Partiel ENSIIE printemps 2011

Nom : **Signature :**

Prénom :

La qualité de la présentation sera prise en compte dans la notation. Aucun document.

Exercice 1 (8 points)

Considérons des données de qualité de l'air mesurées à New York en mai 1973 :

```
> data(airquality)
> data73 <- airquality
```

```
New York Air Quality Measurements

Description:

     Daily air quality measurements in New York, May to September 1973.

Usage:

     data(airquality)

Format:

     A data frame with 154 observations on 6 variables.

       `[,1]'  `Ozone'    numeric  Ozone (ppb)
       `[,2]'  `Solar.R'  numeric  Solar R (lang)
       `[,3]'  `Wind'     numeric  Wind (mph)
       `[,4]'  `Temp'     numeric  Temperature (degrees F)
       `[,5]'  `Month'    numeric  Month (1-12)
       `[,6]'  `Day'      numeric  Day of month (1-31)

Details:

     Daily readings of the following air quality values for May 1, 1973
     (a Tuesday) to September 30, 1973.

        * `Ozone': Mean ozone in parts per billion from 1300 to 1500
          hours at Roosevelt Island

        * `Solar.R': Solar radiation in Langleys in the frequency band
          4000-7700 Angstroms from 0800 to 1200 hours at Central Park

        * `Wind': Average wind speed in miles per hour at 0700 and
          1000 hours at LaGuardia Airport

        * `Temp': Maximum daily temperature in degrees Fahrenheit at
          La Guardia Airport.
```

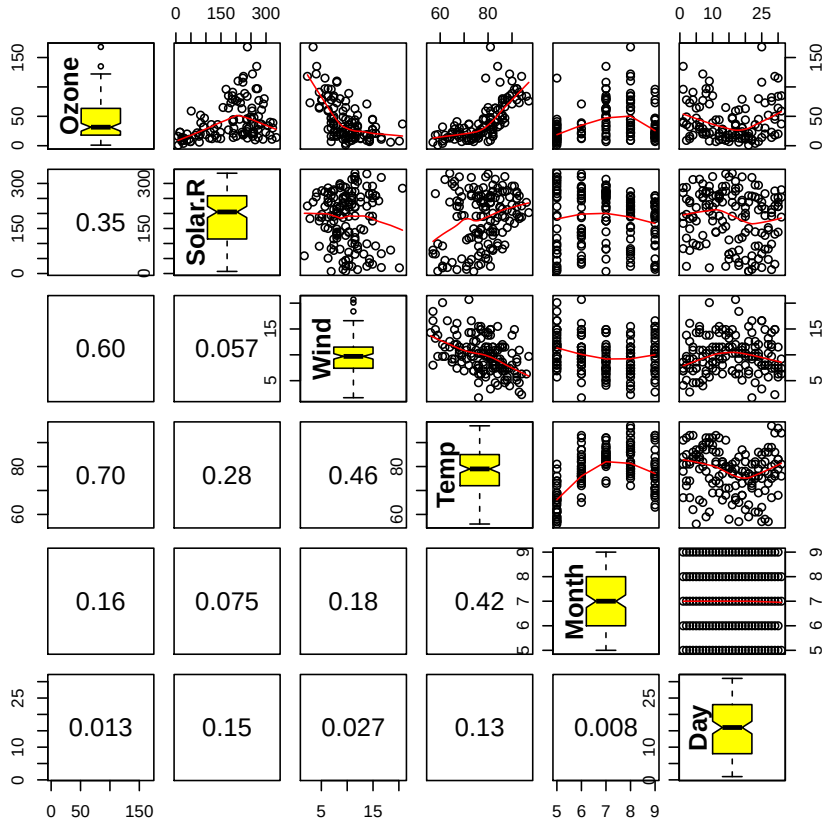

FIGURE 1 – Affichage des valeurs absolues des corrélations (triangle inférieur), des graphiqes bivariables (triangle supérieur), et des boites à moustaches (diagonale)

Un institut de surveillance de la qualité de l'air désire expliquer la concentration d'ozone en fonction des autres variables disponibles en utilisant une régression linéaire.

Etude préalable

1. Faites un commentaire sur les corrélations de la figure 1.
2. Faites un commentaire les boîtes à moustaches de la figure 1.
3. Faites un commentaire sur les graphiques bivariables de la figure 1.

