Master M2, SITN et Data Science, Université Claude Bernard, Lyon1
Modèles de régression
année 2017-2018

Examen du 15 janvier 2018,
Documents admis. Calculatrice autorisée
Appareil connectables interdits.
Durée 1h.

Avant de commencer à rédiger vos réponses, lisez avec attention ces consignes: *Les deux exercices ont été traités avec les logiciels R et SAS. Vous trouvez les codes et les sorties associées sur les feuilles suivantes. Sauf mention spéciale, vous pouvez utiliser le code et les sorties qui vous conviennent: soit ceux de R, soit ceux de SAS, soit les deux.*
S'il faut faire des tests d'hypothèse, ils sont à faire pour un seuil $\alpha = 0.05$.

Exercice 1.
Les données pour cet exercice proviennent de l'adresse internet:
http://stat.ethz.ch/Teaching/Datasets/airpullutionfiltersdat.html
pour étudier le bruit, en décibels, émis par des véhicules (*noise*) fonction de la dimension de la voiture (*size*) et le type de silencieux (*type*) avec lequel chaque voiture est équipée.
Les variables considérées sont les suivantes:
noise: le niveau du bruit émis (en décibels)
size: dimension du véhicule, avec les valeurs: 1 petite, 2 moyenne, 3 grande
type: prend deux valeurs: 1 pour silencieux standard, 2 pour filtre Octel

1) 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, conclusion)
3) Si le modèle (M1) est significatif, quelles sont les variables qui influent le niveau du bruit fait par un véhicule? (spécifiez pour une variable explicative: les hypothèses H_0, H_1, les modèles correspondants, statistique de test et sa loi sous H_0, conclusion. Pour l'autre variable explicative, donnez seulement la conclusion)
4) Quel type de silencieux permet qu'un véhicule soit moins bruyant? Justification.
5) Classez, selon leur taille, les véhicules du moins bruyant au plus bruyant. (Sorties du logiciel SAS conseillées).
6) Donnez la qualité globale du modèle (M1).

Exercice 2.
L'objectif de cette étude est de dépister si un bébé est atteint de la Dystrophie musculaire de Duchenne (DMD), en mesurant certains marqueurs sériques. Une autre question intéressante est de savoir si l'âge de la mère devrait être prise en compte.
Les variables suivantes ont été considérées pour la mère:
age: l'âge de la mère;

ck: concentration de la Creatine Kinase ;
h: concentration de la hémopexine;
pk: concentration de la pyruvate kinase;
ld: concentration de la lactate déshydrogénase.
carrier: variable qualitative pour indiquer si bébé est porteur de DMD. Les valeurs de la variable carrier sont: 0: non porteur, 1: porteur.

On veut modéliser la probabilité qu'un bébé soit atteint de la DMD, en utilisant le modèle (M2).
1) En utilisant les sorties du logiciel SAS, donnez: le nombre d'observations du fichier *dmd.txt*, le nombre d'observations utilisées pour le modèle (M2). Quel est le nombre de bébés atteints de la DMD?
2) Ecrivez le modèle statistique (M2). Il s'agit de quel type de modèle?
3) Testez si le modèle (M2) est significatif. (spécifiez: les hypothèses H_0, H_1, les modèles correspondants, conclusion)
4) Pour le modèle (M2), quelles variables ont une influence sur la probabilité qu'un bébé soit atteint de la DMD. (spécifiez pour une 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)
5) Quelles sont les variables qui augmentent la probabilité qu'un bébé soit atteint de la DMD? Justification.
6) Quelles sont les variables qui diminuent la probabilité qu'un bébé soit atteint de la DMD? Justification.
7) Quel est le taux d'erreur de prévision par le modèle (M2) de la présence de la DMD? (sorties du logiciel SAS à utiliser)

Code et Sorties du Logiciel R


```
###############      EXERCICE 1

exo1=read.table('air_pollution.txt',col.names=c("noise", "size", "type", "side"));
attach(exo1)
size=factor(size);
type=factor(type);
m1=lm(noise~size+type,contrasts=list(size=contr.sum, type=contr.sum))
print(summary(m1))
cat("\n ANOVA DE TYPE III \n ")
print(Anova(m1,type="III"))
```

