Université Paris I, Panthéon - Sorbonne

Première Année Master M.A.E.F. 2020 – 2021

Econométrie II

Contrôle continu n°2, avril 2021

Examen de 2h00. Tout document ou calculatrice est interdit.

1. **Exercice 1 (Sur 19 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 $\sigma^2 > 0$.

(a) 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)**.

(b) 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 $R(\widehat{Y}) = \mathbb{E}(\|\widehat{Y} - X\,\theta\|^2)$, où $\|\cdot\|$ désigne la norme euclidienne classique. Déterminer $R(\widehat{Y})$ en justifiant votre réponse **(1.5pts)**.

(c) A partir du modèle (1), on veut tester l'hypothèse H_0: $\theta_i = 0$ pour tout $i = p - p_0 + 1, \cdots, p$, 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)**.

(d) 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)**

(e) On suppose jusqu'à la fin du problème que $\theta_i = 0$ pour tout $i = p - p_0 + 1, \cdots, p$. 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)**?

(f) 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 $\mathbb{E}[(\widehat{\sigma}^2 - \sigma^2)^2]$ et $\mathbb{E}[(\widehat{\sigma}_0^2 - \sigma^2)^2]$ **(1.5pt)**. Quel estimateur de σ^2 vaut-il mieux choisir entre les deux? **(0.5pts)**

(g) On note $\widehat{R}^2$ et $\widehat{R}^2_0$ les coefficients de détermination R^2 respectifs pour les modèles avec $\widehat{\theta}$ et avec $\widehat{\theta}^0$. Montrer que $\widehat{R}^2 \geq \widehat{R}^2_0$ **(1.5pts)**. Par rapport à ce critère, quel estimateur choisiriez-vous? **(0.5pts)**

Proof. (a) On a $Y = X\,\theta + \varepsilon$ et $\varepsilon \overset{\mathcal{L}}{\sim} \mathcal{N}\big(0\,,\,\sigma^2\,I_n\big)$ où I_n est la matrice identité.

(b) On a $\widehat{\theta} = \big({}^t X\,X\big)^{-1}{}^t X\,Y$.
On a $\widehat{Y} = X\,\theta + P_{[X]}\,\varepsilon$ d'où $R(\widehat{Y}) = \mathbb{E}(\|P_{[X]}\,\varepsilon\|^2)$. D'après Cochran, ε étant un vecteur gaussien centré isotrope, $[X]$ étant un sev de $\mathbf{R}^n$ de dimension n, alors $\|P_{[X]}\,\varepsilon\|^2 \overset{\mathcal{L}}{\sim} \sigma^2\,\chi^2(p+1)$, donc $R(\widehat{Y}) = \sigma^2\,(p+1)$.

(c) Sous H_0 on a tout de même $\|Y - \widehat{Y}\|^2 = \|P_{[X]^\perp}\,\varepsilon\|^2$, d'où d'après Cochran, $\|Y - \widehat{Y}\|^2 \overset{\mathcal{L}}{\sim} \sigma^2\,\chi^2(n-(p+1))$ car $n-(p+1)$ est la dimension de $[X]^\perp$. Donc comme sous H_1 on a $\widehat{\sigma}^2 \overset{\mathcal{L}}{\sim} \frac{\sigma^2}{n-(p+1)}\,\chi^2(n-(p+1))$ sous H_0.

