Projet de statistique: la transformation de Box-cox

Romain HATON et Leonardo MARTINS BIANCO

November 2020

1 Introduction

La régression est un modèle connu pour leur robustesse et sa stabilité face à des fluctuations des échantillons. Elles consistent à modéliser la réponse par une combinaison linéaire de paramètres et des conditions d'expérience plus un bruit. Cependant, certaines hypothèses sont nécessaires afin de pouvoir appliquer cette modélisation Parmi ces conditions, l'hypothèse d'homoscédasticité (variance constante) et la gaussienité des données.
Dans les cas où l'une des hypothèses précédentes ne sont pas vérifiées, une transformation non linéaire des variables peut être envisagé afin de proposer une modélisation plus adaptée au jeu de données étudié. Par exemple, la transformation de Box-Cox étudie les transformations en puissance de la variable à expliquer.
L'étude suivante consiste à étudier cette transformation sur un aspect théorique, ensuite les résultats théorique sont appliquées sur des données simulés dans un premier temps, puis finalement une application sur des données réelles vient conclure notre étude.

2 La transformation de Box-Cox : étude théorique

Le principe de la méthode est le suivant : y étant la variable à expliquer, on définit une famille de transformations paramétriques (h_λ) de y en $h_\lambda(y)$ et on considère le modèle d'observation

$$(x_i, Y_i) : h_\lambda(Y_i) = Z_i = x_i\theta + \epsilon_i,\ \epsilon_i \sim_{iid} \mathcal{N}(0, \sigma^2) \qquad (1)$$

où x_i est le vecteur ligne des conditions d'expérience et ϵ_i le bruit iid gaussien.
L'objectif est de déterminer λ_{opt} permettant un meilleur ajustement en régression de y que celui obtenu en régression linéaire (pour λ_{lin} tel que $h_{\lambda_{lin}}$ est l'identité).

Box et Cox ont proposé la famille de transformations $(\tilde{h}_\lambda)$ suivante, pour $\lambda \in \mathbb{R}$:

$$\forall y > 0,\ \tilde{h}_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

Tandis que Bickel et Doksum définissent, pour $\lambda > 0$:

$$\forall y > 0, \tilde{h}_\lambda(y) = \frac{\mathrm{sgn}(y)|y|^{\lambda-1}}{\lambda}$$

où $\mathrm{sgn}(y)$ est le signe de y (1 si $y > 0$, -1 si $y < 0$, 0 si $y = 0$).

Pour le cas $\lambda = 0$ on peut voir que la transformation est égale à une transformation logarithmique qui est une transformation particulièrement efficace pour normaliser les distributions.

Le modèle linéaire n'est pas en général compatible avec la transformation de Box-Cox car on assume $y > 0$. On observe alors que si $\lambda > 0$,

$$\lim_{y \to 0} \frac{y^\lambda - 1}{\lambda} = -\frac{1}{\lambda}$$

De plus, pour un λ fixé,

$$\frac{d}{dy} \frac{y^\lambda - 1}{\lambda} = y^{\lambda - 1} > 0$$

D'où on conclut que $-\frac{1}{\lambda}$ est la valeur la plus petite que la transformation peut atteindre ; c'est-à-dire, la transformation est bornée inférieurement. Or cela implique que les valeurs de Y ne peuvent pas être normalement distribuées autour d'une valeur pour Y petit, c'est-à-dire, on vérifie la non-normalité de la transformation, lorsque la régression linéaire satisfait une condition de normalité.

Cependant, de façon pratique on peut utiliser cette transformation si les valeurs de Y ne sont pas trop petites. C'est-à-dire, on assumera la normalité de la transformation quand même.

Maintenant, on souhaite calculer la vraisemblance des paramètres $(\lambda, \theta, \sigma^2)$ à l'observation de $Y = (Y_1, ..., Y_n)$. À partir de (1) on a que $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ indépendants et identiquement distribuées. Donc la fonction densité $f(\varepsilon_i)$ s'écrit :

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\varepsilon_i^2}{2\sigma^2}\right)$$

On peut écrire $Z_i = x_i\theta + \varepsilon_i$ par (1) donc $h_\lambda(Y_i) = Z_i \sim \mathcal{N}(x_i\theta, \sigma^2)$ donc la fonction de densité de Z_i s'écrit :

$$f_1(Z_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-Z_i^T Z_i}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(h_\lambda(Y) - x_i\theta)^2}{2\sigma^2}\right)$$

En plus h_λ est continûment différentiable et monotone. On peut écrire que $Y = h_\lambda^{-1}(Z)$. On en déduit que la fonction de densité f_2 de Y est obtenue grâce à :

$$f_2(Y) = |\tilde{J}(h_\lambda(Y))| f_1(h_\lambda(Y))$$

où $J(h(Y))$ est la jacobienne de la fonction $h_\lambda(Y)$. Dans notre cas, la jacobienne pour "une seule observation" y est donnée par

$$\tilde{J}(h(Y)) = \frac{\partial}{\partial Y}\left(\frac{\text{sgn}(Y)|Y|^\lambda - 1}{\lambda}\right) = |Y|^{\lambda - 1}$$

Ainsi on obtient que la densité de y (pour "une seule observation" y) est (ici x est le vecteur de variables explicatives correspondant à y)

$$f_2(y) = \frac{\tilde{J}(\lambda; y)}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(h_\lambda(y) - x\theta)^2}{2\sigma^2}\right)$$

On obtient la vraisemblance en faisant le produit des densités pour un ensemble d'observations $(Y_1, \ldots, Y_n)$. On note $J = \Pi_i \tilde{J}(\lambda; Y_i)$, et on obtient

$$L\left(\lambda,\theta,\sigma^2;Y\right)=\frac{J(\lambda;y)}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\exp\left(-\frac{\sum_i\left(h_\lambda\left(y_i\right)-x_i\theta_i\right)^2}{2\sigma^2}\right)$$

et on note que l'on peut écrire la somme $\sum_i\left(h_\lambda\left(y_i\right)-x_i\theta_i\right)^2$ comme le produit matriciel $\left(h_\lambda(Y)-X\theta\right)'\left(h_\lambda(Y)-X\theta\right)$, où $h_\lambda(Y)$ est le vecteur de composantes $h_\lambda(Y_i)$ et X est la matrice du plan d'expérience. Observons que x_i denote le i-ème vecteur observé de variables explicatives.

$$\boxed{L(\lambda,\theta,\sigma^2,Y)=\frac{J(\lambda,Y)}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\exp\left(\frac{-(h_\lambda(Y)-X\theta)^T(h_\lambda(Y)-X\theta)}{2\sigma^2}\right)}$$

L'objectif est maintenant de déterminer les estimateurs $\hat{\theta}(\lambda)$ de θ et $\hat{\sigma}^2(\lambda)$ de σ^2. On a obtenu :

$$L(\lambda,\theta,\sigma^2,Y)=\frac{J(\lambda,Y)}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\exp\left(\frac{-1}{2\sigma^2}(h_\lambda(Y)-X\theta)^T(h_\lambda(Y)-X\theta)\right)$$

$$\begin{aligned}\log(L(\lambda,\theta,\sigma^2,Y))&=\log\left(\frac{J(\lambda,Y)}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\right)-\frac{1}{2\sigma^2}(h_\lambda(Y)-X\theta)^T(h_\lambda(Y)-X\theta)\\&=\log\left(\frac{\prod_{i=1}^N|Y_i|^{\lambda-1}}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\right)-\frac{1}{2\sigma^2}(h_\lambda(Y)-X\theta)^T(h_\lambda(Y)-X\theta)\\&=(\lambda-1)\sum_{i=1}^N\log(|Y_i|)-\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}(h_\lambda(Y)-X\theta)^T(h_\lambda(Y)-X\theta)\end{aligned}$$

On cherche maximiser cette quantité par rapport aux paramètres en dérivant et en utilisant les formules de dérivations matricielle :

$$\begin{aligned}\frac{\partial\log(L)}{\partial\theta}&=-\frac{1}{2\sigma^2}(-2X^Th_\lambda(Y)+2X^tX\theta)=0\\&\implies-2X^Th_\lambda(Y)+2X^tX\theta=0\\&\implies\hat{\theta}(\lambda)=(X^TX)^{-1}X^Th_\lambda(Y)\end{aligned}$$

Puis de la même façon, en substituant $\hat{\theta}$ dans $\log(L)$ et en dérivant :

$$\begin{aligned}\frac{\partial\log(L)}{\partial\sigma^2}&=-\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{1}{\left(\sigma^2\right)^2}(h_\lambda(Y)-X\hat{\theta})^\top(h_\lambda(Y)-X\hat{\theta})=0\\&\implies\hat{\sigma}^2(\lambda)=\frac{1}{n}(h_\lambda(Y)-X\hat{\theta})^\top(h_\lambda(Y)-X\hat{\theta})\end{aligned}$$

On note que on a vu que cet estimateur est biaisé à distance finie, mais asymptotiquement sans biais ; une alternative possible, sans biais même à distance finie est

$$\hat{\sigma}^2(\lambda)=\frac{1}{n-p}\left(h_\lambda(Y)-X\hat{\theta}\right)^\top\left(h_\lambda(Y)-X\hat{\theta}\right)$$

Bien sûr, on pourrait obtenir les mêmes expressions en remarquant que l'expression trouvée dessus pour la vraisemblance est à un facteur $J(\lambda;y)$ près la même que la vraisemblance trouvée dans le diapo 18 de l'amphi 5. Mais comme $J(\lambda;y)$ ne dépend pas de θ, σ^2, les estimateurs du maximum de vraisemblance restent essentiellement les mêmes.

