Etude et utilisation de la transformation de Box-Cox appliquée au modèle linéaire

Yue Zhang et Gérémi Bridonneau

Introduction

On va étudier une transformations non linéaires et en particulier la transformation de Box-Cox régulièrement utilisé pour stabiliser la variance et corriger les asymétries des données. On va dans un premier temps étudier théoriquement cette transformation et comment obtenir les paramètres optimaux (dans ce travail on met en oruvre l'estimation maximum vraisemblance) de cette transformation, et puis étudier l'intervalle de confiance et tests statistiques sur l'estimateur. Dans un second temps on testera notre transformation sur des données simulées puis nous terminerons par un cas pratique sur l'étude du nombre de cycles à rupture d'un fil peigné en fonction de certains paramètres, en étudiant la regression modèle linéaire d'ordre 1, d'ordre 2 et les choix des variables.

1 La transformation de Box-Cox

On s'intéresse au modèle d'observation (x_i, Y_i):

$$h_\lambda(Y_i) = Z_i = x_i\theta + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

où (h_λ) est une famille de transformations paramétrées par λ.

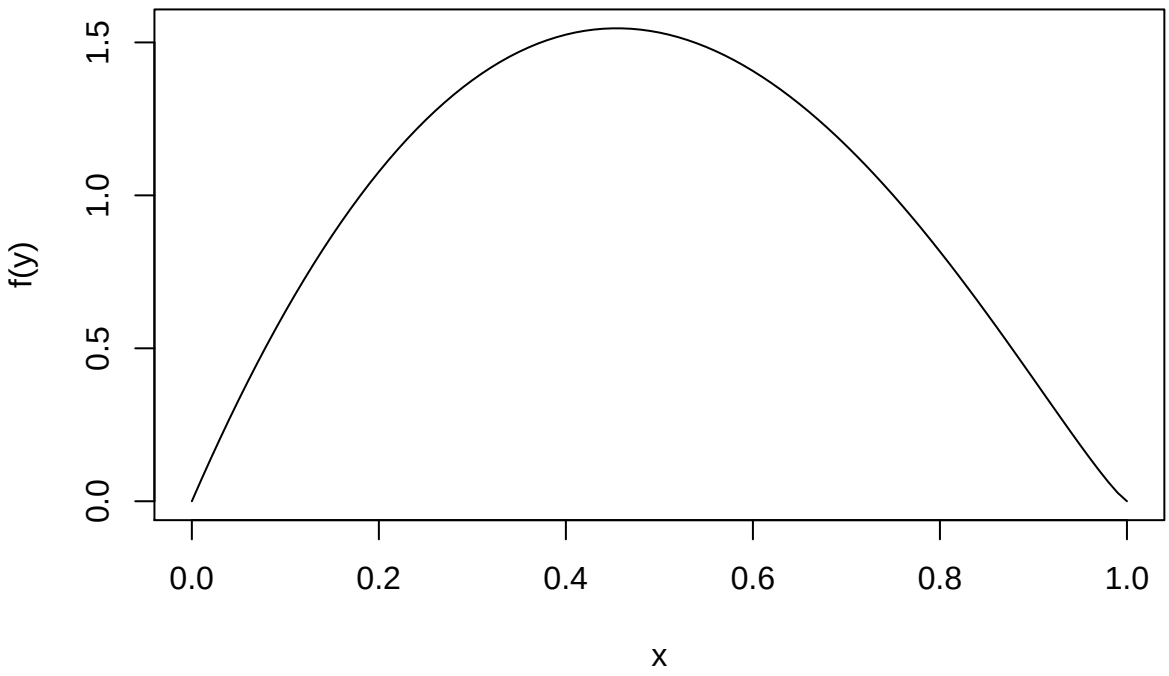
Etude de la tranformation de Box-Cox.

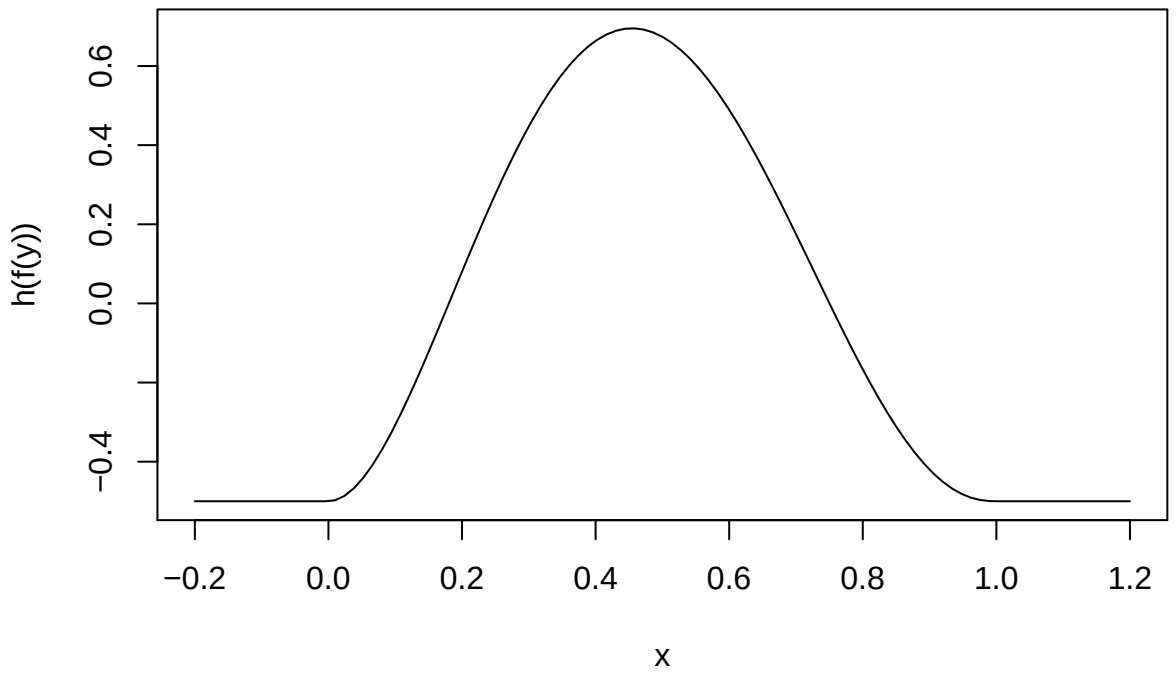
Ici on s'intéresse à la transformation de Box-Cox définie par

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

On remarque que théoriquement cette transformation est incompatible avec le modèle 1. En effet la transformation $\tilde{h}_\lambda$ est valable seulement pour $y > 0$. De plus pour tout $\lambda \neq 0$, la transformation $\tilde{h}_\lambda$ est borné et donc la transformation ne peut pas être gaussienne. Pour $\lambda = 0$ on n'a pas ce problème grâce à la surjectivité du logarithme.

Si toute les observations sont positives on peut quand même utiliser cette transformation car on perdra qu'une faible partie des données normalement dans la queue à gauche de la répartition. Par exemple si les données ne suivent pas une loi normale mais une loi beta de paramètre $\alpha = 2, \beta = 2.2$ et qu'on utilise la transformation de Box et Cox avec $\lambda = 2$ on obtient:


Loi beta de paramètres alpha=2, beta=2.2
f(y)
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
x



On voit qu'on a aucune valeur négative et que donc la gaussianisation n'est pas parfaite mais cette transformation reste raisonnable.

2. Déterminer la fonction de vraisemblance

Supposons que pour $\beta = (\theta, \lambda, \sigma^2)'$ a $p \times 1$ vecteur de paramètres, on ait $h_\lambda(Y_i) = Z_i = x_i\theta + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, ε_i suivent une loi gaussien i.i.d. Donc par la définition de vraisemblance:

$$
\begin{aligned}
L(\lambda, \theta, \sigma^2; Y) &= \prod_{i=1}^{n} \frac{\partial F}{\partial Y_i} \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(h_\lambda(Y_i) - x_i\theta)^2}{2\sigma^2}\right) \left|\frac{\partial h_\lambda(Y_i)}{\partial Y_i}\right| \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(h_\lambda(Y_i) - x_i\theta)^2}{2\sigma^2}\right) \prod_{i=1}^{n} \left|\frac{\partial h_\lambda(Y_i)}{\partial Y_i}\right| \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{(h_\lambda(Y_i) - x_i\theta)'(h_\lambda(Y_i) - x_i\theta)}{2\sigma^2}\right) \prod_{i=1}^{n} \left|Y_i^{\lambda-1}\right|
\end{aligned}
\tag{2}
$$

Donc le terme $J(\lambda; Y) = \prod_{i=1}^{n} \left|\frac{\partial h_\lambda(Y_i)}{\partial Y_i}\right| = \prod_{i=1}^{n} \left|Y_i^{\lambda-1}\right|$, est la transformation de Jacobian du terme $(h_\lambda(Y) - X\theta)$ à Y.

