

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

Econométrie II

Contrôle continu n°2, avril 2017

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

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

$$Y = Z\theta + \varepsilon,$$

où $\varepsilon = {}^t(\varepsilon_1, \dots, \varepsilon_n)$ un vecteur gaussien composé de n v.a. indépendantes centrées. On désignera comme hypothèse H_0 le fait que $\text{var}(\varepsilon_i) = \sigma_\varepsilon^2$ pour tout $i = 1, \dots, n$. On note $\|\cdot\|$ la norme euclidienne classique de \mathbf{R}^n .

- (a) On considère un sous-échantillon de taille n_1 de l'échantillon (Y_1, \dots, Y_n) , avec $n_1 \geq p+1$. Rappeler la formule de l'estimateur non biaisé $\hat{\sigma}^2(n_1)$ de σ_ε^2 par moindres carrés, ainsi que sa loi sous H_0 (**1pt**). On rappelle que sous H_0 , $\sqrt{n_1}(\hat{\sigma}^2(n_1) - \sigma_\varepsilon^2) \xrightarrow[n_1 \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 2\sigma_\varepsilon^4)$.
- (b) On veut tester l'homoscédasticité du modèle. Pour ce faire, on découpe l'échantillon en deux sous-échantillons de tailles respectives n_1 et n_2 tels que $n = n_1 + n_2$, avec $\min(n_1, n_2) \geq p+1$. On définit $\hat{F}_{n_1, n_2} = \hat{\sigma}^2(n_1) / \hat{\sigma}^2(n_2)$. Déterminer la loi de \hat{F}_{n_1, n_2} sous H_0 (**1.5pts**) et montrer que $\hat{F}_{n_1, n_2} \xrightarrow[n_1, n_2 \rightarrow +\infty]{\mathcal{P}} 1$ sous H_0 (**1.5pts**).
- (c) On suppose désormais que $n_1 = [\lambda n]$ avec $\lambda \in]0, 1[$. Donner, en justifiant, le théorème de la limite centrale vérifié par le vecteur ${}^t(\hat{\sigma}^2(n_1), \hat{\sigma}^2(n_2))$ en fonction de n , λ et σ_ε^2 lorsque $n \rightarrow \infty$ (**2pts**). En déduire en utilisant la Delta-méthode que $\sqrt{n}(\hat{F}_{[\lambda n], n-[\lambda n]} - 1) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 2(\lambda(1-\lambda))^{-1})$ (**2.5pts**). Déterminer le seuil asymptotique à partir duquel on considèrera qu'il n'y a pas homoscédasticité avec un risque de premier espèce de 5% (**2.5pts**).
- (d) On pose comme hypothèse alternative H_1 : $\text{var}(\varepsilon_i) = \gamma_\varepsilon^2 \neq \sigma_\varepsilon^2$ pour tout i appartenant à l'échantillon de taille n_2 . Déterminer le théorème de la limite centrale vérifié par $\hat{F}_{[\lambda n], n-[\lambda n]}$ sous H_1 (**3pts**). Montrer que la puissance du test tend vers 1 avec n pour ce problème de test (toujours avec un risque de premier espèce de 5%) (**3pts**). A n fixé pour quelle valeur de λ est-elle maximale (**1pt**)?
- (e) Déduire de ce qui précède une procédure permettant de trouver les deux sous-échantillons de taille n_1 et n_2 qui maximise la possibilité d'être avec 2 régimes d'hétéroscédasticité (**1pt**). Qu'en pensez-vous numériquement (**1pt**)?

2. **(Sur 8 points)** Exercice avec le logiciel R:

On s'intéresse à la perte de poids, variable **Perte**, de différents cafés lors de leur torréfaction à partir d'une base de données issues de nombreuses mesures (189 en tout). Plusieurs variables interviennent sur cette perte: les différents types de cafés considérés (7 différents, numérotés de 1 à 7, mais avec plusieurs mesures effectuées pour chaque café), variable **Cafe**, l'humidité mesurée, variable **Humi**, la luminosité du café, variable **Lumi**, et 4 différentes mesures de la couleur du café, variables **A**, **B**, **Y** et **Gn**. On veut mettre en place un modèle expliquant la perte de poids du café.

- (a) On lance les commandes suivantes:

```
Cafe=as.factor(Cafe)
reg1=lm(Perte~Cafe+Humi+Lumin+A+B+Y+Gn)
summary(reg1)
plot(reg1)
```

Les résultats numériques et graphiques sont les suivants:

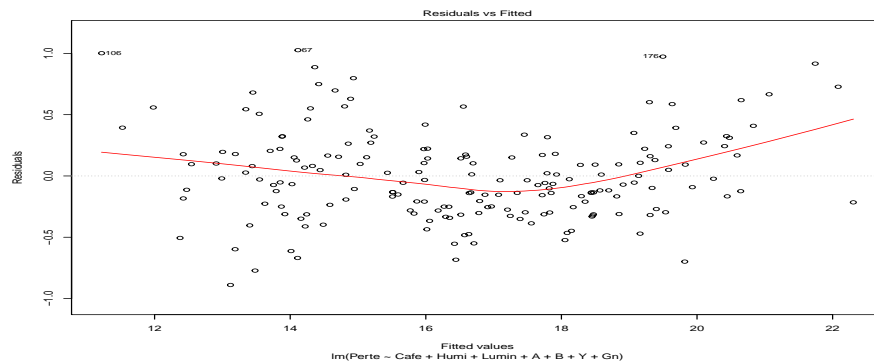
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	27.25102	1.89136	14.408	< 2e-16 ***
Cafe2	1.52527	0.15511	9.834	< 2e-16 ***
Cafe3	1.06378	0.08691	12.240	< 2e-16 ***
Cafe4	0.59288	0.26452	2.241	0.02625 *
Cafe5	0.43903	0.13883	3.162	0.00184 **
Cafe6	0.89410	0.14159	6.315	2.14e-09 ***
Cafe7	1.19981	0.16809	7.138	2.39e-11 ***
Humi	0.91816	0.04423	20.757	< 2e-16 ***
Lumin	-1.01589	0.17688	-5.744	4.00e-08 ***
A	-1.43841	0.14532	-9.898	< 2e-16 ***
B	0.84786	0.12939	6.553	6.02e-10 ***
Y	0.36049	0.17612	2.047	0.04216 *
Gn	-0.01598	0.18574	-0.086	0.93152