On note $L_{max} := \log(L(\lambda, \hat{\theta}, \hat{\sigma}^2))$, c'est-à-dire, on remplace les estimateurs $\hat{\theta}$ et $\hat{\sigma}^2$ trouvés ci-dessus dans la formule de la log-vraisemblance. On trouve :

$$\begin{aligned}
L_{max} &= (\lambda - 1)\sum_{i=1}^{N} \log(|Y_i|) - \frac{n}{2}\left(\log(2\pi) + \log(\hat{\sigma}^2)\right) - \frac{1}{2\hat{\sigma}^2}(h_\lambda(Y) - X\hat{\theta})^\top (h_\lambda(Y) - X\theta) \\
&= (\lambda - 1)\sum_{i=1}^{N} \log(|Y_i|) - \frac{n}{2}\left(\log(2\pi) + \log(\hat{\sigma}^2)\right) \\
&- \frac{1}{2\frac{1}{n}(h_\lambda(Y) - X\theta)^\top (h_\lambda(Y) - X\theta)} (h_\lambda(Y) - X\hat{\theta})^\top (h_\lambda(Y) - X\theta) \\
&= -\frac{n}{2}\log(\hat{\sigma}^2) + (\lambda - 1)\sum_{i=1}^{N} \log(|Y_i|) - \frac{n}{2}\log(2\pi) - \frac{n}{2} \\
&= -\frac{n}{2}\log(\hat{\sigma}^2) + (\lambda - 1)\sum_{i=1}^{N} \log(|Y_i|) + a(n)
\end{aligned}$$

avec $a(n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}$.
Or, l'équation à vérifier par $\hat{\lambda}$ est

$$\frac{\partial L_{max}}{\partial \lambda} = 0$$

Cherchons une forme explicite.

$$\boxed{-\frac{n}{2}\frac{(\hat{\sigma}^2(\lambda))'}{\hat{\sigma}^2} + \sum_{i=1}^{N} \log(|Y_i|) = 0}$$

On peut voir que cette équation vérifié par $\hat{\lambda}$ est difficile à résoudre de façon exacte. Une méthode pour calculer la valeur de $\hat{\lambda}$ est d'utiliser des techniques itératives pour résoudre l'équation précèdent sachant qu'elle s'écrit sous la forme $g(x) = 0$ parmi méthodes comme la Méthode de Newton, la Méthode de la sécante, ou la méthode des parties proportionnelles.

Une autre équation peut être trouvée si l'on note que dans l'expression de L_{max}, si on veut la maximiser, il faut juste minimiser $\log \hat{\sigma}^2$. Mais log croissante implique que on doit juste minimiser $\hat{\sigma}^2$. Or

$$\hat{\sigma}^2 = \frac{1}{n}(h_\lambda(Y) - X\hat{\theta})^\top (h_\lambda(Y) - X\hat{\theta})$$

donc, en utilisant l'expression trouvé dessus pour $\hat{\theta}$:

$$\hat{\lambda} = \text{argmin}_\lambda (h_\lambda(Y)^\top - h_\lambda(Y)^\top X(X^\top X)^{-1})(h_\lambda(Y) - X(X^\top X)^{-1}h_\lambda(Y))$$

Si l'on note la propriété suivante

$$(\mathbb{I} - X(X^\top X)^{-1})^2 = (\mathbb{I} - X(X^\top X)^{-1})$$

et si l'on appelle cette matrice K, on obtient alors

$$\hat{\lambda} = \text{argmin}_\lambda = h_\lambda(Y)^\top K h_\lambda(Y)$$

Ainsi on obtient les condition suivantes :

$$\frac{\partial h_\lambda(Y)^\top K h_\lambda(Y)}{\partial \lambda} = 0$$

$$\frac{\partial^2 h_\lambda(Y)^\top K h_\lambda(Y)}{\partial \lambda^2} > 0$$

La formule de dérivation matricielle suivante est valide :

$$\alpha = \mathbf{x}^\top \mathbf{A}\mathbf{x} \implies \frac{\partial \alpha}{\partial z} = \mathbf{x}^\top \left(\mathbf{A} + \mathbf{A}^\top\right) \frac{\partial \mathbf{x}}{\partial \mathbf{z}}$$

On l'utilise et on note aussi que $K = K^\top$. Donc

$$h_\lambda(Y)^\top K \frac{\partial h_\lambda(Y)}{\partial \lambda} = 0$$

Normalement l'estimateur du maximum de vraisemblance $\hat{\theta}$ est normal à distance finie (voir poly du cours, théorème 23, chapitre 6). Cependant, ici on a un troisième paramètre λ à estimer, c'est-à-dire, et θ dépend de λ (on rappelle que $\hat{\theta} = \left(X^T X\right)^{-1} X^T h_\lambda(Y)$), et λ agit sur le système de manière compliqué. Donc il n'y aucune indication que l'estimateur du maximum de vraisemblance soit toujours normal à distance finie.

Maintenant, sous les hypothèses de normalité asymptotique de l'estimateur du maximum de vraisemblance (faites à l'énoncé), on peut estimer la variance de $\hat{\lambda}$ en utilisant le théorème 13 du chapitre 3 du polycopié du cours, qui affirme que (en notant que $L_{\max}$ est la vraisemblance de λ, i.e., pour chaque λ fixé, la densité de probabilité de l'échantillon)

$$\mathrm{Var}(\hat{\lambda}) = \left(-\mathbb{E}_{\hat{\lambda}}\left[\frac{\partial^2 L_{\max}}{\partial \lambda^2}\right]\right)^{-1}$$

Pour construire un intervalle de confiance pour λ, comme on estime la moyenne $\bar{\lambda}$ et aussi la variance, on sait qu'on doit utiliser la loi de Student (qui est symétrique), dans ce cas de $N - p$ degrés de liberté :

$$\sqrt{N}\frac{\bar{\lambda} - \mu_\lambda}{\widehat{\sigma}} \sim \mathcal{T}(N-1)$$

c'est-à-dire, un intervalle de confiance de niveau α est donné avec les quantiles de la loi de Student, qui est symétrique. L'estimateur du maximum de vraisemblance du paramètre λ représente la troisième composante de l'estimateur du maximum de vraisemblance total des paramètres. Alors, en suivant la démonstration du théorème 24 du polycopié du cours,

$$\frac{\widehat{\lambda} - \lambda}{\widehat{\sigma}\sqrt{\left[(X'X)^{-1}\right]}} \sim \mathcal{T}(N-p)$$

où p est le nombre de variables explicatives. Ainsi, (en suivant le théorème 25 du polycopié, chapitre 6) on construit l'intervalle de confiance de niveau α :

$$\left[\hat{\lambda} - t_{N-p}(1-\alpha/2)\widehat{\sigma}\sqrt{(X'X)^{-1}}; \hat{\lambda} + t_{N-p}(1-\alpha/2)\widehat{\sigma}\sqrt{(X'X)^{-1}}\right]$$

où $t_{N-p}(1-\alpha/2)$ dénote le quantile $(1-\alpha/2)$ de la loi de Student avec $N-p$ degrés de liberté.

Pour réaliser un test de Wald $(H_0) : \lambda = \lambda_0$ contre $(H_1) : \lambda \neq \lambda_0$, on doit d'abord écrire ces hypothèses comme $(H_0) : Ap = 0$ contre $(H_1) : Ap \neq 0$, où $p = (\theta, \sigma^2, \lambda - \lambda_0)^\top$. On note que cette restriction définit un sous-espace ω emboîté. La choix la plus simple pour la matrice A est $A = (0, 0, 1)$; on pourrait choisir une matrice carrée, mais ici ça ne fait pas de différence. Avec cette matrice, on construit la statistique de Wald dans le cas linéaire :

$$W = \frac{1}{\widehat{\sigma}^2}(A\widehat{p})' \left[A\left(X'X\right)^{-1} A'\right]^{-1} (A\widehat{p})\Big)$$

On a vu en cours que la loi de cette statistique est $F(1, N-p)$, une distribution de Fisher. Or, on a une loi pivotale, donc on peut construire la région de rejet de niveau α pour le test :

$$R_\alpha = \left\{ \lambda \mid W(\lambda) > q_{1-\alpha}^{F(1,N-p)} \right\}$$

où p est le nombre de variables explicatives. Le premier argument de F vaut 1 car cela est le rang de la matrice A.

Finalement, on rappelle que le rapport de vraisemblances, on rappelle que sous les hypothèses de régularité, l'asymptotique du rapport de vraisemblance était

$$-2\log(RV) \xrightarrow{\mathcal{L}} \chi^2(p-q)$$

où $p = \dim \Theta$ l'espace entier des paramètres, dans ce cas $p = 3$, et $q = \dim \Theta_0$, la dimension d'un espace restreint de paramètres, linéaire. Dans ce cas, comme on teste $\lambda = 0$ contre $\lambda \neq 0$, $\lambda = 0$ définit un hyperplan de dimension 2 dans l'espace de paramètres, et on peut définir le test par rapport à ce plan. C'est-à-dire, $q = \Theta_0 = 2$. Ainsi, $p - q = 1$.