3. Estimation du maximum de vraisemblance

A λ fixé, on souhaite déterminer l'estimateur du maximum de vraisemblance $\widehat{\theta}(\lambda)$ et $\widehat{\sigma}^2(\lambda)$. Donc tout d'abord, depuis l'équation 2 on calcule la log-vraisemblance.

$$\begin{aligned}\ell &= \log L(\lambda, \theta, \sigma^2; Y) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{(h_\lambda(Y) - X\theta)'(h_\lambda(Y) - X\theta)}{2\sigma^2} + \sum_{i=1}^{n}\log\left|Y_i^{\lambda-1}\right| \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\|h_\lambda(Y) - X\theta\|^2}{2\sigma^2} + (\lambda-1)\sum_{i=1}^{n}\log|Y_i|\end{aligned} \tag{3}$$

Ensuite, étant donné que la log-vraisemblance ℓ l'équation (3) est une transformation monotone de la vraisemblance L dans l'équation (2), on maximise la log-vraisemblance ℓ respectivement pour θ, σ^2 et λ, donc on obtient le premier ordre dérivation ci-dessous:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{(h_\lambda(Y) - X\theta)'(h_\lambda(Y) - X\theta)}{2\sigma^4} = 0 \tag{4}$$

Donc on a $\widehat{\sigma}^2 = \dfrac{(h_\lambda(Y) - X\theta)'(h_\lambda(Y) - X\theta)}{n} = \dfrac{h_\lambda(Y)'(I_n - H)h_\lambda(Y)}{n}$, avec $H = X(X'X)^{-1}X'$ et I_n matrice identité.

$$\begin{aligned}\frac{\partial \ell}{\partial \theta} &= -\frac{2(-X)'(h_\lambda(Y) - X\theta)}{2\sigma^2} \\ &= \frac{X'(h_\lambda(Y) - X\theta)}{\sigma^2} = 0\end{aligned} \tag{5}$$

Donc, $\widehat{\theta} = (X'X)^{-1}X'h_\lambda(Y)$, par $X'h_\lambda(Y) = X'X\theta$.

Pour vérifier la formule avec $L_{max}(\lambda)$, on remplace nos emv $\widehat{\sigma}^2$ et $\widehat{\theta}$ calculés dans les équations (4) et (5) dans la log-vraisemblance ℓ:

$$\begin{aligned}L_{max}(\lambda) := \ell &= \log L(\lambda, \widehat{\theta}(\lambda), \widehat{\sigma}^2(\lambda)) \\ &= -\frac{n}{2}\log\left(\frac{\|h_\lambda(Y) - X\theta\|^2}{n}\right) - \frac{\|h_\lambda(Y) - X\theta\|^2 n}{2\|h_\lambda(Y) - X\theta\|^2} + (\lambda-1)\sum_{i=1}^{n}\log|Y_i| - \frac{n}{2}\log(2\pi) \\ &= -\frac{n}{2}\log(\widehat{\sigma}^2(\lambda)) + (\lambda-1)\sum_{i=1}^{n}\log|Y_i| - \frac{n}{2} - \frac{n}{2}\log(2\pi)\end{aligned} \tag{6}$$

Donc $a(n) = -\dfrac{n}{2} - \dfrac{n}{2}\log(2\pi)$ qui est bien une constante ne dépendant que de n. Maintenant on calcule l'emv $\widehat{\lambda}$:

$$\frac{\partial \ell}{\partial \lambda} = -\frac{2(h_\lambda(Y) - X\theta)\left|\frac{\partial h_\lambda(Y)}{\partial \lambda}\right|}{2\sigma^2} + \sum_{i=1}^{n}\log|Y_i| = 0 \tag{7}$$

Et

$$\begin{aligned}\frac{\partial L_{max}}{\partial \lambda} &= -\frac{n}{2}\frac{1}{\widehat{\sigma}^2(\lambda)}\left|\frac{\partial \widehat{\sigma}^2(\lambda)}{\partial \lambda}\right| + \sum_{i=1}^{n}\log|Y_i| \\ &= -\frac{n}{2}\frac{1}{\widehat{\sigma}^2(\lambda)}\frac{2(h_\lambda(Y)-X\theta)\left|\frac{\partial h_\lambda(Y)}{\partial \lambda}\right|}{n} + \sum_{i=1}^{n}\log|Y_i| \\ &= -\frac{(h_\lambda(Y)-X\theta)\left|\frac{\partial h_\lambda(Y)}{\partial \lambda}\right|}{\widehat{\sigma}^2} + \sum_{i=1}^{n}\log|Y_i| = 0\end{aligned} \tag{8}$$

On peut bien vérifier que $\frac{\partial \ell}{\partial \lambda}$ et $\frac{\partial L_{max}}{\partial \lambda}$ sont égaux par calcule de l'équation du maximum de vraisemblance. Par l'équation (4), on sait que $\widehat{\sigma}^2(\lambda) = \dfrac{h_\lambda(Y)^{'}(I_n - H)h_\lambda(Y)}{n} = \dfrac{SCR(\lambda)}{n}$ avec $H = X(X^{'}X)^{-1}X^{'}$, est la somme de carrés résiduels de variance $h_\lambda(Y^{)}$ divisée par n. Depuis l'équation (8), on peut continuer ce calcul en remplaçant $\widehat{\sigma}^2$ avec pour rappel $h_\lambda(Y) = \dfrac{Y^\lambda - 1}{\lambda}$:

$$\begin{aligned}\frac{\partial L_{max}}{\partial \lambda} &= -\frac{n}{2}\frac{n}{h_\lambda(Y)^{'}(I_n-H)h_\lambda(Y)}\frac{2h_\lambda(Y)^{'}(I_n-H)}{n}\left(\frac{Y^\lambda \log Y}{\lambda} - \frac{Y^\lambda - 1}{\lambda^2}\right) + \sum_{i=1}^{n}\log|Y_i| \\ &= -n\frac{h_\lambda(Y)^{'}(I_n-H)}{h_\lambda(Y)^{'}(I_n-H)h_\lambda(Y)}\left(\frac{Y^\lambda \log Y}{\lambda} - \frac{h_\lambda(Y)}{\lambda}\right) + \sum_{i=1}^{n}\log|Y_i| \\ &= -n\frac{h_\lambda(Y)^{'}(I_n-H)\lambda^{-1}Y^\lambda \log Y}{h_\lambda(Y)^{'}(I_n-H)h_\lambda(Y)} + n\frac{h_\lambda(Y)^{'}(I_n-H)h_\lambda(Y)}{h_\lambda(Y)^{'}(I_n-H)h_\lambda(Y)\lambda} + \sum_{i=1}^{n}\log|Y_i| \\ &= -n\frac{h_\lambda(Y)^{'}(I_n-H)u_\lambda(Y)}{h_\lambda(Y)^{'}(I_n-H)h_\lambda(Y)} + \frac{n}{\lambda} + \sum_{i=1}^{n}\log|Y_i|\end{aligned} \tag{9}$$

avec $u_\lambda(Y) = \lambda^{-1}Y^\lambda \log Y$. Le numérateur dans l'équation (9) est la somme résiduelle des produits dans l'analyse de la covariance de $h_\lambda(Y)$ et $u_\lambda(Y)$. Maintenant on utilise la transformation normalisée afin de simplifier le résultat, on définit $z_\lambda(Y)$ ci-dessous:

$$\begin{aligned}z_\lambda(Y) &= \frac{h_\lambda(Y)}{J(\lambda;Y)^{1/n}} \\ &= \frac{h_\lambda(Y)}{(\prod_{i=1}^{n}|Y_i|)^{\lambda-1/n}}\end{aligned} \tag{10}$$

Donc $\widehat{\sigma}^2$ devient $\widehat{\sigma}^2(\lambda;z) = \dfrac{z_\lambda(Y)^{'}(I_n-H)z_\lambda(Y)}{n} = \dfrac{SCR(\lambda;z)}{n}$, $SCR(\lambda;z)$ est la somme des carrées résiduelle de $z_\lambda(Y)$. De plus, $L_{max} = -\dfrac{n}{2}\log(\widehat{\sigma}^2(\lambda;z)) + a(n)$, donc on propose de trouver $\widehat{\lambda}$ qui maximize $L_{max}(\lambda)$, c'est à dire minimize $\widehat{\sigma}^2(\lambda;z)$. Donc on cherche l'emc (estimateur des moindres carrées)

$$\widehat{\lambda} = \arg\min_{\lambda} SCR(\lambda;z) \tag{11}$$

Par le théorème du cours, l'emv est asymptotiquement normal, donc la distribution de $\sqrt{n}(\widehat{\beta}-\beta)$, quand $n \to \infty$, elle converge en une loi normale.

$$\widehat{V}^{-1/2}\sqrt{n}(\widehat{\beta}-\beta) \to \mathcal{N}(0, I_{dp}) \tag{12}$$

$I_1(\beta)^{-1}$ est la matrice de l'information de Fisher, noté que $\widehat{V} = I_1(\beta)^{-1}$ et I_{dp} est la matrice identité de taille p. Quand $n \geq 30$, par le théorème TCL, $\widehat{\beta}$ tends vers un vecteur gaussien, donc à distance finie la distribution de $\widehat{\beta}$ approche la loi gaussienne.

