Statistique avec le logiciel R, Master TIDE
15/01/2014, Durée : deux heures
Documents et téléphones portables ne sont pas autorisés.
L'usage de la calculatrice est autorisé.

Exercice 1. Régression linéaire

Nous disposons d'un jeu de données (X_1, X_2, X_3, Y) qui présente le taux de décès par attaque cardiaque chez les hommes de 55 à 59 ans dans différents pays industrialisés. Les variables sont $Y = 100 \times \log$ (nombre de décès par crise cardiaque pour 100 000 hommes), $X_1 = 1000\times$ téléphones par habitants, $X_2 =$ calories grasses en pourcentage du total des calories et $X_3 =$ calories provenant de protéines animales en pourcentage du total des calories. Le nombre des observations est 22.

Nous considérons le modèle à trois variables explicatives suivant

$$Y = \mu + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

Les résultats d'estimation sont présentés dans le tableau suivant.

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.04472    4.02490   7.465 6.48e-07
x1           0.05184    0.01975   2.624   0.0172
x2           0.21540    0.19594   1.099   0.2861
x3           5.79202    1.05345   5.498 3.20e-05
---

Residual standard error: 4.317 on 18 degrees of freedom
Multiple R-squared:  0.9391,Adjusted R-squared:  0.929
F-statistic: 92.56 on 3 and 18 DF,  p-value: 3.935e-11
```

1. Réécrivez le modèle en utilisant les coefficients estimés.
2. Le modèle est il globalement significatif ?
3. Y a t il une variable qui doit être éliminée ? Laquelle ?
4. Testez l'hypothèse $H_0 : \beta_1 = 0.05$ contre $H_1 : \beta_1 \neq 0.05$.
5. Testez l'hypothèse $H_0 : \beta_3 = 6$ contre $H_1 : \beta_3 < 6$.

Exercice 2. Valeurs extrêmes

Soit $Y_0, Y_1, Y_2, \ldots$ une suite des variables aléatoires i.i.d. avec la fonction de répartition commune

$$F_Y(y) = \exp\left(-\frac{1}{(a+1)y}\right), \quad y > 0,$$

où $a \in [0,1]$ est un paramètre. Définissons le processus $X_0, X_1, X_2, \ldots$ comme suit

$$X_0 = Y_0, \quad X_i = \max(aY_{i-1}, Y_i), \quad i = 1, \ldots, n.$$

1. Montrez que la loi marginale du processus $X_0, X_1, X_2, \ldots$ est Fréchet de paramètre 1, c.a.d. pour $x > 0$,

$$\mathbb{P}(X_i \leq x) = \exp(-1/x), \quad i = 1, \ldots, n.$$

2. Soit $X_0^*, X_1^*, \ldots$ une suite des variables aléatoires i.i.d. de loi Fréchet de paramètre 1. Notons $M_n^* = \max(X_0^*, \ldots, X_{n-1}^*)$. Montrez que pour $z > 0$,

$$\mathbb{P}(M_n^* \leq nz) = \exp(-1/z).$$

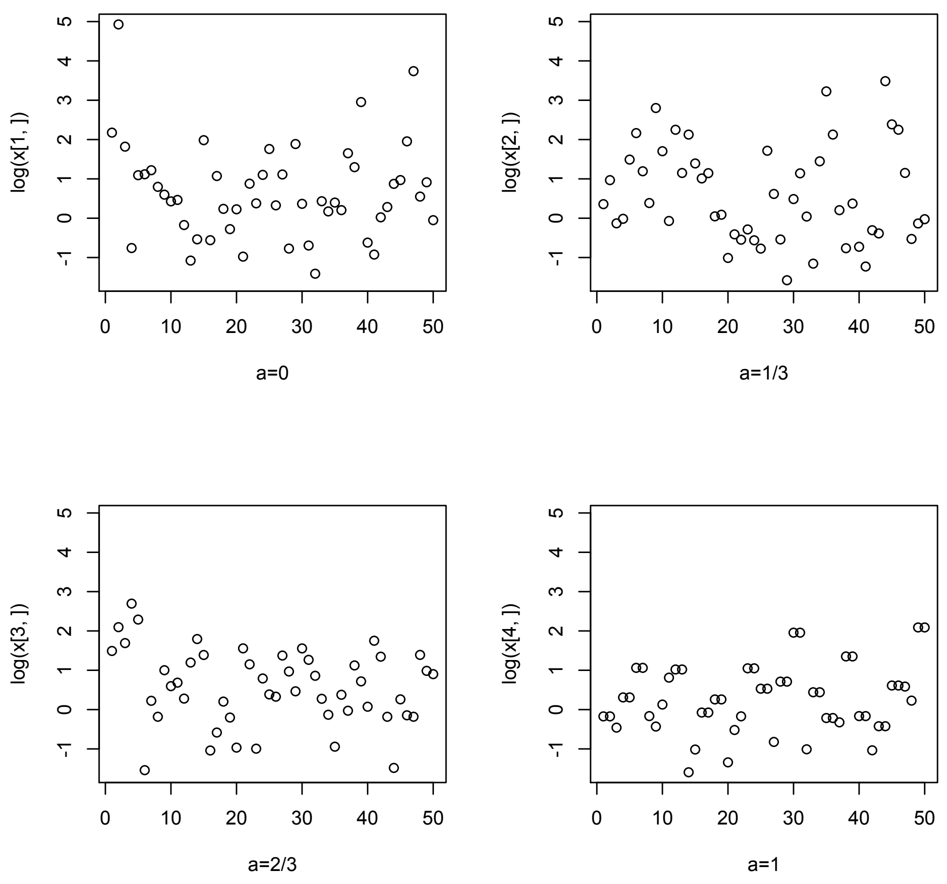
3. Notons $M_n = \max(X_0, \ldots, X_{n-1})$. Montrez que pour $z > 0$,

$$\mathbb{P}(M_n \leq nz) = (\exp(-1/z))^{\frac{1}{a+1}}.$$

4. La figure ci-dessous présente les nuages de points du logarithme de la simulation de série $X_1, \ldots, X_{50}$ pour $a = 0, 1/3, 2/3$, et 1. Le programme utilisé est fourni à la page suivante.

Pouvez vous expliquer pourquoi les valeurs extrêmes apparaissent en groupe quand a augmente ? En particulier, si $a = 1$, les observations les plus grandes apparaissent toujours en couple.

5. En utilisant les résultats obtenus pour les questions 1-3, pouvez vous expliquer pourquoi le maximum des 50 observations diminue quand a augmente ?



Le programme suivant est utilisé pour simuler la série $X_1, \ldots, X_{50}$ pour $a = 0, 1/3, 2/3,$ et 1. `a` est le vecteur $(0, 1/3, 2/3, 1)$, `x` est une matrice de dimension 4×50.

