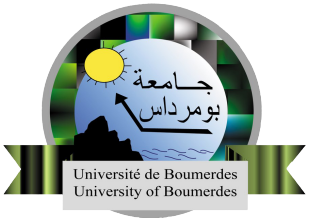
République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université M'hamed Bougara de Boumerdès



Faculté des Sciences
Département des Mathématiques

Mémoire présenté,

Par

M^lle. **Gariti Kenza** & M^lle. **Haddad Nawal**

Pour l'obtention du diplôme de Master en Modélisation Stochastiques et Statistiques

"Régression linéaires aux valeurs extrêmes "

Soutenu publiquement le 06/07/2022, devant le jury composé de :

Présidente	M^me	Chemrik.H	M.A.A	U.M.B.B.
Examinateur	M^r	Tazerouti.M	M.C.B	U.M.B.B.
Encadreur	M^me	Benmansour.M	M.A.A	U.M.B.B.

Année Universitaire 2021 – 2022

Remerciements

Tout d'abord, nous remercions le bon Dieu qui nous a donnée la force, la santé ,la patience, la volonté et le courage qui nous ont permis de terminer nos études par la réalisation de ce modeste travail.

"الحمد لله حمدا كثيرا طيبا مباركا فيه"

Nous tenons a remercie vivement nos trés chéres parents pour leur amour inconditionnel et leur soutien permanent et sans faille.

Nous remercions tout particuliérement notre promotrice Mme M.Benmansour pour avoir accepté de nous encadrer, et nous avoir encourager, consacrer son temps précieux et nous avoir accordé son attention avec une extrême patience et pour sa disponibilité et ses conseils,

Nous souhaite également remercie les membres du jury Mme H.Chemrik et Mer M.Tazerouti d'avoir accepté d'examiner et d'évaluer notre travail .

Nos remerciements s'adressent également a l'ensemble des enseignants du Département des Mathématiques et spécialement spécialité Modélisation Stochastiques et Statistiques.

Pour finir, nous souhaitons adresser nos sincères remerciements à tous ceux qui ont contribué, de près ou de loin à la réalisation de ce mémoire.

Dédicace

A mon très cher père.

A ma très chère mère.

A mes trés chers grands parents.

A mes frères et soeurs.

A ma binôme Nawal et Nabila qui m'ont soutenue jusqu'au bout. A toute ma famille et mes amies.

Je dédié ce modeste travail

KENZA

Dédicace

A mon très cher père.

A ma très chère mère.

A mes trés chers grands parents.

A mes frères .

A ma binôme Kenza et Mon fiancé Mohammed qui m'ont soutenue jusqu'au bout. A toute ma famille et mes amies.

Je dédié ce modeste travail

NAWAL

Table des matières

Liste des figures

Introduction

L'analyse de régression peut être définie comme la recherche de la relation stochastique que lie deux ou plusieurs variables. Son champ d'application recouvre de multiples domaines, parmi lesquels on peut citer la physique, l'astronomie, la biologie, la chimie, la médecine, la géographie, la sociologie, l'histoire, l'économie ...
La régression est l'une des méthodes les plus connues et les plus appliquées en statistique pour l'analyse de données quantitatives. Elle est utilisée pour établir une liaison entre une variable quantitative et une ou plusieurs autres variables quantitatives sous la forme d'un modèle. Si on s'intéresse a la relation entre deux variables, on parlera de régression simple en exprimant une variable en fonction de l'autre. Si la relation porte entre une variable et plusieurs autres variables, on parlera de régression multiple.
La régression linéaire simple et multiple est une classe particulière de modèle de régression. Le but est d'expliquer une variable Y, appelée variable endogène par une ou plusieurs variables explicatives dites exogènes a travers une fonction affine.
Les modèles linéaires classiques et régression linéaire sont basées sur la comparaison des moyennes entre différents groupes ou différentes valeurs d'un prédicteur, avec une incertitude basée sur le calcul de la variance résiduelle. Les coefficients d'un modèle linéaire sont révélés par la méthode des moindres carrés, qui vise à minimiser cette variance résiduelle.
Ces méthodes sont conçues pour être optimales lorsque la variation résiduelle convient à une distribution normale, qui prévoit relativement peu de valeurs extrêmes. La présence de quelques valeurs extrêmes exerce une forte influence sur les produits émis par ces méthodes et rend difficile la détection des effets représentés par la plus grande partie des données.

En particulier, les estimations des moindres carrés pour les modèles de régression sont très sensibles aux valeurs aberrantes. Bien qu'il n'y ait pas de définition précise d'une valeur aberrante, les valeurs aberrantes sont des observations qui ne suivent pas le modèle des autres observations. Ce n'est normalement pas un problème si la valeur aberrante est simplement une observation extrême tirée de la queue d'une distribution normale, mais si la valeur aberrante résulte d'une erreur de mesure non normale ou d'une autre violation des hypothèses standard des moindres carrés ordinaires, cela compromet la validité. des résultats de la régression si une technique de régression non robuste est utilisée.

la régression robuste linéaire présentée dans Ronchetti (2006) et Duncan

et Guerrier (2016). Dans ces références, il est indiqué que la grande majorité des modèles statistiques utilisés dans différents domaines allant de la finance à la biologie et à l'ingénierie, par exemple, sont des modèles paramétriques. Sur la base de ces modèles, des hypothèses sont formulées concernant les propriétés des variables d'intérêt (et les modèles eux-mêmes) et des procédures optimales sont dérivées sous ces hypothèses. Parmi ces procédures, les estimateurs des moindres carrés et du maximum de vraisemblance sont des exemples bien connus qui, cependant, ne sont optimaux que lorsque les hypothèses statistiques sous-jacentes sont exactement satisfaites. Si ce dernier cas ne se vérifie pas, alors ces procédures peuvent devenir considérablement biaisées et/ou inefficaces lorsqu'il existe de petits écarts par rapport au modèle. Les résultats obtenus par les procédures classiques peuvent donc être trompeurs lorsqu'ils sont appliqués à des données réelles (voir par exemple-Ronchetti (2006) et Huber et Ronchetti (2009)).
Afin de résoudre les problèmes liés aux hypothèses paramétriques violées, les statistiques robustes peuvent être considérées comme une extension des statistiques paramétriques classiques en considérant directement les écarts par rapport aux modèles.
En effet, alors que les modèles paramétriques peuvent être une bonne approximation de la véritable situation sous-jacente, des statistiques robustes ne supposent pas que le modèle est exactement correct. Une procédure robuste telle qu'énoncée dans Huber et Ronchetti (2009) devrait donc avoir les caractéristiques suivantes :

— Il devrait estimer efficacement le modèle supposé.
— Il doit être fiable et raisonnablement efficace sous de petits écarts par rapport au modèle (par exemple, lorsque la distribution se situe dans un voisinage du modèle supposé).
— Des écarts plus importants par rapport au modèle ne devraient pas affecter excessivement la procédure d'estimation.

Une méthode d'estimation robuste est un compromis par rapport à ces trois caractéristiques. Ce compromis est illustré par Anscombe et Guttman (1960) à l'aide d'une métaphore assurantielle : « sacrifier une certaine efficacité au modèle afin de s'assurer contre les accidents causés par des écarts au modèle ».
Dans la littérature, les méthodes de régression robuste sont parmi plusieurs approches utilisées lorsqu'on est en présence de données aberrantes. Notamment, dans ce mémoire nous verrons cinq de ces méthodes : la régression avec la méthode de moindre carrées pondérés, la régression avec M-estimation, la régression avec MM-estimation, la régression avec T-régression et régression

quantile. Toutefois, pour pouvoir appliquer ces méthodes, il faut supposer que les données suivent le comportement d'un modèle linéaire paramétrique.

Ce mémoire sera principalement consacré à l'étude des méthodes de régression robuste et la dernière partie de celui ci sera dédiée à l'application de ces méthodes aux données obtenues par la simulation d'un modèle de régression linéaire simple sous logiciel **R**.

Le premier chapitre de ce mémoire rappellera quelques notions de la régression multiple. L'estimation par la méthode des moindres carrées ordinaires y sera présentée ainsi que la définition de point de rupture et d'efficacité relative.

Dans le deuxième chapitre, nous aborderons quelques notion de base sur les valeurs extrêmes, ainsi que la notion de régression multiple robuste. Quelques méthodes seront étudiées, notamment la modélisation par M-estimation, par MM-estimation, par T-régression et par régressions quantiles.

Le dernier chapitre de ce mémoire sera consacré à la comparaison des méthodes de régressions robustes qui seront appliquer sur un échantillon de données obtenues par simulation de modèle de régression linéaire simple sous le logiciel **R**.
Finalement on termine par une conclusion générale.

Chapitre 1

Régression multiple avec estimateur non robuste

Dans la première section de ce premier chapitre, on rappellera le concept de régression linéaire multiple. Par la suite, la deuxième section sera consacrée à la présentation d'une méthode très populaire pour l'estimation des paramètres de régression multiple : l'estimation par la méthode des moindres carrés ordinaires(MCO). On présentera également quelques propriétés des estimateurs obtenus par cette méthode. Finalement, dans la dernière section de ce chapitre , on donnera les définition de rupture, d'efficacité relative ainsi que quelques exemples illustrant les limites de l'estimation par la méthode des MCO.

1.1 Régression linéaire multiple

La modélisation par régression linéaire multiple a pour but d'expliquer ou de prédire une variable réponse Y par une combinaison linéaire de plusieurs variables explicatives X_j, j= 1,...,p. Certain variables X_j peuvent être des transformation des autres variables initialement utilisées dans le modèle pour prendre en considération les effets non linéaires d'une variable. Par exemple, si on désire prendre en considération un effet quadratique de la variable X_j. Soit l'échantillon de n observations suivant :

$$E = \{x_{i,1}, x_{i,2}, ..., x_{i,p}, y_i\}_{i=1}^{n} \tag{1.1}$$

où $x_{i,j}$ est la i^{ieme} observation de la variable X_j et y_i est la i^{ieme} observation de la variable Y. Soit ϵ_i , i =1,...,n, un bruit aléatoire, qui représentera le

terme d'erreur lors de la modélisation. Le modèle de régression linéaire multiple prend alors la forme suivante :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_p x_{i,p} + \epsilon_i \quad \text{i=1,...,n}$$

où β_j , j= 0,...,p sont les paramètres à estimer du modèle et ϵ_i , i=1,...,n sont les erreurs issues de la modélisation. Dans le modèle de régression linéaire multiple,habituellement, on considère les hypothèses suivantes :
Hypothèses :

- H_1 : Les variable X_j sont aléatoire ou fixes.
- H_2 : $E(\epsilon_i) = 0$ pour tout i = 1,...,n .
- H_3 : $Var(\epsilon_i) = \sigma^2$ pour tout i = 1,...,n (homoscédasticité des erreurs).
- H_4 : $Cov(\epsilon_i, \epsilon_j) = 0 \quad i \neq j$ pour tout i, j = 1,...,n .
- H_5 : X_i et ϵ_i sont indépendantes,pour tout i = 1,...,n .
- H_6 : Pour l'inférence,on supposera que $\epsilon_i \sim N(0, \sigma^2)$, pour tout i = 1,...,n .

Pour simplifier les notations, nous utiliserons la notation matricielle. Le modèle de régression linéaires avec notation matricielle prend alors la forme suivante :

$$Y = X\beta + \epsilon \tag{1.2}$$

où

- Y=$(y_1, ..., y_n)^t$ est une matrice de dimension $n \times 1$, c'est-à-dire, les valeurs issues de la variable réponse.
- X est une matrice de dimension $n \times (p+1)$ contenant les valeurs des variables X_j

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

- β=$(\beta_0, ..., \beta_p)^t$ est un vecteur de $(p+1)$ paramètres inconnus de la régression.

— $\epsilon = (\epsilon_1, ..., \epsilon_n)^t$ est une matrice de dimension $n \times 1$ des erreurs du modèle de variance constante et inconnue.

Pour trouver le meilleur modèle linéaire possible, on cherche à estimer les valeurs de β qui minimiseront les erreurs.cette méthode appelée méthode des moindres carrées ordinaires, sera présentée dans la section suivante.

1.2 Estimation des paramètres β par la méthode des MCO

Pour estimer les paramètres du modèle (1.2), une méthode très utilisée est la méthode des moindres carrés ordinaires, cette méthode consiste à minimiser l'erreur quadratique moyenne relative aux termes d'erreur du modèle. Posons $\| A \|_2$ la norme L_2 d'un vecteur A. On cherche alors à résoudre le problème d'optimisation suivant :

$$\begin{aligned} \hat{\beta} &= argmin_\beta \sum_{i=1}^{n} \epsilon_i^2 \\ &= argmin_\beta [(Y - X\beta)^t (Y - X\beta)] \end{aligned} \tag{1.3}$$

$$= argmin_\beta \| Y - X\beta \|_2^2 \tag{1.4}$$

Remarque
Sous les hypothèses H_4 et H_6 du modèle (1.2), l'estimateur obtenu par les MCO est exactement l'estimateur obtenu par la méthodes du maximum de vraisemblance.
En effet, pour le modèle (1.2), avec les hypothèses H_4 et H_6, on a que :

$$Y - X\beta = \epsilon \sim N(0_{n\times 1}, \Sigma = \sigma^2 1_{n\times n})$$

Ainsi, la vraisemblance est donnée par :

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\Pi\sigma^2}} \exp\left(\frac{-\epsilon_i^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\Pi\sigma^2}}\right)^n \exp\left(\sum_{i=1}^{n} \frac{-\epsilon_i^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\Pi\sigma^2}}\right)^n \exp\left(\frac{-(Y - X\beta)^t (Y - X\beta)}{2\sigma^2}\right) \end{aligned}$$

Il est alors évident que maximiser cette vraisemblance par rapport à β est équivalent à minimiser $(Y - X\beta)^t(Y - X\beta)$ par rapport à β. On se retrouve avec le problème de minimisation (1.3).

On peut récrire (1.3) de la façon suivante :

$$\begin{aligned}\hat{\beta} &= argmin_\beta[((Y - X\beta)^t(Y - X\beta))] \\ &= argmin_\beta \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}))^2\end{aligned}$$

Posons $S(\beta) = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}))^2$. On utilise ensuite le calcule différentiel pour obtenir les estimateurs $\hat{\beta}_j$, j=0,...,p, de problème ci-dessus. Ceci consiste à évaluer les dérivées partielles suivantes :

$$\begin{aligned}\frac{\partial S(\beta)}{\partial \beta_0} &= \sum_{i=1}^{n} 2(y_i - (\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p})) \\ \frac{\partial S(\beta)}{\partial \beta_1} &= \sum_{i=1}^{n} 2(y_i - (\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}))x_{i,1} \\ &\vdots \\ \frac{\partial S(\beta)}{\partial \beta_p} &= \sum_{i=1}^{n} 2(y_i - (\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}))x_{i,p}\end{aligned}$$

Les estimateurs $\hat{\beta}_j$, sont les valeurs telles que $\frac{\partial S(\beta)}{\partial \beta_p} \mid_{\beta=\hat{\beta}} = 0$
pour tout j = 0,...,p.
On cherche alors à résoudre le système d'équation linéaire suivant pour trouver les estimateurs $\hat{\beta}_j$:

$$\begin{aligned}\sum_{i=1}^{n} y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i,2} + ... + \hat{\beta}_p \sum_{i=1}^{n} x_{i,p} \\ \sum_{i=1}^{n} y_i x_{i,1} &= \sum_{i=1}^{n} x_{i,1}\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i,2}x_{i,1} + ... + \hat{\beta}_p \sum_{i=1}^{n} x_{i,p}x_{i,1} \\ &\vdots \\ \sum_{i=1}^{n} y_i x_{i,p} &= \sum_{i=1}^{n} x_{i,p}\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1}x_{i,p} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i,2}x_{i,p} + ... + \hat{\beta}_p \sum_{i=1}^{n} x_{i,p}^2\end{aligned}$$

En ramenant ce système sous forme matricielle, on obtient alors :

$$\begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_{i,1} \\ \vdots \\ \sum_{i=1}^{n} y_i x_{n,p} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i,1} & \sum_{i=1}^{n} x_{i,2} & \cdots & \sum_{i=1}^{n} x_{i,p} \\ \sum_{i=1}^{n} x_{i,1} & \sum_{i=1}^{n} x_{i,1}^2 & \sum_{i=1}^{n} x_{i,2}x_{i,1} & \cdots & \sum_{i=1}^{n} x_{i,p}x_{i,1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^{n} x_{i,p} & \sum_{i=1}^{n} x_{i,1}x_{i,p} & \sum_{i=1}^{n} x_{i,2}x_{i,p} & \cdots & \sum_{i=1}^{n} x_{i,p}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$

Ce qui est équivalent à :

$$X^t Y = X^t X \hat{\beta}$$

Si la matrice $X^t X$ est inversible, on trouve alors que l'estimateur $\hat{\beta}$ de β par la méthode des moindres carrés ordinaire est défini par :

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

On peut vérifier que $\hat{\beta}$ est bien un minimum de $S(\beta)$. En effet, considérons la matrice hessienne $H(S) = \left(\frac{\partial^2}{\partial \beta_i \partial \beta_j} S(\beta) \right)_{i,j}$, avec i = 0,1,...,p et j = 0,1,...,p cette matrice était définie positive , c'est-à-dire que pour tout vecteur colonne v à $p+1$ composantes, on a que $vH(S)v^t > 0$. On conclu ainsi que $\hat{\beta}$ est un minimum de $S(\beta)$.

1.2.1 Propriétés de l'estimateur obtenu par la méthode des MCO

Sous les hypothèses H_1 à H_6, on peut démontrer certaines propriétés de l'estimateur des moindres carrés ordinaire $\hat{\beta}$.

Théorème 1(Estimateur sans biais)
$\hat{\beta}$ est un estimateur sans biais, c'est-à-dire que $E(\hat{\beta}) = \beta$.

