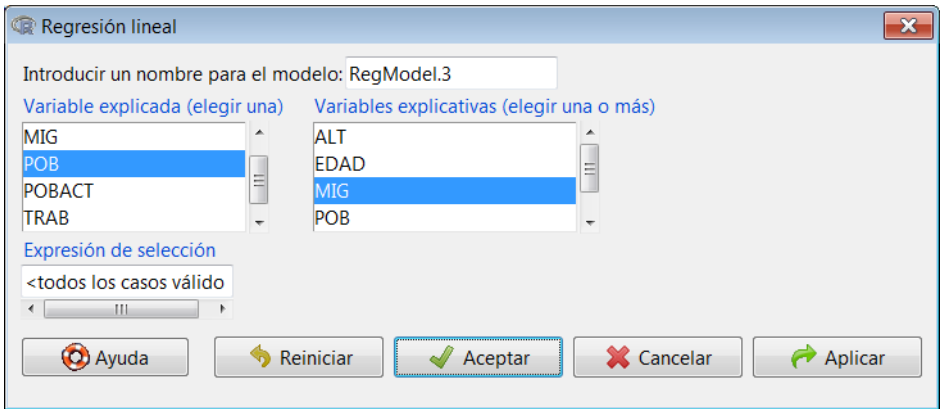
En tots dos casos, la presència d'*outliers* té conseqüències negatives per a l'estimació del model economètric, ja que els errors estàndard dels estimadors són més grossos i empitjora l'ajust global del model (R^2 i *F de Snedecor*).

Estudiarem aquest fet a partir del conjunt de dades *Demografia*, introduït en l'apartat anterior. En aquest cas, estimarem un MRLS en què el percentatge d'immigració explica la població total de cada municipi:

$$POB_i = \beta_0 + \beta_1 MIG_i + e_i$$

Igual que en el cas anterior, valorem el model accedint a la ruta següent:

Estadístics / Ajust de models / Regressió lineal

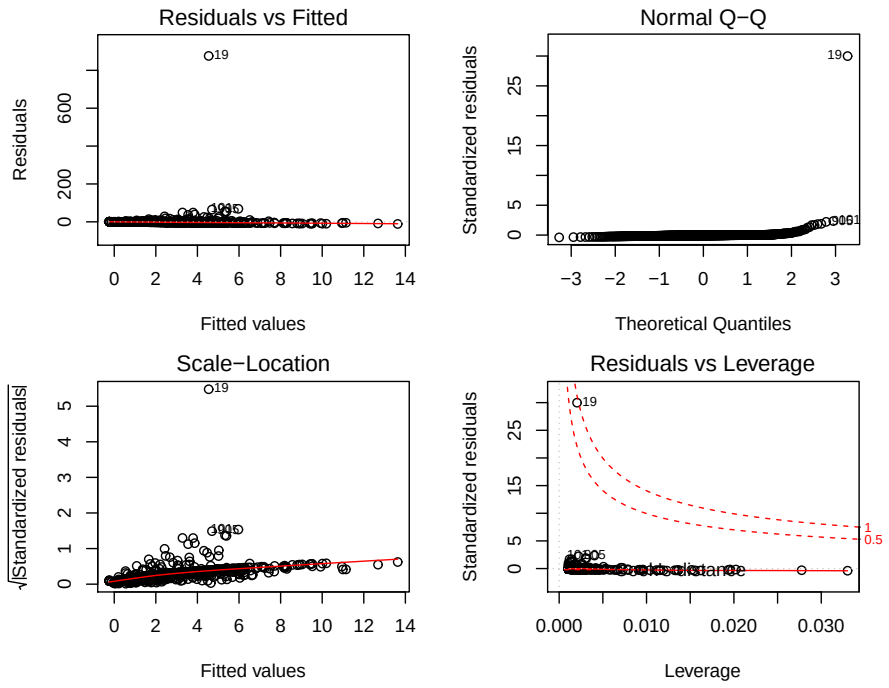


El resultat es mostra a continuació i s'obté un efecte positiu i estadísticament significatiu del regressor sobre la variable dependent:

```
> RegModel.3 <- lm(POB~MIG, data=Demografia)

> summary(RegModel.3)
```

