Université Paris I, Panthéon - Sorbonne

Première Année Master M.A.E.F. 2024 – 2025

Econométrie II

Examen terminal, mai 2025

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

Exercice théorique (Sur 14 points)

Soit $Y = {}^t(Y_1, \ldots, Y_{2n})$ une variable réelle observée pour n individus et p variables exogènes $X^{(j)} = {}^t(X_1^{(j)}, \ldots, X_{2n}^{(j)})$ observées pour ces $2n$ individus (avec $n > p \geq 1$). En notant $X^{(0)} = {}^t(1, \ldots, 1)$ ($2n$ fois), on supposera que les matrices $Z_1 = (X_i^{(j)})_{1 \leq i \leq n, 0 \leq j \leq p}$ et $Z_2 = (X_i^{(j)})_{n+1 \leq i \leq 2n, 0 \leq j \leq p}$ sont de rang $p+1$. On supposera également qu'il existe un vecteur $\theta = (\theta_j)_{0 \leq j \leq p} \in \mathbf{R}^{p+1}$, vecteur inconnu, tel que:

$$Y_i = \theta_0 + \sum_{j=1}^{p} \theta_j \, X_i^{(j)} + \varepsilon_i \quad \text{pour tout } 1 \leq i \leq 2n,$$

avec $(\varepsilon_i)_{i \in \mathbf{N}}$ une suite de variables gaussiennes centrées indépendantes telles que $\mathrm{var}(\varepsilon_i) = \sigma_i^2 < \infty$ pour tout i.
Dans la suite, on veut tester l'homoscédasticité du modèle, donc trouver une statistique de test pour décider entre:

$$\mathrm{H}_0 : \ \exists \sigma^2 > 0 \text{ tel que pour tout } i \in \mathbf{N}, \, \sigma_i^2 = \sigma^2 \qquad \text{contre} \qquad \mathrm{H}_1 : \ \text{les } \sigma_i^2 \text{ ne sont pas toutes égales}$$

1. Un premier test va être mis en place de la manière suivante: on effectue deux régressions par moindres carrés des Y_i par les $X_i^{(j)}$, la première pour $Y^{(1)} = {}^t(Y_1, \ldots, Y_n)$ et cela fournit l'estimateur $\widehat{\theta}^{(1)}$, la seconde pour $Y^{(2)} = {}^t(Y_{n+1}, \ldots, Y_{2n})$ et cela fournit l'estimateur $\widehat{\theta}^{(2)}$.

   (a) Sous l'hypothèse H_0, en justifiant, donner la loi de $Y^{(1)} - Z_1 \, \widehat{\theta}^{(1)}$, puis celle de $\left\| Y^{(1)} - Z_1 \, \widehat{\theta}^{(1)} \right\|^2$ **(1pt)**.

   (b) On considère la statistique de test
   $$\widehat{S}_n = \frac{\left\| Y^{(1)} - Z_1 \, \widehat{\theta}^{(1)} \right\|^2}{\left\| Y^{(2)} - Z_2 \, \widehat{\theta}^{(2)} \right\|^2}.$$
   Sous l'hypothèse H_0, en justifiant, donner la loi de $\widehat{S}_n$ **(2pts)**.

   (c) Lorsque $n \to \infty$, déterminer la limite en loi de $\widehat{S}_n$ sous H_0 **(1pt)**.

   (d) Si on suppose que $\sigma_i^2 = \sigma^2$ pour $i = 1, \ldots, n$ et $\sigma_i^2 = \gamma^2$ pour $i = n+1, \ldots, 2n$, quelle serait la limite en probabilité de $\widehat{S}_n$ **(1pt)**?

   (e) Expliquer concrètement quand accepter H_0 avec un risque de 5% en utilisant $\widehat{S}_n$ **(1pt)**.