} (M1)

```
###############      EXERCICE 2

exo2=read.table('dmd.txt',col.names=c("ind", "hospid", "age", "sdate", "ck", "h", "pk", "ld",
"carrier", "obsno"),na.strings=".");
attach(exo2)
m2=glm(carrier ~ age+ck+h+pk+ld, family="binomial")
summary(m2);
```

} (M2)

```
###############      SORTIES  ####################################

############## SORTIES POUR L'EXERCICE 1
Call:
lm(formula = noise ~ size + type, contrasts = list(size = contr.sum,
    type = contr.sum))

Residuals:
    Min      1Q  Median      3Q     Max
-19.583  -7.292   1.250   6.250  15.833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  810.139      1.550 522.766  < 2e-16 ***
size1         14.028      2.192   6.401 3.42e-07 ***
size2         23.611      2.192  10.773 3.55e-12 ***
type1          5.417      1.550   3.495  0.00141 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.298 on 32 degrees of freedom
Multiple R-squared:  0.9074,    Adjusted R-squared:  0.8987
F-statistic: 104.5 on 3 and 32 DF,  p-value: < 2.2e-16


 ANOVA DE TYPE III
 Anova Table (Type III tests)

Response: noise
             Sum Sq Df    F value    Pr(>F)
(Intercept) 23627701  1 273284.249 < 2.2e-16 ***
size           26051  2    150.659 < 2.2e-16 ***
type            1056  1     12.217  0.001411 **
Residuals       2767 32
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
############## SORTIES POUR L'EXERCICE 2

Call:
glm(formula = carrier ~ age + ck + h + pk + ld, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8992  -0.3800  -0.1668   0.0211   2.3543

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -19.548879   3.305677  -5.914 3.34e-09 ***
age           0.147786   0.044747   3.303 0.000958 ***
ck            0.046698   0.014843   3.146 0.001655 **
h             0.082587   0.027238   3.032 0.002429 **
pk            0.116341   0.048199   2.414 0.015788 *
ld            0.012505   0.005743   2.177 0.029445 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 250.077  on 193  degrees of freedom
Residual deviance:  84.038  on 188  degrees of freedom
  (15 observations deleted due to missingness)
AIC: 96.038

Number of Fisher Scoring iterations: 8
```

```
/* Exercice 1 */


data exo1;
 infile 'air_pollution.txt' expandtabs;
 input noise size type side; run;

 proc glm data=exo1 ;
      class size type;
      model noise =size type  / solution ;
       run;


  /* ********************************************************************** */
/* Exercice 2 */

data exo2;
infile "dmd.txt" ;
 input ind   hospid   age   sdate   ck   h   pk   ld   carrier   obsno;
 run;
 proc logistic data=exo2 descending  covout;
    model carrier=age ck h pk ld;
        output out=dmd1 p=prev predprob=(individual crossvalidate); run;
run;
proc freq data=dmd1;
  table _FROM_*_INTO_; run;
```

(M1)

(M2)

Exercice 1

Le Système SAS 15:41 Wednesday, January 10, 2018 1

The GLM Procedure

Informations sur les niveaux de classe		
Classe	Niveaux	Valeurs
size	3	1 2 3
type	2	1 2

Number of Observations Read	36
Number of Observations Used	36

The GLM Procedure

Dependent Variable: noise

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	3	27107.63889	9035.87963	104.51	<.0001
Error	32	2766.66667	86.45833		
Corrected Total	35	29874.30556			

R-carré	Coef de Var	Racine MSE	noise Moyenne
0.907390	1.147741	9.298297	810.1389

Source	DDL	Type I SS	Carré moyen	Valeur F	Pr > F
size	2	26051.38889	13025.69444	150.66	<.0001
type	1	1056.25000	1056.25000	12.22	0.0014

Source	DDL	Type III SS	Carré moyen	Valeur F	Pr > F
size	2	26051.38889	13025.69444	150.66	<.0001
type	1	1056.25000	1056.25000	12.22	0.0014

