STA201: projet

Victor PRISER

12 novembre 2020

Table des matières

1 Introduction

J'ai à ma disposition un jeu de données qui donne le nombre de cycles de rupture d'un fil peigné en fonction de plusieurs variables. Mon but est de proposer un modèle le plus cohérent possible donnant le nombre de cycles de rupture en fonction de différents paramètres. Je peux penser à appliquer une régression linéaire. Mais le modèle est-il linéaire ? Je vais voir que non et ainsi, je vais chercher une façon de décrire le modèle de façon non-linéaire avec la transformation de Box-Cox. Puis j'essayerai de comparer le modèle linéaire et non linéaire afin de conclure sur le meilleur modèle. Mon étude se déroulera de la manière suivante :

— Je définirai le modèle de la transformation Box-Cox.

— Je vérifierai que le modèle est cohérent en le testant sur des données artificielles.

— J'appliquerai ce modèle à mes données et je le comparerai avec le modèle linéaire.

2 La transformation de Box-Cox

Pour un jeu de n données (Y_i) à expliciter en fonction de conditions d'expérience (x_i) je vais considérer le modèle suivant :

$$\forall i \quad h_\lambda(Y_i) = x_i\theta + \varepsilon_i \quad \varepsilon_i \hookrightarrow \mathcal{N}(0, \sigma^2) \tag{1}$$

Le paramètre du modèle est donc le vecteur $\mu = (\lambda, \theta, \sigma^2)$.

2.1 La fonction h_λ

Je peux considérer une fonction $\tilde{h}_\lambda$ pour $\lambda \in \mathbb{R}$ de la forme :

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

D'après (1), ε_i est une variable aléatoire à valeur dans $\mathbb{R}$ (loi normale). Donc $x_i\theta + \varepsilon_i$ peut prendre toutes les valeurs possibles dans $\mathbb{R}$. Alors que quand $\lambda \neq 0$, $\tilde{h}_\lambda$ est à valeur dans $]\frac{1}{\lambda}, \infty[$. Donc théoriquement, quand $\lambda \neq 0$, $\tilde{h}_\lambda$ n'est pas valide avec le modèle.

Cependant, quand les données sont positives donc quand $\tilde{h}_\lambda(Y_i)$ est défini, je peut pratiquement utiliser $\tilde{h}_\lambda$ dans le modèle (1) car il me suffit de fixer les ε_i en fonction des données. Et même si les ε_i ne peuvent par être très négatifs (par définition du modèle), je pourrais quand même justifier que l'échantillon ε_i suit une loi gaussienne ou non.

Mais en raison de cette légère incompatibilité théorique, je vais considérer dans la suite h_λ pour $\lambda > 0$:

$$\boxed{\forall y \in \mathbb{R} \quad h_\lambda(y) = \frac{sgn(y)|y|^\lambda - 1}{\lambda}}$$

2.2 L'estimateur du maximum de vraisemblance $\hat{\mu} = (\hat{\lambda}, \hat{\theta}, \hat{\sigma}^2)$ de μ

Je cherche d'abord l'expression de la vraisemblance L du modèle.

Je sais que si pour toute fonction u bornée on a :

$$\mathbb{E}(u(Y)) = \int u(y)L(y)\mathrm{d}y$$

Alors Y est une variable aléatoire à densité de loi L.

Je sait que les ε_i sont des variable aléatoires i.i.d de loi $\mathcal{N}(0, \sigma^2)$ donc le vecteur $\varepsilon = (\varepsilon_i)$ est un vecteur gaussien de loi $\mathcal{N}(0, \sigma^2 I_n)$. Il vient donc que le vecteur $Z = (Z_i)$ est un vecteur gaussien de loi $\mathcal{N}(X\theta, I_n\sigma^2)$.

Donc la loi de Z est :

$$f_Z(z) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(\frac{1}{2\sigma^2}|z - X\theta|^2\right)$$

Je pose :

$$H_\lambda : \begin{cases} \mathbb{R}^n \to \mathbb{R}^n \\ Y \mapsto (h_\lambda(Y_i)) \end{cases}$$

Je vois que h_λ est continue sur $\mathbb{R}$ et dérivable sur $\mathbb{R}^*$. De plus $h'_\lambda(y) = |y|^{\lambda-1}$ est positive et continue sur $\mathbb{R}^*$. J'en déduit que h_λ est de classe $\mathcal{C}^1$ dans $\mathbb{R}^*$. Donc par définition H_λ l'est aussi dans $\mathbb{R}^n\backslash\Gamma$ avec Γ l'ensemble des vecteurs qui ont au moins une composante nulle.

Comme h_λ est continue sur $\mathbb{R}$ et h'_λ est positive sur $\mathbb{R}^*$ et que $\lim_{y\to+\infty} h_\lambda(y) = +\infty$ et $\lim_{y\to-\infty} h_\lambda(y) = -\infty$, h_λ est bijective de $\mathbb{R}^*$ dans $\mathbb{R}\backslash\{\frac{1}{\lambda}\}$. Donc H_λ est bijective de $\mathbb{R}\backslash\Gamma$ dans $\mathbb{R}\backslash H_\lambda(\Gamma)$. Le determinant du jacobien est :

$$\det J = (Y_1 Y_2 ... Y_n)^{\lambda-1}$$

Donc il ne s'annule pas sur $\mathbb{R}^n\backslash\Gamma$. J'en conclue donc que H_λ est un $\mathcal{C}^1$ difféomorphisme de $\mathbb{R}^n\backslash\Gamma$ dans $\mathbb{R}^n\backslash H_\lambda(\Gamma)$ De plus, je vois que Γ est négligeable dans $\mathbb{R}^n$, donc les valeurs des intégrales ne seront pas modifiées entre les domaines $\mathbb{R}^n$ et $\mathbb{R}^n\backslash\Gamma$.

Je viens donc de montrer que le changement de variable $z = H_\lambda(y)$ est licite ainsi :

$$\begin{aligned}
\mathbb{E}(u(Y)) &= \mathbb{E}(u \circ H_\lambda^{-1}(Z)) \\
&= \int_{\mathbb{R}^n} u \circ H_\lambda^{-1}(z) f_Z(z)\mathrm{d}z \\
&= \int_{\mathbb{R}^n\backslash H_\lambda(\Gamma)} u \circ H_\lambda^{-1}(z) f_Z(z)\mathrm{d}z \\
&= \int_{\mathbb{R}^n\backslash\Gamma} u(y) f_Z(H_\lambda(y)) \left|\prod_{i=1}^n Y_i\right|^{\lambda-1} \mathrm{d}y \\
&= \int_{\mathbb{R}^n} u(y) f_Z(H_\lambda(y)) \left|\prod_{i=1}^n Y_i\right|^{\lambda-1} \mathrm{d}y
\end{aligned}$$

Ainsi :

$$\boxed{L(Y) = f_Z(H_\lambda(Y)) \left|\prod_{i=1}^{n} Y_i\right|^{\lambda-1} = \frac{|\prod_{i=1}^{n} Y_i|^{\lambda-1}}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}|H_\lambda(Y) - X\theta|^2\right)}$$

Je veux ensuite calculer l'estimateur de vraisemblance $\hat{\theta}$ et $\hat{\sigma}^2$ à λ fixé. Pour cela je les cherche tel que :

$$\frac{\partial \log L}{\partial \sigma}(\hat{\sigma}^2, \hat{\theta}) = 0 \quad \text{et} \quad \frac{\partial \log L}{\partial \theta}(\hat{\sigma}^2, \hat{\theta}) = 0$$

Le système devient :

$$\begin{cases} -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}|H_\lambda(Y) - X\hat{\theta}|^2 = 0 \\ -\frac{1}{\hat{\sigma}^3}{}^tX(H_\lambda(Y) - X\hat{\theta}) = 0 \end{cases}$$

Je suppose que le modèle est identifiable. Ainsi, j'ai que ${}^tXX$ est inversible et j'obtiens :

$$\boxed{\hat{\theta} = ({}^tXX)^{-1}{}^tXH_\lambda(Y) \quad \text{et} \quad \hat{\sigma}^2 = \frac{|H_\lambda(Y) - X\hat{\theta}|^2}{n}}$$

J'admet ici que $(\hat{\theta}, \hat{\sigma}^2)$ est le maximum de $(\theta, \sigma^2) \mapsto L(\theta, \sigma^2)$ (je peux le montrer en montrant que la Hessienne est définie positive).

Lorsque que j'évalue $\log L$ en fonction des estimateurs obtenus :

$$\boxed{\log L_{max}(\lambda) = -\frac{n}{2}\log(\hat{\sigma}^2) + (\lambda - 1)\sum_{i=1}^{n}|Y_i| + a(n)}$$

Avec :

$$\boxed{a(n) = -\frac{n}{2}(\log 2\pi + 1)}$$

Maintenant pour calculer le maximum de vraisemblance de λ il faut résoudre l'équation :

$$\frac{\partial \log L_{max}}{\partial \lambda}(\hat{\lambda}) = 0$$

Qui devient :

$$n\frac{\partial \hat{\sigma}^2}{\partial \lambda}(\hat{\lambda}) = \hat{\sigma}^2(\hat{\lambda})\sum_{i=1}^{n}|Y_i|$$

C'est une équation (d'inconnu $\hat{\lambda}$) difficile à résoudre analytiquement mais qui est peut être résolu numériquement par dichotomie par exemple.

