Université Paris I, Panthéon - Sorbonne

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

Econométrie I

Contrôle continu n°2, 13 décembre 2016

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

Exercice 1 (Sur 12 points)

On effectue une étude de marketing sur l'impact du flacon sur la vente d'un parfum. On conçoit donc I différents flacons et pour chaque flacon on demande à n individus différents de noter (de 0 à 10) l'appréciation du parfum. On notera N_{ik} la note obtenue pour le flacon i par le k-ème individu.

1. Si $(x_1, \dots, x_m)$ est une famille de m réels, que vaut $\widehat{a} = \text{argmin}_{a\in\mathbf{R}} \sum_{j=1}^m (x_j - a)^2$ **(1pt)** ?
2. On suppose que l'on peut écrire que $N_{ik} = \mu_i + \varepsilon_{ik}$ pour tout $i = 1, \dots, I$ et $k = 1, \dots, n$ où on suppose que les erreurs (ε_{ik}) forment une famille de v.a.i.i.d. centrées de variance σ^2. Montrer que l'estimation par moindres carrés des μ_i revient à minimiser $\sum_{i=1}^I \sum_{k=1}^n (N_{ik} - \mu_i)^2$ **(0.5pts)** et en déduire que pour tout $i = 1, \dots, I$, $\widehat{\mu}_i = \frac{1}{n}\sum_{k=1}^n N_{ik}$ **(0.5pts)**. Déterminer en fonction de $\widehat{\mu}_i$, σ^2 et n, un intervalle de confiance asymptotique à 95% pour μ_i **(2pts)**.
3. Expliciter un estimateur non biaisé $\widehat{\sigma}^2$ de σ^2, en fonction des N_{ik}, des $\widehat{\mu}_i$, de n et de I **(1pt)**. En déduire (justifier !) un intervalle de confiance asymptotique à 95% pour μ_i en fonction de $\widehat{\mu}_i$, $\widehat{\sigma}^2$ et n **(2pts)**.
4. On veut tester si le flacon a un réel impact sur l'appréciation du parfum. Expliquer pourquoi cela revient à tester $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ contre sa contraposée H_1 **(0.5pts)**. On note N le vecteur colonne tel que $N = {}^t\big(N_{11}, \dots, N_{1n}, N_{21}, \dots, N_{In}\big)$. Écrire le modèle sous forme matricielle, en notant $\mu = {}^t(\mu_1, \dots, \mu_I)$ **(0.5pts)**. Déterminer une statistique de Fisher $\widehat{F}$ permettant de réaliser ce test **(1pt)**. Montrer qu'asymptotiquement $\widehat{F} \xrightarrow[n\to\infty]{\mathcal{L}} \frac{1}{I-1}\chi^2(I-1)$ **(3pts)**.

Démonstration.

1. On montre facilement que $\widehat{a} = \frac{1}{m}\sum_{j=1}^m x_j$ (voir exercice traité en TD).
2. Pour chaque $i \in \{1, \dots, I\}$, la somme des moindres carrés est $\sum_{k=1}^n (N_{ik} - \mu_i)^2$, donc la somme des moindres carrés totale est bien $\sum_{i=1}^I \sum_{k=1}^n (N_{ik} - \mu_i)^2$.
   Pour minimiser cette somme, il revient à minimiser chaque sous-somme en i, et grâce au résultat de la question 1. on obtient bien l'expression de $\widehat{\mu}_i$.
   Comme les ε_{ik} forment une famille de v.a.i.i.d. de variance finie σ^2, il en est de même pour la famille des $N_{ik} - \mu_i$. On peut donc appliquer un TLC pour $\widehat{\mu}_i$ et on obtient :
   $$\sqrt{n}\,(\widehat{\mu}_i - \mu_i) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0\,,\,\sigma^2).$$
   On en déduit donc, grâce au quantile à 97.5% d'une loi $\mathcal{N}(0,1)$ qui vaut à peu près 1.96 qu'un intervalle de confiance à 95% est :
   $$\Big[\widehat{\mu}_i - 1.96\,\frac{\sigma}{\sqrt{n}}\,,\,\widehat{\mu}_i + 1.96\,\frac{\sigma}{\sqrt{n}}\Big].$$
3. On a $\widehat{\sigma}^2 = \frac{1}{nI-I}\sum_{i=1}^I \sum_{k=1}^k (N_{ik} - \widehat{\mu}_i)^2$.
   On sait d'après le cours que $\widehat{\sigma}^2 \xrightarrow[n\to+\infty]{\mathcal{P}} \sigma^2$. On peut donc utiliser le Lemme de Slutski et obtenir que :
   $$\frac{\sqrt{n}}{\widehat{\sigma}}\,(\widehat{\mu}_i - \mu_i) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0\,,\,1).$$
   On en déduit donc, grâce au quantile à 97.5% d'une loi $\mathcal{N}(0,1)$ qui vaut à peu près 1.96 qu'un intervalle de confiance à 95% est :
   $$\Big[\widehat{\mu}_i - 1.96\,\frac{\widehat{\sigma}}{\sqrt{n}}\,,\,\widehat{\mu}_i + 1.96\,\frac{\widehat{\sigma}}{\sqrt{n}}\Big].$$

