Pràctica 10

Reg. múltiple: tests de lligadures

Aprofundirem en la significació de les variables en el model lineal. Més específicament, analitzarem si una o més les variables poden sortir d'un model donat, tot i mesurant la variació de variància explicada pel model (SSR) "amb i sense" les variables. Equivalentment, avaluant la variació en la suma dels quadrats dels residus (SSE. Cal tenir present que, *traient variables d'un model disminueix la variància explicada al mateix temps i pel mateix volum que s'incrementa la suma de quadrats dels residus.* Són tests-F, basats en un estadístic quocient amb llei *F-de-Fisher-Snedecor*, una metodologia que permet contrastar restriccions lineals més generals, els anomenats **tests de lligadures lineals**.

Continuem amb la base de dades `cases.txt` amb $n = 26$.

1. Introduïm al model les variables d'una manera seqüencial, les que a l'investigador li semblen més rellevants. Aixo té sentit perquè *alguns resultats depenen de l'ordre d'entrada de les regressores.*

```
dades<-read.table("cases.txt",header=T)

mod<-lm(PREU~SUPERF+BANYS+STORM+FP+SUPMTERR+HABIT+DORMIT+PL_GAR,data=dades)
# ordre que decideix l'analista dels preus
n<-nrow(mod$model)
p<-mod$rank  # model amb intercept: rang (p) = nº de variables (k) + 1:
k<-p-1       # 8 variables regressores
gl<-n-k-1    # graus de llibertat dels residus
```

2. Fem el resum del model.

```
smod<-summary(mod); smod;
```

El test global, en què la hipòtesi nul·la diu que tots els coeficients de les variables són zero ($H_0 : \beta_1 = \ldots = \beta_8 = 0$) és significatiu (p-valor=$3.147 \cdot 10^{-7} \leq \alpha = 0.05$). Per tant, acceptem l'alternativa: el model és significatiu (no és nul).

3. Els tests individuals per a $j = 1, \ldots, 8$, que veiem a l'apartat de coeficients del sumari, són els tests bilaterals amb hipòtesi nul·la $\mathbf{H_0 : \beta_j = 0}$, basats en l'estadístic **t-de student**:

$$\mathbf{T_j} = \frac{\hat{\beta}_\mathbf{j}}{\hat{\sigma}_{\hat{\beta}_\mathbf{j}}} \sim \mathbf{t_{n-k-1}} \qquad \text{sota } \mathbf{H_0}$$

Estem contrastant si, **tenint en compte que al model hi ha totes les altres regressores, la presència de X_j és significativa o no**. En aquest test **no** importa l'ordre de les regressores.

```
coefs<-smod$coefficients; coefs
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 18.6376638  5.2409572  3.5561565 0.0024293843
## SUPERF      -0.2459520  0.1353177 -1.8175893 0.0867952842
## BANYS        2.3745908  2.5578653  0.9283486 0.3662212991
## STORMsí     10.8186631  2.3002029  4.7033516 0.0002047449
## FP           6.9097646  3.0835831  2.2408232 0.0386797459
## SUPMTERR     0.2635222  0.1351086  1.9504466 0.0678084452
## HABIT        3.9043738  1.6156173  2.4166451 0.0271943397
## DORMIT      -7.6974440  1.8294263 -4.2075727 0.0005918719
## PL_GAR       1.7708610  1.4043102  1.2610184 0.2243339796
```

Prova: Calculem directament el valor observat de t_{ob} dividint l'estimació puntual (`Estimate`, col1) per l'error típic (Std. Error, col2), i comprovem que els p-valors es calculen fent

$$\mathbf{2(1-P\{t_{n-k-1} > |t_{ob}|\})}$$

```
all.equal(coefs[,3],coefs[,1]/coefs[,2])          # comprovem que la col3 de coefs és col1/col2
all.equal(coefs[,4],2*(1-pt(abs(coefs[,3]),gl)))  # comprovem el càlcul del p-valor
```

Recordeu: `all.equal()` comprova la igualtat, llevat d'arrodoniments.