2. On propose une autre démarche. On note $Y = {}^t(Y_1, \ldots, Y_{2n})$ et $Z = (X_i^{(j)})_{1 \leq i \leq 2n, 0 \leq j \leq p}$, $\widehat{\theta}$ l'estimateur par moindres carrés de θ par régression de Y par rapport aux $X^{(j)}$, $\widehat{Y} = (\widehat{Y}_i)_{1 \leq i \leq 2n} = Z \, \widehat{\theta}$ et $\widehat{\varepsilon} = (\widehat{\varepsilon}_i)_{1 \leq i \leq 2n} = Y - \widehat{Y}$.
   L'idée est de tester H_0 contre H_1': Il existe un vecteur $\beta = {}^t(\beta_1, \ldots, \beta_p) \in \mathbf{R}^p$, $\beta \neq 0$, tel que
   $$\mathbb{E}\big[\varepsilon_i^2\big] = \sigma^2 + \sum_{j=1}^{p} \beta_j \, X_i^{(j)} \quad \text{pour tout } i = 1, \ldots, 2n.$$

   (a) Expliquer en quoi l'hypothèse H_1' est une hypothèse d'hétéroscédasticité. Sous H_1', que vaut $\mathbb{E}[\widehat{\theta}]$ **(1pt)**?

   (b) Soit la matrice $Z \left({}^t Z \, Z\right)^t Z = (p_{ij})_{1 \leq i,j \leq p+1}$. Soit $C1$ la condition: $\max_{1 \leq i \leq 2n}(|p_{ii}|) \underset{n \to \infty}{\longrightarrow} 0$. Montrer que l'hypothèse H_0 et $C1$ induisent que $\mathrm{var}\big(\widehat{\varepsilon}_i\big) \underset{n \to \infty}{\longrightarrow} \sigma^2$ pour tout $i = 1, \ldots, n$ **(1pt)**.

   (c) Sous $C1$ et quand n est grand, si une linéarité apparaît sur le nuage de points des $\widehat{\varepsilon}_i^2$ en fonction des $\widehat{Y}_i$, cela caractérise plutôt H_0 ou H_1' **(1.5pts)**?

   (d) Montrer que sous l'hypothèse H_0, $\mathbb{E}\big[(\varepsilon_i^2 - \sigma^2)^2\big] = 2 \, \sigma^4$ **(1pt)**. En déduire que pour tout $j \in \{0, \ldots, p\}$,
   $$\frac{1}{\sqrt{\sum_{i=1}^{2n} (X_i^{(j)})^2}} \sum_{i=1}^{2n} \big(\varepsilon_i^2 - \sigma^2\big) \, X_i^{(j)} \ \underset{n \to \infty}{\overset{\mathcal{L}}{\longrightarrow}} \ \mathcal{N}\big(0 \, , \, 2 \, \sigma^4\big) \quad \textbf{(2pts)}.$$

Plus généralement on acceptera (ne pas le démontrer) sous H_0 le TLC multidimensionnel

$$\big({}^t Z\,Z\big)^{-1/2}\sum_{i=1}^{2n}\big(\varepsilon_i^2-\sigma^2\big)\,\big(X_i^{(j)}\big)_{0\leq j\leq p}\ \xrightarrow[n\to\infty]{\mathcal{L}}\ \mathcal{N}\big(0\,,\,2\,\sigma^4\,I_{p+1}\big). \tag{1}$$

(e) Soit la condition $C2$: $\displaystyle \frac{1}{\sqrt{2n}}\,\frac{\sum_{i=1}^{2n}X_i^{(j)}}{\sqrt{\sum_{i=1}^{2n}(X_i^{(j)})^2}}\ \xrightarrow[n\to\infty]{}\ 0$ pour $j\in\{1,\dots,p\}$. Montrer que si $(U_i)_{i\in\mathbf{N}}$ est une suite de v.a.i.i.d. centrées de variance finie alors la convergence $C2$ a lieu en probabilité pour $(X_i^{(j)})=(U_i)$ **(3pts)**. Montrer que sous H_0 et $C2$, avec $Z^-=(X_i^{(j)})_{1\leq i\leq 2n,1\leq j\leq p}$ et $\widehat\sigma^2=\frac{1}{2n-p-1}\sum_{i=1}^{2n}\widehat\varepsilon_i^2$,

$$\big({}^t Z^-\,Z^-\big)^{-1/2}\sum_{i=1}^{2n}\big(\varepsilon_i^2-\widehat\sigma^2\big)\,\big(X_i^{(j)}\big)_{1\leq j\leq p}\ \xrightarrow[n\to\infty]{\mathcal{L}}\ \mathcal{N}\big(0\,,\,2\,\sigma^4\,I_p\big)\quad \textbf{(2.5pts)}.$$

