Université Paris I, Panthéon - Sorbonne

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

Econométrie I

Examen final, janvier 2017

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

Exercice 1 (Sur 20 points) **Rappel (Lemme de Slutsky):** Si (X_n) et (Y_n) sont deux suites de v.a. sur $(\Omega, \mathcal{A}, \mathbb{P})$ telles que $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$ et $Y_n \xrightarrow[n\to+\infty]{\mathcal{P}} c$, où $c \in \mathbf{R}$, alors $X_n Y_n \xrightarrow[n\to\infty]{\mathcal{L}} c\,X$.

On dispose des données donnant l'espérance de vie E en France en fonction du sexe S et de l'année depuis 1900. Plus précisément, on notera $E_{1,i}$ (respectivement $E_{2,i}$) pour $i = 1, \dots, n$, l'espérance de vie moyenne des femmes (resp. des hommes) l'an $1900 + i$. On posera $F = (F_j)_{1\le j\le 2n} = {}^t(E_{1,1}, \cdots, E_{1,n}, E_{2,1}, \cdots, E_{2,n})$.

1. Dans un premier temps, sans tenir compte du sexe, on veut savoir si l'espérance de vie dépend linéairement de l'année considérée.

   (a) Ecrire le modèle vectoriel sous-jacent vérifié par F en détaillant la matrice X.

   (b) Montrer que $({}^t X\, X)^{-1} = \frac{1}{n(n^2-1)} \begin{pmatrix} (n+1)(2n+1) & -3(n+1) \\ -3(n+1) & 6 \end{pmatrix}$ pour tout $n \in \mathbf{N}^*$ (on rappelle que $\sum_{i=1}^k i^2 = \frac{1}{6}\,(2k+1)(k+1)k$). En déduire que l'estimateur par moindres carrés $\widehat{\theta} = {}^t(\widehat{\theta}_0, \widehat{\theta}_1)$ des paramètres $\theta = {}^t(\theta_0, \theta_1)$ du modèle vaut:

   $$\widehat{\theta}_0 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \big((2n+1) - 3i\big)(E_{1,i} + E_{2,i}) \quad \text{et} \quad \widehat{\theta}_1 = \frac{3}{n(n^2-1)} \sum_{i=1}^{n} \big(2i - (n+1)\big)(E_{1,i} + E_{2,i}),$$

   où θ_0 désigne l'*intercept* du modèle .

   (c) Si on suppose que les erreurs pour ce modèle forment une suite de v.a.i.i.d. centrées de variance σ^2, l'estimateur $\widehat{\theta}$ est-il sans biais? Déterminer sa matrice de variance en fonction de n et de σ^2. Montrer qu'il est asymptotiquement gaussien et en déduire en particulier que

   $$\sqrt{\frac{n^3}{6}}\,(\widehat{\theta}_1 - \theta_1) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\big(0\,,\,\sigma^2\big).$$

   (d) Donner l'expression de $\widehat{\sigma}^2$ estimateur sans biais de σ^2 en fonction des $E_{1,i}$ et $E_{2,i}$, de $\widehat{\theta}_0$ et $\widehat{\theta}_1$. Est-ce un estimateur convergeant?

   (e) Déterminer le comportement asymptotique de $\widehat{T}_n = \sqrt{\frac{n^3}{6\,\widehat{\sigma}^2}}\,\widehat{\theta}_1$ dans le cas où $\theta_1 = 0$, puis dans le cas où $\theta_1 \neq 0$. Est-ce que $\widehat{T}_n$ est la statistique de Student d'un test? Si on prend les données jusqu'à l'an 2000, avec $\widehat{\sigma} \simeq 5$, on obtient $\widehat{\theta}_1 \simeq 0.37$. Peut-on dire que le modèle est légitime avec un risque de 5%?

2. Dans un deuxième temps, sans tenir compte de l'année, on veut savoir si l'espérance de vie dépend du sexe sous la forme

   $$E_{k,i} = \alpha_k + \varepsilon_{k,i} \quad \text{pour } k = 1, 2 \text{ et } i = 1, \dots, n,$$

   où les $(\varepsilon_{k,i})$ forment une suite de v.a.i.i.d. centrées de variance σ^2.

   (a) Ecrire le modèle sous forme vectorielle et en déduire les estimateurs par moindres carrés de α_1 et α_2.

   (b) Montrer que $\widehat{\alpha}_1$ et $\widehat{\alpha}_2$ vérifient des théorèmes de la limite centrale que l'on précisera.

