Examen Régression Linéaire, rattrapage de M. Abdi Aoures (GIS3)

Cristian Preda

25/6/2018

Pas de documents autorisés. Calculette autorisée. Une feuille A4 recto_verso avec des formules autorisée.

Temps de travail : 2h

Exercice 1 (7p)

Soit (X, Y) un couple de variables aléatoires avec la distriburion de probabilité jointe donnée par :

		X				
		0	1	2	3	
Y	1	.08	.10	.10	.02	.30
	2	.08	.05	.22	.05	.40
	3	.04	.04	.04	.18	.30
		.20	.19	.36	.25	1.0

Figure 1: Distribution jointe de X et Y

On demande :

1. Sont X et Y indépendentes ?
2. Tracer la fonction de régression $r(x) = \mathbb{E}(Y|X = x)$. Quelle est la valeur moyenne de Y prédite par cette fonction pour $X = 2$?
3. On approche $r(x)$ par une expression linéaire,

$$r_{lin}(x) = \beta_0 + \beta_1 x,$$

avec $\beta_1 = \dfrac{Cov(X,Y)}{V(X)}$ et $\beta_0 = \mathbb{E}(Y) - \beta_1 \mathbb{E}(X)$. Tracer r_{lin} sur le même graphe que r et donner la valeur moyenne de Y prédite par r_{lin} pour $x = 2$.
4. Laquelle des deux prédictions données en 2. et 3. garderiez vous ? Justifier.

Exercice 2 (3p)

On réalise une régression linéaire simple entre les variables X et Y (Y est la variable à expliquer) à partir d'un échantillon de taille n, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. On note avec β_0 et β_1 le coefficients du modèle linéaire expliquant Y en fonction de X. Précisez la bonne réponse aux questions suivantes. L'ajustement linéaire est fait en minimisant les moindres carrés.

1. La somme des résidus calculés vaut?

   A. 0 ; B. approximativement 0 ; C. Parfois 0.

2. Dans le cas où (X,Y) est un vecteur distribué selon une loi normale bivariée, y-a-t-il une difference entre les estimateurs des coefficients β_0 et β_1 par la technique de moindres carrés et ceux obtenus par maximum de vraisemblance ?

   A. Oui ; B. Non ; C. Pas toujours, cela depend de la loi des résidus.

Exercice 3 (5p)

Soit (X, Y) un couple de variables aléatoires quantitatives à valeurs $\mathbb{R}^2$. On dispose d'un échantillon i.i.d. de taille $n > 3$ de (X, Y), $E = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

On considère le modèle de régression suivant :

$$Y|_{X=x} = \beta x + \varepsilon,$$

où $\beta \in \mathbb{R}$ est le coefficient de régression et ε l'erreur d'ajustement (le résidu) telle que $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. On considère que ε est indépendant de Y.

1. Déterminer l'estimateur de moindres carrés de β à partir de l'échantillon $E = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

2 Application. Deux observateurs pèsent chacun 5 objets différents et obtiennent les mesures suivantes (notons avec Y les mesures du premier observateur et avec X ceux du deuxième observateur) :

$$E = \{(2.0; 2.2); (1.0; 0.8); (4.2; 3.8); (3.2; 2.5); (6.0; 6.0)\}.$$

On pense utiliser le modèle de régression précédent pour étudier la concordance entre les deux observateurs.

- Tracer graphiquement le nuage des points E.

- Tracer les deux droites de régression obtenues avec les moindres carrées et par le maximum de vraissemblance.

- Quelle devrait etre la valeur de β pour que les deux observateurs concordent (donnent les memes mesures) ?

- En utilisant l'estimation donnée en 1.), proposer une statistique de test pour vérifier l'hypothèse $\beta = 1$ versus $\beta \neq 1$.

Exercice 4 (5p).

Les données suivantes représentent le temps passé par n=28 personnes dans diverses activités : PROFession, TRANsport, MENAge, avec les ENFAnts, faire les COURses, pour faire sa TOILette, pour le REPAs, SOMMeil, TELEvision et pour les LOISirs. Ces variables sont quantitatives et exprimées dans une certaine unité de temps.