On remarque aussi que dans notre cas,

$$\log(RV) = \log \lambda_0 - \log \hat{\lambda}$$

c'est-à-dire,

$$-2\log(RV) = 2\{L_{\max}(\hat{\lambda}) - L_{\max}(\lambda_0)\}$$

Ainsi, le test du rapport de vraisemblance de niveau α correspond à l'intervalle de rejet dessus :

$$\left\{ 2(L_{\max}(\hat{\lambda}) - L_{\max}(\lambda_0)) > q_{\chi^2(1)}(1-\alpha) \right\}$$

3 La transformation de Box-Cox : étude sur des données simulées

Dans cette partie, nous allons mettre en oeuvre la méthodologie de la partie précédente. Pour cela, on se donne le modèle suivant :

$$h_{0.3}(Y_i) = Z_i = a + bx_i + \epsilon_i, \;\; \epsilon_i \sim_{iid} \mathcal{N}(0, \sigma^2)$$

avec $a = 5$, $b = 1$, $\sigma^2 = 2$, $i = 1, ..., n$ et $n = 50$. Les x_i sont issues d'une loi gaussienne centrée réduite.

On génère les x_i avec la fonction `rnorm(n)` que l'on conserve tout au long de la section. On rappelle que la graine est fixée grâce à la fonction `set.seed(999)`.
Pour vérifier que la condition de convergence est juste, on doit vérifier si $\frac{X'X}{n}$ tend vers une matrice définie positive quand $n \to \infty$. Comme X est un vecteur de taille n, $X'X$ est juste un scalaire, il faut vérifier si ce scalaire est positif. Dans notre cas, ce scalaire est bien positif donc $X'X$ tend bien vers une matrice définie positive.

On génère les $\epsilon_i \sim_{iid} \mathcal{N}(0, 2)$, on en déduit :

$$Z_i = 5 + x_i + \epsilon_i \;\; \texttt{et} \;\; Y_i = h_{0.3}^{-1}(Z_i) = (Z_i + 1)^{\frac{1}{0.3}}$$

Pour estimer la régression linéaire simple de Z en fonction de x, on trouve : $\hat{Z}_i = \hat{a} + \hat{b}x_i + \epsilon_i$ avec $\hat{a} = 5.098$ et $\hat{b} = 1.044$. Pour la régression linéaire simple de Y en fonction de x, on trouve : $\hat{Y}_i = \hat{a} + \hat{b}x_i + \epsilon_i$ avec $\hat{a} = 26.58$ et $\hat{b} = 8.37$. Nous traçons les figures suivantes pour savoir si nous avons de bonnes régressions linéaires.

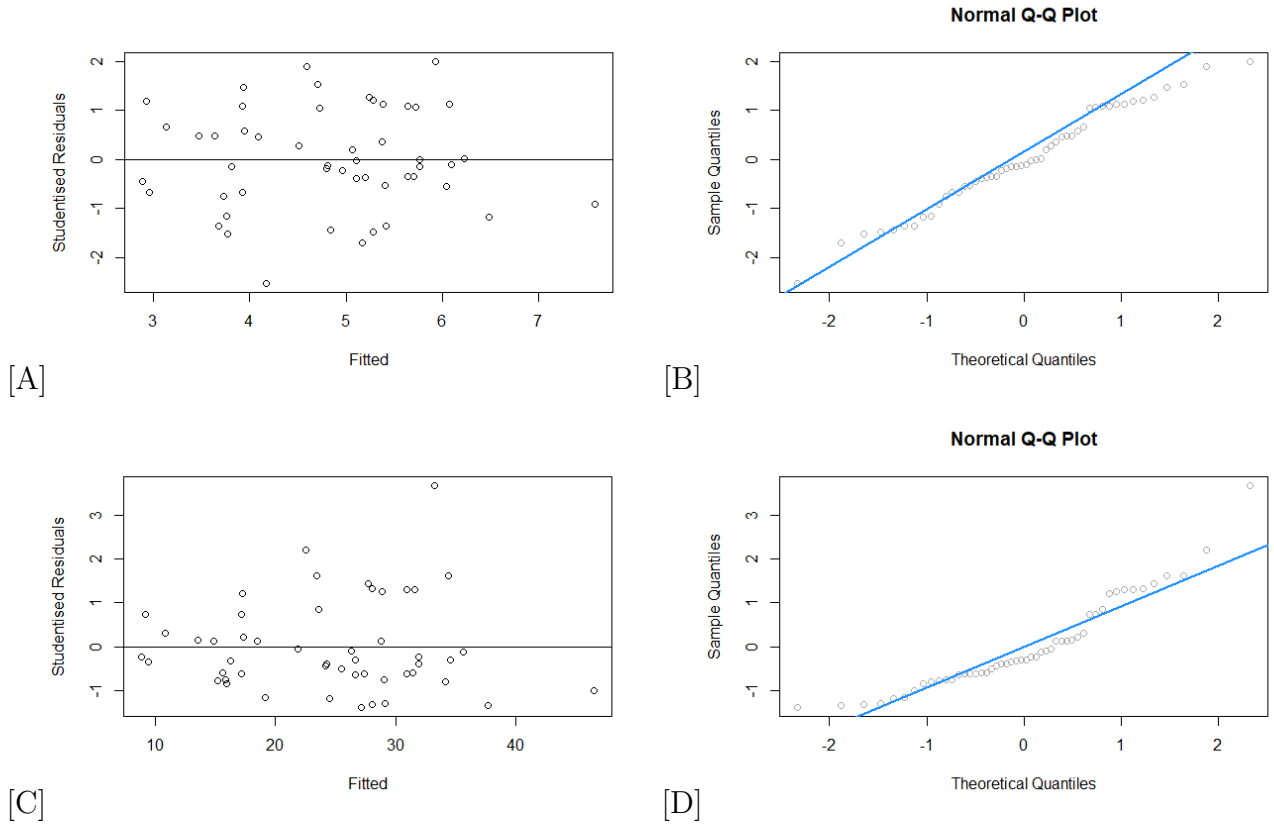


FIGURE 1 – [A] Représentation des résidus studentisés de la régression de Z [B] QQ-plot des résidus de la régression de Z [C] Représentation des résidus studentisés de Y [D] QQ-plot des résidus de la régression de Y

Nous pouvons observer sur la figure [A] que la distribution des résidus est symétrique, ce qui nous dit que les résidus sont centrés. La figure [B] nous donne une droite linéaire qui suit la première bissectrice, ce qui nous permet de supposer que la distribution des résidus se rapproche d'une loi normale. De ces deux figures, nous pouvons dire que nous avons une bonne régression linéaire pour Z. La figure [C] nous montre les résidus ne sont pas répartis de façon symétrique donc nous ne pouvons pas supposer que la distribution des résidus est symétrique. La figure [D] nous permet de dire que la distribution des résidus n'est pas gaussienne parce que la droite représentée ne se rapproche pas de la première bissectrice.

Nous créons la matrice `X` du plan d'expérience avec pour première colonne les valeurs des x_i et à comme deuxième colonne l'intercepte a. Le code suivant :

```
Q = diag(1,n) - X%*%solve(t(X)%*%X)%*%t(X)
Lmle = function(Z){
n = length(Z)
sig2 = ( t(Z)%*%Q%*%Z )/n
-n/2*log(sig2)
}
```

nous permet de calculer $\frac{-n}{2}log(\hat{\sigma}^2)$ qui est utilisé dans la formule de l'estimateur $\hat{\lambda}$.
Nous codons la fonction `lmin(lambda, Y)` qui calcule $L_{max}(\lambda)$ avec la fonction expliquée ci-dessus. Nous utilisons la fonction `lmin: Vlmin = Vectorize(lmin,"lambda")` pour ne pas faire une boucle sur les différentes valeurs de λ. Ces fonctions nous permettent de tracer $L_{max}(\lambda)$ pour $\lambda \in [0; 2]$, nous obtenons la figure suivante :

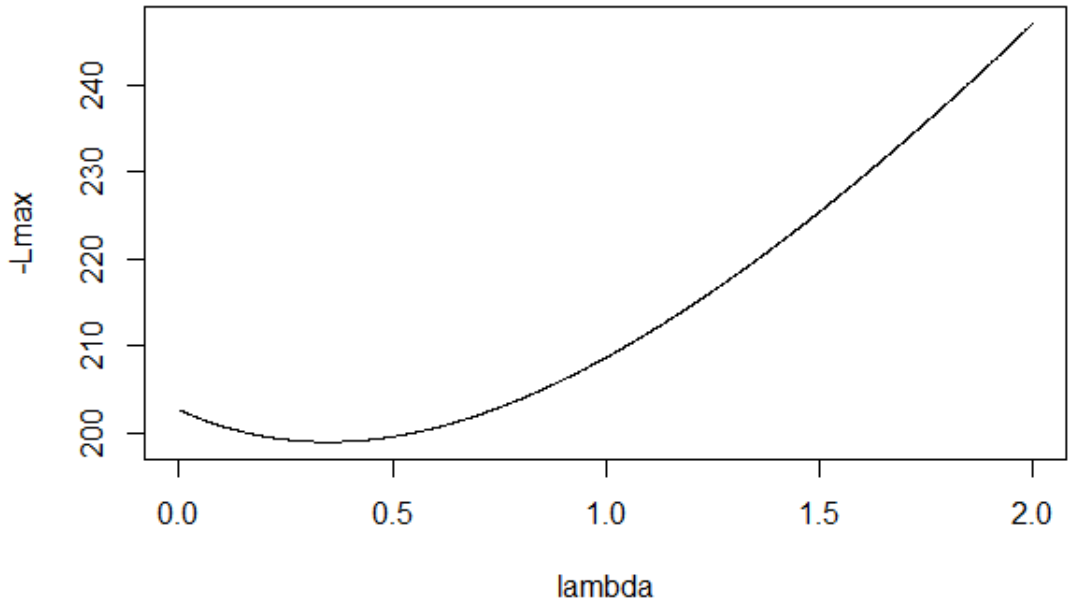


FIGURE 2 – Représentation de $-L_{max}$

Nous pouvons lire graphiquement une valeur de $\hat{\lambda}$ égale à 0.4.

