Pràctica 11

Regressió múltiple: Bandes. Noves observacions.

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

```
library(car);library(lmtest);library(lawstat)
library(nortest);library(doBy);library(foreign)
dades<-read.table("cases.txt",header=T)
```

1. Gestió inicial de la base de dades.

   - Camviem la variable amb valors alfanumèrics `(STORM)` a numèrica `(st)` i reduïm la base de dades que només contingui les 5 variables explicatives següents (aquest model resulta de la pràctica anterior): `SUPERF, st, FP, DORMIT, HABIT`, amb aquest ordre, que formaran el model bàsic:

     ```
     # transformació de la variable STORM
     n<-nrow(dades)
     st<-numeric(n)
     st[dades$STORM=="no"]<-0
     st[dades$STORM=="sí"]<-1
     dades2<-data.frame(PREU=dades$PREU, SUPERF=dades$SUPERF,st=st,FP=dades$FP,
                        DORMIT=dades$DORMIT, HABIT=dades$HABIT);   head(dades2)
     ```

   - Considerem com a model bàsic `mod` el de les 5 variables explicatives `SUPERF, STORM, FP, DORMIT` i `HABIT`. *Recordeu:* L'opció `x=TRUE , y=TRUE` permet obtenir x i y a l'output del model: `x i y es recuperen fent mod$x i mod$y`

     ```
     mod<-lm(PREU ~ SUPERF+ st+ FP+ DORMIT + HABIT ,data=dades2,x=TRUE,y=TRUE)
     smod<-summary(mod); smod
     ```

     Es veu que totes les variables tenen una aportació **significativa**, fins i tot quan les altres variables són dins del model, i que globalment expliquen un **86.69%** de la variabilitat de la resposta (un **83.36%** ajustat).

2. **Noves observacions: Bandes de confiança per a la resposta mitjana i de predicció**.

   Fem una funció `band_conf.mult()` per calcular les bandes, utilitzant la funció `predict()`. Després ho representem gràficament.

   Remarca!: En regressió múltiple no es poden representar les bandes respecte del conjunt d'explicatives multivariant. Fem una gràfica alternativa de les prediccions $\hat{y}_i = \hat{\mu}_i$ ordenades de menor a major, tot i afegint-hi les bandes i una gràfica del núvol de punts de y, per mirar si "majoritàriament" estan dins de les bandes de predicció.

```
band_conf.mult<-function(mod, alpha=.05)
{
y<-mod$y
output<-cbind(y, predict(mod,interval="confidence",level=1-alpha),
              predict(mod,interval="prediction",level=1-alpha)[,-1]) # y,  pred,  muCI, predCI
colnames(output)<-c("y","haty","linf.mu","lsup.mu","linf.pr","lsup.pr")
output<-output[order(output[,2]),]          # ordena les dades segons la predicció prediccions
round(output,digits=2)
}
bandes<-band_conf.mult(mod=mod); head(bandes)

matplot(bandes[,2], bandes[,-1],
        lty = c(1,1,1,2,2), col=c(1,2,2,4,4), type = "l",
        ylab = "prediccions, intervals i observacions", xlab="prediccions",
        main="bandes de conf. (vermell) i de predic. (blau)", cex.main=.8)
points(bandes[,2], bandes[,1],pch=19 )  # valors observats de la resposta
```

Nota: Observem que la majoria (tots menys 6) de respostes (punts) estan dins de les bandes de confiança i tots menys 1 punt estan dins de les bandes de predicció. *Nota!:* Pot ser que la gràfica no surti bé si treballeu amb molts punts.

Què és la línia negra central de la figura? Per què les bandes són irregulars?

3. **Interval de confiança i de predicció per a un vector donat de noves observacions.**

   Nota: En l'apartat anterior no ens hem preocupat de donar un vector de noves observacions per obtenir les bandes, ho hem fet amb funcions de R que ho apliquen sobre les mateixes dades observades com si fossin noves dades. Ara considerarem dades noves.

   Sigui `xh` un data frame de noves observacions. *Atenció !:* cal anomenar les variables amb el mateix nom que a l'arxiu inicial.

