UNIVERSITÉ DE LYON | Lyon 1 | département Mathématiques

Master MAS, parcours M2 SMSD, Université Claude Bernard Lyon 1
Régressions et Grande Dimension,
Année 2024-2025

Examen du 19 février 2025,
Documents écrits et calculatrice autorisés,
Appareils connectables interdits.
Durée 2h30.

Note: Les cinq exercices utilisent des données qui se trouvent dans les packages R carData, AER 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, 4 utilisent les données *Soils* du package R *carData*. Les caractéristiques du sol ont été mesurées sur des échantillons provenant de trois types de contours (sommet, pente, dépression) et à quatre profondeurs. La zone a été divisée en 4 blocs, selon un plan de blocs aléatoires.
Les variables mesurées et utilisées dans les exercices 1, 2, 3, 4 sont:

- pH: le pH du sol
- N: la concentration d'azote
- *Dens*: la densité apparente an gm/cm^3
- P: total du phosphore in ppm
- *Ca, Mg, K, Na*: les concentrations de calcium, magnésium, phosphore, sodium
- *Conduc*: la conductivité
- La variable *Group* prend douze valeurs, de 1 à 12.
- La variable *Block* prend quatre valeurs: 1, 2, 3, 4.

Pour les exercices 1, 2 et 4, la variable expliquée sera pH ou son logarithme.
Remarque: Si parmi les modèles (M1), ... , (M7) certains ont la même forme statistique, il faut le spécifier, sans réécrire la forme statistique.

Exercice 1. (*9 points*)
1) En vous aidant du code R, donnez la forme statistique du modèle (M1). Il s'agit de quel type de modèle? (*1 point*)
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). (*1.25 points*)
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)? (*1.5 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). (*1 point*)
8) Donnez la forme statistique du modèle (M2) et du modèle (M3). (*0.25 points*)
9) Par quelle méthode ont été estimés les coefficients du modèle (M3)? Donnez la forme du processus aléatoire qui a permis d'obtenir les estimations. (*1 point*)
10) Commentez le nuage de points des coefficients estimés par les modèles (M2) et (M3). (*0.5 points*)
11) Donnez la forme statistique du modèle (M4). Les résidus standardisés du (M4) sont-ils de loi normale (spécifiez seulement la pvalue et la conclusion)? (*0.75 points*)

Exercice 2. (*4 points*)
1) En vous aidant du code R, donnez la forme statistique du modèle (M5). Il s'agit de quel type de modèle? (*1.25 points*)
2) Testez si le modèle (M5) 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) Donnez les estimations des paramètres du modèle (M5). Interprétation. (*1 point*)
4) Si le modèle (M5) est significatif, quelles sont les variables qui influent la variable expliquée? (*1 point*)

Exercice 3. (*3.5 points*)
Dans cet exercice on calcule une nouvelle variable, notée *ilpH*, qui prend deux valeurs: 0 si $\log(pH) \leq 1.5$ et 1 si $\log(pH) > 1.5$.
1) En vous aidant du code R, donnez la forme statistique du modèle (M6). Il s'agit de quel type de modèle? (*1.25 points*)
2) Pour le modèle (M6), 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). (*1.25 points*)
3) Commentez la prévision (notée dans le code R par *prev6*) faite par le modèle (M6). (*0.5 points*)

Exercice 4. (*2.25 points*)
1) Donnez la forme statistique du modèle (M7). 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. (*1 point*)
2) Pour le modèle (M7), 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 (M7)? (*1.25 points*)

Exercice 5. (*1.25 points*)
Dans cet exercice on utilise les données *"NMES1988"* du package *AER*. Les données concernent une enquête nationale américaine sur des dépenses médicales en 1987 et 1988.
Les variables utilisées sont:

- *visits*: le nombre de visites (consultations) faites par une personne au cabinet médical;
- *age*: l'âge (divisée par 10) de la personne ;
- *chronic*: le nombre de maladies chroniques d'une personne.