4. Seguidament fem els mateixos tests, a partir d'estadístics F, pensant-los com a **tests de lligadures lineals que comparen les sumes de quadrats del model sense cap restricció (totes les regressores) i "amb" la restricció (coef. de la regressora igual a zero)**. Hi ha una única lligadura lineal: $\beta_1 = 0$.

   Calculem les sumes de quadrats del model total (`mod`) i sumes de quadrats del model que té totes les variables "menys la primera" (`mod_1`). Al treure una variable la suma de quadrats explicada pel model és més petita, nmentre que la suma de quadrats residual o dels errors s'incrementa en la mateixa quantitat (la suma de quadrats totals és constant). Mirem si aquestes diferències en les sumes de quadrats són significatives, a partir dels quocients de les mitjanes quadràtiques respectives; aquests quocients tenen llei F:

$$\mathbf{F\_1} = \frac{\mathbf{(SSR - SSR\_1)/1}}{\mathbf{SSE/(n-k-1)}} = \frac{\mathbf{(SSE\_1 - SSE)/1}}{\mathbf{SSE/(n-k-1)}} \sim \mathbf{F_{1,n-k-1}} \quad \text{sota} \quad H_0 : \beta_1 = 0.$$

   Esquemàticament, ho fem per al coeficient de la primera regressora SUPERF:

   (a) Per al model amb totes les regressores, `mod`, calculem la suma de quadrats: SSE.

   (b) Per al model sense la primera regressora, `mod_1`, calculem la suma de quadrats: SSE_1.

   (c) Els graus de llibertat del numerador son: `gln=nombre de lligadures`=1, i els del denominador $gld = gl = n - p = n - k - 1$.

   (d) Calculem les mitjanes quadràtiques corresponents, l'estadístic F_1 i el p-valor:

```
# mod i sumes de quadrats dels residus de mod:
mod<-lm(PREU~SUPERF+BANYS+STORM+FP+SUPMTERR+HABIT+DORMIT+PL_GAR,data=dades)
sse<-sum(mod$residuals^2)
# mod_1 i sumes de quadrats dels residus de mod£\_£1:
mod_1<-lm(PREU~BANYS+STORM+FP+SUPMTERR+HABIT+DORMIT+PL_GAR,data=dades) # sense SUPERF: beta1=0
sse_1<-sum(mod_1$residuals^2)
#
gln<-1       # graus de llibertat del numerador: num de lligadures
gld<-n-k-1;  # graus de llibertat del denominador: gl de mod (n-k-1)
F_1<-((sse_1-sse)/gln)/(sse/gld)  # test amb lligadures: H0: beta1=0
#
p_1<-1-pf(F_1,gln,gld)
p_1; coefs[2,4]     # comprovem que és el mateix p-valor del t-test

## [1] 0.08679528
## [1] 0.08679528
```

Interpretació: Algunes variables semblen susceptibles de sortir del model, però d'una en una, perquè els tests només ens parlen de la significació d'una variable en un model on hi ha les altres. *Nota:* Podeu comprovar que el valor observat de F_1 és el valor observat de t al quadrat.

Resum: Quan es fa el test $H_0 : \beta_j = 0$ mitjançant l'estadístic $\mathbf{F}\_1$, aquests diem que estem fent **ANOVAS basats en les anomenades sumes de quadrats de tipus II (*type II sums of squares*)**[1] Amb R, es calculen al fer **anova(mod)**. . Hem vist que amb R apareixen al **sumary(mod)** com a t-tests. Esquemàticament, contrasten:

- **X1 | intercept, X2, X3, ..., Xk :** ***És significatiu l'efecte d'afegir*** X_1 ***a un model on ja hi ha*** $X_2, \ldots, X_k$***?***
- **X2 | intercept, X1, X3, ..., Xk :** ***És significatiu l'efecte d'afegir*** X_2 ***a un model on ja hi ha*** $X_1, X_3, \ldots, X_k$***?***
- ⋮
- **Xk-1 | intercept, X1, X2, ..., Xk-2, Xk**
- **Xk | intercept, X1, X2 X1, X2, ..., Xk-1**

[1] També hi ha les sumes de quadrats de tipus III, pero són equivalents en regressió múltiple, difereixen en "disseny d'experiments" quan hi ha interacció entre els factors.

