EISTI – Département Mathématiques

Rattrapage Modèle Linéaire : MI

Février 2019 – durée 2h00

La consultation et l'échange de documents et de calculatrices sont interdits.
Deux feuilles manuscrites R/V A4 et les calculatrices sont autorisées.

Exercice 1 [5 points]

Nous utilisons le modèle de régression linéaire multiple :

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon.$$

Nous obtenons le tableau d'analyse de la variance suivant :

Source de variation	Somme des carrés	d.d.l.	Carrés moyens	F_{obs}
Régression	1504,4	2		
Résiduelle		17		
Totale	1680,8	19		

1) Combien y-a-t-il d'observations dans le jeu de données ?
2) Quelle est la somme des carrées résiduelle ?
3) En déduire le pourcentage de variabilité de Y expliqué par ce modèle ?
4) Justifier les nombres 2, 17 et 19 de la colonne d.d.l.
5) Calculer les carrés moyens
6) En déduire F_{obs}
7) Tester l'hypothèse nulle

   $$H_0 : \beta_1=\beta_2= 0$$

   contre l'hypothèse alternative

   $$H_1 : \text{au moins un des } \beta\neq 0.$$

8) Donner une estimation de la variance de ε.

Exercice 2 [5 points]

On étudie l'influence des heures de travail sur la production industrielle. Pour cela, on dispose des observations de 3 entreprises résumées dans le tableau ci-dessous:

T_1 =Travail (M heures)	Y = Production (M tonnes)
11	45
12	65
14	75

N.B. Le jeu de données est très petit pour les besoins de l'exercice

1) Ecrire le modèle en précisant les hypothèses sur les résidus.
2) Ecrire la matrice du modèle X.

On obtient les résultats suivants :

$$({}^tXX)=\begin{pmatrix}3 & 37\\ 37 & 461\end{pmatrix} \quad \text{et} \quad ({}^tXX)^{-1}=\begin{pmatrix}32,9 & -2,6\\ -2,6 & 0,2\end{pmatrix}$$

3) Vérifier que β_0=41,5 et β_1=-16.
4) Calculer la prévision obtenue par le modèle pour la première ligne du tableau (t_1=11).
5) A partir de la matrice (^tXX), calculer la moyenne et la variance de T_1.
6) Déterminer un intervalle avec un niveau de confiance de 5% pour la vraie valeur de Y sachant t_1=11.

Exercice 3 [9 points]

Considérons le jeu de données sur le Titanic. L'objectif est de prédire si un individu va survivre en fonction de son age, son sexe et sa classe.

Class	Age	Sex	Survived
first	adult	male	yes
first	adult	male	yes
first	adult	male	yes
first	adult	male	Yes
…	…	…	…

```
summary(Mydata)
    Class        Age           Sex       Survived
 crew  :885   adult:2092   female: 470   no :1490
 first :325   child: 109   male  :1731   yes: 711
 second:285
 third :706
```

1) Modèle en fonction du sexe

```
Call:
glm(formula = Survived ~ Sex, family = binomial, data = Mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6226  -0.6903  -0.6903   0.7901   1.7613

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0044     0.1041   9.645   <2e-16 ***
Sexmale      -2.3172     0.1196 -19.376   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2769.5  on 2200  degrees of freedom
Residual deviance: 2335.0  on 2199  degrees of freedom
AIC: 2339

Number of Fisher Scoring iterations: 4
```

1.1. Ecrire le modèle obtenu.
1.2. En déduire la probabilité de survivre pour un homme et pour une femme.
1.3. Calculer l'odd des hommes et celui des femmes.
1.4. Quel est l'odd-ratio entre les hommes et les femmes ?

2) Modèle en fonction de la classe

```
Call:
glm(formula = Survived ~ Class, family = binomial, data = Mydata)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3999  -0.7623  -0.7401   0.9702   1.6906

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.15516    0.07876 -14.667  < 2e-16 ***
Classfirst   1.66434    0.13902  11.972  < 2e-16 ***
Classsecond  0.80785    0.14375   5.620 1.91e-08 ***
Classthird   0.06785    0.11711   0.579    0.562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2769.5  on 2200  degrees of freedom
Residual deviance: 2588.6  on 2197  degrees of freedom
AIC: 2596.6

Number of Fisher Scoring iterations: 4
```

2.1. Ecrire le modèle obtenu.
2.2. Est-ce que tous les coefficients du modèle sont significativement non nuls ?
2.3. Quelle est la probabilité de survivre pour une personne en seconde classe ?
Pour un personnel d’équipage ?

3) Comparaison de modèles

La troisième possibilité est d’expliquer Y en fonction de l’âge.

```
Call:
glm(formula = Survived ~ Age, family = binomial, data = Mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2166  -0.8659  -0.8659   1.5250   1.5250

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.78790    0.04716 -16.705  < 2e-16 ***
Agechild     0.87971    0.19748   4.455  8.4e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2769.5  on 2200  degrees of freedom
Residual deviance: 2749.9  on 2199  degrees of freedom
AIC: 2753.9

Number of Fisher Scoring iterations: 4
```

Quel est le meilleur des trois modèles. Justifier votre réponse ?

Exercice 4 [6 points]

Considérons le jeu de données mtcars disponible sous R.

```
The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).
```

```
[, 1] mpg  Miles/(US) gallon
[, 2] cyl  Number of cylinders
[, 3] disp Displacement (cu.in.)
[, 4] hp   Gross horsepower
[, 5] drat Rear axle ratio
[, 6] wt   Weight (1000 lbs)
```

```
[, 7] qsec 1/4 mile time
[, 8] gear Number of forward gears
[, 9] carb Number of carburetors
```

L'objectif est de prédire la consommation en fonction des autres variables. On ajuste un modèle complet.

```
lm(formula = mpg ~ ., data = Mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0230 -1.6874 -0.4109  0.9640  5.4400

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.88964   17.81996   1.004   0.3259
cyl         -0.41460    0.95765  -0.433   0.6691
disp         0.01293    0.01758   0.736   0.4694
hp          -0.02085    0.02072  -1.006   0.3248
drat         1.10110    1.59806   0.689   0.4977
wt          -3.92065    1.86174  -2.106   0.0463 *
qsec         0.54146    0.62122   0.872   0.3924
gear         1.23321    1.40238   0.879   0.3883
carb        -0.25510    0.81563  -0.313   0.7573
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.622 on 23 degrees of freedom
Multiple R-squared:  0.8596, Adjusted R-squared:  0.8107
F-statistic:  17.6 on 8 and 23 DF,  p-value: 4.226e-08
```

1) Quel est le pourcentage de variance de la consommation expliqué par ce modèle ?
2) A quel test correspond la dernière ligne du tableau ? Quelle est votre conclusion ?
3) Que pouvez-vous dire sur la pertinence de ce modèle ?
4) Quelle est la consommation prédite par le modèle pour une Mazda 4xR ayant les caractéristiques suivantes (justifier votre réponse) :

mpg	cyl	disp	hp	drat	wt	qsec	gear	carb
21.0	6	160.0	110	3.90	2.62	16.46	4	4

5) Calculer le résidu quadratique.