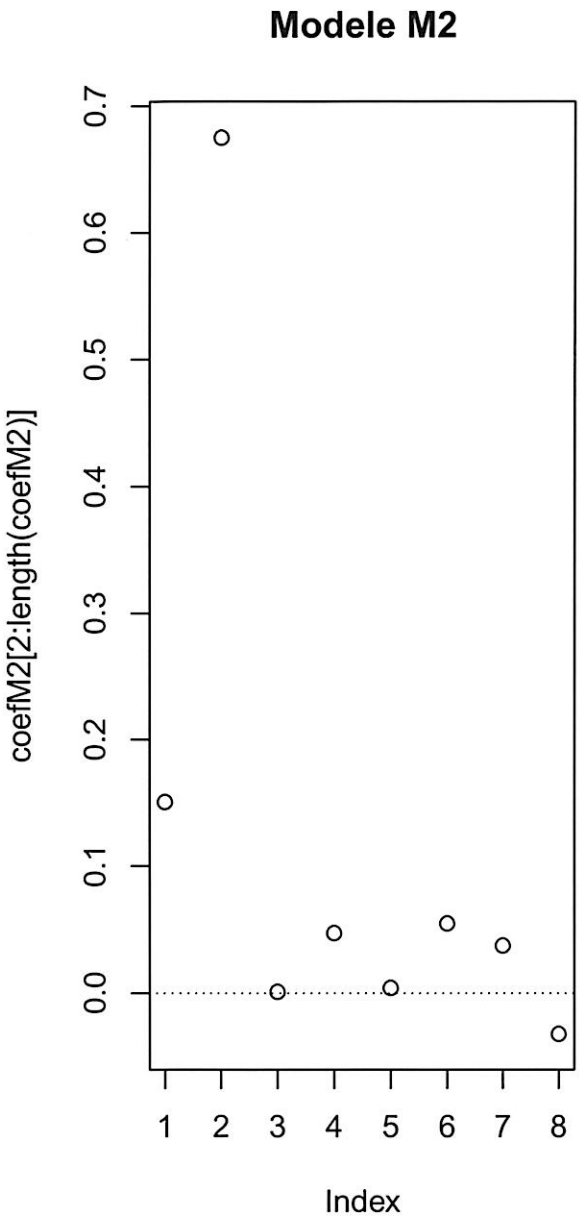
1) Pour le modèle (M8), quelle est la loi de la variable expliquée? (*0.5 points*)
2) Donnez les estimations des paramètres du modèle (M8). Spécifier la valeur du critère AIC. (*0.5 points*)
3) Pour le modèle (M8), quelles variables explicatives ont une influence sur la variable expliquée? (donnez seulement les pvalues et la conclusion). (*0.25 points*)

