Master MAS, parcours M2 SMSD, Université Claude Bernard Lyon 1
Régressions et Grande Dimension,
Année 2023-2024

Examen du 14 février 2024,
Documents écrits et calculatrice autorisés,
Appareils connectables interdits.
Durée 2h30.

Note: Les trois exercices utilisent des données qui se trouvent dans le packages R emplik et elles ont été traitées avec le logiciel R. Vous trouvez les codes R et les sorties associées sur les feuilles suivantes.
Lire avec attention les consignes spécifiées entre parenthèses pour répondre à certaines questions.
Pour les tests d'hypothèse il faut prendre le risque $\alpha = 0.05$.

Les exercices 1, 2, 3, utilisent les données *myeloma* du package R *emplik*. Plus précisément, une étude de mélanome multiple est considérée pour 65 patients dont 48 sont décédés et 17 sont en vie à la fin de l'étude. La variable (continue) *temps* donne la durée de survie et elle est mesurée en mois. Dans l'expérience, sept autres variables continues sont mesurées: $X1 = AGE$ (âge au diagnostic) et les expressions de six gènes : $X2 = LOGBUN$ (le log du BUN au moment du diagnostic), $X3 = HGB$ (hémoglobine au moment du diagnostic), $X4 = LOGWBC$ (le log du WBC au moment du diagnostic), $X5 = LOGPBM$ (le log du pourcentage de plasmocytes dans la moelle osseuse), $X6 = PROTEIN$ (protéinurie au moment du diagnostic) et $X7 = SCALC$ (calcium sérique au moment du diagnostic).
Trois variables qualitatives sont mesurées:

- La variable Yv=*VSTATUS* prend deux valeurs, 0 et 1, indiquant si le patient est en vie ou décédé, respectivement, à la fin de l'étude.
- *PLATELET* qui concerne les plaquettes au moment du diagnostic et prend deux valeurs : 0=anormal, 1=normal. Cette variable est notée par F1 dans le code R.
- *FRAC* concerne les fractures au moment du diagnostic et prend deux valeurs : 0=aucune, 1=présente. Cette variable est notée par F2 dans le code R.

Exercice 1. (*14 points*)
Remarque: Si parmi les modèles (M1), ... , (M9) certains ont la même forme statistique, il faut le spécifier, sans réécrire la forme statistique.

1) En vous aidant du code R, donnez la forme statistique du modèle (M1). Il s'agit de quel type de modèle? (*0.75 points*)
2) Quels sont les paramètres du modèles (M1) et par quelle méthode ils ont été estimés? (*0.75 points*)
3) Testez si le modèle (M1) est significatif. (spécifiez: les hypothèses H_0, H_1, les modèles correspondants, statistique de test et sa loi sous H_0, valeur de la statistique de test, conclusion). (*1 point*)
4) Si le modèle (M1) est significatif, quelles sont les variables qui influent la variable expliquée? (il faut donner les détails seulement pour une seule variable. Ces détails sont: hypothèses H_0, H_1, modèles correspondants, statistique de test et sa loi sous H_0, valeurs de la statistique, conclusion. Pour les autres variables explicatives, donnez seulement la conclusion.) Donc, quelles sont les variables qu'il faut enlever du modèle (M1)? (*1.25 points*)
5) Donnez les estimations des paramètres du modèle (M1). Interprétez ces estimations. (*0.75 points*)
6) Quelle est la qualité globale d'ajustement du modèle (M1)? Interprétation. (*0.25 points*)
7) Donnez la définition des résidus pour le modèle (M1), plus précisément comment ils sont calculés. Pour le modèle (M1), est-ce que ces résidus sont de loi Normale? Justification pas un test d'hypothèse (écrire les deux hypothèses