Je remarque que $\hat{\lambda}$ est à valeur dans $\mathbb{R}_+^*$ car par définition de h_λ, $\lambda > 0$. Donc $\hat{\lambda}$ ne peut pas suivre de loi gaussienne. Et donc l'estimateur du maximum de vraisemblance $\hat{\mu}$ ne peut pas être un vecteur gaussien à distance finie.

Par contre si $\frac{1}{n}{}^tXX$ tend vers une matrice définie positive, il est possible de montrer (et on l'admettra ici) que $\hat{\mu}$ est asymptotiquement gaussien de matrice de variance $\Sigma(\mu)$. Ainsi j'ai la convergence en loi suivante :

$$\boxed{\sqrt{n}(\hat{\mu} - \mu) \underset{n\to\infty}{\hookrightarrow} \mathcal{N}(0, \Sigma(\mu))}$$

2.3 Intervalle de confiance pour λ

Quand je parlerai de convergence en loi ca sera la convergence en loi sous $\mathbb{P}_\lambda$

Par ce qui précède, j'ai la convergence en loi :

$$\sqrt{n}(\hat{\lambda} - \lambda) \underset{n\to\infty}{\hookrightarrow} \mathcal{N}(0, \Sigma_{11}(\mu))$$

Une propriété du rapport de vraisemblance donne la convergence suivante :

$$-\frac{n}{\frac{\partial^2 \log L_{\max}}{\partial \lambda^2}(\lambda)} \to \Sigma_{11}$$

Comme $\hat{\lambda}$ converge vers λ. Je déduit que $\hat{V} = -\frac{n}{\frac{\partial^2 \log L_{\max}}{\partial \lambda^2}(\hat{\lambda})}$. Et un estimateur convergent de da variance asymptotique Σ_{11}.

Je vais donc choisir $\frac{\hat{V}}{n}$ comme estimateur de la variance de $\hat{\lambda}$.

J'ai les résultats suivants :

$$\hat{V} \xrightarrow{\mathcal{L}} \Sigma_{11}(\mu) \quad \text{et} \quad \sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{11}(\mu))$$

Par le lemme de Slutsky pour $G \hookrightarrow \mathcal{N}(0, \Sigma_{11}(\mu))$:

$$U = (\hat{V}, \sqrt{n}(\hat{\lambda} - \lambda)) \xrightarrow{\mathcal{L}} (\Sigma_{11}(\mu), G)$$

J'applique à ce vecteur la fonction $h : (x, y) \mapsto \frac{y}{\sqrt{x}}$ qui est telle que pour $C = \mathbb{R}_+^* \times \mathbb{R}$ qui est un ensemble de points où h est continue :

$$\mathbb{P}((\Sigma_{11}(\mu), G) \in C) = 1$$

Donc :

$$h(U) = \sqrt{n}\frac{\hat{\lambda} - \lambda}{\sqrt{\hat{V}}} \xrightarrow{\mathcal{L}} \frac{G}{\sqrt{\Sigma_{11}(\mu)}} \hookrightarrow \mathcal{N}(0, 1)$$

Si je note k_α le quantile d'ordre $1 - \frac{\alpha}{2}$ le la loi normale $\mathcal{N}(0, 1)$ alors un intervalle de confiance asymptotique de niveau $1 - \alpha$ pour λ est :

$$\boxed{IC = \left[-k_\alpha\sqrt{\frac{\hat{V}}{n}} + \hat{\lambda}, k_\alpha\sqrt{\frac{\hat{V}}{n}} + \hat{\lambda}\right]}$$

Et donc pour n grand :

$$\mathbb{P}_\lambda(IC \ni \lambda) \approx 1 - \alpha$$

2.4 Construction de tests asymptotiques pour tester $\lambda = \lambda_0$

2.4.1 Test de Wald

1. **Modèle :** Le modèle à déjà était décrit plus haut. Je rappelle juste que c'est un modèle paramétrique de paramètre $\mu = (\lambda, \theta, \sigma^2) \in \Theta = \mathbb{R}_+^* \times \mathbb{R}^p \times \mathbb{R}_+^*$
2. **Hypothèses :** Je pose la fonction g qui à un vecteur μ associe $\lambda - \lambda_0$. Ainsi je choisis comme hypothèse nulle :
$$H_0 = \{\mu \in \Theta \quad g(\mu) = 0\}$$
Fâce à l'hypothèse alternative :
$$H_1 = \{\mu \in \Theta \quad g(\mu) \neq 0\}$$
3. **Statistique de test :** Comme j'utilise un test de Wald, la statistique est (car $\hat{\lambda}$ est l'emv de λ) :
$$\boxed{W_n = n\frac{g(\hat{\lambda})^2}{\hat{V}} = n\frac{(\hat{\lambda} - \lambda_0)^2}{\hat{V}}}$$
4. **Comportement sous H_0 :** J'ai déjà justifier dans la partie intervalle de confiance que :
$$]\sqrt{n\frac{(\hat{\lambda} - \lambda_0)^2}{\hat{V}}} \xrightarrow{\mathcal{L} \text{ sous } \mathbb{P}_{\lambda_0}} \mathcal{N}(0,1)$$
Comme le carré d'une v.a suivant une loi normal suit une loi du khi2 à 1 degrés de liberté,
Sous H_0, W_n converge en loi vers $W \hookrightarrow \chi^2(1)$.
5. **Comportement sous H_1 :** J'utilise les convergences en loi sous $\mathbb{P}_\lambda$ où $\lambda \neq \lambda_0$ (je suis donc sous H_1). J'ai par le théorème de Slutsky, $(n^{-1}, n(\hat{\lambda} - \lambda)^2)$ converge en loi vers $(0, G)$ où G est une variable aléatoire fini. Ainsi $\hat{\lambda} - \lambda$ converge en loi vers 0. Et encore $(\hat{\lambda} - \lambda_0)^2$ converge en loi vers $(\lambda - \lambda_0)^2$. Si je pose pour $\varepsilon > 0$ la fonction $h : x \mapsto \mathbb{1}_{|x-(\lambda-\lambda_0)^2|>\varepsilon}$ qui est continue en $(\lambda - \lambda_0)^2$. Donc :
$$\mathbb{E}_\lambda(h(\hat{\lambda} - \lambda_0)^2)) = \mathbb{P}_\lambda(|(\hat{\lambda} - \lambda_0)^2 - (\lambda - \lambda_0)^2| > \varepsilon) \to \mathbb{E}_\lambda(h(\lambda - \lambda_0)^2)) = 0$$
Donc $(\hat{\lambda} - \lambda_0)^2$ converge en probabilité vers $(\lambda - \lambda_0)^2 > 0$.
Il est donc clair que W_n converge en probabilité vers $+\infty$ sous H_1.
6. **La région critique :** Comme sous H_1 la statistique prend des valeurs élevées, je vais choisir une région de rejet de la forme :
$$\boxed{\mathcal{R} = \{W_n > k_\alpha\}}$$
7. **Erreur de première espèce :** Je veux une érreur de première espèce inférieur à α. Soit, pour μ vérifiant H_0 :
$$\mathbb{P}_\mu(\mathcal{R}) \leq \alpha$$
Comme sous H_0 la statistique suit asymptotiquement une loi du khi2 à 1 degrés de liberté, je peux choisir k_α comme étant le quantile de niveau $1 - \alpha$ de la loi $\chi^2(1)$.
J'aurais bien un test asymptotique de niveau α.

8. **Erreur de seconde espèce :** Si H_1 est vérifiée alors W_n converge en probabilité vers $+\infty$ donc :
$$\lim \mathbb{P}_{\mu \in H_1}(W_n > k_\alpha) = 1$$
Et donc l'erreur de seconde espèce :
$$\mathbb{P}_{\mu \in H_1}(W_n < k_\alpha) = 1 - \mathbb{P}_{\mu \in H_1}(W_n > k_\alpha)$$
Tend vers 0 à l'infini.
Le test est donc convergent
9. **La p-valeur :** Si j'observe une valeur de la statistique W_n^{obs} la p-valeur sera définie comme le plus petit α tel que :
$$W_n^{\text{obs}} > k_\alpha$$
Donc :
$$k_{p_{\text{val}}} = W_n^{\text{obs}}$$
Si je pose F la fonction de répartition de la loi du khi2 à 1 degrés de liberté :
$$\boxed{p_{\text{val}} = 1 - F(W_n^{\text{obs}})}$$

2.4.2 Test du rapport de vraisemblance

1. **Modèle :** Le modèle à déjà était décrit plus haut. Je rappelle juste que c'est un modèle paramétrique de paramètre $\mu = (\lambda, \theta, \sigma^2) \in \Theta = \mathbb{R}_+^* \times \mathbb{R}^p \times \mathbb{R}_+^*$
2. **Hypothèses :** Je pose la fonction g qui à un vecteur μ associe $\lambda - \lambda_0$. Ainsi je choisis comme hypothèse nulle :
$$H_0 = \{\mu \in \Theta \quad g(\mu) = 0\}$$
Fâce à l'hypothèse alternative :
$$H_1 = \{\mu \in \Theta \quad g(\mu) \neq 0\}$$
3. **Statistique de test :** Comme je veux un test du rapport de vraisemblance asymptotique, la statistique est (car $\hat{\lambda}$ est l'emv de λ) :
$$\boxed{\zeta_n = -2 \log \frac{\sup_{\mu \in H_0} L}{\sup_{\mu \in \Theta} L} = 2(\log L_{\max}(\hat{\lambda}) - \log L_{\max}(\lambda_0))}$$
4. **Comportement sous H_0 :** Une propriété de ce test dit que comme H_0 peut être vue comme une variété de dimension $\dim(\Theta) - 1$:
Sous H_0, ζ_n converge en loi vers $\zeta \hookrightarrow \chi^2(1)$.
5. **Comportement sous H_1 :** Une propriété du test de rapport de vraisemblance donne aussi que :
ζ_n converge en probabilité vers $+\infty$ sous H_1.