4. Si $\mu_1 = \mu_2 = \cdots = \mu_I$ alors le modèle s'écrit $N_{ik} = \mu_1 + \varepsilon_{ik}$ pour tout i et tout k : il n'y a donc plus d'influence du parfum i sur la loi de N_{ik} : la variable flacon n'est plus significative.
Le modèle s'écrit vectoriellement $N = X\,\mu + \varepsilon$, où X est une matrice de taille (nI, I), où la colonne i est composée uniquement de 0, sauf entre les lignes $(i-1)n+1$ et in où il y a des 1.
Soit C la matrice de taille $(I-1, I)$ telle que ${}^tC = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \\ 1 & 0 & 0 & 0 & \cdots & -1 \end{pmatrix}$. L'hypothèse H_0 s'écrit donc ${}^tC\,\mu = 0$.
On peut donc définir la statistique de Fisher pour la nullité de $I-1$ combinaisons linéaires comme :

$$\widehat{F} = \frac{1}{I-1}\,\frac{{}^t\widehat{\mu}C\big({}^tC({}^tX\,X)^{-1}C\big)^{-1}{}^tC\widehat{\mu}}{\widehat{\sigma}^2}.$$

On sait que lorsque les ε_{ik}, donc lorsque $\widehat{\mu}$ est un vecteur gaussien, alors la distribution de $\widehat{F}$ est une loi de Fisher à $(I-1, nI-(I-1))$ degrés de liberté. Comme $\widehat{\mu}$ est un vecteur asymptotiquement gaussien, on a ${}^tC\widehat{\mu}$ qui est aussi asymptotiquement gaussien et $\big({}^tC({}^tX\,X)^{-1}C\big)^{-1/2}{}^tC$ est asymptotiquement un vecteur centré réduit gaussien de taille $I-1$. Ceci implique que ${}^t\widehat{\mu}C\big({}^tC({}^tX\,X)^{-1}C\big)^{-1}{}^tC\widehat{\mu} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(I-1)$. De plus, on est bien sous les hypothèses permettant de dire que $\widehat{\sigma}^2 \xrightarrow[n\to+\infty]{\mathcal{P}} \sigma^2$. En utilisant le Lemme de Slutsky, on en déduit le résultat demandé.

□

Exercice 2 (Sur 7 points)

On considère le modèle de régression suivant :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta x_{i,2} + \varepsilon_i, \quad 1 \le i \le n,$$

les $x_{i,j}$ étant des variables explicatives observées du modèle et les ε_i des v.a.i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. On note et on calcule :

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \Longrightarrow \quad {}^tXX = \begin{bmatrix} 30 & 20 & 0 \\ 20 & 20 & 0 \\ 0 & 0 & 10 \end{bmatrix}, \quad {}^tXY = \begin{bmatrix} 15 \\ 20 \\ 10 \end{bmatrix}, \quad \text{et} \quad {}^tYY = 59.5.$$

1. Déterminer n, la moyenne de $(x_{i,2})_i$, le coefficient de corrélation des $(x_{i,1})_i$ et des $(x_{i,2})_i$ **(1.5pts)**.
2. Calculer numériquement les estimateurs par moindres carrés ordinaires $\widehat{\theta}$ et $\widehat{\sigma}^2$ de $\theta = {}^t(\beta_0, \beta_1, \beta_2)$ et de σ^2 (on montrera que $\|Y - X\widehat{\theta}\|^2 = \|Y\|^2 - \|X\widehat{\theta}\|^2$) **(2pts)**.
3. Calculer pour β_1 un intervalle de confiance à 95% (on utilisera des valeurs approchées des quantiles d'une loi de Student) **(1pt)**. Tester également l'hypothèse $\beta_2 = 0.8$ au niveau 10% **(1pt)**.
4. Déterminer la moyenne empirique des y_i : $\bar{y}_n$ et en déduire le coefficient de détermination R^2 **(1.5pt)**.

