*Université Paris I, Panthéon - Sorbonne*

**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 $Y = {}^t(Y_1, \cdots, Y_n)$ ainsi que de $p$ variables explicatives observées $X^{(j)}$ avec $1 \leq j \leq p$, telles que $X^{(j)} = {}^t(X_1^{(j)}, \cdots, X_n^{(j)})$. On suppose également que la matrice $Z = (\mathbb{I}, X^{(1)}, \cdots, X^{(p)})$ de taille $(n, p+1)$ est de rang $p+1$, avec $\mathbb{I} = {}^t(1, \cdots, 1)$, et qu'il existe $\theta = {}^t(\theta_0, \theta_1, \cdots, \theta_p)$ et tel que:
$$Y = Z\,\theta + \varepsilon,$$
où $\varepsilon = {}^t(\varepsilon_1, \cdots, \varepsilon_n)$ un vecteur gaussien composé de $n$ v.a. indépendantes centrées. On désignera comme hypothèse $H_0$ le fait que $\mathrm{var}(\varepsilon_i) = \sigma_\varepsilon^2$ pour tout $i = 1, \ldots, n$. On note $\|\cdot\|$ la norme euclidienne classique de $\mathbf{R}^n$.

   (a) On considère un sous-échantillon de taille $n_1$ de l'échantillon $(Y_1, \cdots, Y_n)$, avec $n_1 \geq p+1$. Rappeler la formule de l'estimateur non biaisé $\widehat{\sigma}^2(n_1)$ de $\sigma_\varepsilon^2$ par moindres carrés, ainsi que sa loi sous $H_0$ **(1pt)**. On rappelle que sous $H_0$, $\sqrt{n_1}\big(\widehat{\sigma}^2(n_1) - \sigma_\varepsilon^2\big) \xrightarrow[n_1\to\infty]{\mathcal{L}} \mathcal{N}\big(0, 2\,\sigma_\varepsilon^4\big)$.

   (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 $\widehat{F}_{n_1,n_2} = \widehat{\sigma}^2(n_1)/\widehat{\sigma}^2(n_2)$. Déterminer la loi de $\widehat{F}_{n_1,n_2}$ sous $H_0$ **(1.5pts)** et montrer que $\widehat{F}_{n_1,n_2} \xrightarrow[n_1,n_2\to+\infty]{\mathcal{P}} 1$ 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 ${}^t\big(\widehat{\sigma}^2(n_1), \widehat{\sigma}^2(n_2)\big)$ en fonction de $n$, $\lambda$ et $\sigma_\varepsilon^2$ lorsque $n \to \infty$ **(2pts)**. En déduire en utilisant la Delta-méthode que $\sqrt{n}\big(\widehat{F}_{[\lambda\, n],n-[\lambda\, n]} - 1\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\big(0\,,\, 2(\lambda(1-\lambda))^{-1}\big)$ **(2.5pts)**. Déterminer le seuil asymptotique à partir duquel on considérera qu'il n'y a pas homoscédasticité avec un risque de premier espèce de 5% **(2.5pts)**.

   (d) On pose comme hypothèse alternative $H_1$: $\mathrm{var}(\varepsilon_i) = \gamma_\varepsilon^2 \neq \sigma_\varepsilon^2$ pour tout $i$ appartenant à l'échantillon de taille $n_2$. Déterminer le théorème de la limite centrale vérifié par $\widehat{F}_{[\lambda\, n],n-[\lambda\, n]}$ 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 $\lambda$ 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)**?

*Proof.* (a) On a $\widehat{\sigma}^2(n_1) = \frac{1}{n_1-(p+1)} \sum_{j\in I(n_1)} \widehat{\varepsilon}_j^2$.
On sait d'après Cochran que $\widehat{\sigma}^2(n_1) \overset{\mathcal{L}}{\sim} \frac{\sigma_\varepsilon^2}{n_1-(p+1)}\,\chi^2(n_1-(p+1))$.

