Application de transformation de Box-Cox

LIAO puwei

2020/11/10

Introduction

Le but de ce projet est de étudier la régression linéaire entre 2 type d'observations, mais une avec la transformation , la raison pour laquelle on fait cette transformation est la significativité de régression linéaire des données d'observations d'origine n'est pas très fort, c'est à dire avec la régression linéaire d'observations d'origine, ce n'est pas très évident de prédire les observations de futur, mais après avoir la transformation,la régression linéaire est plus significative , ici il est plus facile à prédire les données de futur.Les 3 sections suivantes, on étudie d'abord la théorie de transformation de box cox, dans cette section, on cherche le meilleur coefficient de transformations box cox et ainsi les estimateurs de moyen et variance à partir de transformation, et puis , on peut construire la test pour tester si on conserve la valeur de coefficient de transformation ou pas,à partir de ce test , on peut facilement à trouver ce coefficient optimal.Dans la 2ème section, on simule les données, et on essaye de trouver le coefficient optimal dans cette exemple.et dans la dernière section, on étudie un exemple pratique, à partir des mesures Box et Draper, en faisant un test pour comparer des différentes modèles, le but c'est de voir quel paramètres de régression linéaire et ainsi quel modèle est plus pertinent pour cette régression linéaire.

Transformation de Box-Cox

à partir de cette section, on applique la transformation Box Cox h_λ de observations y pour faciliter à chercher une regréssion linéaire avec les données x, en considérons le modéle d'observations (x_i, Y_i) cherchons la regrésion linéaire entre $x_i \ et \ h_\lambda(y_i)$ plutôt que celle de $x_i \ et \ Y_i$ on peut obtenir l'équation de la regréssion linéaire suivante

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

où x_i est i ème observation d'expérience et ϵ_i le bruit iid gaussien. L'objectif est de déterminer λ_{opt} permettant un meilleur ajustement en régression de $h(y)$ Definissons la transformation de Box-Cox:

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

1.1

$\tilde{h}_\lambda(y)$ n'est pas compatible théoriquement, dans ce cas si on a y negative, alors dans ce cas on ne peut pas utiliser tout les λ théorique, par exemple, si $\lambda = \frac{1}{2}$, dans ce cas, λ ne peut pas etre trouver car y negative, c'est pour cette raison, on applique la tranformation h_λ mais pas $\tilde{h}_\lambda(y)$

1.2

pour estimer les paramètres calculons d'abord maximum de vraisemblance ,d'après la formule (1), on a

$$h_\lambda(Y_i) - x_i\theta = \epsilon_i \sim_{iid} \mathcal{N}(0, \sigma^2)$$

soit $\Phi = h_\lambda - x\theta$ on cherche la fonction repartition $F_t(\Phi(y_i))$,et on la derive par rapport a y_i,on obtiendra la fonction densite $f(\Phi(y_i))$ on obtient alors:

$$\frac{\partial F_{y_i}(\Phi(y_i))}{\partial y_i} = \frac{\partial h_\lambda(y_i)}{\partial y_i} f(\Phi(y_i))$$

maintenant on peut calculer sa vraisemblance

$$L(\lambda, \theta, \sigma^2; y) = \prod_{i=1}^{n} \left| \frac{\partial h_\lambda(y_i)}{\partial y_i} \right| \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp(-\frac{1}{2\sigma^2}(h_\lambda(Y) - X\theta)^{'}(h_\lambda(Y) - X\theta))$$

avec $J(\lambda, Y) = \prod_{i=1}^{n} \left| \frac{\partial h_\lambda(y_i)}{\partial y_i} \right|$, car $L(\lambda, \theta, \sigma^2; y)$ doit être positive.

1.3

On va maintenant estimer $\widehat{\theta}$ l'EMV de θ et $\widehat{\sigma^2}$ l'EMV de σ^2

$$logL = -\frac{n}{2}log(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}(h_\lambda(Y) - X\theta)^{'}(h_\lambda(Y) - X\theta) + \sum_{i}^{n} log(\left| \frac{\partial h_\lambda(y_i)}{\partial y_i} \right|)$$

$$\frac{\partial logL}{\partial \theta} = \frac{1}{\sigma^2} X^{'}(h_\lambda(Y) - X\theta)$$

Soit $\frac{\partial logL}{\partial \theta} = 0 \;\; \Leftrightarrow \;\; \frac{1}{\sigma^2}X^{'}(h_\lambda(Y) - X\theta) = 0 \;\; \Leftrightarrow \;\; \widehat{\theta} = (X^{'}X)^{-1}X^{'}h_\lambda(Y)$

Vérification: $\frac{\partial^2 logL}{\partial \theta^2} = -\frac{1}{\sigma^2}X^{'}X \leq 0$ Donc

