Master M2, SITN et Data Science, Université Claude Bernard, Lyon 1
Modèles de régression
année 2019-2020

Examen du 29 janvier 2020,
Calculatrice autorisée
Appareils connectables interdits.
Durée 1h30.

Note: Les trois exercices ont été traités avec le logiciel R. L'exercice 2 a été traité également avec le logiciel SAS. Vous trouvez les codes et les sorties associées sur les feuilles suivantes. Pour les tests d'hypothèse, il faut prendre le risque $\alpha = 0.05$.

Exercice 1.

Les données pour cet exercice proviennent du package MASS du logiciel R. Plus précisément, le tableau de données *coop* contient des expériences sur des essais en chimie analytique. Sept échantillons ont été envoyés à 6 laboratoires en 3 lots distincts. Chaque analyse a été dupliquée. Les variables mesurées sont:
Conc: la concentration du produit chimique étudié
Lab: le laboratoire, valeurs prises: L1, L2, L3, L4, L5, L6
Spc: l'échantillon, valeurs: S1, S2, S3, S4, S5, S6, S7
Bat: le lot, valeurs: B1, B2, B3.

1) En vous aidant du code R, donnez la forme statistique du modèle (M1). Il s'agit de quel type de modèle?
2) 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)
3) Si le modèle (M1) est significatif, quelles sont les variables qui influent la concentration du produit chimique étudié? (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?
4) Donnez les estimations des paramètres du modèle (M1). Interprétez ces estimations.
5) Donnez la qualité globale du modèle (M1). Interprétation.

Exercice 2

Les données pour cet exercice proviennent du package MASS du logiciel R, le tableau *biopsy* contenant des données sur le cancer du sein suite à une biopsie pour 699 patientes.
Les variables mesurées sont:
class: prend deux valeurs: 1 pour cancer bénin et 2 pour cancer malin;
V1: épaisseur de la tumeur;
V2: uniformité de la taille des cellules;
V3: uniformité de la forme des cellules.

Pour cet exercice les logiciels R et SAS ont été utilisés. Pour répondre aux questions ils faudrait utiliser les sorties associées à ces deux logiciels.

1) Pour le modèle (M2), écrivez le modèle statistique correspondant. Il s'agit de quel type de modèle?
2) Testez si le modèle (M2) est significatif. (spécifiez: les hypothèses H_0, H_1, les modèles correspondants, statistiques de test utilisées, conclusion).
3) Pour le modèle (M2), quelles variables ont une influence sur la probabilité que le cancer soit bénin? (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).
4) Donnez les estimations de tous paramètres du modèle (M2).
5) En considérant le modèle (M2), quelle est la probabilité que le cancer soit malin si on a mesuré les valeurs suivantes pour les variables explicatives: $V1 = 3$, $V2 = 1$, $V3 = 1$? (vous donnez seulement l'expression de la probabilité, sans faire les calculs)
6) Considérons deux patientes qui ont les mêmes valeurs pour V1 et V2. En échange, la première patiente a $V3 = 2$ et la deuxième patiente a $V3 = 1$. Pour laquelle des deux patientes, la probabilité d'avoir un cancer malin est plus grande? Justifiez votre réponse.
7) Réalisez une comparaison entre les vraies valeurs de la variable *class* et les valeurs prévues pour cette variable par le modèle (M2). Commentez la qualité du modèle (M2).

Exercice 3

On utilise le même tableau de données de l'Exercice 2, complété avec plus les variables V4, ... V9, qui sont toutes numériques. Par les modèles (M3) et (M4), la variable V1 est modélisée fonction de V2, V3, ...V9.
1) Ecrivez la forme statistique du modèle (M3), qui est le même que le (M4). Il s'agit de quel type de modèle?
2) Quels sont les paramètres de ces modèles? Quelles sont les méthodes d'estimation des paramètres utilisées en (M3) et (M4)?
3) En se basant sur les sorties de la fonction "lm" et des deux graphiques obtenus pour les estimations des paramètres des deux modèles, comparez les estimations.

