Master M2, SITN et Data Science, Université Claude Bernard, Lyon1
Modèles de régression
année 2016-2017

Examen du 14 décembre 2016,
Documents admis. Calculatrice autorisée
Appareil connectables interdits.
Durée 1h30.

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. 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 $\alpha = 0.05$. Ecrire les hypothèses à tester.*

Exercice 1.

On veut étudier la présence des nids de chouettes dans des sites forestiers. Pour ceci, on a mesuré les variables:
rand: une variable qui prend deux valeurs: "N" si il y a des nids de chouettes dans la forêt, "R" sinon;
km91: le % de forêt mature sur un rayon de 0.91km;
km118: le % de forêt mature sur une couronne comprise entre deux cercles de rayons 0.91km et de 1.18km;
km140: le % de forêt mature sur une couronne comprise entre deux cercles de rayons 1.18km et 1.40km;
km160: le % de forêt mature sur une couronne comprise entre deux cercles de rayons 1.40km et 1.60km;
km177: le % de forêt mature sur une couronne comprise entre deux cercles de rayons 1.60km et 1.77km;
km241: le % de forêt mature sur une couronne comprise entre deux cercles de rayons 1.77km et 2.41km;
km338: le % de forêt mature sur une couronne comprise entre deux cercles de rayons 2.41km et 3.38km;

1) On modélise la variable *rand* fonction des autres variables spécifiées. Quel type de modèle a été utilisé?
2) Quel est le nombre de sites considérés? Dans combien de sites il y a des nids de chouettes?
3) Est-ce que la présence des nids de chouettes est influencée par les variables considérées? (faire des tests d'hypothèse et donner des indicateurs quantitatifs)
4) Si la réponse à la question précédente est positive, quelles sont les variables qui influencent la présence de nids?
5) Pour le modèle qui contient seulement des variables explicatives significatives, donnez les estimations des paramètres de ce modèle. Interprétez ces estimations. Quelles sont les prévisions réalisées pas ce modèle? Quel est le taux d'erreur de prévision?

Exercice 2.

Dans cet exercice il s'agit d'une étude sur la croissance des Pins. Le fichier *ch5.txt* contient des observations pour les variables suivantes (sont spécifiées seulement celles qu'on a utilisé):
site: le site (pépinière);
faam: famille génétique des pins;
biomass: total biomasse d'un pin, après deux années de croissance.
On modélise la variable *biomass* fonction des variables *site* et *faam*.

1) Quel type de modèle utilise-t-on? Justification.
2) Donnez le nombre total d'observations et le nombre d'observations utilisées. Quelles sont les valeurs des variables *site, faam*. Ecrivez la forme du modèle statistique (M3).
3) Testez si le modèle (M3) est significatif. En faisant des tests d'hypothèse, répondez à la question: quelles sont les variables significatives du modèle (M3)?
4) On considère maintenant le modèle (M4). Ecrivez sa forme statistique associée.
5) Pour le modèle (M4), quelles sont les variables influentes sur *biomass*?
6) Donnez les estimations des paramètres du modèle (M4). Interprétez ces estimations.
7) Il y a-t-il des sites pour lesquels le total de biomasse, après des années de croissance, est le même?
8) Comparez les R^2 pour les modèles (M3) et (M4). Interprétation.

CODE : SAS


```
/* Exercice 1 */


 data ramsey ;
infile 'C:\Users\Gabriela.Ciuperca\M2PRO\Modele_regr\exam2016\ch10.txt' expandtabs;
input rand $  km91 km118 km140 km160 km177 km241 km338 ;
run;
proc logistic data=ramsey covout descending;
     model rand=km91 km118 km140 km160 km177 km241 km338 ;
        run;
  proc logistic data=ramsey covout descending;
     model rand=km91 km118 km140  km241  ;
        output out=outlog p=prev predprob=(individual crossvalidate); run;
run;
proc freq data=outlog;
  table _FROM_*_INTO_; run;

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

data exo2;
 infile 'C:\Users\Gabriela.Ciuperca\M2PRO\Modele_regr\exam2016\ch5.txt';
 input site block rep ozone rain faam ppmhrs vwpph biomass diam dma dmb d2ha dwhb dmot; run;

 proc glm data=exo2 ;
      class site faam;
      model biomass =site faam  site*faam  ;
       run;
     proc glm data=exo2 ;
      class site ;
      model biomass =site  / solution ;
       run;
```

(M0) et (M1)

(M2)

(M3)

(M4)

SORTIE.SAS

Exercice 1

(M0) et (M1)

Procédure LOGISTIC

Model Information	
Table	WORK.RAMSEY
Variable de réponse	rand
Nombre de niveaux de réponse	2
Modèle	logit binaire
Technique d'optimisation	Score de Fisher

Nombre d'observations lues	60
Nombre d'observations utili	60

Response Profile		
Ordered Value	rand	Total Frequency
1	R	30
2	N	30

La probabilité modélisée est rand='R'.

