

Première Année Master M.A.E.F. 2016 – 2017

## Econométrie II

Contrôle continu n°1, mars 2017

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

1. **(Sur 20 points)** On dispose d'une variable  $Y = {}^t(Y_1, \dots, Y_n)$  observée pour  $n$  individus ainsi que de  $p + 1$  variables potentiellement explicatives observées  $X^{(j)}$  avec  $1 \leq j \leq p$ , tel que  $X^{(j)} = {}^t(X_1^{(j)}, \dots, X_n^{(j)})$ . On suppose également que la matrice  $Z = (\mathbb{I}, X^{(1)}, \dots, X^{(p)})$  de taille  $(n, p + 1)$  est de rang  $p + 1$ , avec  $\mathbb{I} = {}^t(1, \dots, 1)$ , et qu'il existe  $\theta = {}^t(\theta_0, \theta_1, \dots, \theta_p)$  et  $\theta_{p+1} \neq 0$  tels que:

$$Y = Z\theta + \theta_{p+1}X^{(p+1)} + \varepsilon,$$

où  $X^{(p+1)} = {}^t(X_1^{(p+1)}, \dots, X_n^{(p+1)})$  est un vecteur composé de  $n$  v.a.i.i.d. absolument continues par rapport à la mesure de Lebesgue et de variance  $\sigma_{p+1}^2$  et  $\varepsilon = {}^t(\varepsilon_1, \dots, \varepsilon_n)$  est un vecteur gaussien composé de  $n$  v.a.i.i.d. centrées de variance  $\sigma_\varepsilon^2$ ,  $X^{(p+1)}$  et  $\varepsilon$  étant indépendants. On note  $\|\cdot\|$  la norme euclidienne classique de  $\mathbf{R}^n$ .

- (a) Montrer que presque sûrement la famille des vecteurs  $(\mathbb{I}, X^{(1)}, \dots, X^{(p+1)})$  est de rang  $p + 2$  (on pourra montrer que ces vecteurs sont presque sûrement linéairement indépendants) **(2 pts)**.
- (b) On effectue la régression linéaire multiple par moindres carrés de  $Y$  par rapport aux variables  $X^{(j)}$  avec  $1 \leq j \leq p + 1$ . Donner une expression matricielle de  $\hat{Y}$  **(1 pt)**. Démontrer qu'il existe une matrice carrée  $M$  de taille  $n$  telle que  $\|Y - \hat{Y}\|^2 = \|M\varepsilon\|^2 \leq \|\varepsilon\|^2$  **(2 pts)**, puis que  $\mathbb{E}(\|\varepsilon\|^2 - \|M\varepsilon\|^2) = (p + 2)\sigma_\varepsilon^2$  (utiliser l'espérance conditionnelle par rapport à  $X^{(p+1)}$  car  $M$  est aléatoire!) **(3 pts)**. Après avoir déterminé la limite de  $\frac{1}{n}\|\varepsilon\|^2$ , en déduire que  $\frac{1}{n}\|Y - \hat{Y}\|^2 \xrightarrow[n \rightarrow +\infty]{P} \sigma_\varepsilon^2$  **(2 pts)**. Déterminer ainsi qu'un équivalent asymptotique (en probabilité) du critère BIC, noté  $\hat{BIC}_{p+1}$ , pour ce modèle de régression est  $n \log(\sigma_\varepsilon^2)$  **(1 pt)**.
- (c) Rappeler quelle est la statistique du test de Student sur la significativité de  $X^{(p+1)}$  **(0.5 pts)**. Quelle est, sous l'hypothèse  $H_0$  (que l'on précisera), sa loi conditionnelle sachant  $X^{(p+1)}$  **(0.5 pts)**? Si on suppose que, sachant  $X^{(p+1)}$ , l'estimateur du vecteur  $\gamma = {}^t(\theta, \theta_{p+1})$  est convergent dans  $\mathbb{L}^2$ , montrer qu'asymptotiquement le test de Student décide la significativité de  $X^{(p+1)}$  quelque soit l'erreur de première espèce  $\alpha$  fixée **(3 pts)**.
- (d) On effectue maintenant une régression linéaire multiple par moindres carrés de  $Y$  par rapport aux variables  $X^{(j)}$  avec  $1 \leq j \leq p$ . Donner l'expression matricielle de  $\hat{Y}$  en utilisant la matrice  $Z$  **(0.5 pts)**. On note  $m = \mathbb{E}(X_1^{(p+1)})$  et  $U_i = X_i^{(p+1)} - m$  pour  $1 \leq i \leq n$ . Montrer que  $Y - \hat{Y} = (I_n - H)(\theta_{p+1}U + \varepsilon)$  avec  $U = {}^t(U_1, \dots, U_n)$ ,  $I_n$  la matrice identité de taille  $n$  et  $H$  une matrice que l'on précisera **(1.5 pts)**. En déduire que  $n \log(\theta_{p+1}^2 \sigma_{p+1}^2 + \sigma_\varepsilon^2)$  est un équivalent asymptotique (en probabilité) du critère BIC, noté  $\hat{BIC}_p$ , pour ce modèle de régression **(2 pts)**.
- (e) Démontrer qu'avec le critère BIC, la probabilité de choisir le vrai modèle avec  $p + 1$  variables plutôt que le sous-modèle avec  $p$  variables tend asymptotiquement vers 1 **(1 pt)**.