LE CODE DU LOGICIEL R POUR LES EXERCICES 1, 2, 3

```
library(MASS)
library(car);
###############  EXERCICE 1   ######################
data(coop)
attach(coop)
Lab=factor(Lab); Spc=factor(Spc); Bat=factor(Bat);
m1=lm(Conc~Lab+Spc+Bat+Lab:Bat, contrasts = list(Bat=contr.sum, Spc=contr.sum,
Lab=contr.sum))
cat("\n SORTIES EXERCICE 1 \n")
print(Anova(m1,type="III"))
print(summary(m1))

###############  EXERCICE 2   ######################
data("biopsy")
attach(biopsy)
m2=glm(class ~ V1+V2+V3, family="binomial")
cat("\n SORTIES EXERCICE 2 \n")
summary(m2);

###############  EXERCICE 3   ######################
library(lqa)
pp=9/20; # c est la puissance de lambda
g=2/5;
n=nrow(biopsy)
la=n^{pp};
m3=lm(V1~V2+V3+V4+V5+V6+V7+V8+V9-1)  #  modele classique
cat("\n SORTIES EXERCICE 3 \n")
summary(m3)
c3=coef(m3)  # estimations par moindres carrées
plot(c3)

wj=(1/abs(c3))^g;   # les poids pour LASSO adaptatif
m4=lqa(V1~V2+V3+V4+V5+V6+V7+V8+V9-1,penalty=adaptive.lasso(lambda=la,
        al.weights=wj),standardize=TRUE)
c4=m4$coef;  #  estimations par LASSO adaptatif
plot(c4)

par(mfrow=c(1,2))
plot(c3, xlab="numero Variable",main="Estimations par moindres carrées" );
plot(c4, xlab="numero Variable",main="Estimations par LASSO adaptatif")
```

LE CODE SAS POUR L'EXERCICE 2

```
data exo2;
infile "exo2.txt" firstobs=2;;
input class V1 V2 V3;  run;

proc logistic data=exo2 descending outest=tab1 covout;
    model class=V1 V2 V3;
      output out=outlog p=prev predprob=(individual crossvalidate); run;
run;
 proc print data=outlog; run;
proc print data=tab1; run;
proc freq data=outlog;
  table _FROM_*_INTO_; run;
```

SORTIES SAS pour l'EXERCICE 2

La procédure LOGISTIC

Informations sur le modèle	
Table	WORK.EXO2
Variable de réponse	class
Nombre de niveaux de réponse	2
Modèle	logit binaire
Technique d'optimisation	Score de Fisher

Nb d'observations lues	699
Nb d'observations utilisées	699

Profil de réponse		
Valeur ordonnée	class	Fréquence totale
1	2	241
2	1	458

La probabilité modélisée est class=2.

Etat de convergence du modèle
Critère de convergence (GCONV=1E-8) respecté.

Statistique d'ajustement du modèle		
Critère	Constante uniquement	Constante et Covariables
AIC	902.527	197.997
SC	907.077	216.195
-2 Log L	900.527	189.997

Test de l'hypothèse nulle globale : BETA=0			
Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	710.5309	3	<.0001
Score	522.4929	3	<.0001
Wald	134.5168	3	<.0001

La procédure LOGISTIC

Analyse des valeurs estimées du maximum de vraisemblance					
Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept	1	-7.5916	0.6672	129.4460	<.0001
V1	1	0.5975	0.0988	36.5478	<.0001
V2	1	0.5535	0.1483	13.9320	0.0002
V3	1	0.7177	0.1542	21.6473	<.0001

Estimation du rapport de cotes			
Effet	Estimation du point	Intervalle de confiance de Wald à 95%	
V1	1.817	1.497	2.206
V2	1.739	1.301	2.326
V3	2.050	1.515	2.773

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	98.8	D de Somers	0.977
Pourcentage discordant	1.1	Gamma	0.977
Pourcentage lié	0.1	Tau-a	0.442
Paires	110378	c	0.988

La procédure FREQ

Fréquence
Pourcentage
Pct de ligne
Pct de col.

