Université Paris I, Panthéon - Sorbonne

Première Année Master M.A.E.F. $2013-2014$

Econométrie II

Contrôle continu n°2, avril 2014

Examen de 1h30. Tout document ou calculatrice est interdit.

1. **(Sur 9 points)** Soit $(Y_i)_{i\in\mathbf{N}}$ une famille de variables aléatoires définie par:

$$Y_i = \theta_0 + \theta_1 X_i + \varepsilon_i \qquad \text{pour tout } i\in\mathbf{N}, \qquad \text{où:} \tag{1}$$

- θ_0 et θ_1 sont des réels inconnus;
- les variables ε_i sont des variables aléatoires indépendantes centrées et telles que pour tout $i\in\mathbf{N}$, $\mathbb{E}\varepsilon_i^2=\sigma^2$, avec $\sigma^2>0$ un réel inconnu.

(a) On suppose connu $(Y_1,\cdots,Y_n)$ et $(X_1,\cdots,X_n)$ avec $n\in\mathbf{N}^*$ (on suppose que le vecteur ${}^t(X_1,\cdots,X_n)$ n'est pas colinéaire au vecteur unité ${}^t(1,\cdots,1)$). Déterminer l'estimateur $\widehat{\theta}=(\widehat{\theta}_0,\widehat{\theta}_1)$ de (θ_0,θ_1) par moindres carrés ordinaires, où l'on notera $X=\begin{pmatrix}1 & X_1\\ \vdots & \vdots\\ 1 & X_n\end{pmatrix}$ **(0.5pts)**. L'écrire en fonction des X_i, des Y_i, de $\overline{X}_n=(X_1+\cdots+X_n)/n$ et de $\overline{Y}_n=(Y_1+\cdots+Y_n)/n$ **(1pt)**.

(b) On suppose que l'on connait une nouvelle observation X_{n+1} et l'on considère la prédiction de Y_{n+1} à l'aide du modèle obtenu, soit $\widetilde{Y}_{n+1}=\widehat{\theta}_0+\widehat{\theta}_1X_{n+1}$. Déterminer l'espérance de l'erreur de prédiction $\widetilde{\varepsilon}_{n+1}=Y_{n+1}-\widetilde{Y}_{n+1}$ **(0.5pts)**. Montrer que sa variance vaut: $\sigma^2\big(1+(1,X_{n+1})({}^tXX)^{-1}{}^t(1,X_{n+1})\big)$ **(2pts)**.

(c) Montrer qu'une autre expression de cette variance est $\sigma^2\big(1+\frac{1}{n}+\frac{(X_{n+1}-\overline{X}_n)^2}{\sum_{i=1}^n(X_i-\overline{X}_n)^2}\big)$ **(1.5pts)**. En déduire que l'erreur de prévision minimale (au sens quadratique) est obtenue pour $X_{n+1}=\overline{X}_n$ **(0.5pts)**.

(d) Lorsque les ε_i sont gaussiennes, en déduire (en justifiant) la région de confiance à 95% de l'erreur de prédiction en fonction de $x=X_{n+1}$ **(2pts)**. Quelle figure géométrique obtient-t-on quand x décrit $\mathbf{R}$ **(1pt)**?

2. **(Sur 9 points)** On suppose maintenant que l'on se trouve dans le cadre de la régression multiple, soit

$$Y_i = \theta_0 + \theta_1 X_i^{(1)} + \cdots + \theta_1 X_i^{(p)} + \varepsilon_i \qquad \text{pour tout } i\in\mathbf{N}, \tag{2}$$

les $(X_i^{(j)})_{1\le i\le n,\,1\le j\le p}$ étant connues et telles que $X=\begin{pmatrix}1 & X_1^{(1)} & \cdots & X_1^{(p)}\\ \vdots & \vdots & & \vdots\\ 1 & X_n^{(1)} & \cdots & X_n^{(p)}\end{pmatrix}$ soit une matrice de rang p.

(a) Déterminer la prédiction optimale de Y_{n+1} au sens du risque quadratique, soit $\widehat{Y}_{n+1}=\mathbb{E}\big(Y_{n+1}\mid(Y_1,\cdots,Y_n)\big)$ **(1pt)**, puis l'erreur de prédiction $Y_{n+1}-\widehat{Y}_{n+1}$ **(0.5pts)**. Que pensez-vous de cette prédiction? **(0.5pts)**

(b) En supposant connues $(X_{n+1}^{(1)},\cdots,X_{n+1}^{(p)})$ et en reprenant les notations précédentes et avec $\widehat{\theta}=(\widehat{\theta}_0,\cdots,\widehat{\theta}_p)$ estimateur de $(\theta_0,\cdots,\theta_p)$ par moindres carrés ordinaires à partir de $(Y_1,\cdots,Y_n)$, déterminer l'espérance **(0.5pts)** et la variance de $\widetilde{\varepsilon}_{n+1}$ **(2pts)**.