4. Distribution asymptotique de l'emv

Estimer la variance de $\widehat{\lambda}$

Par la propriété de l'emv, quand $\widehat{\beta}$ tend à devenir gaussien, on peut prendre pour loi approchée à distance finie la loi asymptotique

$$\begin{aligned} \widehat{V}^{-1/2}(\widehat{\beta}-\beta) &\overset{appr}{\sim} \mathcal{N}(0, I_{dp}) \\ \widehat{\beta} &\overset{appr}{\sim} \mathcal{N}(\beta, I_1(\beta)^{-1}) \end{aligned} \tag{13}$$

Par la définition, la matrice de l'information de Fisher est écrite ci-dessous:

$$\begin{aligned} I_1(\beta) &= \mathbb{E}_\beta[\dot{\ell}\dot{\ell}'] \\ &= -\mathbb{E}_\beta[\ddot{\ell}] \end{aligned} \tag{14}$$

où $\ddot{\ell}$ est la matrice Hessienne $\ddot{\ell} = \dfrac{\partial^2 \ell}{\partial\beta\partial\beta'}$ (pour rappelle on a définit $\beta = (\theta, \lambda, \sigma^2)'$ a $p \times 1$ vecteur de paramètres). En particulier, on n'a pas forcément besoin d'estimer σ^2 simultanément avec θ et λ, donc pour simplifier les calculs, on décide de calculer la matrice Hessienne de $L_{max}(\lambda)$. Dans l'équation (8), on a calculé $\dfrac{\partial L_{max}}{\partial\lambda} = -\dfrac{(h_\lambda(Y) - X\theta)\left|\dfrac{\partial h_\lambda(Y)}{\partial\lambda}\right|}{\widehat{\sigma}^2} + \sum_{i=1}^n \log|Y_i|$, et on obtient sans souci $\dfrac{\partial L_{max}}{\partial\theta} = \dfrac{X'(h_\lambda(Y) - X\theta)}{\widehat{\sigma}^2}$.

$$\begin{aligned} \ddot{\ell} := H(\beta) &= \frac{\partial^2 L_{max}}{\partial\beta\partial\beta'} \\ &= -\widehat{\sigma}^{-2}\begin{bmatrix} X'X & -X'\left|\dfrac{\partial h_\lambda(Y)}{\partial\lambda}\right| \\ -\left|\dfrac{\partial h_\lambda(Y)}{\partial\lambda}\right| X & \left|\dfrac{\partial h_\lambda(Y)}{\partial\lambda}\right|'\left|\dfrac{\partial h_\lambda(Y)}{\partial\lambda}\right| + \left|\dfrac{\partial^2 h_\lambda(Y)}{\partial^2\lambda}\right|(h_\lambda(Y) - X\theta) \end{bmatrix} \end{aligned} \tag{15}$$

$H(\beta)$ est bien une matrice définie négative. Etant donnée la distribution asymptotique normale de l'emv, on peut conclure que $\widehat{Var(\widehat{\beta})} = -[H(\widehat{\beta})]^{-1}$, maintenant on calcul $\widehat{Var(\widehat{\lambda})}$:

$$\widehat{Var(\widehat{\lambda})} = -H(\widehat{\lambda})^{-1} \tag{16}$$

où $H(\lambda) = \dfrac{\partial^2 L_{max}}{\partial^2\lambda} = \dfrac{\left|\dfrac{\partial h_\lambda(Y)}{\partial\lambda}\right|'\left|\dfrac{\partial h_\lambda(Y)}{\partial\lambda}\right| + \left|\dfrac{\partial^2 h_\lambda(Y)}{\partial^2\lambda}\right| h_\lambda(Y)'(I_n - H)}{-\widehat{\sigma}^2}$.

Intervalle de confiance

L'emv est asymptotiquement normalement distribué, par la propriété dans l'équation (13), on peut construire le test $T = \dfrac{\widehat{\beta} - \beta}{\sqrt{\widehat{Var(\widehat{\beta})}}} \sim \mathcal{N}(0, I_{dp})$. Par définition, $P(q_{\alpha/2} < \dfrac{\widehat{\beta} - \beta}{\sqrt{\widehat{Var(\widehat{\beta})}}} < q_{1-\alpha/2}) = 1 - \alpha$, donc on peut obtenir l'intervalle de confiance $[\widehat{\beta} - q_{1-\alpha/2}\sqrt{\widehat{Var(\widehat{\beta})}}, \widehat{\beta} - q_{\alpha/2}\sqrt{\widehat{Var(\widehat{\beta})}}]$, où $q_{\alpha/2}$ et $q_{1-\alpha/2}$ sont quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ sous la loi normale $\mathcal{N}(0,1)$. La distribution est symétrique par rapport à 0, donc l'IC estimateur de β est également $[\widehat{\beta} - q_{1-\alpha/2}\sqrt{\widehat{Var(\widehat{\beta})}}, \widehat{\beta} + q_{1-\alpha/2}\sqrt{\widehat{Var(\widehat{\beta})}}]$.

L'intervalle de confiance de λ est donc $[\widehat{\lambda} - q_{1-\alpha/2}\sqrt{\widehat{Var(\widehat{\lambda})}}, \widehat{\lambda} + q_{1-\alpha/2}\sqrt{\widehat{Var(\widehat{\lambda})}}]$, où $\widehat{Var(\widehat{\lambda})}$ est calculé dans l'équation (16).

Test de Wald

On définit $A = [0, 1, 0]$, $\beta = (\theta, \lambda, \sigma^2)'$, $\beta_0 = (\theta_0, \lambda_0, \sigma_0^2)'$

$H_0 : A(\beta - \beta_0) = 0$ (i.e. $\lambda - \lambda_0 = 0$), contre $H_1 : A(\beta - \beta_0) \neq 0$ (i.e. $\lambda - \lambda_0 \neq 0$)

Sous H_0:

$$T = [AVA']^{-1/2} A(\widehat{\beta} - \beta_0) \to \mathcal{N}(0, I_{dp}) \tag{17}$$

En utilisant la propriété de la statistique de Wald, W est la carré de la norme de T et sa loi asymptotique sous H_0 est:

$$\begin{aligned} W &= (A\widehat{\beta} - A\beta_0)(A\widehat{V}A')^{-1}(A\widehat{\beta} - A\beta_0)' \\ &= \frac{(\widehat{\lambda} - \lambda_0)^2}{\widehat{Var(\widehat{\lambda})}} \to \chi^2(1) \end{aligned} \tag{18}$$

où $\widehat{V} = I_1(\widehat{\beta})^{-1}$, et $W \geq 0$, la région de rejet est unilatère à droite de niveau asymptotique α pour une hypothèse bilatère est $\mathcal{R} = \left\{ W > q_{1-\alpha}^{\chi^2(1)} \right\}$ avec $P_{(H_0)}(\mathcal{R}) \to \alpha$.

5. Test du rapport vraisemblance

Par le théorème asymptotique du RV, sous H_0:

$$TRV = -2\log(RV) \to \chi^2(1) \tag{19}$$

$TRV \geq 0$, la région de rejet $\mathcal{R} = \left\{ TRV > q_{1-\alpha}^{\chi^2(1)} \right\}$ du test de rapport de vraisemblances maximales est asymptotiquement de niveau α, $P_{(H_0)}(\mathcal{R}) \to \alpha$.

Par la définition de rapport de vraisemblance:

$$RV = \frac{L(\lambda_0; Y)}{L(\widehat{\lambda}; Y)} \tag{20}$$

où L est la fonction de vraisemblance.

$$\begin{aligned} TRV &= -2\log(\frac{L(\lambda_0;Y)}{L(\widehat{\lambda};Y)}) \\ &= -2(\log L(\lambda_0;Y) - \log L(\widehat{\lambda};Y)) \\ &= 2(L_{max}(\widehat{\lambda};Y) - L_{max}(\lambda_0;Y)) \\ &= 2(-\frac{n}{2}\log(\widehat{\sigma}^2(\widehat{\lambda})) + \frac{n}{2}\log(\widehat{\sigma}^2(\lambda_0))) \\ &= n\log(\frac{\widehat{\sigma}^2(\lambda_0)}{\widehat{\sigma}^2(\widehat{\lambda})}) \end{aligned} \tag{21}$$

2 Test de la méthode sur des données simulées

1. Modélisation la regression linéaire simple

Condition convergence

La condition de convergence indiquée dans la section 1 est bien vérifiée.