Table de _FROM_ par _INTO_			
FROM(Valeur formatée de la réponse observée)	_INTO_(Valeur formatée de la réponse prédite)		
	1	2	Total
1	446 63.81 97.38 95.50	12 1.72 2.62 5.17	458 65.52
2	21 3.00 8.71 4.50	220 31.47 91.29 94.83	241 34.48
Total	467 66.81	232 33.19	699 100.00

SORTIES R : pour les Exercices 1, 2, 3

SORTIES EXERCICE 1 : Modèle (M1)

Anova Table (Type III tests)

Response: Conc

	Sum Sq	Df	F value	Pr(>F)	
(Intercept)	930.43	1	7685.0867	< 2.2e-16	***
Lab	18.59	5	30.7140	< 2.2e-16	***
Spc	1485.52	6	2044.9885	< 2.2e-16	***
Bat	0.41	2	1.7114	0.182919	
Lab:Bat	3.18	10	2.6226	0.004884	**
Residuals	27.60	228			

Signif. codes: 0 0.001 0.01 0.05 0.1 1

```
Call:
lm(formula = Conc ~ Lab + Spc + Bat + Lab:Bat, contrasts =
list(Bat = contr.sum,
   Spc = contr.sum, Lab = contr.sum))
```

Residuals:

Min	1Q	Median	3Q	Max
-1.04631	-0.16175	0.00417	0.13053	1.85226

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.92151	0.02192	87.665	< 2e-16	***
Lab1	-0.32794	0.04901	-6.691	1.70e-10	***
Lab2	0.04564	0.04901	0.931	0.3528	
Lab3	-0.32722	0.04901	-6.676	1.84e-10	***
Lab4	0.45230	0.04901	9.228	< 2e-16	***
Lab5	0.01135	0.04901	0.232	0.8171	
Spc1	-1.41345	0.05369	-26.326	< 2e-16	***
Spc2	-1.55567	0.05369	-28.975	< 2e-16	***
Spc3	-0.84456	0.05369	-15.730	< 2e-16	***
Spc4	-1.27956	0.05369	-23.832	< 2e-16	***
Spc5	5.83988	0.05369	108.770	< 2e-16	***
Spc6	-0.13567	0.05369	-2.527	0.0122	*
Bat1	-0.03068	0.03100	-0.990	0.3234	
Bat2	0.05730	0.03100	1.849	0.0658	.
Lab1:Bat1	-0.02218	0.06931	-0.320	0.7492	
Lab2:Bat1	0.12282	0.06931	1.772	0.0777	.
Lab3:Bat1	0.02853	0.06931	0.412	0.6810	
Lab4:Bat1	0.12829	0.06931	1.851	0.0655	.
Lab5:Bat1	0.00496	0.06931	0.072	0.9430	
Lab1:Bat2	0.02484	0.06931	0.358	0.7204	
Lab2:Bat2	-0.16087	0.06931	-2.321	0.0212	*
Lab3:Bat2	0.03270	0.06931	0.472	0.6376	
Lab4:Bat2	0.07603	0.06931	1.097	0.2738	
Lab5:Bat2	-0.05587	0.06931	-0.806	0.4210	

Signif. codes: 0 0.001 0.01 0.05 0.1 1

Residual standard error: 0.348 on 228 degrees of freedom
Multiple R-squared: 0.982, Adjusted R-squared: 0.9802
F-statistic: 541.4 on 23 and 228 DF, p-value: < 2.2e-16

SORTIES EXERCICE 2 : Modèle (M2)

```
Call:
glm(formula = class ~ V1 + V2 + V3, family = "binomial")
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.5115	-0.1972	-0.0808	0.0257	2.8107

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-7.59180	0.66727	-11.377	< 2e-16	***
V1	0.59749	0.09883	6.046	1.49e-09	***
V2	0.55355	0.14830	3.733	0.00019	***
V3	0.71769	0.15425	4.653	3.28e-06	***

Signif. codes: 0 0.001 0.01 0.05 0.1 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 900.53 on 698 degrees of freedom
Residual deviance: 190.00 on 695 degrees of freedom
AIC: 198

Number of Fisher Scoring iterations: 7

SORTIES EXERCICE 3

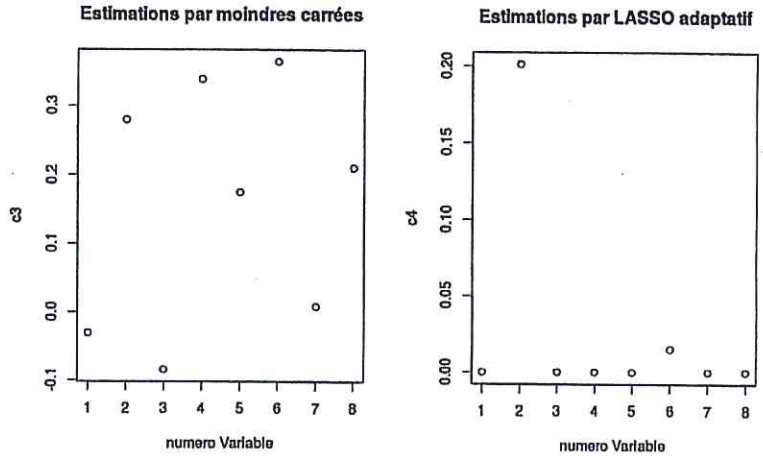
modèle (M3)