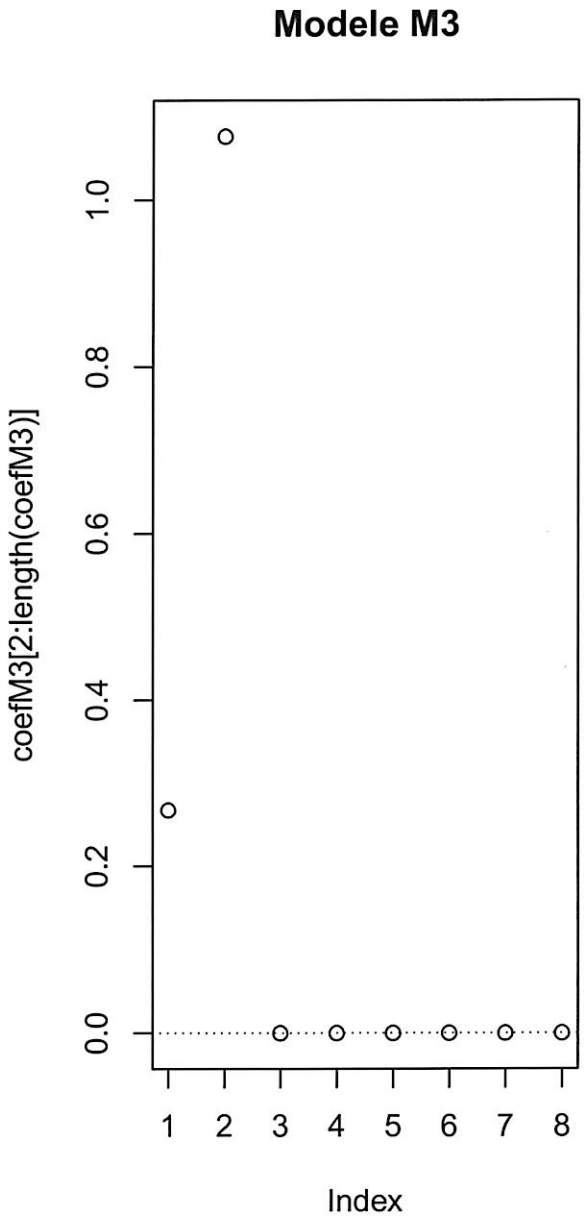
CODE R

```
library(glmnet)
library(AER)
library(quantreg)
library(carData)

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

data("Soils")
attach(Soils)
lpH=log(pH)
M1=lm(lpH~N+Dens+P+Ca+Mg+K+Na+Conduc)
summary(M1)
shapiro.test(residuals(M1))
coefM1=coef(M1)
###################
n=length(lpH)
la=n^{-3/5}
g=2/5;  #  gamma
xx=cbind(N,Dens,P,Ca,Mg,K,Na,Conduc)
M2=glmnet(x=xx,y=lpH,family = "gaussian",intercept = F,lambda = 0)
coefM2=coef(M2)
cat("coef modele M2 \n")
print(coefM2)
plot(coefM2[2:length(coefM2)], main="Modele M2")
abline(h=0, lty=3);
###################
wj=(1/abs(coefM2[2:length(coefM2)]))^g;   # les poids pour LASSO adaptatif
lam=la*wj;
M3=glmnet(x=xx,y=lpH,family = "gaussian",intercept = F,penalty.factor = lam,lambda = 1)
coefM3=coef(M3)
cat("coef modele M3 \n")
print(coefM3)
plot(coefM3[2:length(coefM3)], main="Modele M3")
abline(h=0, lty=3);
###################
xM3=xx[, coefM3[2:length(coefM3)]!=0];  # var X qui ont des estimations differentes de 0
M4=lm(lpH~xM3-1);
summary(M4)
shapiro.test(residuals(M4))

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

Group=factor(Group)
Contour=factor(Contour)
Depth=factor(Depth)
Block=factor(Block)

M5=lm(lpH~Group+Block,
      contrasts=list(Group=contr.sum,Block=contr.sum))
shapiro.test((residuals(M5)))
summary(M5)
cat("\n ANOVA DE TYPE III \n ")
print(Anova(M5,type="III"))

###############  EXERCICE 3 #################

ilpH=vector(mode="numeric",length = n)
ilpH[lpH>1.5]=1
M6=glm(ilpH ~ Ca+Na+Conduc+Block, family="binomial")
summary(M6);
pi6=predict(M6,type = "response")
prev6=vector(mode="numeric",length=length(pi6))
prev6[pi6>0.5]=1   ### pr?vision de "remiss" par le mod?le "m2"
cat("Tableau de contingence obtenu par le mod?le M6 \n")
```

```
table(ilpH, prev6)

###############  EXERCICE 4 #################

M7=rq(pH~xx, tau=0.5)
summary(M7,se="iid")

###############  EXERCICE 5 #################

data("NMES1988")
attach(NMES1988)
M8 = glm(visits ~chronic+age, data = NMES1988, family = poisson)
summary(M8)
```

Graphiques des estimations





EXERCICE 1

SORTIES R

```
Call:
lm(formula = lpH ~ N + Dens + P + Ca + Mg + K + Na + Conduc)

Residuals:
      Min        1Q    Median        3Q       Max
-0.112383 -0.047214 -0.008883  0.043852  0.206348

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.3115474  0.2153750   6.090 3.89e-07 ***
N           -0.4227133  0.4709317  -0.898 0.374901
Dens         0.1502142  0.1065492   1.410 0.166521
P            0.0003225  0.0002768   1.165 0.251014
Ca           0.0299127  0.0079005   3.786 0.000516 ***
Mg          -0.0108828  0.0099520  -1.094 0.280872
K           -0.0853177  0.0797989  -1.069 0.291572
Na           0.0350157  0.0167175   2.095 0.042756 *
Conduc      -0.0444470  0.0149712  -2.969 0.005091 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.07375 on 39 degrees of freedom
Multiple R-squared:  0.7716,    Adjusted R-squared:  0.7247
F-statistic: 16.47 on 8 and 39 DF,  p-value: 2.477e-10


        Shapiro-Wilk normality test

data:  residuals(M1)
W = 0.96418, p-value = 0.1489
```

(M1)

```
coef modele M2
9 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  .
N            0.150625395
Dens         0.675272167
P            0.001099349
Ca           0.047493381
Mg           0.004249844
K            0.055027391
Na           0.037630623
Conduc      -0.032213110
```

(M2)

```
coef modele M3
9 x 1 sparse Matrix of class "dgCMatrix"
                    s0
(Intercept) .
N           0.2673168
Dens        1.0767881
P           .
Ca          .
Mg          .
K           .
Na          .
Conduc      .
```

(M3)

```
Call:
lm(formula = lpH ~ xM3 - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29588 -0.09155 -0.00638  0.12448  0.35926
```