(f) Sous H_0, $C1$ et $C2$, on peut aussi montrer (ne pas le faire) que:

$$\big({}^t Z^-\,Z^-\big)^{-1/2}\sum_{i=1}^{2n}\big(\widehat\varepsilon_i^2-\widehat\sigma^2\big)\,\big(X_i^{(j)}\big)_{1\leq j\leq p}\ \xrightarrow[n\to\infty]{\mathcal{L}}\ \mathcal{N}\big(0\,,\,2\,\sigma^4\,I_p\big).$$

On considère la statistique de test : $\widehat T_n={}^tE_n\,\big({}^tZ^-\,Z^-\big)^{-1}E_n$ où $E_n=\sum_{i=1}^{2n}\big(\widehat\varepsilon_i^2-\widehat\sigma^2\big)\,\big(X_i^{(j)}\big)_{1\leq j\leq p}$. Déterminer la loi limite de $\widehat T_n$ sous H_0 **(1.5pts)**. Comment l'utiliser pour tester H_0 **(1pt)**? Intuitivement, que se passe-t-il pour $\widehat T_n$ sous H_1' **(0.5pts)**?

Proof. 1. (a) $Y^{(1)}-Z_1\,\widehat\theta^{(1)}=P_{[Z_1]^\perp}(\varepsilon_i)_{1\leq i\leq n}\ \overset{\mathcal{L}}{\sim}\ \mathcal{N}(0\,,\,\sigma^2\,P_{[Z_1]^\perp})$.
D'après le Théorème de Cochran, $\big\|Y^{(1)}-Z_1\,\widehat\theta^{(1)}\big\|^2\ \overset{\mathcal{L}}{\sim}\ \sigma^2\,\chi^2(\dim([Z_1]^\perp))=\sigma^2\,\chi^2(n-p-1)$.

(b) On peut écrire que $\widehat S_n=\frac{\frac{1}{(n-p-1)\sigma^2}\,\|Y^{(1)}-Z_1\,\widehat\theta^{(1)}\|^2}{\frac{1}{(n-p-1)\sigma^2}\,\|Y^{(2)}-Z_2\,\widehat\theta^{(2)}\|^2}$. Comme on a également $\big\|Y^{(2)}-Z_2\,\widehat\theta^{(2)}\big\|^2\ \overset{\mathcal{L}}{\sim}\ \sigma^2\,\chi^2(n-p-1)$, et $Y^{(1)}-Z_1\,\widehat\theta^{(1)}=P_{[Z_1]^\perp}(\varepsilon_i)_{1\leq i\leq n}$ alors que $Y^{(2)}-Z_2\,\widehat\theta^{(2)}=P_{[Z_2]^\perp}(\varepsilon_i)_{n+1\leq i\leq 2n}$, donc $Y^{(1)}-Z_1\,\widehat\theta^{(1)}$ et $Y^{(2)}-Z_2\,\widehat\theta^{(2)}$ indépendants, alors le numérateur et le dénominateur de $\widehat S_n$ sont indépendants: par définition $\widehat S_n$ suit une loi de Fisher $F(n-p-1,n-p-1)$.

(c) Il est clair, par la loi des grands nombres que $\frac{1}{n-p-1}\,\big\|Y^{(1)}-Z_1\,\widehat\theta^{(1)}\big\|^2\ \xrightarrow[n\to+\infty]{\mathcal{P}}\ \sigma^2$, de même $\frac{1}{n-p-1}\,\big\|Y^{(2)}-Z_2\,\widehat\theta^{(2)}\big\|^2\ \xrightarrow[n\to+\infty]{\mathcal{P}}\ \sigma^2$, et comme la fonction $g(x,y)$ est continue sur $]0,\infty[^2$, alors $\widehat S_n\ \xrightarrow[n\to+\infty]{\mathcal{P}}\ 1$.

(d) De la même manière, on montre facilement que sous cette hypothèse $\widehat S_n\ \xrightarrow[n\to+\infty]{\mathcal{P}}\ \frac{\sigma^2}{\gamma^2}$.