6. **La région critique :** Comme sous H_1 la statistique prend des valeurs élevées, je vais choisir une région de rejet de la forme :
$$\boxed{\mathcal{R} = \{\zeta_n > k_\alpha\}}$$
7. **Erreur de première espèce :** Je veux une érreur de première espèce inférieure à α. Soit, pour μ vérifiant H_0 :
$$\mathbb{P}_\mu(\mathcal{R}) \leq \alpha$$
Comme sous H_0 la statistique suit asymptotiquement une loi du khi2 à 1 degrés de liberté, je peux choisir k_α comme étant le quantile de niveau $1-\alpha$ de la loi $\chi^2(1)$.
J'aurais bien un test asymptotique de niveau α.
8. **Erreur de seconde espèce :** Si H_1 est vérifiée alors ζ_n converge en probabilité vers $+\infty$ donc :
$$\lim \mathbb{P}_{\mu \in H_1}(\zeta_n > k_\alpha) = 1$$
Et donc l'erreur de seconde espèce :
$$\mathbb{P}_{\mu \in H_1}(\zeta_n < k_\alpha) = 1 - \mathbb{P}_{\mu \in H_1}(\zeta_n > k_\alpha)$$
Tend vers 0 à l'infini.
Le test est donc convergent
9. **La p-valeur :** Si j'observe une valeur de la statistique ζ_n^{obs} la p-valeur sera définie comme le plus petit α tel que :
$$\zeta_n^{\text{obs}} > k_\alpha$$
Donc :
$$k_{p_{\text{val}}} = \zeta_n^{\text{obs}}$$
Si je pose F la fonction de répartition de la loi du khi2 à 1 degrés de liberté :
$$\boxed{p_{\text{val}} = 1 - F(\zeta_n^{\text{obs}})}$$

3 Test de la méthode sur des données simulées

Dans la partie théorique, j'ai imposé $\lambda > 0$ mais dans toute la suite, je vais considérer que les résultats sont les mêmes pour $\lambda \in \mathbb{R}$ où $h_0 = \log$

Je souhaite simuler des données suivant le modèle :
$$h_{0.3}(Y_i) = a + bx_i + \varepsilon_i \quad \varepsilon_i \hookrightarrow_{iid} \mathcal{N}(0, \sigma^2)$$
Avec $a = 5$, $b = 1$, $\sigma^2 = 2$ et les x_i qui sont issues d'une loi normale centrée réduite. Il faut vérifier que x coïncide avec mon hypothèse $\frac{1}{n}{}^t XX \to A$ avec A une matrice définie positive.

Dans mon cas :
$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Donc :

$$\frac{1}{n}{}^tXX = \begin{pmatrix} 1 & \bar{x}_n \\ \bar{x}_n & \sum_{i=1}^n \frac{x_i^2}{n} \end{pmatrix}$$

Comme les x_i sont issues de loi gaussienne centrées réduites i.i.d par la loi forte des grands nombres :

$$\bar{x}_n \to 0 \quad \text{et} \quad \sum_{i=1}^n \frac{x_i^2}{n} \to 1$$

Ainsi :

$$\frac{1}{n}{}^tXX \to I_2$$

I_2 est trivialement définie positive donc pour mes valeurs, je reste dans le cadre de la section précédente.

3.1 Calcule d'un Y simulé

Dans R, je vais donc définir tous les paramètres du modèle :

```
> set.seed(999) #initialisation de l'aléa
> n=50
> x=rnorm(n)
> a=5
> b=1
> sigma2=2
```

Il faut maintenant simuler les ε_i qui sont issues d'une loi normale $\mathcal{N}(0,2)$. Pour cela j'utilise le fait que $\sqrt{2}x$ est issu d'une loi normale $\mathcal{N}(0,2)$ si x est un échantillon issue d'une loi normale centrée réduite.

```
> epsilon=sqrt(2)*rnorm(n)
```

Donc l'échantillon $Z = H_{0.3}(Y)$ s'obtient :

```
> Z = a + b*x + epsilon
```

Pour calculer Y il faut distinguer 2 cas :

— $Z_i \geq -\frac{1}{0.3} \implies Y_i \geq 0$ donc :

$$Y_i = \left(\frac{3}{10}Z_i + 1\right)^{\frac{10}{3}}$$

— $Z_i \leq -\frac{1}{0.3} \implies Y_i \leq 0$ donc :

$$Y_i = -\left(-\frac{3}{10}Z_i - 1\right)^{\frac{10}{3}}$$

Donc je calcule Y de la façon suivante :

```
> f<-function(x,lambda=0.3){
+     if(x>-1/lambda){
+         y=(lambda*x+1)^(1/lambda)
+     }else{
+         y=-(-lambda*x-1)^(1/lambda)
+     }
+     return(y)}
> Yobs=sapply(Z,f) #application de f à toutes les composantes de Z
```

Pour la suite, je fixe les deux matrices H et Q pour éviter de les recalculer à chaque fois :

```
> X=cbind(1,x)
> H=solve(t(X)%*%X)%*%t(X)
> Q=diag(1,n) - X%*%H
```

Maintenant je vais faire comme si je ne connaît que X et Y.

3.2 Justification de l'utilité d'un modèle non linéaire

Je fais une régression linéaire de $Z = H_{0.3}(Y)$ en fonction de x :

```
> dfZx = cbind.data.frame(data.frame(x),data.frame(Z))
> resZx=lm(Z~.,data=dfZx)
> summary(resZx)


Call:
lm(formula = Z ~ ., data = dfZx)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9616 -0.7887 -0.1404  1.1620  2.3727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0693     0.1835  27.626  < 2e-16 ***
x             1.0314     0.1811   5.695 7.32e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.255 on 48 degrees of freedom
Multiple R-squared:  0.4032,        Adjusted R-squared:  0.3908
F-statistic: 32.43 on 1 and 48 DF,  p-value: 7.315e-07
```

Puis si je trace la régression :

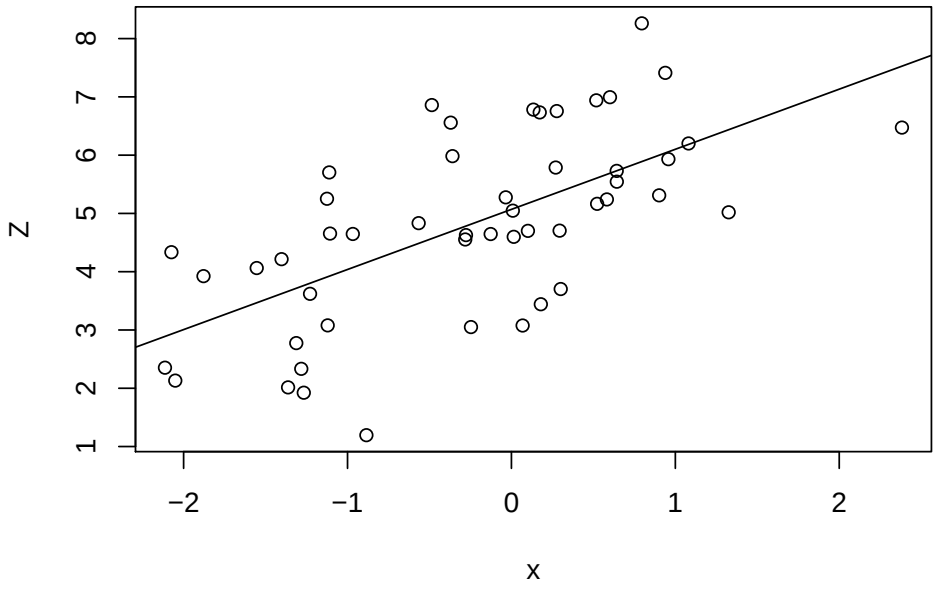

FIGURE 1 – $H_{0.3}(Y)$ en fonction de x

Maintenant si je fais une régression linéaire de Y en fonction de x

```
> dfYx = cbind.data.frame(data.frame(x),data.frame(Yobs))
> resYx=lm(Yobs~.,data=dfYx)
> summary(resYx)


Call:
lm(formula = Yobs ~ ., data = dfYx)

Residuals:
    Min      1Q  Median      3Q     Max
-16.281  -6.672  -2.398   7.100  32.707

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.566      1.573  15.620  < 2e-16 ***
x              8.187      1.552   5.274 3.15e-06 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.75 on 48 degrees of freedom
Multiple R-squared:  0.3669,        Adjusted R-squared:  0.3537
F-statistic: 27.82 on 1 and 48 DF,  p-value: 3.15e-06
```