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

maintenant on calcule $\widehat{\sigma^2}$, $\frac{\partial logL}{\partial \sigma^2} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}(h_\lambda(Y) - X\theta)^{'}(h_\lambda(Y) - X\theta)$ Si $\frac{\partial logL}{\partial \sigma^2} = 0$

$$\frac{n}{\sigma} = \frac{1}{\sigma^3}(h_\lambda(Y) - X\theta)^{'}(h_\lambda(Y) - X\theta) \;\; \Leftrightarrow \;\; \widehat{\sigma^2} = \frac{(h_\lambda(Y) - X\theta)^{'}(h_\lambda(Y) - X\theta)}{n}$$

en remplacant θ par son estimateur:

$$\begin{aligned} n\widehat{\sigma^2} &= (Z - X(X'Z(X'X)^{-1}))'(Z - X(X'Z(X'X)^{-1})) \\ &= Z'Z - X'(X'Z(X'X)^{-}Z)Z - Z'X(X'Z(X'X)^{-}) + XX'(X'Z(X'X)^{-})^2 \\ &= Z'Z - 2Z'X(X'Z(X' \\ &= Z'Z - Z'X(X'X)^{-}(X'Z) \\ &= Z'(I_n - X(X'X)^{-}X')Z \end{aligned}$$

on obtient alors

$$\widehat{\sigma^2} = \frac{Z'(I_n - X(X'X)^{-}X')Z}{n}$$

Vérification:

$$\frac{\partial^2 logL}{\partial \sigma^4} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(h_\lambda(Y) - X\theta)^{'}(h_\lambda(Y) - X\theta)$$

ce qui a la meme signe que:

$$\frac{n}{2} - \frac{\epsilon'\epsilon}{\sigma^2}$$

En remplaçant σ^2 par $\widehat{\sigma^2}$, on obtient

$$\frac{n}{2} - \frac{n^2}{\widehat{\epsilon'}\widehat{\epsilon}} \leq 0 \ \ avec\ n\ assez\ grand$$

Donc

$$\widehat{\sigma^2} = \frac{(h_\lambda(Y) - X\theta)^{'}(h_\lambda(Y) - X\theta)}{n} = \frac{Z'(I_n - X(X'X)^{-}X')Z}{n}$$

On pourra calculer $L_{max}(\lambda)$ maintenant en fixant $\widehat{\theta}\ et\ \widehat{\sigma^2}$

$$L_{max}(\lambda) = logL(\lambda, \widehat{\theta}(\lambda), \widehat{\sigma}^2(\lambda))$$

$$= -\frac{n}{2}log(2\pi) - \frac{n}{2}log\widehat{\sigma}^2(\lambda) - \frac{n}{2(h_\lambda(Y) - X\theta)^{'}(h_\lambda(Y) - X\theta)}(h_\lambda(Y) - X\theta)^{'}(h_\lambda(Y) - X\theta) + \sum_i log|\frac{\partial h_\lambda(Y_i)}{\partial Y_i}|$$

ici

$$\sum_i log|\frac{\partial h_\lambda(Y_i)}{\partial Y_i}| = (\lambda - 1)\sum_i log|Y_i|$$

car

$$|\frac{\partial h_\lambda(Y_i)}{\partial Y_i}| = |Y_i|^{\lambda - 1}$$

donc on obtient finalement que:

$$L_{max}(\lambda) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log\widehat{\sigma}^2(\lambda) - \frac{n}{2} + (\lambda - 1)\sum_i log|Y_i|$$

$$= -\frac{n}{2}\ log(\widehat{\sigma}^2(\lambda)) + (\lambda - 1)\sum_i log|Y_i| + a(n)$$

Avec

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

Pour calculer $\widehat{\lambda}$, on utilise la méthode de l'EMV. Soit $g = (\prod_{i=1}^n |y_i|)^{\frac{1}{n}}$, et $y(x,g) = \frac{h_\lambda(y)}{g^{\lambda-1}}$ Dans ce cas

$$L_{max} = a(n) - \frac{n}{2}log\frac{g^{\lambda-1}y(\lambda,g)^{'}(I_n - H)g^{\lambda-1}y(\lambda,g)}{n} + (\lambda - 1)log(g^n)$$

avec $H = X(X^{'}X)^{-1}X^{'}$

$$= a(n) - \frac{n}{2}log((g^{\lambda-1})^2) - \frac{n}{2}log(S_\lambda^2) + nlog(g^{\lambda-1})$$
$$= a(n) - \frac{n}{2}log(S_\lambda^2)$$

avec $S_\lambda^2 = \frac{y(\lambda,g)^{'}(I_n-H)y(\lambda,g)}{n}$ Ici, pour maximiser L_{max}, c'est la meme chose de minimiser S_λ^2, ce qui est la somme des carrés residuelle de $y(\lambda, g)$, notons $SCR(\lambda)$, dans ce cas

$$\widehat{\lambda} \in argmin_\lambda SCR(\lambda)$$

1.4

notons $\beta = \begin{pmatrix} \theta \\ \sigma^2 \\ \lambda \end{pmatrix}$

définissons

$$\widehat{J}(\beta) = -\frac{\partial^2 logL(\beta|y)}{\partial\beta\partial\beta'}$$

définissons l'information de Fisher

$$J(\beta) = \mathbb{E}(\widehat{J}(\beta))$$

ici

$$(\frac{(J(\beta))}{n})^{\frac{1}{2}}\sqrt{n}(\widehat{\beta}-\beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,I)$$

