Master M2, SITN et Data Science, Université Claude Bernard, Lyon 1
Modèles de régression
année 2018-2019

Examen du 30 janvier 2019,
Calculatrice autorisée
Appareils connectables interdits.
Durée 1h30.

Note: Le premier exercice est traité avec le logiciel SAS, pendant que l'exercice 2 est traité avec le logiciel R. 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

Dans le fichier "*examen_exo1.txt*, on considère 241 patients souffrant de *"Gammapothy monoclonal de signification indéterminée"* pour lesquels on a mesuré les variables:
age: l'âge;
sex: homme ou femme;
futime: le nombre de jours à partir du diagnostic jusqu'au dernier suivi;
death: =1 si décès, 0 si en vie;
alb: niveau albumine ;
creat: niveau créatinine;
hgb: hémoglobine;
mspike: taille de la pointe protien monoclonale au diagnostic.

On veut modéliser la probabilité de décès fonction des variables: *age, sex, alb, creat, hgb, pspike.*

1) Pour le modèle (M1), écrivez le modèle statistique correspondant.
2) Testez si le modèle (M1) est significatif. (spécifiez: les hypothèses H_0, H_1, les modèles correspondants, statistiques de test utilisées, conclusion).
3) Pour le modèle (M1), quelles variables ont une influence sur la probabilité que le patient décède? (pour les variables explicatives *age, sex* donnez: les hypothèses H_0, H_1, les modèles correspondants, statistique de test et sa loi sous H_0, conclusion. Pour les quatre autres variables explicatives, donnez seulement la conclusion).
4) Donnez les estimations de tous paramètres du modèle (M1).
5) En considérant le modèle (M1), quelle est la probabilité qu'un patient soit en vie, sachant que le patient a 70 ans, c'est un homme, diagnostiqué il y a 300 jours, avec les niveaux: d'albumine égal à 2, de créatinine égale à 1, d'hémoglobine égale 10 et il a une taille de la pointe protien monoclonale égale à 2? (vous donnez seulement l'expression de la probabilité, sans faire les calculs)
6) On considère maintenant le modèle (M2). Donnez la forme statistique de ce modèle.
7) Comment les variables explicatives du modèle (M2) influent la probabilité de décès? Explications.
8) Avec un risque de *0.05*, peut-on dire que les estimations des paramètres obtenues par le modèle (M2) sont les mêmes que celles obtenues par le modèle (M1)?

9) Réalisez une comparaison entre les vraies valeurs de la variable *death* et les valeurs prévues pour cette variable par le modèle (M2). Commentez la qualité du modèle (M2).

Exercice 2.

Les données pour cet exercice proviennent du package MASS du logiciel R. Plus précisément, le tableau de données *oats* contient des données sur le rendement de la culture de l'avoine sur un champ divisé en parcelles et traité avec quatre doses d'engrais. Les variables mesurées sont:
Y: le rendement
B: le block de la parcelle, valeurs prises: 1, 2, 3, 4, 5, 6
V: variété de l'avoine, valeurs: 1, 2, 3
N: niveau de l'engrais, valeurs 1 (pour 0 engrais), 2 (pour 0.2 cwt engrais), 3 (pour 0.4 cwt engrais), 4 (pour 0.6 cwt engrais)

Pour information, "cwt" est l'abréviation du quintal (en anglais "hundredweight"), une unité de mesure du poids. En Amérique du Nord, un quintal équivaut à 100 pounds.

1) Est-ce que la variable rendement suit une loi normale? Justification par test d'hypothèse.
2) Donnez la forme statistique du modèle (M3). Il s'agit de quel type de modèle?
3) Testez si le modèle (M3) 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)
4) Si le modèle (M3) est significatif, quelles sont les variables qui influent le rendement de la culture d'avoine? (*Pour chaque type de variable explicative* 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?
5) On considère maintenant le modèle (M4). Donnez sa forme statistique.
6) Quel est l'effet sur le rendement si le niveau d'engrais est de 0.6 cwt?
7) Est-ce que le rendement est affecté si on n'utilise pas d'engrais? Si oui, de quelle manière?
8) Il y a-t-il des variétés d'avoine qui ont un rendement plus faible?
9) Donnez la qualité globale du modèle (M4).

```
########################   CODE SAS, EXERCICE 1 ###################

 data exo1;
  infile "C:examen_exo1.txt";
   input   age    sex $    futime death alb creat  hgb mspike;
 run;

/*  MODELE (M1)  */
  proc logistic  data=exo1  ;
 class sex;
 model death=age sex alb creat  hgb mspike;
run;

/*  MODELE (M2)  */
proc logistic  data=exo1  ;
 class sex;
 model death=age    hgb ;
 output out=outlog p=prev predprob=(individual crossvalidate);
run;
proc freq data=outlog;
title "Sorties pour PROC FREQ ";
  table _FROM_*_INTO_; run;


########################   CODE R, EXERCICE 2 ###################

library(car)
library(MASS)
data(oats)
attach(oats)
shapiro.test(Y)
B=factor(B)
V=factor(V)
N=factor(N)
######  MODELE (M3) #####
m3=lm(Y~B+V+N+B*N+V*N+B*V,contrasts=list(B=contr.sum,V=contr.sum, N=contr.sum))
print(summary(m3))
cat("\n ANOVA DE TYPE III pour le modèle (m3) \n ")
print(Anova(m3,type="III"))

###### MODELE (M4)   ####
m4=lm(Y~B+V+N+B*V,contrasts=list(B=contr.sum,V=contr.sum, N=contr.sum))
print(summary(m4))
cat("\n ANOVA DE TYPE III pour le modèle (m4) \n ")
print(Anova(m4,type="III"))
```