(c) On pose $X=[\mathbb{I}\ \ Z]$, où $\mathbb{I}$ est le vecteur colonne ${}^t(1,\cdots,1)$. Montrer que ${}^tXX=n\begin{pmatrix}1 & {}^t\overline{X}_n\\ \overline{X}_n & \frac{1}{n}{}^tZZ\end{pmatrix}$, où ${}^t\overline{X}_n=\big(\frac{1}{n}\sum_{i=1}^n X_i^{(1)},\cdots,\frac{1}{n}\sum_{i=1}^n X_i^{(p)}\big)$ **(1pt)**. En déduire que $({}^tXX)^{-1}=\frac{1}{n}\begin{pmatrix}1+{}^t\overline{X}_n\Gamma^{-1}\overline{X}_n & -{}^t\overline{X}_n\Gamma^{-1}\\ -{}^t\overline{X}_n\Gamma^{-1} & \Gamma^{-1}\end{pmatrix}$, où $\Gamma=\frac{1}{n}{}^tZZ-\overline{X}_n{}^t\overline{X}_n$ (que l'on supposera inversible) **(2pts)**.

(d) En déduire que l'erreur de prédiction est de variance minimale pour $(X_{n+1}^{(1)},\cdots,X_{n+1}^{(p)})={}^t\overline{X}_n$ **(1pt)** et déterminer alors sa variance **(0.5pts)**.

3. **(Sur 9 points)** Exercice de TP utilisant le logiciel R

   (a) On a tapé les commandes suivantes:

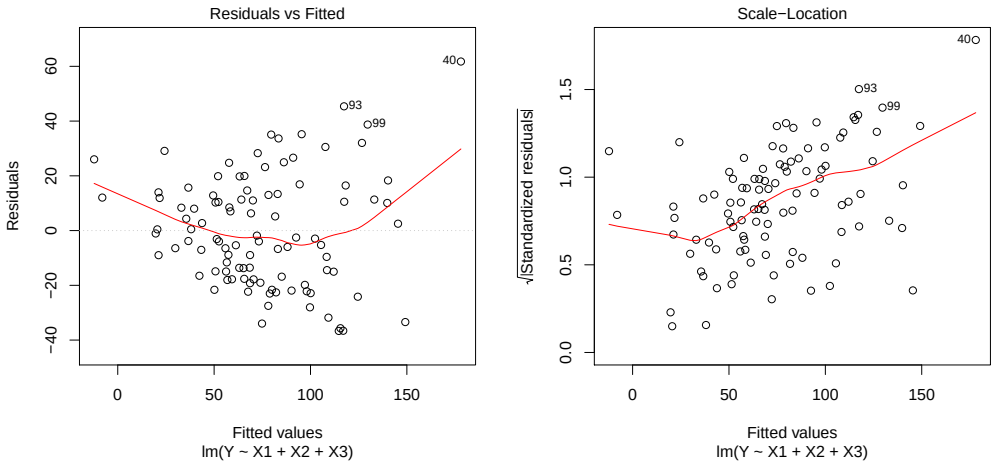
```
X1=sqrt(c(1:100))
X2=3-2*rnorm(100)
X3=c(1:100)
epsi=2*runif(100,-1,1)
Y=(10+0.2*X1-X2+epsi)^2
reg1=lm(Y~X1+X2+X3)
plot(reg1); summary(reg1);
```

   Voici les résultats:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.5210    13.5726   7.406 5.02e-11 ***
X1            5.2903     4.6419   1.140    0.257
X2          -17.7901     1.0871 -16.365  < 2e-16 ***
X3           -0.1792     0.3741  -0.479    0.633

Residual standard error: 20.66 on 96 degrees of freedom
Multiple R-squared: 0.7421,     Adjusted R-squared: 0.7341
F-statistic: 92.09 on 3 and 96 DF,  p-value: < 2.2e-16
```

   On obtient également les figures:



   Questions 1: Expliquer ce qui a été fait. Dire ce que représente les 2 *figures. Expliquer pourquoi l'estimation de l'intercept est d'environ* 100*. Que concluez-vous de ces résultats et des 2 graphes? S'attendait-on à de tels résultats?*

   (b) On tape ensuite les commandes:

```
MatEpsi=matrix(nrow=100,ncol=100, 0)
for (i in c(1:100)) MatEpsi[i,i]=Y[i]
Z=cbind(rep(1,100),X1,X2,X3)
theta=solve(t(Z)%*%solve(MatEpsi^0.5)%*%Z)%*%t(Z)%*%solve(MatEpsi^0.5)%*%Y
theta
```

   Voici les résultats:

```
       [,1]
   93.0702521
X1   5.3632824
X2 -16.2537992
X3  -0.1746591
```

   Questions 2: Qu'a-t-on fait avec ces commandes et pourquoi l'a-t-on fait? Que conclure de ces résultats?

   (c) On a ensuite tapé les commandes:

```
library(MASS)
X=as.data.frame(matrix(c(X1,X2,X3),ncol=3));
Y.lm=lm(Y~.,data=X);
Y.bic=stepAIC(Y.lm,k=log(100))
summary(Y.bic); plot(Y.bic)
```

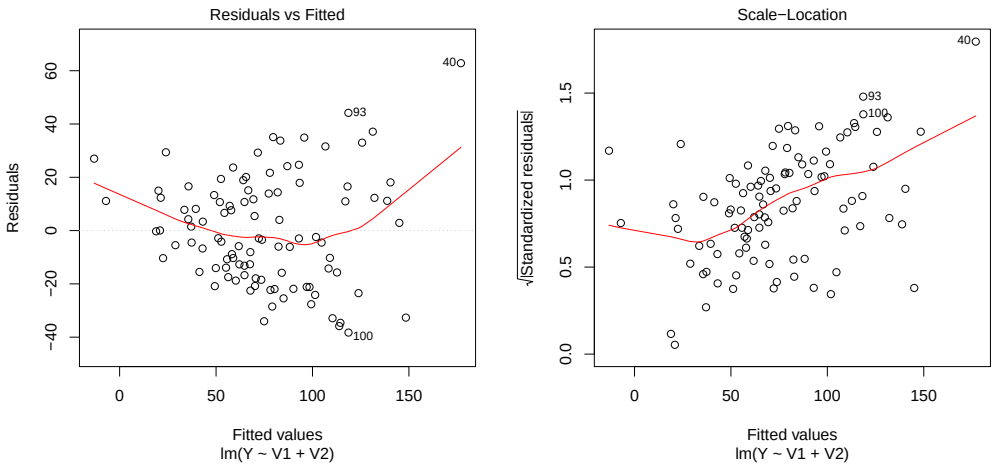
Voici les résultats:

```
Start:  AIC=619.95
Y ~ V1 + V2 + V3
       Df Sum of Sq    RSS    AIC
- V3    1        98  41063 615.59
- V1    1       554  41519 616.69
<none>               40965 619.95
- V2    1    114287 155253 748.58
Step:  AIC=615.59
Y ~ V1 + V2
       Df Sum of Sq    RSS    AIC
<none>               41063 615.59
- V1    1      5220  46283 622.95
- V2    1    114208 155272 743.99

*****************************************
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 106.1019     6.9396  15.289  < 2e-16 ***
V1            3.1073     0.8849   3.512 0.000678 ***
V2          -17.7816     1.0826 -16.425  < 2e-16 ***

Residual standard error: 20.58 on 97 degrees of freedom
Multiple R-squared: 0.7415,     Adjusted R-squared: 0.7362
F-statistic: 139.1 on 2 and 97 DF,  p-value: < 2.2e-16
```

On obtient également les figures:



Questions 3: Qu'a-t-on fait avec ces commandes et qu'a-t-on obtenu? Est-on satisfait du résultat obtenu et expliquer pourquoi on pouvait s'y attendre?

(d) Enfin on a tapé les commandes:

```
boxcox(Y~X1+X2,lambda = seq(0,1,0.05))
YY=abs(Y)^0.45
reg2=lm(YY~X1+X2)
plot(reg2); summary(reg2)
```

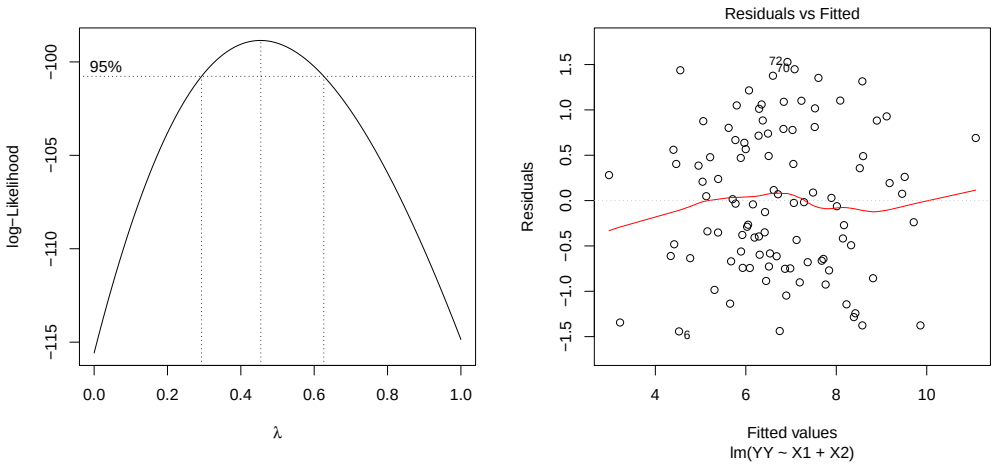
Voici les résultats:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.10133    0.26958  30.052  < 2e-16 ***
X1           0.12643    0.03437   3.678 0.000386 ***
```

```
X2          -0.75811    0.04205 -18.027  < 2e-16 ***

Residual standard error: 0.7993 on 97 degrees of freedom
Multiple R-squared: 0.775,      Adjusted R-squared: 0.7703
F-statistic:   167 on 2 and 97 DF,  p-value: < 2.2e-16
```

On obtient également les figures:



Questions 4: Que représente le graphe de gauche? Expliquer les commandes. Que conclure quant au modèle obtenu? Pouvait-on s'y attendre? Et la valeur `0.7993` *pouvait-on s'y attendre?*