Model Convergence Status
Critère de convergence (GCONV=1E-8) respecté.

Model Fit Statistics		
Criterion	Intercept Only	Constante et Covariables
AIC	85.178	68.134
SC	87.272	84.888
-2 Log	83.178	52.134

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DDL	Pr > Khi-2
Rapport de vrais	31.0440	7	<.0001
Score	23.5778	7	0.0014
Wald	13.0857	7	0.0700

Procédure LOGISTIC

Analysis of Maximum Likelihood Estimates					
Parameter	DDL	Estimate	Standard Error	Wald Chi-Square	Pr > Khi-2
Intercept	1	9.8071	3.4005	8.3174	0.0039
km91	1	-0.0570	0.0371	2.3547	0.1249
km118	1	0.1179	0.0497	5.6299	0.0177
km140	1	-0.1220	0.0523	5.4309	0.0198
km160	1	0.0159	0.0437	0.1324	0.7160
km177	1	-0.0329	0.0391	0.7070	0.4004
km241	1	-0.1088	0.0665	2.6772	0.1018
km338	1	0.0518	0.0362	2.0435	0.1529

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
km91	0.945	0.878	1.016
km118	1.125	1.021	1.240
km140	0.885	0.799	0.981
km160	1.016	0.933	1.107
km177	0.968	0.896	1.045
km241	0.897	0.787	1.022
km338	1.053	0.981	1.131

Association of Predicted Probabilities and Observed Responses			
Pourcentage concordant	88.2	D de Somers	0.768
Pourcentage discordant	11.4	Gamma	0.770
Pourcentage lié	0.3	Tau-a	0.390
Paires	900	c	0.884

(M2)

Procédure LOGISTIC

Model Information	
Table	WORK.RAMSEY
Variable de réponse	rand
Nombre de niveaux de réponse	2
Modèle	logit binaire
Technique d'optimisation	Score de Fisher

Procédure LOGISTIC

Nombre d'observations lues	60
Nombre d'observations utili	60

Response Profile		
Ordered Value	rand	Total Frequency
1	R	30
2	N	30

La probabilité modélisée est rand='R'.

Model Convergence Status
Critère de convergence (GCONV=1E-8) respecté.

Model Fit Statistics		
Criterion	Intercept Only	Constante et Covariables
AIC	85.178	65.046
SC	87.272	75.518
-2 Log	83.178	55.046

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DDL	Pr > Khi-2
Rapport de vrais	28.1317	4	<.0001
Score	20.9999	4	0.0003
Wald	13.3694	4	0.0096

Analysis of Maximum Likelihood Estimates					
Parameter	DDL	Estimate	Standard Error	Wald Chi-Square	Pr > Khi-2
Intercept	1	10.2758	3.0299	11.5020	0.0007
km91	1	-0.0643	0.0348	3.4145	0.0646
km118	1	0.1199	0.0476	6.3342	0.0118
km140	1	-0.1233	0.0470	6.8850	0.0087
km241	1	-0.0811	0.0385	4.4385	0.0351

Procédure LOGISTIC

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
km91	0.938	0.876	1.004
km118	1.127	1.027	1.238
km140	0.884	0.806	0.969
km241	0.922	0.855	0.994

Association of Predicted Probabilities and Observed Responses			
Pourcentage concordant	86.0	D de Somers	0.721
Pourcentage discordant	13.9	Gamma	0.722
Pourcentage lié	0.1	Tau-a	0.367
Paires	900	c	0.861

The FREQ Procedure

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)		
	N	R	Total
N	24 40.00 80.00 80.00	6 10.00 20.00 20.00	30 50.00
R	6 10.00 20.00 20.00	24 40.00 80.00 80.00	30 50.00
Total	30 50.00	30 50.00	60 100.00

