

**Master M2, SITN et Data Science, Université Claude Bernard, Lyon 1**

Modèles de régression

année 2020-2021

**Examen du 27 janvier 2021,**  
Calculatrice autorisée  
Appareils connectables interdits.  
Durée 1h15.

*Note: Les deux exercices ont été traités 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 1.**

Les données pour cet exercice proviennent du package MASS du logiciel R. Plus précisément, le tableau de données *Rabbit* contient des expériences sur des lapins. Les variables d'intérêt mesurées sont:

*BPchange*: le changement de la pression artérielle par rapport au début de l'expérience

*Dose*: la dose de phénylbiguanide, qui prend 6 valeurs (non numériques): 1, 2, 3, 4, 5, 6.

*Treatment*: le traitement, prend 2 valeurs (non numériques): 1 et 2.

- 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 le changement de la pression artérielle? (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 concernent une étude sur la rémission du cancer (Lee 1974). La variable expliquée, notée par *remiss*, prend deux valeurs: 1 pour la rémission et 0 pour la non rémission.

Trois variables explicatives sont considérées: *cell*, *li*, *blast*.

Le logiciel R est utilisé.

- 1) Quels sont les effectifs pour les deux valeurs de la variable *remiss*?
- 2) Pour le modèle (M2), écrivez le modèle statistique correspondant. Il s'agit de quel type de modèle?
- 3) Pour le modèle (M2), quelles variables ont une influence sur la probabilité de rémission? (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). Interprétation.

5) Donnez et commentez la prévision faite par le modèle (M2) pour la rémission ou la non rémission.

6) Donnez la forme statistique du modèle (M3).

7) Comparez les prévisions faites pour la variable *remiss* par les modèles (M2) et (M3). Interprétation.

```

library(car)
#### exercice 1
library(MASS)
data("Rabbit")
attach(Rabbit)
Dose=factor(Dose)
Treatment=factor(Treatment)
m1=lm(BPchange~Dose+Treatment+Dose:Treatment,contrasts=list(Dose=contr.sum, Treatment=contr.sum))
summary(m1)
print(Anova(m1,type="III"))

```

```
##### exercice 2
```

```

exo2=read.table('exo2_exam2021.txt', col.names=c("remiss", "cell", "li", "blast"));
attach(exo2)
cat("\n Les effectifs pour les valeurs de la variable remiss \n")
table(remiss)
m2=glm(remiss ~ cell+li+blast, family="binomial")
summary(m2);
pi2=predict(m2,type = "response")
prev2=vector(mode="numeric",length=length(pi2))
prev2[pi2>0.5]=1  ### prévision de "remiss" par le modèle "m2"
cat("Tableau de contingence obtenu par le modèle m2 \n")
table(remiss, prev2)

m3=glm(remiss ~ li, family="binomial")
summary(m3);
pi3=predict(m3,type = "response")
prev3=vector(mode="numeric",length=length(pi3))
prev3[pi3>0.5]=1  ### prévision de "remiss" par le modèle "m3"
cat("Tableau de contingence obtenu par le modèle m3 \n")
table(remiss, prev3)

```

```

##### SORTIES #####
##### SORTIES EXERCICE 1 #####

```

```

Call:
lm(formula = BPchange ~ Dose + Treatment + Dose:Treatment, contrasts = list(Dose = contr.sum,
  Treatment = contr.sum))

```

```

Residuals:
    Min     1Q  Median     3Q     Max
-9.200 -1.450 -0.365  1.090 10.800

```

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.2183    0.5902   19.008 < 2e-16 ***
Dose1          -9.8783    1.3197   -7.485 1.33e-09 ***
Dose2          -9.1983    1.3197   -6.970 8.17e-09 ***
Dose3          -7.2683    1.3197   -5.508 1.40e-06 ***
Dose4          -0.4183    1.3197   -0.317 0.752625
Dose5          11.2817    1.3197    8.549 3.31e-11 ***
Treatment1     2.3400    0.5902    3.965 0.000244 ***
Dose1:Treatment1 -2.6800    1.3197   -2.031 0.047838 *
Dose2:Treatment1 -2.0100    1.3197   -1.523 0.134303
Dose3:Treatment1 -0.6900    1.3197   -0.523 0.603487
Dose4:Treatment1  4.2600    1.3197    3.228 0.002250 **
Dose5:Treatment1  2.9600    1.3197    2.243 0.029554 *
---

```

```

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

```

```

Residual standard error: 4.572 on 48 degrees of freedom
Multiple R-squared:  0.8709,    Adjusted R-squared:  0.8414
F-statistic: 29.45 on 11 and 48 DF,  p-value: < 2.2e-16

```

```

Anova Table (Type III tests)

```

```

Response: BPchange
            Sum Sq Df  F value    Pr(>F)
(Intercept) 7551.1  1 361.3055 < 2.2e-16 ***
Dose        6021.5  5  57.6240 < 2.2e-16 ***

```

```
Treatment      328.5  1  15.7199 0.0002436 ***
Dose:Treatment  419.9  5   4.0186 0.0039856 **
Residuals      1003.2 48
```

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

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

```
Les effectifs pour les valeurs de la variable remiss
remiss
```

```
 0 1
18 9
```

```
Call:
glm(formula = remiss ~ cell + li + blast, family = "binomial")
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9939  -0.6241  -0.3260   0.6974   1.6312
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.6703     7.1314  -1.496  0.1346
cell          7.4644     7.0579   1.058  0.2902
li            3.2610     1.6477   1.979  0.0478 *
blast        -0.5734     1.4925  -0.384  0.7008
```

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

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 34.372 on 26 degrees of freedom
Residual deviance: 24.186 on 23 degrees of freedom
AIC: 32.186
```

```
Number of Fisher Scoring iterations: 6
```

```
Tableau de contingence obtenu par le modèle m2
```

```
      prev2
remiss 0 1
      0 16 2
      1 5 4
```

```
Call:
glm(formula = remiss ~ li, family = "binomial")
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9448  -0.6465  -0.4947   0.6571   1.6971
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.777     1.379  -2.740  0.00615 **
li            2.897     1.187   2.441  0.01464 *
```

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

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 34.372 on 26 degrees of freedom
Residual deviance: 26.073 on 25 degrees of freedom
AIC: 30.073
```

```
Number of Fisher Scoring iterations: 4
```

```
Tableau de contingence obtenu par le modèle m3
```

```
      prev3
remiss 0 1
      0 16 2
      1 5 4
```

modèle (M2)

modèle (M3)