(b) Il est clair que sous $H_0$ la loi de $\widehat{F}_{n_1,n_2}$ est une loi de Fisher $F(n_1-(p+1), n_2-(p+1))$, puisque $\widehat{\sigma}^2(n_1)$ et $\widehat{\sigma}^2(n_2)$ ont des lois $\sigma_\varepsilon^2\,\chi^2$ renormalisées par leurs degrés de libertés respectifs, et sont clairement indépendants car calculés à partir de deux sous-échantillons disjoints donc indépendants, puisque les $\varepsilon_i$ sont des v.a. indépendantes.
Sous $H_0$, on a $\widehat{\sigma}^2(n_1) \xrightarrow[n_1,n_2\to+\infty]{\mathcal{P}} \sigma_\varepsilon^2$ et $\widehat{\sigma}^2(n_2) \xrightarrow[n_1,n_2\to+\infty]{\mathcal{P}} \sigma_\varepsilon^2$ d'après le cours (ou comme conséquence des TCL). La convergence en loi vers une constante équivalent à la convergence en probabilité vers cette même constante, on peut utiliser le Lemme de Slutsky et on bien $\widehat{F}_{n_1,n_2} \xrightarrow[n_1,n_2\to+\infty]{\mathcal{P}} \sigma_\varepsilon^2/\sigma_\varepsilon^2 = 1$.

(c) D'après le TLC rappelé, on a $\sqrt{n_1}\big(\widehat{\sigma}^2(n_1) - \sigma_\varepsilon^2\big) \xrightarrow[n_1\to\infty]{\mathcal{L}} \mathcal{N}(0\,,\, 2\sigma_\varepsilon^4)$. Or comme $n_1/n \xrightarrow[n\to\infty]{} \lambda$, en utilisant Slutsky on montre facilement que $\sqrt{n}\big(\widehat{\sigma}^2(n_1) - \sigma_\varepsilon^2\big) \xrightarrow[n_1\to\infty]{\mathcal{L}} \mathcal{N}(0\,,\, 2\sigma_\varepsilon^4/\lambda)$. De la même manière, $\sqrt{n}\big(\widehat{\sigma}^2(n_1) - \sigma_\varepsilon^2\big) \xrightarrow[n_1\to\infty]{\mathcal{L}} \mathcal{N}(0\,,\, 2\sigma_\varepsilon^4/(1-\lambda))$. De plus $\widehat{\sigma}^2(n_1)$ et $\widehat{\sigma}^2(n_2)$ sont clairement indépendants car calculés à partir de deux sous-échantillons disjoints donc indépendants, puisque les $\varepsilon_i$ sont des v.a. indépendantes. Ainsi:
$$\sqrt{n}\Big(\begin{pmatrix} \widehat{\sigma}^2(n_1) \\ \widehat{\sigma}^2(n_2) \end{pmatrix} - \begin{pmatrix} \sigma_\varepsilon^2 \\ \sigma_\varepsilon^2 \end{pmatrix}\Big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, 2\sigma_\varepsilon^4 \begin{pmatrix} 1/\lambda & 0 \\ 0 & 1/(1-\lambda) \end{pmatrix}\Big).$$

On applique alors la Delta-méthode avec la fonction $g(x,y) = x/y$, soit une matrice jacobienne $J_g(x,y) = (1/y, -x/y^2) = \sigma_\varepsilon^{-2}(1,-1)$ pour $x = y = \sigma_\varepsilon^2$. En l'appliquant sur le TLC multidimensionnel précédent, on obtient le résultat voulu.
On décidera qu'il y a hétéroscédasticité lorsque les variances des deux sous-échantillons ne sont pas égales, donc quand $\widehat{F}(n_1,n_2)$ ne tend pas vers 1. Cela advient lorsque $|\widehat{F}(n_1,n_2)-1| > s$, où $s$ est un seuil à calculer en fonction du risque de premier espèce. En utilisant la loi asymptotique de $\widehat{F}(n_1,n_2)-1$ et le quantile à $97,5\%$ de la loi normale centrée réduite, qui vaut environ 1.96 on montre que $s = 1.96\sqrt{2(\lambda(1-\lambda))^{-1}/n}$.