   Creem un dataframe amb les noves observacions, només amb els valors de les explicatives que entren al model. En aquest exemple hi ha dos casos nous:

```
names(dades)
xh<-data.frame(SUPERF=c(800,900),st=c(0,1),FP=c(1,0), DORMIT=c(2,3),HABIT=c(5,5)); xh  ## copiem noms i ordre
```

   Fem les prediccions amb els intervals de confiança i de predicció. Ho guardem en un nou data frame.

```
mod<-lm(PREU ~ SUPERF+ st+ FP+ DORMIT + HABIT ,data=dades2,x=TRUE,y=TRUE)

muh.CI<-predict(mod,newdata=xh,interval=c("confidence"))
prxh.CI<-predict(mod,newdata=xh,interval=c("prediction"))
xh.pred<-data.frame(xh,muh.CI,prxh.CI); xh.pred
# la columna de les prediccions  està repetida (fit=fit.1) i es podria eliminar
```

Exercici 1 Considereu un model de regressió múltiple (mo4) per al preu en el qual hi hagi les 4 primeres regressores de `mod`, és a dir, sense HABIT.
- Repetiu les bandes de confiança i de predicció per al nou model, amb les seves grafiques.
- Estimeu la resposta mitjana i la predicció dels mateixos valors `xh` amb aquest nou model.

Exercici 2 Considereu un model de regressió múltiple (fit) per la variable transformada $\frac{100}{\texttt{mpg}}$ respecte les explicatives `disp, hp, wt, am`, del fitxer `mtcars` de la llibreria base de R.
- Feu les bandes de confiança i de predicció per al model, amb les seves grafiques.
- Estimeu la resposta mitjana i la predicció d'un nou cas on: `disp=100, hp=20, wt=2.5, am=1`.

Pràctica 12

Anàlisi dels residus

Continuem amb la base de dades *preuscases.sav* amb $n = 26$ i considerem el model lineal del `PREU` amb les 5 regressores:`SUPERF, st, FP, DORMIT, HABIT`, on `st` és la transformada numèrica de `STORM`.

```
library(car);library(lmtest);library(lawstat);library(nortest);library(doBy);library(foreign)
dades<-read.table("cases.txt",header=T)
n<-nrow(dades);  st<-numeric(n)
st[dades$STORM=="no"]<-0;  st[dades$STORM=="sí"]<-1
dades2<-data.frame(dades[c(1,3)],st,dades[c(4,2,5)]) # head(dades2)
mod<-lm(PREU ~. , data=dades2, x=TRUE, y=TRUE)
k<-mod$rank-1; ## n<-length(mod$y) ## el tenim d'abans
p<-mod$rank    ## p=k+1, pel terme independent (intercept)
X<-mod$x       ## matriu del disseny
```

recomano especificar totes les regressores i NO posar ~.

Exercicis. - Per resoldre'ls, es recomana revisar la pràctica 6 (regressió simple).

1. Calculeu els residus (bruts e_i, estandarditzats r_i i estudentitzats dr_i) d'aquest model.
2. Feu un diagnòstic de les hipòtesis del model:
   - Estudieu la linealitat del model mitjançant una grafica de residus. Comenteu els aspectes destacables de la gràfica.
   - Es pot fer el test de manca de linealitat per aquest model? Raoneu la resposta.
     El test de *lack-of-fit* o manca d'ajust lineal es pot fer sempre que **hi hagi rèpliques**, és a dir, casos amb idèntics valors de les variables regressores. En primer lloc, es crea un nou factor que té per valors les combinacions de valors de les regressores (fixeu-vos bé en el codi !!!):

```
 paste(dades2[1,-1],collapse="")   # ho apliquem al primer cas (fila 1)
# ara ho apliquem a tot el data.frame
 unir<-function(v){paste(v,collapse="")}  # funció que col.lapsa un vector en un únic valor
 fact<-apply(dades2[-1],1,FUN="unir")
 dades2$fact<-fact
 taula<-table(fact)
 any(taula>1)       # !!
lm.lof<-lm(PREU~0+as.factor(fact),data=dades2)
summary(lm.lof)     # !!
anova(mod,lm.lof)   # !!
```