Démonstration
Si X est fixe, on a :

$$\begin{aligned} E[\hat{\beta}] &= E[(X^tX)^{-1}X^tY] \\ &= (X^tX)^{-1}X^tE[Y] \\ &= (X^tX)^{-1}X^tE[X\beta+\epsilon] \\ &= (X^tX)^{-1}X^t(X\beta+E[\epsilon]) \\ &= (X^tX)^{-1}X^tX\beta \\ &= \beta \end{aligned}$$

Si X est aléatoire ,

$$\begin{aligned} E[\hat{\beta}] &= E[(X^tX)^{-1}X^tY] \\ &= E[E[(X^tX)^{-1}X^tY|X]] \\ &= E[(X^tX)^{-1}X^tE[Y|X]] \\ &= E[(X^tX)^{-1}X^tXE[X\beta+\epsilon|X]] \\ &= E[(X^tX)^{-1}X^t(X\beta+E[\epsilon|X])] \end{aligned}$$

En vertu de H2 et H5 $E[\epsilon|X]=0$ on obtient ainsi :

$$\begin{aligned} E[\hat{\beta}] &= E[(X^tX)^{-1}X^tX\beta] \\ &= \beta \end{aligned}$$

Théorème 2(variance)
La matrice de variance covariance conditionnelle à X de l'estimateur $\hat{\beta}$ est :

$$\Sigma(\hat{\beta}|X) = \sigma^2(X^tX)^{-1}$$

.
Démonstration. En effet

$$\begin{aligned} \Sigma(\hat{\beta}|X) &= \Sigma((X^tX)^{-1}X^tY|X) \\ &= (X^tX)^{-1}X^t\Sigma(Y|X)((X^tX)^{-1}X^t)^t \\ &= (X^tX)^{-1}X^t\Sigma(X\beta+\epsilon|X)X(X^tX)^{-1} \\ &= (X^tX)^{-1}X^t\Sigma(\epsilon|X)X(X^tX)^{-1} \\ &= (X^tX)^{-1}X^t\sigma^2X(X^tX)^{-1} \\ &= \sigma^2(X^tX)^{-1} \end{aligned}$$

Théorème 3 (Normalité)
Sous les six hypothèses H_1 à H_6,on a

$$\hat{\beta}|X \sim N(\beta, \sigma^2(X^tX)^{-1})$$

.
Démonstration.
Soit Z un vecteur aléatoire $d \times 1$ tel que

$$Z = (Z_1, Z_2, ..., Z_d) \sim N(\mu, H)$$

où H est une matrice définie positive de dimension $d \times d$ et N est la loi normale multidimensionnelle. Alors

$$AZ + B \sim N(A\mu + B, AHA^t)$$

où A est une matrice de dimension $d \times d$ et B est un vecteur de d éléments. Ainsi, puisque :

$$Y = X\beta + \epsilon \sim N(X\beta, \sigma^2 1_{d\times d})$$

il vient

$$\hat{\beta} = (X^tX)^{-1}X^tY \sim N((X^tX)^{-1}X^tX\beta, (X^tX)^{-1}X^t\sigma^2((X^tX)^{-1}X^t)^t)$$

.
On obtient alors

$$\hat{\beta} \sim N(\beta, \sigma^2(X^tX)^{-1})$$

Théorème 4 (Théorème de Gauss-Markov)
L'estimateur $\hat{\beta}$ est un estimateur **BLUE** (Best Linear Unbiaised Estimator) pour les paramètres β du modèle (1.2)

Démonstration.
Soit $\hat{\beta} = CY$ un autre estimateur linéaire sans biais pour le vecteur des paramètres β du modèle(1.2) avec $C = (X^tX)^{-1}X^t + D$.
où D est une matrice de dimension $(p + 1) \times n$ et on supposera que X est

non aléatoire. On a donc

$$\begin{aligned}
E[\hat{\beta}] &= E[CY] \\
&= E[((X^tX)^{-1}X^t + D)Y] \\
&= E[((X^tX)^{-1}X^t + D)(X\beta + \epsilon)] \\
&= ((X^tX)^{-1}X^t + D)X\beta + ((X^tX)^{-1}X^t + D)E[\epsilon] \\
&= (X^tX)^{-1}X^tX\beta + DX\beta \\
&= (1_{(p+1)\times(p+1)} + DX)\beta
\end{aligned}$$

Pour que β' soit un estimateur sans biais, il faut que DX=0. Ainsi

$$\begin{aligned}
\Sigma(\hat{\beta}) &= \Sigma(CY) \\
&= C\Sigma(Y)C^t \\
&= C\sigma(X\beta + \epsilon)C^t \\
&= C\Sigma[\epsilon]C^t \\
&= \sigma^2CC^t \\
&= \sigma^2((X^tX)^{-1}X^t + D)(X(X^tX)^{-1} + D^t) \\
&= \sigma^2[(X^tX)^{-1} + (X^tX)^{-1}X^tD^t + DX(X^tX)^{-1} + DD^t] \\
&= \sigma^2[(X^tX)^{-1} + (X^tX)^{-1}(DX)^t + DX(X^tX)^{-1} + DD^t] \\
&= \sigma^2[(X^tX)^{-1} + DD^t] \\
&= \Sigma(\beta') + \sigma^2DD^t
\end{aligned}$$

Comme les termes diagonaux de la matrice σ^2DD^t sont positifs ou nuls, on trouve ainsi que$[\Sigma(\beta')]_{i.i} \geq [\Sigma(\hat{\beta})]_{i.i}$, pour tout i=0,...,p.

Exemple

Notre objectif est de trouver s'il existe une relation entre la taille et le poids d'un individu.
Ce modèle a mis en évidence une relation linéaire entre ces deux variables. Nous disposons pour faire cette étude d'un jeu de données obtenu à partir d'un échantillon de 60 étudiantes de la résidence Ziani Lounes de l'umbb et utilisant un modèle de régression linéaire simple. (cas particulière de régression linéaire multiple).

Nuage de points

Un nuage de points est une représentation graphique de données, éventuellement interprétable par l'identification de relation, le nuage de points utilisé pour présenter la mesure de deux ou plusieurs variables liées. Le nuage de points est particulièrement utile lorsque les valeurs de variables sur l'axe des y dépendent des valeurs de la variable de l'axe des x. Dans un nuage de points les points sont placé sans être reliés. La tendance qui en résulte indique le type de la force de la relation entre deux ou plusieurs variables.

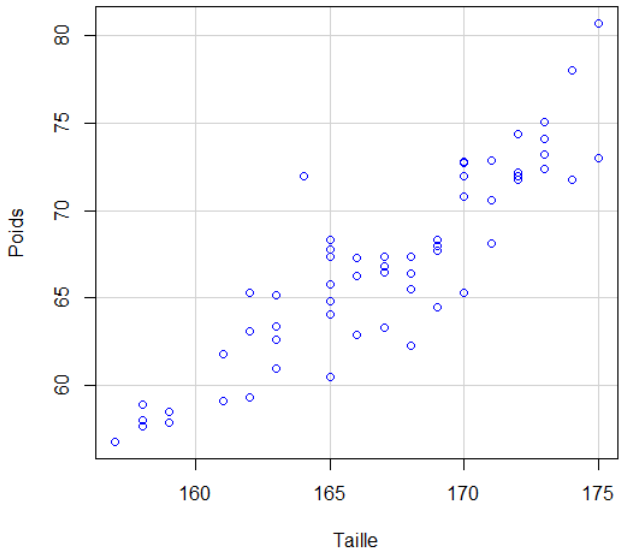


FIGURE 1.1 – le nuage de points des poids en fonction de taille des étudiantes

A partir de ce nuage, une simple régression linéaire semble appropriée pour expliquer cette relation.

Résumé le modèle :

coefficient	t value	pr(>| t |)	R^2
-103.32419	-9.05	1.1e-12	0.7934
1.01993	14.92	$<$ 2e-16	

A l'aide de logiciel **R** nous avons estimé les coefficients de modèle de régression linéaire simple, en utilisant la méthode MCO.
La droite de régression est définie par l'équation suivante :
Poids = -103.32419 + 1.01993 Taille .

La droite d'ajustement représentée sur le même plan que le nuage de points est donnée par :

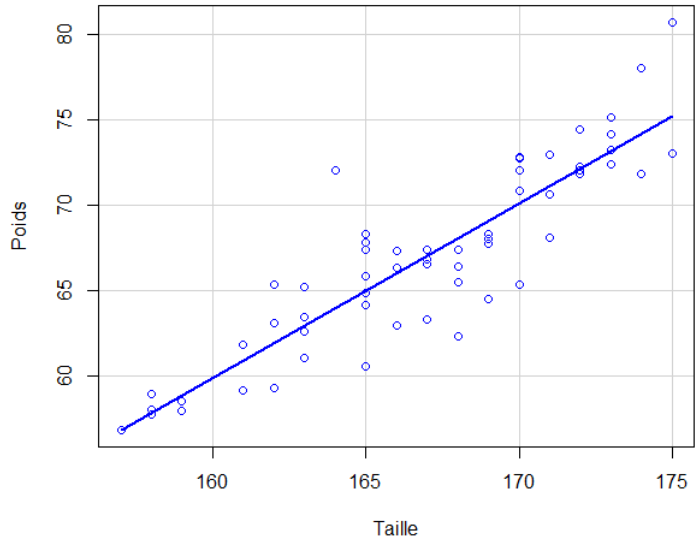


FIGURE 1.2 – Nuage de points avec la droite d'ajustement

On remarque que la droite des moindres carrés ordinaire est proche de la majorité des points de nuage. Le coefficient de détermination $R^2 = 0.7934$ est proche de 1 ce qui confirme une bonne relation linéaire entre les deux variables le poids et la taille.

Test sur les résidus :

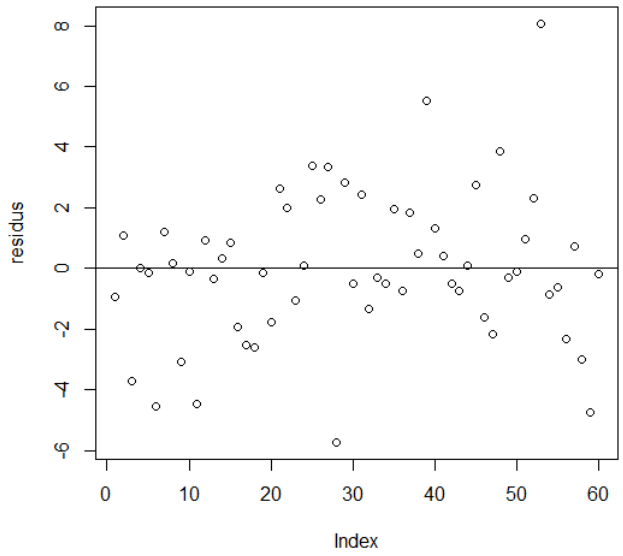


FIGURE 1.3 – Un graphique des valeurs résiduelles par rapport aux valeurs de prédicteur

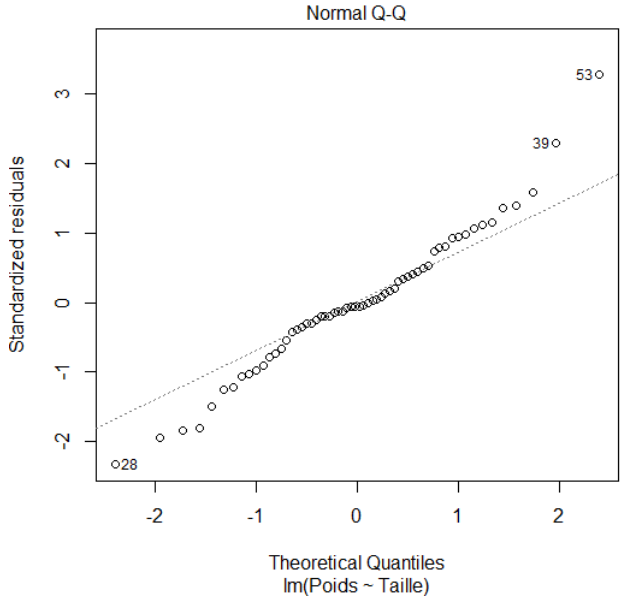


FIGURE 1.4 – Un graphique pour vérifier la normalité des résidus

On remarque que tous les points sont présenter sur la droite ce qui prouve que les résidus sont de loi normal.

À partir du tracé et du test de des hypothèses suivantes :
Hypothèse nulle - L'homoscédasticité est présente : les résidus sont distribués avec des variances égales.
Contre l'hypoyhèse alternative l'hétéroscédasticité est présente : les résidus ne sont pas distribués avec des variances égales.

```
        studentized Breusch-Pagan test

data:  model
BP = 0.51977, df = 1, p-value = 0.4709
```

Nous obtenons une valeur de p de 0.4709 supérieur à 0.05, nous acceptons donc l'hypothèse nulle et concluons que l'homoscédasticité est présente dans le modèle.

1.3 Point de rupture et efficacité relative d'un estimateur

1.3.1 Point de rupture

Le point de rupture est une notion très importante lorsqu'on aborde le concept de robustesse pour les estimateurs. Il est défini comme étant la plus petite fraction de contamination dans un échantillon qui déstabilise complètement un estimateur. Une définition simplifiée du point de rupture et permettant de travailler avec des échantillons finis a été introduit par Donoho et Huber en en 1983
Définition 1.1 (Point de rupture)[Donoho et Huber[1983]].
Soit $E = (x_i, y_i)_{i=1}^n$ un échantillon de n points. Considérons un éstimateur $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$ pour les paramètres de régression $\beta = (\beta_0, \beta_1, ..., \beta_p)$ et $\lambda(m, \hat{\beta}, E) = \sup \parallel \hat{\beta}(X') - \hat{\beta}(X) \parallel_2$ pour tous échantillons corrompus X' dans lequel m point de l'échantillon original X ont été remplacés par des valeurs arbitraires. Alors, on définit le point de ruptue de l'estimateur $\hat{\beta}$ par :

$$\theta_n = min_m \left\{ \frac{m}{n}; \lambda(m, \hat{\beta}, E) = \infty \right\}$$

Pour l'estimateur des MCO celui ci utilise chaque point de l'échantillon pour le calcul de l'estimateur $\hat{\beta}$. Ainsi comme on peut le voir dans la figure 1.5, le remplacement d'un seul point de l'échantillon original par un point

aberrant modifie l'estimateur. Alors si on remplace un seul point de l'échantillon original par un point arbitrairement loin de cet échantillon nous donne $\lambda(1, \hat{\beta}, E) = \infty$.

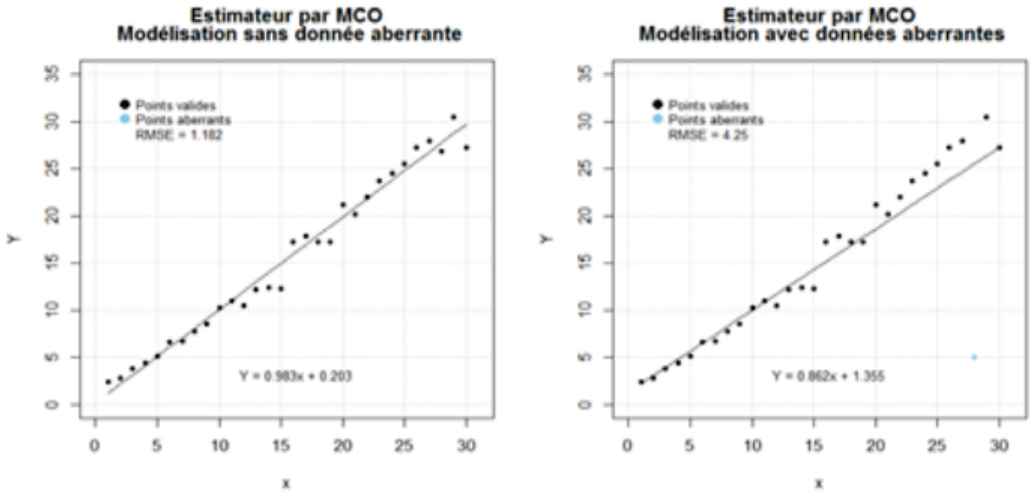


FIGURE 1.5 – Simulation de données avec l'insertion d'un point arbitrairement loin de cet échantillon (graphique de droite)pour illustrer le point de rupture de l'estimation des MCO

Le point de rupture pour l'estimateur de MCO est donc $\theta_n = \frac{1}{n}$, car l'insertion d'un seul point arbitrairement loin de cet échantillon modifie notre estimation des paramètres.
Lorsque n devient très grand, le point de rupture tend vers 0%

1.3.2 Efficacité relative (EFF)

L'efficacité relative d'un estimateur $\hat{\beta}_2$ est le quotient entre son erreur quadratique moyenne et celle du meilleur estimateur possible $\hat{\beta}_1$ connu pour le paramètre qu'on cherche à estimer et qui dans le cas unidimensionnel s'écrit sous la forme :

$$Eff(\hat{\beta}_1, \hat{\beta}_2) = \frac{E[(\hat{\beta}_1 - \beta)^2]}{E[(\hat{\beta}_2 - \beta)^2]}$$

Dans le contexte de la régression linéaire multiple, l'estimateur des moindres carrés ordinaire est considéré comme étant **BLUE**(en vert du **théorème 4**).

Quand les hypothèses sont respectées, il offre une variance minimale, donc on le considère comme l'estimateur linéaire le plus efficace connu.

Exemple
Voici quelque exemples simulés utilisant l'estimateur obtenu par la méthode des MCO et qui contiennent des taux différents de données aberrantes.

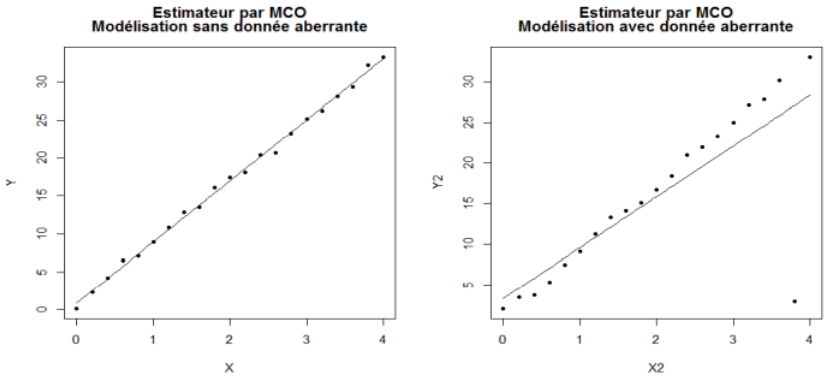


FIGURE 1.6 – Modélisation sans et avec une donnée aberrante

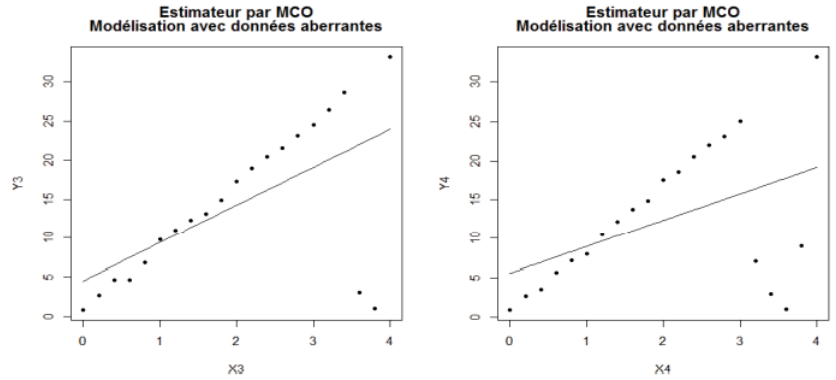


FIGURE 1.7 – Modélisation avec 10% et 20% de données aberrantes

On remarque que la présence d'une seule donnée aberrante peut affecter considérablement la régression linéaire et a chaque fois on augment le nombre des donnés aberrantes les résultats peuvent affecter significativement. C'est pourquoi nous aborderons la notion d'estimateur robuste dans le prochaine chapitre. Ce type d'estimation permettra la construction d'un modèle qui ne sera pas affecté par les données aberrantes.

En résume, dans ce chapitre nous avons d'abord rappelé quelques notion de la régression linéaire multiple. Nous avons ensuite présenté l'estimation des paramètres de régression par la méthode des MCO avec un exemple simple. Nous avons terminé ce chapitre en donnant la définition de point de rupture et d'efficacité relative.