Car $J(\beta)$ on ne le connait pas, on le remplace par un estimateur $\widehat{J}(\widehat{\beta})$ Donc ici,

$$(\widehat{J}(\widehat{\beta}))^{\frac{1}{2}}(\widehat{\beta}-\beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,I)$$

et $\widehat{\beta} \sim \mathcal{N}(\beta, (\widehat{J}(\widehat{\beta}))^{-1})$

$$-\hat{J}(\beta) = \begin{pmatrix} \frac{\partial^2\ell(\boldsymbol{\beta}|\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} & \frac{\partial^2\ell(\boldsymbol{\beta}|\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\sigma^2} & \frac{\partial^2\ell(\boldsymbol{\beta}|\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\lambda} \\ \left(\frac{\partial^2\ell(\boldsymbol{\beta}|\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\sigma^2}\right)' & \frac{\partial^2\ell(\boldsymbol{\beta}|\boldsymbol{y})}{\partial(\sigma^2)^2} & \frac{\partial^2\ell(\boldsymbol{\beta}|\boldsymbol{y})}{\partial\sigma^2\partial\lambda} \\ \left(\frac{\partial^2\ell(\boldsymbol{\beta}|\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\lambda}\right)' & \left(\frac{\partial^2\ell(\boldsymbol{\beta}|\boldsymbol{y})}{\partial\sigma^2\partial\lambda}\right)' & \frac{\partial^2\ell(\boldsymbol{\beta}|\boldsymbol{y})}{\partial\lambda^2} \end{pmatrix}$$

$$= \frac{1}{\sigma^2}\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \frac{\boldsymbol{X}'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\theta})}{\sigma^2} & -\boldsymbol{X}'\boldsymbol{u} \\ \frac{(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\theta})'\boldsymbol{X}}{\sigma^2} & \frac{(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\theta})'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\theta})}{\sigma^4} - \frac{n}{\sigma^2} & -\frac{\boldsymbol{u}'(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\theta})}{\sigma^2} \\ -\boldsymbol{u}'\boldsymbol{X} & -\frac{(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\theta})'\boldsymbol{u}}{\sigma^2} & \boldsymbol{u}'\boldsymbol{u} - \boldsymbol{v}'(z-\boldsymbol{X}\boldsymbol{\theta}) \end{pmatrix}$$

$$= \frac{1}{\sigma^2}\begin{pmatrix} \widehat{J}(\beta)_{11} & \hat{J}(\beta)_{12} \\ \widehat{J}(\beta)_{21} & \hat{J}(\beta)_{22} \end{pmatrix}$$

on peut pour ce moment estimer $Var(\widehat{\lambda})$ par

$$\widehat{\text{Var}}(\widehat{\lambda}) = \sigma^2\left(\widehat{J}(\boldsymbol{\beta})_{22} - \widehat{\boldsymbol{J}}(\boldsymbol{\beta})_{21}\left[\widehat{\boldsymbol{J}}(\boldsymbol{\beta})_{11}\right]^{-1}\widehat{\boldsymbol{J}}(\boldsymbol{\beta})_{12}\right)^{-1}$$

alors a partir de ce moment ,on peut calculer cette estimateur:

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{\lambda}) &= \hat{\sigma}^2\left(\widehat{J}(\widehat{\beta})_{22}-\widehat{J}(\widehat{\beta})_{21}\left[\widehat{J}(\widehat{\beta})_{11}\right]^{-}\widehat{J}(\widehat{\beta})_{12}\right)^{-1}\\ &= \hat{\sigma}^2\left[\hat{u}'\hat{u}-\hat{v}'(I-H)\hat{z}-\left(\hat{u}'H\hat{u}+\frac{2\left(\hat{u}'(I-H)\hat{z}\right)^2}{n\hat{\sigma}^2}\right)\right]^{-1}\\ &= \hat{\sigma}^2\left[\hat{u}'(I-H)\hat{u}-\hat{v}'(I-H)\hat{z}-\frac{2\left(\hat{u}'(I-H)\hat{z}\right)^2}{n\hat{\sigma}^2}\right]^{-1}\\ &= -\left[\frac{\hat{v}'(I-H)\hat{z}-\hat{u}'(I-H)\hat{u}}{\hat{\sigma}^2}+\frac{2}{n}\left(\frac{\hat{u}'(I-H)\hat{z}}{\hat{\sigma}^2}\right)^2\right]^{-1}\\ &= -H(\widehat{\lambda})^{-1}\end{aligned}$$

$avec\ H = X(X'X)^{-}X'\ ,\ v_i = -\frac{\partial \boldsymbol{u}_i}{\partial\lambda}\ , u_i = \frac{\partial \boldsymbol{z}_i}{\partial\lambda}$

$$\widehat{SE}(\widehat{\lambda}) = \frac{1}{\sqrt{-H(\widehat{\lambda})}}$$

$d'après\ ce\ qu'on\ a\ obtenu,\ on\ peut\ donc\ trouver\ l'intervalle\ de\ confiance:$

$$\lambda \in [\widehat{\lambda} \pm q_{1-\frac{\alpha}{2}} * \widehat{SE}(\widehat{\lambda})]$$