   (c) En déduire que:

   $$\sqrt{\frac{n}{2\,\widehat{\sigma}^2}}\,\big((\widehat{\alpha}_1 - \widehat{\alpha}_2) - (\alpha_1 - \alpha_2)\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\big(0\,,\,1\big),$$

   où $\widehat{\sigma}^2$ est l'estimateur non biaisé de σ^2 que l'on précisera en fonction des $E_{1,i}$, $E_{2,i}$, $\widehat{\alpha}_1$ et $\widehat{\alpha}_2$.

   (d) Donner l'expression de la statistique de Student permettant de tester si le sexe est un facteur explicatif significatif de l'espérance de vie. En l'an 2000 on trouve que $\widehat{\alpha}_1 \simeq 67$, $\widehat{\alpha}_2 \simeq 60$ et $\widehat{\sigma} \simeq 12$. Grâce au théorème précédent, que peut-on conclure avec un risque de 5%?

3. On considère désormais un modèle où l'espérance de vie dépend linéairement à la fois du sexe et de l'année considérée, soit:
$$E_{k,i} = \beta_k + i\,\gamma_k + \varepsilon_{k,i} \quad \text{pour } k = 1, 2 \text{ et } i = 1, \dots, n,$$
toujours avec les $(\varepsilon_{k,i})$ formant une suite de v.a.i.i.d. centrées de variance σ^2.

   (a) A partir de ce qui précède donner l'expression des estimateurs par moindres carrés $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\gamma}_1$ et $\widehat{\gamma}_2$.

   (b) Démontrer que:
$$\sqrt{\frac{n^3}{24\,\widehat{\sigma}^2}}\,\big((\widehat{\gamma}_1 - \widehat{\gamma}_2) - (\gamma_1 - \gamma_2)\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\big(0\,,\,1\big),$$
   où $\widehat{\sigma}^2$ est l'estimateur non biaisé de σ^2 que l'on précisera en fonction des $E_{1,i}$, $E_{2,i}$, $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\gamma}_1$ et $\widehat{\gamma}_2$.

   (c) On trouve numériquement $\widehat{\sigma} \simeq 4$, $\widehat{\gamma}_1 \simeq 0.382$ et $\widehat{\gamma}_2 \simeq 0.365$. Peut-on alors admettre que $\gamma_1 = \gamma_2$ avec un risque de 5%? Quel autre modèle peut-on proposer?

Proof. 1. (a) On a $F = X\theta + \varepsilon$ où $X = {}^t\begin{pmatrix} 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & n & 1 & 2 & \cdots & n \end{pmatrix}$.

   (b) Du calcul avec aussi $1 + \cdots + n = n(n+1)/2$.
   On utilise la formule $\theta = ({}^t X\,X)^{-1}{}^t X\,F$.

   (c) $\mathbb{E}(\varepsilon) = 0$ implique que $\widehat{\theta}$ est sans biais.
   On sait aussi que $\mathrm{var}(\widehat{\theta}) = \sigma^2\,({}^t X\,X)^{-1}$.
   Question calculatoire: on détermine $H = X\,= ({}^t X\,X)^{-1}{}^t X$ puis $\max_i(|H_{ii}|)$. On trouve que $H_{ii} = 6\big((i-(n+1)/2)^2 + (n^2-1)/12\big)$ pour $1 \le i \le n$ (pour $n+1 \le i \le 2n$, on retrouve la même chose en remplaçant i par $i-n$). D'où $\max_i(|H_{ii}|) = (2n-1)/(n(n+1)) \xrightarrow[n\to\infty]{} 0$. L'estimateur est donc asymptotiquement gaussien.
   Comme $\mathbb{E}(\widehat{\theta}_1) = \theta_1$, $\mathrm{var}(\widehat{\theta}_1) = \sigma^2\frac{6}{n(n^2-1)}$ pour $n \ge 2$ et $\widehat{\theta}_1$ asymptotiquement gaussien, on en déduit que $\big(\mathrm{var}(\widehat{\theta}_1)\big)^{-1/2}\big(\widehat{\theta}_1) - \theta_1\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0\,,\,1)$, d'où le résultat.