Chapitre 2

Régression linéaire avec estimation robuste

Dans les statistiques robustes, la régression robuste est une forme d'analyse de régression conçue pour surmonter certaines limites des méthodes paramétriques et non paramétriques traditionnelles. L'analyse de régression cherche à trouver la relation entre une ou plusieurs variables indépendantes et une variable dépendante. Certaines méthodes de régression largement utilisées, telles que les moindres carrés ordinaires, ont des propriétés favorables si leurs hypothèses sous-jacentes sont vraies, mais peuvent donner des résultats trompeurs si ces hypothèses ne sont pas vraies, on dit donc que les moindres carrés ordinaires ne sont pas robustes aux violations de ses hypothèses. Les méthodes de régression robustes sont conçues pour ne pas être trop affectées par les violations des hypothèses par le processus de génération de données sous-jacent.
La régression linéaire robuste est moins sensible aux valeurs aberrantes que la régression linéaire standard. La régression linéaire standard utilise l'ajustement des moindres carrés ordinaires pour calculer les paramètres du modèle qui relient les données de réponse aux données de prédicteur avec un ou plusieurs coefficients. Par conséquent, les valeurs aberrantes ont une grande influence sur l'ajustement, car la quadrature des résidus amplifie les effets de ces points de données extrêmes. Les modèles qui utilisent la régression linéaire standard, reposent sur certaines hypothèses, telles qu'une distribution normale des erreurs. Si la distribution des erreurs est asymétrique ou sujette aux valeurs aberrantes, les hypothèses du modèle sont invalidées et les estimations des paramètres, les intervalles de confiance et les autres statistiques calculées perdent leur fiabilité.

La régression robuste basée sur certains méthodes d'estimation telle que : Moindres carrés pondérés, M-estimation, MM-estimation, t-régression et régression quantile.

2.1 Notion de base sur les valeurs extrêmes

2.1.1 Définitions :

Définition 1 :
Les valeurs extrêmes sont des points de données éloignés des autres points de données. En d'autres termes, ce sont des valeurs inhabituelles dans un ensemble de données.

Définition 2 :
Une valeur extrême est une valeur très petite ou très grande dans une distribution de probabilité. Ces valeurs extrêmes se trouvent dans les queues d'une distribution de probabilité (c'est-à-dire les extrémités de la distribution).
Certains auteurs utilisent le terme "valeur extrême" comme un autre nom pour la valeur minimale et/ou la valeur maximale d'une fonction (c'est à- dire le plus petit et/ou le plus grand nombre dans l'ensemble), et d'autres l'utilisent comme synonyme d'une valeur aberrante. Cependant, dans la plupart des cas, lorsque les gens parlent de valeurs extrêmes, ils parlent généralement de valeurs associées à **la théorie des valeurs extrêmes**.

2.1.2 Théorie des valeurs extrêmes

la théorie des valeurs extrêmes a été lancé par Fisher et Tippett (1928) ; Fréchet (1927) ; Gnedenko (1943) ; Gumbel (1958),.. la théorie des valeurs extrêmes propose un cadre statistique pour traiter les événements rares : estimer un quantile extrême et estimer la probabilité d'occurrence d'un événement qui n'a pas été observé.

La théorie des valeurs extrêmes ou l'analyse des valeurs extrêmes (EVA) est une branche de la statistique traitant des écarts extrêmes par rapport à la médiane des distributions de probabilité. Il cherche à évaluer, à partir d'un échantillon ordonné d'une variable aléatoire, la probabilité d'événements plus extrêmes que ceux précédemment observés. L'analyse des valeurs extrêmes est largement utilisée dans de nombreuses disciplines :

En hydrologie : crues consécutives à des pluies torrentielles : aux Pays-Bas, digues menacées, par l'effet conjoint des grandes marées et des conditions climatiques en Mer du Nord (Inondation de 1953).

En assurance : survenue des sinistres d'intensité exceptionnelle (ouragan Katrina en 2005, importants incendies en risques industriels, sinistres graves en responsabilité civile auto-mobile) qui peuvent avoir des conséquences négatives sur les résultats et la solvabilité des organismes d'assurance.

En Finance : fortes variations du cours d'actifs financiers, gestions du risque opérationnel des banques (la crise f des années 2000).

En Climatologie : étude des évènements climatiques extrêmes (précipitations, températures, chutes de neige), modélisation des grands feux de forêt.

En météorologie [Coles etWalshaw, 1994 ; Smith, 2001] où l'étude de la vitesse du vent, par exemple, permet d'évaluer le degré de résistance des matériaux face à la pression exercée par le vent (au cours d'une tempête par exemple) sur les bâtiments ou les structures de génie civil.

2.1.3 les causes des valeurs extrêmes

- Erreurs humaine, ex : Erreurs de saisie.
- Erreurs d'instrument, ex : Erreurs de mesure.
- Erreurs de traitement des données, ex : Manipulation des données.
- Erreurs d'échantillonnage, ex : Extraction des données de mauvaises sources

2.1.4 L'importance des valeurs extrêmes

De nombreuses analyses statistiques étudient le corps principal des données et examinent son comportement en termes de moyennes. Dans de nombreux cas, cependant, les valeurs extrêmes des données sont plus intéressantes. Par exemple, si nous étudions le niveau d'une rivière au fil du temps, les seules valeurs qui nous intéressent vraiment sont celles qui sont vraiment élevées ou vraiment basses. S'ils sont trop élevés, nous pourrions avoir des inondations et s'ils sont trop bas, la rivière pourrait s'assécher. Un autre exemple est celui des températures sur les banquises. Trop haut et on aura de la fonte, trop bas et les étagères grossiront.

Les valeurs extrêmes peuvent avoir un impact important sur les analyses statistiques et fausser les résultats de tout test d'hypothèse s'ils sont inexacts. Ces valeurs extrêmes peuvent également avoir un impact sur la puissance statistique, ce qui rend difficile la détection d'un véritable effet s'il y en a un.

2.1.5 Détection des valeurs extrêmes

il existe plusieurs façons pour détecter les valeurs extrêmes. Toutes ces méthodes utilisent différentes approches pour trouver des valeurs inhabituelles par rapport au reste de l'ensemble de données. tel que :

Inspection visuelle :

— Boite à moustache.

— Histogramme et nuage de points.

— Méthode de tri.

Méthodes statistiques :

— Méthode de l'intervalle interquartile.

— Scores z.

— Fixer des seuils pour identifier les valeurs extrêmes.

— Transformer la variable pour induire la normalité.

2.1.6 Traitement des valeurs extrêmes

Trois principales méthodes sont utilisées pour gérer les valeurs extrêmes, hormis leur suppression des données :
1- Réduire la pondération des valeurs extrêmes(pondération de césure).
2- Changer les valeurs des valeurs extrêmes (Winsorisation, Césure, imputation par exemple via la régression quantile).

winsorisation est une méthode qui remplace initialement les valeurs les plus petites et les plus grandes par les observations les plus proches d'elles. Ceci est fait pour limiter l'effet des valeurs extrêmes sur le calcul.
La technique Winsorize a été introduite pour la première fois par Dixon (1960), qui l'a attribuée à Charles P. Winsor.
3- Utiliser les techniques d'estimation robustes(M-estimation).

2.2 Sensibilité aux valeurs extrêmes :

2.2.1 Mesures de tendance centrale :

Une mesure de tendance centrale vise à identifier le centre d'une distribution, la moyenne et la médiane en sont deux exemples bien connus. Le centre défini par la moyenne équilibre la somme des écarts de part et d'autre de la valeur moyenne, tandis que celui défini par la médiane équilibre le nombre d'observations de part et d'autre. Pour cette raison, l'ajout d'une valeur extrême à un échantillon peut affecter fortement sa moyenne, mais très peu sa médiane.
Par exemple, prenons un échantillon de des 10 valeurs suivantes :
18 29 30 40 43 44 48 49 56 83
La moyenne est de 44 et la médiane égale 43.5 sont approximativement égales :
Si on ajoutait la valeur 580 à cet échantillon, la nouvelle médiane serait de 44, tandis que la moyenne serait d'environ 93 et ne représenterait plus une valeur "typique" de l'échantillon.

2.2.2 Point de rupture :

Le point de rupture (breakdown point) d'un estimateur est défini par la question suivante : combien de valeurs extrêmes, si elles sont assez extrêmes, peuvent affecter sans limite la valeur de l'estimé ? On l'exprime généralement comme une fraction du nombre d'observations.
Avec n observations, la moyenne a un point de rupture de $\frac{1}{n}$, car une seule observation extrême suffit à l'entraîner vers des valeurs extrêmes. Dans l'exemple précédent, si on augmentait la valeur extrême ajoutée, la moyenne pourrait augmenter sans limite.
Dans le cas de la médiane, elle réagirait de la même façon à toute valeur extrême ajoutée d'un côté de la distribution, peu importe la magnitude de cette valeur extrême (la nouvelle médiane serait de 44 peu importe si la donnée ajoutée était de 100 ou 300 ou 1000). Pour faire augmenter la médiane sans limite, c'est toute la moitié supérieure du jeu de données qu'il faudrait faire augmenter, la médiane a donc un point de rupture de 0.5.

2.2.3 Précision des estimés et valeurs extrêmes :

Nous avons vu que la valeur de la moyenne est sensible à l'ajout de valeurs extrêmes d'un côté de la distribution (cas asymétrique). Si les valeurs extrêmes apparaissent de façon symétrique de part et d'autre de la moyenne, sa valeur reste inchangée. Cependant, puisque l'écart-type de la distribution est aussi sensible aux valeurs extrêmes, la précision avec laquelle on peut estimer de la moyenne est affectée.
Dans le graphique ci-dessous, la courbe verte représente une distribution normale centrée réduite, $y \sim N(0, 1)$. La courbe orange représente le mélange de deux distributions : 95% des observations proviennent de la distribution $N(0, 1)$ et 5% proviennent d'une distribution avec un écart-type plus grand : $N(0, 5)$. Ce mélange représente le cas où la plupart des observations suivent une distribution normale, sauf une petite fraction dont les valeurs sont plus extrêmes qu'attendu. Sur une échelle linéaire de la densité de probabilité $f(\mid y \mid)$ (à gauche), les deux distributions apparaissent très semblables. Sur une échelle logarithmique (à droite), on voit clairement que les valeurs extrêmes sont beaucoup plus probables pour la distribution de mélange (ex : environ 30 fois plus probable d'obtenir y=-4 ou y=4).

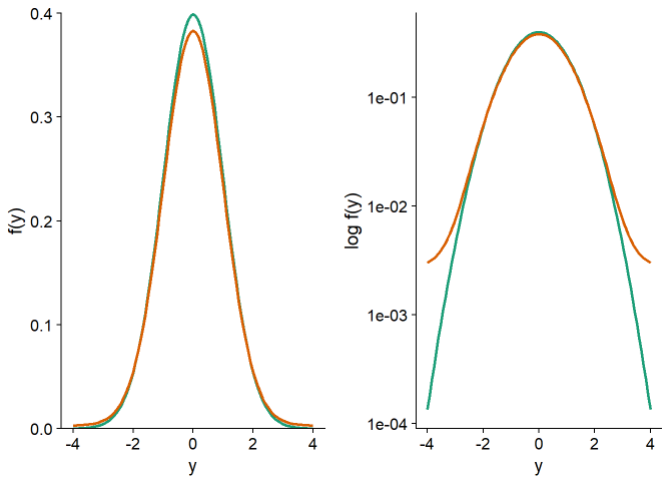


FIGURE 2.1 – Graphique d'une distribution normale réduite

Comparons maintenant les erreurs-types pour la moyenne et la médiane de ces distributions. Pour ce faire, nous simulons 1000 échantillons de 100 observations de chacune des deux distributions, dans le cas de la distribution de mélange, l'écart-type est de 1 pour les 95 premières observations et de 5 pour les 5 dernières.

Pour la distribution normale, l'erreur-type de la moyenne obtenue par simulation est d'environ 0.10, tel que prévu par la formule $\frac{\sigma}{\sqrt{n}}$. Pour la distribution de mélange, l'erreur-type est environ 50% plus élevée (0.15).
Quant à la médiane, son erreur-type est supérieure à celle de la moyenne pour la distribution normale, mais étant moins sensible (plus robuste) aux valeurs extrêmes, elle est estimée plus précisément pour la distribution de mélange.
Supposons que nous comparons deux groupes entre lesquels la moyenne et la médiane d'une variable réponse diffèrent, si la distribution de la variable est symétrique, alors la moyenne est identique à la médiane pour chaque groupe. Si la variable suit une distribution normale, il est plus facile de détecter une différence entre les moyennes qu'entre les médianes ; un test basé sur les moyennes, comme le test t, a une plus grande puissance. En présence de valeurs extrêmes, l'erreur-type de la moyenne augmente et un test basé sur la différence entre médianes pourrait être plus puissant.
Les M-estimateurs, que nous verrons plus loin dans un contexte de régression, sont des mesures de tendance centrale qui font un compromis entre l'efficacité de la moyenne pour une distribution normale et la robustesse aux valeurs extrêmes de la médiane. Lorsque la distribution est normale, la précision de ces estimateurs s'approche de celle de la moyenne, mais ils ont un point de rupture plus élevé et peuvent donc mieux conserver leur précision en présence de plusieurs valeurs extrêmes.

2.2.4 Valeurs extrêmes et régression :

Dans une régression linéaire simple, la moyenne de la réponse y correspond à une fonction linéaire du prédicteur x, tandis que la variation aléatoire autour de cette moyenne est représentée par un résidu ϵ qui suit une distribution normale

$$y = \beta_0 + \beta_1 x + \epsilon \qquad avec \qquad \epsilon \sim N(0, \sigma^2)$$

Les concepts présentés ici s'appliquent autant à une régression linéaire multiple, mais le cas d'un prédicteur unique est plus simple à illustrer.
Les coefficients β_0 et β_1 sont estimés par la méthode des moindres carrés ordinaire, c'est-à-dire qu'on vise à minimiser la somme des résidus au carré pour les n observations :

$$\sum_{i=1}^{n} \hat{\epsilon_i}^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x\right)^2$$

Ici $\hat{\epsilon}$ est l'estimé de la valeur du résidu i en fonction de la valeur estimée des coefficients.
Pour une régression linéaire, l'influence d'une observation sur l'estimé des coefficients dépend de deux facteurs : la taille du résidu de cette observation, $\hat{\epsilon}$, ainsi que le positionnement de x_i. Pour un même x_i, les résidus $\hat{\epsilon}$ plus extrêmes ont une plus grande influence ; pour une même taille de résidu, ceux correspondant à une valeur de x_i plus extrême ont aussi une plus grande influence, comme le montre le graphique ci-dessous.

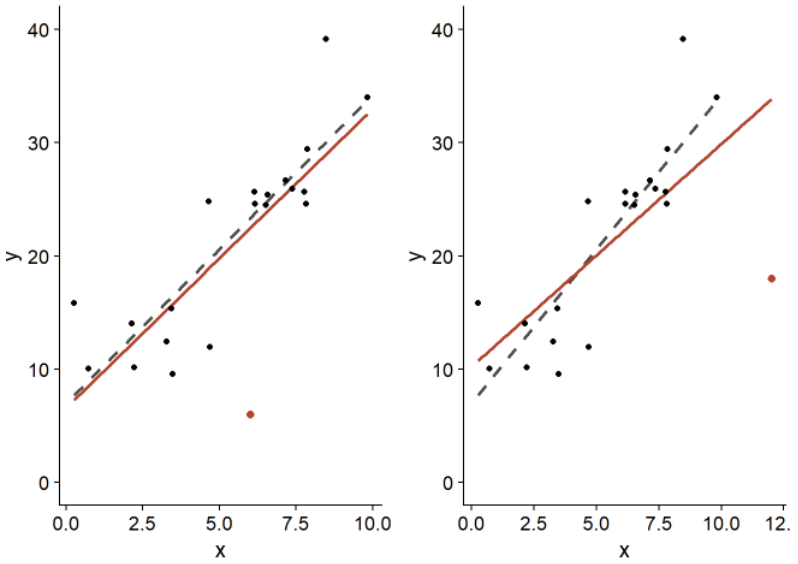


FIGURE 2.2 – L'influence d'une valeur extrême sur la droite d'ajustement

Dans les deux cas, le point en orange possède le même résidu, soit ϵ=-20. Cependant, celui placé près de la limite supérieure de x_i (panneau droit) affecte davantage l'estimé de la pente $\hat{\beta}$ (ligne orange = avec ce point ; ligne grise pointillée = sans ce point).
Les résidus situés près des extrêmes de x_i exercent un grand effet de levier (leverage) sur la droite de régression. Puisque la droite de régression passe toujours par le centre de gravité du nuage de points, $(\bar{x}, \bar{y})$, un résidu situé plus loin du centre fait davantage "pivoter" la droite dans sa direction.
La distance de Cook mesure l'influence d'un point sur l'ajustement du modèle de régression, elle tient compte à la fois de la magnitude de $\hat{\epsilon}$ et de son

effet de levier en fonction de la position en x_i. Généralement, une distance de Cook supérieure à 1 indique une observation ayant une grande influence.

Exemple
Le jeu de données Animals2 inclus avec le package robustbase contient des mesures de la masse corporelle (body, en kg) et de la masse du cerveau (brain, en g) pour 65 espèces animales.

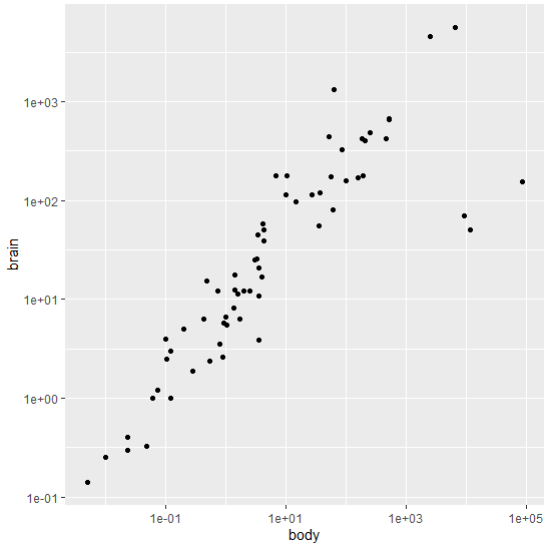


FIGURE 2.3 – Nuage de points pour le jeu de données animals2

Tous les animaux dans ce jeu de données sont des mammifères, excepté trois qui sont des dinosaures. Il s'agit des trois observations avec la masse corporelle la plus grande, mais dont la masse du cerveau se retrouve sous la tendance générale.

Dans une analyse statistique, les valeurs aberrantes (outliers) peuvent être exclues si nous avons des informations indépendantes indiquant que les mesures sont incorrectes, ou qu'elles proviennent d'une population différente du reste des observations. Puisqu'il est raisonnable de croire que la relation allométrique diffère entre les mammifères et les dinosaures, il serait justifié d'exclure ces derniers avant d'effectuer la régression.
Une régression linéaire basée sur l'ensemble des données donne une pente de 0.59 pour log(brain) en fonction de log(body).

Dans les graphiques de diagnostic d'une régression, R indique automatiquement les numéros ou noms des rangées correspondant aux valeurs ex-

trêmes. Dans ce cas-ci, chaque rangée du jeu de données est identifiée du nom de l'animal.