En effet on a que si $x_1, x_2, \ldots \overset{i.i.d}{\sim} \mathcal{N}(0,1)$,

$$\frac{X'X}{n} = \frac{1}{n}\begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{bmatrix}\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{n}\sum_{i=1}^n x_i \\ \frac{1}{n}\sum_{i=1}^n x_i & \frac{1}{n}\sum_{i=1}^n {x_i}^2 \end{bmatrix} \tag{22}$$

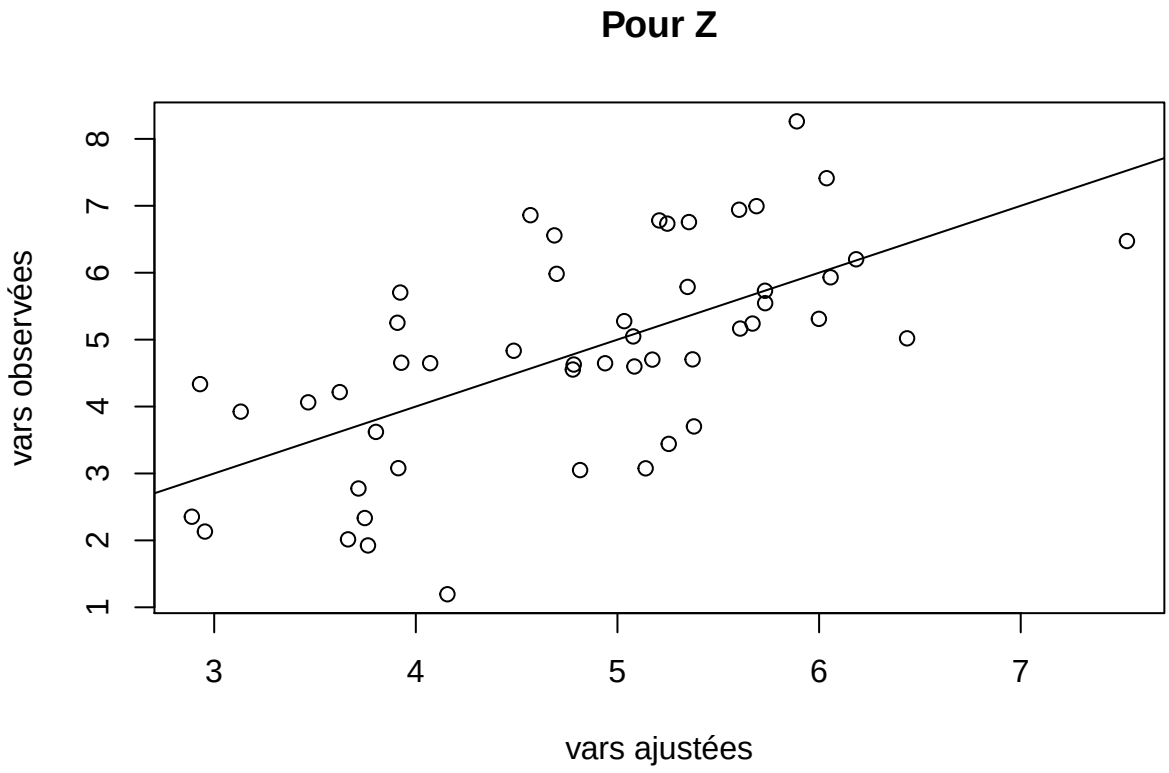
On a de plus $\mathbb{E}[{x_i}^2] = \mathrm{Var}(x_i) + \mathbb{E}[x_i]^2 = 1$. Ainsi d'après la loi des grands nombres on a $(X'X)/n$ qui converge vers la matrice identité de taille 2×2 qui est bien définie positive.

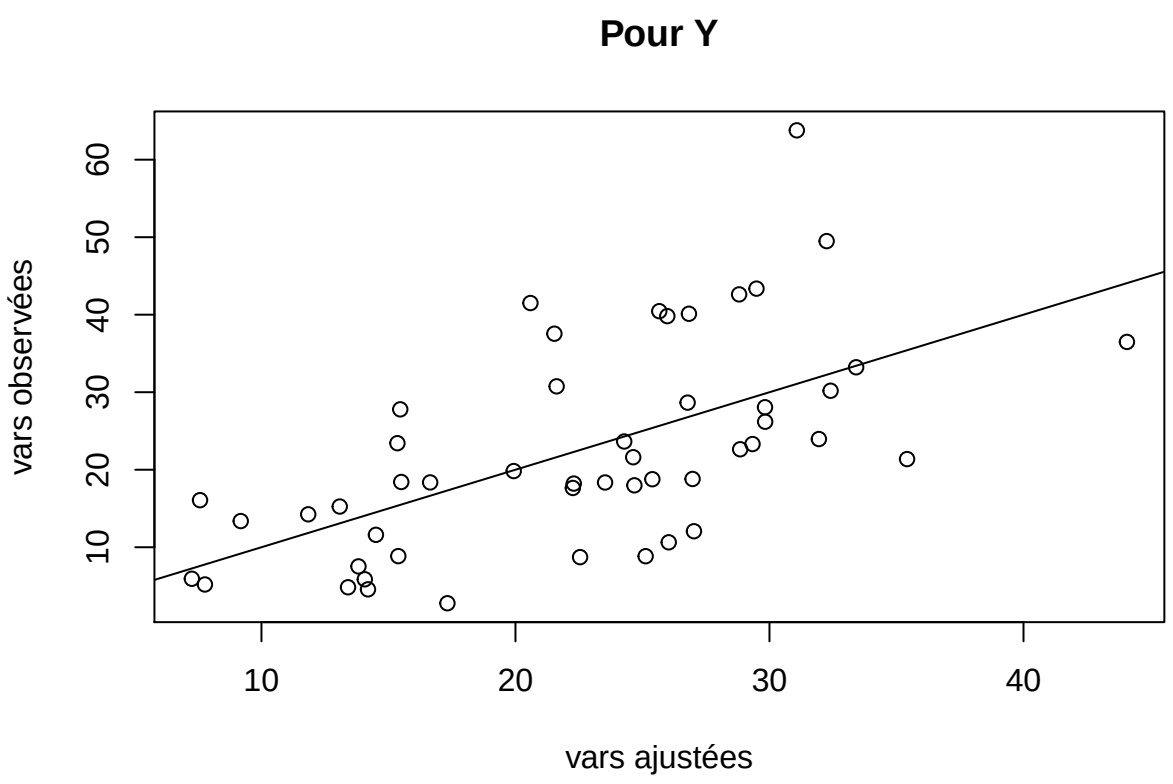
Estimation de la regression linéaire simple

Le modèle linéaire simple donc est $z_i = \mu + \theta x_i + \sigma\varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0,1)$ i.i.d. Le plan d'expérience X est en taille $[n,p]$ i.e. $[50 \times 2]$, et $\widehat{\theta} = (X'X)^{-1}X'Z$, $\widehat{\sigma}^2 = \dfrac{1}{n-p}||Z - X\widehat{\theta}||^2$.

Par la définition, $\dfrac{y_i^\lambda - 1}{\lambda} = h_\lambda(y_i) = z_i$, $y_i = (\lambda z_i + 1)^{1/\lambda}$.

En Traçant la profondeur ajustée en fonction de la profondeur observée, on peut observer que le modèle n'est pas bien ajustée, les points s'allongent autour de la première bissectrice, mais les ajustements sont parfois assez éloigné des valeurs observées, Y plus grave que Z visuellement, donc ce qui implique aussi ses bruts résidus fortes et potentiellement celle de Y est plus forte que celle de Z.