   (d) On a $\widehat{\sigma}^2 = \frac{1}{2n-2}\sum_{i=1}^n \big((E_{1,i} - \widehat{\theta}_0 - \widehat{\theta}_1\,i)^2 + (E_{2,i} - \widehat{\theta}_0 - \widehat{\theta}_1\,i)^2\big)$.
   D'après le cours, toutes les hypothèses sont vérifiées pour assurer la convergence en probabilité de $\widehat{\sigma}^2$ vers σ^2.

   (e) On montrer facilement d'après la question (c) et le Lemme de Slutsky que sous $H_0:\ \theta_1 = 0$, alors $\widehat{T}_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0\,,\,1)$. Et si $\theta_1 \neq 0$, alors $\widehat{T}_n - \sqrt{\frac{n^3}{6\,\widehat{\sigma}^2}}\,\theta_1 \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0\,,\,1)$, donc $\widehat{T}_n \xrightarrow[n\to\infty]{\mathcal{L}} \mp\infty$. Cela permet donc de tester H_0.
   Asymptotiquement $\widehat{T}_n$ se comporte comme la statistique de Student sous H_0 permettant de tester si $\theta_1 = 0$, la seule différence est le fait que l'on a remplacé $n(n^2-1)$ par n^3.
   Avec les données numériques, on a $\widehat{T}_n \simeq 30$ nettement plus grand que 1.96 donc on en conclut que le modèle est légitime car on rejette H_0.

2. (a) On a $F = X\,\alpha + \varepsilon$ avec $X = {}^t\begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}$. On en déduit que
$$\widehat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^n E_{1,i} \quad \text{et} \quad \widehat{\alpha}_2 = \frac{1}{n}\sum_{i=1}^n E_{2,i}.$$

   (b) En utilisant directement le TLC classique, on a $\sqrt{n}\,\big(\widehat{\alpha}_1 - \alpha_1\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0\,,\,\sigma^2)$ et $\sqrt{n}\,\big(\widehat{\alpha}_2 - \alpha_2\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0\,,\,\sigma^2)$.

   (c) Il est facile de voir que $\widehat{\alpha}_1$ et $\widehat{\alpha}_2$ sont indépendants, et comme $\mathbb{E}(\widehat{\alpha}_1 - \widehat{\alpha}_2) = \alpha_1 - \alpha_2$ et $\mathrm{var}(\widehat{\alpha}_1 - \widehat{\alpha}_2) = 2\sigma^2/n$, on en déduit que $\sqrt{\frac{n}{2\,\sigma^2}}\,\big((\widehat{\alpha}_1 - \widehat{\alpha}_2) - (\alpha_1 - \alpha_2)\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\big(0\,,\,1\big)$. De plus on sait que $\widehat{\sigma}^2 = \frac{1}{2n-2}\sum_{i=1}^n \big((E_{1,i} - \widehat{\alpha}_1)^2 + E_{2,i} - \widehat{\alpha}_2)^2\big)$ est un estimateur sans biais convergent vers σ^2. En utilisant le Lemme de Slutsky, on obtient le théorème de la limite centrale proposé.

   (d) La statistique de Student pour ce test est bien $\widehat{T}_n = \sqrt{\frac{n}{2\,\widehat{\sigma}^2}}\,\big((\widehat{\alpha}_1 - \widehat{\alpha}_2)$. Le calcul numérique donne $\widehat{T}_n \simeq 4 > 1.96$: on rejette donc l'hypothèse H_0, l'espérance de vie dépend bien du sexe.

3. (a) Le modèle peut s'écrire plutôt comme deux modèles distincts: $F_1 = X\,{}^t(\beta_1, \gamma_1) + \varepsilon_1$ et $F_2 = X\,{}^t(\beta_2, \gamma_2) + \varepsilon_2$ pour l'estimation des paramètres $\beta_1, \gamma_1, \beta_2, \gamma_2$. On obtient ainsi dans ce cas $({}^t X\,X)^{-1} = \frac{2}{n(n^2-1)}\begin{pmatrix} (n+1)(2n+1) & -3(n+1) \\ -3(n+1) & 6 \end{pmatrix}$ pour chaque modèle et on obtient ainsi:
$$\widehat{\beta}_k = \frac{2}{n(n-1)}\sum_{i=1}^n \big((2n+1) - 3i\big)E_{k,i} \quad \text{et} \quad \widehat{\gamma}_k = \frac{6}{n(n^2-1)}\sum_{i=1}^n \big(2i - (n+1)\big)E_{k,i},$$