(d) On a $\widehat{Y}^0 = P_{[X^0]}\,Y = X^0\theta^0 + P_{[X^0]}\,\varepsilon$ et sous H_0, $X^0\theta^0 = X\,\theta$, d'où $\widehat{Y} - \widehat{Y}^0 = P_{[X]}\,\varepsilon - P_{[X^0]}\,\varepsilon$. Comme $[X^0] \subset [X]$, l'inclusion étant stricte, on peut noter $A = [X^0]^\perp \cap [X]$, A étant donc un sev de dimension $p+1-(p-p_0+1) = p_0$ et on a $[X] = [X^0] \oplus A$, ces deux sev étant orthogonaux l'un l'autre. Par suite, $P_{[X]}\,\varepsilon = P_{[X^0]}\,\varepsilon + P_A\,\varepsilon$, d'où $\|\widehat{Y} - \widehat{Y}^0\|^2 = \|P_A\,\varepsilon\|^2$.
Sous H_0, on a d'après Cochran, $\|P_A\,\varepsilon\|^2 \overset{\mathcal{L}}{\sim} \sigma^2\,\chi^2(p_0)$. Comme $A \subset [X]$ et $(n-(p+1))\,\widehat{\sigma}^2 = P_{[X]^\perp}\varepsilon$, on déduit d'après Cochran que $P_A\,\varepsilon$ et $P_{[X]^\perp}\varepsilon$ sont indépendants, donc le numérateur et le dénominateur de $\widehat{F}$ sont indépendants. Au final on a $\widehat{F} \overset{\mathcal{L}}{\sim} F(p_0, n-(p+1))$.
Si $\widehat{F} \leq q_{F(p_0,n-(p+1))}(1-\alpha)$, quantile d'ordre $1-\alpha$ pour la loi $F(p_0, n-(p+1))$, alors on accepte H_0, sinon on la rejette.

(e) Comme précédemment on a $R(\widehat{Y}) = (p+1)\,\sigma^2$ alors que $R(\widehat{Y}) = (p+1-p_0)\,\sigma^2$. Il est donc clair qu'en terme de minimisation du risque quadratique on préférera utiliser $\widehat{\theta}^0$.

(f) On a d'après ce qui précède $\widehat{\sigma}^2_0 \overset{\mathcal{L}}{\sim} \frac{\sigma^2}{n-(p-p_0+1)}\,\chi^2\big(n-(p-p_0+1)\big)$.
On a $\mathbb{E}[Z^4] = \frac{2}{\sqrt{2\pi}}\int_0^\infty t^4 e^{-t^2/2}dt = \frac{2}{\sqrt{2\pi}}\Big(\Big[-t^3e^{-t^2/2}\Big]_0^\infty + 3\int_0^\infty t^2e^{-t^2/2}dt\Big) = 3\,\mathbb{E}[Z^2] = 3$. D'où $\mathrm{var}(Z) = \mathbb{E}[Z^4] - (\mathbb{E}[Z^2])^2 = 3-1 = 2$.
On a $\mathbb{E}\big[(\widehat{\sigma}^2-\sigma^2)^2\big] = \mathrm{var}\big(\frac{\sigma^2}{n-(p+1)}\,\chi^2(n-(p+1))\big) = \frac{\sigma^4}{(n-(p+1))^2}\,(n-(p+1))\,\mathrm{var}(Z) = \frac{2\,\sigma^4}{n-(p+1)}$ car un χ^2 ayant la même loi que la somme de carrés de gaussiennes centrées réduites, sa variance est la somme des variances de ces carrés de gaussiennes. De même $\mathbb{E}\big[(\widehat{\sigma}^2_0-\sigma^2)^2\big] = \frac{2\,\sigma^4}{n-(p-p_0+1)} = \frac{2\,\sigma^4}{n+p_0-(p+1)}$.
On préférera donc choisir $\widehat{\sigma}^2_0$ qui est également non biaisé mais de variance plus petite que celle de $\widehat{\sigma}^2$.

(g) On sait le $\widehat{R}^2$ augmente quand on augmente le nombre de variables du modèle. Ceci se retrouve par la norme de la projection orthogonale sur un sous-espace A qui est plus petite que celle sur un sous-espace B contenant A. Par conséquent $\widehat{R}^2 \geq \widehat{R}^2_0$. En utilisant ce critère on préfère $\widehat{\theta}$ à $\widehat{\theta}^0$.