Commenter les résulatats de cette analyse statistique :

```
d =read.table("http://math.univ-lille1.fr/~preda/GIS4/temps.csv", header=TRUE, sep=",", row.names=1)

# on garde que les variables d'interet
d = d[, 1:10]

head(d)
```

```
     PROF TRAN MENA ENFA COUR TOIL REPA SOMM TELE LOIS
haus  610  140   60   10  120   95  115  760  175  315
faus  475   90  250   30  140  120  100  775  115  325
fnau   10    0  495  110  170  110  130  785  160  130
hmus  615  141   65   10  115   90  115  765  180  305
fmus  179   29  421   87  161  112  119  776  143  373
hcus  585  115   50    0  150  105  100  760  150  385
```

```
# quelques analyses
summary(d)
```

```
      PROF            TRAN            MENA            ENFA
 Min.   : 10.0   Min.   :  0.00   Min.   : 50.0   Min.   :  0.00
 1st Qu.:356.8   1st Qu.: 47.50   1st Qu.: 96.5   1st Qu.: 10.00
 Median :535.0   Median : 95.50   Median :256.0   Median : 22.00
 Mean   :448.9   Mean   : 86.07   Mean   :277.0   Mean   : 33.32
 3rd Qu.:630.8   3rd Qu.:127.00   3rd Qu.:423.5   3rd Qu.: 56.00
 Max.   :655.0   Max.   :148.00   Max.   :710.0   Max.   :110.00
      COUR            TOIL            REPA            SOMM
 Min.   : 52.0   Min.   : 77.00   Min.   : 85.0   Min.   :745.0
 1st Qu.: 85.0   1st Qu.: 89.50   1st Qu.:100.0   1st Qu.:761.5
 Median :112.0   Median : 92.00   Median :110.0   Median :775.0
 Mean   :108.7   Mean   : 94.86   Mean   :118.1   Mean   :785.6
 3rd Qu.:131.0   3rd Qu.: 96.25   3rd Qu.:132.5   3rd Qu.:808.2
 Max.   :170.0   Max.   :130.00   Max.   :180.0   Max.   :848.0
      TELE            LOIS
 Min.   : 40.00   Min.   :130.0
 1st Qu.: 64.75   1st Qu.:308.8
 Median : 91.50   Median :346.5
 Mean   : 99.43   Mean   :338.4
 3rd Qu.:122.75   3rd Qu.:385.5
 Max.   :180.00   Max.   :475.0
```

```
round(cor(d),3)
```

```
       PROF   TRAN   MENA   ENFA   COUR   TOIL   REPA   SOMM   TELE   LOIS
PROF  1.000  0.939 -0.906 -0.865 -0.654 -0.112 -0.461 -0.558 -0.056  0.074
TRAN  0.939  1.000 -0.870 -0.810 -0.503 -0.077 -0.610 -0.705 -0.041  0.126
MENA -0.906 -0.870  1.000  0.861  0.500 -0.040  0.358  0.438 -0.206 -0.212
ENFA -0.865 -0.810  0.861  1.000  0.542  0.118  0.364  0.281  0.122 -0.426
COUR -0.654 -0.503  0.500  0.542  1.000  0.591 -0.183 -0.022  0.219 -0.074
TOIL -0.112 -0.077 -0.040  0.118  0.591  1.000 -0.353 -0.211  0.325 -0.141
REPA -0.461 -0.610  0.358  0.364 -0.183 -0.353  1.000  0.818  0.318 -0.020
SOMM -0.558 -0.705  0.438  0.281 -0.022 -0.211  0.818  1.000  0.020  0.236
TELE -0.056 -0.041 -0.206  0.122  0.219  0.325  0.318  0.020  1.000 -0.288
LOIS  0.074  0.126 -0.212 -0.426 -0.074 -0.141 -0.020  0.236 -0.288  1.000
```