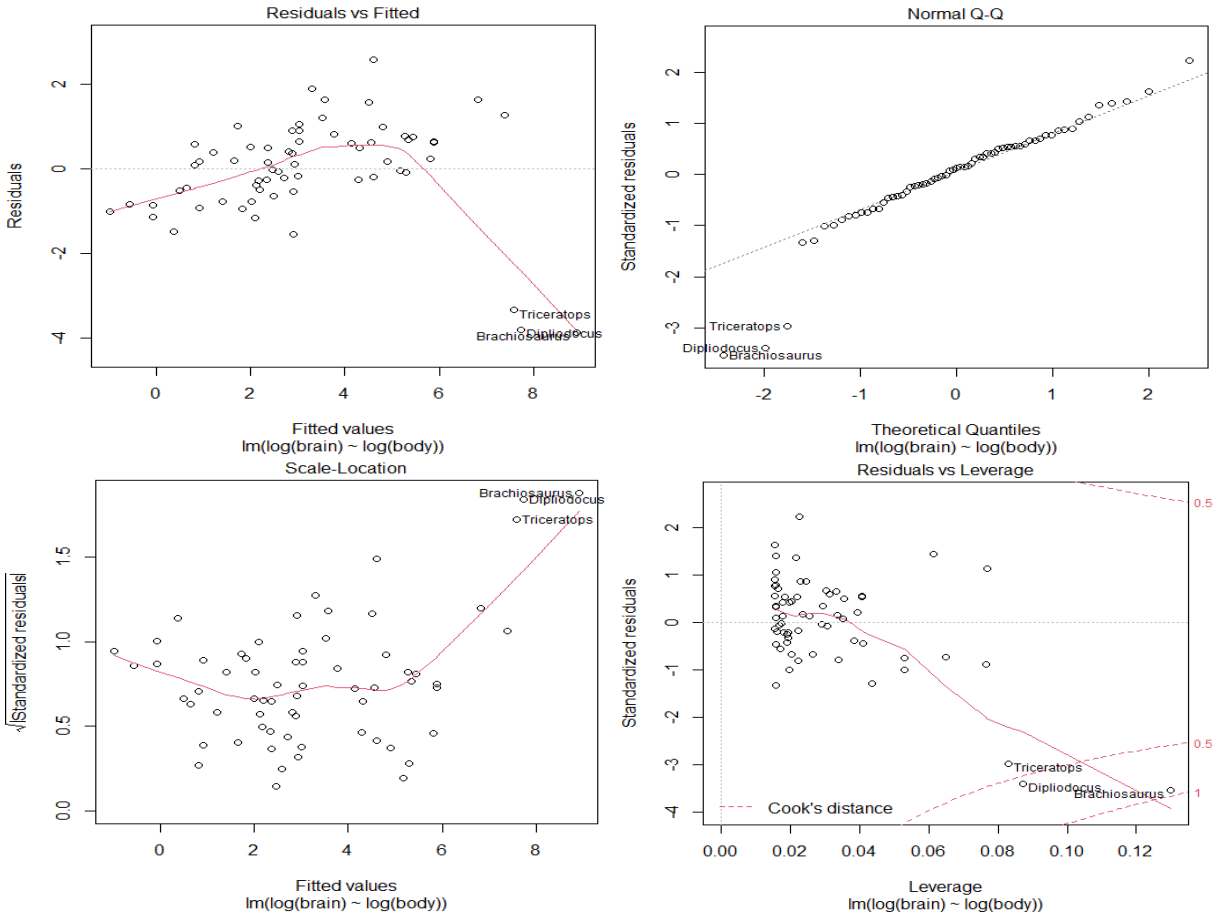


FIGURE 2.4 – Graphique des résidus

Le 4*eme* graphique, Residuals vs. Leverage, permet d'identifier les points avec une forte influence. Les lignes pointillées démarquent les seuils de 0.5 et 1 pour la distance de Cook. Ici, aucun des trois points extrêmes ne dépasse 1, mais leur influence est supérieure de beaucoup à celle du reste des points. En comparaison, la régression ignorant les trois données extrêmes donne une pente de 0.75.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13479    0.09604   22.23   <2e-16 ***
log(body)    0.75169    0.02846   26.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared:  0.9208, Adjusted R-squared:  0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

Le résultat des deux régressions est illustré dans le graphique suivant (courbe orange : avec dinosaures, courbe grise pointillée : sans dinosaures, région ombragée : intervalle de confiance).

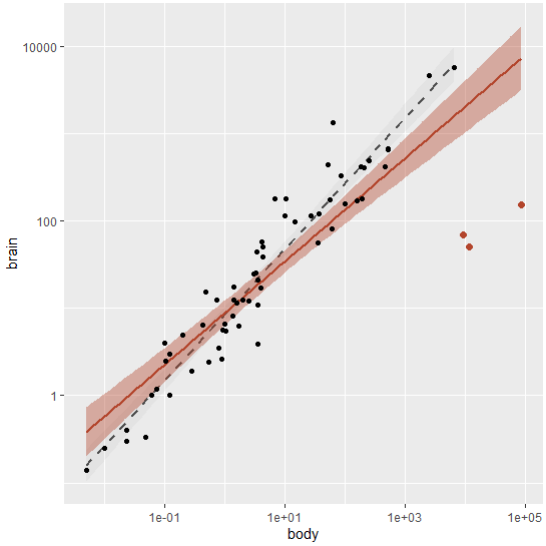


FIGURE 2.5 – Graphe d'intervalle de confiance

Dans les prochaines sections, nous verrons comment réduire l'influence des valeurs extrêmes sans les exclure complètement de l'analyse.

2.3 Méthodes d'estimations robuste :

2.3.1 Moindres carrés pondérés

La méthode des moindres carrés ordinaires suppose qu'il existe une variance constante dans les erreurs (ce que l'on appelle **l'homoscédasticité**).

La méthode des moindres carrés pondérés peut être utilisée lorsque l'hypothèse des moindres carrés ordinaires de la variance constante des erreurs est violée (appelée **hétéroscédasticité**). Le modèle considéré est

$$Y = X\beta + \varepsilon^*$$

où ε^* est supposée être multivariée normalement distribuée avec un vecteur moyen 0 et une matrice de variance-covariance non constante :

$$\begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

Si nous définissons l'inverse de chaque variance σ_i^2, comme le poids $w_i = \frac{1}{\sigma_i^2}$ alors soit la matrice W une matrice diagonale contenant ces poids :

$$\begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}$$

L'estimation des moindres carrés pondérés est alors

$$\hat{\beta}_{WLS} = argmin_s \sum_{i=1}^{N} \varepsilon_i^{*2} = (X^T W X)^{-1} X^T W Y$$

Avec ce réglage, nous pouvons faire quelques observations :

— Étant donné que chaque poids est inversement proportionnel à la variance de l'erreur, il reflète l'information contenue dans cette observation. Ainsi, une observation avec une petite variance d'erreur a un poids important car elle contient relativement plus d'informations qu'une observation avec une grande variance d'erreur (petit poids).

— Les poids doivent être connus (ou plus généralement estimés) jusqu'à une constante de proportionnalité.

2.3.2 Exemple : Ensemble de données d'apprentissage assisté par ordinateur

L'apprentissage assisté par ordinateur de nouvelles données ont été recueillies à partir d'une étude sur l'apprentissage assisté par ordinateur auprès de n = 12 élèves.

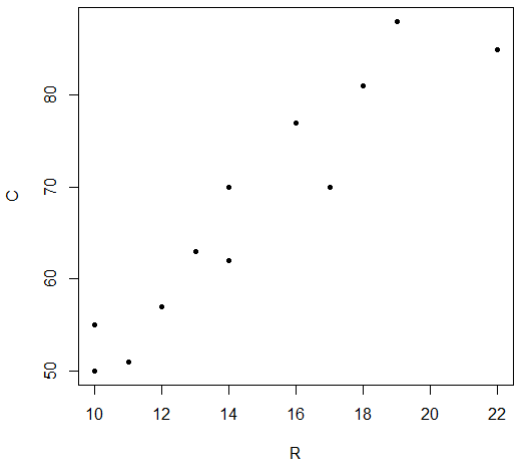


FIGURE 2.6 – Nuage de points des coût en fonction des réponses

où
C= Le coût
R=réponses

A partir de ce nuage de points, une simple régression linéaire semble appropriée pour expliquer cette relation.
Une ligne des moindres carrés ordinaires est d'abord ajustée à ces données.

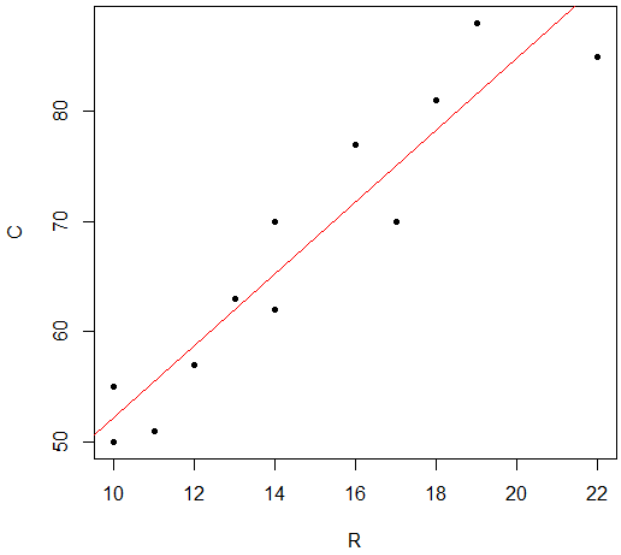


FIGURE 2.7 – Nuage de points avec la droite d'ajustement

On trouvent ci-dessous le résumé de l'ajustement de régression linéaire simple pour ces données.

Résumé du modèle :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.4727     5.5162   3.530  0.00545 **
R             3.2689     0.3651   8.955 4.33e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.598 on 10 degrees of freedom
Multiple R-squared:  0.8891,    Adjusted R-squared:  0.878
F-statistic: 80.19 on 1 and 10 DF,  p-value: 4.33e-06
```

Test d'hétéroscédasticité

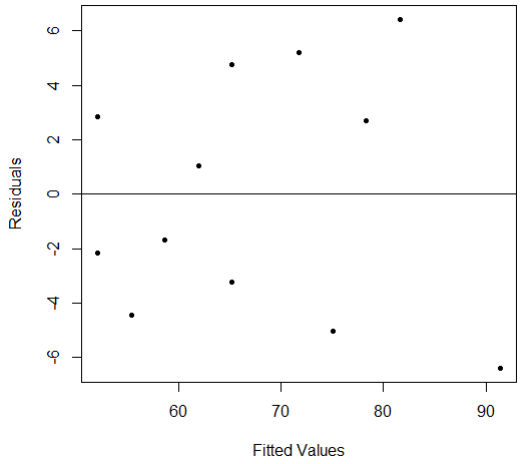


FIGURE 2.8 – Un graphique des valeurs résiduelles par rapport aux valeurs de prédicteur

```
data:  model
BP = 6.5323, df = 1, p-value = 0.01059
```

À partir du tracé et du test ci-dessus, les résidus ne sont pas distribués avec des variances égales donc L'hétéroscédasticité est présente. Nous obtenons une valeur de p= 0,01059 qui inférieur à 0.05.

Régression des moindres carrés pondérés

Le modèle des moindres carrés pondéré ci-dessus conclut que l'estimation du coefficient "Coût" a changé, et l'ajustement du modèle est effectivement amélioré. Le modèle des moindres carrés pondérés donne un écart-type résiduel (RSE) de 1.159, ce qui est bien meilleur que celui d'un modèle de régression linéaire simple qui est de 4.598. Ce qui implique que les valeurs prédites sont beaucoup plus proches des valeurs réelles lorsqu'elles sont ajustées sur un modèle des moindres carrés pondérés par rapport à un modèle de régression simple. La valeur R au carré ne montre pas beaucoup de différence, dans le modèle pondéré des moindres carrés 0.8951 par rapport à la régression linéaire simple 0.8891. Ces deux changements dans les valeurs des mesures de performance dans les deux modèles concluent que les moindres carrés pondérés sont meilleurs par rapport au modèle de régression linéaire simple.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.3006     4.8277   3.584  0.00498 **
data$R        3.4211     0.3703   9.238 3.27e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.159 on 10 degrees of freedom
Multiple R-squared:  0.8951,    Adjusted R-squared:  0.8846
F-statistic: 85.35 on 1 and 10 DF,  p-value: 3.269e-06
```

2.4 M-estimation des paramètres du modèle

Le M-estimation est une généralisation de l'approche de maximum de vraisemblance.
Elle repose sur le principe de minimiser une fonction de coût qui permettra de pénaliser les résidus ϵ_i les plus grands.La fonction de coût s'écrit :

$$C(\epsilon) = \sum_{i=1}^{n} \rho(\epsilon_i) \tag{2.1}$$

où la fonction ρ est une fonction continue et symétrique appelée "fonction objective".
Cette fonction satisfait les conditions dans la définition 2.2 ci-dessous donnée par Maronna, Martin et Yohai. Pour cela, nous allons introduire, dans la définition suivante, la notion de ψ-fonction.

Définition 2.1 Une ψ-fonction est une fonction continue par morceaux définie dans R telle que :

1. ψ est symétrique par rapport à : $\psi(-x) = -\psi(x)$ pour tout x dans R.
2. $\psi(x) \leq 0$ pour $x \leq 0$ et $\psi(x) > 0$ pour $0 < x < x_r$, où $x_r = sup\{x : \psi(x) > 0\}(x_r > 0)$.
3. $\psi'(0) = 1$.

Remarque :
Le point 3 n'est pas strictement requis, mais on l'utilise pour la normalisation dans les cas où ψ est continue en $x = 0$. Il en découle aussi, du point 1, que

$\psi(0)$=0 et on impose que $\psi(0)$=0 pour tous les cas où ψ est discontinue en $x = 0$.

Définition 2.2 : Une ρ fonction, au point x, est représentée par l'intégrale d'une ψ-fonction sur [0,x].

$$\rho(x) = \int_0^x \psi(u)du$$

D'aprés la définition 2.1 et 2.2 on voit que $\rho(0)= 0$, que ρ est une fonction paire et que $\rho'(x) = \psi$. Dans la littérature, plusieurs fonctions ρ ont été proposées. Elles dépendent de certaines constantes qui permettent d'augmenter la robustesse des estimateurs lorsqu'il y a présence de données aberrantes, mais au déteriment de leur l'éfficacité. Les deux fonctions ρ les plus utilisées sont celles proposées par Huber et par Tukey.

Exemple 1 :
La fonction ρ proposée par "Huber" est définie pour une constant $c > 0$ par :

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{si } |x| \leq c \\ c(|x| - \frac{c}{2}) & \text{si } |x| > c \end{cases}$$

La dérivée associée à cette fonction est donnée par :

$$\psi(x) = \begin{cases} \frac{1}{x} & \text{si } |x| \leq c \\ csing(x) & \text{si } |x| > c \end{cases}$$

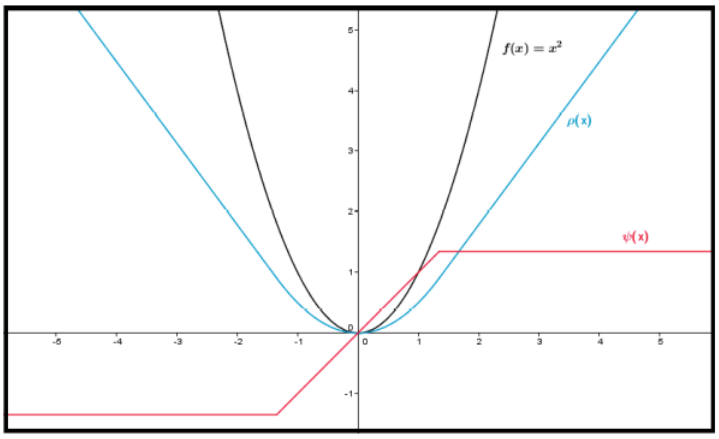


FIGURE 2.9 – Fonction de "Huber" utilisant un paramètre c=1.345

Exemple 2 :
La fonction ≪ Tukey's Biweight ≫ est définie pour une constante $c > 0$ par :

$$\rho(x) = \begin{cases} \frac{c}{6}[1-(1-(\frac{x}{c})^2)^3] & \text{si } |x| \leq c \\ \frac{c^2}{6} & \text{si } |x| > c \end{cases} \tag{2.2}$$

La dérivée associée à cette fonction est :

$$\psi(x) = \begin{cases} x[1-(\frac{x}{c})^2]^2 & \text{si } |x| \leq c \\ 0 & \text{si } |x| > c \end{cases}$$

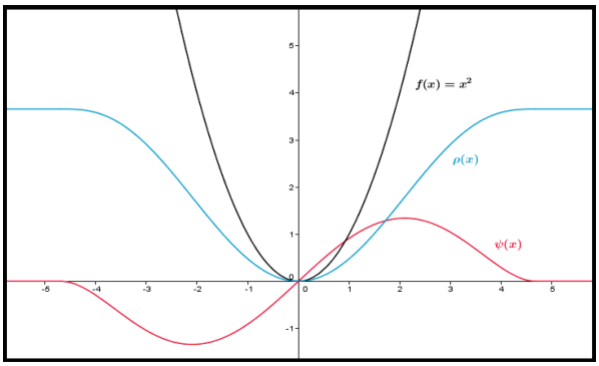


FIGURE 2.10 – Fonction de≪ Tukey's Biweight ≫ c=4.685

Dans les figures 2.9 et 2.10, les fonction proposées par Huber et Tukey ont été illustrés en utilisant des constante c égale à 1.345 et 4.685 respectivement. Nous verrons plus loin qu'utiliser ces deux valeurs comme constante c permet d'obtenir une efficacité relative de 95%. Aussi, on peut remarquer que les fonction proposées par Huber et Tukey permettent de pénaliser les plus grandes erreurs comparativement à la fonction objective utilisée par les MCO, c'est-à-dire $f(x) = x^2$, en leur donnant moins de poids, En effet , pour ces fonctions, on remarque à partir de x=c les deux fonctions sont croissantes, mais elles augmentent moins rapidement que la fonction $f(x) = x^2$.
Considérons le modèle de régression linéaire multiple (1.2). La M-estimation des paramètres du modèle β_j, j=0,...,p consiste à résoudre :

$$\hat{\beta}^M = argmin_\beta \sum_{i=1}^{n} \rho[y_i - (\beta_0 + \beta_1 x_{i,1} + ... + \beta_1 x_{i,p})]$$

L'estimateur obtenue par la méthode de MCO (1.3) est invariant par changement d'échelle. Par exemple, supposons qu'on cherche à modéliser le

poids des toro en fonction de leur âge, l'ajustement du modèle (c'est -à-dire le coefficient de détermination $\mathbf{R}^2$) ne sera pas changé par le fait d'utiliser des poids mesurés en kilogramme. Cette propriété bien pratique n'est pas de l'unité utilisée, on doit introduire un estimateur de la dispersion des résidus. Posons $\hat{\epsilon} = y_i - (\hat{\beta}_0^M + \hat{\beta}_1^M x_{i,1} + ... + \hat{\beta}_p^M x_{i,p})$, pour tous i=1,...,n.
Un estimateur robuste et très populaire pour la dispersion des résidus est le $\ll re - scalesMAD \gg$:

$$\hat{\sigma} = 1.4826MAD \tag{2.3}$$

où MAD signifie " Median absolute deviation " et est calculé de façons suivante :

$$MAD = médiane|\hat{\epsilon_i}|$$

Cet estimateur est peu sensible aux données aberrantes et possède un point de rupture de 50%, car il utilise la médiane au lieu de la moyenne. La multiplication par le facteur 1.4826 permet, lorsque l'échantillon est grand et que $\epsilon_i \sim N(0, \sigma^2)$, d'obtenir une estimation robuste pour l'écart-type .
Ainsi, on définit le M-estimateur $\hat{\beta}^M$ de la façon suivante :

$$\hat{\beta}^M = argmin_\beta \sum_{i=1}^{n} \rho\left(\frac{\epsilon_i}{\hat{\sigma}}\right) \tag{2.4}$$

$$= argmin_\beta \sum_{i=1}^{n} \rho\left(\frac{1}{\hat{\sigma}}[y_i - (\beta_0 + \beta_1 xi, 1 + ... + \beta_p xi, p)]\right)$$

