Université Paris I, Panthéon - Sorbonne

Première Année Master M.A.E.F. $2014-2015$

Econométrie II

Examen final, mai 2015

Examen de 3h00. Tout document ou calculatrice est interdit.

1. **(16 points)** On veut modéliser le prix P d'un smartphone d'occasion en fonction de son âge A qui est donné en mois. On dispose d'un échantillon de n annonces sur le "Le Bon Coin", à partir desquelles on observe $(P_1,\cdots,P_n)$ en fonction de $(A_1,\cdots,A_n)$. Pour simplifier, on suppose que l'on a choisi des annonces telles que $A_i=i$. On propose un premier modèle:
$$P_i=\theta_0+\theta_1\,i+\varepsilon_i \qquad \text{pour } i=1,\cdots,n, \text{ avec} \tag{1}$$
$\theta={}^t(\theta_0,\theta_1)\in\mathbf{R}^2$ le vecteur des paramètres inconnus et $(\varepsilon_i)_{1\leq i\leq n}$ une suite de v.a.i.i.d. centrées, de variance σ^2 et de moment d'ordre 4 $\mu_4<\infty$.

   (a) Ecrire le modèle sous une forme matricielle, et rappeler l'expression de l'estimateur $\widehat{\theta}$ de θ par moindres carrés ordinaires sous forme matricielle. Cet estimateur est-il convergeant dans $\mathbb{L}^2$? Est-il asymptotiquement normal? **(3pts)**.

   (b) Le modèle (1) admet un clair défaut: quand i devient très grand, le prix prévu devient négatif. Pour remédier à cela, on propose le second modèle suivant:
$$P_i=\theta_2\,i^{-p}+\varepsilon_i \qquad \text{pour } i=1,\cdots,n, \text{ avec} \tag{2}$$
$p>0$ supposé connu, et $\theta_2\in\mathbf{R}$ inconnu. Donner l'expression de l'estimateur $\widehat{\theta}_2$ par moindres carrés ordinaires en fonction de n, p et des P_i. Expliquer pourquoi $\widehat{\theta}_2>0$. A-t-on résolu le problème précédent? L'estimateur est-il asymptotiquement convergeant dans $\mathbb{L}^2$? Est-il asymptotiquement normal? On suppose que $p=1/2$. Déterminer un Théorème de la Limite Centrale vérifié par $\widehat{\theta}_2$. Si p n'est pas connu, proposer une méthode pour estimer p **(4pts)**.

   (c) Le modèle (2) est un peu frustre et ne modélise pas suffisamment bien les données (comment peut-on s'en apercevoir?). On observe qu'il existe une différence de comportement des prix entre les appareils de plus de 3 ans ($i>36$) et ceux de moins de 3 ans ($i\leq 36$). On suppose ainsi que pour $i=1,\cdots,36$, on modélise P par un modèle linéaire de type (1) et pour $i=37,\cdots,n$, un modèle linéaire de type (2) avec $p=1/2$. Ecrire matriciellement le modèle, et vérifier la convergence de l'estimateur de $(theta_0,\theta_1,\theta_2)$ par moindres carrés. Pouvez-vous proposer un test permettant de tester s'il y a une rupture ou non? **(3pts)**.

   (d) Un défaut majeur du modèle précédent est qu'il crée un grand saut de prix qui n'a pas lieu d'être entre la prédiction de prix d'un smartphone de 36 mois et celle d'un smartphone de 37 mois. On rajoute donc une containte au modèle précédent (toujours avec $p=1/2$) pour garantir la continuité. Quelle relation existe-t-il entre θ_0, θ_1 et θ_2 pour obtenir la continuité du modèle en $i=36$? En remplaçant θ_2 par θ_0 et θ_1, réécrire le modèle matriciellement. Si on note $(\widehat{\theta}_0,\widehat{\theta}_1)$ l'estimateur par moindres carrés obtenu, montrer que $\widehat{\theta}_0$ et $\widehat{\theta}_1$ ne convergent pas vers θ_0 et θ_1 quand $n\to\infty$. Proposer un test pour tester ce modèle et donner sa loi asymptotique. Si 36 n'était pas exactement le moment du changement de modèle, déterminer une procédure permettant de retrouver automatiquement cette date **(6pts)**.

2. **(10 points)** Exercice de TP utilisant le logiciel R

   (a) On a tapé les commandes suivantes avec le logiciel R4:

```
x1=4+0*c(1:100); x2=-3*rnorm(100,0); x3=x2^2; x4=5*runif(100)-2; eps=rnorm(100,0);
y=5*x1-3*x2+2*x3+20*eps;
```

   Les réalisations des variables y, $x1$, $x2$, $x3$ et $x4$ seront supposées connues, alors que celle de la variables *eps* seront inconnues.