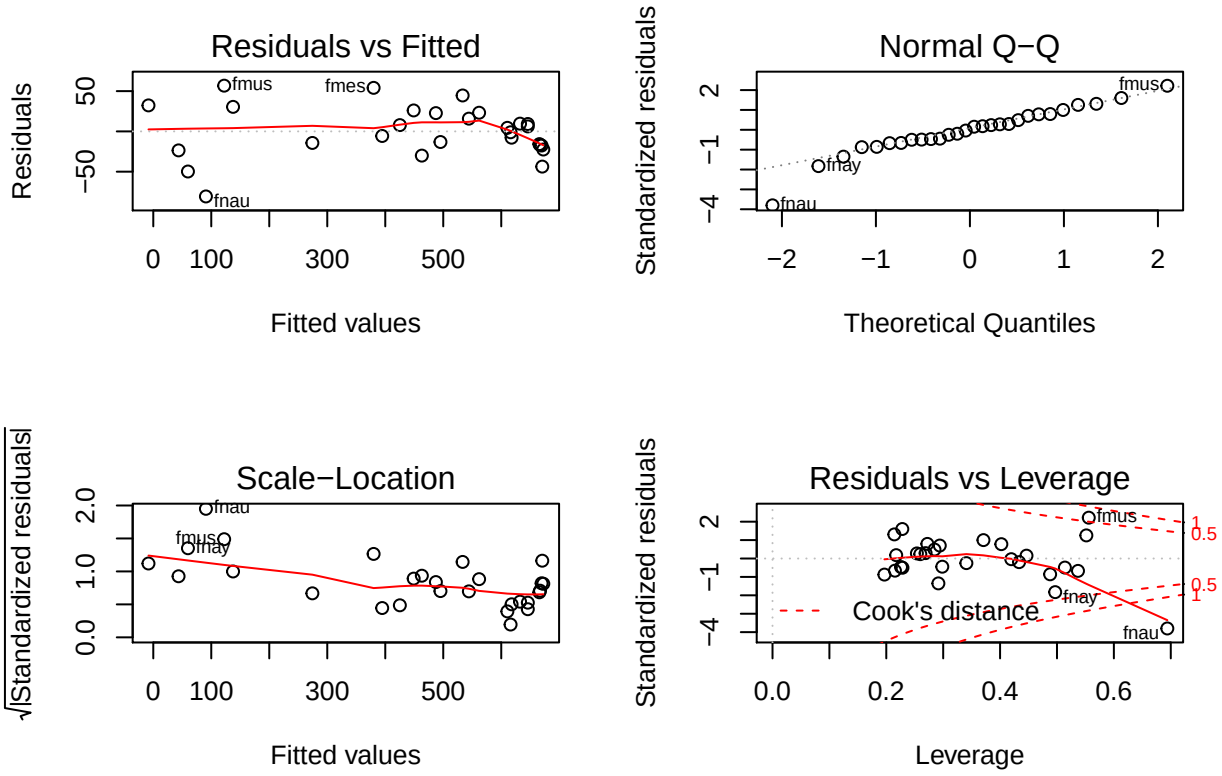
```
#un modèle expliquant le temps passé au travail en fonction des autres variables

m0 = lm(PROF~., data = d)
summary(m0)
```

```

Call:
lm(formula = PROF ~ ., data = d)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-80.789 -17.079   1.662  22.806  56.760

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1936.0293   671.9613   2.881  0.00994 **
TRAN           0.7454     0.8610   0.866  0.39807
MENA          -0.3921     0.1387  -2.827  0.01118 *
ENFA          -2.3741     0.7490  -3.170  0.00530 **
COUR          -1.7578     0.7224  -2.433  0.02562 *
TOIL           0.9179     1.0922   0.840  0.41172
REPA           0.1227     1.2249   0.100  0.92133
SOMM          -1.3635     0.7802  -1.748  0.09756 .
TELE          -0.4910     0.5055  -0.971  0.34429
LOIS          -0.4556     0.1526  -2.986  0.00792 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.43 on 18 degrees of freedom
Multiple R-squared:  0.9809,    Adjusted R-squared:  0.9713
F-statistic: 102.7 on 9 and 18 DF,  p-value: 1.295e-13
```

```r
par(mfrow=c(2,2))
plot(m0)

library(lmtest)
```

```
Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```