   - Estudieu la igualtat de variàncies del model: gràficament i amb algun test. Interpreteu-ho.
   - Estudieu la normalitat del model: gràficament i amb algun test. Interpreteu-ho.
3. Feu un diagnòstic de *punts especials, atípics o anòmals*: *ouliers*, de *palanca*, *influents*. A cada apartat, calculeu els indicadors de l'anomalia, feu el recompte de casos que superen els llindars i feu les gràfiques on es vegin els valors i els llindars. Per a cada anomalia, doneu un llistat dels casos que superen els llindars. *Nota:* Per localitzar els casos que satisfan una condició

```
subset(dataframe,condicio)   ## vegeu el codi de la pag. 4
```

també es pot fer amb la funció which(condició), entre d'altres

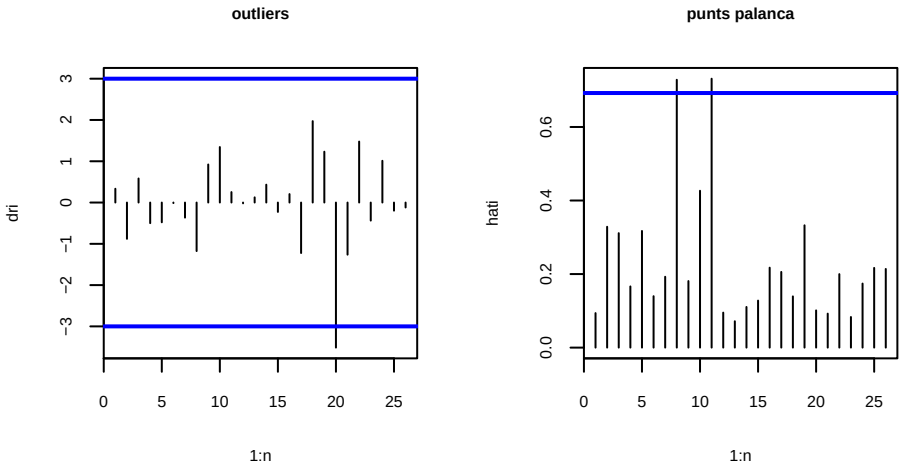
Per simplificar notations, diguem $\mathbf{p = k + 1}$ al nombre de coeficients quan hi ha intercept.

- *Outliers*: Casos amb residu estandarditzat i/o estudentitzat amb valor absolut més gran o igual que **3**. Podeu utilitzar la funció `outlierTest()`. Recordeu que el *p-valor* del test de Bonferroni per contrastar la significació del màxim outlier en valor absolut és:

$$\mathbf{2n(1 - pt(drmax, n - p - 1))} \qquad \text{on} \quad \mathbf{drmax} = \max_i |dri|$$

- Efecte *palanca* d'un cas (fila, $\mathbf{x_i^t}$). Utilitzeu la funció `hatvalues()` per calcular els efectes de palanca i comproveu que són els elements diagonals de la matriu $\mathbf{H}$ (*hatmatrix*). Apliqueu els llindars:

$$\mathbf{h_{ii} = x_i^t (X^t X)^{-1} x_i \geq \frac{3p}{n}}$$



- *Influència* del cas i *en el coeficient* $\hat{\beta}_j$, $j = 1, \ldots, p$. Utilitzeu la funció `dfbetas()` per calcular les influències en els coeficients i apliqueu els llindars:

$$\mathbf{|DFBETAS_{j(i)}| \geq \frac{2}{\sqrt{n}}}$$

- Influència del cas i en la predicció del propi valor. Podeu utilitzar la funció `dffits()` per calcular les influències en les prediccions i aplicar els llindars:

$$\mathbf{|DFFITS_{j(i)}| \geq 2\sqrt{\frac{p}{n}}}$$

error: ha de ser qf(0.5,p,n-p)

- Distància de Cook: cas influent en el model. El llindar més utilitzat és **1**. Alguns experts recomanen Un altre llindar, que depén de la mida de la mostra i del nombre de regressores és **F**=`qf(0.5,p,1-p)` és la quantila 0.5 de la llei de Fisher-Snedecor). Podeu utilitzar la funció `cooks.distance()` per calcular les influències en el model i aplicar els llindars:

