Pràctica 9

Regressió múltiple: inferència

Disposem d'una base de dades `cases.txt` amb $n = 26$ casos, en els que es relaciona la resposta $Y =$ *PREU, preu de la vivenda*, amb les $k = 8$ regressores següents:

SUPERF de la vivenda, en m^2	*PL_GAR* $= n^o$ *de places d'aparcament*
HABIT $= n^o$ *d'estances*	*SUPTERR del terreny, en* m^2
DORMIT $= n^o$ *de dormitoris*	*FP* $= n^o$ *de llocs de foc (*$1 =$ *sí,* $0 =$ *no)*
BANYS $= n^o$ *de banys*	*STORM = finestres anti-temporal (sí, no)*

Les variables FP i són qualitatives de tipus dicotòmiques o binàries; en anglès, *dummy*.

```
library(car);library(lmtest);library(lawstat);library(nortest);library(doBy)
dades<-read.table("cases.txt",header=T)
head(dades,4); dim(dades)
```

1. Exploreu la matriu de corelacions i la gràfica `pairs()`, sense les variables qualitatives. Valoreu la possibilitat de multicolinealitat. Feu també gràfiques apropiades per avaluar el possible efecte de les dicotòmiques sobre la resposta.

```
pairs(dades[,-c(4,6)],panel=function(x,y){points(x,y);abline(lm(y~x),lwd=2)}) # traiem qualitatives
print(round(cor(dades[,-c(4,6)]),digits=2)) # traiem qualitatives
attach(dades)
```

```
par(mfrow=c(1,2))
boxplot(PREU~FP);
boxplot(PREU~STORM)
table(FP); table(STORM)   # veiem el nombre de casos
```

2. Feu el model lineal de PREU respecte de les **8** regressores incloses les dicotòmiques i doneu el resultat del test anova de significació del model. *Nota:* `mod1<-lm(PREU~.,data=dades)` fa el model amb totes les variables. Per veure quines conclusions traiem, responeu els apartats següents:
a) Té sentit (és significatiu) el model amb les 8 regressores (diguem-li model complet)? Raoneu la resposta. Escriviu les hipòtesis del test que esteu fent i la hipòtesi que accepteu. Vol dir aixo que el model lineal és la millor opció?
b) Alguna de les variables és susceptible de sortir del model? Quines? Raoneu la resposta.
c) Doneu les estimacions dels coeficients del model complet. Algún coeficient us sembla incoherent? Justifiqueu la resposta.
d) Doneu els errors típics dels coeficients.
e) Doneu l'estimació de la variància del model. Assigneu el seu valor a MSE.

```
mod<-lm(PREU ~ .,data=dades); names(mod)
smod<-summary(mod); smod; names(smod)
coefs<-smod$coefficients[,1]; coefs
e.tips<-smod$coefficients[,2]; e.tips
MSE<-smod$sigma^2; print(paste("MSE= ",MSE))
```

3. Calculem:
a) L'interval de confiança (95%) per al coeficient de la varible DORMIT.
b) L'interval de confiança (95%) per a la variància del model.

c) El valor observat de l'estadístic F del test de significació del model: $fob = \frac{MSR}{MSE} = \frac{SSR/k}{MSE}$. *Nota:* Podeu calcular el numerador usant $SST = (n-1)\text{Var}(y)$ i $SSR = SST - SSE$.
d) El p-valor de fob. Comproveu que els valor de F i el p-valor coincideixen amb els del sumari del model.

```
n<-nrow(mod$model)   # nombre de casos
k<-mod$rank-1            # nombre de variables (quan hi ha intercept)
gl<-n-k-1                # graus de llibertat dels residus
alpha<-.05
(b1<-smod$coefficients[2,1])
(sb1<-smod$coefficients[2,2])  # estim. de la desviació típica del coef. (error típic)
li_b1<-b1-qt(1-alpha/2,gl)*sb1
ls_b1<-b1+qt(1-alpha/2,gl)*sb1
(int.beta1<-list(linf=li_b1,lsup=ls_b1))    # interval beta1

a<-qchisq(alpha/2,n-k-1)
b<-qchisq(1-alpha/2,n-k-1)
li_var<-(n-k-1)*MSE/b
ls_var<-(n-k-1)*MSE/a
(int.var<-list(linf=li_var,lsup=ls_var ))   # interval variància model

SST<-(n-1)*var(PREU)
SSR<-SST-(MSE)*(n-k-1)
MSR<-SSR/k
(fob<-MSR/MSE)                              # valor observat de l'estadistic F
(pvalor<-1-pf(fob,k,n-k-1))                   # p-valor
## una altra manera d'obtenir la mateixa informació
fob<-smod$fstatistic; fob
pval<-1-pf(fob[1],fob[2],fob[3]); print(paste("pvalor= ",pval))
```

4. Feu una funció `betaCI(mod,i,alpha)` per obtenir l'interval de confiança per al coeficient $i = 0, 1, 2, \ldots$. Apliqueu-la al model dels preus, per al coeficients següents: en primer lloc al coeficient β_4 i després al terme independent (*intercept*) β_0.
5. Feu la funció `var.CI(mod,alpha)` per a l'interval de confiança de la variància. Apliqueu-la al model dels preus.
6. Calculeu les estimacions dels paràmetres del model, les prediccions i els residus utilitant les expressions matricials amb les matrius A, H i M respectivament. **Nota:** La variable STORM ha de ser transformada (Sí a 1 i No a 0) abans de crear la matriu de disseny $\mathbf{X}$. [No cal que mostreu les matrius completes, només: les estimacions dels coeficients i les diagonals de les matrius H i M. On intervenen aquestes diagonals?]

```
# transformació de la variable STORM
dades$st[dades$STORM=="no"]<-0
dades$st[dades$STORM=="sí"]<-1
head(dades,3)

##   PREU DORMIT SUPERF FP HABIT STORM BANYS PL_GAR SUPMTERR st
## 1   53      2    967  0     5    no   1.5      0     1006  0
## 2   55      2    815  1     5    no   1.0      2      848  0
## 3   56      3    900  0     5    sí   1.5      1      935  1
```

```
# matriu de disseny  amb "st" en el lloc de "STORM"
n<-nrow(dades)
X<-as.matrix(cbind(rep(1,n),dades[c(2:5,10,7:9)]))  # posem "st" on hi havia STORM
head(X,3)

##   rep(1, n) DORMIT SUPERF FP HABIT st BANYS PL_GAR SUPMTERR
## 1         1      2    967  0     5  0   1.5      0     1006
## 2         1      2    815  1     5  0   1.0      2      848
## 3         1      3    900  0     5  1   1.5      1      935
```

```
list(coef=A%*%Y, diag_H=diag(H), diag_M=diag(M))
```

L'exercici de la pràctica consisteix en mostrar tot el codi que falta i respondre totes les preguntes formulades.