Université Paris I, Panthéon - Sorbonne

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

Econométrie II

Examen terminal, mai 2024

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

1. **Exercice 1 (Sur 13 points)**

Soit $Y = {}^t(Y_1, \ldots, Y_n)$ une variable réelle observée pour n individus et p variables exogènes $X^{(j)} = {}^t(X_1^{(j)}, \ldots, X_n^{(j)})$ observées pour ces n individus pour $1 \leq j \leq p$ (avec $n > p \geq 1$). On note X la matrice constituée des p vecteurs colonnes $X^{(j)}$ et on suppose que le rang de cette matrice est p.
On désire tester s'il existe bien une relation linéaire entre Y et $X^{(j_0)}$, où $j_0 \in \{1, \ldots, p\}$. Dans la suite, on considère X^{-j_0} le sous-espace vectoriel engendré par les vecteurs $X^{(j)}$ pour $j \in \{1, \ldots, p\}$ et $j \neq j_0$.

(a) On note $\big(X^{-j_0}\big)^{\perp}$ le sous-espace vectoriel orthogonal de X^{-j_0}. Déterminer sa dimension d **(0.5pts)**.

(b) Pour $(u_1, \ldots, u_d)$ une base orthonormale de $\big(X^{-j_0}\big)^{\perp}$, on note $U = [u_1, \ldots, u_d]$ la matrice (n, d) constituée par les vecteurs colonnes u_i. Soit $\varepsilon = {}^t(\varepsilon_1, \ldots, \varepsilon_n)$, où $(\varepsilon_k)_{k \in \mathbf{N}}$ est une suite de v.a.i.i.d. de loi $\mathcal{N}(0, \sigma_*^2)$ avec $\sigma_*^2 > 0$ est inconnu. On note $\xi = {}^t U\, \varepsilon$. Déterminer la loi de ξ **(1.5pts)**.

(c) On suppose que $Y = X\,\theta^* + \varepsilon$ avec $\theta^* = {}^t(\theta_1^*, \ldots, \theta_p^*) \in \mathbf{R}^p$. On note $Z = {}^t U\, Y$ et $W^{(j_0)} = {}^t U\, X^{(j_0)}$. Démontrer que $W^{(j_0)} \neq 0$ **(0.5pts)**, puis que $Z = \theta_{j_0}^*\, W^{(j_0)} + \xi$ **(1pt)**.

(d) En déduire l'expression de l'estimateur linéaire $\widehat{\theta}_{j_0}$ sans biais de variance minimale de $\theta_{j_0}^*$ **(1pt)** et démontrer que sa loi est $\mathcal{N}\big(\theta_{j_0}^*\,,\, \sigma_*^2\, \|W^{(j_0)}\|^{-2}\big)$ **(1pt)**.

(e) En déduire un estimateur non biaisé $\widehat{\sigma}_{j_0}^2$ de σ^2 et préciser sa loi **(1pt)**. Que peut-on dire par rapport à l'estimateur sans biais par moindres carrés de σ^2 du modèle initial $Y = X\,\theta^* + \varepsilon$ **(1pt)**?

(f) On souhaite tester $H_0 :\ \theta_{j_0}^* = 0$ contre $H_1 :\ \theta_{j_0}^* \neq 0$. Déduire de ce qui précède une statistique de test dont on précisera la région critique de niveau 5% **(1.5pts)**.

(g) On considère le cas particulier où $p = 2$, $X^{(1)} = {}^t(1, 1, \ldots, 1)$ et $X^{(2)} = {}^t(1, 2, \ldots, n)$ et on choisit $j_0 = 2$. Vérifier que l'on peut choisir $(u_1, \ldots, u_{n-1})$ telle que $u_i = v_i / \|v_i\|$ pour $i = 1, \ldots, n-1$ et

$$v_1 = {}^t\big(1, -1, 0, \ldots, 0\big), \quad v_2 = {}^t\big(1, 1, -2, 0, \ldots, 0\big), \quad \ldots \quad , v_j = {}^t\big(1, 1, \ldots, 1, -j, 0, \ldots, 0\big) \quad \textbf{(1pt)}\ .$$

On rappelle que $\sum_{k=1}^m k^2 = m(m+1)(2m+1)/6$. En déduire la loi de $\widehat{\theta}_2$ **(1pt)** et sa normalité asymptotique quand $n \to \infty$ **(0.5pts)**. Comparez avec ce que l'on aurait obtenu avec l'estimateur par moindres carrés classique **(1.5pts)**.