(d) Sous $H_1$ tout le raisonnement précédent peut être reproduit, à la nuance près que $\widehat{\sigma}^2(n_2)$ converge maintenant vers $\gamma_\varepsilon^2$. Aussi a-t-on

$$\sqrt{n}\Big(\begin{pmatrix}\widehat{\sigma}^2(n_1)\\ \widehat{\sigma}^2(n_2)\end{pmatrix} - \begin{pmatrix}\sigma_\varepsilon^2\\ \gamma_\varepsilon^2\end{pmatrix}\Big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\Big(\begin{pmatrix}0\\0\end{pmatrix}, 2\begin{pmatrix}\sigma_\varepsilon^4/\lambda & 0\\ 0 & \gamma_\varepsilon^4/(1-\lambda)\end{pmatrix}\Big).$$

Puis $J_g(x,y) = (1/y, -x/y^2) = (1/\gamma_\varepsilon^2, -\sigma_\varepsilon^2/\gamma_\varepsilon^4)$. La Delta-méthode implique alors:

$$\sqrt{n}(\widehat{F}(n_1,n_2) - \sigma_\varepsilon^2/\gamma_\varepsilon^2) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\Big(0, 2\,(1/\gamma_\varepsilon^2, -\sigma_\varepsilon^2/\gamma_\varepsilon^4)\begin{pmatrix}\sigma_\varepsilon^4/\lambda & 0\\ 0 & \gamma_\varepsilon^4/(1-\lambda)\end{pmatrix}\begin{pmatrix}1/\gamma_\varepsilon^2\\ -\sigma_\varepsilon^2/\gamma_\varepsilon^4\end{pmatrix}\Big) = \mathcal{N}\Big(0, \frac{2}{\lambda(1-\lambda)}\frac{\sigma_\varepsilon^4}{\gamma_\varepsilon^4}\Big).$$

La puissance du test est donnée par $1 - \mathbb{P}_{H_1}(\text{Choisir } H_0)$. On choisit $H_0$ dès que $\sqrt{n}\,|\widehat{F}(n_1,n_2)-1| > 1.96\sqrt{2(\lambda(1-\lambda))^{-1}}$, soit $\big|\sqrt{n}\,(\widehat{F}(n_1,n_2) - \sigma_\varepsilon^2/\gamma_\varepsilon^2) + \sqrt{n}\,(\sigma_\varepsilon^2/\gamma_\varepsilon^2 - 1)\big| > 1.96\sqrt{2(\lambda(1-\lambda))^{-1}}$. Avec la loi asymptotique de $\sqrt{n}\,(\widehat{F}(n_1,n_2) - \sigma_\varepsilon^2/\gamma_\varepsilon^2)$, et en notant $Z_n$ une variable telle que $Z_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$, cela revient à écrire:

$$\Big|\frac{\sigma_\varepsilon^2}{\gamma_\varepsilon^2} Z_n + \sqrt{2n(\lambda(1-\lambda))}\Big(\frac{\sigma_\varepsilon^2}{\gamma_\varepsilon^2} - 1\Big)\Big| > 1.96.$$

Il est bien clair que la probabilité d'un tel événement tend vers 1 quand $n \to \infty$.
Cette probabilité est maximale quand $\lambda(1-\lambda)$ est maximale, donc quand $\lambda = 1/2$.

(e) Une méthode pour détecter l'hétéroscédasticité pourrait être de déterminer le découpage en deux sous-échantillon qui maximise la distance $\sqrt{n}\,|\widehat{F}(n_1,n_2)-1|$. Il faudrait pour cela faire varier $n_1$ et $n_2$, mais aussi les individus des sous-échantillons.
La limite de cette méthode serait que $2^n$ tests devraient être effectuer, ce qu devient vite prohibitif dès que $n \geq 30$; ce n'est donc pas une méthode possible numériquement.

□

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.