On cherche alors à résoudre le problème d'optimisation :

$$\begin{aligned} L(\beta) &= argmin_\beta \sum_{i=1}^{n} \rho\left(\frac{\epsilon_i}{\hat{\sigma}}\right) \\ &= argmin_\beta \sum_{i=1}^{n} \rho\left(\frac{1}{\hat{\sigma}}[y_i - (\beta_0 + \beta_1 xi, 1 + ... + \beta_p xi, p)]\right) \end{aligned}$$

Le calcule différentiel nous permettra encore une fois de résoudre ce problème de minimisation et d'obtenir les estimateurs $\left\{\hat{\beta}_i^{\,M}\right\}_{j=0}^{p}$. On calculant les dérivées partielles de la fonction ρ par rapport à chaque paramètre β_i du modèle recherché et en posant $\psi(u) = \dfrac{d\rho(u)}{du}$, on se retrouve avec un système à p+1

équation :

$$\frac{\partial L(\beta)}{\partial \beta_0} \mid_{\beta=\hat{\beta}^M} = \sum_{i=1}^{n} \psi \left(\frac{y_i - (\hat{\beta}_0^{\,M} + \hat{\beta}_1^{\,M} x_{i,1} + \hat{\beta}_p^{\,M} x_{i,p})}{\hat{\sigma}} \right) \left(\frac{-1}{\hat{\sigma}} \right) = 0$$

$$\frac{\partial L(\beta)}{\partial \beta_1} \mid_{\beta=\hat{\beta}^M} = \sum_{i=1}^{n} \psi \left(\frac{y_i - (\hat{\beta}_0^{\,M} + \hat{\beta}_1^{\,M} x_{i,1} + \hat{\beta}_p^{\,M} x_{i,p})}{\hat{\sigma}} \right) \left(\frac{-x_{i,1}}{\hat{\sigma}} \right) = 0$$

$$\vdots$$

$$\frac{\partial L(\beta)}{\partial \beta_p} \mid_{\beta=\hat{\beta}^M} = \sum_{i=1}^{n} \psi \left(\frac{y_i - (\hat{\beta}_0^{\,M} + \hat{\beta}_1^{\,M} x_{i,1} + \hat{\beta}_p^{\,M} x_{i,p})}{\hat{\sigma}} \right) \left(\frac{-x_{i,p}}{\hat{\sigma}} \right) = 0 \quad (2.5)$$

Rappelons que $\hat{\epsilon} = y_i - (\hat{\beta}_0^{\,M} + \hat{\beta}_1^{\,M} x_{i,1} + \hat{\beta}_p^{\,M} x_{i,p})$. Pour résoudre ce système d'équation, Draper et Smith ont défini la fonction de poids suivantes :

$$w(u) = \frac{\psi(u)}{u}$$

Posons $w_i = w\left\{\frac{\hat{\epsilon_i}}{\hat{\sigma}}\right\}$ pour tout i= 1,2,...,n , avec $w_i = 1$ si $\hat{\epsilon_i} = 0$ .On peut alors déduire l'égalité suivante :

$$w_i = \frac{\psi\left(\frac{\hat{\epsilon_i}}{\hat{\sigma}}\right)}{\frac{\hat{\epsilon_i}}{\hat{\sigma}}} \Rightarrow \psi\left(\frac{\hat{\epsilon_i}}{\hat{\sigma}}\right) = \frac{w_i \hat{\epsilon_i}}{\hat{\sigma}}$$

En substituent $\psi\left(\frac{\hat{\epsilon_i}}{\hat{\sigma}}\right) = \frac{w_i \hat{\epsilon_i}}{\hat{\sigma}}$ dans le système d'équation (2.5), on peut réécrire le système de la façons suivante :

$$\frac{\partial L(\beta)}{\partial \beta_0} \mid_{\beta=\hat{\beta}^M} = \sum_{i=1}^{n} \psi\left(\frac{\hat{\epsilon_i}}{\hat{\sigma}}\right) = \frac{1}{\hat{\sigma}} \sum_{i=1}^{n} w_i \hat{\epsilon_i} = 0$$

$$\frac{\partial L(\beta)}{\partial \beta_1} \mid_{\beta=\hat{\beta}^M} = \sum_{i=1}^{n} \psi\left(\frac{\hat{\epsilon_i}}{\hat{\sigma}}\right) x_{i,1} = \frac{1}{\hat{\sigma}} \sum_{i=1}^{n} w_i \hat{\epsilon_i} x_{i,1} = 0$$

$$\vdots$$

$$\frac{\partial L(\beta)}{\partial \beta_p} \mid_{\beta=\hat{\beta}^M} = \sum_{i=1}^{n} \psi\left(\frac{\hat{\epsilon_i}}{\hat{\sigma}}\right) x_{i,p} = \frac{1}{\hat{\sigma}} \sum_{i=1}^{n} w_i \hat{\epsilon_i} x_{i,p} = 0$$

En utilisant le fait que $\hat{\sigma} = y_i - (\hat{\beta_0}^M + \hat{\beta_1}^M x_{i,1} + \hat{\beta_p}^M x_{i,p})$, on obtient le système d'équation suivant :

$$\begin{aligned}
\sum_{i=1}^{n} w_i(y_i - (\hat{\beta_0}^M + \hat{\beta_1}^M x_{i,1} + \hat{\beta_p}^M x_{i,p})) &= 0 \\
\sum_{i=1}^{n} x_{i,1} w_i(y_i - (\hat{\beta_0}^M + \hat{\beta_1}^M x_{i,1} + \hat{\beta_p}^M x_{i,p})) &= 0 \\
&\vdots \\
\sum_{i=1}^{n} x_{i,p} w_i(y_i - (\hat{\beta_0}^M + \hat{\beta_1}^M x_{i,1} + \hat{\beta_p}^M x_{i,p})) &= 0
\end{aligned}$$

En réorganisant les équations, on trouve les égalités :

$$\begin{aligned}
\sum_{i=1}^{n} w_i y_i &= \sum_{i=1}^{n} w_i \hat{\beta_0}^M + \sum_{i=1}^{n} w_i \hat{\beta_1}^M x_{i,1} + ... + \sum_{i=1}^{n} w_i \hat{\beta_p}^M x_{i,p} \\
\sum_{i=1}^{n} w_i x_{i,1} y_i &= \sum_{i=1}^{n} x_{i,1} w_i \hat{\beta_0}^M + \sum_{i=1}^{n} x_{i,1} w_i \hat{\beta_1}^M x_{i,1} + ... + \sum_{i=1}^{n} x_{i,1} w_i \hat{\beta_p}^M x_{i,p} \\
&\vdots \\
\sum_{i=1}^{n} w_i x_{i,p} y_i &= \sum_{i=1}^{n} x_{i,p} w_i \hat{\beta_0}^M + \sum_{i=1}^{n} x_{i,p} w_i \hat{\beta_1}^M x_{i,1} + ... + \sum_{i=1}^{n} x_{i,p} w_i \hat{\beta_p}^M x_{i,p}
\end{aligned}$$

En ramenant ce système d'équation sous forme matricielle, on obtient alors :

$$\begin{pmatrix} \sum_{i=1}^{n} w_i y_i \\ \sum_{i=1}^{n} w_i x_{i,1} y_i \\ \vdots \\ \sum_{i=1}^{n} w_i x_{i,p} y_i \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} w_i & \sum_{i=1}^{n} x_{i,1} w_i & \sum_{i=1}^{n} x_{i,2} w_i & \cdots & \sum_{i=1}^{n} x_{i,p} w_i \\ \sum_{i=1}^{n} w_i x_{i,1} & \sum_{i=1}^{n} x_{i,1}^2 w_i & \sum_{i=1}^{n} x_{i,2} w_i x_{i,1} & \cdots & \sum_{i=1}^{n} x_{i,p} w_i x_{i,1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^{n} w_i x_{i,p} & \sum_{i=1}^{n} x_{i,1} w_i x_{i,p} & \sum_{i=1}^{n} x_{i,2} w_i x_{i,p} & \cdots & \sum_{i=1}^{n} x_{i,p}^2 w_i \end{pmatrix} \begin{pmatrix} \hat{\beta_0}^M \\ \hat{\beta_1}^M \\ \vdots \\ \hat{\beta_p}^M \end{pmatrix}$$

Ce qui est équivalent à :

$$X^t W Y + X^t W X \hat{\beta}^M$$

où la matrice W est une matrice diagonale $n \times n$ contenant tous les poids w_i :

$W = diag(w_1, w_2, ..., w_n) = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}$ Si la matrice $X^t W X$ est inversible, on en déduit que

$$\hat{\beta}^M = (X^t W X)^{-1} X^t W Y$$

On trouve que le M-estimateur des paramètres du modèle est très similaire à celui obtenu par la méthode des moindres carrés ordinaire. L'introduction d'une matrice de poids permet de réduire l'influence des données aberrantes. Toutefois, cette matrice de poids dépend des résidus, lesquels dépendent de l'estimation des paramètres de la régression.
On utilise alors IRLS(Iteratively Reweighted Least squares), une procédure itérative, pour trouver le M-estimateur des paramètres du modèle. Voici algorithme proposé par Susanti, Pratiwi, Sulistijowati et Liana pour trouver un M-estimateur des paramètres du modèle.

Algorithm :M-estimateur
itération 0

- On calcule les estimateurs $\hat{\beta}^{M^{(0)}}$ en utilisant la méthode des MCO .
- On calcule ensuite les résidus $\left\{\hat{\epsilon_i}^{(0)}\right\}_{i=1}^{n}$ en utilisant les paramètres estimés $\left\{\hat{\beta}^{M^{(0)}}\right\}_{i=0}^{p}$
- On trouve $\hat{\sigma}^{(0)}$ =1.4826 MAD,où MAD est obtenu avec $\left\{\hat{\epsilon_i}^{(0)}\right\}_{i=1}^{n}$.
- On choisit une fonction de poids W(u)(associée à une fonction objective $\rho(u)$)
- On calcule ensuite les poids $w_i^{(0)} = w\left(\frac{\hat{\epsilon_i}^{(0)}}{\hat{\sigma}^{(0)}}\right)$
- On obtient alors la matrice $W^{(0)} = \begin{pmatrix} w_1^{(0)} & 0 & \cdots & 0 \\ 0 & w_2^{(0)} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & w_n^{(0)} \end{pmatrix}$

itération 1

- On calcule $\hat{\beta}^{M^{(1)}} = (X^t W^{(0)} X)^{-1} X^t W^{(0)} Y$ en utilisant la méthode des MCO.
- On déduit les résidus $\left\{\hat{\epsilon_i}^{(1)}\right\}_{i=1}^{n}$ en utilisant les paramètres estimés $\left\{\hat{\beta}^{M^{(1)}}\right\}_{i=0}^{p}$
- On trouve $\hat{\sigma}^{(1)}$ =1.4826 MAD,où MAD est obtenu avec $\left\{\hat{\epsilon_i}^{(1)}\right\}_{i=1}^{n}$.
- On calcule ensuite les poids $w_i^{(1)} = w\left(\frac{\hat{\epsilon_i}^{(1)}}{\hat{\sigma}^{(1)}}\right)$

— On obtient alors la matrice $W^{(1)} = \begin{pmatrix} w_1^{(1)} & 0 & \cdots & 0 \\ 0 & w_2^{(1)} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & w_n^{(1)} \end{pmatrix}$

itération L

— On calcule $\hat{\beta}^{M^{(L)}} = (X^t W^{(L-1)} X)^{-1} X^t W^{(L-1)} Y$.

— On calcule ensuite les résidus $\left\{\hat{\epsilon_i}^{(L)}\right\}_{i=1}^{n}$ en utilisant les paramètres estimés $\left\{\hat{\beta}^{M^{(L)}}\right\}_{i=0}^{p}$

— On trouve $\hat{\sigma}^{(L)}$ =1.4826 MAD, où MAD est obtenu avec $\left\{\hat{\epsilon_i}^{(L)}\right\}_{i=1}^{n}$.

— On calcule ensuite les poids $w_i^{(L)} = w\left(\frac{\hat{\epsilon_i}^{(L)}}{\hat{\sigma}^{(L)}}\right)$

— On obtient alors la matrice $W^{(L)} = \begin{pmatrix} w_1^{(L)} & 0 & \cdots & 0 \\ 0 & w_2^{(L)} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & w_n^{(L)} \end{pmatrix}$

On arrête le processus à la L*ieme* itération lorsque :

$$\frac{\| \hat{\beta}^{M^{(L)}} - \hat{\beta}^{M^{(L-1)}} \|_2}{\| \hat{\beta}^{M^{(L)}} \|_2} < \xi$$

où ξ est un nombre positif aussi petit que l'on veut, fixé à l'avance ,par exemple ξ =0.0001. Dans le logiciel R, la fonction rlm utilise plutôt le pourcentage de changement entre les résidus de l'itération L et ceux de l'itération L-1 pour un ξ fixé :

$$\frac{\| \hat{\beta}^{M^{(L)}} - \hat{\beta}^{M^{(L-1)}} \|_2}{\| \hat{\beta}^{M^{(L)}} \|_2} < \xi$$

2.4.1 Propriétés d'un M-estimateur des paramètres du modèle

Point de rupture

Comme l'estimateur des MCO, le M-estimateur des paramètres du modèle défini précédemment est construit en supposant qu'il n'y ait pas d'erreur dans les variables indépendantes. Donc il ne considèrent pas les points de levier. On appelle point de levier un point qui semble aberrant par rapport aux autres points obtenus pour la variable explicative.

En fait, un seul mauvais point de levier peut influencer considérablement

l'estimation des paramètres comme on peut le remarquer dans le graphique ci-dessous : Dans le premier graphique à gauche de la figure 2.3, on peut voir

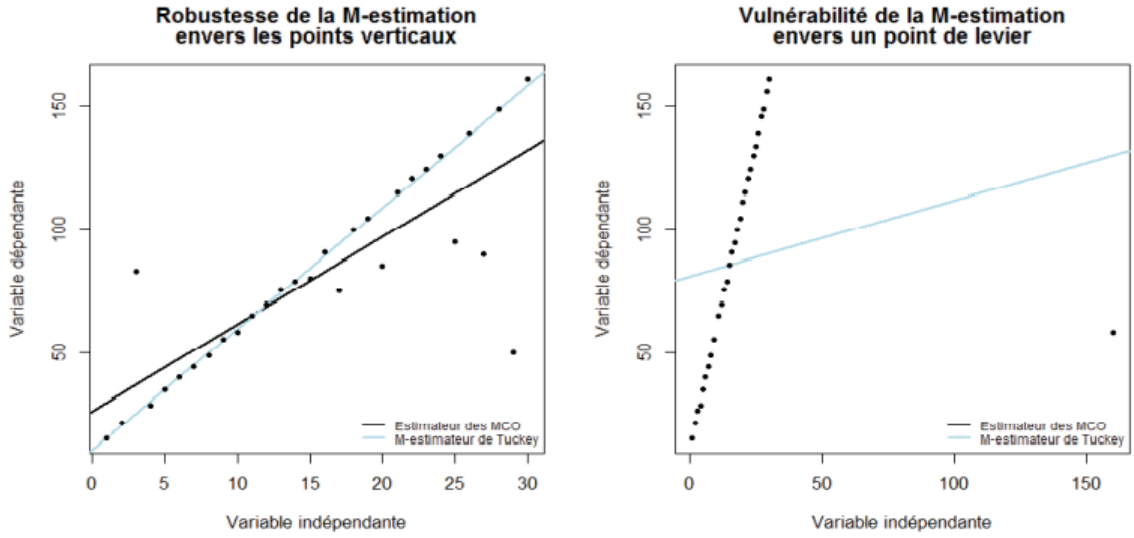


FIGURE 2.11 – Influence des points verticaux et d'un point de levier sur M-estimation

robustesse d'un M-estimateur des paramètres du modèle, construit à partir de la fonction ρ proposée par Tukey,devant les points verticaux. On définit un point vertical comme étant un point qui semble aberrant par rapport aux autres points obtenus pour la variable réponses. L'estimateur des MCO quant è lui, est fortement influencé par ces points aberrants.
Toutefois, dans le graphique de droite, on remarque que la présences d'un seul point de levier influence considérablement la M-estimation.
Le point de rupture d'un M-estimateur des paramètres du modèle est donc $\theta_n = \frac{1}{n}$. Lorsque n devient très grand, le point de rupture tend vers 0% .

Efficacité relative Comme dit précédemment, un M-estimateur des paramètres du modèle dépend de certaines constantes qui permettent d'augmenter sa résistance lorsqu'il y a présences de données aberrantes,mais cela au déterminent de son efficacité. Dans le cas d'un M-estimateur des paramètres du modèle construit avec la fonction $\rho(u)$ proposée par Huber, on utilise habituellement une constant c= 1.345 ,ce qui permet d'obtenir une efficacité de 95% par rapport à l'estimation par la méthode des MCO. Dans le cas d'un M-estimateur des paramètres du modèle construit avec la fonction $\rho(u)$ proposée par Tukey, on utilise habituellement c=4.685, ce qui permet d'obtenir aussi une efficacité de 95% par rapport à la méthode des MCO.

Remarque :
Si on diminue la valeur des constantes c, ci-dessus, on obtiendra des estimateurs plus robustes lorsqu'il y a présence de données aberrantes, car on pénalisera les erreurs les plus grandes. Toutefois, nous aurons un estimateur moins efficace .