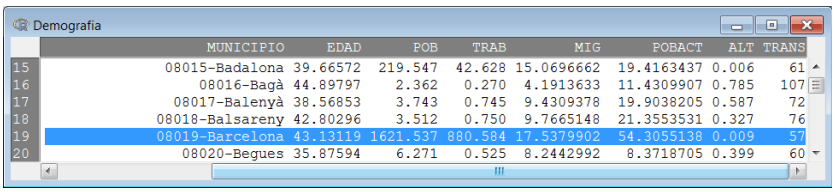
```
Call:
lm(formula = POB ~ MIG, data = Demografia)

Residuals:
    Min      1Q  Median      3Q     Max
 -31.31   -8.01   -3.91   -0.86 1607.93

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3875     3.1626  -0.123   0.9025
MIG           0.7979     0.2474   3.226   0.0013 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.98 on 939 degrees of freedom
Multiple R-squared:  0.01096, Adjusted R-squared:  0.009906
F-statistic: 10.41 on 1 and 939 DF,  p-value: 0.0013
```

És possible que hi hagi algun *outlier* en les variables? Vegem els gràfics de diagnòstic de l'estimació efectuada:



En tots els gràfics observem que el residu associat a l'observació 19 s'allunya considerablement de la resta dels residus. Comprovem quina observació ocupa aquesta posició visualitzant el conjunt de dades *Demografia*. Veiem que l'observació atípica correspon al municipi de Barcelona. Aquest resultat és lògic: aquest municipi té molts més habitants que la resta dels municipis catalans, amb la qual cosa el mesurament

d'aquesta observació no és errònia, ja que és lògic que aquest valor sigui tan alt comparat amb la resta de les observacions.



	MUNICIPIO	EDAD	POB	TRAB	MIG	POBACT	ALT	TRANS
15	08015-Badalona	39.66572	219.547	42.628	15.0696662	19.4163437	0.006	61
16	08016-Bagà	44.89797	2.362	0.270	4.1913633	11.4309907	0.785	107
17	08017-Balenyà	38.56853	3.743	0.745	9.4309378	19.9038205	0.587	72
18	08018-Balsareny	42.80296	3.512	0.750	9.7665148	21.3553531	0.327	76
19	08019-Barcelona	43.13119	1621.537	880.584	17.5379902	54.3055138	0.009	57
20	08020-Begues	35.87594	6.271	0.525	8.2442992	8.3718705	0.399	60

Com es pot identificar la presència d'*outliers*? A partir d'un model estimat, una opció és el test de valors atípics de Bonferroni, el qual reporta el p-valor per als residus estudentitzats absoluts, usant la distribució t. En R-Commander, això es fa accedint a la ruta següent del menú desplegable:

Models / Diagnòstics numèrics / Test de valors atípics de Bonferroni

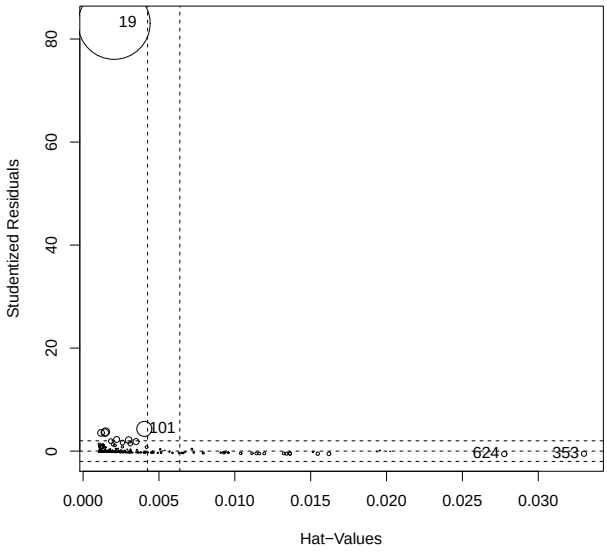
El resultat mostra dos valors atípics, el més destacat dels quals és l'observació 19, corresponent a Barcelona.

```
> outlierTest(RegModel.3)
     rstudent unadjusted p-value Bonferonni p
19   83.048751          0.0000e+00      0.000000
101   4.322249          1.7094e-05      0.016086
```

Alternativament, es pot calcular el gràfic d'influències, que compara en un gràfic bidimensional els valors estimats del model (*hat values*) i els residus estudentitzats. Es fa accedint a la ruta següent:

Models / Gràfiques / Gràfica d'influències

Aquesta acció mostra dos resultats. El primer és gràfic, en el qual es veu com el valor de l'observació 19 està clarament apartada de la resta de les observacions:



El segon apareix en la consola. Ens mostra una llista de possibles *outliers* i la distància de Cook (*CookD*). Aquesta mesura permet detectar l'estranyesa d'una observació, i serveix per a detectar aquelles observacions que tenen un efecte més gran en l'ajust que la resta, i que poden fer canviar els valors estimats pels paràmetres del model d'una manera substancial.