$q_{1-\frac{\alpha}{2}}\ le\ quantile\ de\ loi\ normale\ de\ niveau\ 1-\frac{\alpha}{2}$

on voudrait construire un test de wald: $H_0 : \lambda = \lambda_0\ contre\ H1 : \lambda \neq \lambda_0\ A = (0,0,1)\ \beta_0 = \begin{pmatrix}\theta_0\\ \sigma_0^2\\ \lambda_0\end{pmatrix}$

posons un test de wald:

$$H_0 : A(\beta-\beta_0) = 0\ contre\ H1 : A(\beta-\beta_0) \neq 0$$

$on\ peut\ alors\ definir\ la\ statistique\ de\ Wald:$

$$W = [A(\hat{\beta}-\beta_0)'(A\hat{J}(\hat{\beta})^{-}A')^{-}A(\hat{\beta}-\beta_0)]$$

$ce\ qui\ suit\ une\ loi\ \chi^2(1)\ sous\ H_0$

$son\ intervalle\ de\ rejet\ est\ donc:$

$$R_\alpha = \left\{x;\quad W(x) > q_{1-\alpha}^{\chi^2(1)}\right\}$$

1.5

calculons leurs rapport de vraisemblance:

$$LRV = -2log(\frac{L(\lambda)}{L(\widehat{\lambda})})$$

$$= -2(-logL(\lambda) - logL(\widehat{\lambda}))$$

$$= -2(L_{max}(\lambda) - L_{max}(\widehat{\lambda}))$$

$$= 2(L_{max}(\widehat{\lambda}) - L_{max}(\lambda))$$

ce qui converge vers une loi de $\chi^2(1)$ posons un test hypothese : $H_0 : \lambda = \lambda_0$ *contre* $H1 : \lambda \neq \lambda_0$ *et sous* H_0 l'intervalle de rejet est la suivante:

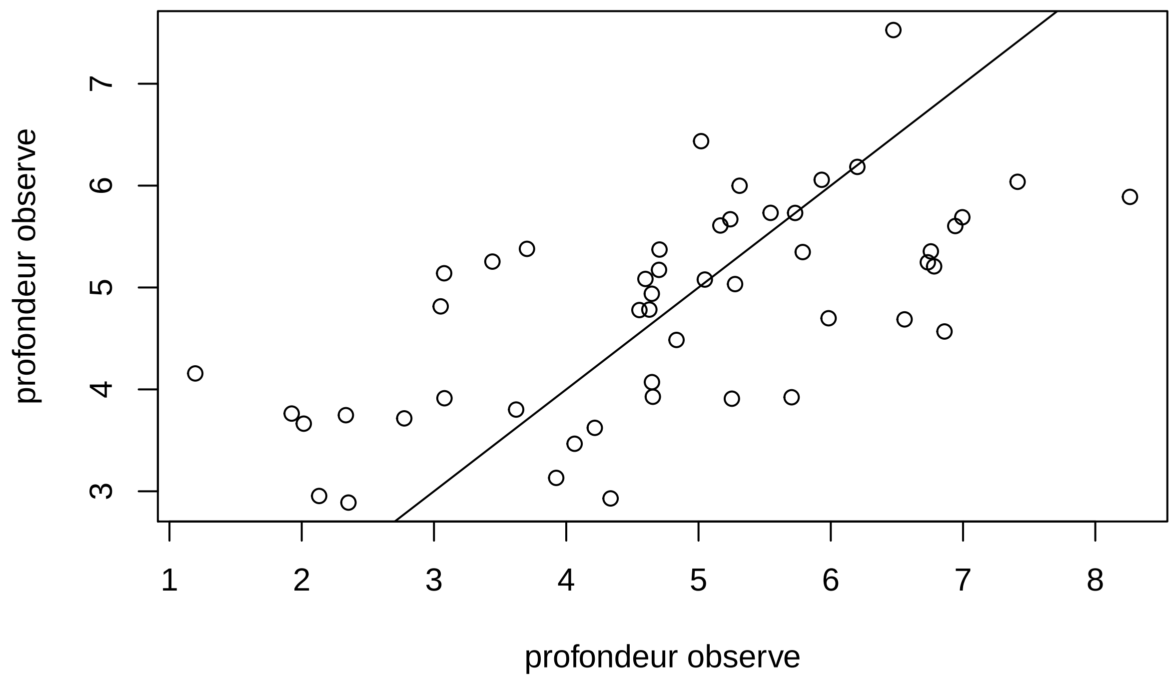
$$R_\alpha = \left\{ TRV > q_{1-\alpha}^{\chi_1^2} \right\}$$