2. **(Sur 6 points)** Exercice avec le logiciel R:

- (a) On compile les commandes suivantes:

```
n=500;i=c(1:n)
X1=i; X2=runif(n,3,10); X3=log(i); epsilon=rnorm(n)
Y=40+10*X3+2*X2+i^0.5*epsilon
reg1=lm(Y~X1+X2)
summary(reg1)
par(mfrow=c(2,2)); plot(reg1)
```

Les résultats numériques et graphiques sont les suivants:

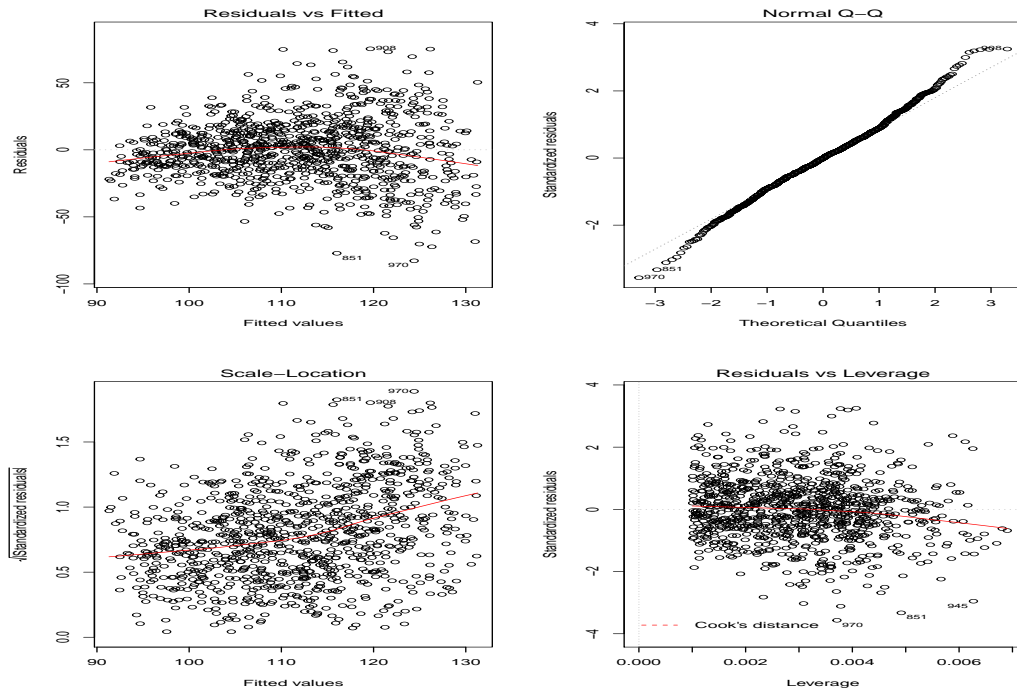
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	84.900889	2.783674	30.500	<2e-16 ***
X1	0.029195	0.002546	11.467	<2e-16 ***
X2	1.862415	0.364945	5.103	4e-07 ***

Residual standard error: 23.24 on 997 degrees of freedom

Multiple R-squared: 0.1371, Adjusted R-squared: 0.1353

F-statistic: 79.17 on 2 and 997 DF, p-value: < 2.2e-16



Question 1: Dire, avec des formules mathématiques, ce à quoi correspondent 84.900889, 0.364945 11.467, 4e-07 et 79.17. Que peut-on conclure quant à la régression effectuée? Est-ce surprenant? (3.5pts)

(b) On poursuit avec les commandes qui suivent:

```
BX=boxcox(Y~X1+X2,plotit = TRUE,lambda = seq(-3,3))
ind=which(BX$y==max(BX$y))
(lambda=BX$x[ind])
```

Question 2: On obtient comme résultat > [1] 0.5151515. Qu'a-t-on fait et qu'obtient-on? Pouvait-on s'y attendre? (1.5pts)

(c) Les commandes suivantes:

```
Z=Y^0.5; reg2=lm(Z~X1+X2+X3)
summary(reg2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	6.9998093	0.3343708	20.934	< 2e-16 ***
X1	-0.0002147	0.0002418	-0.888	0.375
X2	0.0874513	0.0168894	5.178	2.71e-07 ***
X3	0.5138798	0.0708845	7.250	8.38e-13 ***

Residual standard error: 1.075 on 996 degrees of freedom

Multiple R-squared: 0.1732, Adjusted R-squared: 0.1707

F-statistic: 69.56 on 3 and 996 DF, p-value: < 2.2e-16

Question 3: Expliquer ce qui a été fait et pourquoi. A-t-on gagné par rapport à la première régression? Quelle devrait être la démarche suivante? (1pt)