Démonstration. 1. La valeur de n est donnée par le coefficient (1,1) de la matrice ${}^tXX$ donc $n = 30$. Puis de la même façon on trouve :

$$\overline{x}_2 = \frac{1}{30}\sum_{i=1}^{30} x_{i,2} = \frac{({}^tXX)_{1,3}}{30} = 0.$$

Alors les $x_{i,2}$ sont centrés et le coefficient de corrélation entre les deux variables x_1 et x_2 est :

$$\frac{\sum_{i=1}^{30} x_{i,1}x_{i,2}}{\sqrt{\sum_{i=1}^{30}(x_{i,1} - \overline{x_1})^2}\sqrt{\sum_{i=1}^{30} x_{i,2}^2}} = 0.$$

2. L'estimateur des MCO est donné par

$$\widehat{\theta} = ({}^tXX)^{-1}\,{}^tXY = \begin{bmatrix} 0.1 & -0.1 & 0 \\ -0.1 & 0.15 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}\begin{bmatrix} 15 \\ 20 \\ 10 \end{bmatrix} = \begin{bmatrix} -0.5 \\ 1.5 \\ 1 \end{bmatrix}.$$

Un estimateur non biaisé de σ^2 est $\widehat{\sigma}^2$ qui s'écrit :

$$\widehat{\sigma}^2 = \frac{\|Y - X\widehat{\theta}\|^2}{n-3} = \frac{\|Y\|^2 - \|X\widehat{\theta}\|^2}{27} = \frac{{}^tYY - {}^tYX({}^tXX)^{-1}{}^tXY}{27} = 1.$$

3. On sait que
$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\beta_1}} = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}\sqrt{({}^tXX)^{-1}_{2,2}}} \sim \mathcal{T}_{n-3} = \mathcal{T}_{27}$$
(loi de Student à 27 degrés de liberté), alors on en déduit un intervalle de confiance à 95% pour β_1 est :
$$\begin{aligned} I(\beta_1) &= \left[\widehat{\beta}_1 - q_{27}(0.975)\widehat{\sigma}\sqrt{({}^tXX)^{-1}_{2,2}}; \widehat{\beta}_1 + q_{27}(0.975)\widehat{\sigma}\sqrt{({}^tXX)^{-1}_{2,2}}\right] = [1.5 - 2\sqrt{0.15}; 1.5 + 2\sqrt{0.15}] \quad \text{en utilisant l'approximation} \\ &\approx [0.71; 2.29]. \end{aligned}$$

4. Pour tester l'hypothèse : $H_0 : \beta_2 = 0.8$ contre $H_1 : \beta_2 \neq 0.8$ au niveau 10%, on calcule la statistique de test :
$$\widehat{T}_2 = \frac{\widehat{\beta}_2 - \beta_2}{\widehat{\sigma}_{\beta_2}} = \frac{\widehat{\beta}_2 - \beta_2}{\widehat{\sigma}\sqrt{({}^tXX)^{-1}_{3,3}}} \sim \mathcal{T}_{n-3} = \mathcal{T}_{27}$$
$$\widehat{T}_2 = \frac{1 - 0.8}{\sqrt{0.1}} \approx 0.6 < q_{27}(0.95) \simeq 1.65$$
en utilisant encore l'approximation du quantile par celui d'une $\mathcal{N}(0,1)$, donc on accepte l'hypothèse H_0 au risque $\alpha = 10\%$.