```
for(j in 1:4){
   u = runif(51)
   y = -1/(a[j]+1)/log(u)
   for(i in 1:50){x[j,i] = max(a[j]*y[i],y[i+1])}
}
```

Questions de cours

1. En utilisant trois fonctions ou façons différentes, écrivez le code R qui permet de créer le vecteur $(1, 2, 3, 4, 5, 6)$.

2. Que fait l'expression R suivante

`pnorm(3,2,sqrt(5))-pnorm(1,2,sqrt(5))` ?

3. Soient `x` un vecteur avec 5 “levels” et `y` un vecteur numérique de dimension 5. Par exemple on a ci-dessous les réponses dans la fenêtre de console R. Qu'obtient on si l'on tape `y[x]` ?

```
> x
[1] 3 1 2 5
Levels: 1 2 3 4 5
> y
[1] 0 1 2 3 4
```

4. Soient `A` et `B` deux matrices numériques de même dimension. Qu'obtient on si l'on tape `A*B` et `A%*%B` en R ?

5. Que fait le programme suivant ? Pouvez vous prédire la valeur de `y[5000]` ?

```
n = 5000
x = rnorm(n)
y = cumsum(x)/(1:n)
plot(y,type="l")
```

(Indication sur la fonction `cumsum` : si $x = (2, 5, 3)$, alors `y=cumsum(x)` donne $y = (2, 7, 10)$.)