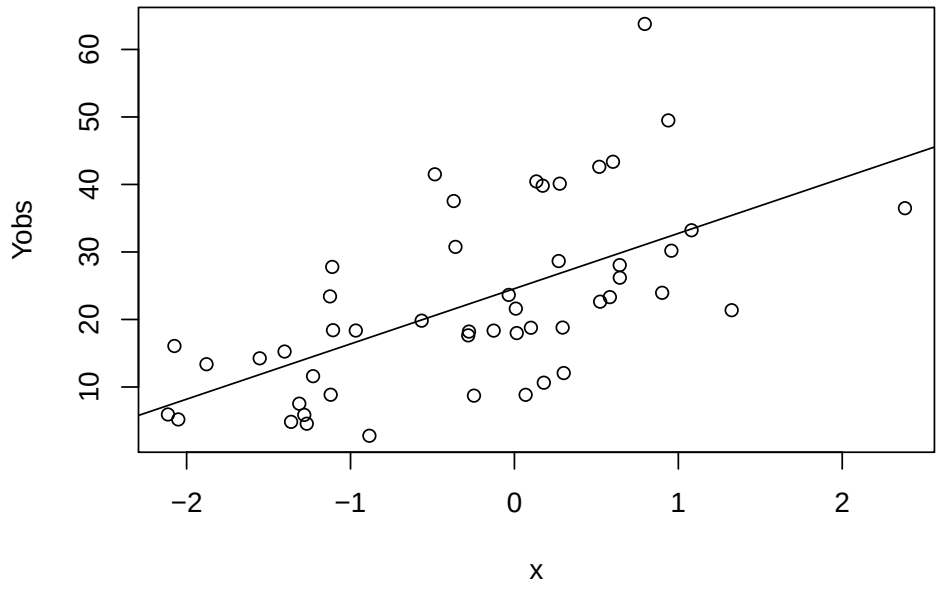

FIGURE 2 – Y en fonction de x

A vue d'oeil, les deux régréssion ont le même ajustement. En effet, elles ont toutes les deux un R^2 similaire. Pour pouvoir mieux comparer les deux régréssion, il faut étudier les résidus. Je calcule donc les résidus de la regréssion de Z et de Y :

```
> resY= Yobs - resYx$coeff[1] - resYx$coeff[2]*x
> resZ= Z - resZx$coeff[1] - resZx$coeff[2]*x
```

Et j'obtiens :

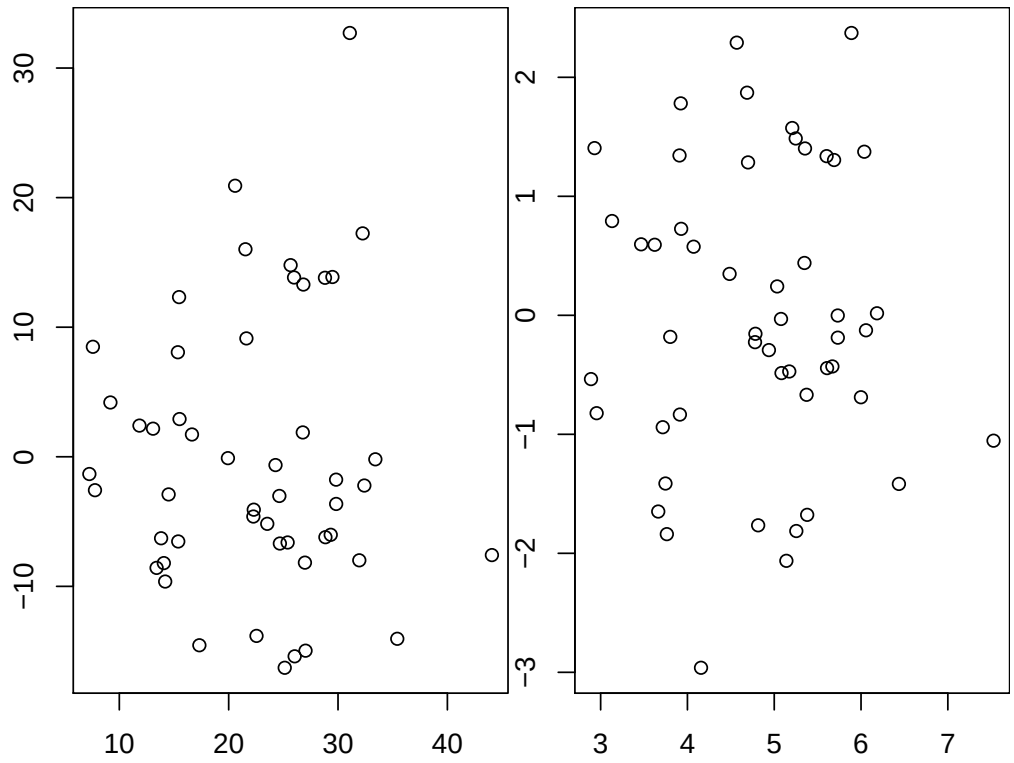

FIGURE 3 – résidu de Y à gauche et de Z à droite de Y-ajusté à gauche et Z-ajusté à droite

Pour mieux les étudier (il faut que les résidus soit de variance environ égal à 1), je vais les studentiser :

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

Je vais donc calculer la matrice de projection :

$$H = X({}^tXX)^{-1t}X$$

Et j'ai $\hat{\sigma}$ qui s'obtient par la formule :

$$\hat{\sigma} = \sqrt{\frac{1}{n}}||Y - X\hat{\theta}|| = \sqrt{\frac{1}{n}}||\varepsilon||$$

```
> sigY=sqrt(1/n)*norm(as.matrix(resY),"2")
> sigZ=sqrt(1/n)*norm(as.matrix(resZ),"2")
> PH = X%*%H
> for(i in 1:n){
+     resY[i] = resY[i]/(sigY*sqrt(1-PH[i,i]))
+     resZ[i] = resZ[i]/(sigZ*sqrt(1-PH[i,i]))
+ }
```

Je les affiches avec un délimiteur qui permêt d'observer les valeurs de résidu aberrantes (>2).

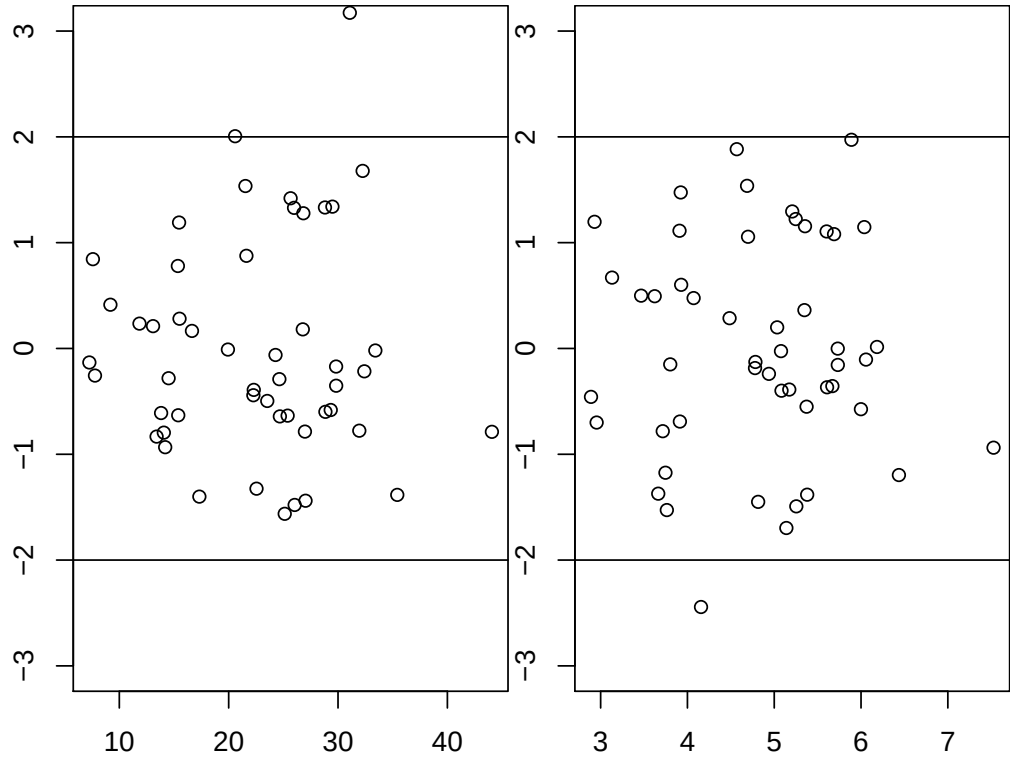

FIGURE 4 – résidu de Y à gauche et de Z à droite en fonction de Y-ajusté à gauche et Z-ajusté à droite

Je devine de l'hétéroscédasticité pour les résidus de la régression avec Y car les points s'étalent plus quand la valeur de l'estimation de Y augmente. Et donc la transformation semble enlever cette hétéroscédasticité.