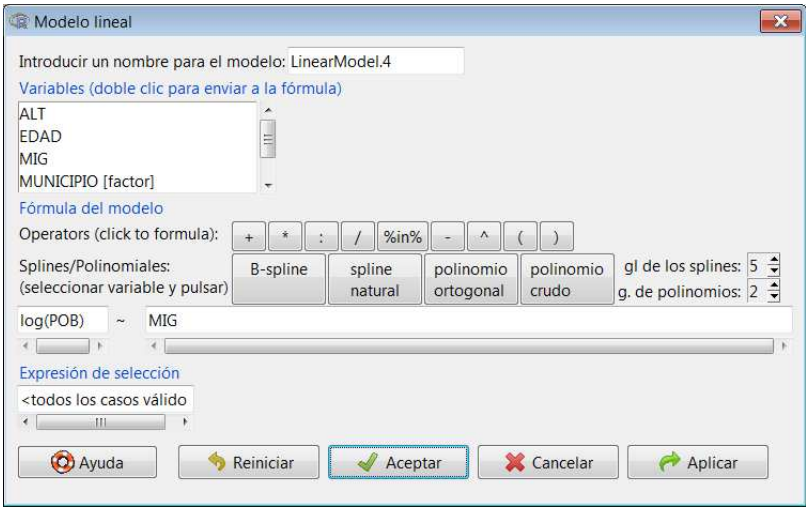
```
> influencePlot(RegModel.3, id.method="noteworthy", id.n=2)
       StudRes          Hat       CookD
19   83.0487509 0.002045796 0.92045949
101   4.3222485 0.004034997 0.19272734
353  -0.5104403 0.033011959 0.06671540
624  -0.5668982 0.027758536 0.06775765
```

Segons aquesta mesura, el principal *outlier* segueix essent l'observació 19. Quina pot ser la solució a la presència d'aquesta observació tan particular? Excloure-la del model estimat podria ser una solució, però l'observació no és errònia, i obviar-la significa no considerar la principal ciutat de Catalunya en un estudi sobre aquest territori. No sembla, doncs, una solució recomanable. Una solució alternativa és canviar la *forma funcional* de l'especificació, que pot passar per transformar alguna variable. Optarem per expressar la variable dependent en logaritmes, això és:

$$log(POB)_i = \beta_0 + \beta_1 MIG_i + e_i$$

Es donen dues conseqüències en produir-se aquesta transformació. La primera és que els valors de la variable *POB* es comprimeixen i, en conseqüència, hi ha menys distància entre el valor 19 i la resta. D'altra banda, també canvia la interpretació dels coeficients. Per a fer aquesta estimació, s'ha d'accedir a la ruta d'un model lineal, en el quadre de diàleg del qual podem especificar la relació funcional entre les variables:

Estadístics / Ajust de models / Model lineal



El resultat mostra una millora significativa de l'ajust del model i de la significació individual dels coeficients respecte al model anterior.

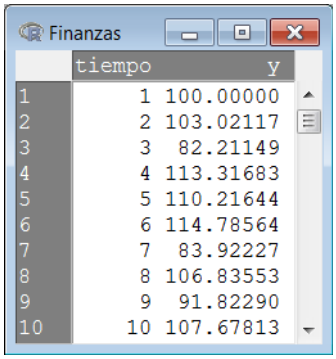
```
> LinearModel.4 <- lm(log(POB) ~ MIG, data=Demografia)

> summary(LinearModel.4)

Call:
lm(formula = log(POB) ~ MIG, data = Demografia)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3525 -1.1869 -0.2454  0.9859  6.5293

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -0.729731   0.087897  -8.302 3.54e-16 ***
MIG          0.090751   0.006875  13.200  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.556 on 939 degrees of freedom
Multiple R-squared:  0.1565,  Adjusted R-squared:  0.1556
F-statistic: 174.3 on 1 and 939 DF,  p-value: < 2.2e-16
```