□

2. **(10 points)** Exercice de TP utilisant le logiciel `R`

On considère un jeu de données réelles mettant en relation une mesure quantitative du degré d'avancement du diabete (variable `prog`) et 10 variables cliniques dans une cohorte de 442 patients. Chacun des 442 patients est décrit par les variables suivantes :

- `prog`: l'indice de progression de la maladie (le plus grand, le plus en progrès)
- `age`: l'âge du patient
- `sex`: le sexe du patient, codé numériquement (1, homme ou 2, femme)
- `bmi`: l'indice de masse corporelle
- `map`: la pression artérielle moyenne
- `ser1` - `ser6`: 6 mesures sérologiques

On a séparé au préalable et aléatoirement le jeu de données en deux sous-ensembles: `diabetes1` et `diabetes2`.

(a) On cherche à construire un modèle linéaire avec de bonnes propriétés prédictives pour la variable `prog`. On a ainsi tapé les commandes suivantes avec le logiciel R:

```
Diab1=read.table("C:/Donnees/diabetes1.txt",header=TRUE)
Diab1$sex=as.factor(Diab1$sex)
attach(Diab1)
reg1=lm(prog~age+bmi+map+ser1+ser2+ser3+ser4+ser5+ser6+sex)
summary(reg1)
```

Voici le début de `Diab1`:

```
> Diab1
    prog age sex  bmi    map ser1  ser2 ser3 ser4   ser5 ser6
1    151  59   2 32.1 101.00  157  93.2 38.0 4.00 4.8598   87
3    141  72   2 30.5  93.00  156  93.6 41.0 4.00 4.6728   85
6     97  23   1 22.6  89.00  139  64.8 61.0 2.00 4.1897   68
9    110  60   2 32.1  83.00  179 119.4 42.0 4.00 4.4773   94
10   310  29   1 30.0  85.00  180  93.4 43.0 4.00 5.3845   88
:     :    :   :  :     :      :    :     :    :     :      :
```

Une partie des résultats obtenus est présente ci-dessous:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -319.8486    84.3700  -3.791 0.000202 ***
age            0.1468     0.3097   0.474 0.635981
bmi            6.2159     1.0607   5.860 2.02e-08 ***
map            1.1904     0.3172   3.753 0.000232 ***
ser1          -0.3994     0.7728  -0.517 0.605918
ser2          -0.2271     0.7468  -0.304 0.761353
ser3           0.2344     0.9817   0.239 0.811551
ser4          20.6894     8.1506   2.538 0.011943 *
ser5          41.5733    21.7245   1.914 0.057174 .
ser6           0.1199     0.3858   0.311 0.756264
sex2         -22.9688     8.2902  -2.771 0.006154 **

Residual standard error: 52.35 on 189 degrees of freedom
Multiple R-squared:  0.5669,    Adjusted R-squared:  0.544
F-statistic: 24.74 on 10 and 189 DF,  p-value: < 2.2e-16
```

Questions I.1: Pourquoi a-t-on tapé la deuxième commande? Montrer mathématiquement que dans le cadre de cette variable à deux modalités il n'était pas obligatoire de la taper. Combien y-a-t-il d'invidus dans cette base de données? Ecrire précisément le modèle sous forme matricielle en précisant exactement la première ligne de la matrice X. En notant les paramètres θ_j, $j = 0, \ldots, p$, dans l'ordre donné ci-dessus, que vaut numériquement $\widehat{\theta}_4$? Ecrire mathématiquement ce qu'est le nombre 0.3097*. De ces résultats, pourriez-vous conclure statistiquement que plus on est en surpoids, plus la maladie progresse? Que pensez-vous de cette régression?* **(6pts)**

(b) On tape ensuite les commandes:

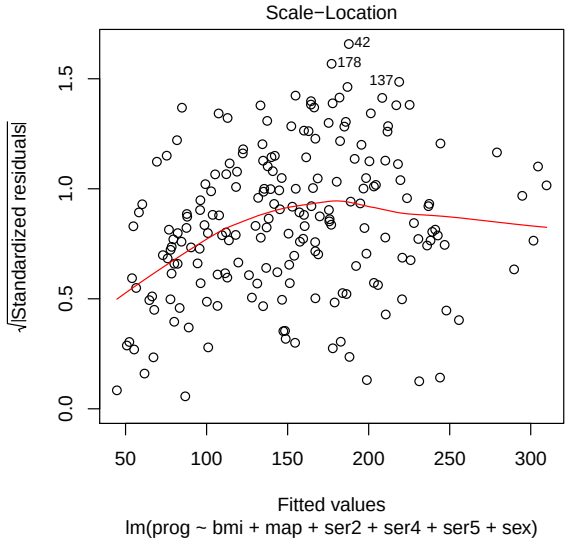
```
library(leaps)
Z=matrix(c(age,bmi,map,ser1,ser2,ser3,ser4,ser5,ser6,sex),ncol=10);
colnames(Z)=c("age","bmi","map","ser1","ser2","ser3","ser4","ser5","ser6","sex");
r=leaps(Z,prog)
t=(r$Cp==min(r$Cp))
colnames(Z)[r$whi[t]]
reg2=lm(prog~bmi+map+ser2+ser4+ser5+sex)
summary(reg2); plot(reg2)
```

Voici les résultats obtenus:

```
> colnames(Z)[r$whi[t]]
[1] "bmi"  "map"  "ser2" "ser4" "ser5" "sex"
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -291.5707    39.0335  -7.470 2.71e-12 ***
bmi            6.4745     0.9778   6.621 3.44e-10 ***
map            1.2302     0.3008   4.089 6.35e-05 ***
ser2          -0.6142     0.1653  -3.716 0.000265 ***
ser4          22.0946     4.6624   4.739 4.16e-06 ***
ser5          31.5945     9.1086   3.469 0.000645 ***
sex2         -21.9458     8.1080  -2.707 0.007405 **

Residual standard error: 51.92 on 193 degrees of freedom
Multiple R-squared:  0.5651,    Adjusted R-squared:  0.5516
F-statistic:  41.8 on 6 and 193 DF,  p-value: < 2.2e-16
```



Questions I.2: Qu'a-t-on fait en tapant ces commandes? Que concluez vous quant aux résultats et graphe obtenus? **(1pt)**

On tape maintenant les commandes suivantes:

```
library(MASS)
X=as.data.frame(Z)
reg3=stepAIC(lm(prog~.,data=X),k=log(200))
summary(reg3); plot(reg3)
```

Voici ce que l'on obtient à la fin des résultats:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -311.7870    36.4236  -8.560 3.19e-15 ***
bmi            8.1540     0.9918   8.221 2.67e-14 ***
ser5          53.7323     8.1519   6.591 3.90e-10 ***

Residual standard error: 55.9 on 197 degrees of freedom
Multiple R-squared:  0.4854,    Adjusted R-squared:  0.4801
F-statistic:  92.9 on 2 and 197 DF,  p-value: < 2.2e-16
```

Questions I.3: Qu'a-t-on fait en tapant ces commandes? Que concluez vous quant aux résultats obtenus? **(1pt)**

On effectue maintenant ceci:

```
Diab2=read.table("C:/Donnees/diabetes2.txt",header=TRUE)
Diab2$sex=as.factor(Diab2$sex)
new=data.frame(age=Diab2$age,bmi=Diab2$bmi,map=Diab2$map,ser1=Diab2$ser1,ser2=Diab2$ser2,
ser3=Diab2$ser3,ser4=Diab2$ser4,ser5=Diab2$ser5,ser6=Diab2$ser6,sex=Diab2$sex)
pred1=predict(reg1,new); pred2=predict(reg2,new); pred3=predict(reg3,new)
MSE1=sqrt(mean((Diab2$prog-pred1)^2)); MSE2=sqrt(mean((Diab2$prog-pred2)^2));
MSE3=sqrt(mean((Diab2$prog-pred3)^2)); MSE1; MSE2; MSE3
```

Et on obtient:

```
> MSE1; MSE2; MSE3
[1] 56.98467
[1] 57.47944
[1] 57.73582
```

Questions I.4: Qu'a-t-on fait en tapant ces commandes? Que concluez vous à partir des résultats obtenus? Est-ce surprenant? **(2pts)**