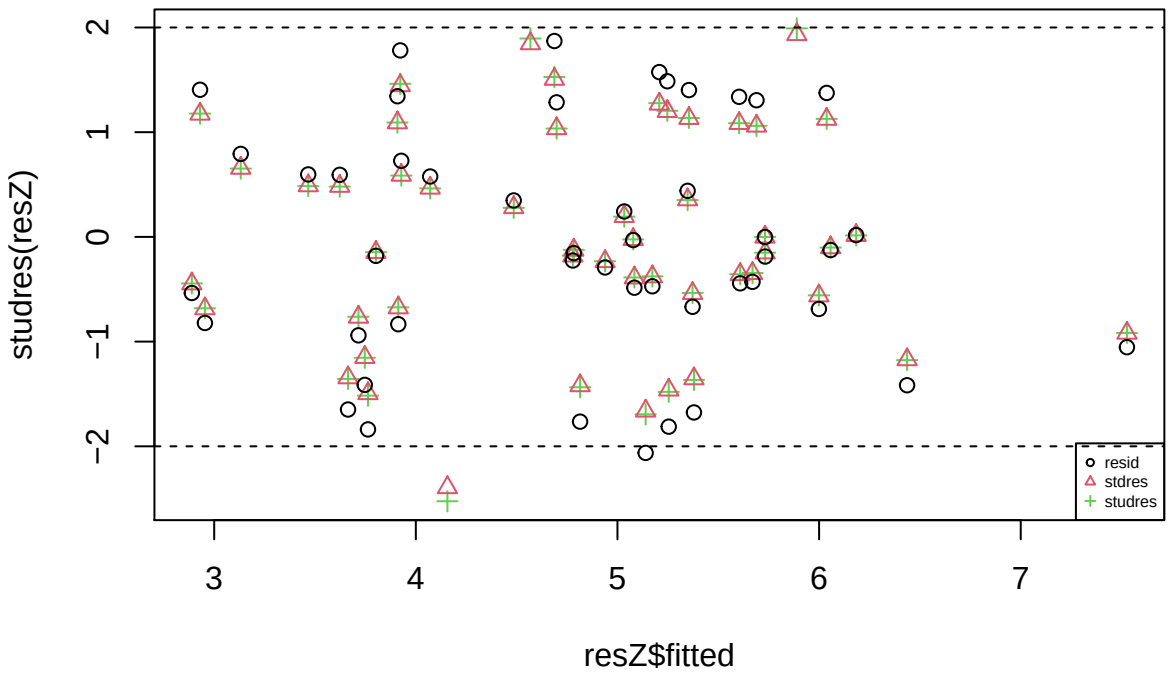

Pour Z
vars observées
vars ajustées
1
2
3
4
5
6
7
8
3
4
5
6
7



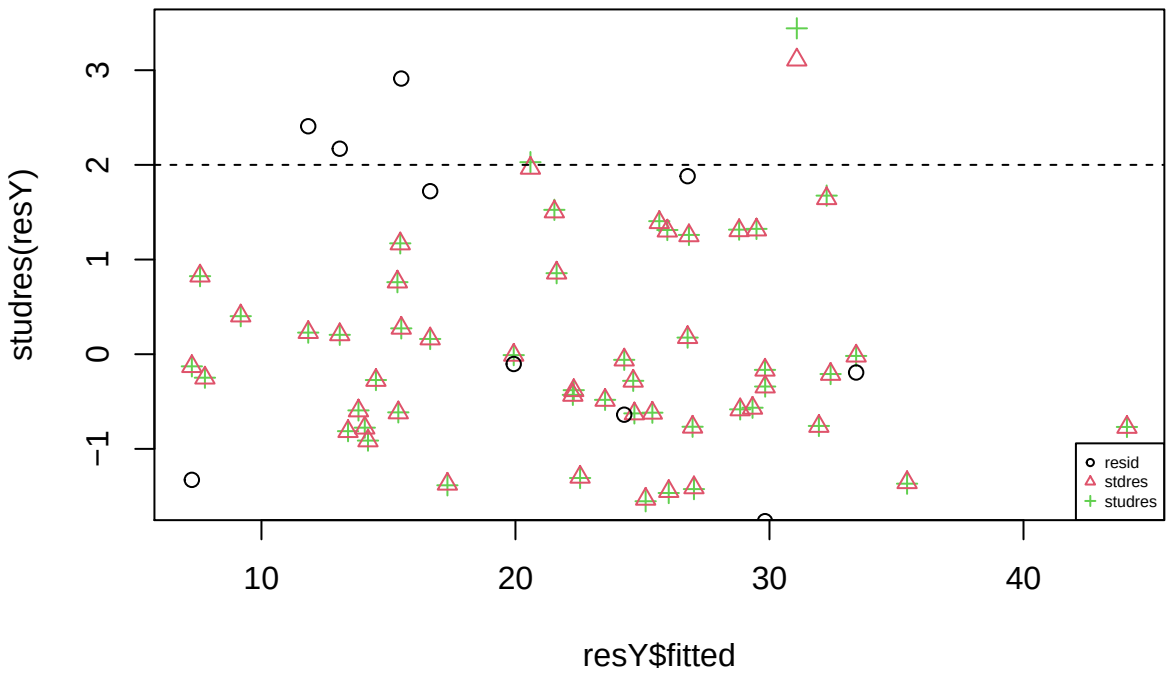
Etude des résidus

Par la définition, les résidus $\widehat{\varepsilon} = Z - X\widehat{\theta}$. Et les résidus studentisés donc $t_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1-h_{i,i}}}$ où $h_{i,i}$ sont les éléments diagonaux de $H = X(X^{'}X)^{-1}X^{'}$.

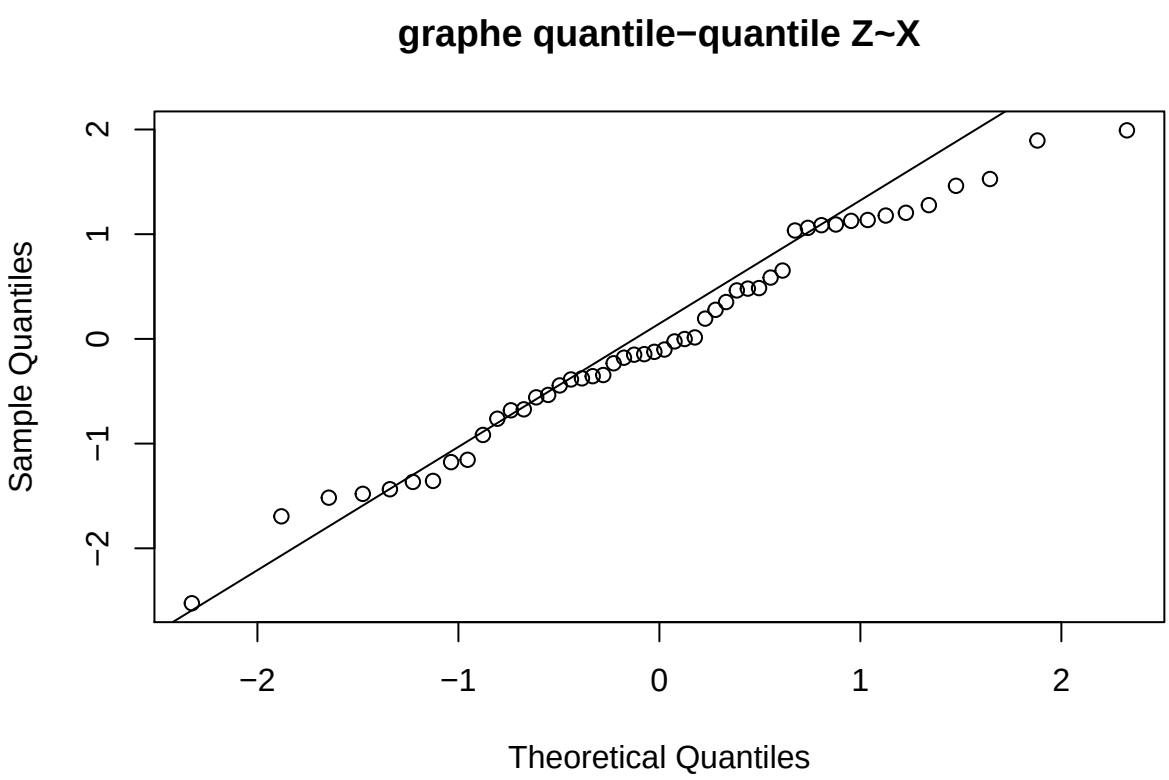
Pour $Z \sim X$, on voit que les résidus bruts sont forts parce que les valeurs observées sont elles-mêmes assez éloignés de 0 et estimées avec une faible précision. La plupart d'entre eux sont raisonnablement compris entre -2 et 2 sauf un seul point, en respectant la règle empirique d'appartenance de 95% des résidus à l'intervalle -2 et 2.

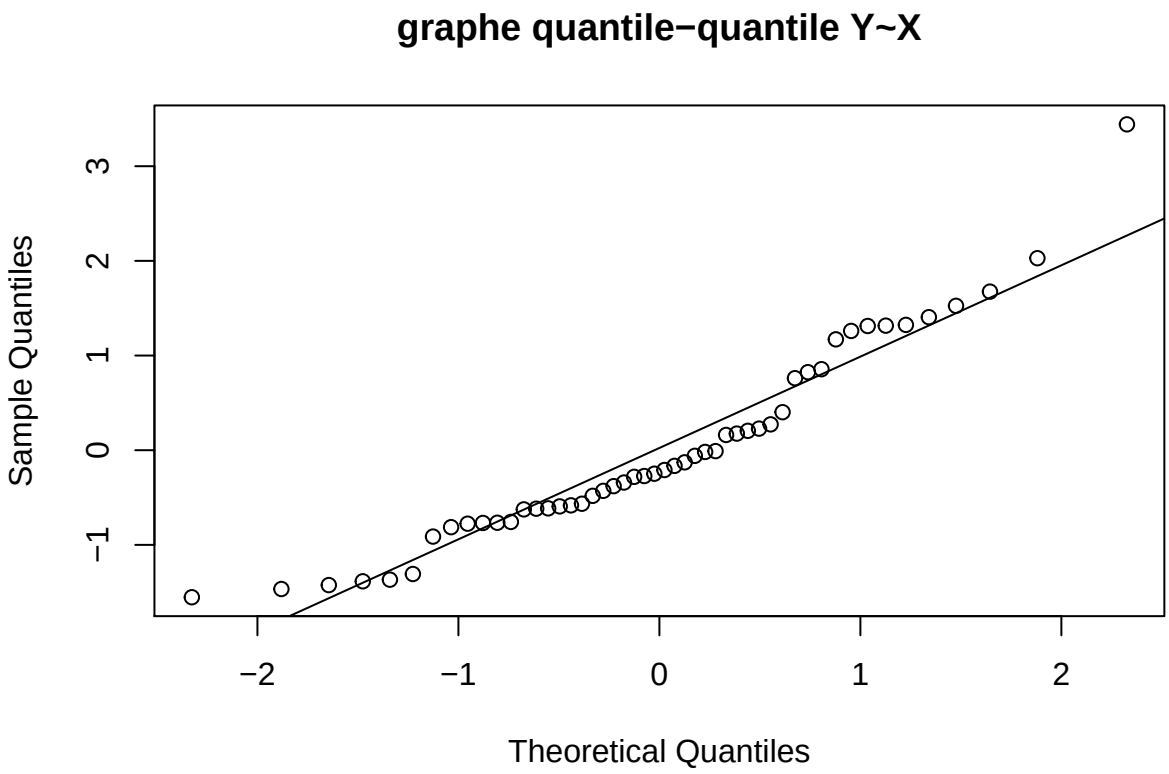


Pour $Y \sim X$, on voit que les résidus bruts sont plus fortes que celles de Z, parce que les valeurs observées sont elles-mêmes beaucoup plus éloignés de o et estimées avec une très faible précision. La plupart d'entre eux sont raisonnablement compris entre -2 et 2 sauf trois points, en respectant aussi la règle empirique d'appartenance de 95% des résidus à l'intervalle -2 et 2.



