Partiel TIAD - Master 1 IM
Mardi 10 Mars 2015

Durée 2 h. Calculatrices autorisées. Seul un formulaire sur feuille double est autorisé.

Tran Viet Chi, `chi.tran@univ-lille1.fr`, bureau 316 (bâtiment M3).

Exercice 1 (Pression et température du mercure)
Nous analysons des données pour obtenir des relations entre la température en degrés Celsius et la pression en millimètre de mercure du mercure gazeux. La base de données, sous **R**, s'appelle `pressure`. Elle contient deux variables, T (`temperature`) et P (`pressure`).

1. On cherche graphiquement une relation entre la pression et la température. On trace successivement les graphiques de la Figure 1. Commenter.

(a) (b) (c)

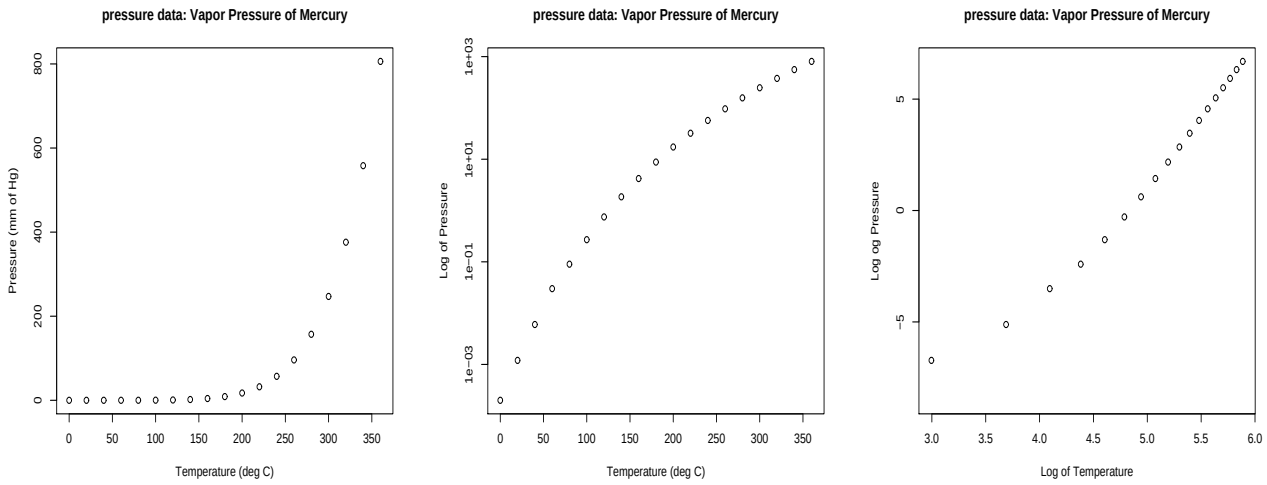

FIGURE 1 – *(a) : Pression en fonction de la température. (b) : Log de la pression en fonction de la température. (c) : Log de la pression en fonction du log de la température.*

2. Afin de quantifier la relation entre pression et température, on réalise les deux régressions linéaires `mco` et `mco2` (voir sorties ci-dessous) :

```
> mco<-lm(log(pressure$pressure) ~ pressure$temperature)
> summary(mco)

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -6.068144   0.483831  -12.54 5.10e-10 ***
pressure$temperature  0.039792   0.002296   17.33 3.07e-12 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.096 on 17 degrees of freedom
Multiple R-Squared: 0.9464,     Adjusted R-squared: 0.9433
F-statistic: 300.3 on 1 and 17 DF,  p-value: 3.070e-12

> mco2<-lm(log(pressure$pressure[2:19]) ~ log(pressure$temperature[2:19]))
> summary.lm(mco2)

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -23.5908     1.0742  -21.96 2.25e-13 ***
log(pressure$temperature[2:19])   5.0260     0.2115   23.76 6.62e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6994 on 16 degrees of freedom
Multiple R-Squared: 0.9724,     Adjusted R-squared: 0.9707
F-statistic: 564.5 on 1 and 16 DF,  p-value: 6.624e-14
```

Dans la seconde régression, on n'utilise les données que de 2 à 19, car la donnée numéro 1 correspond à la température 0 (et son log n'est pas défini).

Comment peut-on comparer ces deux régressions ? Laquelle vous semble la meilleure ?

3. Nous obtenons les résultats suivants sous **R** :

```
mean(log(pressure$temperature[2:19]))=5.02
mean(log(pressure$temperature[2:19])^2)=25.78

mean(log(pressure$pressure[2:19]))=1.63
mean(log(pressure$pressure[2:19])^2)=18.43

mean(log(pressure$pressure[2:19])*log(pressure$temperature[2:19]))=11.22
```

Calculer les variances des variables `log(temperature)` et `log(pressure)`. Calculer la covariance de ces deux variables.

4. On s'intéresse à la régression de $\log(P)$ sur $\log(T)$:

$$\log(P) = a\log(T) + b + \varepsilon \tag{1}$$

Donner les expressions littérales des estimateurs MCO $\widehat{a}$ et $\widehat{b}$, puis retrouver à l'aide de la question **3** les résultats de la question **2**. D'où viennent les différences à votre avis ?