4. Permanència estructural

Aquest problema sorgeix quan es trenca una de les hipòtesis bàsiques del model de regressió estàndard, que és la hipòtesi de permanència estructural. El problema sorgeix quan, en una sèrie temporal, en un punt del temps canvia la relació entre la variable dependent i un dels regressors. Per estudiar aquest problema amb un exemple senzill, analitzarem l'efecte del temps sobre l'evolució del preu d'un actiu financer fictici, que denominarem y. És a dir, estudiarem el model següent:

$$y_t = \beta_0 + \beta_1 t + e_t$$

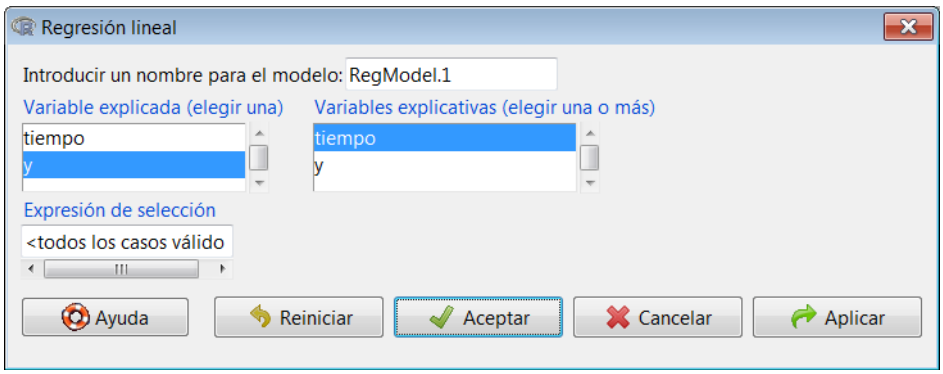
El primer pas és importar i visualitzar les dades.



Encara que aquí només es mostren 10 observacions, el conjunt de dades conté $T = 1000$ observacions temporals.

El primer pas és valorar el model de regressió:

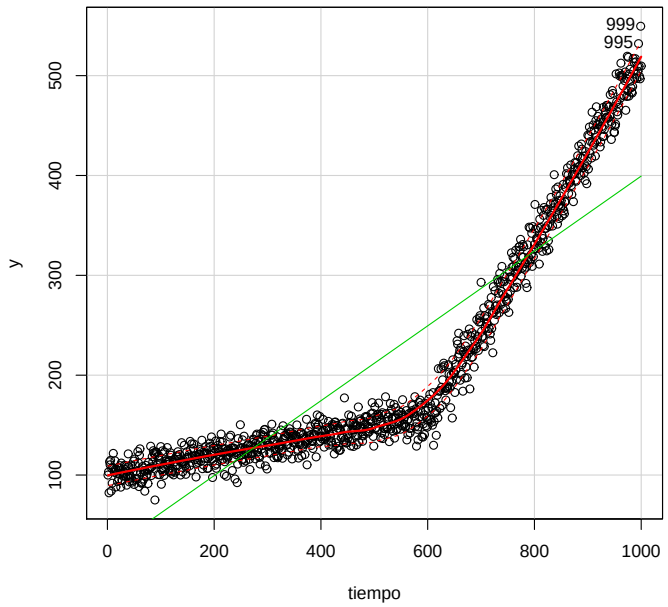
Estadístics / Ajust de models / Regressió lineal



El resultat de l'estimació és el següent:

```
> RegModel.1 <- lm(y~tiempo, data=Finanzas)

> summary(RegModel.1)

Call:
lm(formula = y ~ tiempo, data = Finanzas)

Residuals:
     Min       1Q   Median       3Q      Max
-123.724  -46.900   -2.406   44.058  150.465

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) 24.443812   3.499777   6.984 5.22e-12 ***
tiempo       0.374961   0.006057  61.903  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.29 on 998 degrees of freedom
Multiple R-squared:  0.7934,  Adjusted R-squared:  0.7932
F-statistic:  3832 on 1 and 998 DF,  p-value: < 2.2e-16
```

Com veiem, és un ajust bastant bo, i tant els coeficients valorats com el model estimat global són significatius estadísticament. Aquesta estimació dóna un coeficient $hat\beta_1 = 0,37$. Fins a quin punt és aquesta estimació correcta? Per entendre millor el concepte de permanència estructural, vegem en un pla cartesià el diagrama de dispersió de les dues variables: el temps en l'eix horitzontal i el preu de l'actiu financer en l'eix vertical. Aquest gràfic s'obté anant a l'opció *Gràfiques* del menú desplegable.