2.5 MM-estimation des paramètres du modèle

La MM-estimation des paramètres du modèle est une combinaison de la S-estimation et de la M-estimation des paramètres du modèle. On estime d'abord les paramètres régression en utilisant la S-estimation (qui minimise la dispersion des résidus) et on applique ensuite la M-estimation en utilisant à la première itération les paramètres du modèle obtenus par S-estimation. De plus, pour chaque étape de la M-estimation, on utilise la même estimation de l'écart-type, $\hat{s}^{MM}$, aussi obtenue avec la S-estimation. Cela permet d'obtenir des estimation avec un haut point de rupture et plus efficace que la M-estimation et S-estimation . Ainsi, le MM-estimateur est la solution de :

$$\begin{aligned}
\frac{\partial L(\beta)}{\partial \beta_0} \mid_{\beta=\hat{\beta}^M} &= \sum_{i=1}^{n} \psi\left(\frac{y_i - (\hat{\beta}_0^{\,M} + \hat{\beta}_1^{\,M} x_{i,1} + \hat{\beta}_p^{\,M} x_{i,p}}{\hat{s}^{MM}}\right) = \frac{-1}{\hat{s}^{MM}} = 0 \\
\frac{\partial L(\beta)}{\partial \beta_1} \mid_{\beta=\hat{\beta}^M} &= \sum_{i=1}^{n} \psi\left(\frac{y_i - (\hat{\beta}_0^{\,M} + \hat{\beta}_1^{\,M} x_{i,1} + \hat{\beta}_p^{\,M} x_{i,p}}{\hat{s}^{MM}}\right) = \frac{-x_{i,1}}{\hat{s}^{MM}} = 0 \\
&\vdots
\end{aligned}$$

$$\frac{\partial L(\beta)}{\partial \beta_p} \mid_{\beta=\hat{\beta}^M} = \sum_{i=1}^{n} \psi\left(\frac{y_i - (\hat{\beta}_0^{\,M} + \hat{\beta}_1^{\,M} x_{i,1} + \hat{\beta}_p^{\,M} x_{i,p}}{\hat{s}^{MM}}\right) = \frac{-x_{i,p}}{\hat{s}^{MM}} = 0 \quad (2.6)$$

où ρ satisfait aux condition C1,C2 et C3 ainsi qu'à la définition 2.2 et ψ est la dérivée de ρ .Un choix usuel pour la fonction ρ est la fonction proposée par Tukey (2.2).
On a donc que

$$\hat{\beta}^{MM} = (X^t W X)^{-1} X^t W Y$$

où $\hat{s}^{MM}$ est l'estimation de la dispersion obtenue par la S-estimation. Voici un algorithme proposé par Susanti, Pratiwi, Sulistijowati et Liana pour trouver un MM-estimateur des paramètres du modèle :

Algorithme : MM-estimateur

Itération 0 :

- On effectue l'algorithme pour la S-estimation. Les paramètres du modèle trouvés à la dernière itération deviennent les paramètres initiaux de régression $\hat{\beta}^{MM(0)}$ pour cet algorithme et $\hat{s}^{MM}$, obtenu à partir des erreurs à la dernière itération de l'algorithme pour la S-estimation, sera l'estimation de la dispersion à chaque itération de l'algorithme .
- On calcule ensuite les résidus $\left\{\hat{\epsilon_i}^{(0)}\right\}_{i=1}^{n}$ en utilisant les paramètres estimés $\hat{\beta}^{MM(0)}$.
- On pose $u_i = \frac{\hat{\epsilon_i}^{(0)}}{\hat{s}^{MM}}$ pour i= 1,...,n.
- On calcule ensuite les poids (pour l'estimation des paramètres du modèle) $w_i^{(0)} = w(u_i) = \frac{\psi(u_i)}{u_i}$.
- On obtient alors la matrice $W^{(0)} = \begin{pmatrix} w_1^{(0)} & 0 & \cdots & 0 \\ 0 & w_2^{(0)} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & w_n^{(0)} \end{pmatrix}$

Itération 1 :

- On calcule $\hat{\beta}^{MM(1)} = (X^t W^{(0)} X)^{-1} X^t W^{(0)} Y$
- On calcule ensuite les résidus $\left\{\hat{\epsilon_i}^{(1)}\right\}_{i=1}^{n}$ en utilisant les paramètres estimés $\hat{\beta}^{MM(1)}$.
- On pose $u_i = \frac{\hat{\epsilon_i}^{(1)}}{\hat{s}^{MM}}$ pour i= 1,...,n.
- On calcule ensuite les poids (pour l'estimation des paramètres du modèle) $w_i^{(1)} = w(u_i) = \frac{\psi(u_i)}{u_i}$.
- On obtient alors la matrice $W^{(0)} = \begin{pmatrix} w_1^{(1)} & 0 & \cdots & 0 \\ 0 & w_2^{(1)} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & w_n^{(1)} \end{pmatrix}$

Itération L :

- On calcule $\hat{\beta}^{MM(L)} = (X^t W^{(L-1)} X)^{-1} X^t W^{(L-1)} Y$

— On calcule ensuite les résidus $\left\{\hat{\epsilon}^{(L)}\right\}_{i=1}^{n}$ en utilisant les paramètres estimés $\hat{\beta}^{MM(L)}$

— On pose $u_i = \frac{\hat{\epsilon_i}^{(L)}}{\hat{s}^{MM}}$ pour i= 1,...,n.

— On calcule ensuite les poids (pour l'estimation des paramètres du modèle) $w_i^{(L)} = w(u_i) = \frac{\psi(u_i)}{u_i}$.

— On obtient alors la matrice $W^{(0)} = \begin{pmatrix} w_1^{(L)} & 0 & \cdots & 0 \\ 0 & w_2^{(L)} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & w_n^{(L)} \end{pmatrix}$

On arrête le processus à la L^{ieme} itération lorsque :

$$\frac{\| \hat{\beta}^{MM(L)} - \hat{\beta}^{MM(L-1)} \|_2}{\| \hat{\beta}^{MM(L)} \|_2} < \xi$$

où ξ est un très petit nombre positif fixé à l'avance, par exemple $\xi = 0.0001$. Dans le logiciel R, la fonction rlm utilise plutôt le pourcentage de changement entre les résidus de l'itération L et ceux de l'itération L-1 pour un ξ fixé :

$$\frac{\| \hat{\epsilon}^{MM(L)} - \hat{\epsilon}^{MM(L-1)} \|_2}{\| \hat{\epsilon}^{MM(L)} \|_2} < \xi$$

2.5.1 Propriétés des MM-estimateurs

Point de rupture

Dans la première itération du processus de MM-estimation des paramètres du modèle, on utilise la S estimation des paramètres du modèle pour trouver les paramètres initiaux de régression ainsi qu'un estimateur de la dispersion des erreurs. Cette première itération du processus permet donc d'obtenir un haut point de rupture pour le MM-estimateur résultant, tandis que les itérations suivantes visent à obtenir une forte efficacité asymptotique.Il a été prouvé que si le S-estimateur des paramètres du modèle trouvé à la première itération du processus possède un point de rupture de 50%, alors le MM-estimateur des paramètres du modèle aura également un point de rupture de 50%. Ainsi,lors de l'application de l'algorithme de S-estimation, on utilise une constante c=1.547 si on désire obtenir un estimateur avec un point de rupture de 50%.

Efficacité relative

Alors que le point de rupture d'un MM-estimateur des paramètres du modèle dépend du choix des constantes dans la recherche du S-estimateur dans la

première itération du processus, l'efficacité relative d'un MM-estimateur des paramètres de régression est déterminée par le choix des constantes dans les itérations suivantes . Par conséquent, contrairement à la M-estimation des paramètres du modèle, le point de rupture et l'efficacité relative ,pour la MM-estimation des paramètres du modèle, sont indépendants l'un de l'autre, de sort que si si le oint de rupture reste fixé à 50 % , l'efficacité relative peut - être réglée aussi prés qu'on le souhaite de 100%. Ainsi, lors de l'application de l'algorithme de M-estimation (sur les paramètres initiaux trouvés par S-estimation), on utilise une constante c=4.685 pour obtenir une efficacité relative de 95% par rapport à l'estimation par la méthode des MCO.

Exemple

Voici le résultat de "lmrob" appliquée au jeu de données Animals2 vu précédemment. La première partie du sommaire des résultats ressemble au tableau obtenu avec "lm "(estimé des coefficients, erreur-type et test de signification).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.11749    0.09146   23.15   <2e-16 ***
log(body)    0.74603    0.02065   36.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

avec $\beta_1 = 0.74603$,$\beta_0 = 2.11749$,RSE=0.721 et $R^2 = 0.9229$.

On remarque que la valeurs de RSE diminue et R^2 augment ce qui signifie que la MM-estimation est une methode robust par rapport à la méthode MCO.

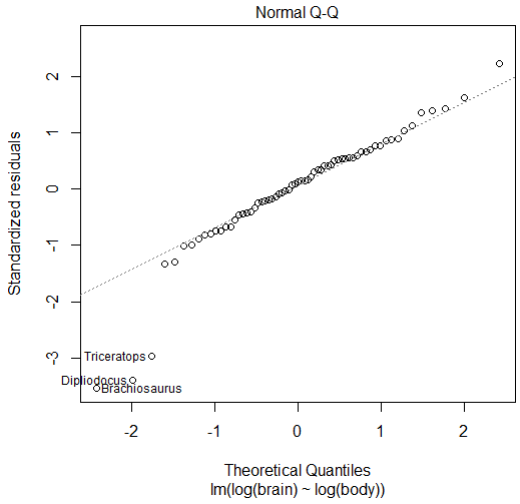


FIGURE 2.12 – Q-QNormal

ici, les points sont plutôt bien alignés sur la droite, donc les résidus sont bien distribués le long de la droite figurant sur le plot.

2.6 T-Régression

la distribution t de Student est d'abord présentée comme une façon d'estimer la distribution de la moyenne d'un échantillon $\bar{x}$ lorsque la variance de la population est inconnue. Pour un échantillon de n observations, si $\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}$ suit une distribution normale centrée réduite et qu'on remplace σ par son estimé s à partir de l'échantillon, alors $\frac{\sqrt{n}(\bar{x}-\mu)}{s}$ suit une distribution t avec $n-1$ degrés de liberté.
Le modèle de régression linéaire standard s'écrit

$$Y = X\beta + \varepsilon$$

où X est la matrice de conception, β est un vecteur de paramètres à estimer, ε est un vecteur d'erreurs généralement supposées normalement distribuées avec une moyenne nulle.

$$\varepsilon_i \backsim N(0, \sigma^2)$$

La distribution normale a des queues assez fines, ce qui rend la régression vanille extrêmement sensible aux valeurs aberrantes. Par exemple, nous trouverons ci-dessous des lignes de régression pour deux ensembles de données,

identiques à l'exception de la présence d'une seule valeur aberrante. La seule valeur aberrante ci-dessus fait que le modèle sous-estime gravement la vraie pente. Si nous essayons d'extrapoler même légèrement au-delà de la plage des données, nous sommes susceptibles d'être loin. La solution habituelle serait de supprimer l'observation, mais les valeurs aberrantes dans les ensembles de données réels ne sont généralement pas aussi évidentes et il est donc probablement préférable de trouver une solution qui réduit l'influence des valeurs aberrantes sans nous obliger à jeter des données.

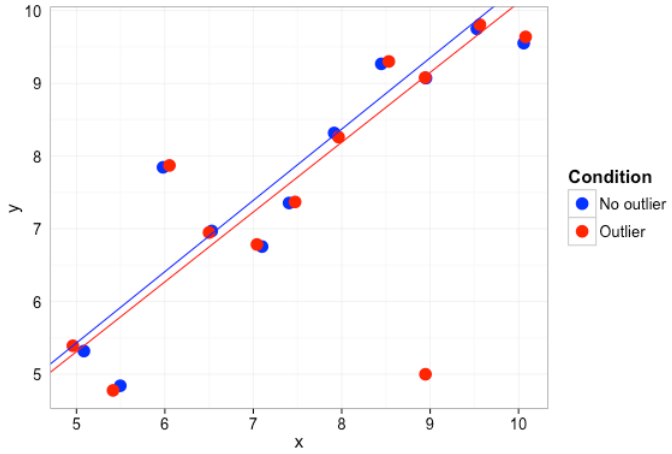


FIGURE 2.13 – Influence les valeurs extrêmes sur le droite d'ajustement

Une solution au problème des valeurs extrêmes consiste à remplacer les erreurs normales dans le modèle de régression par celles qui suivent une distribution avec des queues plus épaisses, comme la t distribution. La distribution familière t utilisée dans les tests d'hypothèses ne fonctionnera pas, car nous ne pouvons pas contrôler sa moyenne et sa variance, mais on peut facilement lui donner des paramètres d'emplacement et d'échelle en écrivant sa fonction de densité comme suit :

$$f(x, \mu, \sigma, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}\left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

Sous cette forme, la t distribution a une moyenne μ(quand $\nu > 1$), et une variance $\frac{\sigma^2\nu}{\nu-2}$ (quand $\nu > 2$).
L'avantage de cette paramétrisation est que nous pouvons contrôler la moyenne et la variance comme avec la distribution normale, et le paramètre DOF ν donne un moyen simple de contrôler la graisse des queues et le poids accordé aux valeurs extrêmes.

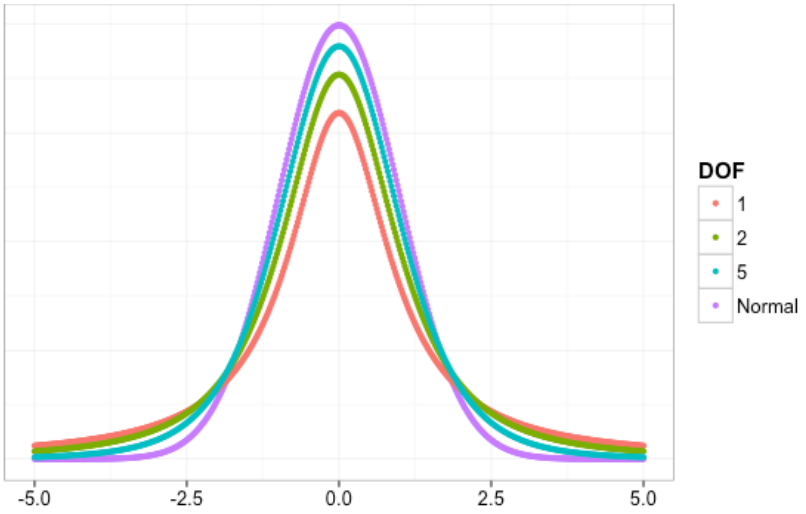


FIGURE 2.14 – Distribution t avec DOF varaiable

Estimation

Nous avons ici deux options : nous pouvons traiter les degrés de liberté ν comme une donnée et les utiliser comme paramètre de réglage pour contrôler la graisse des queues, ou nous pouvons estimer ν à partir des données avec tout le reste. Le second cas donne au modèle une sorte de robustesse adaptative, tandis que le premier simplifie considérablement le modèle mais nécessite de choisir une valeur ν priori dont le chercheur pense qu'elle donnera un niveau de robustesse approprié (des valeurs inférieures, en théorie, entraînent une plus grande robustesse à valeurs aberrantes en créant une distribution à queue très épaisse).
La vraisemblance, avec ν inclus, est donnée par :

$$L(\beta, \sigma, \nu) = \left(\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma} \right)^n \prod_{i=1}^{n} \left(1 + \frac{e_i^2}{\nu\sigma^2} \right)^{-\frac{\nu+1}{2}}$$

Dans ce cas, les estimations du maximum de vraisemblance sont connues pour n'avoir aucune forme fermée, et donc pour s'adapter au modèle, nous devrons brancher le log-vraisemblance

$$logL(\beta, \sigma, \nu) = n \left(ln\Gamma\left(\frac{\nu+1}{2}\right) - ln\Gamma\left(\frac{\nu}{2}\right) - ln\frac{\sqrt{\nu}}{\sigma} \right) - \frac{\nu+1}{2} \sum_{i=1}^{n} ln\left(1 + \frac{e_i^2}{\nu\sigma^2}\right)$$

dans notre optimiseur préférons. nous avons essayer de dériver des expressions pour les MLE lorsque ν est maintenu constant, mais on ne trouve pas non plus de forme fermée pour celles-ci, donc ce sont des méthodes numériques jusqu'au bout. En fait, cela ne fonctionne pas si bien à moins qu'il ne ν soit corrigé ; la probabilité est trop forte. Sans utiliser de méthodes plus compliquées (par exemple, la maximisation des attentes), il est préférable de choisir une valeur ν a priori.

Exemple

En appliquant ce modèle au jeu de données Animals2, nous obtenons des coefficients de régression comparables (en tenant compte de la marge d'erreur) à ceux obtenus dans la section précédente avec "lmrob". Ces résultats de retrouvent dans la section "Location model" du sommaire de "tlm" .

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2244     0.2745  -4.461 8.17e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Scale parameter taken to be  2 )