Pour $Z \sim X$ les points s'alignent, mais pas pour les points qui se trouvent près des deux extrémité, donc on doute de la gaussiannité des résidus. On même l'observation pour $Y \sim X$.


graphe quantile−quantile Z~X
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2



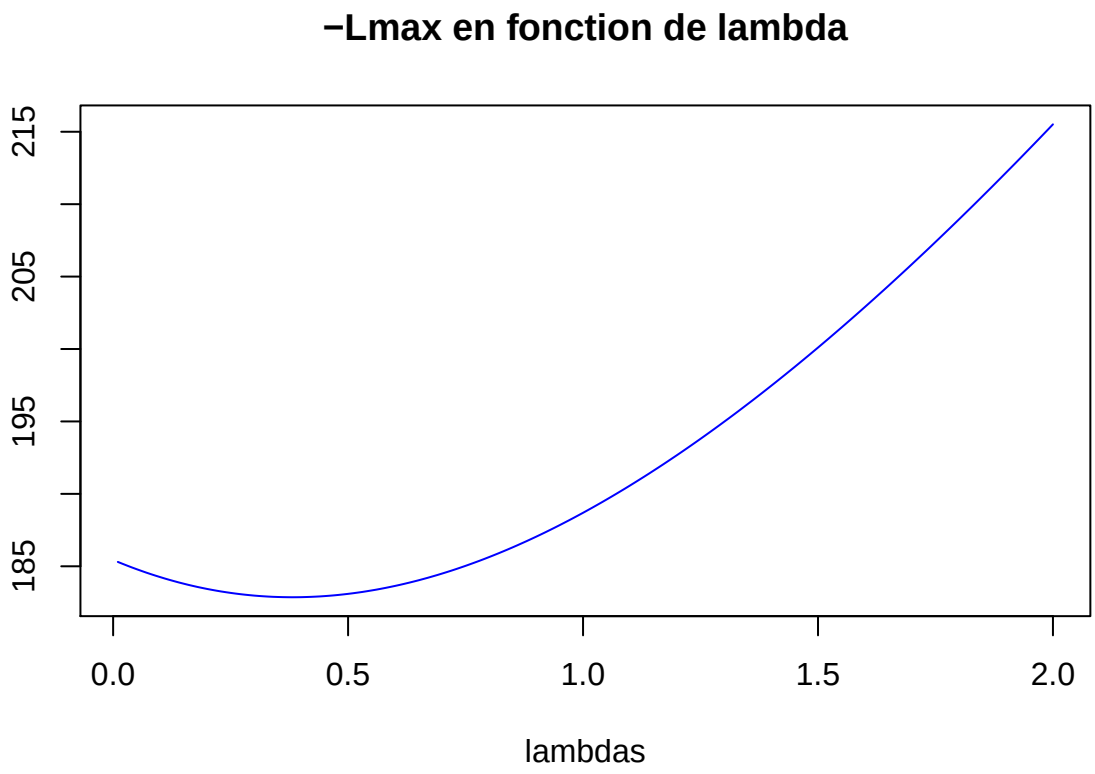
Donc on s'intéresse à faire le test Shapiro pour la normalité. Pour $Z \sim X$, la p-valeur est égale à $0.6019 > \alpha = 0.05$, donc on garder l'hypothèse de gaussianité avec une risque de seconde espèce inconnue. Pour $Y \sim X$, p-valeur égale à $0.007836 < \alpha = 0.05$, donc on rejette l'hypothèse de gaussianité avec une risque de première espèce $\alpha = 0.05$.

```
##
##  Shapiro-Wilk normality test
##
## data:  studres(resZ)
## W = 0.98116, p-value = 0.6019
```

```
##
##  Shapiro-Wilk normality test
##
## data:  studres(resY)
## W = 0.93397, p-value = 0.007836
```

2. Mise en oeuvre le calcul $\widehat{\lambda}$

La variable Q est $(I_n - H)$, où I_n est la matrice identité de taille 50×50, $H = X(X'X)^{-1}X'$. Le variable $sig2$ égale à $\dfrac{h_\lambda(Y)'(I_n - H)h_\lambda(Y)}{n}$ qui est exactement $\widehat{\sigma}^2$ où on a démontré dans l'équation (4). La fonction *Lmle* retourne le terme $-\dfrac{n}{2}\log(\widehat{\sigma}^2)$.



On voit bien que la fonction $-L_{max}$ est une fonction quadratique convexe, à environ $\lambda = 0.4$ $-L_{max}$ atteint le minimum.

3. Calcul $\widehat{\lambda}$ et $\widehat{Var(\widehat{\lambda})}$

Etant donné que $Z = h_\lambda(Y)$ est la transformation de Y, par la définition donc $\lambda \neq 0$, pour la Méthodes Newton, on commence l'itération à partir de 2. En utilisant la fonction optimisation nlm, on obtient la valeur estimée $\widehat{\lambda}$ est $resopt\$estimate = 0.3817253$. Comme démontré dans l'équation (16), la variance de $\widehat{\lambda}$ est l'inverse de la hessienne donc est 0.02948455. (Pour la minimisation $-L_{max}$, la matrice hessienne pour λ est définie positive 33.91607)

4. Tests hypothèses

Intervale de confiance

Comme on a montré dans le premier section, l'intervalle de confiance pour λ est $[\widehat{\lambda} - q_{1-\alpha/2}\sqrt{\widehat{Var(\widehat{\lambda})}}, \widehat{\lambda} + q_{1-\alpha/2}\sqrt{\widehat{Var(\widehat{\lambda})}}]$ au niveau asymptotiquement $1-\alpha$. Ici on fixe α à 0.05, donc $q_{1-\alpha/2}$ est 1.959964 la quantile d'ordre $1-\alpha/2$ sous loi normale, donc on obtient l'IC $[0.04517863, 0.718272]$.

Test de Wald

Das l'équation (18), on a montré $W = \dfrac{(\widehat{\lambda} - \lambda_0)^2}{\widehat{Var(\widehat{\lambda})}} \rightarrow \chi^2(1)$. Donc on veut tester $H_0 : \lambda = \lambda_0$, contre $H_1 : \lambda \neq \lambda_0$ par la statistique de Wald pour les quatre cas ci-dessous:

1) Par la définition de $h_\lambda(Y)$, quand $\lambda = 1$, les données Y ne nécessitent pas de transformation, donc dans le test $\lambda_0 = 1$. On a $W_{obs} > q^{\chi^2}_{1-\alpha}$ qui se trouve dans la région de rejet (unilatère à droite), et p-valeur $0.0003173887 < \alpha$, donc rejette H_0 et accepte H_1 avec risque d'erreur de première espèce α.

2) Pour la même raison, quand $\lambda = 0.5$, la transformation à appliquer aux observations est en racine carrée. Donc dans le test on veut $\lambda_0 = 0.5$. On a $W_{obs} < q^{\chi^2}_{1-\alpha}$ qui ne se trouve pas dans la région de rejet (unilatère à droite), et p-valeur $0.4909476657 > \alpha$, donc on rejette H_1 et accepte H_0 avec risque d'erreur seconde espèce inconnue.

3) Quand $\lambda_0 = 0.3$, on a $W_{obs} < q^{\chi^2}_{1-\alpha}$ qui ne se trouve pas dans la région de rejet, et p-valeur $0.6341115306 > \alpha$, donc on rejette H_1 et accepte H_0 avec risque de seconde espèce inconnue.

4) Quand $\lambda_0 = 0$, on a $W_{obs} > q^{\chi^2}_{1-\alpha}$ qui se trouve dans la région de rejet (unilatère à droite), et p-valeur $0.0262112618 < \alpha$, donc on rejette H_0 et accepte H_1 avec risque d'erreur premier espèce $\alpha = 5\%$. (Note que dans $h_\lambda(Y)$, $\lambda \neq 0$, donc on test 0.000001 ici)

5. Test de rapport de vraisemblance

```
## [1] "lambda= 1"
## [1] "TRV:  11.6612671671051"
## [1] "p_value:  0.000638148590238264"
## [1] "conserve H_0? FALSE"
## [1] "lambda= 0.5"
## [1] "TRV:  0.466380664297333"
## [1] "p_value:  0.494656961393314"
## [1] "conserve H_0? TRUE"
## [1] "lambda= 0.3"
## [1] "TRV:  0.228985473833632"
## [1] "p_value:  0.632277106764555"
## [1] "conserve H_0? TRUE"
## [1] "lambda= 1e-06"
## [1] "TRV:  5.1484589536189"
## [1] "p_value:  0.0232670132735007"
## [1] "conserve H_0? FALSE"
```

Par l'équation (21), TRV suit une loi $\chi^2(1)$, donc de même façon, on calcule les TRV observés et les p-valeurs. En comparant les p-valeurs et α, les conclusions obtenues sont les mêmes que celles données par le test Wald dans la question précédente.