Il faut maintenant voir si les résidus sont normaux. En effectuant un test de Shapiro-Wilks qui teste l'hypothèse nulle : *L'échantillon est gaussien* fâce à l'hypothèse alternative : *Les échantillon ne sont pas gaussien* :

```
> shapiro.test(resY)


        Shapiro-Wilk normality test

data:  resY
W = 0.94108, p-value = 0.01486


> shapiro.test(resZ)
```

```
        Shapiro-Wilk normality test

data:  resZ
W = 0.97925, p-value = 0.5209
```

Le test de shapiro-Wilks me permet de refuser l'hypothèse de normalité du résidu de Y avec une probabilité égale à 1.5%(= p-val) de se tromper (car on refuse l'hypothèse nulle pour un test de niveau 1.5%).

Le test de shapiro-Wilks ne me permet pas de refuser la normalité (p-val élevée) de Z. Je peux donc accepter l'hypothèse de normalité avec une probabilité inconnue de se tromper.

Donc au final le modèle linéaire avec bruit gaussien de Y n'est pas valide. Il faut un modèle non-linéaire comme Z.

3.3 Détermination de $\hat{\lambda}$

Par définition $\hat{\lambda}$ est le minimum de la fonction $-\log L_{\max}$. Donc il faut déterminer cette fonction.

J'ai besoin de la fonction h_λ.

```
> h<-function(lambda){
+     return(function(y){
+         if(lambda==0){
+             return(log(y))
+         }else{
+         if(y>0){
+             return((y^lambda - 1)/lambda)
+         }else{
+             return((-1-(-y)^lambda)/lambda)
+     }}
+     })
+ }
```

Ensuite j'ai besoin de l'estimateur de θ.

```
> theta.est<-function(lambda,Y=Yobs){
+     Z=sapply(Y,h(lambda))
+     return(H%*%Z)
+ }
```

Puis de l'estimateur de σ^2 (en remarquant que Q est une matrice orthogonale symétrique donc ${}^tQQ = Q^2 = Q$ donc ${}^t(ZQ)ZQ = {}^tZQZ$)

```
> sigma2.est<-function(lambda,Y=Yobs){
+     Z=sapply(Y,h(lambda))
+     return(t(Z)%*%Q%*%Z/n)
+ }
```

Et je peux enfin determiner la fonction $lmin = -\log L_{\max}$

```
> lmin<-function(lambda,Y=Yobs){
+     return(n/2*log(sigma2.est(lambda,Y))-(lambda-1)*sum(log(abs(Y)))
+     + n/2*(log(2*pi)+1))
+ }
```

Et en voici le tracé :

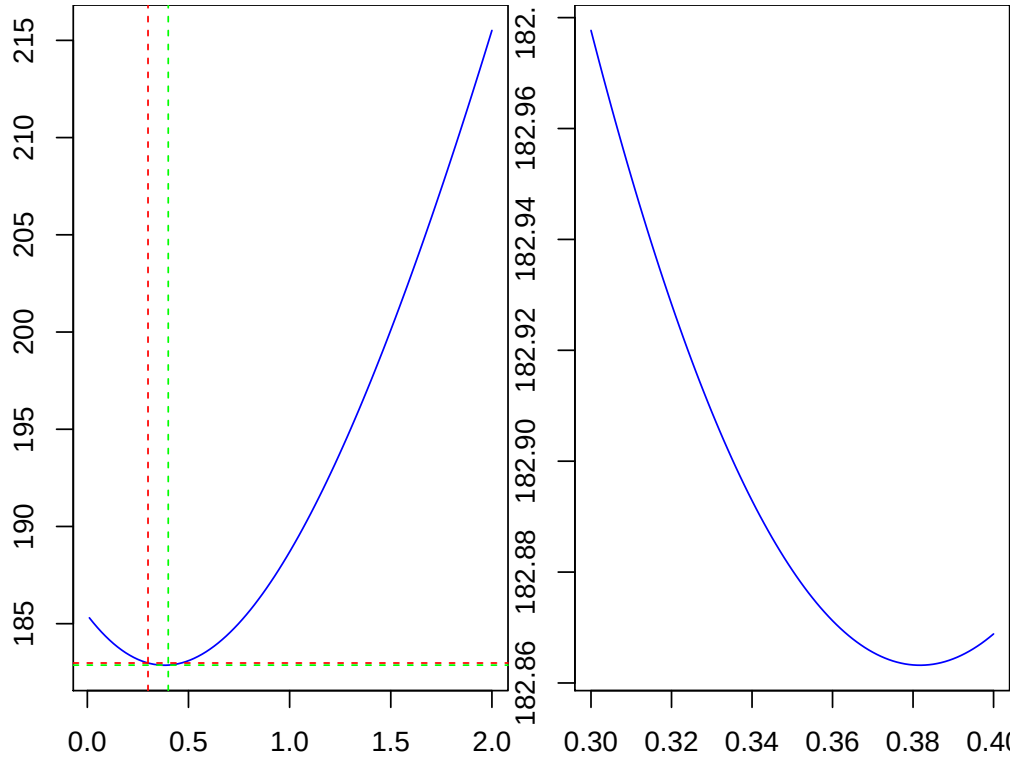

FIGURE 5 – Tracé de $-\log L_{\max}$ en fonction de λ. En rouge c'est le point d'abscisse $\lambda = 0.3$ et en vert c'est le point d'abscisse $\lambda = 0.4$. J'ai fait un zoom entre ces deux valeurs à droite.

Je vois graphiquement que : $\hat{\lambda} \approx 0.38$.

R propose une $\hat{\lambda}$ numérique trouvé par un algorithme de Newton.

```
> resopt=nlm(lmin,p=3,hessian=T)
> lambda.est=resopt$estimate
> print(lambda.est)
```

```
[1] 0.381728
```

Comme d'apres la partie théorique un estimateur de la variance de l'estimateur de $\hat{\lambda}$ est : $\frac{1}{\frac{\partial^2 \log L_{\max}}{\partial \lambda^2}(\hat{\lambda})}$

```
> var.lambda.est=1/resopt$hessian
> print(var.lambda.est)

           [,1]
[1,] 0.02948458
```

Je peux ensuite déduire un intervalle de confiance asymptotique $[a_1, a_2]$:

```
> alpha=0.05
> kalpha=qnorm(1-alpha/2,0,1)
> a1 = lambda.est -kalpha*sqrt(var.lambda.est)
> a2 = lambda.est +kalpha*sqrt(var.lambda.est)
> print(a1)

           [,1]
[1,] 0.04518111


> print(a2)

          [,1]
[1,] 0.7182748
```

Je résume les résultats obtenus dans un tableau :

$\hat{\lambda}$	0.38
$\hat{\sigma}^2$	2.41
$\hat{\theta}$	$(5.9, 1.3)$
$\frac{\hat{V}}{n}$	0.030
IC	$[0.045, 0.718]$

3.4 Test sur λ

3.4.1 Test de Wald

Je vais faire plusieurs test de Wald pour tester $\lambda = \lambda_0$ avec une alternative bilatérale.

Je calcule la statistique pour mes observations :

```
> Wobs<-function(lambda0){
+     return(((lambda.est-lambda0)^2)/var.lambda.est)
+ }
```

Donc je vais calculer la p-valeur pour tous les tests suivants :

— H_0 : Les données Y ne nécessite pas de transformation ($\lambda = 1$)

```
> print(1-pchisq(Wobs(1),1))
```

```
                [,1]
[1,] 0.0003174102
```

— H_0 : La transformé à appliquer est en racine carré ($\lambda = \frac{1}{2}$)

```
> print(1-pchisq(Wobs(1/2),1))
```

```
           [,1]
[1,] 0.4909577
```

— $H_0 : \lambda = 0.3$

```
> print(1-pchisq(Wobs(0.3),01))
```

```
           [,1]
[1,] 0.6341007
```

— $H_0 : \lambda = 0$ (Je vais supposer que les résultats sont les mêmes quand j'ai $\tilde{h}_0$ comme transformation)

```
> print(1-pchisq(Wobs(0),01))
```

```
            [,1]
[1,] 0.02620991
```

3.4.2 Test du rapport de vraisemblance

Je fais la même chose et la statistique observé est :

```
> zetaObs<-function(lambda0,lambda.estime=lambda.est,Y=Yobs){
+     return(2*(lmin(lambda0,Y)-lmin(lambda.estime,Y)))
+ }
```

Je calcule encore la p-valeur :

— H_0 : Les données Y ne nécessite pas de transformation ($\lambda = 1$)

```
> print(1-pchisq(zetaObs(1),1))

             [,1]
[1,] 0.0006381486
```

— H_0 : La transformé à appliquer est en racine carré ($\lambda = \frac{1}{2}$)

```
> print(1-pchisq(zetaObs(1/2),1))

         [,1]
[1,] 0.494657
```

— $H_0 : \lambda = 0.3$

```
> print(1-pchisq(zetaObs(0.3),1))

          [,1]
[1,] 0.6322771
```

— $H_0 : \lambda = 0$

```
> print(1-pchisq(zetaObs(0),01))

           [,1]
[1,] 0.02326665
```

3.4.3 Résumé et interprétation

Je résume mes p-valeurs dans le tableau suivant :

Test \ λ	1	$\frac{1}{2}$	0.3	0
Test de Wald	$3.2 \cdot 10^{-4}$	0.49	0.63	$2.6 \cdot 10^{-2}$
TRV	$6.4 \cdot 10^{-4}$	0.49	0.63	$2.3 \cdot 10^{-2}$

J'observe que les deux tests donnent presque les mêmes p-valeurs. Ceci est rassurant car je travail sur des tests asymptotiques. Donc je veux n suffisamment grand afin d'être proche des valeurs asymptotiques. Comme les deux tests ont les mêmes comportements asymptotique, si les deux tests avait donné des résultats différents cela aurait voulu dire qu'au moins un des deux tests n'est pas dans sa région asymptotique (est donc pas valide).