Estimation

1. L'institut teste 3 modèles :
   (a)
```
> summary(modeleA <- lm(Ozone ~ ., data = data73))

Call:
lm(formula = Ozone ~ ., data = data73)

Residuals:
    Min      1Q  Median      3Q     Max
-37.014 -12.284  -3.302   8.454  95.348

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.11632   23.48249  -2.730  0.00742 **
Solar.R       0.05027    0.02342   2.147  0.03411 *
Wind         -3.31844    0.64451  -5.149 1.23e-06 ***
Temp          1.89579    0.27389   6.922 3.66e-10 ***
Month        -3.03996    1.51346  -2.009  0.04714 *
Day           0.27388    0.22967   1.192  0.23576
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.86 on 105 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared: 0.6249,        Adjusted R-squared: 0.6071
F-statistic: 34.99 on 5 and 105 DF,  p-value: < 2.2e-16
```
   (b)
```
> summary(modeleB <- lm(Ozone ~ Solar.R + Wind + Temp,
+     data = data73))

Call:
lm(formula = Ozone ~ Solar.R + Wind + Temp, data = data73)

Residuals:
    Min      1Q  Median      3Q     Max
-40.485 -14.219  -3.551  10.097  95.619

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Solar.R       0.05982    0.02319   2.580  0.01124 *
Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
Temp          1.65209    0.25353   6.516 2.42e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.18 on 107 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared: 0.6059,        Adjusted R-squared: 0.5948
F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```
   (c)
```
> summary(modeleC <- lm(Ozone ~ Solar.R + I((Wind)^(1/20)) +
+     I(log(Temp)), data = data73))
```

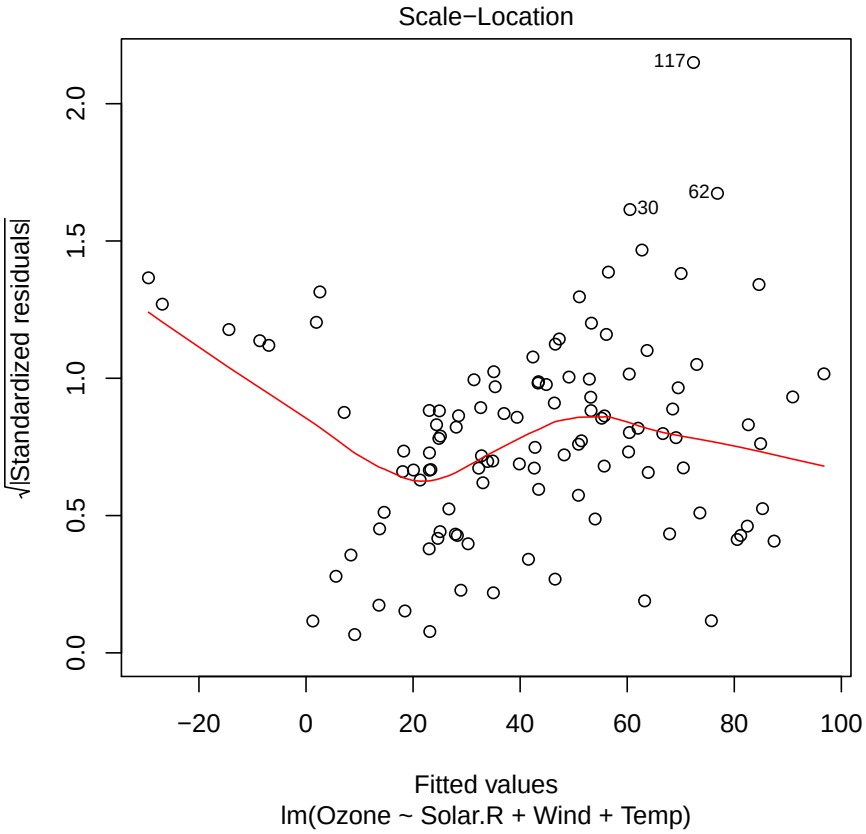

FIGURE 2 – Racine carré de la valeur absolue des résidus en fonction de la prédiction

```
Call:
lm(formula = Ozone ~ Solar.R + I((Wind)^(1/20)) + I(log(Temp)),
    data = data73)