Nous utilisons la commande `nlm(lmin,Y=Y,p=c(0.4),hessian=TRUE)` pour trouver le minimum de la fonction $L_{max}(\lambda)$. Nous mettons $p = c(0.4)$ pour dire que nous partons de 0.4, la valeur que nous venons de lire graphiquement, pour trouver la valeur minimale. La valeur optimale pour $\hat{\lambda}$ est d'environ 0.35 et l'estimation de sa variance est $\text{Var}(\hat{\lambda}) \approx 0.02$.

Nous utilisons la formule suivante pour calculer un intervalle de confiance de niveau 1-α avec $\alpha = 0.05$:

$$\left[\hat{\lambda} - t_{N-p}(1-\alpha/2)\widehat{\sigma}\sqrt{(X'X)^{-1}}; \hat{\lambda} + t_{N-p}(1-\alpha/2)\widehat{\sigma}\sqrt{(X'X)^{-1}}\right]$$

avec $N = 50$, $p = 1$, $\widehat{\sigma}^2$ l'inverse de la matrice hessienne et X le vecteur des x_i.
On obtient donc l'intervalle de confiance de λ suivant :

$$[0,3109005 \ ; \ 0,3864511]$$

La valeur de λ appartient bien a cet intervalle.

3.1 Test de Wald

Nous souhaitons tester plusieurs alternative pour savoir si une transformation est nécessaire et quelle transformation est la plus optimale. Pour tester ces différentes alternatives nous procédons à des tests de Wald bilatéraux. Le test de Wald se formule de la façon suivante :
Nous disposons des ensembles d'observations suivants : $X = (X_1, ..., X_n)$, $Y = (Y_1, ..., Y_n)$ et $Z = (Z_1, ..., Z_n)$, nous disposons également d'une fonction transformation T telle que $Z_i = a + bX_i + \epsilon_i$ avec $\epsilon_i \sim_{iid} \mathcal{N}(0, \sigma^2)$ et $a, b \in \mathbb{R}$. Nous souhaitons tester si une transformation T de Y permet de

retrouver le modèle tel que $T(Y_i) = a + bX_i + \epsilon_i$. Pour cela, on pose : $T(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i'$ avec $\epsilon_i' \sim_{iid} \mathcal{N}(0, \sigma^2)$. Nous estimons les coefficients β_0 et β_1 par $\hat{\beta}_0$ et $\hat{\beta}_1$ grâce à la méthode des moindres carrés. Afin de tester si la transformation T est adaptée, nous testons l'égalité des vecteurs $\beta = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}$ et $\beta^{(0)} = \begin{pmatrix}a\\ b\end{pmatrix} = \begin{pmatrix}5\\ 1\end{pmatrix}$, où β est estimé par $\hat{\beta} = \begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}$.

L'hypothèse nulle pour ce test est $H_0 : \beta = \beta^{(0)}$ contre l'hypothèse alternative $H_1 : \beta \neq \beta^{(0)}$. La statistique de test W sous H_0 s'écrit : $W = (\hat{\beta} - \beta^{(0)})^T V^{-1} (\hat{\beta} - \beta^{(0)})$ avec V la matrice de variance-covariance. Dans ces conditions, $W \sim \chi^2(2)$.
La région de rejet pour ce test de niveau asymptotique α est $\mathcal{R}_\alpha = \{W > q_{1-\alpha}^{\chi^2(2)}\}$ où $q_{1-\alpha}^{\chi^2(2)}$ est le quantile de la loi $\chi^2(2)$ au niveau 1-α.

Pour le premier test, nous souhaitons tester si les données Y ne nécessitent pas de transformation. Dans ce cas, la transformation T est égale à la fonction identité et $\hat{\beta}_0 = 26.58$ et $\hat{\beta}_1 = 8.37$. Nous obtenons $W_{obs} \approx 85.42$. Ensuite, nous calculons la p-value avec la formule suivante :

$$p_{value} = \mathbb{P}(W \in \mathcal{R}_\alpha) = 1 - \mathbb{P}(W_{obs} < q_{1-\alpha}^{\chi^2(2)})$$

Nous trouvons une p-value qui est nulle pour le niveau $\alpha = 0.05$ avec une erreur d'arrondi très faible. Donc nous rejetons l'hypothèse nulle avec une erreur de première espèce de $\alpha = 0.05$.

Pour le second test, nous souhaitons tester si la transformation à appliquer aux observations Y est en racine carrée. Dans ce cas, la transformation T est égale à la fonction racine et $\hat{\beta}_0 = 4.84$ et $\hat{\beta}_1 = 0.92$. Nous obtenons $W_{obs} \approx 0.51$. Ensuite, nous calculons la p-value avec la formule suivante :

$$p_{value} = \mathbb{P}(W \in \mathcal{R}_\alpha) = 1 - \mathbb{P}(W_{obs} < q_{1-\alpha}^{\chi^2(2)})$$

Nous trouvons une p-value $p_{value} = 0.77$ pour un niveau $\alpha = 0.05$. Donc nous acceptons l'hypothèse nulle avec une erreur de seconde espèce inconnue.

Pour le troisième test, nous souhaitons tester si une transformation $T(Y) = h_{0.3}(Y)$ c'est-à-dire $T(y) = \frac{\text{sgn(y)}|y|^{0.3}-1}{0.3}$ à appliquer aux observations Y. Dans ce cas, $\hat{\beta}_0 = 5.10$ et $\hat{\beta}_1 = 1.04$. Nous obtenons $W_{obs} \approx 0.15$. Ensuite, nous calculons la p-value avec la formule suivante :

$$p_{value} = \mathbb{P}(W \in \mathcal{R}_\alpha) = 1 - \mathbb{P}(W_{obs} < q_{1-\alpha}^{\chi^2(2)})$$

Nous trouvons une p-value $p_{value} = 0.93$ pour un niveau $\alpha = 0.05$. Donc nous acceptons l'hypothèse nulle avec une erreur de seconde espèce inconnue.

Pour le dernier test, nous souhaitons tester si une transformation $T(Y) = log(Y)$ à appliquer aux données Y. Dans ce cas, la transformation $T(y) = log(y)$ et $\hat{\beta}_0 = 2.998$ et $\hat{\beta}_1 = 0.47$. Nous obtenons $W_{obs} \approx 290.59$. Ensuite, nous calculons la p-value avec la formule suivante :

$$p_{value} = \mathbb{P}(W \in \mathcal{R}_\alpha) = 1 - \mathbb{P}(W_{obs} < q_{1-\alpha}^{\chi^2(2)})$$

Nous trouvons une p-value qui est nulle pour le niveau $\alpha = 0.05$ avec une erreur d'arrondi très faible. Donc nous rejetons l'hypothèse nulle avec une erreur de première espèce de $\alpha = 0.05$.

Nous pouvons en conclure que de ces quatre hypothèses, la transformation optimale est obtenue pour $\lambda = 0.3$.

3.2 Test de rapport de vraisemblance

Nous souhaitons tester plusieurs alternative pour savoir si une transformation est nécessaire et quelle transformation est la plus optimale. Pour tester ces différentes alternatives nous procédons à des tests de rapport de vraisemblance bilatéraux. Le test de rapport de vraisemblance se formule de la façon suivante :

Nous disposons des ensembles d'observations suivants : $X = (X_1, ..., X_n)$, $Y = (Y_1, ..., Y_n)$ et $Z = (Z_1, ..., Z_n)$, nous disposons également d'une fonction transformation T telle que $Z_i = a + bX_i + \epsilon_i$ avec $\epsilon_i \sim_{iid} \mathcal{N}(0, \sigma^2)$ et $a, b \in \mathbb{R}$. Nous souhaitons tester si une transformation T de Y permet de retrouver le modèle tel que $T(Y_i) = a + bX_i + \epsilon_i$. Pour cela, on pose : $T(Y_i) = \beta_0 + \beta_1 X_i + \epsilon'_i$ avec $\epsilon'_i \sim_{iid} \mathcal{N}(0, \sigma^2)$. Nous estimons les coefficients β_0 et β_1 par $\hat{\beta}_0$ et $\hat{\beta}_1$ grâce à la méthode des moindres carrés.