5. Les coefficients pour a et b sont-ils significatifs au seuil 5% ? Justifier.

6. Montrer que (1) est équivalent au fait que la pression varie comme une puissance de la température, si l'on omet le bruit ε.

7. On étudie les résidus de la régression (1).

```
res2=residuals(mco2)    mean(res2)=8.83e-17    var(res2)=0.46
```

(On rappelle que la commande `var` de **R** donne la variance corrigée (*i.e.* renormalisée par $n-1$))
Commentez ces résultats ainsi que la Figure 2(a). Tracer ce QQ-plot a-t-il un intérêt pour l'exploitation des résultats de la régression ? si oui, lequel ?

(a) (b)

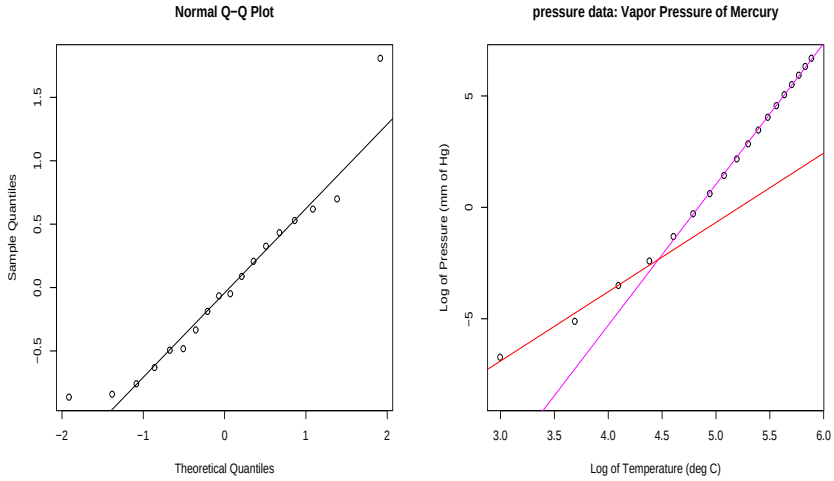

FIGURE 2 – *(a) : QQ-plot des résidus de la régression (1) avec ceux de la loi normale. (b) : Droites de régression de la question 9.*

8. Sur la Figure 2(b), on peut voir que la courbe présente une légère inflexion au niveau de la 5^{e}valeur. On réalise un test de Chow pour savoir s'il y a deux comportements différents.

```
> mcop3<-lm(log(pressure$pressure[6:19]) ~ log(pressure$temperature[6:19]))
> summary.lm(mcop3)

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -30.5862     0.2508  -122.0   <2e-16 ***
log(pressure$temperature[6:19])   6.3237     0.0466   135.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06748 on 12 degrees of freedom
Multiple R-Squared: 0.9993,     Adjusted R-squared: 0.9993
F-statistic: 1.842e+04 on 1 and 12 DF,  p-value: < 2.2e-16

> mcop4<-lm(log(pressure$pressure[2:5]) ~ log(pressure$temperature[2:5]))
> summary.lm(mcop4)

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -16.2226     1.1660  -13.91  0.00513 **
log(pressure$temperature[2:5])   3.1089     0.3048   10.20  0.00947 **
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3173 on 2 degrees of freedom
Multiple R-Squared: 0.9811,     Adjusted R-squared: 0.9717
F-statistic: 104.1 on 1 and 2 DF,  p-value: 0.009474

> resp3=residuals(mcop3)
> resp4=residuals(mcop4)
> var(resp3)=0.004
> var(resp4)=0.067
> qf(0.95,2,14)=3.739
> qf(0.025,12,2)=0.196
> qf(0.975,12,2)=39.415
```

Effectuer le test de Chow. Vérifier les hypothèses du test de Chow.

9. Conclure sur le comportement de la pression en fonction de la température.

Exercice 2 (ACP sur les taux de crime aux US)
La table `Crime`, placée dans une librairie intitulée `malib`, nous fournit les taux de crime en 1977 pour 100 000 individus dans 7 catégories et pour chacun des 50 états américains (ainsi une valeur de 1881.9 pour le nombre de larcins en Alabama signifie par exemple que le nombre total de larcins divisé par le nombre total d'habitants et multiplié par 100 000 est 1881.9). La table est donnée en annexe.

Nous avons réalisé avec la procédure `princomp` une analyse en composante principale pour décrire les données et mieux comprendre la délinquance américaine en 1977. Les résultats se trouvent dans l'annexe.

1. Combien y a-t-il d'individus ? de variables ?

2. Quelles informations nous apportent les statistiques descriptives pour les variables ? Commenter brièvement.

3. Au vu des valeurs propres, justifier pourquoi on peut conserver les 3 premiers axes principaux. Quelle est l'inertie expliquée par ces trois axes ?

4. Lorsque l'on regarde les composantes du vecteur propre associé à la plus grande valeur propre, que peut-on remarquer ? Que traduit cela en général, quant à l'interprétation du premier axe factoriel ?

On se concentre maintenant sur les axes factoriels 2 et 3.

5. Au vu de la matrice de corrélation des composantes principales avec les variables originales, que peut-on dire ?

6. Pour le nuage des individus, les CTR et CO2 sont donnés en annexe. Commenter les axes 2 et 3 en utilisant ces statistiques. Quelle interprétation donner à ces nouveaux axes ?