```
shapiro.test(m0$residuals)
```

```
    Shapiro-Wilk normality test

data:  m0$residuals
W = 0.98094, p-value = 0.8723
```

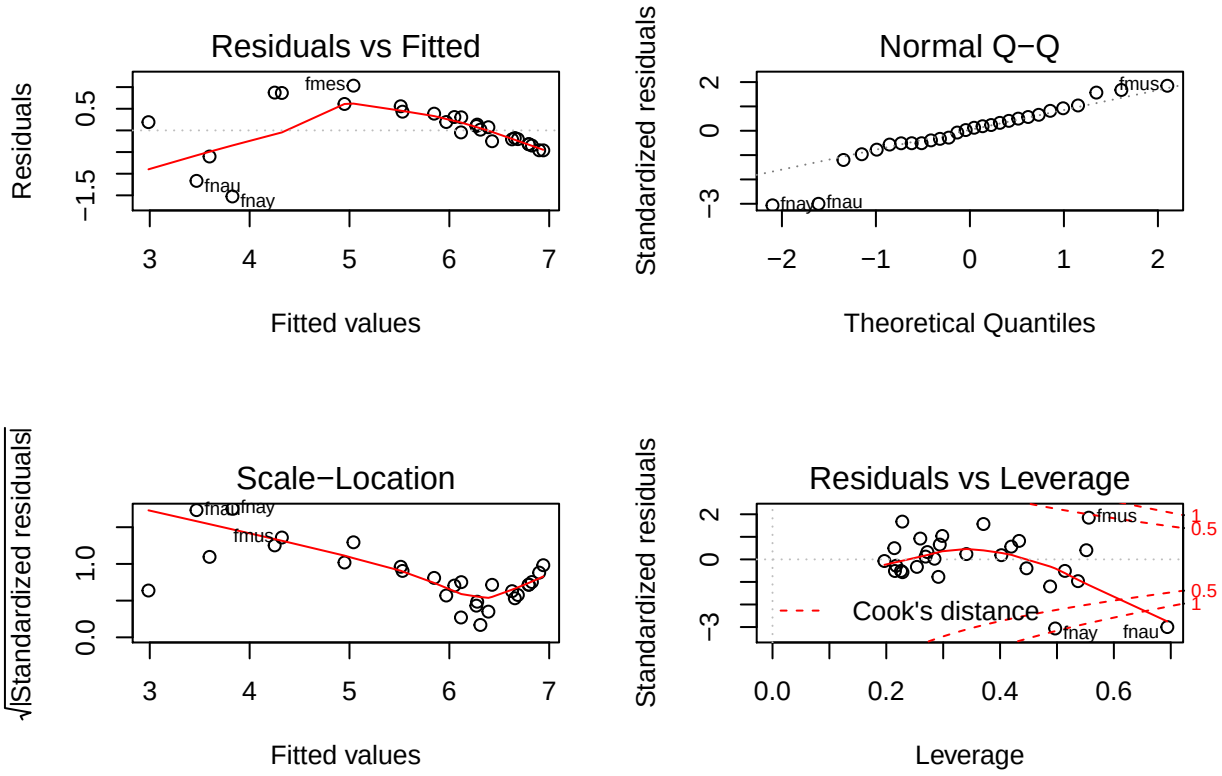
```
bptest(m0)
```

```
    studentized Breusch-Pagan test

data:  m0
BP = 20.269, df = 9, p-value = 0.01633
```

```
dwtest(m0)
```

```
    Durbin-Watson test

data:  m0
DW = 2.3091, p-value = 0.7838
alternative hypothesis: true autocorrelation is greater than 0
```