En aquest gràfic també apareix la recta estimada en el model $(24,44 + 0,37t)$, que és la mateixa per a tots els punts. No obstant això, veiem que la relació funcional entre les dues variables canvia sobre el punt $t = 600$. Veiem que abans i després el pendent canvia de manera significativa, com ho mostra la recta corba que ressegueix les observacions. Així doncs, sembla raonable estimar dos models, partint la mostra en dues parts, amb coeficients estimats diferents.

Estadísticament, com detectem la presència d'un canvi estructural? Un test útil en aquest sentit és el **test de Chow**. Aquest contrast consisteix a estimar dos models separant la mostra en dues submostres a partir d'un punt de tall determinat, per després comparar les SCE de la regressió per a tota la grandària mostral amb les SCE de les regressions per a cada una de les dues submostres fixades. Aquest test és una mica arbitrari, ja que requereix que fixem un punt de tall per endavant de manera aproximada.

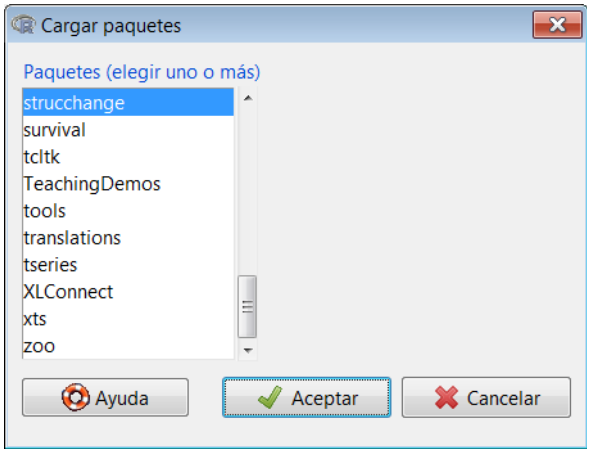
En R-Commander aquest contrast no està disponible en el menú, però això no significa que no es pugui fer mitjançant codi. Per a això, s'ha d'instal·lar el paquet *strucchange* en la consola:

```
> install.packages("strucchange")
```

Una vegada instal·lada aquesta biblioteca, s'ha de carregar. Això ho farem anant a la ruta del menú desplegable:

Opcions / Carregar paquets

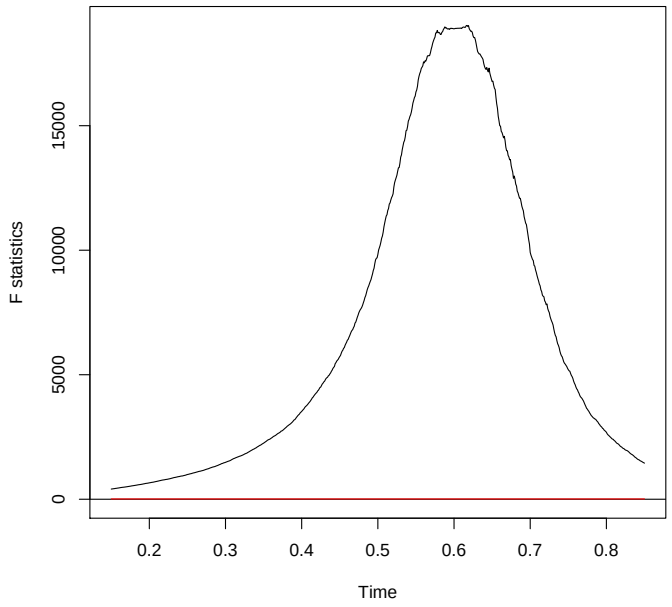
En el quadre de diàleg que ens apareixerà, seleccionem el paquet que acabem d'instal·lar.



La funció de R inclosa en aquest paquet que calcula l'estadístic de Chow és `Fstats`. Un fet positiu és que, opcionalment, podem introduir el període temporal en què sospitem que es produeix el canvi estructural. Si no l'especifiquem, aquesta funció calcula l'estadístic per a tots els punts de tall en la mostra. Les instruccions que hem d'introduir en la finestra d'instruccions són les següents:

```
> Fs <- Fstats(y ~ tiempo, data = Finanzas)
> plot(Fs)
```