Afin de tester si la transformation T est adaptée, nous testons l'égalité des vecteurs $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ et $\beta^{(0)} = \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \end{pmatrix}$, où β est estimé par $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$.
L'hypothèse nulle pour ce test est $H_0 : \beta = \beta^{(0)}$ contre l'hypothèse alternative $H_1 : \beta \neq \beta^{(0)}$.
Ces hypothèses sont équivalentes aux hypothèse suivantes :
L'hypothèse nulle est $H_0 : \theta = \beta - \beta^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ contre contre l'hypothèse alternative $H_1 : \theta \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
On note $\theta^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

La statistique de test TRV sous H_0 s'écrit : $TRV = 2(L(\hat{\theta}) - L(\theta^{(0)}))$ où L est la log-vraisemblance.
Dans ces conditions, $TRV \sim \chi^2(2)$.
La région de rejet pour ce test de niveau asymptotique α est $\mathcal{R}_\alpha = \{TRV > q_{1-\alpha}^{\chi^2(2)}\}$ où $q_{1-\alpha}^{\chi^2(2)}$ est le quantile de la loi $\chi^2(2)$ au niveau 1-α.

Pour le premier test, nous souhaitons tester si les données Y ne nécessitent pas de transformation. Dans ce cas, la transformation T est égale à la fonction identité et $\hat{\theta} = \begin{pmatrix} 21.58 \\ 7.37 \end{pmatrix}$. Nous obtenons $TRV_{obs} \approx 16.53$. Ensuite, nous calculons la p-value avec la formule suivante :

$$p_{value} = \mathbb{P}(TRV \in \mathcal{R}_\alpha) = 1 - \mathbb{P}(TRV_{obs} < q_{1-\alpha}^{\chi^2(2)})$$

Nous trouvons une p-value $p_{value} = 0.0003$ pour le niveau $\alpha = 0.05$. Donc nous rejetons l'hypothèse nulle avec une erreur de première espèce de $\alpha = 0.05$.

Pour le second test, nous souhaitons tester si la transformation à appliquer aux observations Y est en racine carrée. Dans ce cas, la transformation T est égale à la fonction racine et $\hat{\theta} = \begin{pmatrix} -0.16 \\ -0.08 \end{pmatrix}$.
Nous obtenons $TRV_{obs} \approx -2.16$. Ensuite, nous calculons la p-value avec la formule suivante :

$$p_{value} = \mathbb{P}(TRV \in \mathcal{R}_\alpha) = 1 - \mathbb{P}(TRV_{obs} < q_{1-\alpha}^{\chi^2(2)})$$

Nous trouvons une p-value $p_{value} = 1$ pour un niveau $\alpha = 0.05$ avec une erreur d'arrondi très faible. Donc nous acceptons l'hypothèse nulle avec une erreur de seconde espèce inconnue.

Pour le troisième test, nous souhaitons tester si une transformation $T(Y) = h_{0.3}(Y)$ c'est-à-dire $T(y) = \frac{\text{sgn(y)}|y|^{0.3}-1}{0.3}$ à appliquer aux observations Y. Dans ce cas, $\hat{\theta} = \begin{pmatrix} 0.10 \\ 0.04 \end{pmatrix}$. Nous obtenons $TRV_{obs} \approx -1.82$. Ensuite, nous calculons la p-value avec la formule suivante :

$$p_{value} = \mathbb{P}(TRV \in \mathcal{R}_\alpha) = 1 - \mathbb{P}(TRV_{obs} < q_{1-\alpha}^{\chi^2(2)})$$

Nous trouvons une p-value $p_{value} = 1$ pour un niveau $\alpha = 0.05$ avec une erreur d'arrondi très faible. Donc nous acceptons l'hypothèse nulle avec une erreur de seconde espèce inconnue.

Pour le dernier test, nous souhaitons tester si une transformation $T(Y) = log(Y)$ à appliquer aux données Y. Dans ce cas, la transformation $T(y) = log(y)$ et $\hat{\theta} = \begin{pmatrix} -2.01 \\ -0.53 \end{pmatrix}$. Nous obtenons $TRV_{obs} \approx 13.11$. Ensuite, nous calculons la p-value avec la formule suivante :

$$p_{value} = \mathbb{P}(TRV \in \mathcal{R}_\alpha) = 1 - \mathbb{P}(TRV_{obs} < q_{1-\alpha}^{\chi^2(2)})$$

Nous trouvons une p-value $p_{value} = 0.0003$ pour le niveau $\alpha = 0.05$. Donc nous rejetons l'hypothèse nulle avec une erreur de première espèce de $\alpha = 0.05$.
Nous retrouvons bien les mêmes conclusions pour les tests que la partie précédente.

Afin de vérifier les résultats des parties précédentes, nous utilisons la fonction `powerTransform` qui nous donne la valeur optimale de λ. Pour ce modèle ci, nous trouvons $\lambda = 0.348676$. Cette valeur nous montre bien qu'une transformation est nécessaire pour Y. Cette transformation est optimale pour une valeur de λ entre 0.3 et 0.5 ce qui confirme nos résultats des parties précédentes.

4 Cas de données réelles

Dans cette partie, on applique les connaissances théoriques développés dessus à un cas pratique. On utilise le jeu de données `NbCycleRupture.csv` qui consiste de 27 mesures du nombre de cycles à rupture (y) d'un fil peigné en fonction de trois variables `x1, x2, x3`. Notre objectif est d'analyser le modèle linéaire pour ce jeu de données, et appliquer la transformation de Box-Cox pour essayer de trouver un régression plus adéquate. On compare aussi le modèles obtenus pour essayer à choisir "le meilleur" dans un sens.

Tout d'abord on charge les données dans `R` en utilisant les commandes

```
NbCycleRupture <- read.csv("/votre_directoire/NbCycleRupture.csv", sep=";")
# Vérification de la taille.
dim(NbCycleRupture)
```

la troisième ligne vérifie que le jeu de données a 27 mesures.

4.1 Régréssion linéaire multiple

D'abord on ajuste un modèle linéaire multiple à ces données :

$$Y = \theta X + \varepsilon$$

```
# Pour simplifier notation:
x1 <- NbCycleRupture$x1
x2 <- NbCycleRupture$x2
x3 <- NbCycleRupture$x3
y  <- NbCycleRupture$y

model_1 <- lm(y ~ x1+x2+x3)
summary(model_1)
```

Cela est réalisé via le code que suit `plot(model_1)`

On obtient donc les graphes suivants :

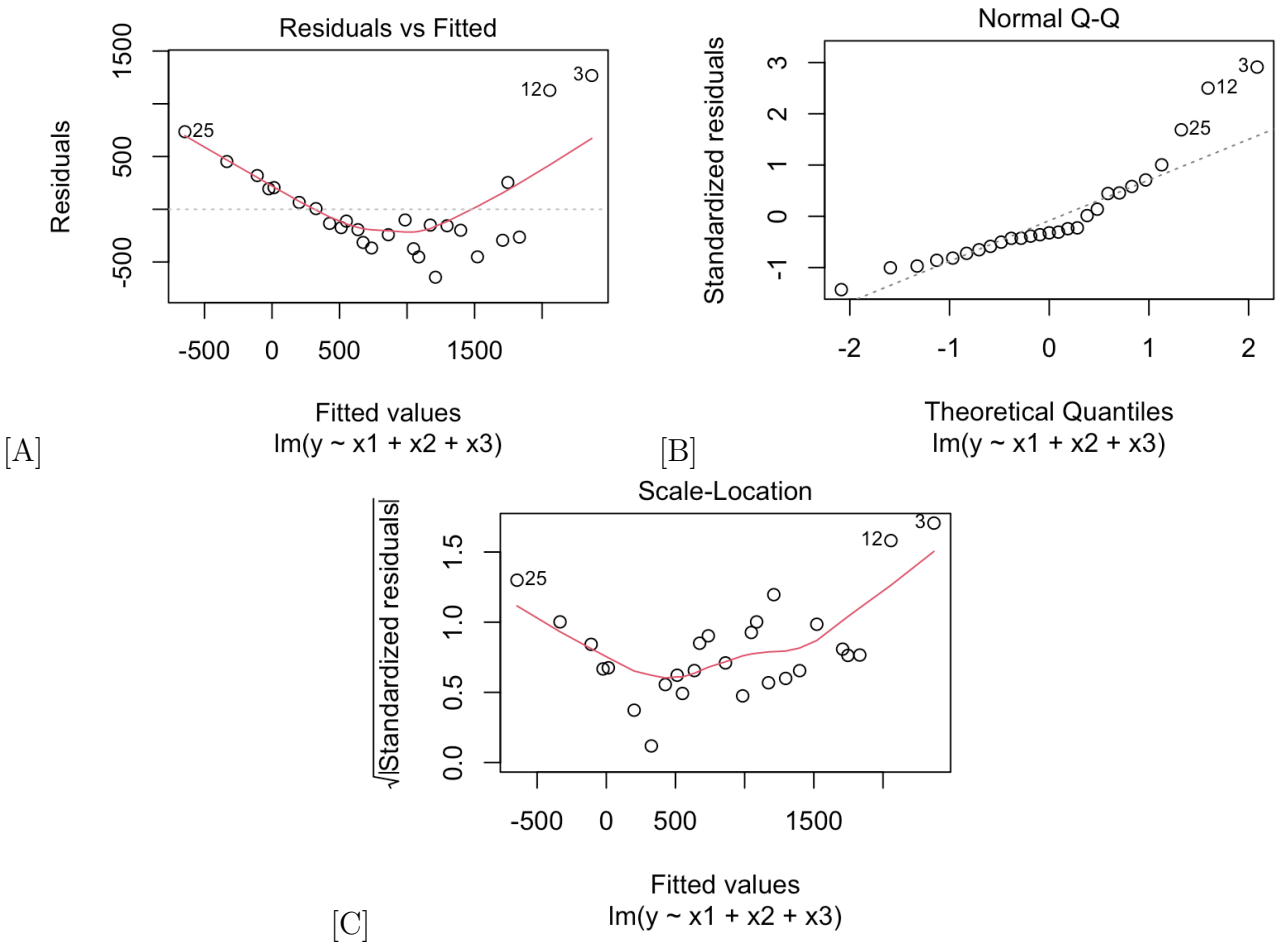


FIGURE 3 – [A] Résidus v. valeurs ajustées ; montre que l'hypothèse de linéarité n'est pas satisfaite. [B] Graphe QQ ; montre que l'hypothèse de normalité des résidus n'est pas vérifié. [C] Scale-Location ; montre que l'hypothèse d'homoscédasticité n'est pas vérifié dans le modèle.