H_0 et H_1, la valeur de la statistique de test, la p-value et l'interprétation). A-t-on eu raison de considérer comme méthode d'estimation des coefficients celle que vous avez spécifié à la question 2)? (*1 point*)
8) Donnez la forme statistique du modèle (M2). (*0.25 points*)
9) Commentez les résultats obtenus pour le modèle (M2). Comparez avec le modèle (M3). (*0.5 points*)
10) Donnez la forme statistique du modèle (M3) et du modèle (M4). Par quelle méthode ont été estimés les coefficients du modèle (M3)? (*0.5 points*)
11) Par quelle méthode ont été estimés les coefficients du modèle (M4)? Donnez la forme du processus aléatoire qui a permis d'obtenir les estimations. (*1 point*)
12) Commentez le nuage de points des coefficients estimés par les modèles (M3) et (M4). (*0.5 points*)
13) Quels sont les éléments de l'ensemble $A1$? Interprétation de l'ensemble $A1$. (*0.5 points*)
14) Donnez la forme statistique du modèle (M5). Le modèle (M5) est-il significatif ?(donnez seulement la pvalue, sans autres détails). (*0.5 points*)
15) Par les modèles (M2) et (M5) on modélise la même variable explicative. Comment expliquez-vous la différence des $R2$ ajustés, sachant que (M2) a été obtenu de (M1) et (M5) de (M4)? (*0.5 points*)
16) Donnez la forme statistique du modèle (M6). Par quelle méthode d'estimation les coefficients de ce modèle sont estimés? Ecrivez la forme du processus aléatoire qui a permis d'obtenir ces estimateurs. (*0.75 points*)
17) Pour le modèle (M6), quelles sont les variables qui influent la variable expliquée? (il faut donner les détails seulement pour une seule variable. Ces détails sont: hypothèses H_0, H_1, modèles correspondants, statistique de test et sa loi sous H_0, valeurs de la statistique, conclusion. Pour les autres variables explicatives, donnez seulement la conclusion.) Donc, quelles sont les variables qu'il faut enlever du modèle (M6)? (*1.25 points*)
18) Comment le modèle (M7) a été obtenu? En spécifiant seulement les pvalues, quelles sont les variables explicatives de (M7) qui sont significatives? (*0.5 points*)
19) Par quelle méthode d'estimation les coefficients du modèle (M8) ont été estimés? (*0.75 points*)
20) Donnez la forme statistique du modèle (M9) et la méthode d'estimation des coefficients de ce modèle. (*0.75 points*)

Exercice 2. (*2.75 points*)
1) En vous aidant du code R, donnez la forme statistique du modèle (M10). Il s'agit de quel type de modèle? (*1 point*)
2) Pour le modèle (M10), quelles variables explicatives ont une influence sur la variable expliquée? (donnez les détails suivants pour une seule variable explicative: les hypothèses H_0, H_1, les modèles correspondants, statistique de test et sa loi sous H_0, conclusion. Pour les autres variables explicatives, donnez seulement la conclusion). (*1 point*)
3) Commentez la prévision (notée dans le code R par *prev*) faite par le modèle (M10). (*0.75 points*)