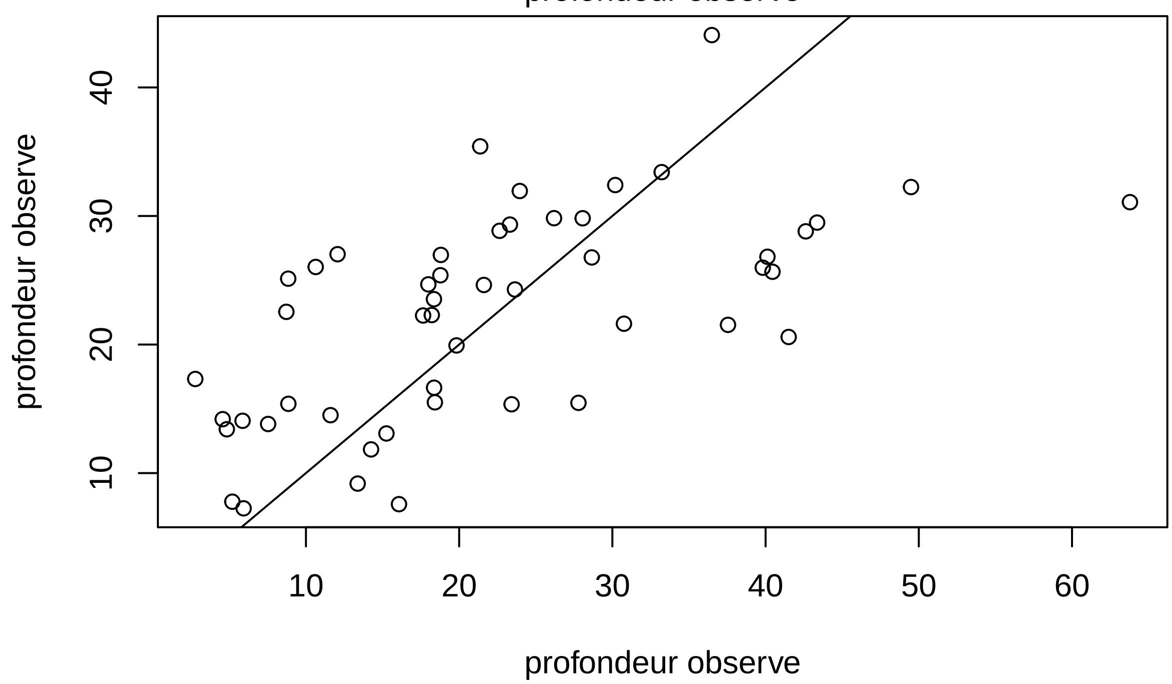
avec $TRV = 2(L_{max}(\widehat{\lambda}) - L_{max}(\lambda_0)),$ *et* $q_{1-\alpha}^{\chi_1^2}$ *quantile de niveau* $1-\alpha$ *de loi de* χ_1^2

Test de la méthode sur des données simulées

Dans cette section, on simule les observation x et celles apres avoir la transformation z ,et celles de ne pas avoir la transformation y,et puis on etudie la regression lineaire entre x et z et celle entre x et y, pour savoir que la transformation est elle significative pour la régression linéaire.

2.1

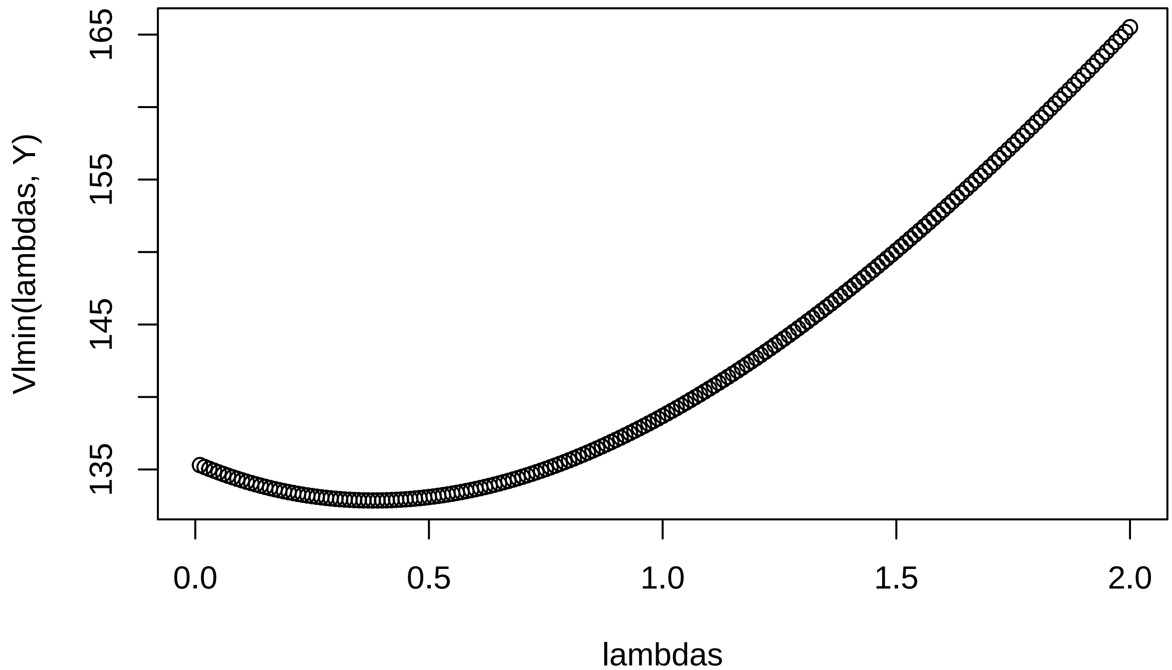




C'est visible que la régression linéaire de x et z, c'est à dire celles après avoir la transformation est plus significative que la régression linéaire de x et y,c'est à dire celles qui ne pas avoir la transformation, parce que la graphe précédentes, dont tout les points sont tous plus proche à cette droite.