On voit facilement à partir des graphes que les hypothèses du modèle linéaire ne sont pas vérifiés. On obtient aussi le "sommaire" du modèle comme suit :

```
    Residuals:
   Min     1Q Median     3Q    Max
-644.5 -279.1 -150.2  199.5 1268.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   861.37      93.94   9.169 3.83e-09 ***
x1            660.00     115.06   5.736 7.66e-06 ***
x2           -535.83     115.06  -4.657 0.000109 ***
x3           -310.83     115.06  -2.702 0.012734 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 488.1 on 23 degrees of freedom
Multiple R-squared:  0.7291,        Adjusted R-squared:  0.6937
F-statistic: 20.63 on 3 and 23 DF,  p-value: 1.028e-06
```

Quelques remarques sur les tests réalisés dans le sommaire :

— R^2 : On voit que les valeurs de R^2 0.7. Le critère dit informellement que c'est une bonne chose si la valeur de R est proche de 1, ce qui signifie que la somme des carrés résiduels des données observés est proche de la somme des carrés résiduels des prédictions du modèle. La valeur 0.7 pourrait être meilleure.

— Signification globale du modèle : cela est vérifié à travers du test de Fisher, qui teste H_0 : "toutes variables explicatives sont zéro" contre H_1 : "quelque variable explicative n'est pas zéro". On observe que le test a une p-valeur $p = 10^{-6}$, c'est-à-dire, on est bien sûr (un peut prend un niveau $\alpha \ll 0.01$) que le modèle est globalement significatif.

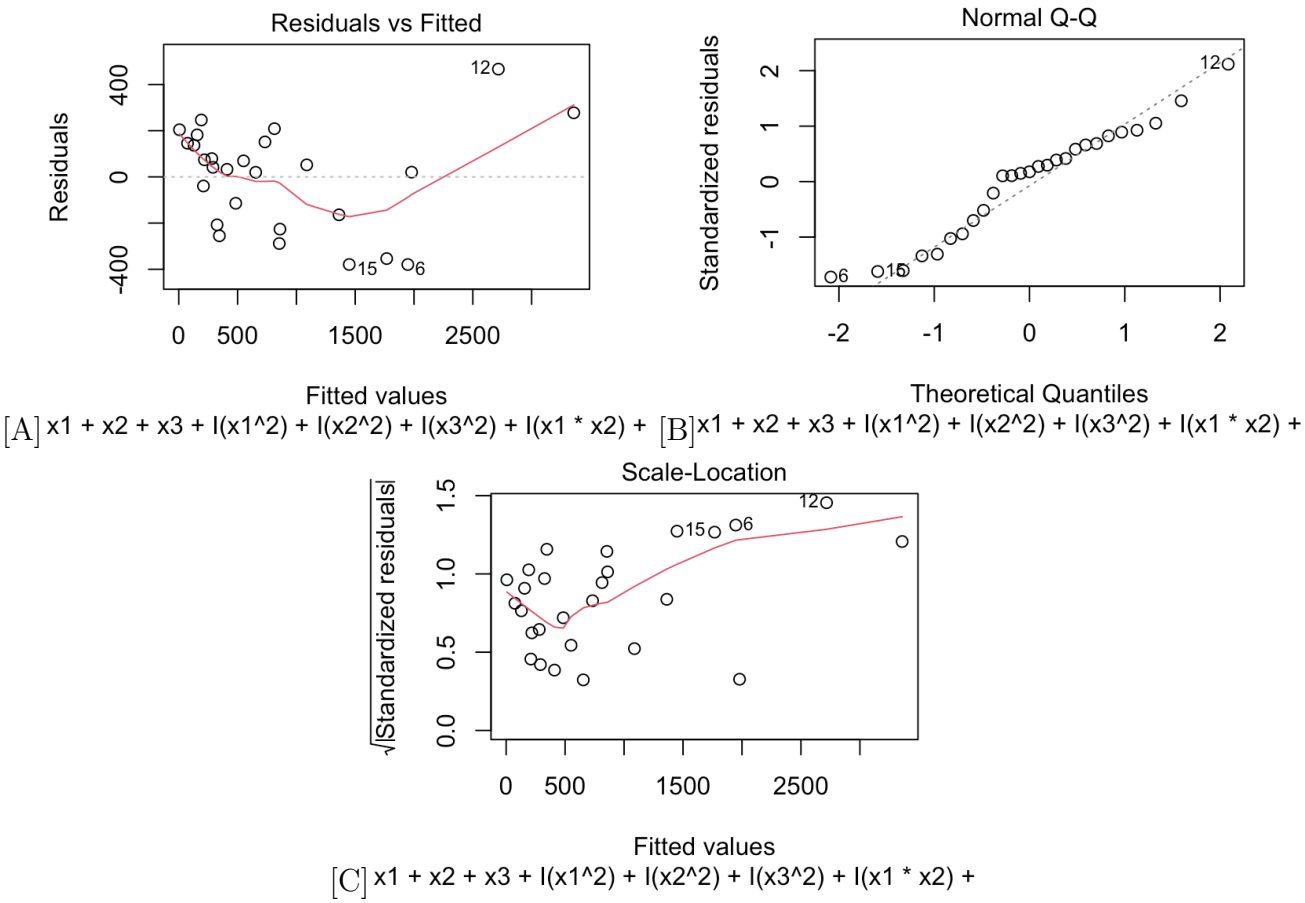
— Signification des variables : cela est vérifié à travers d'un t-test. On voit que avec un niveau $\alpha = 0.01$ le test nous dit ($p < \alpha$) que toutes les variables du modèle sont significatives.

4.2 Introduction de variables d'ordre plus haute

Maintenant, on introduit les variables $x_i x_j, 1 \leq i, j \leq 3$ d'ordre 2 dans le modèle antérieur. En termes de code, la procédure est très similaire :

```
# Maintenant analysons le modèle avec des variables d'ordre plus haute:
model_2 <- lm(y ~ x1 + x2 + x3 + I(x1^2) + I(x2^2)
+ I(x3^2) + I(x1*x2) + I(x1*x3) + I(x2*x3))
summary(model_2)
plot(model_2)
```

On obtient alors les graphes suivantes, qui permettent de vérifier les hypothèses du modèle.



[A] [B] [C]

FIGURE 4 – [A] Résidus v. valeurs ajustées ; montre que l'hypothèse de linéarité n'est pas satisfaite. [B] Graphe QQ ; montre que l'hypothèse de normalité des résidus n'est pas vérifié. [C] Scale-Location ; montre que l'hypothèse d'homoscédasticité n'est pas vérifié dans le modèle.

On voit alors que les hypothèses sont toujours pas vérifiés pour le modèle avec variables de second ordre.

On obtient aussi le sommaire suivant qui permet de vérifier la significance du modèle et ses variables :

```
Residuals:
    Min      1Q  Median      3Q     Max
-379.48 -185.95   41.41  148.48  466.69

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   550.70     138.44   3.978 0.000973 ***
x1            660.00      64.09  10.299 1.00e-08 ***
x2           -535.83      64.09  -8.361 1.99e-07 ***
x3           -310.83      64.09  -4.850 0.000150 ***
I(x1^2)       238.56     111.00   2.149 0.046317 *
I(x2^2)       275.72     111.00   2.484 0.023712 *
I(x3^2)       -48.28     111.00  -0.435 0.669081
I(x1 * x2)   -456.50      78.49  -5.816 2.06e-05 ***
I(x1 * x3)   -235.67      78.49  -3.003 0.008011 **
I(x2 * x3)    142.92      78.49   1.821 0.086278 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 271.9 on 17 degrees of freedom
Multiple R-squared:  0.9379,        Adjusted R-squared:  0.905
F-statistic: 28.51 on 9 and 17 DF,  p-value: 1.564e-08
```

Quelques remarques sur les tests réalisés dans le sommaire, comme avant :

— R^2 : On voit que les valeurs de R^2 0.9. Cette valeur est raisonnablement proche de 1, c'est une bonne chose.

— Signification globale du modèle : On observe que le test a une p-valeur $p = 10^{-8}$, c'est-à-dire, on est bien sûr (un peut prend un niveau $\alpha \ll 0.01$) que le modèle est globalement significatif.