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

(a) On commence par simuler une base de données avec une variable `Y` dépendant d'autres variables `X1, X2, X3, X4, X5` et `X6`, avec une erreur `epsi`:

```
n=100
X1=rexp(n,1); X2=c(1:n); X3=c(1:n)^0.5*cos(c(1:n)^0.5)
X4=runif(n,-1,1); X5=rnorm(n,0,5); X6=exp(c(1:n)/10)
epsi=0.1*c(1:n)*rnorm(n,0,1)
a0=50; a1=4; a2=0; a3=-2; a4=-5; a5=0; a6=0
Y=a0+a1*X1+a2*X2+a3*X3+a4*X4+a5*X5+a6*X6+epsi
Y[1]=0
```

Désormais on oublie le modèle simulé et on effectue un traitement pour retrouver un modèle permettant de prédire `Y` en fonction de la connaissance de `X1, X2, X3, X4, X5` et `X6`:

```
Ylm1=lm(Y~X1+X2+X3+X4+X5+X6); summary(Ylm1)
```

Voici les résultats:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.5212023  1.8621597  25.519  < 2e-16 ***
X1           4.6366753  0.6220122   7.454 4.60e-11 ***
X2           0.0222050  0.0383785   0.579   0.5643
X3          -1.9758512  0.2012987  -9.816 5.03e-16 ***
X4          -3.0637215  1.3694118  -2.237   0.0277 *
X5          -0.1381499  0.1722117  -0.802   0.4245
X6          -0.0002109  0.0002568  -0.821   0.4136

Residual standard error: 7.704 on 93 degrees of freedom
Multiple R-squared:  0.7203,Adjusted R-squared:  0.7023
F-statistic: 39.92 on 6 and 93 DF,  p-value: < 2.2e-16
```

Questions 1: Ecrire formellement et en détail le vrai modèle. Expliquer précisément et formellement ce que sont les valeurs `0.0222050`, `0.0383785`, `0.579` *et comment on calcule* `0.5643`. *Quel test représente la valeur* `39.92` *et qu'en conclure* **(2.5pts)***?*

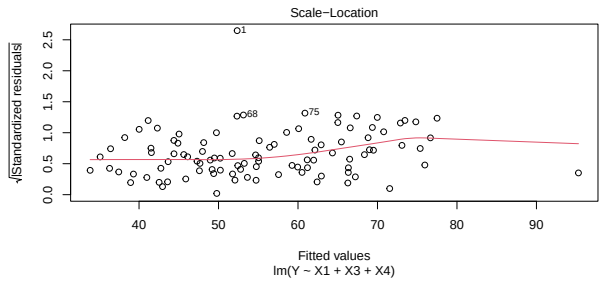
(b) On tape ensuite les commandes:

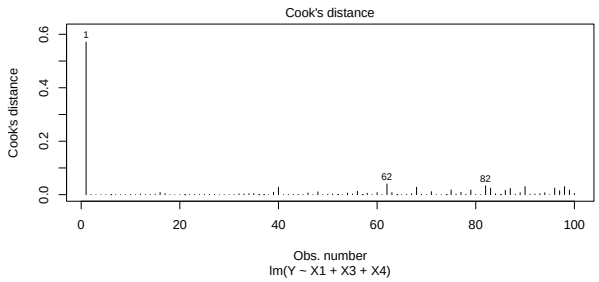
```
library(MASS)
Ylm2=stepAIC(Ylm1,k=log(n),direction="both", trace=FALSE)
summary(Ylm2); plot(Ylm2,3); plot(Ylm2,4)
```

On obtient les résultats numériques et les graphes suivants:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.1574     1.0145  47.469  < 2e-16 ***
X1            4.6445     0.6141   7.563 2.36e-11 ***
X3           -1.9384     0.1445 -13.413  < 2e-16 ***
X4           -3.1184     1.3439  -2.320   0.0224 *

Residual standard error: 7.638 on 96 degrees of freedom
Multiple R-squared:  0.7163,Adjusted R-squared:  0.7074
F-statistic: 80.78 on 3 and 96 DF,  p-value: < 2.2e-16
```





Questions 2: Expliquer ce qui a été fait et pourquoi. Que conclure des résultats numériques? Que penser des deux graphes? Est-on surpris de l'ensemble des résultats obtenus? **(2pts)***?*

(c) On tape ensuite les commandes:

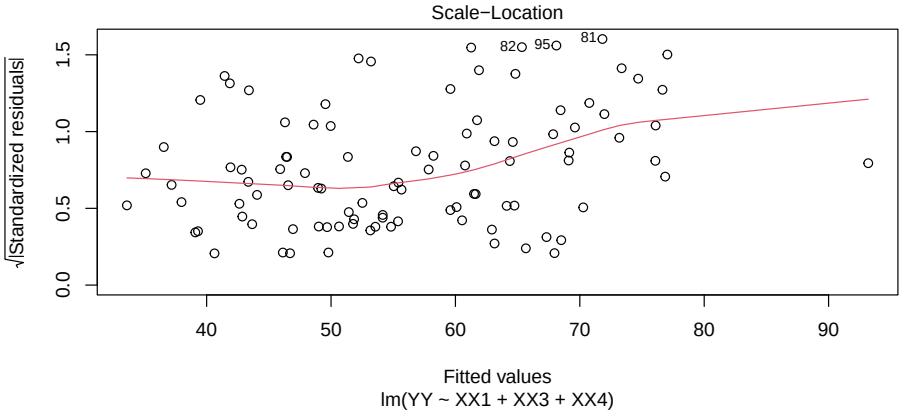
```
YY=Y[-1]; XX1=X1[-1]; XX3=X3[-1]; XX4=X4[-1]
Ylm3=lm(YY~XX1+XX3+XX4); summary(Ylm3)
plot(Ylm3,3);
```