(M4)

```
Coefficients:
        Estimate Std. Error t value Pr(>|t|)
xM3N     3.94113    0.27101   14.54   <2e-16 ***
xM3Dens  0.85299    0.02473   34.50   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1512 on 46 degrees of freedom
Multiple R-squared:  0.9907,    Adjusted R-squared:  0.9903
F-statistic:  2457 on 2 and 46 DF,  p-value: < 2.2e-16


        Shapiro-Wilk normality test

data:  residuals(M4)
W = 0.98524, p-value = 0.8009
```

(M4)

EXERCICE 2.

```
        Shapiro-Wilk normality test

data:  (residuals(M5))
W = 0.9565, p-value = 0.07289


Call:
lm(formula = lpH ~ Group + Block, contrasts = list(Group = contr.sum,
    Block = contr.sum))

Residuals:
      Min        1Q    Median        3Q       Max
-0.126323 -0.033000  0.007479  0.029166  0.200206

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.53121    0.01023 149.719  < 2e-16 ***
Group1       0.14183    0.03392   4.181 0.000201 ***
Group2       0.04604    0.03392   1.357 0.183922
Group3      -0.09611    0.03392  -2.833 0.007795 **
Group4      -0.17261    0.03392  -5.089 1.42e-05 ***
Group5       0.17370    0.03392   5.121 1.29e-05 ***
Group6       0.12198    0.03392   3.596 0.001042 **
Group7      -0.08476    0.03392  -2.499 0.017619 *
Group8      -0.16342    0.03392  -4.818 3.15e-05 ***
Group9       0.14575    0.03392   4.297 0.000144 ***
Group10      0.05176    0.03392   1.526 0.136516
Group11     -0.06007    0.03392  -1.771 0.085816 .
Block1      -0.04131    0.01771  -2.332 0.025965 *
Block2       0.01279    0.01771   0.722 0.475336
Block3      -0.01571    0.01771  -0.887 0.381687
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.07086 on 33 degrees of freedom
Multiple R-squared:  0.8216,    Adjusted R-squared:  0.7459
F-statistic: 10.86 on 14 and 33 DF,  p-value: 1.281e-08
```

(M5)

ANOVA DE TYPE III
Anova Table (Type III tests)

Response: lpH

	Sum Sq	Df	F value	Pr(>F)	
(Intercept)	112.542	1	22415.7546	< 2.2e-16	***
Group	0.714	11	12.9337	5.125e-09	***
Block	0.049	3	3.2441	0.03427	*
Residuals	0.166	33			

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(M5)

EXERCICE 3

Call:
glm(formula = ilpH ~ Ca + Na + Conduc + Block, family = "binomial")

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.71394	-0.33263	0.04593	0.17768	2.46528

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0127	5.8394	-0.516	0.6059	
Ca	0.5189	0.4758	1.090	0.2755	
Na	1.4919	0.9948	1.500	0.1337	
Conduc	-1.8519	0.9726	-1.904	0.0569	.
Block2	4.3650	2.2552	1.936	0.0529	.
Block3	4.2860	2.4408	1.756	0.0791	.
Block4	3.2055	2.0618	1.555	0.1200	

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 66.459 on 47 degrees of freedom
Residual deviance: 21.250 on 41 degrees of freedom
AIC: 35.25

Number of Fisher Scoring iterations: 7

Tableau de contingence obtenu par le mod?le M6

ilpH \ prev6	0	1
0	21	2
1	3	22

(M6)

EXERCICE 4

Call: rq(formula = pH ~ xx, tau = 0.5)

tau: [1] 0.5

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	5.43716	0.78078	6.96373	0.00000
xxN	-4.32324	1.70723	-2.53231	0.01547
xxDens	-0.09391	0.38626	-0.24313	0.80918
xxP	0.00268	0.00100	2.66775	0.01106
xxCa	0.10366	0.02864	3.61924	0.00084
xxMg	-0.10682	0.03608	-2.96075	0.00520
xxK	-0.32616	0.28929	-1.12745	0.26644
xxNa	0.10109	0.06060	1.66805	0.10332
xxConduc	-0.16310	0.05427	-3.00519	0.00462

(M7)

EXERCICE 5

```
Call:
glm(formula = visits ~ chronic + age, family = poisson, data = NMES1988)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.4097  -2.0997  -0.7320   0.8036  17.5450

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.699101   0.074041  22.948  < 2e-16 ***
chronic      0.199526   0.004083  48.870  < 2e-16 ***
age         -0.039542   0.010013  -3.949 7.85e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 26943  on 4405  degrees of freedom
Residual deviance: 24752  on 4403  degrees of freedom
AIC: 37533

Number of Fisher Scoring iterations: 5
```

(M8)