Examen Régression Linéaire (GIS2A4)

Cristian Preda

17/12/2018

Tous documents autorisés. Calculette autorisée.

Temps de travail : 2h

Exercice 1 (5p)

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

$$f(x,y) = \begin{cases} e^{-y} & \text{si } 0 < x < y < \infty \\ \\ 0 & \text{sinon.} \end{cases}$$

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 $X = 2$?

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. Nous recevons une nouvelle observation x_{n+1} et nous calculons la prévision correspondante, $\hat{y}_{n+1}$. La variance de la valeur prévue est minimale lorsque

   A. $x_{n+1} = 0$; B. $x_{n+1} = \overline{x}$; C. aucun rapport avec x_{n+1}.

2. La somme des résidus est ?

   A. négative ; B. positive ; C. nulle.

Exercice 3 (5p)

En juin 2018, on a relevé dans les petites annonces les superficies (en m2) et les prix (en euros) de 108 appartements de type T3 à louer sur l'agglomération de Lille (cf. Figure 1. ci-dessous).

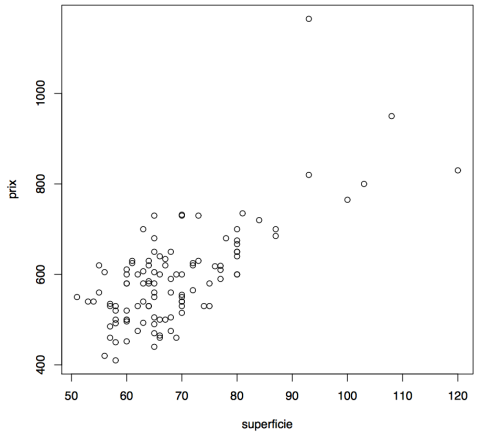


Figure 1: Nuage de points (prix vs superficie)

On dispose de la sortie R suivante obtenue lors d'un ajustement linéaire du prix en fonction de la superficie.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 134.3450    45.4737   2.954  0.00386
Superficie    6.6570     0.6525  10.203  < 2e-16

Residual standard error: 77.93 on 106 degrees of freedom
Multiple R-Squared: 0.4955,     Adjusted R-squared: 0.4907
F-statistic: 104.1 on 1 and 106 DF,  p-value: < 2.2e-16
```

Figure 2: Sortie R

1. Donner une estimation du coefficient de corrélation linéaire entre le prix et la superficie ?
2. Proposer un modèle permettant d'étudier la relation entre le prix des appartements et leur superficie. Préciser les hypothèses de ce modèle.
3. D'après la sortie R, est-ce que la superficie joue un rôle sur le prix des appartements ? Considérez-vous ce rôle comme important ?
4. Quelle est l'estimation du coefficient de la superficie dans le modèle ? Comment interprétez-vous ce coefficient ?
5. Comment interprétez-vous l'intercept du modèle ?

Exercice 4 (7p)

On souhaite étudier la relation entre le prix d'une voiture, son poids et sa puissance (cylindrée). Commenter la sortie R ci-dessous réalisée par un étudiant GIS2A4 et précisez quelles sont vos recommandations pour améliorer cette analyse.

Chargement de la base de données dans R:

```
d = read.table("http://math.univ-lille1.fr/~preda/GIS4/car.txt", header=TRUE, sep="\t", row.names=1)
d=d[, c(1,5,9)]
str(d)
```

```
'data.frame':   18 obs. of  3 variables:
 $ CYL  : int  1350 1588 1294 1222 1585 1297 1796 1565 2664 1166 ...
 $ POIDS: int  870 1110 1050 930 1105 1080 1160 1010 1320 815 ...
 $ PRIX : int  30570 39990 29600 28250 34900 35480 32300 32000 47700 26540 ...
```

```
print(d)
```

```
                  CYL POIDS  PRIX
ALFASUD-TI-1350  1350   870 30570
AUDI-100-L       1588  1110 39990
SIMCA-1307-GLS   1294  1050 29600
CITROEN-GS-CLUB  1222   930 28250
FIAT-132-1600GLS 1585  1105 34900
LANCIA-BETA-1300 1297  1080 35480
PEUGEOT-504      1796  1160 32300
RENAULT-16-TL    1565  1010 32000
RENAULT-30-TS    2664  1320 47700
TOYOTA-COROLLA   1166   815 26540
ALFETTA-1.66     1570  1060 42395
PRINCESS-1800-HL 1798  1160 33990
DATSUN-200L      1998  1370 43980
TAUNUS-2000-GL   1993  1080 35010
RANCHO           1442  1129 39450
MAZDA-9295       1769  1095 27900
OPEL-REKORD-L    1979  1120 32700
LADA-1300        1294   955 22100
```

