Université Paris I, Panthéon - Sorbonne

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

Econométrie I

Contrôle continu n°1, novembre 2016

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

Exercice 1 (Sur 21 points)

Soit $(Y_i)_{1\leq i\leq n}$ une famille de variables aléatoires définie par:

$$Y_i = \theta_0 + \sum_{k=1}^{p} \theta_k \, Z_i^{(k)} + \varepsilon_i \qquad \text{pour tout } i \in \{1, \dots, n\}, \qquad \text{où:} \tag{1}$$

- $\theta = {}^t(\theta_0, \theta_1, \dots, \theta_p)$ est un vecteur composé de $p+1$ réels inconnus.
- pour $1 \leq j \leq p$, les $(Z_i^{(j)})_{1\leq i\leq n}$ sont p familles de réels connues. On note $X = \begin{pmatrix} 1 & Z_1^{(1)} & \cdots & Z_1^{(p)} \\ \vdots & \vdots & & \vdots \\ 1 & Z_n^{(1)} & \cdots & Z_n^{(p)} \end{pmatrix}$ et on suppose que son rang est $p+1$ avec $p+1 \leq n+1$.
- la suite $(\varepsilon_i)_i$ est une suite de v.a.i.i.d. de loi gaussienne centrée de variance $\sigma^2 > 0$.

1. On note $Y = (Y_i)_{1\leq i\leq n}$ et $\varepsilon = (\varepsilon_i)_{1\leq i\leq n}$. Ecrire le modèle (1) sous une forme matricielle, en précisant la loi du vecteur d'erreur ε **(0.5pts)**.
2. Rappeler l'expression de l'estimateur $\widehat{\theta}$ de θ par moindres carrés en fonction de X et Y **(0.5pts)**. On note $\widehat{Y} = X\,\widehat{\theta}$. On mesure la qualité de la prédiction par cet estimateur avec le risque quadratique $R(\widehat{Y}) = \mathbb{E}\big(\|\widehat{Y} - X\,\theta\|^2\big)$, où $\|\cdot\|$ désigne la norme euclidienne classique. Déterminer $R(\widehat{Y})$ en justifiant votre réponse **(1.5pts)**.
3. A partir du modèle (1), on veut tester l'hypothèse H_0: $\theta_i = 0$ pour tout $i = p - p_0, \cdots, p$, où $p_0 \in \mathbf{N}^*$, contre l'hypothèse H_1, son complément. On note $\widehat{\sigma}^2 = \frac{1}{n-(p+1)}\|Y - \widehat{Y}\|^2$. Déterminer sous H_0 la loi de $\widehat{\sigma}^2$ **(1.5pts)**.
4. On note X^0 la matrice extraite de X contenant uniquement ses $p - p_0 + 1$ premières colonnes et $\widehat{Y}^0 = X^0\,\widehat{\theta}^0$, où $\widehat{\theta}^0$ est obtenu par régression par moindres carrés sur les $p - p_0$ premières variables. On définit:
$$\widehat{F} = \frac{\frac{1}{p_0}\,\|\widehat{Y} - \widehat{Y}^0\|^2}{\widehat{\sigma}^2}.$$
Montrer que sous H_0, $\|\widehat{Y} - \widehat{Y}^0\|^2 = \|P_A\varepsilon\|^2$ où A est un sous-espace vectoriel de $\mathbf{R}^n$ de dimension p_0 que l'on précisera et P_A est la matrice de la projection orthogonale sur A **(3.5pts)**. En déduire la loi du numérateur de $\widehat{F}$ **(1pt)**. Montrer que $\widehat{F}$ suit une loi de Fisher à $(p_0, n-p-1)$ degrés de liberté **(1.5pts)**. Quelle règle de décision s'en déduit pour décider de H_0 avec un risque de première espèce $\alpha \in]0,1[$? **(1pt)**
5. On suppose jusqu'à la fin du problème que $\theta_i = 0$ pour tout $i = p - p_0, \cdots, p$. Déterminer alors $R(\widehat{Y})$ et $R(\widehat{Y}^0)$ **(1.5pts)**. Quel estimateur vaut-il mieux choisir entre $\widehat{\theta}$ et $\widehat{\theta}^0$ **(0.5pts)**?
6. Pour estimer σ^2, on utilise les estimateurs par moindres carrés non biaisés $\widehat{\sigma}^2$ et $\widehat{\sigma}_0^2$ construits respectivement à partir de $\widehat{\theta}$ et $\widehat{\theta}^0$. Déterminer en justifiant la loi de $\widehat{\sigma}_0^2$ **(0.5pts)**. Montrer que pour Z une variable de loi $\mathcal{N}(0,1)$, $\mathrm{var}(Z^2) = 2$ **(1.5pts)**. Déterminer alors les risques quadratiques de $\widehat{\sigma}^2$ et $\widehat{\sigma}_0^2$, soit $\mathbb{E}\big[(\widehat{\sigma}^2 - \sigma^2)^2\big]$ et $\mathbb{E}\big[(\widehat{\sigma}_0^2 - \sigma^2)^2\big]$ **(1.5pt)**. Quel estimateur de σ^2 vaut-il mieux choisir entre les deux? **(0.5pts)**
7. Pour A et B deux sous-espaces vectoriels de $\mathbf{R}^n$ tels que $A \subset B$, montrer que pour tout $x \in \mathbf{R}^n$, $\|P_A x\|^2 \leq \|P_B x\|^2$ **(2pts)**. On note $\widehat{R}^2$ et $\widehat{R}_0^2$ les coefficients de détermination R^2 respectifs pour les modèles avec $\widehat{\theta}$ et avec $\widehat{\theta}^0$. Montrer que $\widehat{R}^2 \geq \widehat{R}_0^2$ presque sûrement **(1.5pts)**. Par rapport à ce critère, quel estimateur choisiriez-vous? **(0.5pts)**