— Signification des variables : On note que avec un niveau $\alpha = 0.1$ le test nous dit que toutes la seule variable qui ne passe pas le test c'est x_3^2 ; c'est-à-dire, cette variable ne doit pas rentrer dans le modèle, car la p-value est $0.66 \gg 0.1$. Le risque de seconde espèce de cette décision n'est pas connu puisqu'on ne connaît pas de façon exacte l'alternative.

4.3 Choix de modèle

On se demande quel modèle choisir : l'un avec les variables mixtes ou l'autre, plus simple ? Or, comme le modèle à variables mixtes contient le modèle simple (sans les variables mixtes), on a vu en cours que l'on peut réaliser l'analyse de variance (ANOVA) pour décider cela. Sur R, construire le tableau d'analyse de variance c'est tout simplement le commande `anova(modele_1, modele_2)`, qui nous donne le résultat suivant :

```
    Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3
Model 2: y ~ x1 + x2 + x3 + I(x1^2) + I(x2^2) + I(x3^2) + I(x1 * x2) +
    I(x1 * x3) + I(x2 * x3)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     23 5480593
2     17 1256745  6   4223848 9.5227 0.0001154 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Or, le test réalisé a comme hypothèse H_0 : "choix du modèle 1" contre H_1 : "réjection du modèle 1 et acceptation du modèle 2". On reçoit la p-valeur $p = 0.0001154$, c'est à dire, avec un niveau $\alpha \ll 0.01$, on préfère le modèle 2 (avec des variables mixtes) au modèle 1.

4.4 La transformation de Box-Cox

On cherche maintenant une transformation de type Box-Cox, que l'on a étudié dans les questions précédentes, pour stabiliser la variance du modèle 1.

Cherchons d'abord la bonne valeur de λ. En code R, ça peut être fait comme suit :

```
# On applique Box-Cox au modèle:
transf_boxcox = boxcox(model_1, lambda = seq(-3,3))
```

et on obtient alors le graphe suivant :

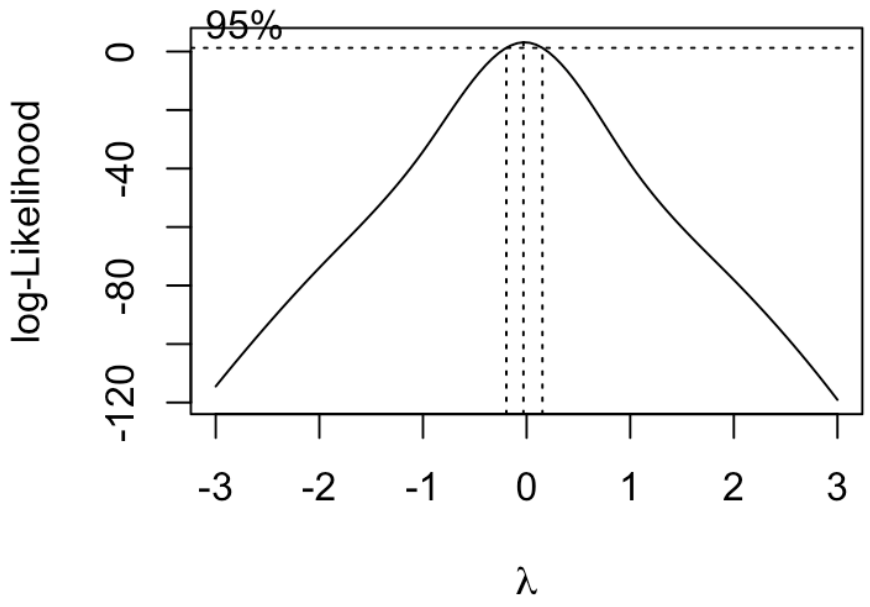


FIGURE 5 – On voit nettement que la bonne valeur de λ est zéro

On voit facilement que la bonne valeur de λ est zéro, et donc la transformation de Box-Cox est juste une transformation logarithmtique. Le nouveau modèle est simplement implémenté par

```
modele_3 <- lm(log(y) ~ x1 + x2 + x3)
summary(modele_3)
plot(modele_3)
```

On obtient pour ce nouveau modèle les graphes suivantes :

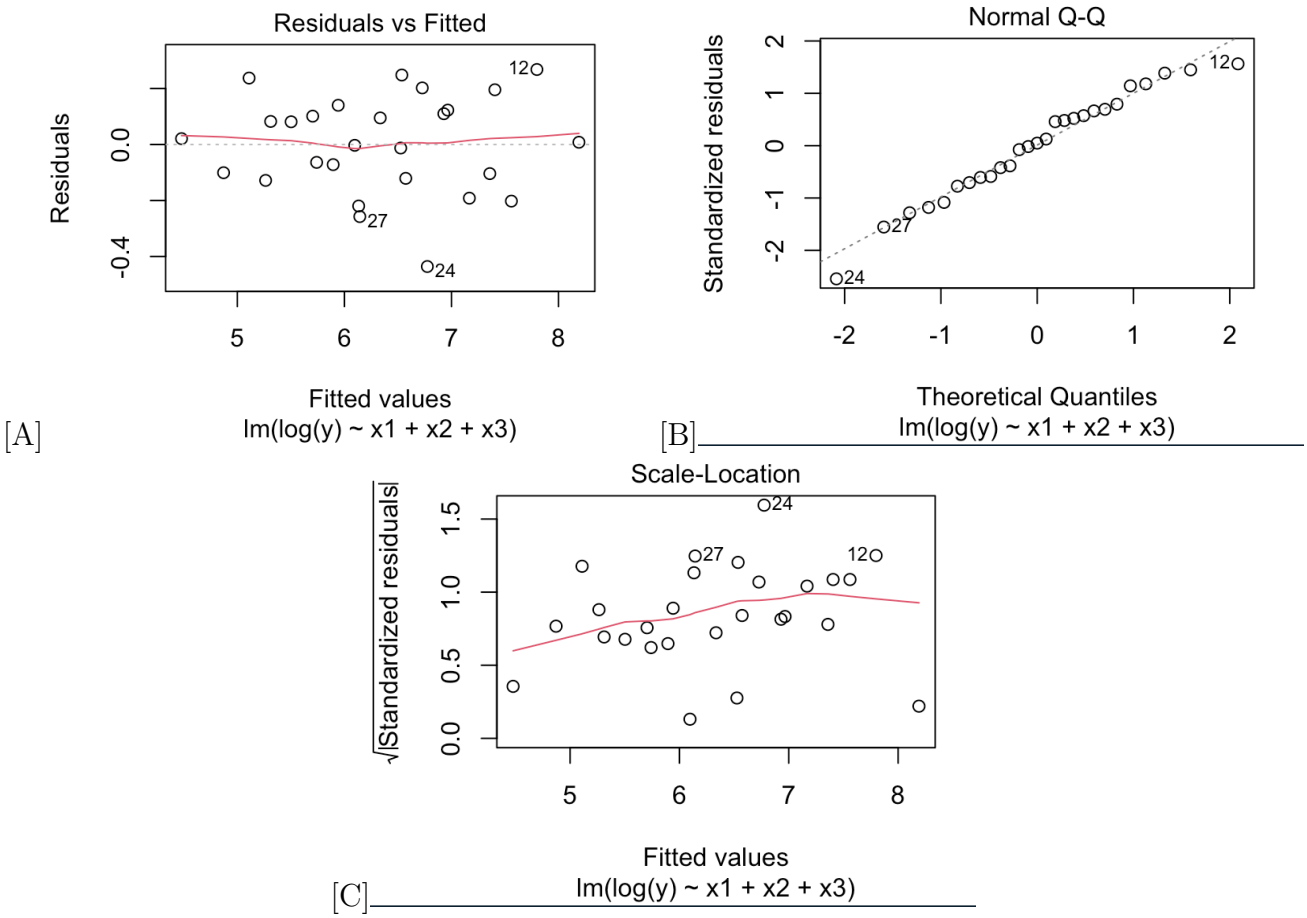


FIGURE 6 – [A] Résidus v. valeurs ajustées ; montre que après la transformation de Box-Cox, l'hypothèse de linéarité est maintenant satisfaite. [B] Graphe QQ ; montre que après la transformation, l'hypothèse de normalité des résidus est vérifié. [C] Scale-Location ; montre que l'hypothèse d'homoscédasticité est vérifié dans le nouveau modèle, c'est-à-dire, la variance a été stabilisé.

Les graphes montrent que le nouveau modèle est beaucoup plus satisfaisant. Passons à une analyse plus rigoureuse de ce fait avec le sommaire de ce modèle :

```
    Residuals:
     Min       1Q   Median       3Q      Max
-0.43592 -0.11250  0.00802  0.11635  0.26790