5. **Amb la funció `anova()` de R es fan els tests ANOVAS "sequencials" (*type I sums of squares*).**

   - **X1 | intercept :** ***És significatiu l'efecte d'afegir*** X_1 ***a un model amb només l'intercept?***
   - **X2 | intercept, X1 :** ***És significatiu l'efecte d'afegir*** X_2 ***a un model amb intercept i*** X_1***?***
   - **X3 | intercept, X1, X2**
   - **...**
   - **Xk-1 | intercept, X1, X2, ..., Xk-2**
   - **Xk | intercept, X1, X2, ..., Xk-1**

   Els tests basats en les sumes de quadrats de tipus I avaluen, de manera seqüencial, si cada nova variable és rellevant en el model, un cop les anteriors ja són dins. L'ordre és fonamental.

   Atenció: La particularitat d'aquests tests és que el denominador sempre és MSE=SSS/(n-k-1), és a dir, l'estimació de la variància obtinguda al model per a totes les regressores (`mod`). Només el test per a l'última variable coincideix en els tipus I i tipus II.

```
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: PREU
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## SUPERF     1 2143.26 2143.26 96.3357 2.034e-08 ***
## BANYS      1  151.20  151.20  6.7962 0.0184133 *
## STORM      1  398.71  398.71 17.9213 0.0005599 ***
## FP         1  327.29  327.29 14.7113 0.0013246 **
## SUPMTERR   1   42.45   42.45  1.9081 0.1850669
## HABIT      1   57.56   57.56  2.5872 0.1261427
## DORMIT     1  423.33  423.33 19.0278 0.0004244 ***
## PL_GAR     1   35.38   35.38  1.5902 0.2243340
## Residuals 17  378.21   22.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Manualment, fem els  primers tests
## Atenció: el denominador sempre és SSE/gl del model global mod amb totes les regressores
sse<-sum(mod$residuals^2)
gld<-gl
# sumes de quadrats de mod1: intercept i X1
mod1<-lm(PREU~SUPERF,data=dades)
sse1<-sum(mod1$residuals^2)
sst<-(n-1)*var(dades$PREU)
sse0<-sst # la constant (intercept) no explica variabilitat de la resposta
gln<-1
F1<-((sse0-sse1)/1)/(sse/gld)
p1<-1-pf(F1,1,gld)          # test de signif. de X1  (respecte del model constant)
#
mod12<-lm(PREU~SUPERF+BANYS,data=dades)
sse12<-sum(mod12$residuals^2)
gln<-1
F2vs1<- ((sse1-sse12)/gln)/(sse/gld)
p2vs1<-1-pf(F2vs1,gln,gld)    # test de signif. de X2, respecte d'un model qua ja té X1
```

```
print(list(SUPERF_F_tipI=F1,SUPERF_pval_tipI=p1,BANYS_F_tipI=F2vs1,BANYS_pval_tipI=p2vs1))
```

```
## $SUPERF_F_tipI
## [1] 96.33565
##
## $SUPERF_pval_tipI
## [1] 2.034084e-08
##
## $BANYS_F_tipI
## [1] 6.796198
##
## $BANYS_pval_tipI
## [1] 0.01841331
```

Interpretació: El resultat de `anova(mod)` suggereixen que "algunes" de les variables 5 (SUPMTERR), 6 (HABIT), 8 (PL_GAR) es podrien treure del model. La pregunta és: totes elles alhora ? : No ho podem assegurar, caldrà comprovar-ho fent un test de lligadures apropiat. Ho veiem tot seguit.

6. Tests generals amb lligadures:

$$\mathbf{H_0} : \beta_5 = \beta_6 = \beta_8 = \mathbf{0} \tag{10.1}$$

són $\mathbf{r = 3}$ lligadures lineals.