Pour ce qui est des résultats, je peux accepter les hypothèses $\lambda = 0.3$ et $\lambda = \frac{1}{2}$ avec une probabilité inconnue de se tromper.

Et je refuse les hypothèses $\lambda = 1$ et $\lambda = 0$ avec une probabilité proche (si on est dans la région asymptotique) de la p-valeur de se tromper. Au vu, de mon étude des résidus, il est logique de refuser l'hypothèse $\lambda = 1$.

Je peux retrouver les résultats du TRV grâce à la fonction *powerTransform* de R :

```
> library(carData)
> library(car)
> dfYx = cbind.data.frame(data.frame(x),data.frame(Yobs))
> resYx=powerTransform(Yobs~.,data=dfYx)
> summary(resYx)


bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.3817         0.5       0.0452       0.7183

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                            LRT df     pval
LR test, lambda = (0) 5.148486  1 0.023267

Likelihood ratio test that no transformation is needed
                            LRT df       pval
LR test, lambda = (1) 11.66127  1 0.00063815
```

Je peux donc voir que l'intervalle de confiance est le même. La p-valeur du TRV aussi la même pour le test d'hypothèse $\lambda = 1$ et $\lambda = 0$.

3.5 Simulation de la puissance

Je vais simuler dans un premier temps le niveau du TRV pour un test asymptotique à 5%. Par définition du niveau α :

$$\alpha = \mathbb{P}_{H_0}(\mathcal{R}_\alpha)$$

Je répète n fois l'expérience suivante :

Je simule des données Y qui suivent H_0 et je pose (X_i) la variable aléatoire qui vaut 1 si le TRV de niveau asymptotique 5% ne valide pas $\lambda = 0.3$ et 0 sinon. Les X_i sont indépendants et de même loi donc par la loi forte des grand nombre :

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p.s} \mathbb{E}_{H_0}(X_1) = \mathbb{P}_{H_0}(X_1 = 1) = \alpha$$

Donc si je répète 100 fois l'expérience je devrais déjà avoir une bonne éstimation du niveau α.

Je vais donc coder la fonction *nivTRV* qui donne la moyenne des X_i pour k X_i :

```
> nivTRV<-function(k){
+     niv=0
+     for (i in 1:k){
+         epsilon=sqrt(2)*rnorm(n)  #nouveau epsilon
+         Z = a + b*x + epsilon
+         newY=sapply(Z,f)
+         resopt=nlm(lmin,p=0.3,hessian=T,Y=newY)
+         lambda.est=resopt$estimate
+         if(1-pchisq(zetaObs(0.3,lambda.est,newY),1)<0.05){ #refus de H_0
+             niv=niv+1
+         }
+     }
+     return(niv/k)
+ }
> print(nivTRV(1000))
```

```
[1] 0.04
```

J'ai donc un niveau simulé de 4% c'est rassurant car je suis proche du niveau asymptotique. Cela ne démenti pas le fait que ma statistique est proche de celle asymptotique.

Je peux conclure que le test refusera plus facilement H_0 pour des échantillons petits (par rapport à ce que préconise le niveau asymptotique).

En répétant cette méthode, il est possible de simuler la puissance du TRV. En effet, pour λ_1 quelconque, si je simule n Y qui suivent le modèle $\lambda = \lambda_1$ et que je les teste à l'aide du TRV d'hypothèse nulle : $\lambda = 0.3$ fâce à l'hypothèses bilaterale. Si je note k le nombre de rejet de H_0 alors en appliquant la même méthode que pour la simulation du niveau, $\frac{k}{n}$ est une bonne approximation de la puissance en λ_1 qui est $\mathbb{P}_{\lambda_1}(\mathcal{R}_\alpha)$.

Donc la fonction suivante donne une bonne approximation de la puissance :

```
> pw<-function(lambda1,k=1000){
+     niv=0
+     for (i in 1:k){
+         epsilon=sqrt(2)*rnorm(n)
+         Z = a + b*x + epsilon
+         newY=sapply(Z,f,lambda=lambda1)
+         resopt=nlm(lmin,p=lambda1,hessian=T,Y=newY)
+         lambda.est=resopt$estimate
+         if(1-pchisq(zetaObs(0.3,lambda.est,newY),1)<0.05){
```

```
+             niv=niv+1
+         }
+     }
+     return(niv/k)
+ }
> print(pw(0.1))
```

```
[1] 0.539
```

Si je trace la courbe de la puissance, j'obtiens :

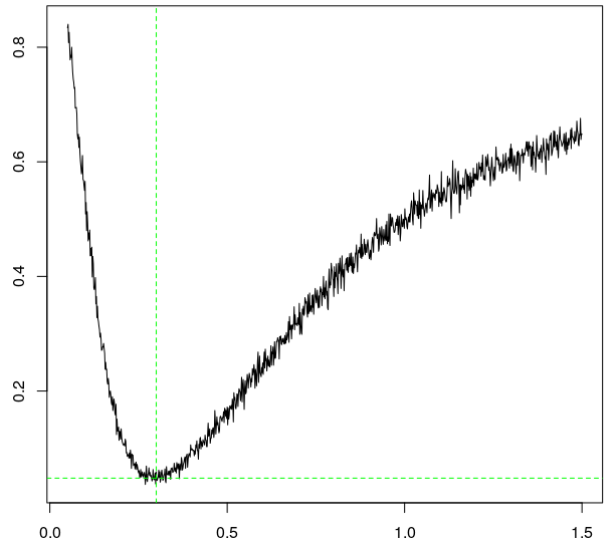

FIGURE 6 – Puissance du test (en vert j'ai le point $\lambda = 0.3$) je répète 1000 fois le test et j'ai 1000 points

Je peux déjà remarquer que la simulation m'indique que le test est sans biais car le minimum de la puissance est atteint en $\lambda = 0.3$. Et je peux remarqué que si le modèle suit λ_1. $\lambda = 0.3$ sera plus facilement accéptée si $\lambda_1 - \lambda > 0$ (pente plus faible à droite du minimum).

Je peux faire le même pour le TRV qui teste $\lambda = 0$ au niveau asymptotique 5%. Et je trâce la puissance sur l'intervalle $[0.1, 0.3]$:

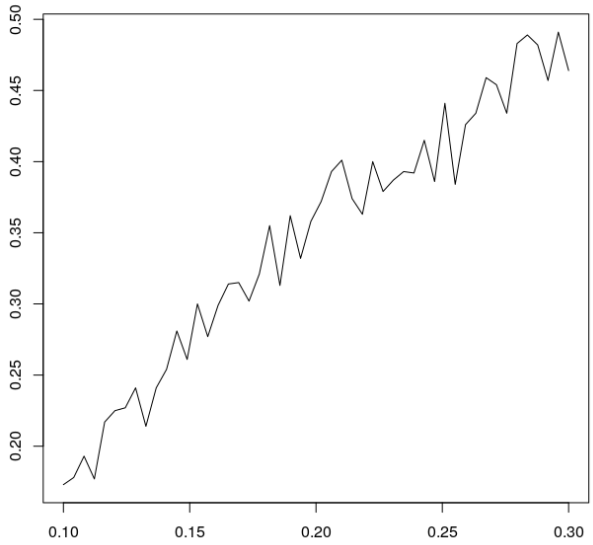

FIGURE 7 – Puissance du test je répète 1000 fois le test et j'ai 50 points

Je remarque que la puissance est donc croissante à droite de $\lambda = 0$. Donc plus le λ du modèle est grand moins le modèle à de chance d'être accépté comme étant logarithmique. Cela coïncide avec l'idée que h_λ peur être prolongé par continuité par $\log(y)$ quand λ tend vers 0.

4 Modèle pratique

Je considère des donées y(cycles de ruptures) en fonction de 3 donées la longueur l, l'amplitude du cycle de chargement a et le chargement c. Puis ta transformation suivante à été appliqué :

$$\begin{cases} x_1 = \dfrac{l - 300}{50} \\ x_2 = a - 9 \\ x_3 = \dfrac{c - 45}{5} \end{cases} \tag{2}$$

Je récupère les données dans un data-frame :

```
> df<- read.table("NbCycleRupture.csv",header=TRUE,sep=";")
```

4.1 Modèle avec effets additifs

Je considère le modèle (M1) suivant :

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \varepsilon$$

Où ε est vecteur gaussien centré de variance σ^2.