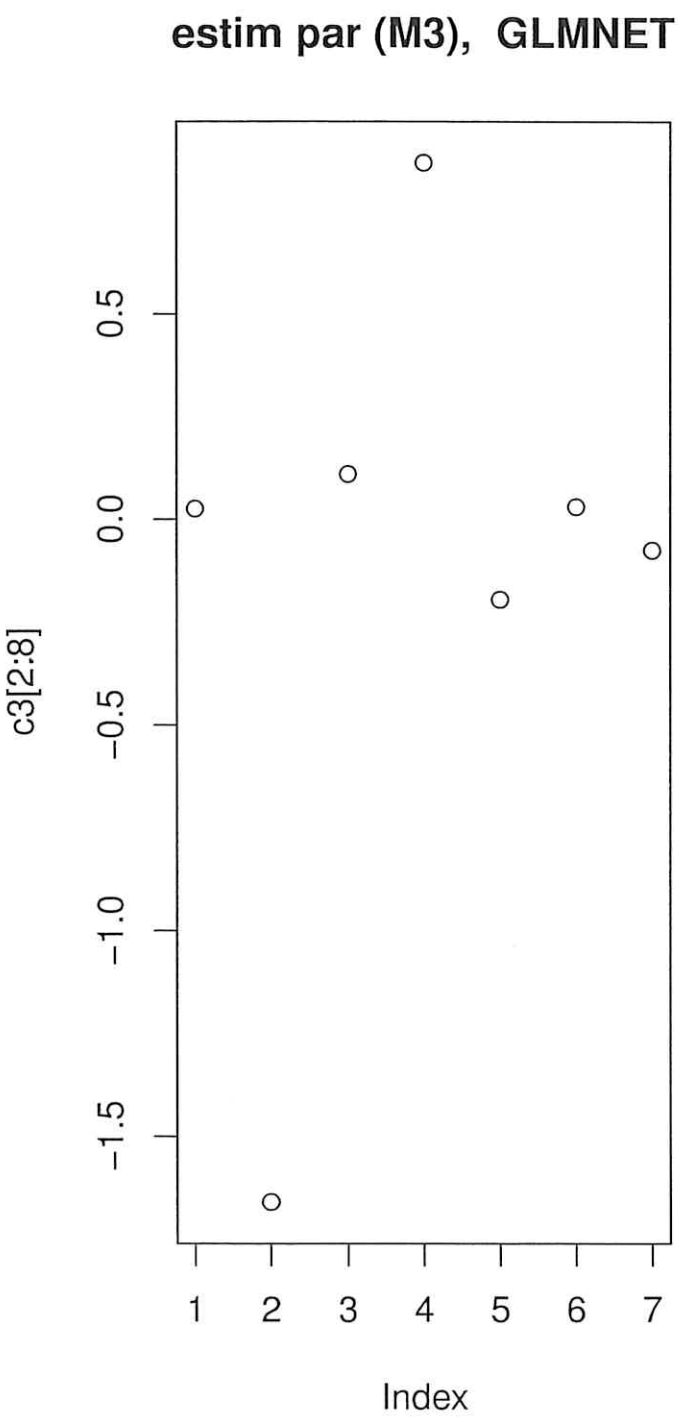
Exercice 3. (*3.25 points*)
1) En vous aidant du code R, donnez la forme statistique du modèle (M11). Il s'agit de quel type de modèle? (*1.5 points*)
2) Testez si le modèle (M11) est significatif. (spécifiez: les hypothèses H_0, H_1, les modèles correspondants, statistique de test et sa loi sous H_0, valeur de la statistique de test, conclusion). (*0.75 points*)
3) Les modèles (M2) et (M11) modélisent la même variable expliquée. Donnez la relation qui indique comment on calcule la prévision de la variable expliquée par ces deux modèles. Lequel de ces deux modèles donne des meilleurs résultats d'ajustement (justification)? (*1 point*)