   Questions I.1: quelles sont les lois des variables $x2$, $x3$ et $x4$? Ecriver le modèle vérifié par y sous forme matricielle $Y=X\,\theta+\varepsilon$, en précisant la matrice X et le vecteur θ.

(b) On fait maintenant comme si θ était inconnu. On tape ainsi les commandes:

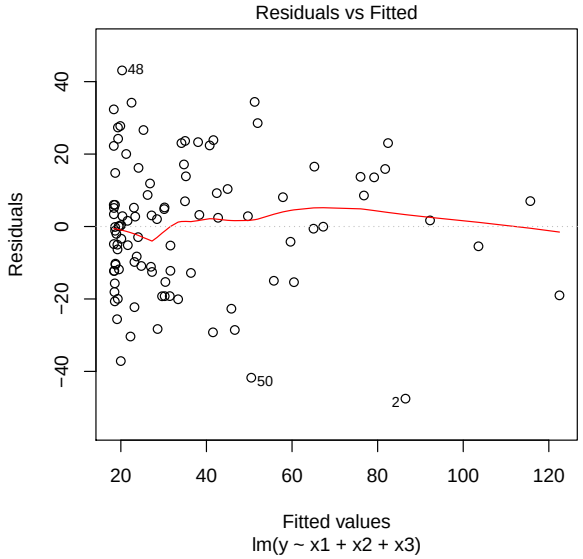
```
reg1=lm(y~x1+x2+x3);
summary(reg1)
plot(reg1)
```

Voici une sélection des résultats et graphes obtenus.

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

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.1229     2.3103   8.277 6.89e-13 ***
x1               NA         NA      NA       NA
x2           -2.4049     0.6019  -3.995 0.000126 ***
x3            1.8708     0.1574  11.888  < 2e-16 ***

Residual standard error: 18.14 on 97 degrees of freedom
Multiple R-squared:  0.6202,    Adjusted R-squared:  0.6123
F-statistic: 79.19 on 2 and 97 DF,  p-value: < 2.2e-16
```



Questions I.1: Qu'a-t-on fait en tapant ces commandes? Que représentent précisément (en fonction des variables) les valeurs numériques 19.1229, 0.1574, −3.995 *et* 79.19. *Pourquoi a-t-on un* NA *à la suite de* $x1$*? Les résultats obtenus pour* $\widehat{\theta}$ *sont-ils conformes à ce que l'on attendait?*

Questions I.2: Calculer uniquement en fonction de X *la valeur de la matrice de covariance* Σ *de* $\widehat{\theta}$. *Si on note* $x1 = (x1_i)$, $x2 = (x2_i)$ *et* $x3 = (x3_i)$, *écrire* Σ *en fonction des* xj_i. *Déterminer* $\mathbb{E}(xj_i \times xk_i)$ *pour* $1 \leq j \leq k \leq 3$. *En déduire en justifiant que*

$$\frac{1}{100}\,{}^t X\, X \simeq \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

En déduire si la valeur numérique 0.6019 *était prévisible. Par le même raisonnement, peut-on en déduire que* $\widehat{\theta}$ *est un estimateur convergent de* θ*?*

Questions I.3: Que pensez-vous de la valeur du R^2, *des tests de Fisher et de student et du graphique? Le modèle linéaire obtenu est-il acceptable?*

On suppose maintenant que l'on ne connaissait pas le modèle et on désire le retrouver en sachant que y peut être expliquée par $x1$, $x2$, $x3$ et $x4$. Pour cela on tape les commandes suivantes:

```
library(MASS)
X=as.data.frame(cbind(x1,x2,x3,x4));
y.lm=lm(y~.,data=X);
y.bic=stepAIC(y.lm,k=log(100))
```

Les résultats sont les suivants:

```
Start:  AIC=594.99
y ~ x1 + x2 + x3 + x4

Step:  AIC=594.99
y ~ x2 + x3 + x4
       Df Sum of Sq   RSS    AIC
- x4    1        14 31930 590.43
<none>               31917 594.99
- x2    1      5154 37071 605.36
- x3    1     46480 78397 680.25

Step:  AIC=590.43
y ~ x2 + x3
       Df Sum of Sq   RSS    AIC
<none>               31930 590.43
- x2    1      5255 37185 601.06
- x3    1     46524 78455 675.72
```

Questions I.4: Qu'a-t-on fait? Quel est le modèle sélectionné? Est-ce surprenant?