   (b) On utilise les mêmes résultats que précédemment pour montrer la normalité asymptotique et on a ainsi
$$\sqrt{\frac{n^3}{12\,\sigma^2}}\,\big(\widehat{\gamma}_k - \gamma_k\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0\,,\,1)$$
   pour $k = 1$ et 2. On utilise ensuite l'indépendance de $\widehat{\gamma}_1$ et $\widehat{\gamma}_2$, le fait que $\widehat{\sigma}^2 = \frac{1}{2n-4}\sum_{i=1}^n \big((E_{1,i} - \widehat{\beta}_1 - \widehat{\gamma}_1\,i)^2 + (E_{2,i} - \widehat{\beta}_2 - \widehat{\gamma}_2\,i)^2\big)$ converge vers σ^2 en probabilité, puis le Lemme de Slutsky pour montrer le théorème de la limite centrale.

(c) On obtient ainsi que $\sqrt{\frac{n^3}{24\,\widehat{\sigma^2}}}\,(\widehat{\gamma_1}-\widehat{\gamma_2}) \simeq 0.85$, ce qui est inférieur à 1.96: on accepte donc H_0, soit $\gamma_1=\gamma_2$.
Ceci nous conduirait à plutôt proposer comme modèle:

$$E_{k,i} = \beta_k + i\,\gamma + \varepsilon_{k,i} \quad \text{pour } k=1,2 \text{ et } i=1,\dots,n,$$

avec le même γ (slope) pour les hommes et les femmes.

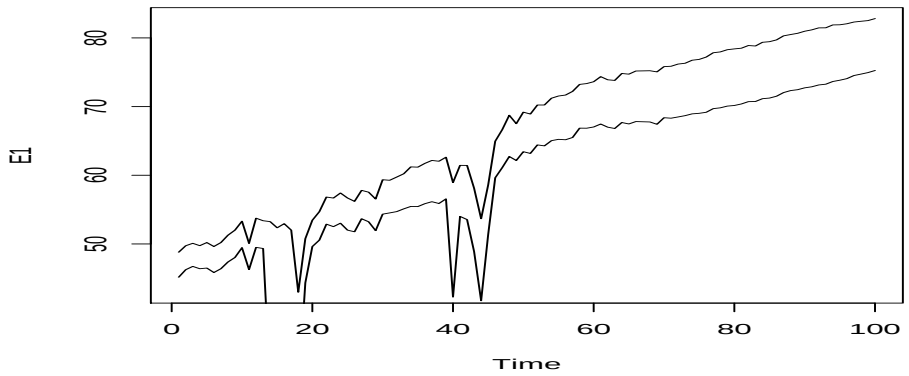
□

Exercice 2 (Sur 7 points)

On reprend les données et notations de l'Exercice 1 et on effectue l'étude évoquée précédemment avec le logiciel `R`. On commence à lire les données à partir d'un fichier (donnant les espérances de vie depuis 1816) puis on les représente graphiquement:

```
Esp=read.table('EsperanceVie.txt',header=TRUE)
E1=Esp$Fem[86:185]
E2=Esp$Ma[86:185]
I=1:100
ts.plot(E1)
lines(I,E2)
```

Soit le graphe:



On reprend le même ordre que les questions précédentes

1. On tape les commandes suivantes:

```
F=c(E1,E2)
J=c(I,I)
Reg1=lm(F~J)
summary(Reg1)
```

On obtient alors les résultats numériques et le graphe suivants:

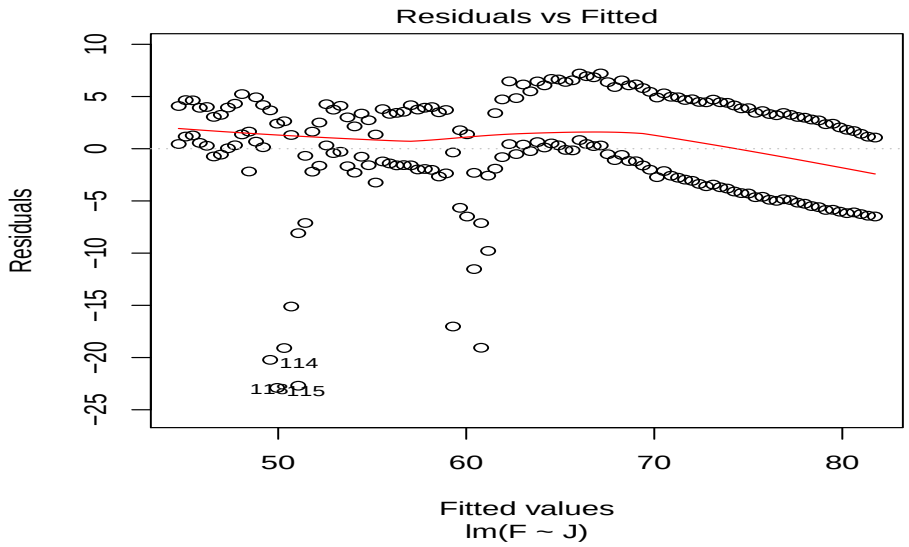
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.33430    0.76904   57.65   <2e-16 ***
J            0.37404    0.01322   28.29   <2e-16 ***

