Université Paris I, Panthéon - Sorbonne

Première Année Master M.A.E.F. 2023 – 2024

Econométrie II

Contrôle continu n°2, avril 2024

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

1. Préambule **(Sur 4 points)**

   (a) Soit $a \in \mathbf{R}$ et $\lambda > 0$. On considère $f_a(u) = u^2 - 2\,a\,u + 2\,\lambda\,|u|$ pour $u \in \mathbf{R}$. Montrer que f_a admet un unique minimum atteint en $u_{\min} = \text{signe}(a)\,(|a| - \lambda)_+$, où $x_+ = \max(x, 0)$ pour $x \in \mathbf{R}$ **(1.5pts)**.

   (b) Soit $Z \in \mathcal{N}\big(0\,,\,1\big)$. Montrer que pour $\eta > 0$, $\mathbb{P}\big(|Z| \geq \eta\big) \leq e^{-\eta^2/2}$. Indication: on pourra par exemple étudier la fonction $g(x) = \mathbb{P}(Z \leq x) - \frac{1}{2}\,e^{-x^2/2}$ **(2.5pts)**.

2. Exercice théorique **(Sur 15 points)** Dans la suite, on notera $\|U\|_2^2 = {}^t\!U\,U$ pour tout $U \in \mathbf{R}^m$, $m \in \mathbf{N}^*$. On considère le modèle linéaire suivant:

$$Y = X\,\theta^* + \varepsilon$$

   où

   - $Y = {}^t(Y_1, \ldots, Y_n)$ est observé ($n \in \mathbf{N}^*$ connu);
   - $\varepsilon = {}^t(\varepsilon_1, \ldots, \varepsilon_n)$, où $(\varepsilon_k)_{k\in\mathbf{N}}$ est une suite de v.a.i.i.d. de loi $\mathcal{N}(0, \sigma_*^2)$ où $\sigma_*^2 > 0$ est inconnu;
   - $\theta^* = {}^t(\theta_1^*, \ldots, \theta_p^*) \in \mathbf{R}^p$ est un vecteur de paramètres inconnus ($p \in \mathbf{N}^*$ connu);
   - X est une matrice (n, p) composée de nombres réels connus et telle que ${}^t\!X\,X = I_p$, matrice identité de taille p et on supposera pour simplifier que si $X^{(j)}$ désigne la colonne j de X, alors $\|X^{(j)}\|_2 = 1$ pour $1 \leq j \leq p$.

   (a) Déterminer l'expression de l'estimateur par moindres carrés de θ et donner sa loi **(0.5pts)**. Déterminer le risque quadratique de cet estimateur **(1pt)**.

   (b) On note $s = \sum_{j=1}^p \mathbb{I}\!\mathrm{I}_{\theta_j^*\neq 0}$, on suppose $0 < s < p$ et on note $J^* = \big\{j \in \{1, \ldots, p\},\ \theta_j^* \neq 0\big\}$. Si J^* était connu, quel serait en fonction de s le risque quadratique de l'estimateur par moindres carrés noté $\widehat{\theta}^{oracle}$ **(1pt)**?

   (c) On suppose désormais J^* inconnu. Pour essayer d'améliorer le risque quadratique de l'estimateur, on définit $\widehat{\theta}_\lambda^L$ tel que:

$$\widehat{\theta}^L \in \underset{\theta={}^t(\theta_1,\ldots,\theta_p)\in\mathbf{R}^p}{\text{argmin}}\ g_\lambda(\theta) \qquad \text{où} \qquad g_\lambda(\theta) = \big\|Y - X\,\theta\big\|_2^2 + 2\,\lambda\,\|\theta\|_1,$$

   où $\lambda > 0$ est fixé et $\|\theta\|_1 = \sum_{j=1}^p |\theta_j|$. Montrer que $g_\lambda(\theta) = \|Y\|^2 + \sum_{j=1}^p f_{a_j}(\theta_j)$ où a_j est à préciser et f_a définie en préambule **(1pt)** et en déduire que $\widehat{\theta}^L = {}^t(\widehat{\theta}_1^L, \ldots, \widehat{\theta}_p^L)$ est unique et pour $1 \leq j \leq p$,

$$\widehat{\theta}_j^L = \text{signe}\big({}^t\!X^{(j)}\,Y\big)\,\big(\big|{}^t\!X^{(j)}\,Y\big| - \lambda\big)_+ \quad \textbf{(1.5pts)}.$$

   (d) On note $\eta = {}^t(\eta_1, \ldots, \eta_p)$ avec $\eta_j = {}^t\!\varepsilon\,X^{(j)}$ pour $j = 1, \ldots, p$. Quelle est la loi de η (justifier!) **(1pt)**? Soit A l'évenement:

$$A = \bigcap_{j=1}^n \big\{|\eta_j| \leq \lambda/2\big\}.$$

   Démontrer que $\mathbb{P}(A) \geq 1 - \sum_{j=1}^p \mathbb{P}\big(|\eta_j| > \lambda/2\big)$ **(1.5pts)**. En déduire avec l'aide du préambule qu'en choisissant $\lambda = \sigma_*\,\sqrt{8\,\kappa\,\ln(p)}$ avec $\kappa > 1$, alors $\mathbb{P}(A) \geq 1 - p^{1-\kappa}$ **(1pt)**.