Modèle (M1)

The LOGISTIC Procedure

Model Information	
Data Set	WORK.EXO1
Response Variable	death
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	241
Number of Observations Used	187

Response Profile		
Ordered Value	death	Total Frequency
1	0	45
2	1	142

Probability modeled is death=0.

Note: 54 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information		
Class	Value	Design Variables
sex	female	1
	male	-1

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	208.380	160.293
SC	211.611	182.910
-2 Log L	206.380	146.293

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	60.0875	6	<.0001
Score	50.7406	6	<.0001
Wald	34.1025	6	<.0001

The LOGISTIC Procedure

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
age	1	31.1036	<.0001
sex	1	0.3765	0.5395
alb	1	0.2609	0.6095
creat	1	0.1881	0.6645
hgb	1	7.7987	0.0052
mspike	1	0.1984	0.6560

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.6495	2.9289	0.3172	0.5733
age		1	-0.1402	0.0251	31.1036	<.0001
sex	female	1	0.1426	0.2324	0.3765	0.5395
alb		1	-0.2548	0.4988	0.2609	0.6095
creat		1	0.2079	0.4794	0.1881	0.6645
hgb		1	0.4943	0.1770	7.7987	0.0052
mspike		1	-0.2296	0.5156	0.1984	0.6560

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	0.869	0.827	0.913
sex female vs male	1.330	0.535	3.308
alb	0.775	0.292	2.060
creat	1.231	0.481	3.151
hgb	1.639	1.159	2.319
mspike	0.795	0.289	2.183

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	85.3	Somers' D	0.707
Percent Discordant	14.7	Gamma	0.707
Percent Tied	0.0	Tau-a	0.260
Pairs	6390	c	0.853

Modèle (M2)

The LOGISTIC Procedure

Model Information	
Data Set	WORK.EXO1
Response Variable	death
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	241
Number of Observations Used	239

Response Profile		
Ordered Value	death	Total Frequency
1	0	57
2	1	182

Probability modeled is death=0.

Note: 2 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	264.583	206.686
SC	268.060	217.115
-2 Log L	262.583	200.686

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	61.8976	2	<.0001
Score	54.8383	2	<.0001
Wald	40.1158	2	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2389	1.8692	0.4393	0.5075
age	1	-0.1123	0.0187	36.0687	<.0001
hgb	1	0.3251	0.1209	7.2299	0.0072

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	0.894	0.862	0.927
hgb	1.384	1.092	1.754

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	82.7	Somers' D	0.655
Percent Discordant	17.2	Gamma	0.656
Percent Tied	0.0	Tau-a	0.239
Pairs	10374	c	0.828

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of _FROM_ by _INTO_			
FROM(Formatted Value of the Observed Response)	_INTO_(Formatted Value of the Predicted Response)		
	0	1	Total
0	18 7.53 31.58 60.00	39 16.32 68.42 18.66	57 23.85
1	12 5.02 6.59 40.00	170 71.13 93.41 81.34	182 76.15
Total	30 12.55	209 87.45	239 100.00