```
summary(d)
```

```
      CYL           POIDS           PRIX
 Min.   :1166   Min.   : 815   Min.   :22100
 1st Qu.:1310   1st Qu.:1020   1st Qu.:29842
 Median :1578   Median :1088   Median :33345
 Mean   :1632   Mean   :1079   Mean   :34159
 3rd Qu.:1798   3rd Qu.:1127   3rd Qu.:38458
 Max.   :2664   Max.   :1370   Max.   :47700
```

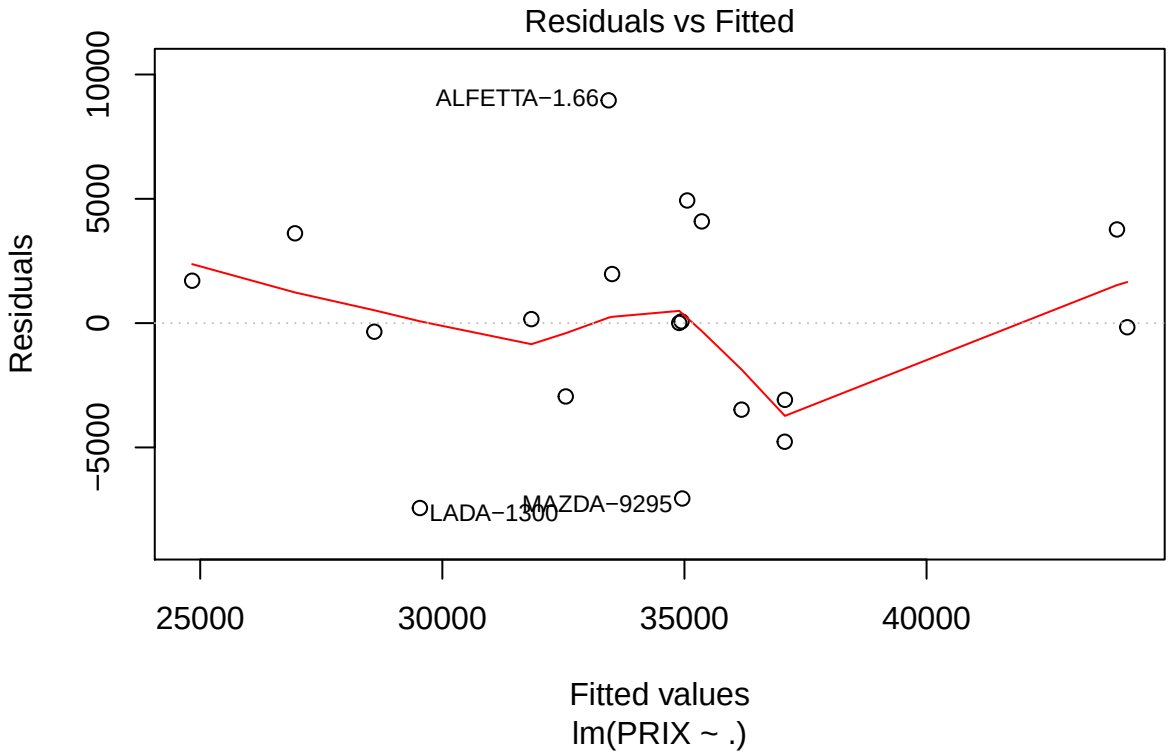
Le modèle :

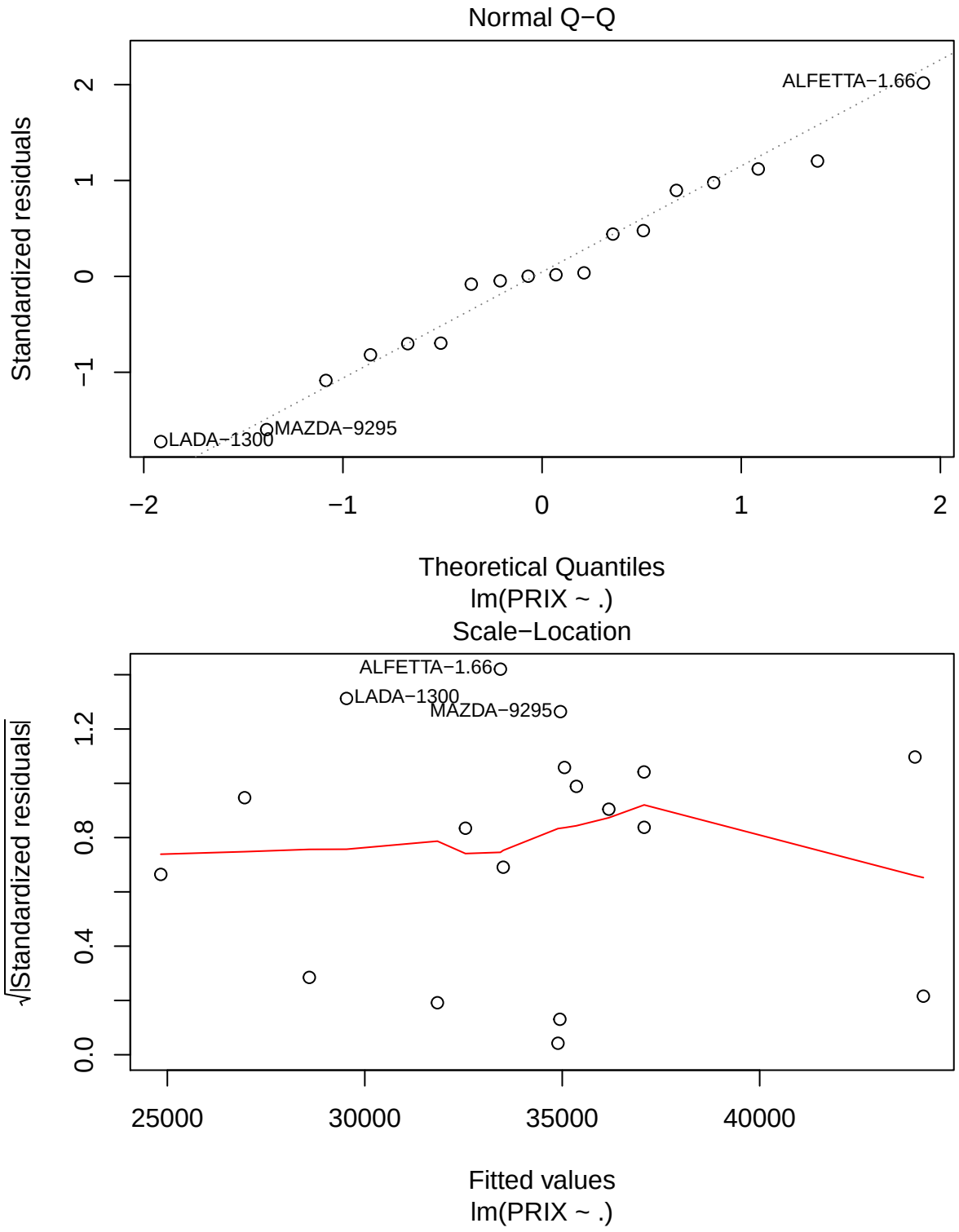
```
m0 = lm(PRIX~., data = d)
summary(m0)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-7436.5 -3050.8    38.9  3203.4  8960.6
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3409.494   9416.603  -0.362   0.7223
CYL             2.061      4.828   0.427   0.6756
POIDS          31.706     13.181   2.405   0.0295 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4573 on 15 degrees of freedom
Multiple R-squared:  0.5726,    Adjusted R-squared:  0.5157
F-statistic: 10.05 on 2 and 15 DF,  p-value: 0.001702
```