Residual standard error: 5.397 on 198 degrees of freedom
Multiple R-squared:  0.8017,    Adjusted R-squared:  0.8007
F-statistic: 800.4 on 1 and 198 DF,  p-value: < 2.2e-16
```

Quelle serait la prédiction d'espérance de vie d'une femme française en 2100? Quelles sont vos conclusions sur la régression?

2. On tape les commandes suivantes:

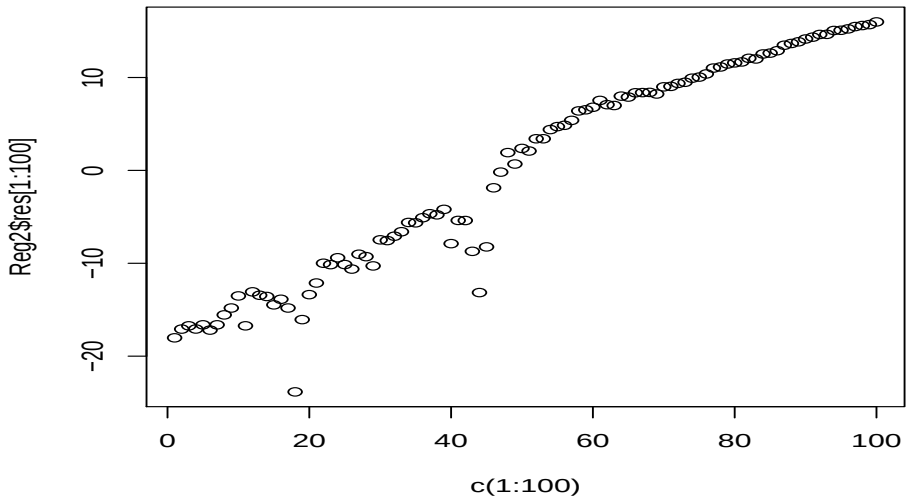
```
S1=c(rep(1,100),rep(0,100))
S2=c(rep(0,100),rep(1,100))
Reg2=lm(F~S1+S2-1)
summary(Reg2)
plot(c(1:100),Reg2$res[1:100])
```

On obtient alors les résultats numériques et le graphe suivants:




```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
S1   66.823      1.157   57.77   <2e-16 ***
S2   59.624      1.157   51.55   <2e-16 ***

Residual standard error: 11.57 on 198 degrees of freedom
Multiple R-squared:  0.968,     Adjusted R-squared:  0.9677
F-statistic:  2997 on 2 and 198 DF,  p-value: < 2.2e-16
```

Quelle serait votre prédiction pour l'espérance de vie des femmes françaises en 2100? Quelles sont vos conclusions



quant à la régression?

3. On tape les commandes suivantes:

```
J1=c(I,rep(0,100))
J2=c(rep(0,100),I)
Reg3=lm(F~S1+S2+J1+J2-1)
summary(Reg3)
```

On obtient alors les résultats suivants:

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
S1 47.49914    0.80961   58.67   <2e-16 ***
S2 41.16945    0.80961   50.85   <2e-16 ***
J1  0.38265    0.01392   27.49   <2e-16 ***
J2  0.36543    0.01392   26.25   <2e-16 ***

Residual standard error: 4.018 on 196 degrees of freedom
Multiple R-squared:  0.9962,    Adjusted R-squared:  0.9961
F-statistic: 1.278e+04 on 4 and 196 DF,  p-value: < 2.2e-16
```

Expliquer pourquoi les écarts-types empiriques sont les mêmes pour $S1$ et $S2$, ainsi que pour $J1$ et $J2$. Quelle serait votre prédiction pour l'espérance de vie des femmes françaises en 2100? Quelles sont vos conclusions par rapport à ce nouveau modèle?