Residuals:
    Min      1Q  Median      3Q     Max
-39.275 -13.679  -2.543  11.375  79.199

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       359.77838  168.08967   2.140  0.03459 *
Solar.R             0.06101    0.02212   2.758  0.00683 **
I((Wind)^(1/20)) -700.45950  102.01120  -6.866 4.49e-10 ***
I(log(Temp))      104.48626   18.40468   5.677 1.19e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.17 on 107 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared: 0.6427,        Adjusted R-squared: 0.6327
F-statistic: 64.16 on 3 and 107 DF,  p-value: < 2.2e-16
```

2. Quel critère maximise la régression linéaire ?
3. Quelles hypothèses statistiques sont utilisées en regression linéaire ?
4. Quel modèle vous semble le plus adapté ? Justifiez votre réponse.
5. Quelles variables vous semblent significatives ? Pourquoi ?
6. Qels problèmes détectez vous sur la figure 2 ?
7. Ecrivez une interprétation (simple) de la regression retenue.

Prédiction

L'institut dispose d'un second jeu de données de format identique au précédent mais relatif à des données mesurées en 1974

1. Ecriver une fonction R (ErreurDePrediction(modele73,data74)) qui calcule l'erreur de prédiction du modèle estimé en 1973 sur les données de 1974 [1].

Exercice 2 (8 points)

Considérons le jeu de données `babyfood` qui est une étude sur le développement de maladie respiratoire dans la première année de vie d'un enfant en fonction de son sexe et son mode de nutrition.

```
> library(faraway)
> data(babyfood)
> xtabs(disease/(disease + nondisease) ~ sex + food,
+     babyfood)

      food
sex        Bottle     Breast      Suppl
  Boy  0.16812227 0.09514170 0.12925170
  Girl 0.12500000 0.06681034 0.12598425

> babyfood

  disease nondisease  sex   food
1      77        381  Boy Bottle
2      19        128  Boy  Suppl
3      47        447  Boy Breast
4      48        336 Girl Bottle
5      16        111 Girl  Suppl
6      31        433 Girl Breast
```

1. Quel modèle est estimé par le code suivant ?

   ```
   > mdl <- glm(cbind(disease, nondisease) ~ sex + food,
   +     family = binomial, babyfood)
   ```

2. Donner la forme de la vraisemblance du vecteur de paramètre.
3. Commenter les statistiques suivantes ?

   ```
   > summary(mdl)

   Call:
   glm(formula = cbind(disease, nondisease) ~ sex + food, family = binomial,
       data = babyfood)

   Deviance Residuals:
         1        2        3        4        5        6
    0.1096  -0.5052   0.1922  -0.1342   0.5896  -0.2284

   Coefficients:
               Estimate Std. Error z value Pr(>|z|)
   (Intercept)  -1.6127     0.1124 -14.347  < 2e-16 ***
   sexGirl      -0.3126     0.1410  -2.216   0.0267 *
   foodBreast   -0.6693     0.1530  -4.374 1.22e-05 ***
   foodSuppl    -0.1725     0.2056  -0.839   0.4013
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   (Dispersion parameter for binomial family taken to be 1)

       Null deviance: 26.37529  on 5  degrees of freedom
   ```

1. La fonction R `predict(model,newdata`permet d'obtenir les prédictions utilisant les variables de newdata et un modèle model disponible

```
Residual deviance:  0.72192  on 2  degrees of freedom
AIC: 40.24

Number of Fisher Scoring iterations: 4
```

4. Donner et commenter l'effet de l'allaitement.
5. Calculer un intervalle de confiance sur l'effet de l'allaitement (le quantile d'ordre 0.975 d'une loi normale centrée réduite est approximativement 1.96).
6. Quel test réaliser pour déterminer l'intérêt du modèle ?
7. A l'aide du résultat suivant justifiez votre opinion sur l'intérêt du modèle.

```
> pchisq(deviance(mdl), df.residual(mdl), lower = FALSE)

[1] 0.6970062
```

8.

Exercice 3 (4 points)

Questions de cours :

1. Quels sont les caractéristiques d'un modèle linéaire généralisé ?
2. Quel est l'intérêt de la fonction de lien. Donner un exemple.
3. Qu'est ce qu'un modèle additif ?
4. Citer et expliquer brièvement deux types de régression non paramètrique.