(e) Soit $q_{0.025}$ et $q_{0.975}$ les quantiles d'ordre 0.025 et 0.975 de la loi $F(n-p-1,n-p-1)$. On acceptera H_0 si $\widehat S_n\in[q_{0.025}\,,\,q_{0.975}]$.

2. (a) Il est clair que H_1' est un cas particulier de H_1.
On a $\mathbb{E}[\widehat\theta]=({}^tZ\,Z)^{-1}\,{}^tZ\,\mathbb{E}[Y]=\theta+({}^tZ\,Z)^{-1}\,{}^tZ\,\mathbb{E}[\varepsilon]=\theta$.

(b) Sous H_0, on sait que $\widehat\varepsilon=P_{[Z]^\perp}\,\varepsilon$, et ainsi $\mathrm{var}(\widehat\varepsilon_i)=\sigma^2(1-p_{ii})$ pour tout $i\in\{1,\dots,2n\}$. D'où le résultat.

(c) On sait que $\widehat Y_i=\widehat\theta_0+\widehat\theta_1\,X_i^{(1)}+\cdots+\widehat\theta_p\,X_i^{(p)}$. S'il y a une linéarité dans le nuage de points, cela signifie que $\widehat\varepsilon_i^2=\mu+\nu\,\widehat Y_i+\xi_i$ pour $i=1,\dots,2n$, avec des (ξ_i) suite de v.a. centrées. De ceci, 2 possibilités: si le slope de cette droite est nul, soit $\nu=0$, car cela signifié que $\mathbb{E}[\widehat\varepsilon_i^2]\simeq\mu$: ce sera plutôt l'hypothèse H_0, sinon ce sera H_1' car on a alors $\mathrm{var}(\widehat\varepsilon_i)\simeq\mathrm{var}(\varepsilon_i)\simeq\mu+\nu\,\widehat Y_i\simeq\mu+\nu\,\big(\widehat\theta_0+\widehat\theta_1\,X_i^{(1)}+\cdots+\widehat\theta_p\,X_i^{(p)}\big)$ ce qui est bien une combinaison linéaire comme dans H_1'.

(d) Sous H_0, $\mathbb{E}\big[(\varepsilon_i^2-\sigma^2)^2\big]=\sigma^4\,\mathrm{var}(U)$ où U suit une loi $\chi^2(1)$ et on sait que pour une une v.a. suivant un $\chi^2(m)$ sa variance est $2\,m$. Donc ici $\mathrm{var}(U)=2$ et $\mathbb{E}\big[(\varepsilon_i^2-\sigma^2)^2\big]=2\,\sigma^4$.
On va appliquer le Théorème de Lindeberg avec $a_i^{(n)}=X_i^{(j)}/\sqrt{\sum_{i=1}^{2n}(X_i^{(j)})^2}$ ce qui fait que l'on a bien $\sum_{i=1}^{2n}\big(a_i^{(n)}\big)^2=1$, $\xi_i=(\varepsilon_i^2-\sigma^2)/\sqrt{2\,\sigma^4}$, ce qui fait que les ξ_i sont indépendants (car les ε_i le sont), et $\mathbb{E}[\xi_i]=0$ puisque sous H_0, $\mathbb{E}[\varepsilon_i^2]=\sigma^2$, et $\mathrm{var}(\xi_i)=1$. Toutes les conditions sont réunies et ainsi $\sum_{i=1}^{2n}a_i^{(n)}\,\xi_i\ \xrightarrow[n\to\infty]{\mathcal{L}}\ \mathcal{N}(0,1)$, d'où le résultat.

(e) Soit $(U_i)_{i\in\mathbf{N}}$ est une suite de v.a.i.i.d. centrées de variance finie ρ^2. Alors en notant $\overline U_{2n}=\frac{1}{2n}\sum_{i=1}^{2n}U_i$, on a $\overline U_{2n}\ \xrightarrow[n\to+\infty]{\mathcal{P}}\ 0$ par la loi forte des grands nombres, donc également $\overline U_{2n}/\rho\ \xrightarrow[n\to+\infty]{\mathcal{P}}\ 0$. De plus $\frac{1}{2n}\sum_{i=1}^{2n}U_i^2\ \xrightarrow[n\to+\infty]{\mathcal{P}}\ \rho^2$, également par la LGN. Donc par le Lemme de Slutsky,

$$\frac{\overline U_{2n}}{\sqrt{\frac{1}{2n}\sum_{i=1}^{2n}U_i^2}}=\frac{1}{\sqrt{2n}}\,\frac{\sum_{i=1}^{2n}U_i}{\sqrt{\sum_{i=1}^{2n}U_i^2}}\ \xrightarrow[n\to+\infty]{\mathcal{P}}\ 0.$$