5. La première composante du vecteur ${}^tXY$ est $\sum_{i=1}^{30} y_i$ donc on en déduit que
$$\overline{y} = \frac{15}{30} = 0.5$$
Le coefficient de détermination est donné par
$$R^2 = 1 - \frac{\sum_{i=1}^{30} \widehat{\varepsilon}_i^{\,2}}{\|Y - \overline{y}\mathbb{1}\|^2} = 1 - \frac{27 \times \widehat{\sigma}^2}{{}^tYY - 30\overline{y}^2} = 1 - \frac{27}{59.5 - 30 \times 0.5^2} = 1 - \frac{27}{59.5 - 7.5} \approx 1 - 0.52 = 0.48$$
(avec $\mathbb{1}$ le vecteur de $\mathbb{R}^n$ composé de 1). Le modèle explique donc seulement 48% de le variation des données.

□

Exercice 3 (Sur 3 points)

Considérons le jeu de données suivant : $X = (1,2,3,4,5), \quad Y = (9,13,2,8,1), \quad Z = (3,4,5,6,8)$.

1. Considérons le modèle à deux variables explicatives où Z est expliqué par X et Y. Quelles commandes faut-il taper dans `R` pour obtenir le résultat de la Figure 1 **(0.5pts)** ?
2. Quel est le modèle estimé en utilisant les résultats de la Figure 1 **(0.5pts)** ? Ce modèle est-il cependant statistiquement satisfaisant ? Que doit-on faire ? **(1pt)**
3. Calculer et comparer les coefficients de détermination partielle associés à X et à Y respectivement, à partir des Figure 2 et 3 et conclure **(1pt)**.

```
Call:
lm(formula = Z ~ X + Y)

Residuals:
      1       2       3       4       5
 0.1559  0.1054 -0.3128 -0.3142  0.3657

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.91629    0.85672   2.237   0.1548
X            1.14851    0.18001   6.380   0.0237 *
Y           -0.02452    0.05659  -0.433   0.7071
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4276 on 2 degrees of freedom
Multiple R-squared:  0.9753,    Adjusted R-squared:  0.9506
F-statistic: 39.47 on 2 and 2 DF,  p-value: 0.02471
```

FIGURE 1 –

```
Analysis of Variance Table

Response: Z
          Df Sum Sq Mean Sq F value Pr(>F)
Y          1 6.9917  6.9917  2.6863 0.1997
Residuals  3 7.8083  2.6028
```

```
Analysis of Variance Table

Response: Z
          Df Sum Sq Mean Sq F value   Pr(>F)
X          1   14.4 14.4000     108 0.001901 **
Residuals  3    0.4  0.1333
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 2 –

```
Analysis of Variance Table

Response: Z
          Df  Sum Sq Mean Sq F value  Pr(>F)
X          1 14.4000 14.4000 78.7586 0.01246 *
Y          1  0.0343  0.0343  0.1877 0.70706
Residuals  2  0.3657  0.1828
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 3 –

Démonstration.

1. Pour obtenir la Figure 1 on fait appelle à la fonction `lm` de **R** puis à la fonction `summary` :

```
regXY <- lm(Z ~ X+Y)
summary(regXY)
```

2. Le modèle s'écrit :
$$Z = 1.916 + 1.149X - 0.025Y.$$

3. Le coefficient de détermination se lit sur la Figure 1 à la ligne `Multiple R-squared`, il vaut 0.9753. Le modèle semble satisfaisant d'après cet indicateur puisqu'il explique plut de 97% de la variance présente dans des données. Cependant, au regard des résultats des tests de Student, donnés dans la table Figure 1, on voir que la variable Y devrait être supprimée. En effet la p-valeur est bien plus grande que le seuil d'erreur de 0.05 fixé par **R** donc ici on accepte l'hypothèse H_0 selon laquelle le coefficient devant Y est nul.

4. On utilise deux fois la commande `lm` puis `anova` avec une seule variable explicative. Le coefficient de détermination partielle associé à Y est
$$R^2_Y = \frac{SCE - SCE_{\text{sans X}}}{SCR_{\text{sans X}}} = \frac{14.43 - 6.9917}{7.8083} \approx 0.08$$
$$R^2_X = \frac{SCE - SCE_{\text{sans Y}}}{SCR_{\text{sans Y}}} = \frac{14.43 - 14.4}{0.4} \approx 0.95$$
on a donc $R^2_X >> R^2_Y$ l'apport de X par rapport à celui de Y pour expliquer le modèle est considérable.

□