2.2

on a deja cree la matrice d'experience Xexp la question precedente. d'apres la section 1, les code suivante nous permet de calculer l'estimation de variance de epsilon ce qui nous donne $Z'(I_n - X(X'X)^{-1}X')Z$ ou la matrice Q est égale à $I_n - X(X'X)^{-1}X'$ dans cette question la fonction Lmle nous permet de calculer les coef de Lmax.



d'après ce qu'on a pour les code, on peut tracer cette courbe $-L_{max}$ en fonction de λ et visiblement, le point qui minimise cette courbe est entre 0 et 0.5

2.3

resopt; $minimum [1] 132.8632

$estimate [1] 0.3817253

$gradient [1] -6.306777e-05

$hessian [,1] [1,] 33.91607

$code [1] 1

$iterations [1] 4

d'apres ces résultat obtenus, on peut voir que ,le min c'est 132.8632 , et le point qui minimise cette courbe est 0.3817253, et d'après la partie théorique , on peut déduire facilement que l'estimateur de variance est 0.02948455, c'est à dire

$$\hat{\lambda} = 0.3817253$$

$$v\hat{a}r(\hat{\lambda}) = 0.02948455$$

2.4

D'après la section théorique, on en deduit donc intervalle de confiance de niveau $1 - \alpha$ avec $\alpha = 0.05$ pour λ est la suivante:

$$\lambda \in [0.04517861\ ,\ 0.718272]$$

On effectue la test de Wald, pour tester les 4 hypothèse suivante(après avoir la traduction):

$$H_0 : \lambda = 1$$

$$H_0 : \lambda = 0.5$$

$$H_0 : \lambda = 0.3$$

$$H_0 : \lambda = 0$$

Calculons alors les 4 statistique de wald et ainsi le quantile:

$$w1 = 12.96488$$

$$w2 = 0.4744487$$

$$w3 = 0.2265263$$

$$w4 = 4.942053$$

$$q = 3.84145$$

Après avoir comparé les quatres statistiques, on rejette l'hypothèse nulle ce qui est plus grand que cette quantile, c'est à dire on conserve $H_0 : \lambda = 0.5$ *et* $H_0 : \lambda = 0.3$ et on rejette les 2 autres sous niveau de $\alpha = 0.05$

2.5

Calculons alors les 4 statistique de rapport de vraisemblance:

$$l1 = 11.66127$$

$$l2 = 0.4663807$$

$$l3 = 0.2289855$$

$$l4 = 5.148459$$

les 2 test suit meme loi sous H0, donc on a le meme quantile les 2 tests, ici le test de rapport de vraisemblance nous dit la meme chose ,c'est qu'on conserve ces 2 hypotheses: H_0 :la transformation a appliquer aux observations est en racine carree $H_0 : \lambda = 0.3$

2.6

On utilise la fonction powertransform pour comparer les resultats precedente ici on obtient:

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

[1] "lambda= 1" [1] "TRV: 11.661267167003" [1] "p_value: 0.000638148590273264" [1] "conserve H_0? FALSE" [1] "lambda= 0.5" [1] "TRV: 0.466380664195356" [1] "p_value: 0.494656961440495" [1] "conserve H_0? TRUE" [1] "lambda= 0.3" [1] "TRV: 0.228985473731598" [1] "p_value: 0.632277106840418" [1] "conserve H_0? TRUE" [1] "lambda= 1e-06" [1] "TRV: 5.14845895351687" [1] "p_value: 0.0232670132748679" [1] "conserve H_0? FALSE"

Ici on obtient tous la même résultat que précédentes, et pour les tests aussi, donc c'est bon

Cas pratique

3.1

Dans cette section, on propose un exemple pratique mais en utilisant des mesures de Box et Draper Definissons le modele suivante:

$$(M1) \qquad Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon = X\theta + \epsilon$$