On peut écrire que l'on obtiendra le résultats demandé si

$$({}^tZ^-\,Z^-)^{-1/2}\sum_{i=1}^{2n}(\sigma^2-\widehat\sigma^2)\,\big(X_i^{(j)}\big)_{1\leq j\leq p}\ \xrightarrow[n\to+\infty]{\mathcal{P}}\ 0\quad\Longrightarrow\quad(\sigma^2-\widehat\sigma^2)\big({}^tZ^-\,Z^-\big)^{-1/2}\sum_{i=1}^{2n}\big(X_i^{(j)}\big)_{1\leq j\leq p}\ \xrightarrow[n\to+\infty]{\mathcal{P}}\ 0$$

$\square$

Exercice de TP utilisant le logiciel `R` (Sur 9 points)

On s'intéresse à une base de données `Guns` relative à la criminalité annuelle dans les 50 états des USA entre 1977 et 1999. Nous allons étudier la variable `murder`,taux d'homicides pour 100000 habitants, en fonction d'autres variables:

- `year`: l'année, soit $1977, 1978, \ldots, 1998, 1999$.
- `state`: le nom d'un des 50 états des USA.
- `density`: la densité de population de l'état.
- `population`: la population totale de l'état.
- `income`: le PNB par habitant de l'état.
- `male`: le pourcentage de jeunes hommes de 10 à 29 ans dans la population de l'état.
- `prisoners`: le taux de prisonniers par état pour 100000 habitants.
- `law`: existence d'une loi sur le port d'armes (`yes` or `no`).

1. Une première étape réside dans la préparation des données:

```
Guns$law=as.factor(Guns$law)
Guns$state=as.factor(Guns$state)
str(Guns);
```

Ce qui amène les résultats numériques:

```
'data.frame': 1150 obs. of  9 variables:
 $ year      : int  1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 ...
 $ murder    : num  14.2 13.3 13.2 13.2 11.9 10.6 9.2 9.4 9.8 10.1 ...
 $ prisoners : int  83 94 144 141 149 183 215 243 256 267 ...
 $ male      : num  18.2 18 17.8 17.7 17.7 ...
 $ population: num  3.78 3.83 3.87 3.9 3.92 ...
 $ income    : num  9563 9932 9877 9541 9548 ...
 $ density   : num  0.0746 0.0756 0.0762 0.0768 0.0772 ...
 $ state     : Factor w/ 50 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ law       : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Questions 1: Qu'a-t-on fait ici et pourquoi **(0.5pts)**?

2. On commence ensuite par étudier l'effet global du temps sur la variable `murder`:

```
lin0=lm(murder ~ year, data = Guns)
summary(lin0)
```

Voici les résultats:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 169.13391   32.84502   5.149 3.07e-07 ***
year         -0.08164    0.01652  -4.941 8.91e-07 ***

Residual standard error: 3.716 on 1148 degrees of freedom
Multiple R-squared:  0.02083,Adjusted R-squared:  0.01997
F-statistic: 24.42 on 1 and 1148 DF,  p-value: 8.906e-07
```

Questions 2: Ecrire formellement le modèle considéré. Qu'indique la valeur `8.91e-07`*? Quelle sera la différence entre le taux d'homicides en 1999 et celui prévu par ce modèle en 2025 (en arrondissant)? Que conclure de cette première étude* **(1.5pts)**?