Exercice 2
(M3)

The GLM Procedure

Class Level Information		
Class	Levels	Values
site	6	1 2 3 4 5 6
faam	4	1 2 3 4

Number of Observations Read	612
Number of Observations Used	450

The GLM Procedure

Dependent Variable: biomass

The GLM Procedure

Dependent Variable: biomass

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Model	17	37326507.99	2195676.94	16.78	<.0001
Error	432	56520384.34	130834.22		
Corrected Total	449	93846892.33			

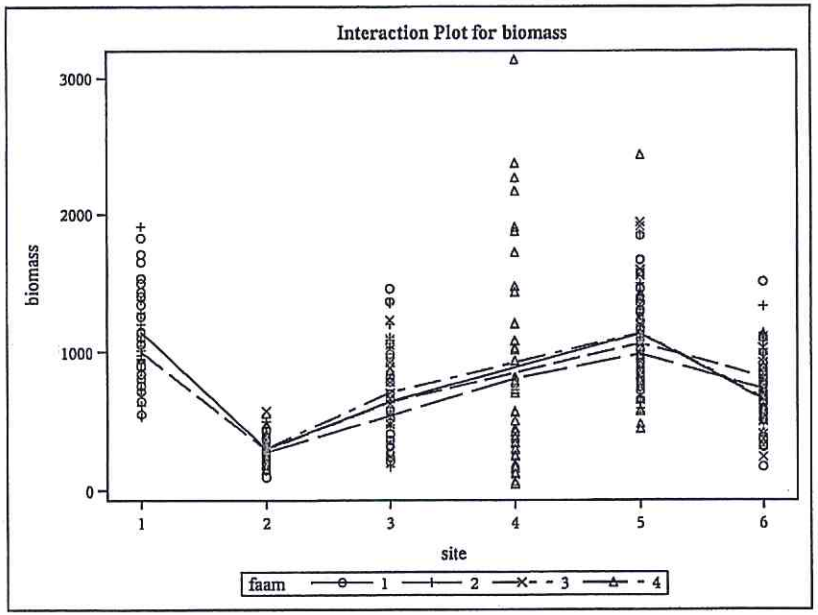
R-Square	Coeff Var	Root MSE	biomass Mean
0.397738	48.42832	361.7101	746.8979

Source	DDL	Type I SS	Moyenne quadratique	Valeur F	Pr > F
site	5	36212484.30	7242496.86	55.36	<.0001
faam	3	114810.11	38270.04	0.29	0.8308
site*faam	9	999213.58	111023.73	0.85	0.5717

Source	DDL	Type III SS	Moyenne quadratique	Valeur F	Pr > F
site	5	35962106.52	7192421.30	54.97	<.0001
faam	3	117428.47	39142.82	0.30	0.8260
site*faam	9	999213.58	111023.73	0.85	0.5717

The GLM Procedure

Dependent Variable: biomass



CM4)

The GLM Procedure

Class Level Information		
Class	Levels	Values
site	6	1 2 3 4 5 6

Number of Observations Read	612
Number of Observations Used	450

The GLM Procedure

Dependent Variable: biomass

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Model	5	36212484.30	7242496.86	55.79	<.0001
Error	444	57634408.03	129807.23		
Corrected Total	449	93846892.33			

The GLM Procedure

Dependent Variable: biomass

R-Square	Coeff Var	Root MSE	biomass Mean
0.385868	48.23788	360.2877	746.8979

Source	DDL	Type I SS	Moyenne quadratique	Valeur F	Pr > F
site	5	36212484.30	7242496.86	55.79	<.0001

Source	DDL	Type III SS	Moyenne quadratique	Valeur F	Pr > F
site	5	36212484.30	7242496.86	55.79	<.0001