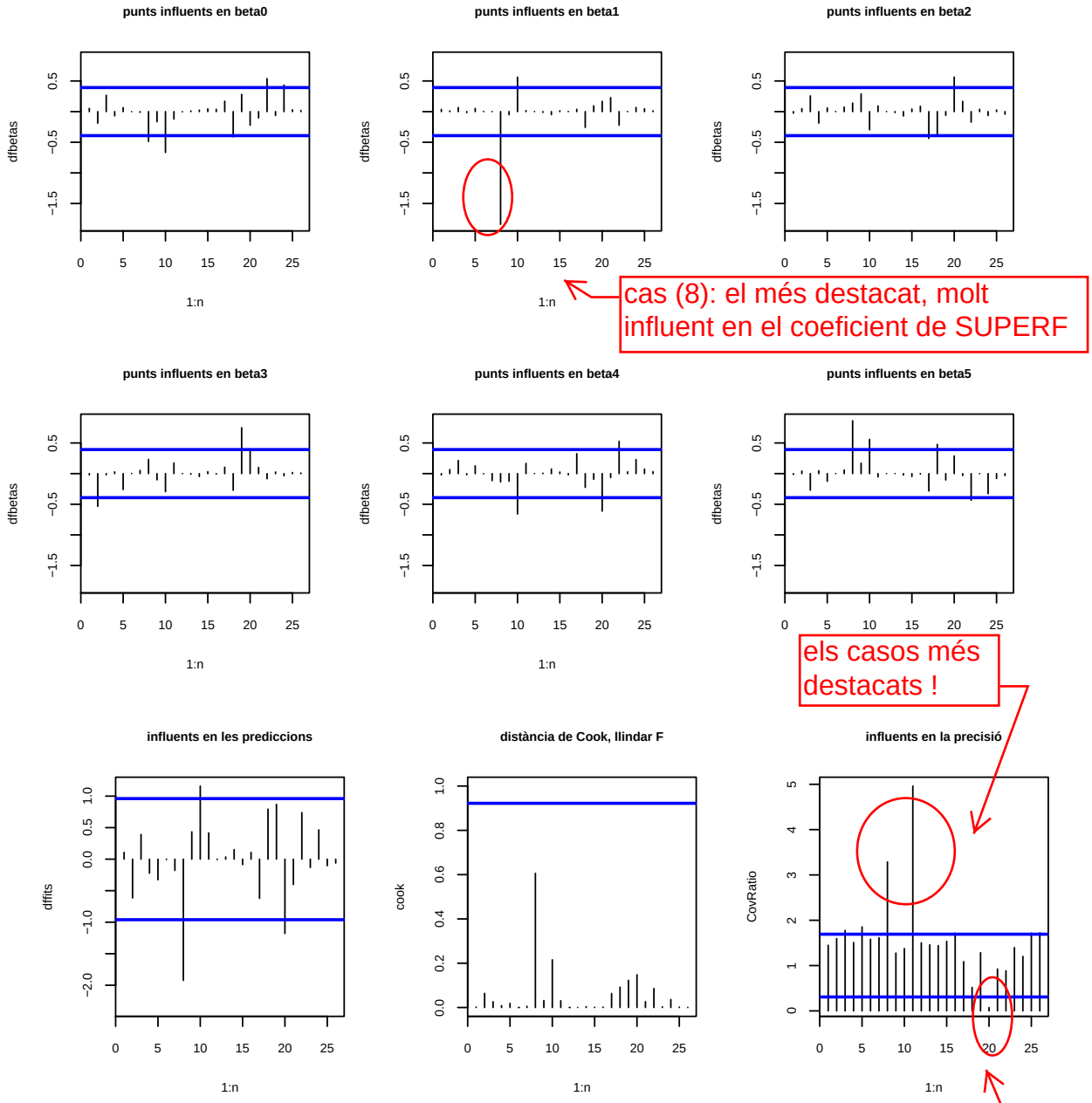
$$\mathbf{\frac{dr_i^2}{p} \cdot \frac{h_{ii}}{1 - h_{ii}} \geq F}$$