```
plot(m0)
```



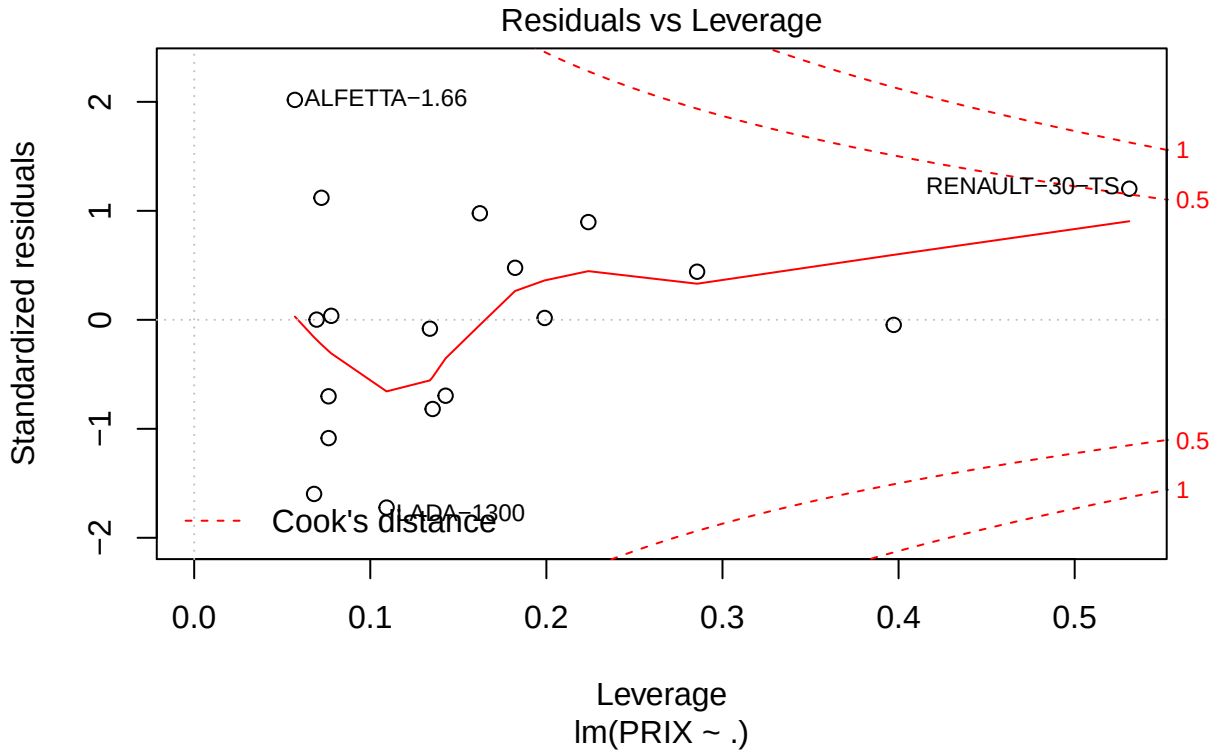



```
library(lmtest)
```

```
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 = 0.97423, p-value = 0.8724
```

```
bptest(m0)
```

```
    studentized Breusch-Pagan test

data:  m0
BP = 0.066819, df = 2, p-value = 0.9671
```

```
dwtest(m0)
```

```
    Durbin-Watson test

data:  m0
DW = 1.7111, p-value = 0.2754
alternative hypothesis: true autocorrelation is greater than 0
```

```
print(influence.measures(m0))
```

```
Influence measures of
     lm(formula = PRIX ~ ., data = d) :
```

	dfb.1_	dfb.CYL	dfb.POID	dffit	cov.r	cook.d	hat	inf
ALFASUD-TI-1350	4.38e-01	0.179436	-0.371260	0.478225	1.343	7.73e-02	0.2238	
AUDI-100-L	-7.71e-02	-0.137241	0.148070	0.315471	1.020	3.26e-02	0.0723	
SIMCA-1307-GLS	2.13e-02	0.214854	-0.146335	-0.279097	1.300	2.69e-02	0.1428	
CITROEN-GS-CLUB	-2.01e-02	0.007924	0.007397	-0.030836	1.418	3.39e-04	0.1339	
FIAT-132-1600GLS	-9.79e-05	-0.000195	0.000205	0.000473	1.322	7.98e-08	0.0695	
LANCIA-BETA-1300	-5.20e-02	-0.182885	0.144939	0.219353	1.437	1.69e-02	0.1822	
PEUGEOT-504	1.25e-01	0.012601	-0.110326	-0.314037	1.042	3.25e-02	0.0763	
RENAULT-16-TL	6.30e-03	0.003181	-0.005276	0.010304	1.333	3.79e-05	0.0778	
RENAULT-30-TS	-3.16e-01	0.966517	-0.294025	1.301006	1.934	5.46e-01	0.5310	*
TOYOTA-COROLLA	2.53e-01	0.055017	-0.189181	0.271360	1.655	2.60e-02	0.2856	*
ALFETTA-1.66	1.18e-01	-0.052384	-0.006872	0.562342	0.505	8.23e-02	0.0572	
PRINCESS-1800-HL	7.86e-02	0.006439	-0.068435	-0.198143	1.205	1.36e-02	0.0763	
DATSUN-200L	3.18e-02	0.015954	-0.030944	-0.036518	2.040	4.76e-04	0.3972	*
TAUNUS-2000-GL	2.94e-03	0.006977	-0.005481	0.008218	1.535	2.41e-05	0.1990	
RANCHO	-1.78e-01	-0.335076	0.322558	0.429417	1.205	6.17e-02	0.1622	
MAZDA-9295	-6.81e-02	-0.189940	0.118975	-0.458014	0.754	6.22e-02	0.0681	
OPEL-REKORD-L	-4.89e-02	-0.237482	0.148401	-0.320074	1.240	3.50e-02	0.1354	
LADA-1300	-4.01e-01	0.147422	0.149021	-0.650850	0.713	1.21e-01	0.1093	