6. Vérification par fonction "powerTransform"

Etant donné que l'on a toujours des problèmes sur l'installation du package "car", on l'exécute sur compilateur en ligne et voici dessous les résultats obtenus.

On peut voir que l'estimateur de la puissance transformation appliquée sur Y est la même valeur que l'on a obtenu par maximisation de le vraisemblance L_{max} calculé par nlm dans la question 3. Puis l'intervalle de confiance $[0.0452, 0.7183]$ est identique à celui que l'on calcule dans question 4 au niveau asymptotique $\alpha = 0.05$. Comme on a montré dans question 5, $TRV \sim \chi^2(1)$, pour $\lambda = 1$ et $\lambda = 0$, notre TRV observés et les p-valeurs sont les mêmes que les résultats donnés par fonction $powerTransform$.

```
res <- powerTransform(Y~X, family="bcPower")

summary(res)
```

Run (Ctrl-Enter)

Any scripts or data that you put into this service are public.

```
Loading required package: carData
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
```

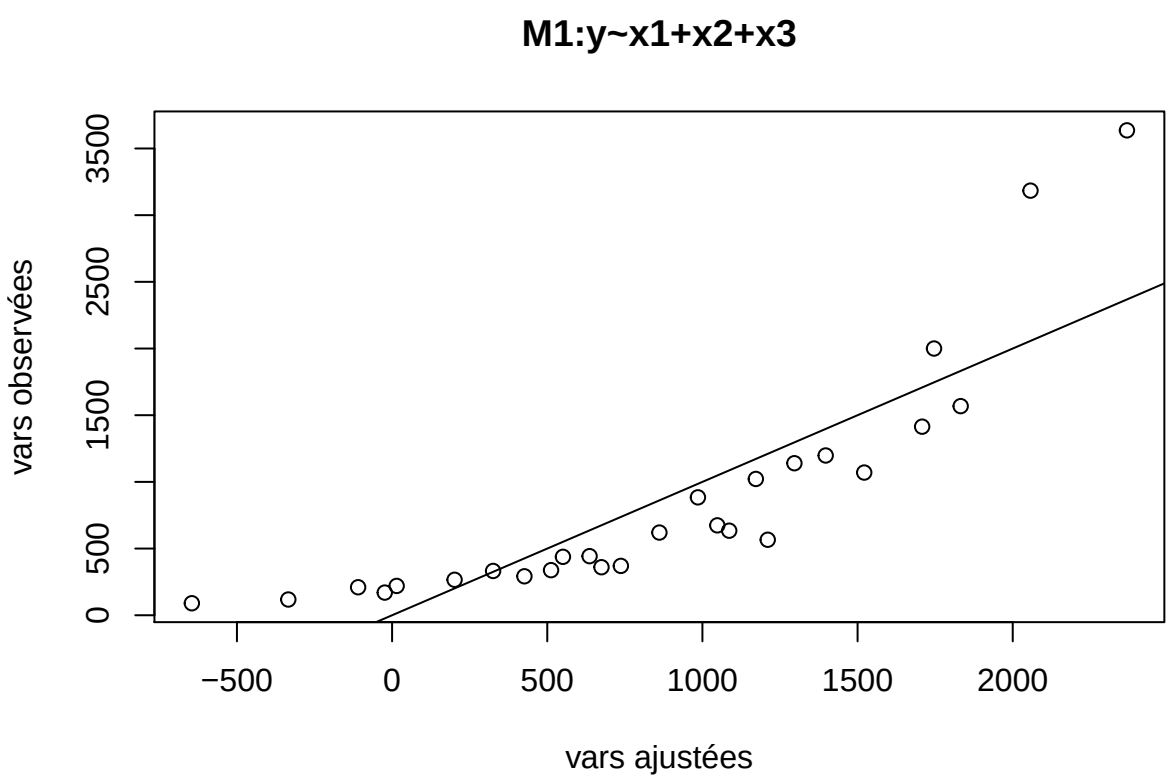
FIGURE 1 – powerTransform

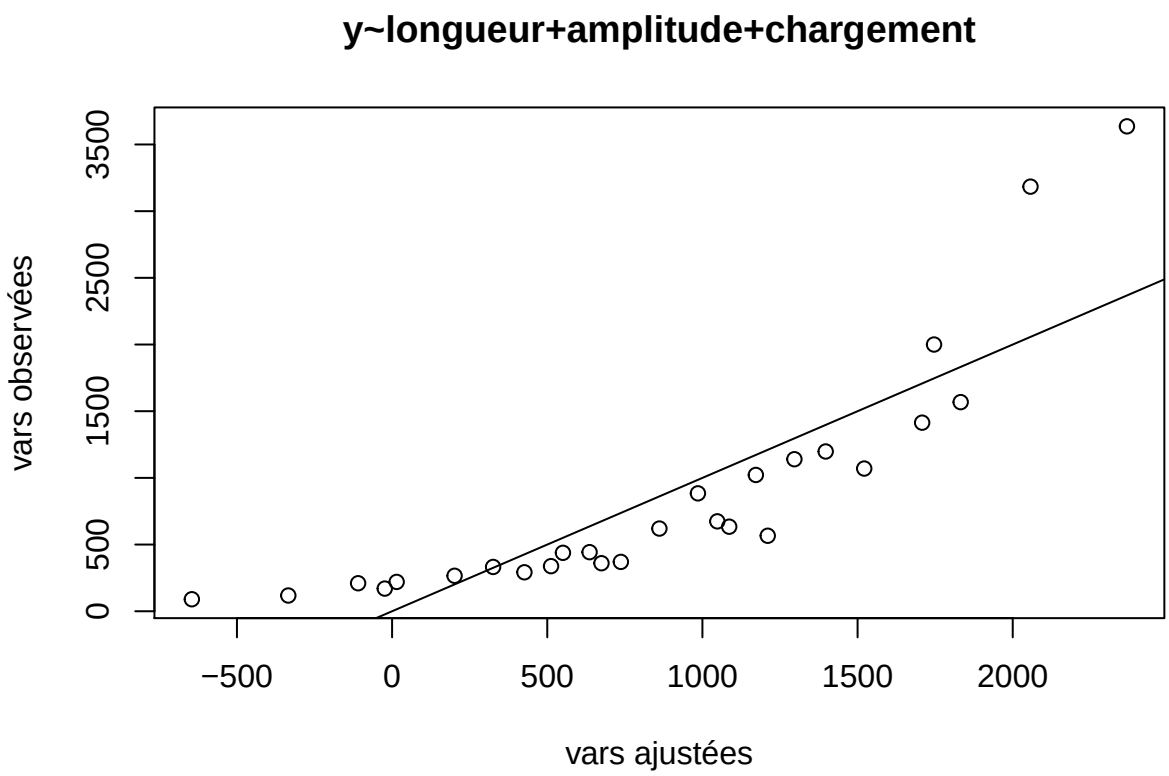
3 Cas pratique

1. Modèle regression linéaire multiple

Analyse regression LM multiple

On définit le modèle linéaire multiple $(M1)$ $y = \mu + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \sigma\varepsilon$, $\varepsilon \sim \mathcal{N}(0, I_n)$ i.i.d, on note que $\theta = [\mu, \beta_1, \beta_2, \beta_3]$.


M1:y~x1+x2+x3
vars observées
0
500
1500
2500
3500
−500
0
500
1000
1500
2000
vars ajustées



D'abord, on voit que le modèle pour les variables transformées $M1 : y = \mu + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \sigma\varepsilon$ n'est pas bien ajusté, comme le montre visuellement la figure ci-dessus: certains des points s'alignent prêsque autour de la première bissectrice, mais les premieres données et les dernieres données s'éloignent assez loin que la droite $y = x$, pour la plupart des données, les ajustements sont un peu plus proches des valeurs observées mais avec certain des bruits assez évidents.

Et puis par la définition dans l'énoncé, on obtient les valeurs pour les variables non transformées (longueur, amplitude et chargement). De même façon, on applique aussi modèle linéaire multiple, mais les deux ajustements sont identiques. Bien sûre que les estimateurs de coefficient ($\widehat{\theta}$) changent, parce que les variables explicatives (i.e. la nouvelle matrice expérience $X = [1, \text{longueur, amplitude , chargement}]$) changent. Mais les valeurs ajustées ne changent pas, donc la performance de modèle ne change pas non plus. Comme on peut vérifier par les données ci-dessous, les estimateurs changent mais le R^2 et la F statistique ne changent pas.

```
##
## 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
```

```
##
## Call:
## lm(formula = df$y ~ longueur + amplitude + chargement)
##
## 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 *
## longueur      13.200      2.301   5.736 7.66e-06 ***
## amplitude   -535.833    115.057  -4.657 0.000109 ***
## chargement   -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
```

A partir d'ici, on analyse que les valeurs pour $y \sim x_1 + x_2 + x_3$. Parce que pour $y \sim longueur + amplitude + chargement$ c'est la même méthode et les conclusions sont les mêmes que pour $y \sim x_1 + x_2 + x_3$.

Analyse la significativité des variables

Ici on veut tester chaque composante θ_i dans θ. $H_0 : \theta_i = 0$ contre $H_1 : \theta_i \neq 0$, on suppose que $\alpha = 0.05$.