avec$\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$ ainsi

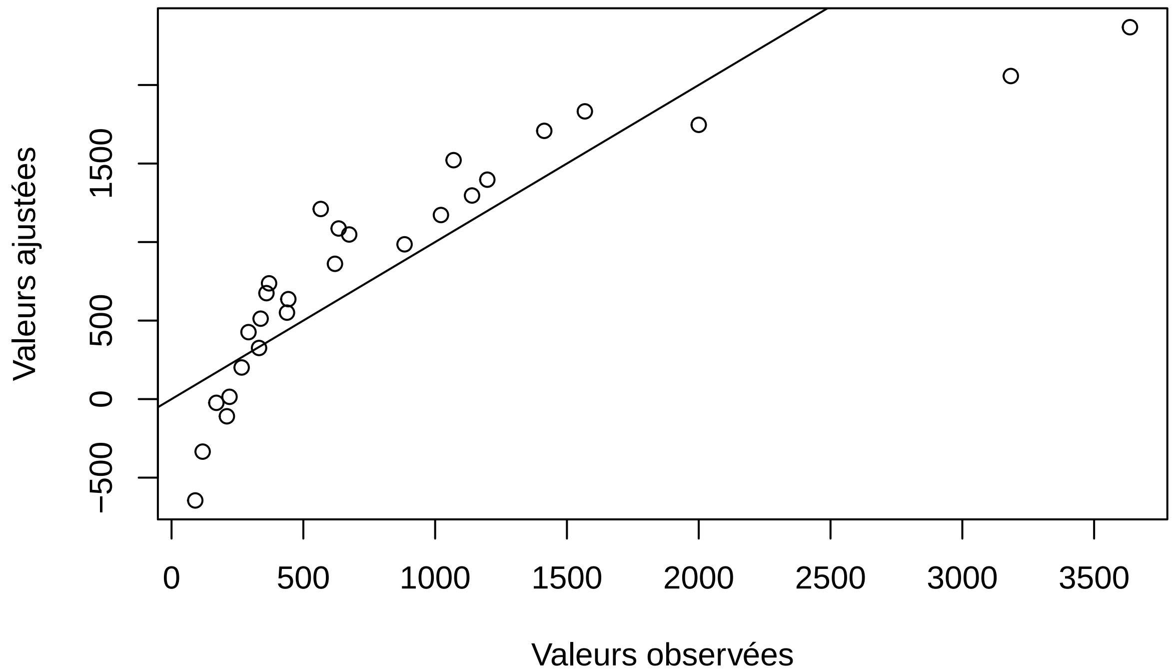
$$X = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & x_{3,1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,27} & x_{2,27} & x_{3,27} \end{pmatrix}$$

$$\theta = \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

le valeur ajuste apres avoir codé c'est

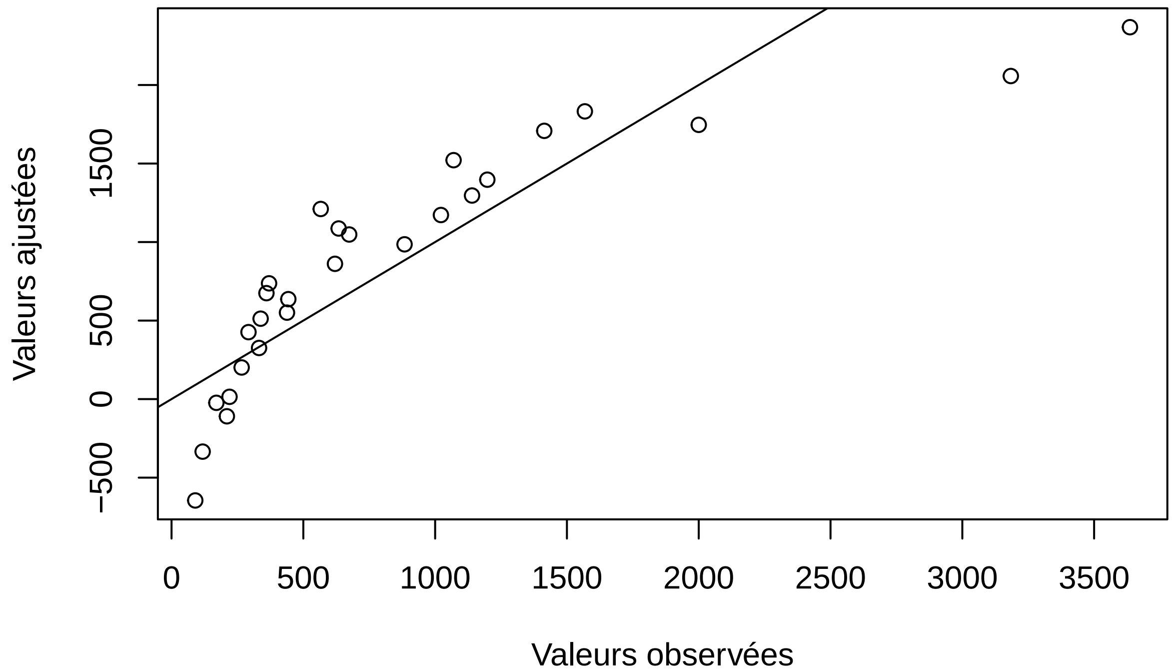
```
##             [,1]
##  [1,] 1048.03704
##  [2,] 1708.03704
##  [3,] 2368.03704
##  [4,]  512.20370
##  [5,] 1172.20370
##  [6,] 1832.20370
##  [7,]  -23.62963
##  [8,]  636.37037
##  [9,] 1296.37037
## [10,]  737.20370
## [11,] 1397.20370
## [12,] 2057.20370
## [13,]  201.37037
## [14,]  861.37037
## [15,] 1521.37037
## [16,] -334.46296
## [17,]  325.53704
## [18,]  985.53704
## [19,]  426.37037
## [20,] 1086.37037
## [21,] 1746.37037
## [22,] -109.46296
## [23,]  550.53704
## [24,] 1210.53704
## [25,] -645.29630
## [26,]   14.70370
## [27,]  674.70370
```

Et la graphe de régression linéaire est la suivante




```
##
## 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
```