Exercice 2 (Sur 4 points)

On dispose de la consommation hebdomadaire de gaz et la température moyenne externe d'une maison test au sud de l'Angleterre pendant une saison. Une régression pour expliquer la consommation de gaz en fonction de la température est réalisée avec **R**. Les résultats numériques sont les suivants:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.97802 -0.11082  0.02672  0.25294  0.63803

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72385    0.12974    ?     < 2e-16 ***
Temp        -0.27793       ?     -11.04 1.05e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3548 on 28 degrees of freedom
Multiple R-Squared: 0.8131,     Adjusted R-squared: 0.8064
F-statistic: 121.8 on 1 and 28 DF,  p-value: 1.046e-11
```

1. Donner le modèle et les hypothèses de la régression **(0.5pts)**.
2. Donner une estimation de la variance du terme d'erreur dans le modèle de régression simple **(0.5pts)**.
3. Compléter le tableau **(1pts)**.
4. Soit A une variable aléatoire de loi de Student de degré de liberté 28. Quelle est la probabilité que A soit inférieure à 11.04 **(0.5pts)**?
5. Préciser les éléments du test correspondant à la ligne "`Temp`" du tableau (H_0, H_1, la statistique de test, sa loi sous H_0 et la règle de décision) **(0.5pt)**.
6. Rappeler la définition et interpréter le nombre "Multiple R-squared" du tableau **(0.5pts)**.
7. Pensez-vous que la température extérieure a un effet sur la consommation de gaz? Justifiez **(0.5pts)**.

Exercice 3 (Sur 2 points)

Nous étudions le nombre d'enfants y_i de 10 ménages en fonction du nombre de fois z_i où le ménage a mangé dans au restaurant en un trimestre. Nous disposons de 10 couples de mesures $= (z_i, y_i)$ et nous savons que: $\overline{z} = 15$, $\overline{y} = 2.4$ et

$$\frac{1}{10}\sum_{i=1}^{10}(z_i - \overline{z})^2 = 34.6, \quad \frac{1}{10}\sum_{i=1}^{10}(y_i - \overline{y})^2 = 4.24, \quad \frac{1}{10}\sum_{i=1}^{10}(z_i - \overline{z})(y_i - \overline{y}) = -10.7$$

1. On note $y = \widehat{\mu} + \widehat{\beta} z$ la droite de régression. Calculer à 10^{-2} près, $\widehat{\mu}$ et $\widehat{\beta}$ **(1pt)**.
2. Donner la formule du coefficient de détermination et montrer que l'application numérique donne 78%. Commentez. **(1pt)**.