3. On tape ensuite les commandes:

```
lin1=lm(murder ~ year+state, data = Guns)
summary(lin1); anova(lin1)
```

On obtient les résultats numériques:

```
> summary(lin1)

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        173.127036  12.403257  13.958  < 2e-16 ***
year                -0.081640   0.006237 -13.089  < 2e-16 ***
stateAlaska         -1.043478   0.413737  -2.522 0.011807 *
stateArizona        -2.330435   0.413737  -5.633 2.25e-08 ***
     :                    :          :        :       :
     :                    :          :        :       :
stateWyoming        -6.273913   0.413737 -15.164  < 2e-16 ***

Residual standard error: 1.403 on 1099 degrees of freedom
Multiple R-squared:  0.8664,Adjusted R-squared:  0.8603
F-statistic: 142.5 on 50 and 1099 DF,  p-value: < 2.2e-16

> anova(lin1)
Analysis of Variance Table

Response: murder
            Df  Sum Sq Mean Sq F value    Pr(>F)
year         1   337.3  337.26  171.32 < 2.2e-16 ***
state       49 13692.6  279.44  141.95 < 2.2e-16 ***
Residuals 1099  2163.4    1.97
```

Questions 3: Préciser le modèle formel qui a été mis en place. Que conclure des résultats numériques? Quel test précisément a été effectué pour obtenir la valeur `141.95` **(1.5pts)***?*

4. On tape ensuite les commandes (on rappelle que la commande `V1*V2` permet d'obtenir un modèle avec interactions):

```
lin2=lm(murder ~ year*state, data = Guns)
summary(lin2); anova(lin2)
BIC(lin0,lin1,lin2); anova(lin0,lin1); anova(lin1,lin2)
```

On obtient les résultats numériques:

```
> summary(lin2)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.982e+02  7.725e+01   3.861 0.000120 ***
year                    -1.446e-01  3.886e-02  -3.721 0.000209 ***
stateAlaska              3.036e+02  1.092e+02   2.779 0.005542 **
stateArizona            -2.557e+02  1.092e+02  -2.341 0.019417 *
     :                        :          :        :       :
     :                        :          :        :       :
stateWyoming             1.022e+02  1.092e+02   0.935 0.349910
year:stateAlaska        -1.533e-01  5.495e-02  -2.789 0.005382 **
year:stateArizona        1.275e-01  5.495e-02   2.320 0.020548 *
     :                        :          :        :       :
     :                        :          :        :       :
year:stateWyoming       -5.455e-02  5.495e-02  -0.993 0.321125

Residual standard error: 1.236 on 1050 degrees of freedom
Multiple R-squared:  0.9009,Adjusted R-squared:  0.8916
F-statistic: 96.45 on 99 and 1050 DF,  p-value: < 2.2e-16

> anova(lin2)
Analysis of Variance Table

Response: murder
             Df  Sum Sq Mean Sq  F value    Pr(>F)
```

```
year          1   337.3  337.26 220.7273 < 2.2e-16 ***
state        49 13692.6  279.44 182.8878 < 2.2e-16 ***
year:state   49   559.1   11.41   7.4679 < 2.2e-16 ***

> BIC(lin0,lin1,lin2)
      df      BIC
lin0   3 6302.055
lin1  52 4356.759
lin2 101 4358.242

> anova(lin0,lin1)
Analysis of Variance Table

Model 1: murder ~ year
Model 2: murder ~ year + state
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   1148 15856.0
2   1099  2163.4 49     13693 141.95 < 2.2e-16 ***

> anova(lin1,lin2)
Analysis of Variance Table

Model 1: murder ~ year + state
Model 2: murder ~ year * state
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1099 2163.4
2   1050 1604.3 49    559.11 7.4679 < 2.2e-16 ***
```

Questions 4: Ecrire formellement le modèle considéré. Interpréter les résultats obtenus, notamment ceux des commandes `anova`*. Quel test a été précisément effectué pour obtenir la valeur numérique* `7.4679`*? Des trois modèles proposés, lequel choisir* **(1.5pts)***?*

5. On tape ensuite les commandes:

```
lin3=lm(murder~.,data=Guns)
library(MASS)
lin4=stepAIC(lin3,k=log(1150))
summary(lin4); BIC(lin0,lin1,lin2,lin3,lin4); plot(lin4)
```

Avec pour résultats numériques et graphe:

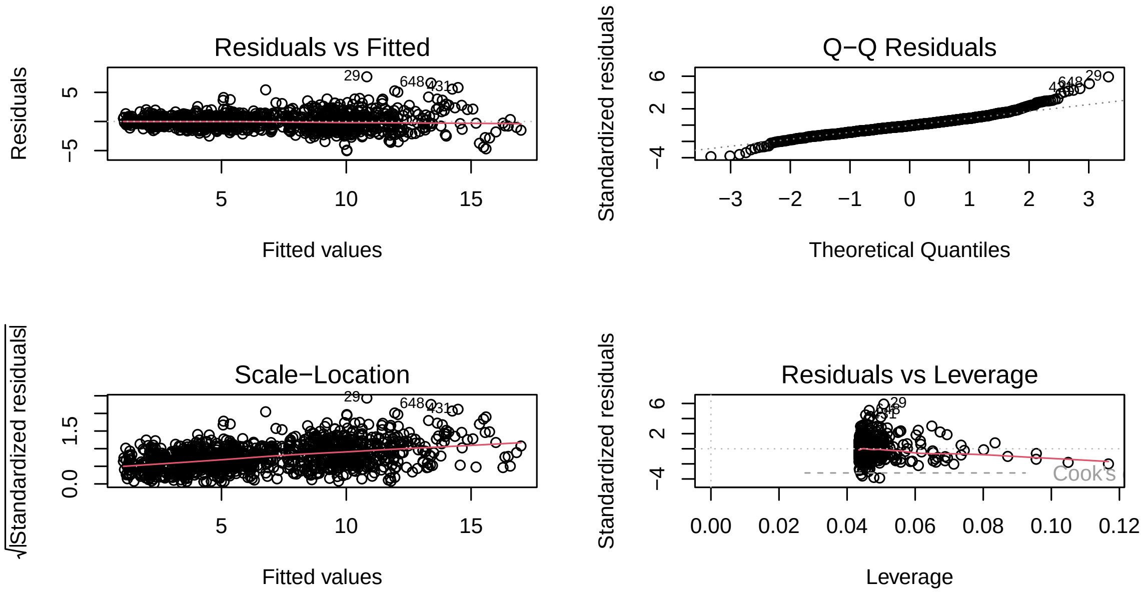
```
> summary(lin4)

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.275e+01  4.736e-01  26.928  < 2e-16 ***
prisoners          -7.602e-03  6.916e-04 -10.992  < 2e-16 ***
population         -3.547e-01  6.865e-02  -5.168 2.81e-07 ***
income              1.610e-04  4.008e-05   4.016 6.33e-05 ***
stateAlaska        -3.530e+00  5.330e-01  -6.623 5.51e-11 ***
stateArizona       -2.725e+00  3.980e-01  -6.846 1.26e-11 ***
     :                     :          :        :       :
     :                     :          :        :       :
stateWyoming       -8.801e+00  4.776e-01 -18.427  < 2e-16 ***

Residual standard error: 1.33 on 1097 degrees of freedom
Multiple R-squared:  0.8801,Adjusted R-squared:  0.8744
F-statistic: 154.8 on 52 and 1097 DF,  p-value: < 2.2e-16

> BIC(lin0,lin1,lin2,lin3,lin4)
      df      BIC
lin0   3 6302.055
lin1  52 4356.759
lin2 101 4358.242
```

```
lin3   58 4272.193
lin4   54 4246.466
```



Questions 5: Qu'est-ce qui a été fait ici? Formaliser le modèle obtenu, préciser ses avantages et défauts **(1.5pts)** *?*

6. On tape ensuite les commandes:

```
BX=boxcox(lin4,plotit = TRUE,lambda = seq(-5,5,0.01),data=Guns)
ind=which(BX$y==max(BX$y)); lambda=BX$x[ind]; lambda
regBC=lm(I(log(murder))~ prisoners + population + income + state, data = Guns)
summary(regBC); plot(regBC)
```