Mais le modèle peut aussi s'écrire sous la forme :

$$y = \mu_0 + \mu_1 l + \mu_2 a + \mu_3 c + \varepsilon$$

Je peux maintenant ajuster le modèle :

```
> res1=lm(y~.,data=df)
> print(summary(res1))


Call:
lm(formula = y ~ ., data = df)

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
```

Sans la transformation de variable j'aurai eu :

```
> l=50*df$x1 + 300
> a=df$x2 + 9
> c=5*df$x3 + 45
> y=df$y
> newdf=data.frame(y,l,a,c)
> res1sansTF=lm(y~.,data=newdf)
> print(summary(res1sansTF))


Call:
lm(formula = y ~ ., data = newdf)
```

```
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) 4521.370   1621.721   2.788 0.010454 *
l             13.200      2.301   5.736 7.66e-06 ***
a           -535.833    115.057  -4.657 0.000109 ***
c            -62.167     23.011  -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
```

Je vois donc que la transformation de variable n'a ni changé la significativité du test ni l'ajustement du test. Ceci est logique car la transformation est linéaire. Cette transformation a seulement permis d'uniformiser les variables x_i (elles ont le même ordre de grandeur donc en fonction des coefficients de la régression, il est plus facile de voir quelles variables ont le plus d'impactes sur y).

J'observe un R^2 de 0.7 donc la régression est bien ajusté. La p-valeur du test de significativité globale de régression (qui teste l'hypothèse d'un modèle i.i.d) est faible. Ainsi je peux conclure q'un modèle i.i.d ($\lambda = 1$) ne conviendrait pas si le modèle $(M1)$ est vérifié. Pour ce qui est du test de significativité de chaque variable je vois que pour chaque variables, la p-valeur du test de student qui teste l'hypothèse $\theta_i = 0$ est faible. Ainsi pour le modèle (M_1) je peux considérer que toutes les variables explicatives sont significatives avec probabilité inférieure à 5% pour chaque variables.

Si j'analyse les résidus :

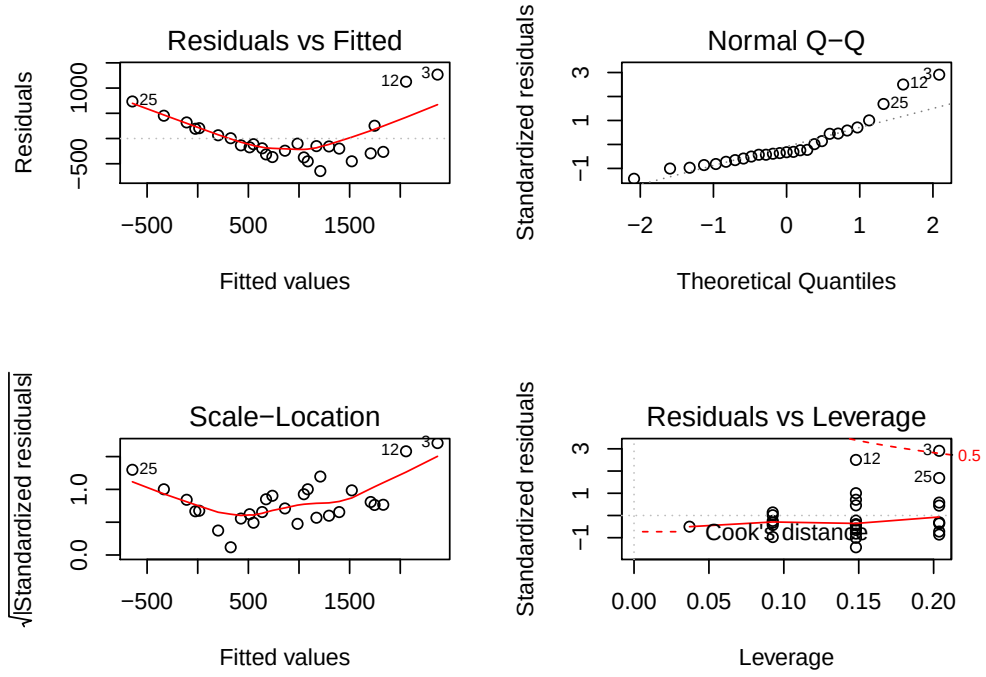

FIGURE 8 – résidus pour le modèle ajusté M1

Quand j'observe la courbe *residuals vs Fitted* je vois des non-linéarités (courbe moyenne non constante égale à 0). Il y a aussi de l'hétéroscédasticité car j'observe une variance faible pour les premiers résidus (proche de la courbe de moyenne) et une variance élevée pour les derniers résidus (loin de la courbe de la moyenne). Il semble aussi y avoir des valeurs influentes (les résidus supérieur à 1000).

La courbe *Normal Q-Q* semble indiquer que les résidus ne sont pas issues d'une loi normale.

Les deux autres courbes indiquent quelques valeurs influente par exemple 3 qui a une distance de Cook élevée

En conclusion, les hypothèses du modèle M1 ne semblent pas réspectés.

4.2 Modèle polynomiale d'ordre 2

Je considère le modèle (M2) :

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_1^2 + \theta_5 x_2^2 + \theta_6 x_3^2 + \theta_7 x_1 \cdot x_2 + \theta_8 x_1 \cdot x_3 + \theta_9 x_2 \cdot x_3 + \varepsilon$$

```
> res2=lm(y~x1+x2+x3+I(x1^2)+I(x2^2)+I(x3^2)+I(x1*x2)+I(x1*x3)+I(x2*x3),data=df)
> print(summary(res2))
```

```
Call:
lm(formula = y ~ x1 + x2 + x3 + I(x1^2) + I(x2^2) + I(x3^2) +
    I(x1 * x2) + I(x1 * x3) + I(x2 * x3), data = df)

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
```

Je vois que avec les nouveaux termes le modèle est mieux ajusté (R^2 plus grand). Le test de student qui teste la significativité de toutes les variables explicatives m'indique que $(x_3)^2$ n'est pas significative. Il serait donc possible d'enlever plusieurs variables au modèle.

Si je regarde les résidus :

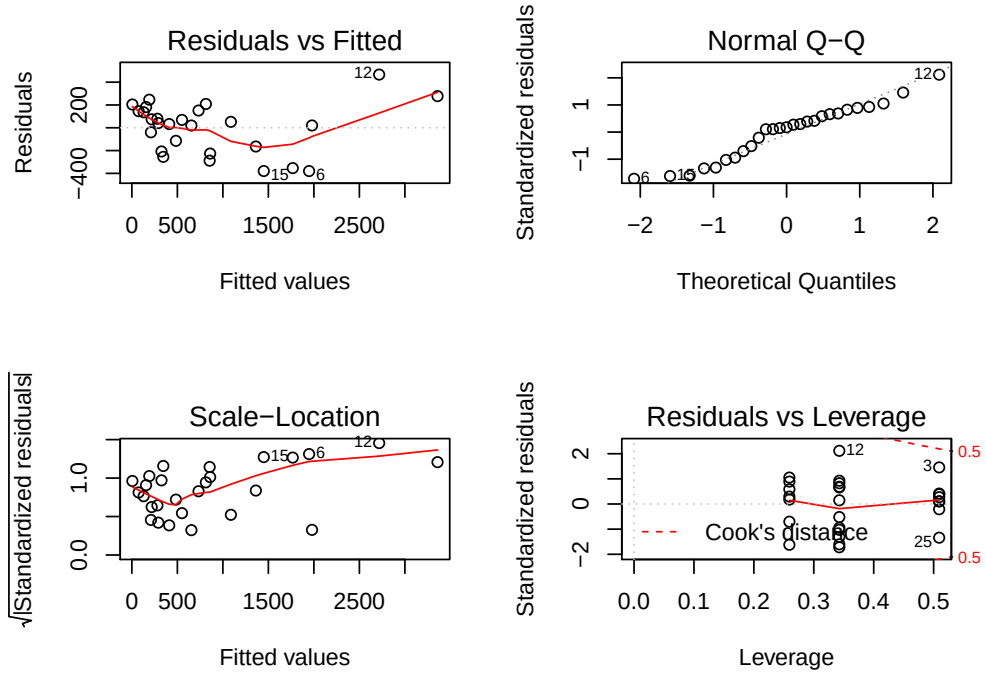

FIGURE 9 – résidus de (M2)

J'observe des non-linéarités sur les résidus (courbe rouge sur le premier graphe non constante égale à 0). Il serait donc peut être préférable de proposer un modèle non-linéaire.

4.3 Comparaison du modèle M2 et M1

Je souhaite effectuer un test de fisher qui permet de choisir entre l'hypothèse (M1) et (M2) La construction est la suivante :

1. **Modèle :** Je considère le modèle (M2).
2. **Hypothèses :** Je choisis hypothèse nulle $(M1)$:

$$H_0 = \{\theta_i = 0 \quad i \geq 4\}$$

Fâce à l'hypothèse alternative :

$$H_1 = \{\exists i \geq 4 \quad \theta_i \neq 0\}$$

3. **Statistique de test :** Je considère la statistique :

$$\boxed{F_n = \frac{|Y_{M1} - Y_{M2}|^2 (n-10)}{|Y - Y_{M2}|^2 (9-3)}}$$

Avec Y_{M2} la solution ajustée au modèle (M2) et Y_{M1} la solution ajustée au modèle (M1).