On obtient les résultats numériques et le graphe suivant:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.9528     0.7169  68.286  < 2e-16 ***
XX1           4.4951     0.4315  10.418  < 2e-16 ***
XX3          -1.8869     0.1016 -18.569  < 2e-16 ***
XX4          -4.8116     0.9589  -5.018 2.43e-06 ***

Residual standard error: 5.364 on 95 degrees of freedom
```

```
Multiple R-squared:  0.8354,Adjusted R-squared:  0.8302
F-statistic: 160.7 on 3 and 95 DF,  p-value: < 2.2e-16
```



Questions 3: Expliquer ce qui a été fait et pourquoi. A-t-on gagné avec ces traitements? **(1pt)***?*

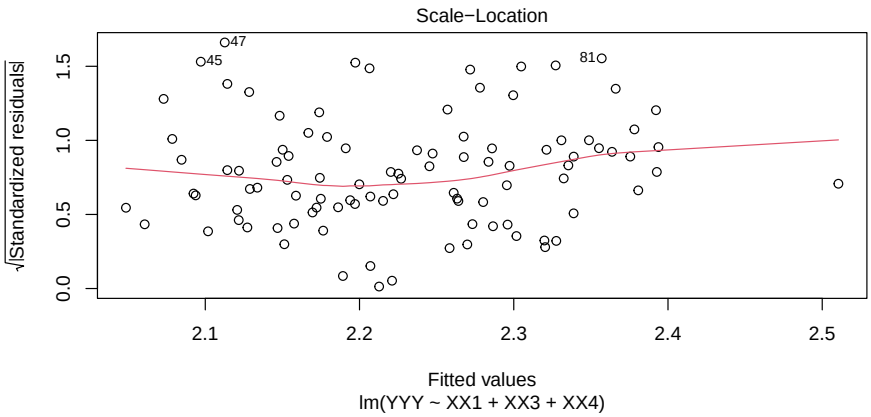
(d) On tape ensuite les commandes:

```
library(MASS)
dataY=data.frame(cbind(YY,XX1,XX3,XX4))
BC=boxcox(YY ~XX1+XX3+XX4,data=dataY,plotit = TRUE,lambda = seq(-3,3, 1/10))
ind=which(BC$y==max(BC$y))
lambda=BC$x[ind]; lambda
YYY=YY^0.5
Ylm4=lm(YYY~XX1+XX3+XX4); plot(Ylm4,3)
1-sum((YY-Ylm4$fit^2)^2)/sum((YY-mean(YY))^2)
```

Avec pour résultats numériques et graphe:

```
>  lambda
[1] 0.5151515

> 1-sum((YY-Ylm4$fit^2)^2)/sum((YY-mean(YY))^2)
[1] 0.836149
```



Questions 4: Qu'est-ce que cette valeur de `lambda`*? Que conclure du graphe? Expliquer ce qu'est la valeur obtenue* `0.836149` *et la manière dont elle est calculée* **(1.5pts)***?*

(e) On tape enfin les commandes:

```
Z=as.numeric(Y>quantile(Y,0.3)); mean(Z)
regZ=glm(Z~X1+X2+X3+X4+X5+X6,family=binomial(link="probit"),na.action=na.pass)
summary(regZ)
predZ=as.numeric(regZ$fitted.values>0.5)
mean((Z-predZ)^2)
```

Voici les résultats:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.858e-01  5.078e-01   1.154  0.24868
X1           5.923e-01  2.539e-01   2.333  0.01965 *
X2          -4.515e-03  1.228e-02  -0.368  0.71303
X3          -3.863e-01  8.466e-02  -4.563 5.03e-06 ***
X4          -1.126e+00  4.086e-01  -2.755  0.00587 **
X5          -3.089e-02  4.941e-02  -0.625  0.53189
X6           1.096e-05  2.284e-04   0.048  0.96172

> mean((Z-predZ)^2)
[1] 0.09
```

Questions 5: Expliquer ce qui a été fait ainsi que le modèle obtenu pour `Z`*. Que représente la valeur* `0.71303`*? Est-on surpris des résultats obtenus? Comment prédire* `Z` *avec ce modèle si tous les* Xi *valent* 1*? Que représente la valeur* `0.09`*?* **(2pts)***?*