   (e) Sachant l'événement A, montrer que $\widehat{\theta}_i^L = 0$ si $i \in \{1, \ldots, p\} \setminus J^*$ **(1.5pts)**.

   (f) Sachant l'événement A, montrer que $\big\|\widehat{\theta}^L - \theta^*\big\|_2 \leq \big\|\widehat{\theta}^{oracle} - \theta^*\big\|_2 + \sqrt{\sum_{j\in J^*}\lambda^2}$ **(1.5pts)**, puis montrer que

$$\big\|\widehat{\theta}^{oracle} - \theta^*\big\|_2 \leq \sqrt{s\,\frac{\lambda^2}{4}} + \sqrt{s\,\lambda^2} \leq \frac{3}{2}\,\lambda\,\sqrt{s} \qquad \text{sachant } A \textbf{ (1.5pts)}.$$

   En déduire, qu'il existe C tel que $\mathbb{P}\Big(\big\|\widehat{\theta}^L - \theta^*\big\|_2^2 \leq C\,\sigma_*^2\,s\,\ln(p)\Big) \geq 1 - p^{1-\kappa}$ **(2pts)**.

3. **(Sur 6 points)** Exercice de TP utilisant le logiciel `R`

(a) Soit la base de données `deathdata`, dans laquelle le taux de mortalité (`death-rate`) ainsi que 15 autres variables quantitatives (températures, pollution, densité de population, données socio-économiques,...) sont répertoriées dans 60 villes des USA en 2010. On commence par effectuer les commandes suivantes

```
death_lm = lm(death_rate ~ .,data=deathdata)
summary(death_lm)
```

Voici les résultats:

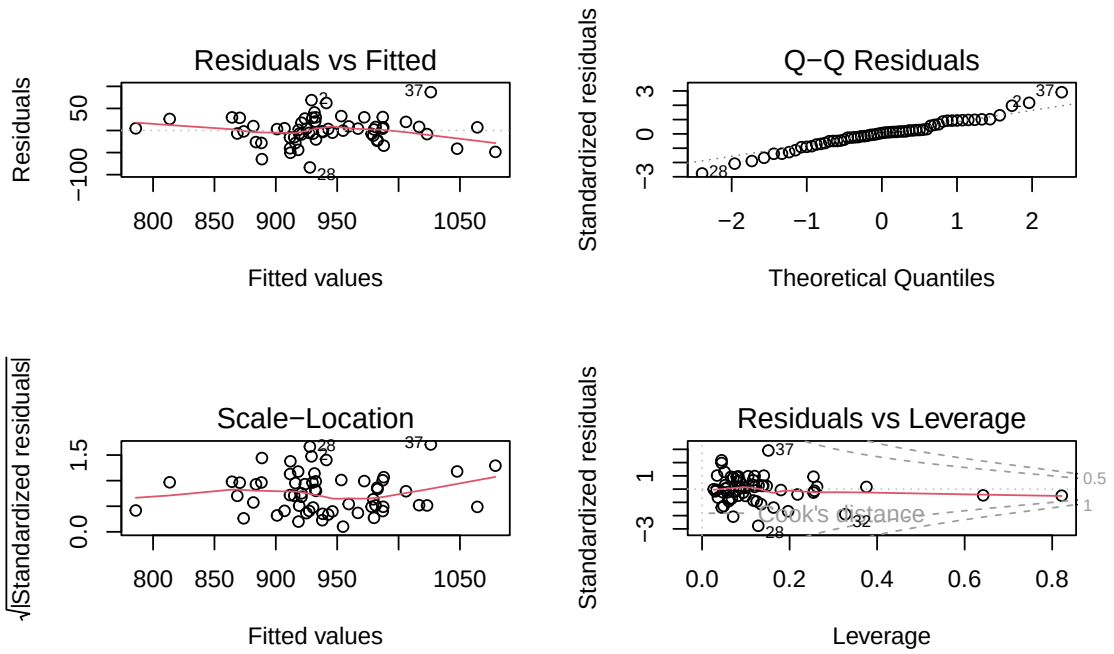
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.863e+03  4.108e+02   4.535  4.4e-05 ***
Precipitation             2.072e+00  8.418e-01   2.462  0.01781 *
January_temperature      -2.178e+00  6.752e-01  -3.225  0.00238 **
July_temperature         -2.834e+00  1.771e+00  -1.600  0.11670
percent_65_or_older      -1.404e+01  7.746e+00  -1.813  0.07670 .
household_size           -1.154e+02  6.200e+01  -1.862  0.06933 .
schooling_over_22        -2.425e+01  1.121e+01  -2.163  0.03605 *
full_kitchens            -1.146e+00  1.467e+00  -0.781  0.43871
urban_population_density  1.004e-02  4.123e-03   2.435  0.01899 *
nonwhite_population       3.533e+00  1.282e+00   2.755  0.00850 **
office_workers            5.229e-01  1.551e+00   0.337  0.73760
poor_families             2.671e-01  2.565e+00   0.104  0.91755
hydrocarbons             -8.890e-01  4.524e-01  -1.965  0.05574 .
oxides_of_Nitrogen        1.866e+00  9.345e-01   1.997  0.05201 .
Sulfur_Dioxide           -3.447e-02  1.423e-01  -0.242  0.80968
humidity                  5.331e-01  1.052e+00   0.507  0.61474