1) μ: p-valeur $3.83e-09 < \alpha$, donc on rejette H_0 et accepte H_1, c'est-à-dire on décide que le coefficient μ est non nul avec risque de première espèce $\alpha = 0.05$, et l'intercepte est utile dans $M1$ donc elle est significative.
2) β_1: p-valeur $7.66e-06 < \alpha$, donc on rejette H_0 et accepte H_1, c'est-à-dire on décide que le coefficient β_1 est non nul avec risque de première espèce $\alpha = 0.05$, et x_1 est utile dans $M1$ donc elle est significative.
3) β_2: p-valeur $0.000109 < \alpha$, donc on rejette H_0 et accepte H_1, c'est-à-dire on décide que le coefficient β_2 est non nul avec risque de première espèce $\alpha = 0.05$, et x_2 est utile dans $M1$ donc elle est significative.
4) β_3: p-valeur $0.012734 < \alpha$, donc on rejette H_0 et accepte H_1, c'est-à-dire on décide que le coefficient β_3 est non nul avec risque de première espèce $\alpha = 0.05$, et x_3 est utile dans $M1$ donc elle est significative.

En conclusion, on considère de garder toutes les composantes variables dans $M1$.

Analyse la significativité globale de la regression

Il s'agit de test Fisher sur un sous modèle linéaire du $M1$. On veut tester $H_0 : \mu = \beta_1 = \beta_2 = \beta_3 = 0$, contre H_1 : l'un des paramètres n'est pas nul.

On observe que le F-statistique observé est 20.63, et sa p-valeur est $1.028e-06 < \alpha = 0.05$, donc on rejette H_0 et accepte H_1, avec risque d'erreur de première espèce α. On peut conclure qu'au moins l'un des quatre coefficient (composante paramètre) n'est pas nul, au moins l'un des quatre variables (intercepte et x_1, x_2, x_3) est significative.

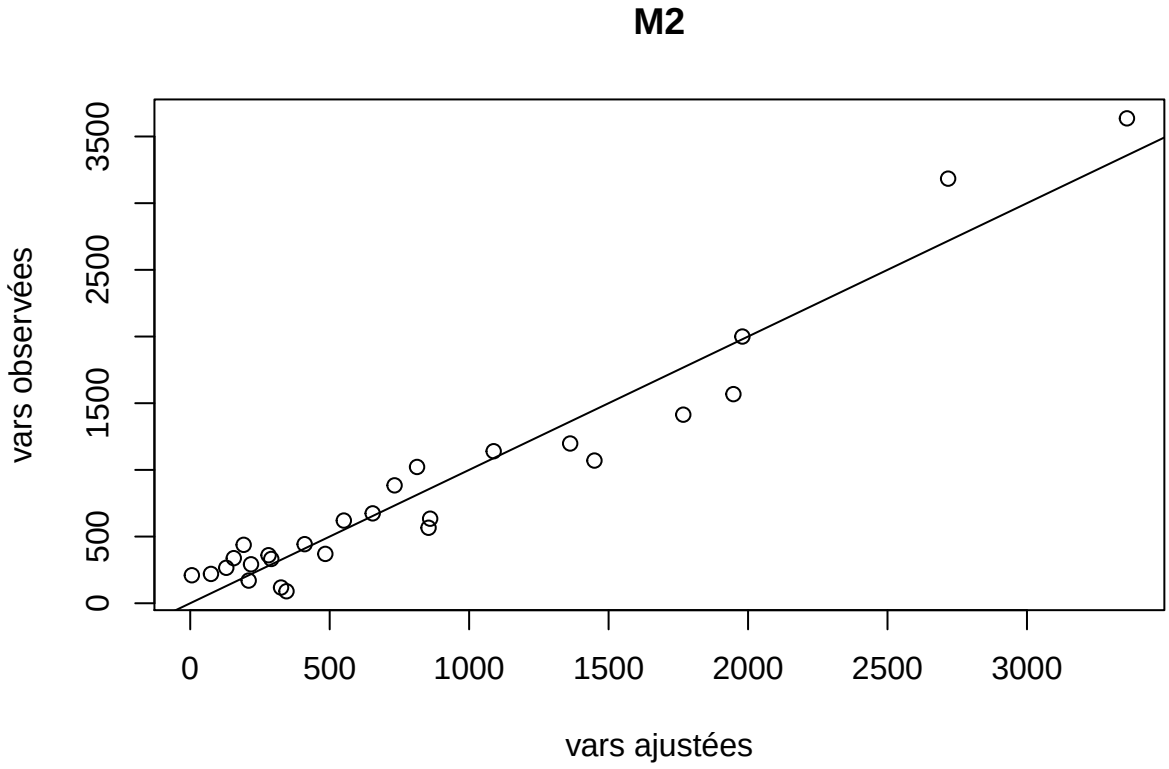
Analyse R squared

R^2 s'interprète comme la part de variance expliquée par les régresseurs supplémentaires. Donc 0.7291 signifie que la population de 72.91% de données peuvent être expliquées par notre modèle $M1$, la mauvaise qualité de l'ajustement et aussi énormément des brutes. En conclusion, $M1$ le LM multiple n'est pas un modèle idéal.

2. Modèle regression linéair d'ordre 2

Etude redression LM d'ordre 2

Maintenant, on voudrais modéliser un modèle régression linéaire d'ordre 2 $M2$: $y = \mu + \sum_i \beta_i x_i + \sum_i \beta_{ii} x_i^2 + \sum_{i<j} \beta_{i,j} x_i x_j + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, I_n)$ i.i.d. Voici dessous le graphe de l'ajustement:



On voit bien visuellement que l'ajustement du modèle $M2$ est largement mieux que $M1$! Les points s'allongent autour de la première bissectrice, les ajustements sont beaucoup plus proches des valeurs observées. En plus, il n'existe plus les valeurs négatives.

Maintenant, on analyse rapidement les valeurs données par R de même façon que l'on a fait dans la question précédente.

```
##
## Call:
```

```
## lm(formula = y ~ df$x1 + df$x2 + df$x3 + I(df$x1^2) + I(df$x2^2) +
##     I(df$x3^2) + I(df$x1 * df$x2) + I(df$x1 * df$x3) + I(df$x2 *
##     df$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 ***
## df$x1               660.00      64.09  10.299 1.00e-08 ***
## df$x2              -535.83      64.09  -8.361 1.99e-07 ***
## df$x3              -310.83      64.09  -4.850 0.000150 ***
## I(df$x1^2)          238.56     111.00   2.149 0.046317 *
## I(df$x2^2)          275.72     111.00   2.484 0.023712 *
## I(df$x3^2)          -48.28     111.00  -0.435 0.669081
## I(df$x1 * df$x2)   -456.50      78.49  -5.816 2.06e-05 ***
## I(df$x1 * df$x3)   -235.67      78.49  -3.003 0.008011 **
## I(df$x2 * df$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
```

Analyse la significativité des variables

Ici on veut tester chaque composante θ_i dans θ. $H_0 : \theta_i = 0$ contre $H_1 : \theta_i \neq 0$, on suppose que $\alpha = 0.05$. Etant donné qu'il y a dix composantes paramètres dans l'estimateur θ, on n'analyse pas un par un les variables ici.

En vérifiant chaque p-valeur correspondantes les variables, on peut conclure que seulement la variable x_3^2 et la variable x_2x_3 ne sont pas significatives. Parce que ses p-valeurs sont supérieures à α donc on regarde H_0 (avec risque de seconde espèce inconnue) et décide que ses coefficients sont nuls, donc les deux variables ne sont pas utiles dans $M2$ (i.e. ne sont pas significatives). Pour la reste, elles sont toutes significatives. En conclusion, on considère de ne pas garder x_3^2 et x_2x_3 dans $M2$.

Analyse la significativité globale de la regression

On veut tester H_0 : tous les composantes de l'estimateur sont égalent à 0, contre H_1 : l'un des paramètres n'est pas nul.

On observe que le F-statistique observé est 28.51, et son p-valeur est $1.564e-08 < \alpha = 0.05$, donc on rejette H_0 et accepte H_1, avec risque d'erreur de première espèce α. On peut conclure qu'au moins l'un des dix coefficients (composantes paramètre) n'est pas nul, au moins l'une des dix variables est significative.

Analyse R squared

R^2 s'interprète comme la part de variance expliquée par les régresseurs supplémentaires. Donc 0.9379 signifie que la population de 93.79% de données sont expliquées par notre modèle $M2$, ce qui implique un très bonne qualité de l'ajustement. En conclusion, $M2$ le LM multiple d'ordre 2 est un modèle assez performant.

Test modèle

On veut construire un test $M1$ contre $M2$ en utilisant "anova" fonction. En théorie, le F statistique est le même que le test de signifigativité globale de la regression. On voit bien le $F_{obs} = 9.5227$, p-valeur égale à $0.000115 < \alpha = 0.05$, donc on conclut que les variables d'interaction ajoutées sont significatives.

3. Test de variance

Conclusion

En conclusion on peut dire que l'on a ici étudié théoriquement la transformation Box-Cox puis l'on a utilisé différents outils pour l'utiliser concrètement sur un étude de cas. On a aussi pu se rendre compte de l'utilité et des limites de cette méthode.