Est. degrees of freedom parameter:  2.071194
Standard error for d.o.f:  0.6678805
No. of iterations of model : 8 in 0.08
Heteroscedastic t Likelihood : -86.3654
```

Ona β_0 = -1.2244 et RSE=0.6678805 plus petit que RSE de MM-estimation. Il est important de spécifier estDof = TRUE pour estimer le nombre de degrés de liberté, plutôt que de supposer une valeur fixe. Cependant, cet estimé risque d'être peu précis sauf si on a beaucoup de données. Ici, le nombre de degrés de liberté est de 2.08, avec une erreur-type de 0.67.

2.7 Régression quantiles

Considérons une variable aléatoire Y ayant pour fonction de répartition $F_Y(y) = P(Y \leq y)$.Soit $\tau = P(Y < q_\tau(Y))$, par définition, le $\tau^{ième}$ quantile est $q_\tau(Y) = inf(y : F_Y(y) \geq \tau)$.
Un quantile très utilisé est la médiane ($\tau = 0.5$). Les régressions quantiles ont pour but d'évaluer comment les quantiles conditionnels $q_\tau(Y|X) = inf(y : F_{Y|X}(y) \geq \tau)$ varient en fonction de X $\in R^p$. Un aspect très intéressant des

régressions quantiles est qu'en changeant la valeur de τ, on peut décider de pénaliser plus les erreurs positives ou les erreurs négatives et ainsi obtenir des intervalles de confiances pour l'estimation des paramètres du modèle. De façon similaire aux estimateurs de ce chapitre, l'estimation des paramètres du modèle par les régressions quantiles ainsi que quelques propriétés seront présentées.

2.7.1 Estimation des paramètres β avec la régression quantile

Considérons l'échantillon E définie en (1.1).Dans la régression quantile standard, on suppose que les quantiles conditionnels s'expriment sous la forme linéaire suivante :

$$q_\tau(Y|X) = X^t\beta_\tau \tag{2.7}$$

où à chaque τ correspond un vecteur de coefficients $\beta_\tau = (\beta_0, \beta_{1,\tau}, .., \beta_{p,\tau})^t$ associé aux p variables explicatives ainsi qu'à l'ordonnée à l'origine. On permet ainsi aux coefficients β_τ de différer d'un quantile à l'autre. On peut réécrire (2.7) de la manière suivante :

$$y = X^t\beta_\tau + \epsilon_\tau \qquad avec \qquad 0 = q_\tau(\epsilon_\tau|X) = inf(\epsilon_\tau : F_{\epsilon_\tau|x}(\epsilon_\tau) \geq \tau),$$

ce qui est très similaire à la régression linéaire standard. L'estimation de la régression quantile est basée sur le fait que le quantile d'ordre τ est le résultat du problème d'optimisation suivant :

$$q_\tau(Y) = argmin_\alpha E[\rho_\tau(Y - \alpha)]$$

, où $\rho_\tau(u) = (\tau - 1(u < 0))u$ et 1 est la fonction indicatrice.
Dans la figure ci-dessous,on peut voir la représentation graphique de quelque fonctions ρ_τ pour différentes valeurs de τ.

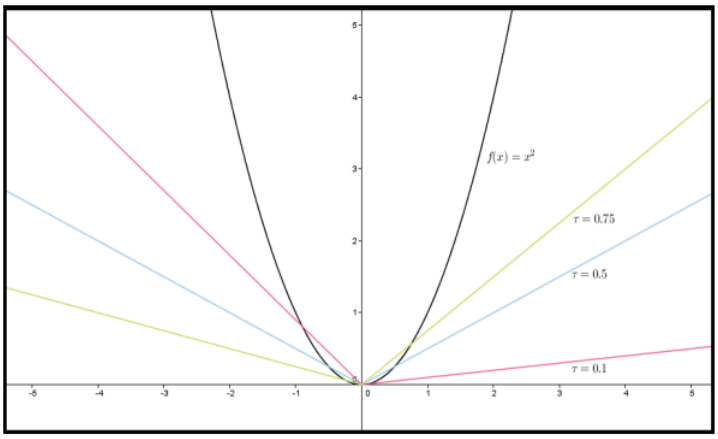


FIGURE 2.15 – Fonction ρ_τ avec différentes valeurs de τ

On voit bien dans la figure précédente que, selon la valeur de τ, on peut obtenir des modélisations qui donnent plus ou moins de poids aux grandes ou aux petites valeurs.
Par exemple, si on prend $\tau = 0.1$, on donne plus de poids aux valeurs négatives et moins de poids aux valeurs positives.
Si on fait la comparaison avec le modèle de régression linéaire (1.3), la fonction de perte quadratique utilisée pour celle -ci lorsqu'on applique la méthode des moindre carrés ordinaire est remplacée, dans la régression quantile, par la fonction $\rho_\tau(.)$. Ainsi, comme cette fonction augmente de façon linéaire avec les résidus, les très grands écarts ont une influence moindre sur les paramètres du modèle, ce qui explique la robustesse des régressions quantiles aux valeurs aberrantes.
L'intérêt de cette définition est qu'elle s'étend simplement au cadre conditionnel :

$$q_\tau(Y|X = x) = argmin_\alpha E([\rho_\tau(Y - \alpha)|X = x] \tag{2.8}$$

Dans le cas de la régression quantile classique, on peut se limiter aux fonctions linéaires puisqu'on suppose $q_\tau(Y|X) = X^t\beta_\tau$. On a alors :

$$\hat{\beta}_\tau = argmin_\beta E[\rho_\tau(Y - X^t\beta)] \tag{2.9}$$

Comme on désire obtenir un estimateur pour un échantillon fini de valeurs, on utilise la moyenne empirique au lieu de la moyenne théorique.Ainsi, en considérant l'échantillon E, on remplace l'espérance obtenue en (2.13) par la moyenne empirique sur l'échantillon :

$$\hat{\beta}_\tau = argmin_\beta \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - (\beta_0) + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_p x_{i,p})) \qquad (2.10)$$

Pour τ= 0.5, la fonction ρ devient la fonction valeur absolue,ce qui revient à minimiser la somme des valeurs absolue des erreurs.Cette méthode ,appelée "estimation des moindres déviations absolues", est utilisée depuis longtemps comme alternative robuste pour l'estimation des moindres carrés ordinaires. la valeur de τ pour obtenir différentes estimations des paramètres de régressions. On peut ainsi obtenir un intervalle de confiance pour les paramètres du modèle.Comme $\tau \in [0.1]$, il existe une infinité de régressions quantiles possibles.Toutefois, il est illusoire de faire l'estimations de mille régressions quantiles pour un échantillon qui ne contient que 10 individus. Ainsi, en pratique, le nombre de régressions quantiles qu'on estime dépendre de la taille de l'échantillon (de l'ordre de n ln(n) où n représente la taille de l'échantillon). Contrairement aux estimateurs du chapitre 2, il n'existe pas de solution explicite à (2.13). Il faut donc résoudre numériquement ce problème. Toutefois, on peut remarquer que la fonction ρ_τ n'est pas dérivable en 0, ce qui implique que plusieurs algorithmes standards (celui de Newton-Raphson en est un exemple) ne peuvent être utilisés. Cependant, définissant u comme la partie positive et v comme la partie négative, on peut reformuler (2.16) avec le programme linéaire suivant :

$$min_{(\beta,u,v)\in R^p \times R^{2n})} \left\{ \tau 1_n^t u + (1-\tau) 1_n^t v | X\beta + u + v = Y \right\}$$

où 1_n est un vecteur de 1 de taille n. Cette reformulation est intéressante, car plusieurs algorithmes, comme celui de la méthode du simplex, permettent la résolution de ce programme linéaire . Toutefois, lorsque le nombre d'observation est très la méthode du simplex devient très coûteuse en temps de calcule. On privilégie alors des méthodes plus performantes, telles que la méthode des point intérieurs, pour résoudre ce problème.

2.7.2 Propriétés des estimateurs des régression quantiles

Supposons que la fonction quantile conditionnelle associée à τ soit linéaire,c'est-à-dire :

$$q_\tau(Y|X) = X^t \beta_\tau$$

Considérons la fonction de distribution conditionnelle

$$P(Y < y|X) = F_{Y|X}(y)$$

et posons

$$q_\tau(Y|X) = F^{-1}_{Y|X}(\tau) \equiv \xi(\tau)$$

Considérons les condition suivantes :
A1. La fonction de répartition F associée à la variable Y est absolument continue, avec une densité continue $f(\xi)$ et uniformément bornée sur $[0, \infty[$ au point $\xi(\tau)$.
A2.Il existe des matrices définies positives $D_0 et D_1$tel que :
(i) $D_0 = E[XX^t]$
(ii) $D_1(\tau) = E[f(\xi(\tau))XX^t]$
(iii) $\frac{\|X\|}{\sqrt{n}} \longrightarrow 0$
La normalité asymptotique de l'estimateur $\hat{\beta}_\tau$ défini en (2.14) est alors donnée par le théorème ci-dessous
Théorème 5 (Normalité asymptotique)
Sous les conditions A1-A2 et lorsque n tend vers l'infini, on a :

$$\sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \longrightarrow^d N(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1})$$

où d est la convergence en distribution et N est la loi normale multidimensionnelle

2.7.2.1 Propriété d'équivariance de l'estimateur $\hat{\beta}_\tau$

On retrouve pour plusieurs estimateurs certaines propriétés d'invariance qui sont très importantes pour l'interprétation des résultats.

Théorème 6 (É*quivariance*)[Koenker and Bassett,1978]
Soit A une matrice de dimension p × p non singulière, $\lambda \in R^p$, et $\alpha > 0$. Considérons $\hat{\beta}_\tau(Y, X)$ la solution (2.14). Alors, pour tout $\tau \in [0, 1]$, on a :

1. $\hat{\beta}_\tau(\alpha Y, X) = \alpha\hat{\beta}_\tau(Y, X)$
2. $\hat{\beta}_\tau(-\alpha Y, X) = -\alpha\hat{\beta}_{1-\tau}(Y, X)$
3. $\hat{\beta}_\tau(Y + X\lambda, X) = \hat{\beta}_\tau(Y, X) + \lambda$
4. $\hat{\beta}_\tau(Y, AX) = A^{-1}\hat{\beta}_\tau(Y, X)$

Remarque
Les propriétés 1 et 2 sont appelées l'invariance au changement d'échelle, la propriété 3 est habituellement appelée l'invariance à la translation et la propriété 4 est l'invariance à la reparamétrisation.

Point de rupture
Similairement à la M-estimation,les régressions quantiles sont robustes aux points verticaux.Cependant, un seul point de levier peut forcer l'ensemble des régressions quantiles à traverser ce point. Ainsi, les régressions quantiles possèdent un point de rupture $\theta_n = \frac{1}{n}$.
Lorsque n devient très grand, le point de rupture tend vers 0%. La figure ci-dessous illustre bien la sensibilité des régressions quantiles devant un seul point de levier.

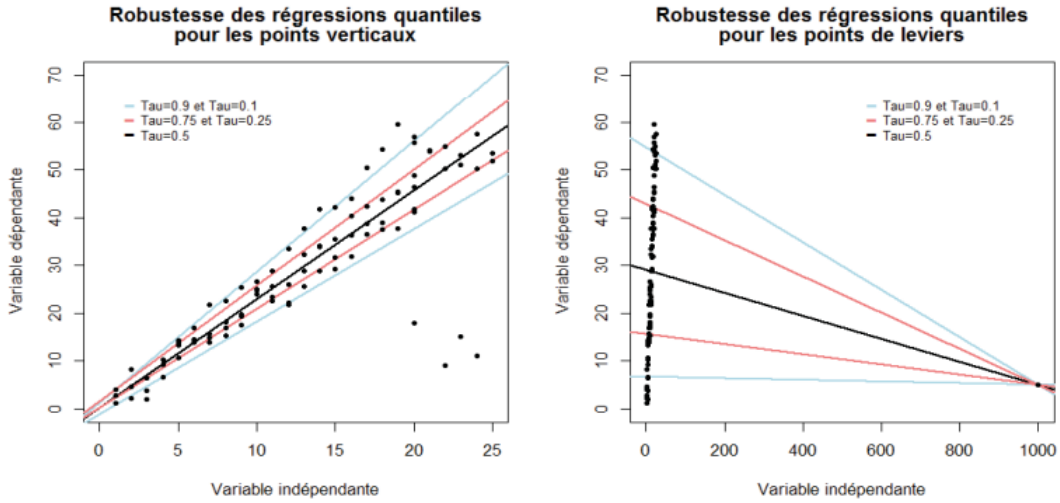


FIGURE 2.16 – Influence de points verticaux et d'un point de levier sur les régressions quantiles

Exemple
Nous utiliserons la fonction "rq" du package "quantreg" pour effectuer une régression quantile.

Le jeu de données "Mammals" inclus avec ce package montre la vitesse maximale connue (en km/h) de mammifères en fonction de leur poids. Puisque l'échelle de poids varie sur plusieurs ordres de grandeur, il est plus utile de représenter son logarithme.

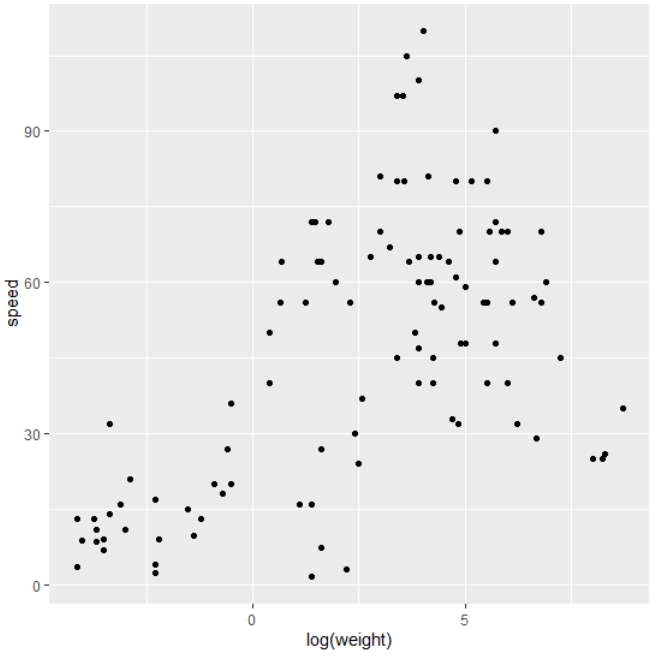


FIGURE 2.17 – Nuage de point des données Mammals

D'après ce graphique, il semble que le poids pourrait agir comme facteur limitant pour la vitesse des mammifères, donc son effet devrait être davantage ressenti sur les quantiles élevés de la distribution.
Pour visualiser rapidement le résultat d'une régression quantile avec un prédicteur, nous pouvons faire appel à la fonction geom quantile de ggplot2.

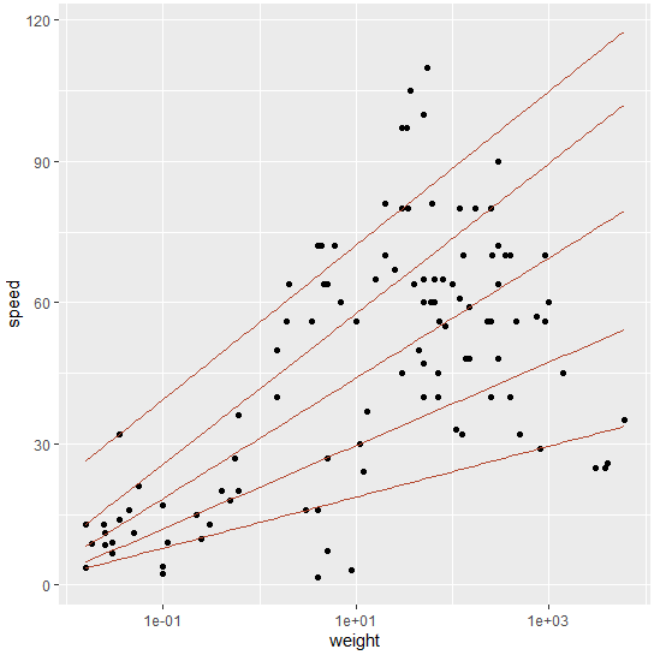


FIGURE 2.18 – Régression quantile avec un prédicteur

En résumé, dans ce chapitre nous avons vu quelque notions de bases sur les valeurs extrême, ainsi cinq méthodes d'estimation robustes des paramètres du modèle : Moindres carrées pondérés ,la M-estimation, la MM-estimation, la t-régression et les régressions quantiles.
Les régressions quantiles nous permettent d'obtenir des intervalles pour les paramètres en modifiant la valeur de τ. Pour les quatre autre méthodes, la t-régression semble être la méthode la moins sensible aux données aberrantes.

Chapitre 3

Simulation et comparaison

Ce chapitre sera consacré à la comparaison entre les différentes méthodes d'estimation robuste citer dans le deuxième chapitre.
Nous comparons les différentes méthodes on utilisons des simulation sans et avec données aberrantes pour un modale de régression linéaire simple, nous illustrons aussi des simulation pour un modèle de régression linéaire simple dont l'écart-type des résidus n'est pas constant, sans et avec valeurs aberrantes.
La modélisation a été faite à l'aide de logiciel **R**.
Pour MCO nous avons utilisé la fonction "lm", pour M-estimation et MM-estimation nous avons utilisé les fonction "glm" et "lmrob" ainsi "rlm" de package "robustbase", pour la t-régression on a utilisé la fonction "tlm" de les packages(hett,MASS) et pour les régression quantiles nous avons utilisé la fonction "rq" de package "quantreg".

3.1 Simulation sans données aberrantes

3.1.1 Avec variance constante

Dans un premier temps ,nous avons introduit un vecteur de résidus ϵ,généré aléatoirement,issus d'une loi normale de moyenne nulle et d'écart-type égale à 1 .
on été introduit le modèle $y_i = 2.2 + 7 * x_i + \epsilon_i$, i=1,2,...,100, est un modèle linéaire avec n = 100 et où les x_i constituent une suite de valeurs dans l'intervalle [0,1]
Application de moindres carrés ordinaires

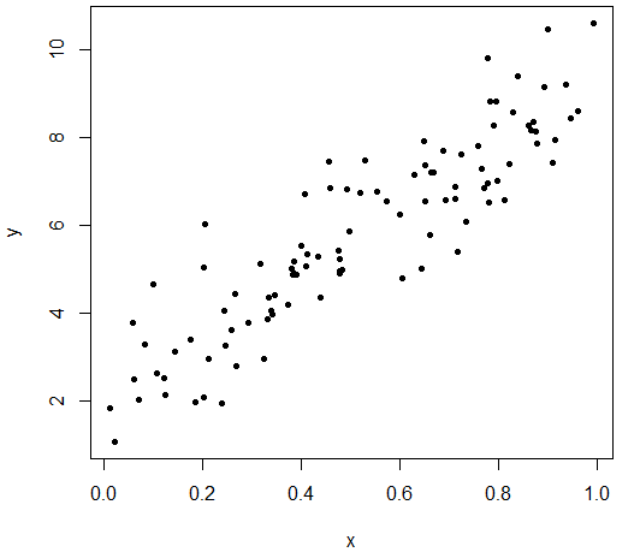


FIGURE 3.1 – Nuage de points des y en fonction des x

On remarque que les points de nuage forment pratiquement une droite donc le modèle proposé est bon.

Estimation des paramètres

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0207     0.2058   9.818    3e-16 ***
x             7.3123     0.3535  20.688   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9411 on 98 degrees of freedom
Multiple R-squared:  0.8137,    Adjusted R-squared:  0.8118
F-statistic:   428 on 1 and 98 DF,  p-value: < 2.2e-16
```

A partir des résultats d'estimation donnés sous **R**, nous avons : $\beta_0 = 2.0207$, $\beta_1 = 7.3123$, RSE $= 0.9411$ et $R^2 = 0.8137$.

Donc la droite de régression est définie par l'équation :

$$\hat{y} = 2.0207 + 7.3123x$$

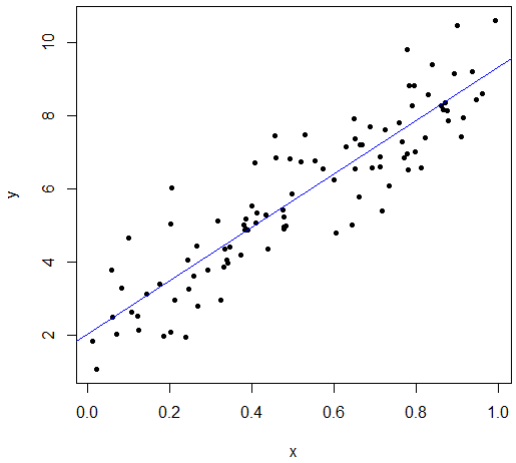


FIGURE 3.2 – Nuage de points avec la droite d'ajustement

On remarque que les points sont bien alignées sur la droite de régression donc le modèle semble bien.

Graphiques des résidus :

— **QQ-Normal** : Ce tracé est particulièrement utile pour tester la normalité des résidus, si les points semblent s'aligner le long de la droite indique que les résidus ont distribués normalement mais si les points forment une courbe et non pas une droite ce qui indique que les données sont normalement distribuées.

— **Valeurs prévues vs Résidus** : Ce tracé est particulièrement utile pour tester l'hypothèse de linéarité concernant la relation entre les variables indépendante et la variable dépendante, plus précisément si la relation linéaire, les résultats des résidus doit former un "nuage" homogène autour de la droite centrale.

— **Résidus standardisés vs valeurs ajustées** : Ce tracé est utilisé pour détecter l'homoscédasticité(hypothèse de variance égale). Montre comment les résidus sont distribués dur la plage de prédicteurs, il est similaire au graphique valeurs résiduelle vs ajusté, sauf qu'il utilise des valeurs résiduelle standardisés, il ne devrait y avoire aucun motif discernable dans l'intrigue. Cela impliquant que les erreurs sont normalement distribués, mais au cas où le graphique montre un motif discernable(probablement forme d'entonnoir), impliquant une distribution d'erreur non normale.

— **La distance de Cook** : Ce tracé mesure l'influence de l'observation i sur l'estimation du paramètre β_j, si la distance de l'observation i est grande alors cette observation influence beaucoup l'estimation de β. En résumé ce graphe permet d'identifier les points avec une forte influence.