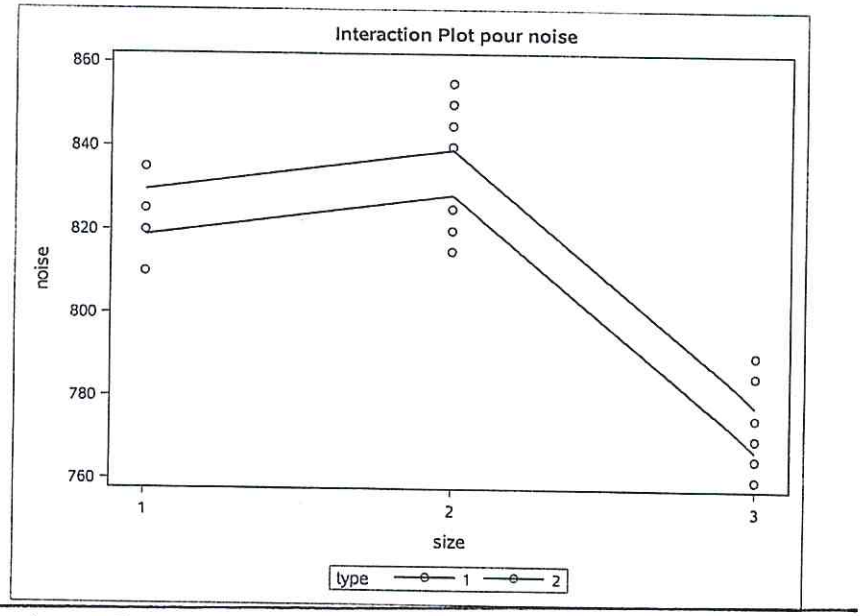
Paramètre	Valeur estimée		Erreur type	Valeur du test t	Pr > |t|
Intercept	767.0833333	B	3.09943245	247.49	<.0001
size 1	51.6666667	B	3.79601399	13.61	<.0001
size 2	61.2500000	B	3.79601399	16.14	<.0001
size 3	0.0000000	B	.	.	.
type 1	10.8333333	B	3.09943245	3.50	0.0014
type 2	0.0000000	B	.	.	.

page 5

The GLM Procedure

Dependent Variable: noise

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.



Exercice 2

Procédure LOGISTIC

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

Nb d'observations lues	209
Nb d'observations utilisées	194

page 6

Procédure LOGISTIC

Profil de réponse		
Valeur ordonnée	carrier	Fréquence totale
1	1	67
2	0	127

La probabilité modélisée est carrier=1.

Note: 15 observations were deleted due to missing values for the response or explanatory variables.

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

Statistiques d'ajustement du modèle		
Critère	Constante uniquement	Constante et Covariables
AIC	252.077	96.038
SC	255.344	115.646
-2 Log L	250.077	84.038

Test de l'hypothèse nulle globale : BETA=0			
Test	Khi-2	DDL	Pr > Khi-2
Rapport de vrais	166.0383	5	<.0001
Score	108.0405	5	<.0001
Wald	35.4505	5	<.0001

Estimations par l'analyse du maximum de vraisemblance					
Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept	1	-19.5489	3.3059	34.9666	<.0001
age	1	0.1478	0.0447	10.9066	0.0010
ck	1	0.0467	0.0148	9.8952	0.0017
h	1	0.0826	0.0272	9.1925	0.0024
pk	1	0.1163	0.0482	5.8261	0.0158
ld	1	0.0125	0.00574	4.7412	0.0294

page 7

Procédure LOGISTIC

Estimations des rapports de cotes			
Effet	Valeur estimée du point	95% Intervalle de confiance de Wald	
age	1.159	1.062	1.266
ck	1.048	1.018	1.079
h	1.086	1.030	1.146
pk	1.123	1.022	1.235
ld	1.013	1.001	1.024

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	96.3	D de Somers	0.925
Pourcentage discordant	3.7	Gamma	0.925
Pourcentage lié	0.0	Tau-a	0.420
Paires	8509	c	0.963

Procédure FREQ

Fréquence
Pourcentage
Pctage en ligne
Pctage en col.

Table de _FROM_ par _INTO_			
FROM(Valeur formatée de la réponse observée)	_INTO_(Valeur formatée de la réponse prédite)		
	0	1	Total
0	122 62.89 96.06 90.37	5 2.58 3.94 8.47	127 65.46
1	13 6.70 19.40 9.63	54 27.84 80.60 91.53	67 34.54
Total	135 69.59	59 30.41	194 100.00
Fréquence manquante = 15			