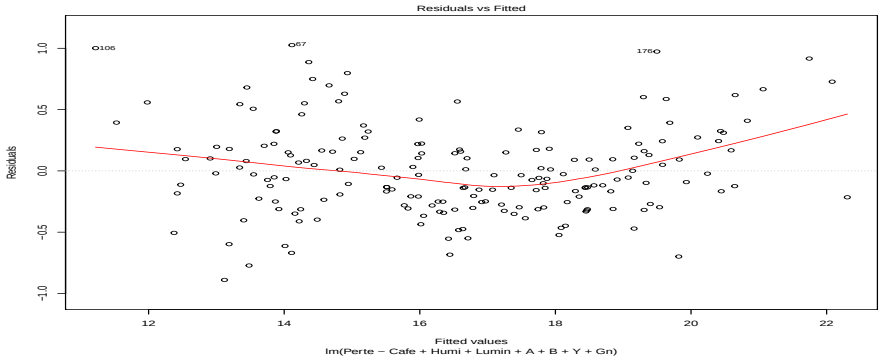
(a) On lance les commandes suivantes:

```
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)**

*Proof.* La variable `Cafe` initialement quantitative est transformée en une variable qualitative avec 7 modalités, c'est-à-dire en variables $\mathbf{1}_{\texttt{Cafe}=j}$. Comme la somme de ces 7 variables est égale à 1, donc ces 7 variables sont colinéaires avec l'intercept, le logiciel choisit de n'en considérer que 6, ommetant `Cafe=1`.
Même si le $R^2$ est très élevé, la régression n'est pas satisfaisante car le graphe des résidus en fonctions des valeurs prédites présente une forme très nette et certaines variables explicatives ne semblent pas vraiment significatives comme en témoignent les p-values données par les tests de Student (notamment pour `Gn`.
Deux démarches semblent nécessaires pour améliorer le modèle: éliminer certaines variables et changer le modèle, par exemple en corrigeant une possible hétéroscédasticité. □

(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)**

*Proof.* On a effectué une sélection des variables explicatives par la minimisation du critère BIC. Au final, deux variables ont été éliminées: `Gn` et `Y`. L'ensemble des variables explicatives sont désormais significatives (la variable qualitative `Cafe` également, même si une modalité, la 4, semble non essentielle). Mais le graphe étant à peu près le même, on ne peut encore être satisfait. □

(c) Les commandes suivantes:

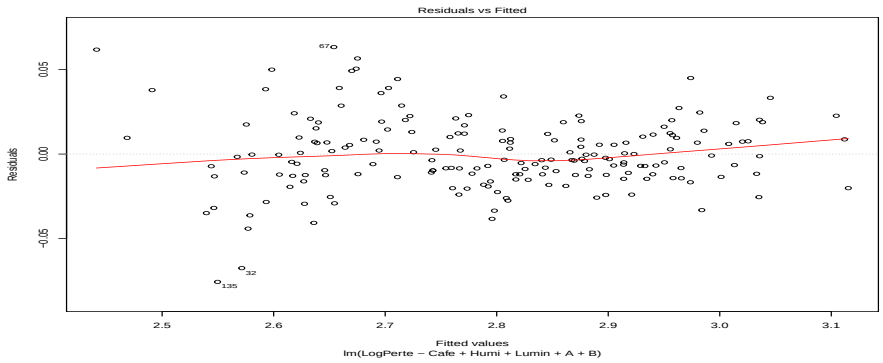
```
library(MASS)
BX=boxcox(Perte ~ Cafe+Humi+Lumin+A+B,plotit=TRUE,lambda = seq(-3,3))
ind=which(BX$y==max(BX$y)); (lambda=BX$x[ind])
LogPerte=log(Perte)
reg3=lm(LogPerte ~ Cafe+Humi+Lumin+A+B)
summary(reg3);  plot(reg3,1)
```

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)**

*Proof.* On a effectué une correction de l'hétéroscédasticité par une transformation de Box-Cox. Le résultat obtenu donne une puissance estimée $\lambda \simeq 0.27$. On choisira donc de transformer la variable à expliquer `Perte` en `log(Perte)` qui est normalement la transformation pour $\lambda = 0$, entier le plus proche de 0.27... Le graphe est désormais clairement plus acceptable car sans forme marquée. □

(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)**

*Proof.* On a calculé le $R^2$ en passant à l'exponentielle du modèle précédent. On a ainsi gagné par rapport aux modèles des questions 1 et 2.
L'exponentielle étant une fonction croissante, on peut reprendre le modèle de la question 3 pour déterminer le café permettant la plus petite perte possible. Celle-ci est obtenue pour le café 1 (le coefficient est 0 alors que tous les autres sont strictement positifs). □