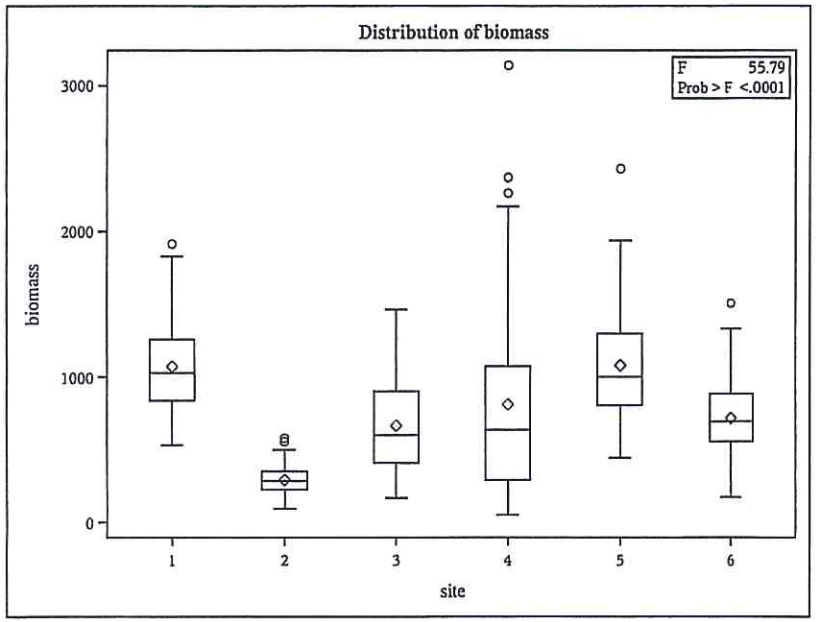
Parameter		Valeur estimée		Erreur type	Valeur du test t	Pr > |t|
Intercept		718.0962500	B	36.77170919	19.53	<.0001
site	1	357.3087500	B	63.69046860	5.61	<.0001
site	2	-424.0183333	B	52.00304985	-8.15	<.0001
site	3	-54.2914167	B	59.29259947	-0.92	0.3603
site	4	93.8930093	B	61.28618199	1.53	0.1262
site	5	361.4890625	B	52.00304985	6.95	<.0001
site	6	0.0000000	B	.	.	.

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.

CODE et SORTIES : R

The GLM Procedure

Dependent Variable: biomass




```
######################   CODE R  ############################################################

library(car)
###############"      EXERCICE 1 ###################################
ramsey=read.table('ch10.txt', col.names=c("rand", "km91", "km118", "km140", "km160", "km177",
"km241", "km338"));
attach(ramsey)
rand=factor(rand);
#  modélé fonction seulement d une constante #
m0=glm(rand ~ 1, family="binomial")
summary(m0);
BIC(m0)
#  modélé fonction de toutes les variabes #

m1=glm(rand ~ km91+km118+km140+km160+km177+km241+km338, family="binomial")
summary(m1);
BIC(m1)

#  modélé fonction de toutes les variabes #

m2=glm(rand ~ km91+km118+km140+km241, family="binomial")
summary(m2);
BIC(m2)
#  tableau de contingence vraies valeurs et prévisions ##
cat("tableau de contingence vraies valeurs et prévisions \n")
table(m2$fitted.value>0.5,rand=="R")

###############"      EXERCICE 2 ###################################
exo2=read.table('ch5.txt', na.strings = ".",
col.names=c("site", "block", "rep", "ozone", "rain", "faam", "ppmhrs", "vwpph", "biomass", "diam",
"dma", "dmb", "d2ha", "dwhb", "dmot"));
attach(exo2)
site=factor(site);
faam=factor(faam);
m3=lm(biomass~site+faam+site*faam,
contrasts=list(site=contr.sum, faam=contr.sum))
print(summary(m3))
cat("\n ANOVA DE TYPE III \n ")
print(Anova(m3,type="III"))

m4=lm(biomass~site,
contrasts=list(site=contr.sum))
print(summary(m4))
cat("\n ANOVA DE TYPE III \n ")
print(Anova(m4,type="III"))

############################################### SORTIES ######################

Call:
glm(formula = rand ~ 1, family = "binomial")

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.177  -1.177   0.000   1.177   1.177

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.867e-17  2.582e-01       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.178  on 59  degrees of freedom
Residual deviance: 83.178  on 59  degrees of freedom
AIC: 85.178
```

(M0)

```
Number of Fisher Scoring iterations: 2

[1] 87.27201
```

(M1)

```
Call:
glm(formula = rand ~ km91 + km118 + km140 + km160 + km177 + km241 +
    km338, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3919  -0.7454  -0.0187   0.7252   1.9051

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.80718    3.40056   2.884  0.00393 **
km91        -0.05698    0.03713  -1.535  0.12490
km118        0.11793    0.04970   2.373  0.01766 *
km140       -0.12199    0.05235  -2.330  0.01978 *
km160        0.01591    0.04374   0.364  0.71598
km177       -0.03286    0.03908  -0.841  0.40044
km241       -0.10880    0.06650  -1.636  0.10179
km338        0.05176    0.03621   1.430  0.15285
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.178  on 59  degrees of freedom
Residual deviance: 52.134  on 52  degrees of freedom
AIC: 68.134

Number of Fisher Scoring iterations: 6

[1] 84.88845
```