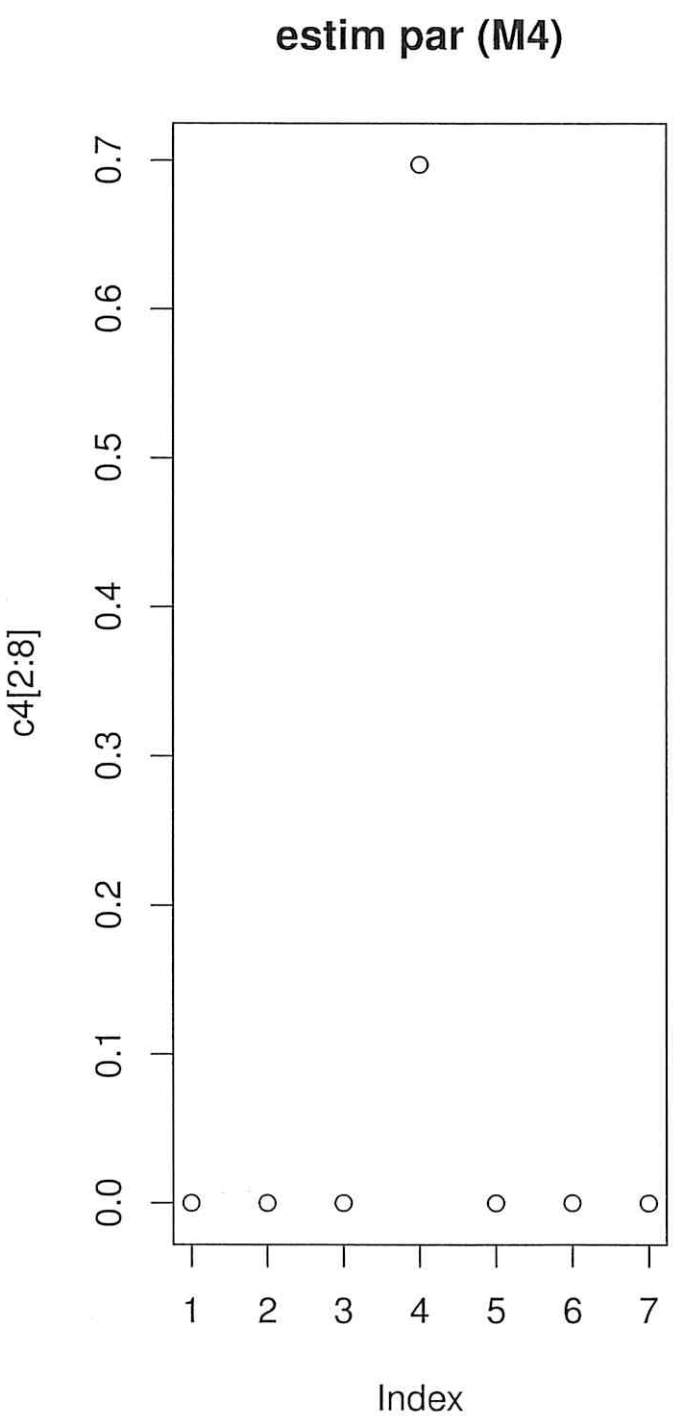
CODE R.

```
library(quantreg)
library(glmnet)
library(car)   # pour la fonction "Anova"
par(mfrow=c(1,2))

###############  EXERCICE 1 #################

cat("Debut  EXERCICE 1 \n \n")

### data myeloma du package emplik
myeloma1=na.omit(myeloma)
nom=c("temps", "vstatus", "logBUN","HGB","platelet","age",
"logWBC","FRAC","logPBM","protein","SCALC")
colnames(myeloma1)=nom

ltime=log(myeloma1[,1])  ## myeloma1[,1] représente la variable "time"="temps"
X1=myeloma1[,6]; X2=myeloma1[,3]; X3=myeloma1[,4]; X4=myeloma1[,7];
X5=myeloma1[,9]; X6=myeloma1[,10]; X7=myeloma1[,11];
F1=myeloma1[,5]; F2=myeloma1[,8];
Yv=myeloma1[,2]

##########
M1=lm(ltime~X1+X2+X3+X4+X5+X6+X7)
summary(M1)
shapiro.test(rstudent(M1))

##########
M2=lm(ltime~X2+X3)
summary(M2)

##########

library(glmnet)
xx=cbind(X1,X2,X3,X4,X5,X6,X7)
M3=glmnet(x=xx,y=ltime,family = "gaussian",intercept = F,lambda = 0)
c3=coef(M3)  # estimations par moindres carrées
plot(c3[2:8], main="estim par (M3),  GLMNET"); # package "glmnet""

n=length(X1)
la=n^{-3/5}  ## lambda
g=2/5;  #  gamma
wj=(1/abs(c3[2:length(c3)]))^g;
lam=la*wj;
M4=glmnet(x=xx,y=ltime,family = "gaussian",intercept = F,penalty.factor =
lam,lambda = 1)
c4=coef(M4)
plot(c4[2:8], main="estim par (M4)")

A1=which(c4[2:length(c4),1]!=0)
cat("\n, Les elements de l ensemble A1 sont: \n")
print(A1)
xm=xx[, c4[2:length(c4),1]!=0];
M5=lm(ltime~xm-1);
summary(M5)

##############
M6=rq(ltime~xx-1, tau=0.5)
c6=coef(M6)
summary(M6,se="iid")
```

```
##############
M7=rq(ltime~xx[,c(1,2)]-1, tau=0.5)
summary(M7,se="iid")

##############
p=ncol(xx);
pp2=2/5;
g1=12.25/10;
hh=vector(mode="numeric",length=p);
for(ii in 1:p)
{
  hh[ii]= abs(c6[ii])**{g1}
}
hh=1/hh;
ll=n^{pp2}*hh;   # les nouveaux lambda
M8=rq(ltime~xx-1,tau=0.5, method="lasso", lambda=ll)
c8=coef(M8)
seuil=0.0001;
c8[abs(c8)<seuil]=0;
A2=which(c8!=0)
cat("Les elements de l ensemble A2 sont: \n")
print(A2)
xn=xx[, c8!=0];

M9=rq(ltime~xn-1, tau=0.5)
summary(M9, se="iid")

###############  EXERCICE 2 #################

##############
M10=glm(Yv ~ xx, family="binomial")
summary(M10);

##############
pi=predict(M10,type = "response")
prev=vector(mode="numeric",length=length(pi))
prev[pi>0.5]=1
cat("Tableau de contingence obtenu par le modèle M10 \n")
table(Yv, prev)

###############  EXERCICE 3 #################
F1=factor(F1)
F2=factor(F2)

##############
M11=lm(ltime~F1+F2+F1:F2,contrasts=list(F1=contr.sum,F2=contr.sum))
shapiro.test((residuals(M11)))
summary(M11)
cat("\n ANOVA DE TYPE III \n ")
print(Anova(M11,type="III"))
```