Residual standard error: 0.3717 on 176 degrees of freedom

Multiple R-squared: 0.978, Adjusted R-squared: 0.9765

F-statistic: 652.5 on 12 and 176 DF, p-value: < 2.2e-16



Question 1: Expliquez comment est prise en compte la variable Cafe dans cette régression et pourquoi elle est déclinée en 6 variables dans le tableau. Que peut-on conclure quant à la régression effectuée? Que faire ensuite? (2.5pts)

(b) On poursuit avec les commandes qui suivent:

```
reg2=stepAIC(lm(Perte~Cafe+Lumin+Humi+A+B+Y+Gn),k=log(189),direction="both", trace=FALSE)
summary(lm(reg2$model))
plot(lm(reg2$model),1)
```

Les résultats numériques sont les suivants (le graphe donne quasiment la même chose que le précédent):

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	25.33704	1.44125	17.580	< 2e-16 ***
Cafe2	1.35577	0.10465	12.955	< 2e-16 ***
Cafe3	1.03435	0.08636	11.977	< 2e-16 ***
Cafe4	0.13436	0.11766	1.142	0.25502
Cafe5	0.31548	0.11266	2.800	0.00567 **
Cafe6	0.78067	0.12016	6.497	7.94e-10 ***
Cafe7	1.04364	0.12724	8.202	4.56e-14 ***
Lumin	-0.79022	0.09347	-8.454	9.89e-15 ***
Humi	0.92859	0.04214	22.036	< 2e-16 ***
A	-1.50531	0.12668	-11.882	< 2e-16 ***
B	0.76551	0.12246	6.251	2.93e-09 ***

Residual standard error: 0.3744 on 178 degrees of freedom

Multiple R-squared: 0.9775, Adjusted R-squared: 0.9762

F-statistic: 771.6 on 10 and 178 DF, p-value: < 2.2e-16

Question 2: Qu'a-t-on fait et qu'obtient-on? Est-on désormais satisfait? (1.5pts)

(c) Les commandes suivantes:

```
library(MASS)
BX=boxcox(Perte ~ Cafe+Humi+Lumin+A+B,plotit=TRUE,lambda = seq(-3,3))
ind=which(BX$y==max(BX$y)); (lambda=BX$x[ind])
LogPerte=log(Perte)
reg3=lm(LogPerte ~ Cafe+Humi+Lumin+A+B)
summary(reg3); plot(reg3,1)
```

Les résultats numériques et graphiques sont les suivants:

[1] 0.2727273

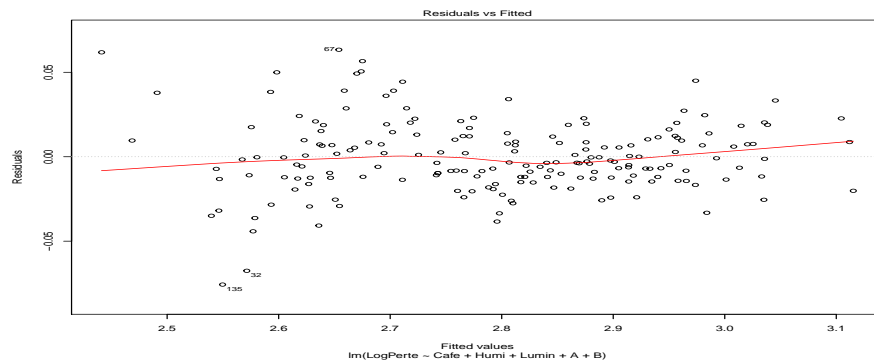
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.117752	0.084277	36.994	< 2e-16 ***
Cafe2	0.103113	0.006119	16.850	< 2e-16 ***
Cafe3	0.066510	0.005050	13.170	< 2e-16 ***
Cafe4	0.012560	0.006880	1.825	0.069603 .
Cafe5	0.031103	0.006588	4.722	4.73e-06 ***
Cafe6	0.062919	0.007026	8.955	4.46e-16 ***
Cafe7	0.082289	0.007440	11.060	< 2e-16 ***
Humi	0.054360	0.002464	22.060	< 2e-16 ***
Lumin	-0.037351	0.005466	-6.834	1.27e-10 ***
A	-0.062178	0.007408	-8.394	1.43e-14 ***
B	0.027577	0.007161	3.851	0.000164 ***

Residual standard error: 0.02189 on 178 degrees of freedom

Multiple R-squared: 0.9791, Adjusted R-squared: 0.9779

F-statistic: 832.9 on 10 and 178 DF, p-value: < 2.2e-16



Question 3: Expliquer ce qui a été fait et pourquoi. A-t-on gagné par rapport à la précédente régression? (2pts)

(d) Les commandes suivantes:

```
1-sum((Perte-exp(reg3$fit))^2)/sum((Perte-mean(Perte))^2)
```

Le résultat numérique est [1] 0.9816623.

Question 4: Expliquer cette commande et le résultat. Quel café choisiriez-vous pour obtenir la plus petite perte possible? (2pts)