(M2)

```
Call:
glm(formula = rand ~ km91 + km118 + km140 + km241, family = "binomial")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.39200  -0.69305  -0.04356   0.83558   1.80758

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.27576    3.02970   3.392 0.000695 ***
km91        -0.06430    0.03479  -1.848 0.064615 .
km118        0.11992    0.04765   2.517 0.011840 *
km140       -0.12332    0.04700  -2.624 0.008690 **
km241       -0.08114    0.03851  -2.107 0.035132 *
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.178  on 59  degrees of freedom
Residual deviance: 55.046  on 55  degrees of freedom
AIC: 65.046

Number of Fisher Scoring iterations: 5

[1] 75.5177
tableau de contingence vraies valeurs et prévisions

       FALSE TRUE
  FALSE    24    6
  TRUE      6   24
```


(M3)

```
Call:
lm(formula = biomass ~ site + faam + site * faam, contrasts = list(site = contr.sum,
    faam = contr.sum))

Residuals:
    Min      1Q  Median      3Q     Max
-761.47 -188.75  -33.02  137.20 2322.28

Coefficients: (6 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   856.34     120.85   7.086 5.65e-12 ***
site1         324.22     236.72   1.370   0.1715
site2        -562.26     124.55  -4.514 8.20e-06 ***
site3        -207.52     165.85  -1.251   0.2115
site4         360.56     361.78   0.997   0.3195
site5         223.24     124.55   1.792   0.0738 .
faam1          53.74      63.94   0.840   0.4011
faam2         -20.91      63.94  -0.327   0.7438
faam3          60.14      63.94   0.941   0.3475
site1:faam1   -86.15     353.32  -0.244   0.8075
site2:faam1   -48.22      90.43  -0.533   0.5941
site3:faam1   -60.09     154.88  -0.388   0.6982
site4:faam1   311.95     469.29   0.665   0.5066
site5:faam1       NA         NA      NA       NA
site1:faam2  -156.99     265.19  -0.592   0.5542
site2:faam2    21.47      90.43   0.237   0.8124
site3:faam2    12.09     154.88   0.078   0.9378
site4:faam2       NA         NA      NA       NA
site5:faam2       NA         NA      NA       NA
site1:faam3   161.30     156.63   1.030   0.3037
site2:faam3   -50.24      90.43  -0.556   0.5788
site3:faam3       NA         NA      NA       NA
site4:faam3       NA         NA      NA       NA
site5:faam3       NA         NA      NA       NA
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 361.7 on 432 degrees of freedom
  (162 observations deleted due to missingness)
Multiple R-squared:  0.3977,    Adjusted R-squared:  0.374
F-statistic: 16.78 on 17 and 432 DF,  p-value: ≤ 2.2e-16
```

ANOVA DE TYPE III

(M4)

```
Call:
lm(formula = biomass ~ site, contrasts = list(site = contr.sum))

Residuals:
    Min      1Q  Median      3Q     Max
-761.47 -196.54  -28.23  151.67 2322.28

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   773.83      17.74  43.623  < 2e-16 ***
site1         301.58      46.02   6.554 1.56e-10 ***
site2        -479.75      34.87 -13.757  < 2e-16 ***
site3        -110.02      41.92  -2.625  0.00897 **
site4          38.16      43.79   0.872  0.38391
site5         305.76      34.87   8.768  < 2e-16 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 360.3 on 444 degrees of freedom
  (162 observations deleted due to missingness)
Multiple R-squared:  0.3859,    Adjusted R-squared:  0.379
F-statistic: 55.79 on 5 and 444 DF,  p-value: ≤ 2.2e-16
```

```
 ANOVA DE TYPE III
 Anova Table (Type III tests)

Response: biomass
              Sum Sq  Df  F value    Pr(>F)
(Intercept) 247019941   1 1902.975 < 2.2e-16 ***
site         36212484   5   55.794 < 2.2e-16 ***
Residuals    57634408 444
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

(M4)