Graphiques des estimations





Debut EXERCICE 1

SORTIES R

```
Call:
lm(formula = ltime ~ X1 + X2 + X3 + X4 + X5 + X6 + X7)

Residuals:
     Min       1Q   Median       3Q      Max
-1.64777 -0.72182  0.01488  0.66173  2.00782

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.806613   2.370889   2.871 0.005735 **
X1           0.004238   0.013188   0.321 0.749134
X2          -1.487401   0.410106  -3.627 0.000614 ***
X3           0.081899   0.048672   1.683 0.097912 .
X4          -0.441547   0.541054  -0.816 0.417847
X5          -0.361496   0.338106  -1.069 0.289495
X6           0.017743   0.021916   0.810 0.421528
X7          -0.097290   0.069639  -1.397 0.167811
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9343 on 57 degrees of freedom
Multiple R-squared:  0.326,     Adjusted R-squared:  0.2432
F-statistic: 3.939 on 7 and 57 DF,  p-value: 0.001434


        Shapiro-Wilk normality test

data:  rstudent(M1)
W = 0.97843, p-value = 0.3142
```

(M1)

```
Call:
lm(formula = ltime ~ X2 + X3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77546 -0.69326  0.07704  0.64310  1.85169

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.21352    0.73291   5.749 2.94e-07 ***
X2          -1.61799    0.37610  -4.302 6.11e-05 ***
X3           0.07061    0.04597   1.536     0.13
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.939 on 62 degrees of freedom
Multiple R-squared:  0.2595,    Adjusted R-squared:  0.2356
F-statistic: 10.86 on 2 and 62 DF,  p-value: 9.023e-05
```

(M2)

```
, Les elements de l ensemble A1 sont:
X4
 4
```

(M4)

```
Call:
lm(formula = ltime ~ xm - 1)
```

(M5)