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.33475    0.03572 177.327  < 2e-16 ***
x1           0.83238    0.04375  19.025 1.43e-15 ***
x2          -0.63087    0.04375 -14.419 5.22e-13 ***
x3          -0.39262    0.04375  -8.974 5.66e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1856 on 23 degrees of freedom
Multiple R-squared:  0.9658,        Adjusted R-squared:  0.9614
F-statistic: 216.8 on 3 and 23 DF,  p-value: < 2.2e-16
```

Analysons ce sommaire.

— R^2 : On voit que les valeurs de R^2 0.96. C'est beaucoup meilleur qu'avant ! (C'était R 0.7)

— Signification globale du modèle : le test de Fisher rend la p-valeur $p = 10^{-16}$, c'est-à-dire, on est très sûr (niveau $\alpha \ll 0.01$) que quelque variable n'est pas zéro, i.e., que globalement le modèle est significatif.

— Signification des variables : avec une précision extrêmement haute, toutes les variables sont significatives ; la plus grande p-valeur vaut environ 10^{-9} !

On peut dire donc que la transformation de Box-Cox a fait un jeu de données initialement pas adapté au modèle linéaire, très bien adapté au modèle linéaire.

Voyons qu'est-ce qu'arrive si on applique cette même transformation de Box-Cox au modèle avec variables d'ordre deux :

```
# On applique la même Box-Cox que pour le modèle 1 dans le modèle 2:
modele_4 <- lm(log(y) ~ x1 + x2 + x3 + I(x1^2) + I(x2^2) + I(x3^2)
+ I(x1*x2) + I(x1*x3) + I(x2*x3))
summary(modele_4)
plot(modele_4)
```

Comme avant, on obtient les graphes :

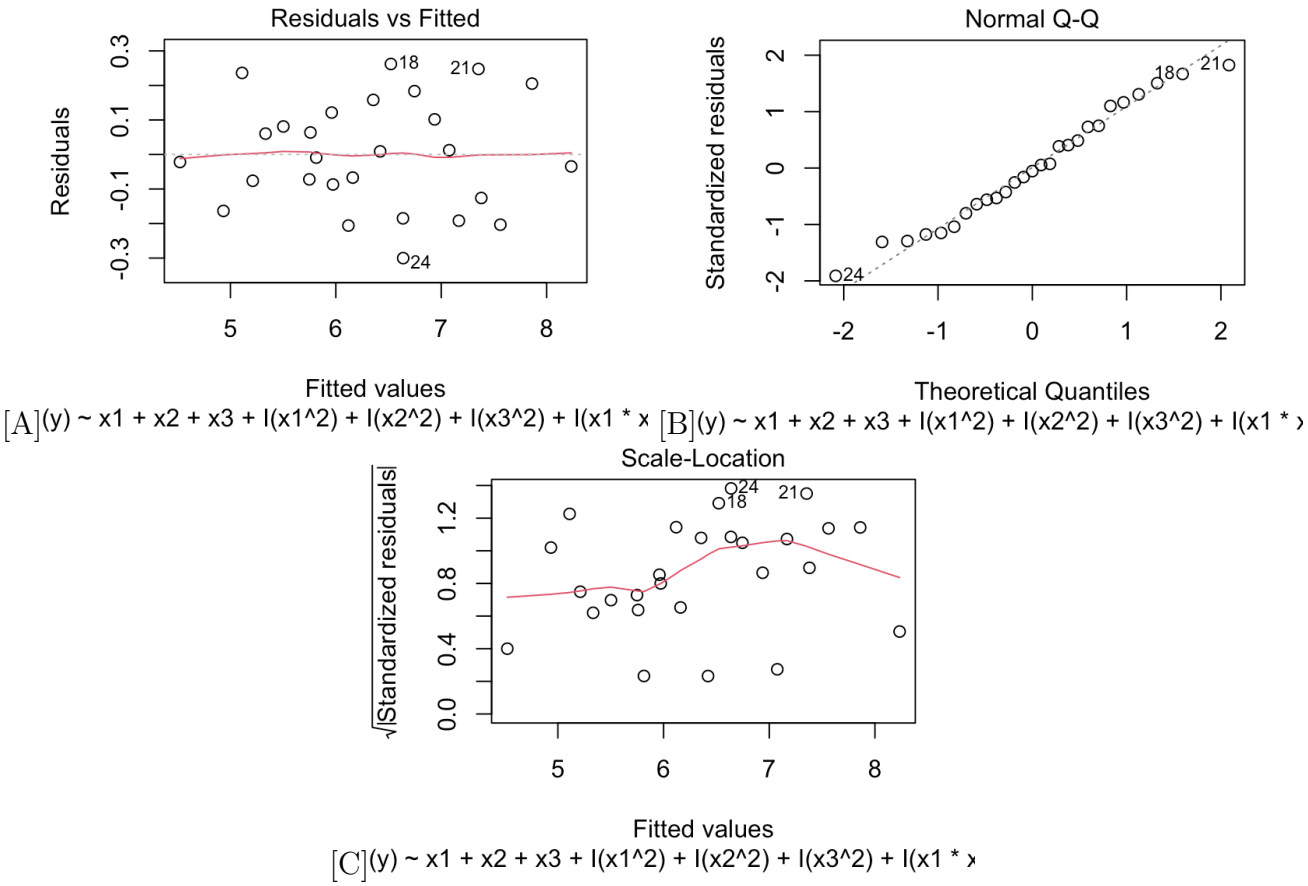


FIGURE 7 – [A] Résidus v. valeurs ajustées ; montre que après la transformation de Box-Cox, l'hypothèse de linéarité est maintenant satisfaite. [B] Graphe QQ ; montre que après la transformation, l'hypothèse de normalité des résidus est plutôt satisfaite, mais n'est pas idéale. [C] Scale-Location ; montre que l'hypothèse d'homoscédasticité ne semble pas être vérifié dans le nouveau modèle.

Voyons le sommaire :

```
    Residuals:
      Min        1Q    Median        3Q       Max
-0.300182 -0.106445 -0.009047  0.111572  0.262081

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.42070    0.09867  65.072  < 2e-16 ***
x1           0.83238    0.04568  18.224 1.36e-12 ***
x2          -0.63087    0.04568 -13.812 1.14e-10 ***
x3          -0.39262    0.04568  -8.596 1.35e-07 ***
I(x1^2)     -0.08595    0.07911  -1.086    0.292
I(x2^2)      0.02434    0.07911   0.308    0.762
I(x3^2)     -0.06733    0.07911  -0.851    0.407
I(x1 * x2)  -0.03824    0.05594  -0.684    0.503
I(x1 * x3)  -0.06841    0.05594  -1.223    0.238
I(x2 * x3)  -0.02102    0.05594  -0.376    0.712
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1938 on 17 degrees of freedom
Multiple R-squared:  0.9725,        Adjusted R-squared:  0.9579
F-statistic: 66.76 on 9 and 17 DF,  p-value: 1.724e-11
```

— R^2 : On voit que les valeurs de R^2 0.95. C'est meilleur qu'avant (C'était R 0.9).
— Signification globale du modèle : le test de Fisher rend la p-valeur $p = 10^{-11}$, c'est-à-dire, on est très sûr (niveau $\alpha \ll 0.01$) que quelque variable n'est pas zéro, i.e., que globalement le modèle est significatif.
— Signification des variables : On obtient que dans le modèle transformé, aucune variable d'ordre deux est significative (si l'on prend $\alpha = 0.1$ par exemple).

4.5 Choix de modèle, de nouveau

Comme avant, on se demande quel est le bon modèle entre les deux modèles transformés. On soupçonne que comme toutes les variables mixtes ont été rejetées, le modèle avec eux est superflu. On réalise de nouveau une analyse de variance et on obtient :

```
 Analysis of Variance Table

Model 1: log(y) ~ x1 + x2 + x3
Model 2: log(y) ~ x1 + x2 + x3 + I(x1^2) + I(x2^2) + I(x3^2) + I(x1 *
    x2) + I(x1 * x3) + I(x2 * x3)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     23 0.7925
2     17 0.6384  6    0.1541 0.6839 0.6651
```

Bien sûr, notre intuition était correcte : le modèle simple sans les variables mixtes est préférable (p-valeur $p = 0.6651 \gg \alpha = 0.1$).

5 Conclusion

L'étude suivante est une étude de la transformation de Box-Cox. La première partie est une section où nous avons étudié cette transformation et ses conditions d'applications. Elle a aussi servi à déterminer les méthodes nécessaire pour accepter ou refuser une transformation. Dans cette même partie, on définit aussi une méthode afin de trouver la valeur du paramètre λ optimale adéquate pour un certain jeu de données.
La seconde partie a été dédié à l'application de la méthode étudiée sur des données simulées. Elle a aussi permis de vérifier les résultats de la transformations et de montrer que cette méthode permet d'améliorer la distribution des données afin de les ramener vers des données normalisées. Nous avons appliqué la méthode sur un jeu de données que nous avons créé le λ initial est donc connu. La transformation de Box-Cox a permis de retrouver une valeur très proche de cette valeur. Cette partie confirme alors les résultats théoriques et les valide.
La dernière partie a été dédié à une application sur des données réelles. On observe alors que les résultats de l'application a cette méthode sur les données `NbCycleRupture`. L'utilisation de la régression linéaire sur les données avant l'utilisation de la transformation donne des résultats peu satisfaisants. Ce résultat est attendu car la distribution initiale des données ne vérifie pas les conditions d'application d'un modèle linéaire comme nous avons pu le voir. Une transformation de Box-Cox a permis d'améliorer significativement ces résultats.
A travers cette étude nous faisons ressortir la caractéristique de la transformation de Box-Cox qui permet de transformer puis quelles soit mieux adapté à des applications de modèles linéaires quand les données ont initialement une distribution non adaptée.