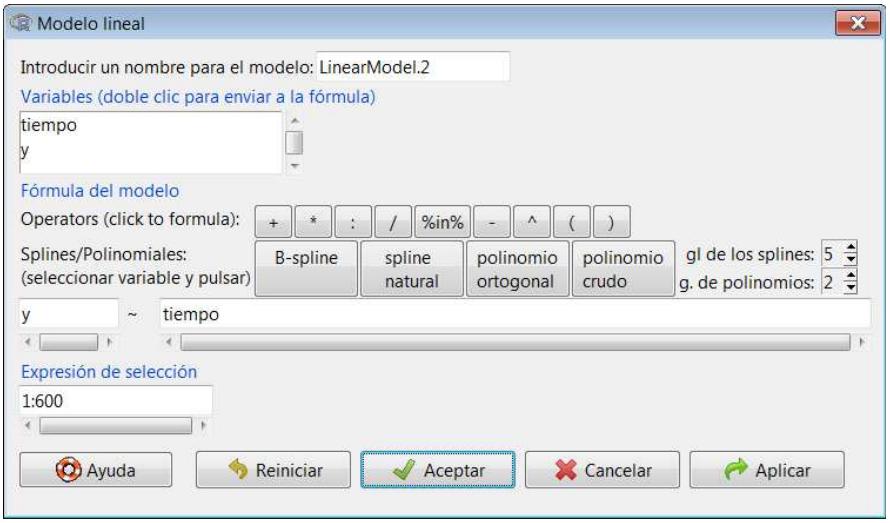
El gràfic resultant té la forma següent:



Què ens diu aquest gràfic? Doncs que el valor de l'estadístic F assoleix el màxim aproximadament en el 60% de la mostra, que coincideix amb el punt $t = 600$. La nostra estratègia serà estimar dos models, un amb la submostra $t = 1, \ldots, 600$ i un altre amb la submostra $t = 601, \ldots, 1000$. Per fer-ho, en el quadre de diàleg del model lineal introduirem, en l'opció *Expressió de la selecció*, la submostra per a la qual volem estimar el model.

Vegem el resultat de la primera estimació per a la submostra $t = 1, \ldots, 600$.

Estadístics / Ajust de models / Model lineal



```
> LinearModel.2 <- lm(y ~ tiempo, data=Finanzas, subset=1:600)

> summary(LinearModel.2)

Call:
lm(formula = y ~ tiempo, data = Finanzas, subset = 1:600)

Residuals:
    Min      1Q  Median      3Q     Max
-33.897  -6.760   0.229   6.522  33.266

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.001e+02  8.268e-01  121.08   <2e-16 ***
tiempo      9.884e-02  2.384e-03   41.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.11 on 598 degrees of freedom
Multiple R-squared:  0.742, Adjusted R-squared:  0.7415
F-statistic:  1720 on 1 and 598 DF,  p-value: < 2.2e-16
```

I ara el resultat de la segona estimació per a la submostra $t = 601, \ldots, 1000$.

```
> LinearModel.3 <- lm(y ~ tiempo, data=Finanzas, subset
    =601:1000)

> summary(LinearModel.3)

Call:
lm(formula = y ~ tiempo, data = Finanzas, subset = 601:1000)

Residuals:
    Min      1Q  Median      3Q     Max
-42.796 -10.810  -0.868  10.745  47.448

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -3.814e+02  5.315e+00  -71.77   <2e-16 ***
tiempo       8.957e-01  6.571e-03  136.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.18 on 398 degrees of freedom
Multiple R-squared:  0.979, Adjusted R-squared:  0.979
F-statistic: 1.858e+04 on 1 and 398 DF,  p-value: < 2.2e-16
```

D'aquestes dues estimacions obtenim importants conclusions. La primera és que els paràmetres estimats són molt diferents, això és, per a la primera submostra obtenim un pendent $\hat{\beta}_1 \backsimeq 0,1$; i per a la segona submostra $\hat{\beta}_1 \backsimeq 0,9$. La relació entre les variables ha canviat, doncs, considerablement en el punt $t = 600$. A més a més, l'ajust dels dos submodels és molt millor que per al model global, ja que les dues rectes estimades s'ajusten molt millor als dos trams d'observacions.

Bibliografia

Artís Ortuño, M.; del Barrio Castro, T.; Clar López, M.; Guillén Estany, M.; Suriñach Caralt, J. (2011). *Econometría*. Barcelona. Material didàctic UOC.

Liviano Solís, D.; Pujol Jover, M. (2013). *Matemáticas y Estadística con R*. Barcelona. Material didàctic UOC.