Frequency Missing = 2

```
        Shapiro-Wilk normality test

data:  Y
W = 0.9838, p-value = 0.4807


Call:
lm(formula = Y ~ B + V + N + B * N + V * N + B * V, contrasts = list(B = contr.sum,
    V = contr.sum, N = contr.sum))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.97222    1.69156  61.465  < 2e-16 ***
B1           31.36111    3.78245   8.291 2.96e-09 ***
B2            3.27778    3.78245   0.867  0.39305
B3           -8.05556    3.78245  -2.130  0.04151 *
B4           -5.80556    3.78245  -1.535  0.13530
B5          -13.05556    3.78245  -3.452  0.00168 **
V1            0.52778    2.39223   0.221  0.82688
V2            5.81944    2.39223   2.433  0.02117 *
N1          -24.58333    2.92987  -8.391 2.30e-09 ***
N2           -5.08333    2.92987  -1.735  0.09300 .
N3           10.25000    2.92987   3.498  0.00148 **
B1:N1         0.25000    6.55139   0.038  0.96981
B2:N1        -7.00000    6.55139  -1.068  0.29383
B3:N1         1.00000    6.55139   0.153  0.87970
B4:N1        -4.25000    6.55139  -0.649  0.52146
B5:N1         2.00000    6.55139   0.305  0.76226
B1:N2        -2.25000    6.55139  -0.343  0.73366
B2:N2         5.50000    6.55139   0.840  0.40782
B3:N2         7.50000    6.55139   1.145  0.26135
B4:N2         0.58333    6.55139   0.089  0.92964
B5:N2        -5.16667    6.55139  -0.789  0.43651
B1:N3        -0.25000    6.55139  -0.038  0.96981
B2:N3        -2.83333    6.55139  -0.432  0.66849
B3:N3         4.83333    6.55139   0.738  0.46639
B4:N3        -2.75000    6.55139  -0.420  0.67765
B5:N3        -0.16667    6.55139  -0.025  0.97987
V1:N1         0.08333    4.14346   0.020  0.98409
V2:N1         1.45833    4.14346   0.352  0.72733
V1:N2        -0.91667    4.14346  -0.221  0.82641
V2:N2         3.79167    4.14346   0.915  0.36744
V1:N3        -0.08333    4.14346  -0.020  0.98409
V2:N3        -2.87500    4.14346  -0.694  0.49311
B1:V1        -2.61111    5.34919  -0.488  0.62901
B2:V1         5.47222    5.34919   1.023  0.31449
B3:V1        -9.69444    5.34919  -1.812  0.07996 .
B4:V1         9.30556    5.34919   1.740  0.09218 .
B5:V1         4.05556    5.34919   0.758  0.45427
B1:V2       -11.40278    5.34919  -2.132  0.04134 *
B2:V2         8.18056    5.34919   1.529  0.13667
B3:V2        16.76389    5.34919   3.134  0.00384 **
B4:V2        -8.98611    5.34919  -1.680  0.10336
B5:V2       -11.48611    5.34919  -2.147  0.03997 *
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 14.35 on 30 degrees of freedom
Multiple R-squared:  0.8811,    Adjusted R-squared:  0.7186
F-statistic: 5.423 on 41 and 30 DF,  p-value: 3.333e-06

 ANOVA DE TYPE III pour le modèle (m3)
 Anova Table (Type III tests)

Response: Y
           Sum Sq Df   F value    Pr(>F)
(Intercept) 778336  1 3777.9738 < 2.2e-16 ***
B            15875  5   15.4114 1.609e-07 ***
V             1786  2    4.3354   0.02219 *
N            20020  3   32.3926 1.540e-09 ***
B:N           1788 15    0.5786   0.86816
V:N            322  6    0.2603   0.95103
B:V           6013 10    2.9188   0.01123 *
Residuals     6181 30
```

```
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

Modèle (M4)

```
Call:
lm(formula = Y ~ B + V + N + B * V, contrasts = list(B = contr.sum,
    V = contr.sum, N = contr.sum))

Residuals:
     Min       1Q   Median       3Q      Max
-22.0000  -7.9375  -0.3333   7.9583  23.3333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.9722     1.5026  69.196  < 2e-16 ***
B1           31.3611     3.3599   9.334 1.28e-12 ***
B2            3.2778     3.3599   0.976 0.333886
B3           -8.0556     3.3599  -2.398 0.020204 *
B4           -5.8056     3.3599  -1.728 0.090057 .
B5          -13.0556     3.3599  -3.886 0.000295 ***
V1            0.5278     2.1250   0.248 0.804846
V2            5.8194     2.1250   2.739 0.008478 **
N1          -24.5833     2.6026  -9.446 8.68e-13 ***
N2           -5.0833     2.6026  -1.953 0.056292 .
N3           10.2500     2.6026   3.938 0.000250 ***
B1:V1        -2.6111     4.7516  -0.550 0.585045
B2:V1         5.4722     4.7516   1.152 0.254831
B3:V1        -9.6944     4.7516  -2.040 0.046519 *
B4:V1         9.3056     4.7516   1.958 0.055662 .
B5:V1         4.0556     4.7516   0.854 0.397365
B1:V2       -11.4028     4.7516  -2.400 0.020095 *
B2:V2         8.1806     4.7516   1.722 0.091195 .
B3:V2        16.7639     4.7516   3.528 0.000896 ***
B4:V2        -8.9861     4.7516  -1.891 0.064287 .
B5:V2       -11.4861     4.7516  -2.417 0.019249 *
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 12.75 on 51 degrees of freedom
Multiple R-squared:  0.8405,    Adjusted R-squared:  0.778
F-statistic: 13.44 on 20 and 51 DF,  p-value: 6.355e-14


 ANOVA DE TYPE III pour le modèle (m4)
 Anova Table (Type III tests)

Response: Y
            Sum Sq Df   F value    Pr(>F)
(Intercept) 778336  1 4788.0271 < 2.2e-16 ***
B            15875  5   19.5317 8.101e-11 ***
V             1786  2    5.4945 0.0069026 **
N            20020  3   41.0528 1.228e-13 ***
B:V           6013 10    3.6992 0.0009032 ***
Residuals     8290 51
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```