```
m1=lm(log(PROF)~., data =d)
par(mfrow=c(2,2))
plot(m1)
```




```
shapiro.test(m1$residuals)
```

```
    Shapiro-Wilk normality test

data:  m1$residuals
W = 0.96515, p-value = 0.4582
```

```
bptest(m1)
```

```
    studentized Breusch-Pagan test

data:  m1
BP = 14.417, df = 9, p-value = 0.1082
```

```
dwtest(m1)
```

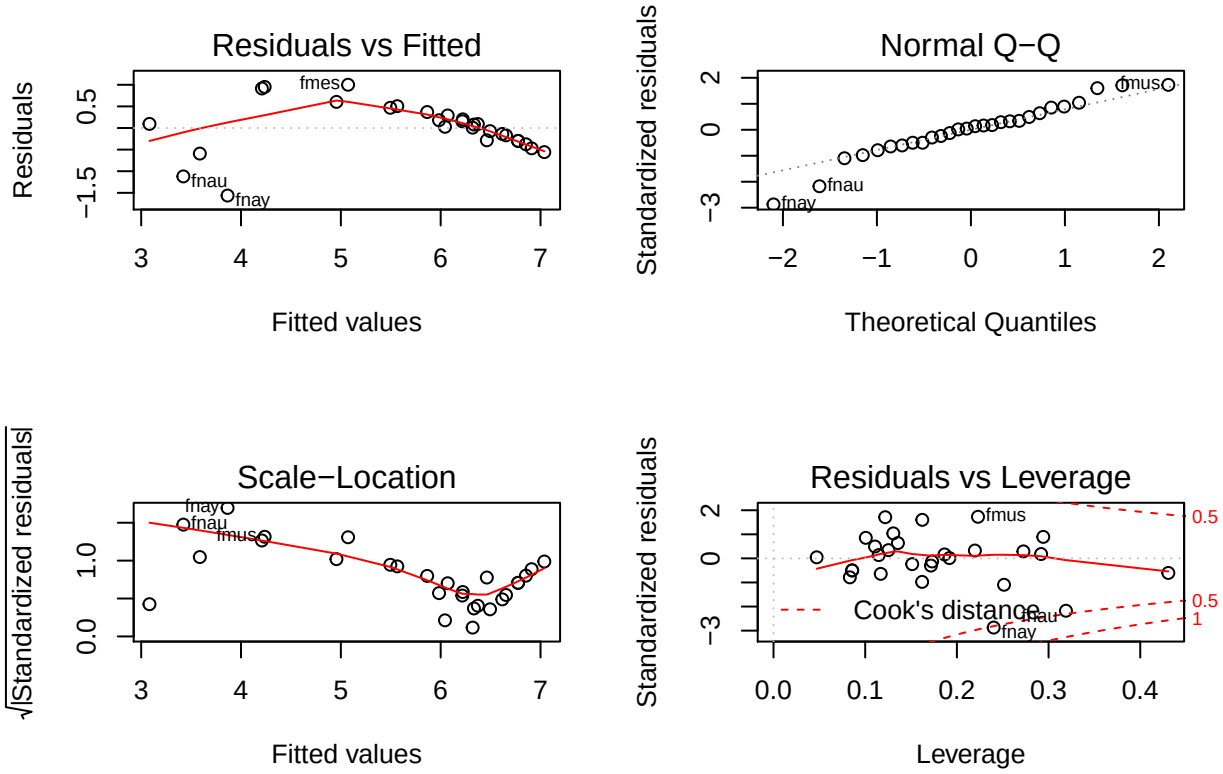
```
    Durbin-Watson test

data:  m1
DW = 2.3621, p-value = 0.8261
alternative hypothesis: true autocorrelation is greater than 0
```

```
library(MASS)
m2 = stepAIC(m1, method="stepwise", trace=F)
summary(m2)
```

```
Call:
lm(formula = log(PROF) ~ ENFA + COUR + TOIL + SOMM, data = d)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.56286 -0.29811  0.05379  0.30950  1.00045

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.985578   3.788783   4.747 8.74e-05 ***
ENFA        -0.019626   0.005147  -3.813 0.000894 ***
COUR        -0.019174   0.005680  -3.376 0.002606 **
TOIL         0.026243   0.013664   1.921 0.067264 .
SOMM        -0.015327   0.004386  -3.494 0.001956 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.624 on 23 degrees of freedom
Multiple R-squared:  0.8061,    Adjusted R-squared:  0.7724
F-statistic: 23.91 on 4 and 23 DF,  p-value: 6.578e-08
```

```
par(mfrow=c(2,2))
plot(m2)
```




```
shapiro.test(m2$residuals)
```

```

    Shapiro-Wilk normality test

data:  m2$residuals
W = 0.96197, p-value = 0.3879
```

```r
bptest(m2)
```

```
    studentized Breusch-Pagan test

data:  m2
BP = 11.47, df = 4, p-value = 0.02176
```

```r
dwtest(m2)
```

```
    Durbin-Watson test

data:  m2
DW = 2.3282, p-value = 0.8047
alternative hypothesis: true autocorrelation is greater than 0
```