Ici on etudie la régression linéaire entre y et longueur, amplitude, chargement




```
##
## 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
```

d'après ces 2 modèles, on va facilement que, il n'y a pas de différences entre ces 2 modèles, et ainsi, multiple R-squared vaut 0.7291, c'est pas assez grand, donc la significativité de ces 2 régression linéaire sont faible, et de plus, ce P value est très petit, c'est à dire on ne garde pas ce modèle

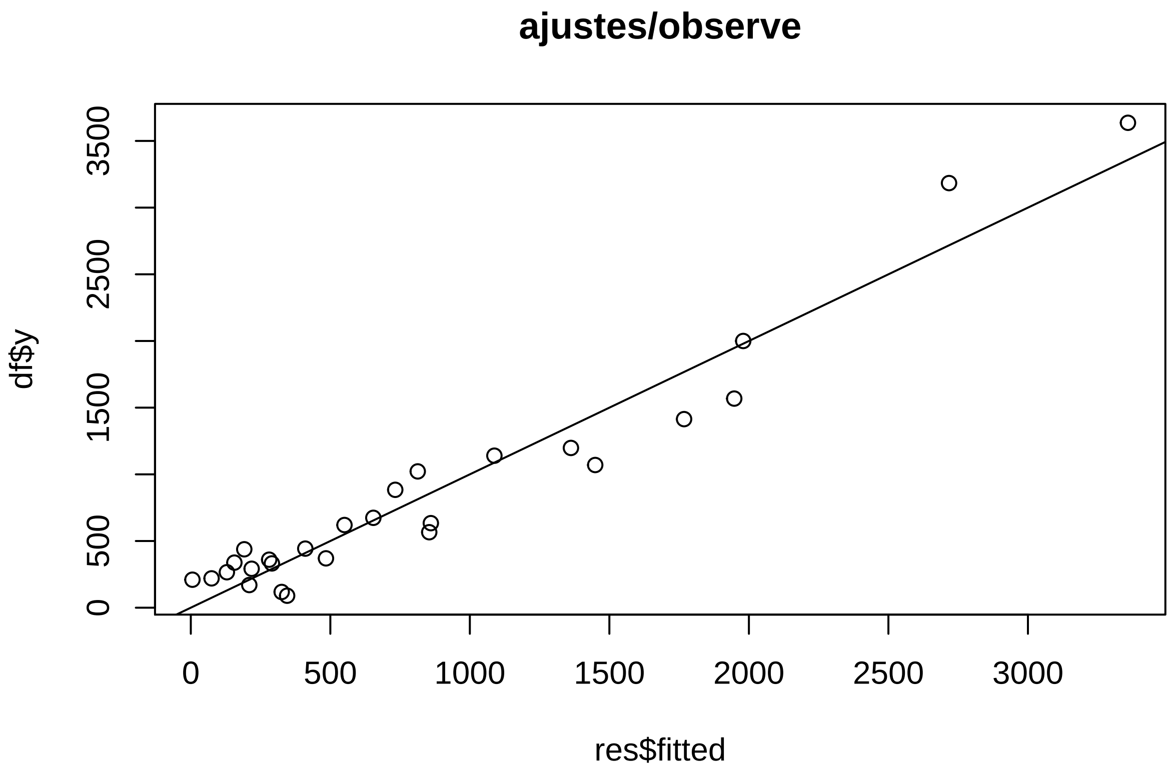
3.2

Construisons un autre modele:

$$(M2) \qquad \begin{aligned} Y = \mu + &\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 \\ &+ \beta_7 x_1 x_2 + \beta_8 x_1 x_3 + \beta_9 x_2 x_3 + \epsilon \end{aligned}$$

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & x_1^2 & x_2^2 & x_3^2 & x_1x_2 & x_1x_3 & x_2x_3 \end{pmatrix}$$

$$\theta = \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \\ \beta_9 \end{pmatrix}$$



D'après cette graphe, on voit que la regression lineaire de ce modèle est plus précise que précédentes.

```
##
## 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$x3 * df$x2) + I(df$x1 *
##     df$x3))
##
## 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$x3 * df$x2)   142.92      78.49   1.821 0.086278 .
## I(df$x1 * df$x3)  -235.67      78.49  -3.003 0.008011 **
## ---
## 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
```

on voit que Multiple R-squared: 0.9379, cela veut dire que cette modèle est plus pertinent

```
## Analysis of Variance Table
##
## Model 1: Y ~ df$x1 + df$x2 + df$x3
## Model 2: Y ~ df$x1 + df$x2 + df$x3 + I(df$x1^2) + I(df$x2^2) + I(df$x3^2) +
##     I(df$x1 * df$x2) + I(df$x3 * df$x2) + I(df$x1 * df$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
```

Ici, p val = 0.0001154,ce qui est plus petit que 0.001, c'est a dire ajouter les termes d'ordres 2 rend la regression lineaire plus precise, donc ici on prefere la modele M2,donc l'ajout des variables d'intersection est preconiser.

3.3

(M1bis)et (M2bis) on propose tous la transformation box cox en (M1) et (M2)

conclusion

La transformation box cox nous permet de faciliter tout le temps la régression linéaire, mais il faut faire attention avec la choix de modèle.Parce que la choix de modele nous permet d'avoir une regression lineaire encore plus precise que les autre.