- El model global és el model amb les 8 regresssores, és a dir,

$$\texttt{mod:} \quad \mathbf{y} = \beta_0 + \beta_1\mathbf{X_1} + \beta_2\mathbf{X_2} + \beta_3\mathbf{X_3} + \beta_4\mathbf{X_4} + \beta_5\mathbf{X_5} + \beta_6\mathbf{X_6} + \beta_7\mathbf{X_7} + \beta_8\mathbf{X_8} + \epsilon$$

  Diguem **SSE** la suma de quadrats dels residus d'aquest model.
- Considerem el model amb la restricció (lligadura) descrita a la hipòtesi nul·la (possiblement amb estimacions dels coeficients diferents)

$$\texttt{modll:} \ \mathbf{y} = \beta_0 + \beta_1\mathbf{X_1} + \beta_2\mathbf{X_2} + \beta_3\mathbf{X_3} + \beta_4\mathbf{X_4} + \beta_7\mathbf{X_7} + \epsilon$$

  Diguem $\mathbf{SSE}_{ll}$ les sumes de quadrats dels residus d'aquest model restringit.
- L'estadístic de contrast per al test (10.1) es basa en l'estadístic:

$$\mathbf{F_{ll}} = \frac{\mathbf{(SSE_{ll} - SSE)/r}}{\mathbf{SSE/(n-k-1)}} \sim \mathbf{F_{r,n-k-1}} \quad \text{sota } H_0 \tag{10.2}$$

Exercici 1: Feu el codi R del test de lligadures anteior. Comproveu que és significatiu i que, per tant, no podem treure totes les variables alhora. Comproveu també quina és la disminució de R2-aj (valor que permet comparar la qualitat de l'ajust en models amb diferent nombre de regressores, sempre i quan tots dos models tinguin intercept). Rebutjar la H_0 ens ve a dir que la disminució en el valor de R2-aj és significativa.

Comproveu que heu obtingut els resultats correctes, aplicant el codi següent:

```
anova(mod,modll)        # fa el test de lligadures
##  anova(modll,mod)    # idem: l'ordre dels models no importa

require(car)
linearHypothesis(mod, c("SUPMTERR=0","HABIT=0","PL_GAR=0"))  # el mateix
```

Exercici 2: Feu el test de lligadures per treure BANYS (2), SUPMTERR (5) i PL_GAR (8). Comproveu que es poden treure aquestes tres variables alhora i acceptar el model restringit. Escriviu la hipòtesi nul·la. Comproveu que R2-aj té un valor "més similar" al R2-aj del model total (el de 8 variables). Escriviu la hipòtesi nul·la. Els signes dels coeficients us semblen "coherents"?

Exercici 3: Partiu del model de l'exercici anterior (eliminades BANYS, SUPMTERR i PL_GAR), que podeu anomenar ara `modf` (oblidem ja `mod`, i fixem aquest model com a base):

(a) Feu el test $\beta_3 = 10$ (coeficient de FP) mitjançant un $t-$test. *Nota:* Recordeu que l'estadístic de de $H_0 : \beta_j = \beta_j^*$ amb el $t-$test és $\mathbf{T_j} = \frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathbf{t_{n-k-1}}$ (sotaH_0). Quina és la conclusió ? (justificada).

(b) Feu el mateix test test de lligadures amb un $F-$test. Com a test de lligadures feu-lo manualment i comproveu que ho heu fet correctament amb la funció `linearHypothesis()` de la llibreria `car`. **Atenció**: La funció `anova(modf,modll)` no es pot aplicar perquè canviem la variable resposta.

(c) Sobre el model `modf`, feu el test $\beta_2 = \beta_3$ (igualtat dels coeficients de STORM i FP) de dues maneres: *(1)* amb la funció `linearHypothesis()` de la llibreria `car` i *(2)* calculant manualment F_{ll} i p_{ll}. **Nota:** Per fer aquest test, caldrà que recodifiqueu la variable STORM com a variable numèrica, diguem-li *st* (fet anteriorment) i utilitzar aquesta variable `st` en els models. Quina és la conclusió del test?