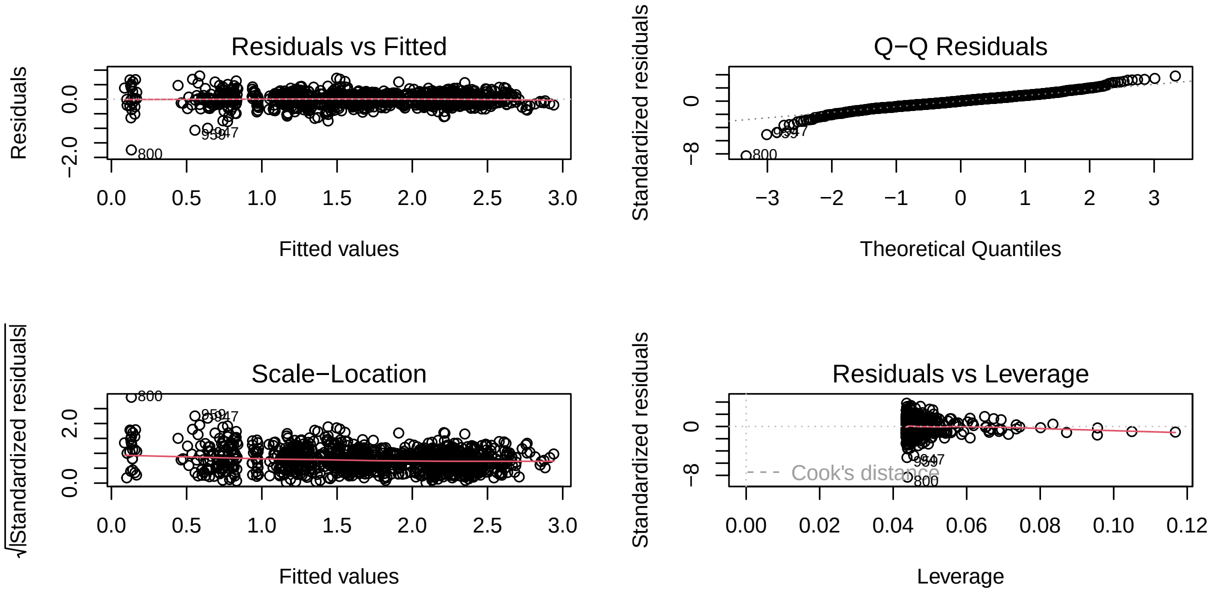
Avec pour résultats numériques et graphe:

```
> ind=which(BX$y==max(BX$y)); lambda=BX$x[ind]; lambda
[1] 0.37

> summary(regBC)

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.645e+00  7.645e-02  34.603  < 2e-16 ***
prisoners           -9.371e-04  1.116e-04  -8.394  < 2e-16 ***
population          -1.908e-02  1.108e-02  -1.722 0.085322 .
income               7.826e-06  6.470e-06   1.210 0.226687
stateAlaska         -2.767e-01  8.605e-02  -3.216 0.001339 **
stateArizona        -2.564e-01  6.425e-02  -3.991 7.02e-05 ***
     :                    :          :         :       :
     :                    :          :         :       :
stateWyoming        -1.136e+00  7.710e-02 -14.733  < 2e-16 ***

Residual standard error: 0.2148 on 1097 degrees of freedom
Multiple R-squared:  0.8945,Adjusted R-squared:  0.8895
F-statistic: 178.8 on 52 and 1097 DF,  p-value: < 2.2e-16
```



Questions 6: Qu'est-ce qui a été effectué ici? Formaliser le modèle obtenu et que conclure des graphes **(1pt)**?

7. On tape enfin les commandes:

```
reglaw1=glm(murder~law,data=Guns)
summary(reglaw1); Anova(reglaw1)
reglaw2=glm(law~murder,family=binomial(link="logit"),na.action=na.pass,data=Guns)
summary(reglaw2)
```

Voici les résultats:

```
> summary(reglaw1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.3434     0.1241  59.191  < 2e-16 ***
lawyes       -2.0595     0.2492  -8.264 3.85e-16 ***

> Anova(reglaw1)
Analysis of Deviance Table (Type II tests)

Response: murder
    LR Chisq Df Pr(>Chisq)
law   68.295  1  < 2.2e-16 ***

> summary(reglaw2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.05314    0.14215  -0.374    0.709
murder      -0.16919    0.02165  -7.813 5.58e-15 ***
```

Questions 7: Préciser les 2 modèles formels considérés. Comment interpréter les résultats obtenus quant à l'influence d'une loi sur le port d'arme et le taux d'homicides **(1.5pts)**?