4. **Comportement sous H_0 :** Sous H_0, F_n suit une loi de Fisher de paramètre 6 et $n-10$.
5. **Comportement sous H_1 :** La statistique à tendance à prendre des valeurs plus élevées sous H_1.
6. **La région critique :** Comme sous H_1 la statistique prend des valeurs plus élevées, je vais choisir une région de rejet de la forme :

$$\boxed{\mathcal{R} = \{F_n > k_\alpha\}}$$

7. **Erreur de première espèce :** Je veux une érreur de première espèce inférieur à α. Soit, pour θ vérifiant H_0 :

$$\mathbb{P}_\theta(\mathcal{R}) \leq \alpha$$

Comme sous H_0 la statistique suit de Fischer à 6 et $n-10$ degrés de liberté, je peux choisir k_α comme étant le quantile de niveau $1-\alpha$ de la loi $\mathcal{F}(6, n-10)$.
J'aurais bien un test asymptotique de niveau α.

8. **Erreur de seconde espèce :**Cette erreur est difficile à quantifier.
9. **La p-valeur :** Si j'observe une valeur de la statistique F_n^{obs} la p-valeur sera définie comme le plus petit α tel que :

$$F_n^{\text{obs}} > k_\alpha$$

Donc :

$$k_{p_{\text{val}}} = F_n^{\text{obs}}$$

Si je pose F la fonction de répartition de la loi $\mathcal{F}(6, n-10)$:

$$\boxed{p_{\text{val}} = 1 - F(F_n^{\text{obs}})}$$

J'effectue ce test dans R :

```
> yM1=fitted.values(res1)
> yM2=fitted.values(res2)
> n=length(yM1)
> F= norm(as.matrix(yM1-yM2),"2")^2*(n-10)/(norm(as.matrix(y-yM2),"2")^2*6)
> print(F)

[1] 9.522668

> print(1-pf(F,6,n-10))

[1] 0.0001154311
```

La fonction *anova* peut donner directement le résultat cherché :

```
> print(anova(res1,res2))

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
```

Comme la p-valeur est faible, je peux refuser l'hypothèse H_0 avec une probabilité égale à la p-valeur de se tromper. Ainsi, si (M2) est vérifié les coefficients d'ordre 2 auront une probabilité très faible d'être tous nuls. Je vais donc préférer le modèle (M2) à (M1).

4.4 Transformation de Box-Cox

Je cherche un λ optimal de la transformation de Box-Cox qui ajuste au mieux le modèle (M1). Ce nouveau modèle avec transformation sera appelé (M1bis) :

$$h_\lambda(y) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \varepsilon$$

Où ε est un vecteur gaussien centré de variance $\sigma^2 I_n$.

J'applique donc la transformation *powerTransform* :

```
> PW1=powerTransform(y~.,data=df)
> print(summary(PW1))

bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1   -0.0592           0      -0.1789       0.0606

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                            LRT df    pval
LR test, lambda = (0) 0.9213384  1 0.33712

Likelihood ratio test that no transformation is needed
                           LRT df       pval
LR test, lambda = (1) 84.07566  1 < 2.22e-16
```

Le TRV qui teste l'hypothèse $\lambda = 0$ me donne une p-valeur supérieure à 5%. Je peux accepter cette hypothèse avec un risque inconnu et le modèle devient :

$$\log(y) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \varepsilon$$

Je peux donc ajuster le modèle :

```
> res1bis=lm(log(y)~.,data=df)
```

Je fais deux même avec le modèle (M2bis) (avec les termes d'ordre 2) :

```
> res2bis=lm(log(y)~x1+x2+x3+I(x1^2)+I(x2^2)+I(x3^2)+I(x1*x2)+I(x1*x3)+I(x2*x3),data=df)
```

4.5 Choix du modèle

Je désire choisir entre le modèle (M1bis),(M2bis) et (M2). Pour cela je vais d'abord comparer les R^2 :

```
> R2.2 =summary(res2)$r.squared
> R2.1bis =summary(res1bis)$r.squared
> R2.2bis =summary(res2bis)$r.squared
> print(c(R2.2,R2.1bis,R2.2bis))
```

```
[1] 0.9378729 0.9658443 0.9724857
```

Puis je compare les variances (notons qu'on a mis à l'échelle log la variance de (M2) pour pouvoir la comparer avec les autres modèles) :

```
> sig2.2 =summary(res2)$sigma^2
> sig2.1bis =summary(res1bis)$sigma^2
> sig2.2bis =summary(res2bis)$sigma^2
> print(c(mean(log(y+sig2.2)-log(y)),sig2.1bis,sig2.2bis))
```

```
[1] 4.88759295 0.03445656 0.03755315
```

Au final, il est claire que je vais rejeter (M2) car sa variance est bien plus élevée que celle des autres modèles et le modèle est bien moins ajusté (il y a en plus des non-linéarités dans le modèle). Par contre, il est difficile de choisir entre (M1bis) et (M2bis) car (M1bis) est légèrement moins ajusté mais présente, par contre, une variance plus faible que (M2bis).

Je vais préférer le modèle le plus simple et celui qui à la variance la plus faible. Je vais donc choisir (M1bis). Il reste plus qu'à vérifier les résidus pour voir s'il n'y a rien d'anormal.

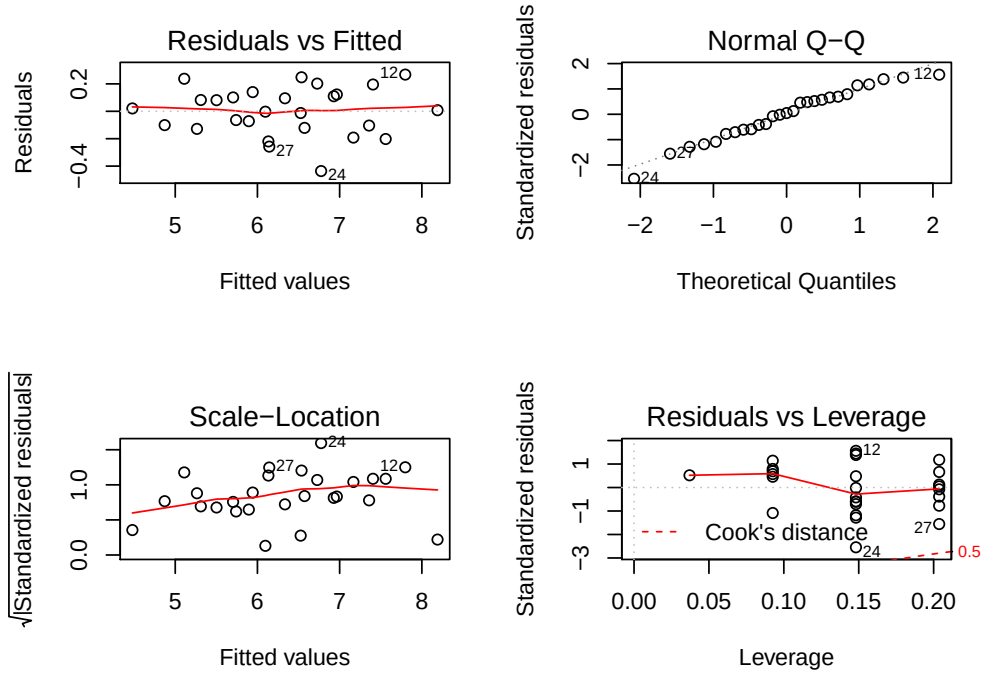

FIGURE 10 – résidus de (M1bis)

Je ne vois donc pas apparaître d'hétéroscédasticité (points répartis équitablement) ni de non linéarités (pas de motif). L'échantillon semble de plus gaussien (points qur la bissectrice dans Normal Q-Q). Il n'y pas non plus de valeurs influentes (distance de Cook faible).

Je peux donc conclure que les hypothèses sur le modèle (M1bis) semblent vérifiées.

Pour comparer M2bis et M1bis, nous pouvons aussi faire un test de Fisher qui compare M1bis et M2bis (car M1bis est inclus dans M2bis). Ce test revient à tester la significativité des régresseurs sui interviennent dans les termes d'ordre 2 pour M2bis. Donc sous l'hypothèse que le modèle M2bis est vérifié. Je vais tester l'hypothèse nulle : M1bis (les régresseurs d'ordre 2 sont nuls) fâce à l'hypothèse contraire :

```
> print(anova(res1bis,res2bis))


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
```

La p-value de ce test est supérieure à 0.05, je vais donc faire l'hypothèse que les termes d'ordre 2 ne sont pas significatifs dans la régression. L'erreur est inconnue. Cela valide donc mon choix du modèle M1bis par rapport à M2bis.

Je vais donc choisir le modèle (M1bis).

5 Conclusion

J'ai donc introduit la transformation de Box&Cox théoriquement et j'ai pu ensuite vérifier que en pratique cette transformation marchait bien. Sachant cela, j'ai pu appliquer cette transformation à mon jeu de donnés et j'ai pu avoir un meilleur modèle que le modèle linéaire.