- El **CovRatio**, mesura els canvis en la matriu de covariàncies dels coeficients. Com que, la matriu de covariàncies dels betas és $\Sigma_{\hat{\beta}} = \sigma^2 (X^t X)^{-1}$, el "determinant" $|\quad|$ d'aquesta matriu és l'anomenada

variància generalitzada, i el **covRatio** mesura el rati de canvi en el determinant entre que el punt sigui al model o no hi sigui:

$$\mathbf{CovRatio} = \frac{\mathbf{MSE_{(i)}}\,|\mathbf{X^t X}|}{\mathbf{MSE}\,|\mathbf{X^t_{(i)}X_{(i)}}|} = \frac{\mathbf{MSE_{(i)}}}{\mathbf{MSE}}\,\frac{\mathbf{1}}{\mathbf{1-h_{ii}}} \notin \mathbf{1} \pm \frac{\mathbf{3p}}{\mathbf{n}}$$

Recordeu que el determinant de la matriu inversa és l'invers del determinant de la matriu i que el subíndex entre parèntesi, (**i**), indica que els càlculs s'han fet sense el cas. Podeu utilitzar la funció `covratio()`. Fora de l'interval, el cas influeix en la precisió de les estimacions (la qual cosa implica errors típics i marges d'error excessius).



Casos atípics: Recomptes i llistats. Fet per als outliers i dfbetas.

Per completar l'exercici 3, feu el mateix per als palanca, dffits, cooks i covratio.

```
outlierTest(mod)                ## no és significatiu
out<-abs(dri)>3
sum(out)
list(outliers=subset(dades2,out))

cont.dfbetas<-function(v){sum(abs(v)>2/sqrt(n))}
list(num.influ.coefs=apply(dfbetas,2,"cont.dfbetas"))

inf0<-abs(dfbetas[,1])>(2/sqrt(n))
#subset(dades2,inf0)
influ.coefs<-list(subset(dades2,inf0))
for(j in (1:k)){
  inf<-abs(dfbetas[,j+1])>(2/sqrt(n))
  influ.coefs[[j+1]]<-subset(dades2,inf)
}
names(influ.coefs)<-paste("infl.beta",0:k,sep="")
influ.coefs


##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 20 -3.514149          0.0023196     0.060311
## [1] 1
## $outliers
##    PREU SUPERF st FP DORMIT HABIT    fact
## 20   35   1137  0  0      4     7 11370047
##
## $num.influ.coefs
## (Intercept)      SUPERF          st          FP      DORMIT       HABIT
##           5           2           2           2           3           4
##
## $infl.beta0
##    PREU SUPERF st FP DORMIT HABIT    fact
## 8    70   2261  0  0      3     6 22610036
## 10   82   2104  0  0      4     9 21040049
## 18   64   1226  0  0      4     8 12260048
## 22   43    596  0  0      3     5  5960035
## 24   46    696  0  0      2     4  6960024
##
## $infl.beta1
##    PREU SUPERF st FP DORMIT HABIT    fact
## 8    70   2261  0  0      3     6 22610036
## 10   82   2104  0  0      4     9 21040049
##
## $infl.beta2
##    PREU SUPERF st FP DORMIT HABIT    fact
## 17   62   1126  1  0      3     7 11261037
## 20   35   1137  0  0      4     7 11370047
##
## $infl.beta3
##    PREU SUPERF st FP DORMIT HABIT    fact
## 2    55    815  0  1      2     5 8150125
## 19   66    929  0  1      2     5 9290125
##
## $infl.beta4
##    PREU SUPERF st FP DORMIT HABIT    fact
## 10   82   2104  0  0      4     9 21040049
## 20   35   1137  0  0      4     7 11370047
## 22   43    596  0  0      3     5  5960035
##
## $infl.beta5
##    PREU SUPERF st FP DORMIT HABIT    fact
## 8    70   2261  0  0      3     6 22610036
## 10   82   2104  0  0      4     9 21040049
## 18   64   1226  0  0      4     8 12260048
## 22   43    596  0  0      3     5  5960035
```

Comentari: Els punts influents o atípics s'han d'analitzar perquè poden ser errades.

Si un o més punts són MOLT destacats i no són erronis, convé fer el model amb i sense ells.