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.58474 -0.67709  0.04372  0.86055  1.97266

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
xm  0.70429    0.03693   19.07   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.124 on 64 degrees of freedom
Multiple R-squared:  0.8504,    Adjusted R-squared:  0.848
F-statistic: 363.7 on 1 and 64 DF,  p-value: < 2.2e-16
```

(M5)

```
Call: rq(formula = ltime ~ xx - 1, tau = 0.5)

tau: [1] 0.5

Coefficients:
     Value    Std. Error t value  Pr(>|t|)
xxX1  0.03578  0.01558    2.29715  0.02524
xxX2 -1.43073  0.58331   -2.45279  0.01720
xxX3  0.08516  0.06860    1.24140  0.21946
xxX4  0.46734  0.42091    1.11032  0.27144
xxX5 -0.15393  0.47960   -0.32096  0.74939
xxX6  0.03672  0.03084    1.19063  0.23865
xxX7 -0.00505  0.09939   -0.05082  0.95964
```

(M6)

```
Call: rq(formula = ltime ~ xx[, c(1, 2)] - 1, tau = 0.5)

tau: [1] 0.5

Coefficients:
                 Value    Std. Error t value  Pr(>|t|)
xx[, c(1, 2)]X1  0.05599  0.01271    4.40405  0.00004
xx[, c(1, 2)]X2 -0.53229  0.54361   -0.97917  0.33124
Les elements de l ensemble A2 sont:
xxX4
   4
```

(M7)

(M8)

```
Call: rq(formula = ltime ~ xn - 1, tau = 0.5)

tau: [1] 0.5

Coefficients:
   Value    Std. Error t value  Pr(>|t|)
xn  0.71651  0.03510   20.41456  0.00000
```

(M9)

Exercice 2

```
Call:
glm(formula = Yv ~ xx, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6621  -1.0586   0.4968   0.7714   1.2542

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.73846    6.39433   1.054    0.292
```

(M10)

```
xxX1         -0.02988    0.03491  -0.856    0.392
xxX2         -0.42788    1.18150  -0.362    0.717
xxX3         -0.21756    0.13611  -1.598    0.110
xxX4         -0.66266    1.45834  -0.454    0.650
xxX5         -0.60159    0.97501  -0.617    0.537
xxX6          0.23131    0.15533   1.489    0.136
xxX7          0.19669    0.21348   0.921    0.357

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.706  on 64  degrees of freedom
Residual deviance: 63.672  on 57  degrees of freedom
AIC: 79.672

Number of Fisher Scoring iterations: 6

Tableau de contingence obtenu par le modèle M10
   prev
Yv   0  1
  0  2 15
  1  3 45
```

(M10)

Exercice 3

```
        Shapiro-Wilk normality test

data:  (residuals(M11))
W = 0.97791, p-value = 0.296


Call:
lm(formula = ltime ~ F1 + F2 + F1:F2, contrasts = list(F1 = contr.sum,
    F2 = contr.sum))

Residuals:
     Min       1Q   Median       3Q      Max
-2.56831 -0.71201 -0.01886  0.92212  1.87251

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.55657    0.29918   8.545  5.1e-12 ***
F11         -0.14157    0.29918  -0.473    0.638
F21          0.02832    0.29918   0.095    0.925
F11:F21      0.12163    0.29918   0.407    0.686
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.085 on 61 degrees of freedom
Multiple R-squared:  0.02688,   Adjusted R-squared:  -0.02097
F-statistic: 0.5617 on 3 and 61 DF,  p-value: 0.6423

 ANOVA DE TYPE III
 Anova Table (Type III tests)

Response: ltime
            Sum Sq Df F value    Pr(>F)
(Intercept) 85.997  1 73.0229 5.098e-12 ***
F1           0.264  1  0.2239    0.6378
F2           0.011  1  0.0090    0.9249
F1:F2        0.195  1  0.1653    0.6858
Residuals   71.838 61
---
```

(M11)