Residual standard error: 32.33 on 44 degrees of freedom
Multiple R-squared:  0.7985,Adjusted R-squared:  0.7298
F-statistic: 11.63 on 15 and 44 DF,  p-value: 9.56e-11
```

Questions 1: Expliquer ce qui a été fait. Expliquer précisément et formellement comment la valeur `0.43871` *est obtenue et la conclusion qui en découle. Donner la formule permettant d'obtenir la valeur* `0.7985` *et expliquer ce que l'on peut en déduire* **(2.5pts)***?*

(b) On tape ensuite les commandes:

```
reg1=stepAIC(death_lm,k=log(60),direction = c("backward"))
summary(reg1)
par(mfrow=c(2,2)); plot(reg1)
```

Voici la figure obtenue:



et les résultats numériques:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.034e+03  8.189e+01  12.628  < 2e-16 ***
Precipitation             1.219e+00  6.206e-01   1.964 0.054846 .
January_temperature      -1.672e+00  4.256e-01  -3.929 0.000253 ***
schooling_over_22        -1.578e+01  6.126e+00  -2.577 0.012849 *
urban_population_density  9.288e-03  3.227e-03   2.878 0.005793 **
nonwhite_population       4.081e+00  5.675e-01   7.191 2.45e-09 ***
hydrocarbons             -7.373e-01  3.083e-01  -2.392 0.020422 *
oxides_of_Nitrogen        1.577e+00  5.888e-01   2.677 0.009903 **

Residual standard error: 32.39 on 52 degrees of freedom
Multiple R-squared:  0.761,Adjusted R-squared:  0.7289
F-statistic: 23.66 on 7 and 52 DF,  p-value: 4.358e-14
```

Questions 2: Qu'a-t-on fait avec ces commandes et pourquoi l'a-t-on fait? Pourquoi la variable `Precipitation` *est-elle considérée alors que sa p-value est supérieure à* 0.05*? Pourquoi choisir ce modèle alors qu'il a un* R^2 *plus faible que le précédent? Que peut-on dire à la suite des graphes?* **(2pts)**

(c) On a enfin tapé les commandes:

```
j=c(c(1:27),c(29:36),c(38:60))
deathdata2=deathdata[j,]
death_lm2 = lm(death_rate ~ .,data=deathdata2)
reg2=stepAIC(death_lm2,k=log(58),direction = c("backward"))
summary(reg2)
```

Voici les résultats obtenus:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.070e+03  7.257e+01  14.737  < 2e-16 ***
Precipitation             1.216e+00  5.317e-01   2.288  0.02643 *
January_temperature      -1.838e+00  3.672e-01  -5.006 7.27e-06 ***
schooling_over_22        -1.794e+01  5.491e+00  -3.267  0.00197 **
urban_population_density  9.426e-03  2.806e-03   3.358  0.00151 **
nonwhite_population       3.553e+00  4.986e-01   7.127 3.80e-09 ***
hydrocarbons             -6.678e-01  2.643e-01  -2.527  0.01471 *
oxides_of_Nitrogen        1.453e+00  5.043e-01   2.882  0.00581 **

Residual standard error: 27.68 on 50 degrees of freedom
Multiple R-squared:  0.7974,Adjusted R-squared:  0.769
F-statistic: 28.11 on 7 and 50 DF,  p-value: 3.009e-15
```

Questions 3: Expliquez ce qui a été fait et ce que l'on peut en conclure **(1.5pts)**