```
Call:
lm(formula = V1 ~ V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9551 -0.9673  0.6678  2.1440  7.0613

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
V2 -0.03019    0.07562  -0.399 0.689862
V3  0.28028    0.07534   3.720 0.000216 ***
V4 -0.08398    0.04862  -1.727 0.084614 .
V5  0.33877    0.05794   5.847 7.81e-09 ***
V6  0.17508    0.03845   4.554 6.26e-06 ***
V7  0.36486    0.05594   6.523 1.36e-10 ***
V8  0.00811    0.04517   0.180 0.857573
V9  0.21077    0.05970   3.530 0.000443 ***
---
Signif. codes:  0  0.001  0.01  0.05  0.1  1

Residual standard error: 2.322 on 675 degrees of freedom
  (16 observations deleted due to missingness)
Multiple R-squared:  0.8075,  Adjusted R-squared:  0.8052
F-statistic: 353.9 on 8 and 675 DF,  p-value: < 2.2e-16
```



(M3) (M4)

Graphique coefficients estimés par (M3) et (M4)

LE CODE DU LOGICIEL R POUR LES EXERCICES 1, 2, 3

```
library(MASS)
library(car);
###############  EXERCICE 1   ######################
data(coop)
attach(coop)
Lab=factor(Lab); Spc=factor(Spc); Bat=factor(Bat);
m1=lm(Conc~Lab+Spc+Bat+Lab:Bat, contrasts = list(Bat=contr.sum, Spc=contr.sum,
Lab=contr.sum))
cat("\n SORTIES EXERCICE 1 \n")
print(Anova(m1,type="III"))
print(summary(m1))

###############  EXERCICE 2   ######################
data("biopsy")
attach(biopsy)
m2=glm(class ~ V1+V2+V3, family="binomial")
cat("\n SORTIES EXERCICE 2 \n")
summary(m2);

###############  EXERCICE 3   ######################
library(lqa)
pp=9/20; # c est la puissance de lambda
g=2/5;
n=nrow(biopsy)
la=n^{pp};
m3=lm(V1~V2+V3+V4+V5+V6+V7+V8+V9-1)  #  modele classique
cat("\n SORTIES EXERCICE 3 \n")
summary(m3)
c3=coef(m3)  # estimations par moindres carrées
plot(c3)

wj=(1/abs(c3))^g;   # les poids pour LASSO adaptatif
m4=lqa(V1~V2+V3+V4+V5+V6+V7+V8+V9-1,penalty=adaptive.lasso(lambda=la,
        al.weights=wj),standardize=TRUE)
c4=m4$coef;  #  estimations par LASSO adaptatif
plot(c4)

par(mfrow=c(1,2))
plot(c3, xlab="numero Variable",main="Estimations par moindres carrées" );
plot(c4, xlab="numero Variable",main="Estimations par LASSO adaptatif")
```

LE CODE SAS POUR L'EXERCICE 2

```
data exo2;
infile "exo2.txt" firstobs=2;;
input class V1 V2 V3;  run;

proc logistic data=exo2 descending outest=tab1 covout;
    model class=V1 V2 V3;
      output out=outlog p=prev predprob=(individual crossvalidate); run;
run;
 proc print data=outlog; run;
proc print data=tab1; run;
proc freq data=outlog;
  table _FROM_*_INTO_; run;
```