Analyse des résidus

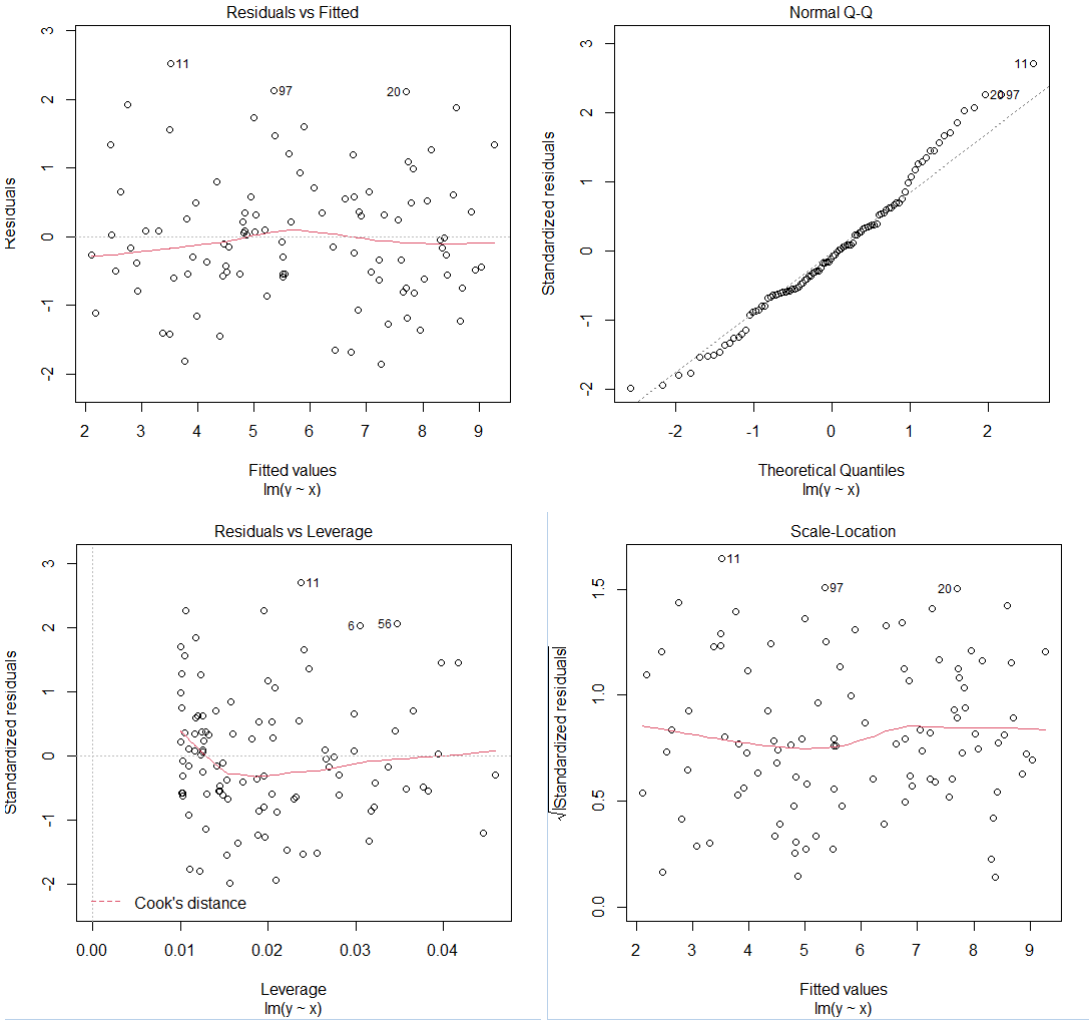


FIGURE 3.3 – Graphiques des résidus

Le premier graphe Residuals vs Fitted : le graphe nous montre que lorsque les réponses prédites par le modèle (fitted values) augmentent, les résidus restent globalement uniformément distribués de part et d'autre de 0. Cela montre, qu'en moyenne, la droite de régression, est bien adaptée aux données, et donc que l'hypothèse de linéarité est acceptable.

Le deuxième graphe Normal Q-Q : les résidus sont bien distribués le long de la droite figurant sur le graphe, alors l'hypothèse de normalité est acceptée dans notre cas .

le troisième graphe Residuals vs Leverage : La ligne pointillée dé-

marque le seuil de 1 pour la distance de cook dans notre cas aucun points est proche ou dépasse le 1. Donc aucune donnée est influente (absence de données aberrantes).
le quatrième graphique : Aucun motif discernable est remarquable ce la impliquant que les erreurs sont normalement distribué avec présence de l'homoscédasticitè.

3.1.2 Avec variance non constante

Dans ce cas nous avons introduit un vecteur de résidus ϵ, généré aléatoirement, issus d'une loi normale de moyenne nulle et de variance non constante. on été introduit le modèle $y_i = 2.2 + 7 * x_i + \epsilon_i$, i=1,2,...,100, est un modèle linéaire avec n = 100 et où les x_i constituent une suite de valeurs dans l'intervalle [0,1].

Estimation par méthode MCO

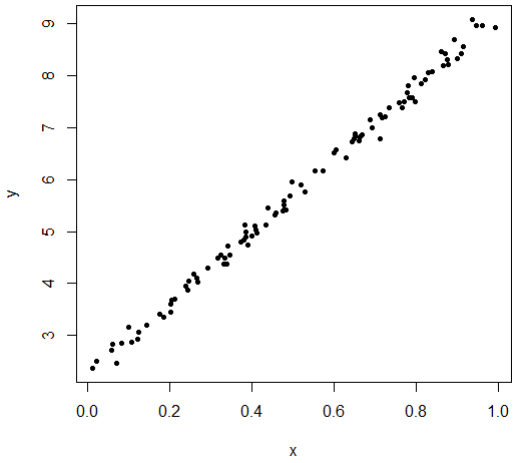


FIGURE 3.4 – Nuage de points des y en fonction des x dans le cas où la variance est non constante

On remarque que les points de nuage forment une droite donc le modèle proposé est bon.

Remarque

— R^2 : coefficient de détermination .

— RSE : Residual standard erreur.

Estimation des paramètres :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.21045    0.02899   76.26   <2e-16 ***
x            6.98409    0.04978  140.30   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1325 on 98 degrees of freedom
Multiple R-squared:  0.995,     Adjusted R-squared:  0.995
F-statistic: 1.968e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

A partir des résultats d'estimation déterminer sous R, nous avons : $\beta_0 = 2.21045$, $\beta_1 = 6.98409$, RSE $= 0.1325$ et $R^2 = 0.995$.

On remarque que les valeur de R^2 est augment par rapport à le premier cas au la variance est constate, tandis que, la valeur de RSE diminue.
La droite de régression est définie par l'équation :

$$\hat{y} = 2.21045 + 6.98409x$$

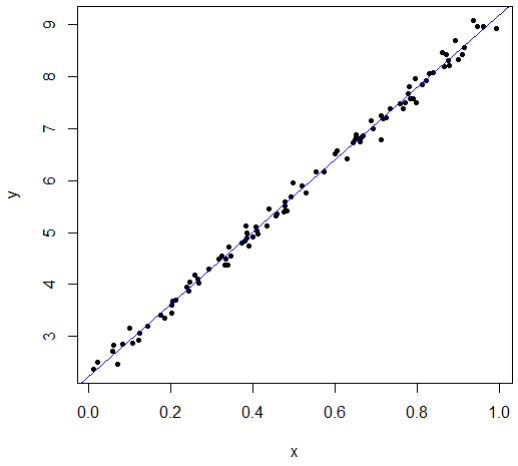


FIGURE 3.5 – Nuage de points avec la droite d'ajustement cas variance non constante

On remarque que les points sont parfaitement alignées sur la droite de régression donc le modèle semble bien.

Analyse des résidus

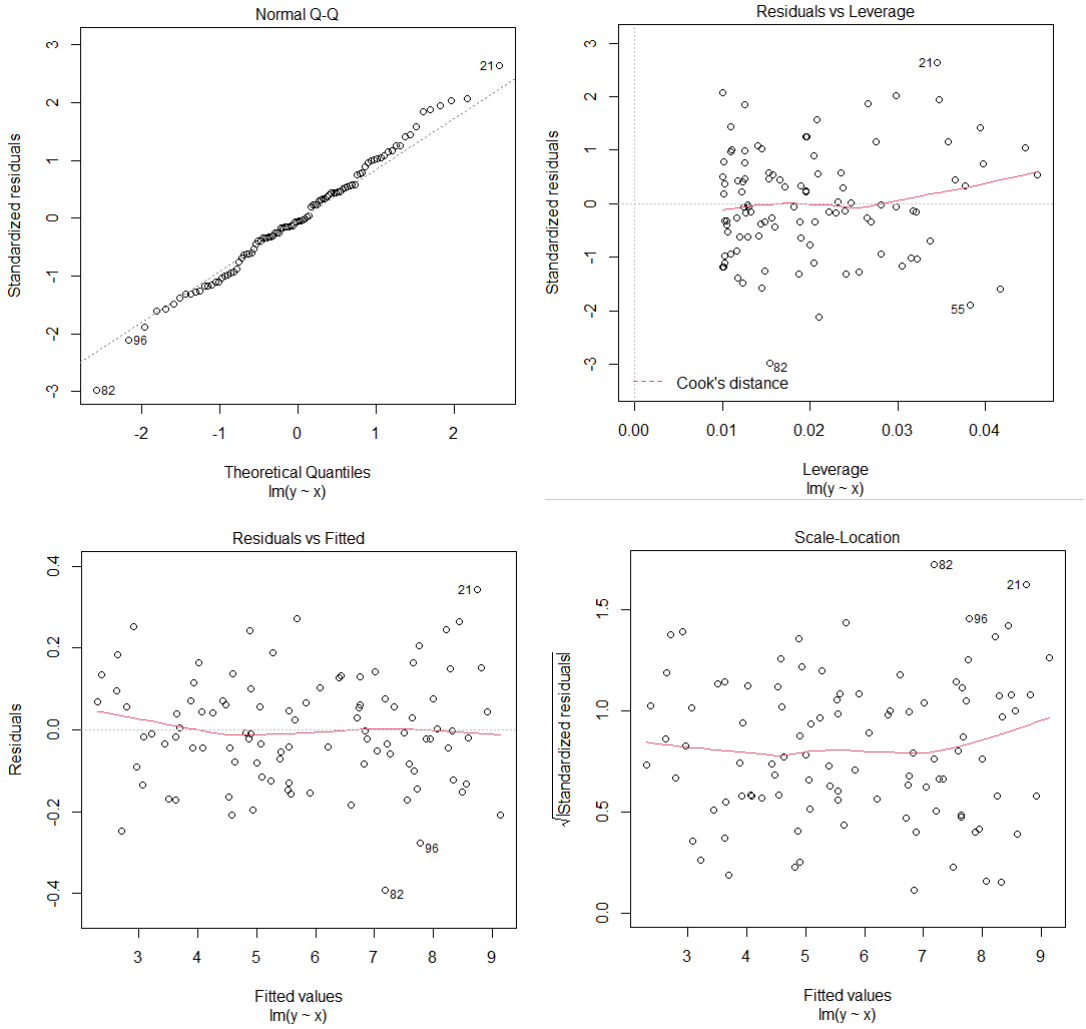


FIGURE 3.6 – Graphique des résidus

Le premier graphe Residuals vs Fitted : le graphe montre, qu'en moyenne, la droite de régression, est parfaitement adaptée aux données, et donc que l'hypothèse de linéarité est acceptable.

Le deuxième graphe Normal Q-Q : les résidus sont bien distribués le long de la droite figurant sur le graphe, alors l'hypothèse de normalité est

acceptée dans notre cas .

le troisième graphe Residuals vs Leverage : la ligne pointillée démarque le seuil de 1 pour la distance de cook dans notre cas aucun points est proche ou dépasse le 1 .donc aucun point est influente.
le quatrième graphique : Aucun motif discernable est remarquable ce la impliquant que les erreurs sont normalement distribué avec présence de l'homoscédasticitè.

Test d'hétéroscédasticité

```
        studentized Breusch-Pagan test

data:  model
BP = 3.5546, df = 1, p-value = 0.05938
```

On remarque que la p-value =0.059 est supérieur à 0.05 donc on rejette l'hypothèse de l'hétéroscédasticité.

Méthode moindre carré pondérer

Estimation des paramètres

```
Weighted Residuals:
    Min      1Q  Median      3Q     Max
-3.5722 -0.8125 -0.0860  0.7224  3.0054

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.21431    0.02688   82.37   <2e-16 ***
data$R       6.97650    0.04903  142.28   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.274 on 98 degrees of freedom
Multiple R-squared:  0.9952,    Adjusted R-squared:  0.9951
F-statistic: 2.024e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

A partir des résultats d'estimation déterminer sous **R**, nous avons :
$\beta_0 = 2.21431$, $\beta_1 = 6.97650$,RSE $= 1.274$ et $R^2 = 0.9952$.
On remarque une augmentation pour les valeurs de RSE et R^2 par rapport à la méthode MCO.

Méthode de MM-estimation

Estimation des paramètres

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.19526    0.02364   92.87   <2e-16 ***
x1           7.03716    0.04717  149.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.1299
Multiple R-squared:  0.9961,    Adjusted R-squared:  0.9961
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.195226$, $\beta_1 = 7.03716$, RSE $= 0.1299$ et $R^2 = 0.9961$.
On remarque que RSE est diminue et R^2 augment par rapport aux résultats de MCP.

Estimation par t-régression

Estimation des paramètres

```
Location model :

Call:
tlm(lform = y1 ~ x1, data = data, estDof = TRUE)

Residuals:
      Min         1Q     Median         3Q        Max
-12.04467   -0.09223   -0.01250    0.07287    8.05675

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.21190    0.01993   111.0   <2e-16 ***
x1           7.01498    0.03027   231.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Scale parameter(s) as estimated below)

Scale Model :

Call:
tlm(lform = y1 ~ x1, data = data, estDof = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-4.4620  -2.5888   0.2659   2.3103   3.8630

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.1467     0.2852  -18.04   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Scale parameter taken to be  2 )


Est. degrees of freedom parameter:  0.863286
Standard error for d.o.f:  0.1362664
No. of iterations of model : 41 in 0.3
Heteroscedastic t Likelihood : -15.27649
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.21190$, $\beta_1 = 7.01498$, RSE $= 0.13626$ t $R^2 = 0.99595$.

Estimation par régression quantiles

Estimation des paramètres

```
Call: rq(formula = y ~ x, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = data)

tau: [1] 0.1

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.03886      1.96479  2.11340
x           7.00406      6.90185  7.09280

Call: rq(formula = y ~ x, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = data)

tau: [1] 0.25

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.14922      2.02107  2.21176
x           6.93481      6.81605  7.14913

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.20122      2.17057  2.27902
x           6.98752      6.86680  7.07590

Call: rq(formula = y ~ x, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = data)

tau: [1] 0.75

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.28147      2.22759  2.34816
x           6.99056      6.89146  7.12785

Call: rq(formula = y ~ x, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = data)

tau: [1] 0.9

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.34402      2.29503  2.43507
x           7.07630      6.94959  7.15951
```

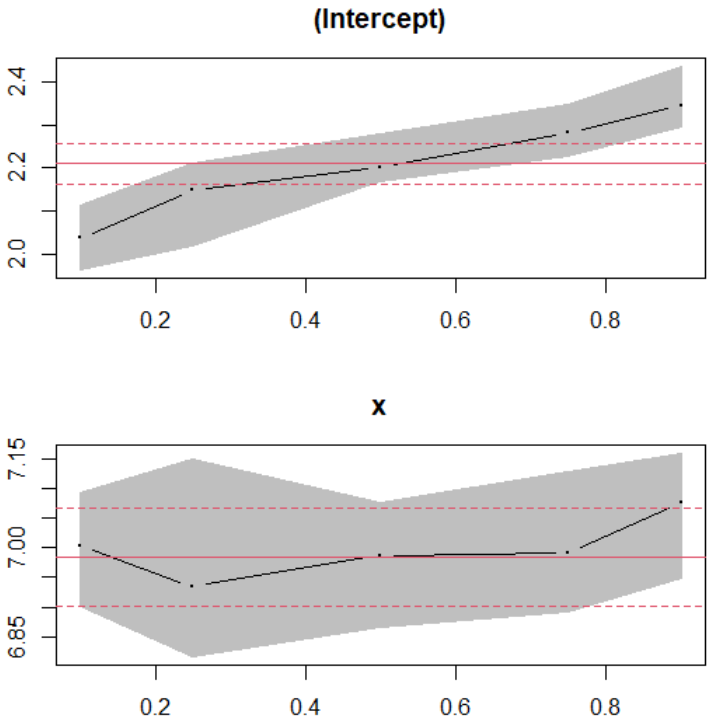


FIGURE 3.7 – Intervalle de confiance pour coefficients de modèle

A partir des résultats d'estimation déterminer sous **R**,et le graphe ci dessus,on remarque que le coefficient constant β_0 avoir une valeur initial 2.2 pour un quantile dont l'ordre est compris entre 0.4 et 0.6.
Pour la pente β_1 est proche de sa vraie valeur pour des quantiles dont l'ordre est compris entre 0.4 et 0.6.

Dans le cas où la variance des résidus est constante, nous avons appliquer que la méthode de moindre carrés ordinaire(MCO). Par contre dans le cas où la variance est variée, nous avons appliquer les différent méthodes d'estimation robustes.
Les résultats sont présenter dans le tableau que suit :

	MCO	MCP	MM	t-reg	reg-quant
R^2	0.9950	0.9952	0.9961	0.9959	0.9060
RSE	0.1325	1.274	0.1299	0.1362	0.1194

D'aprés le tableau ci dessus on remarque la méthode qui admet le plus petit RSE est la méthode de régression quantile, et la méthode qui a le plus

grand R^2 est la de méthode MM-estimation.

3.2 Simulation avec une seul valeur aberrante :

3.2.1 Avec variance constante :

moindre carrés ordinaire :

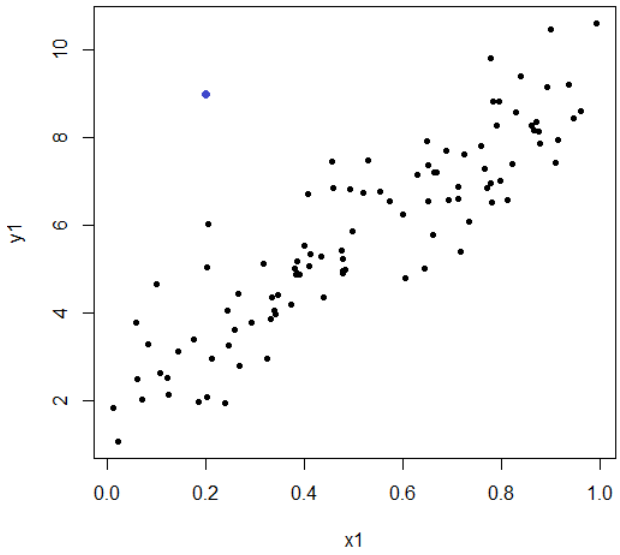


FIGURE 3.8 – Nuage de point avec 1% valeurs aberrantes

Un nuage de points peut également montre si les valeurs extrême sont présentes dans l'ensemble de données, les valeurs extrêmes sont celles qui sont éloignées des autres données de l'ensemble de données, comme le point en blue au figure précédant.

Estimation des paramètres

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1996     0.2346   9.377 2.49e-15 ***
x1            7.0708     0.4046  17.476  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.085 on 99 degrees of freedom
Multiple R-squared:  0.7552,    Adjusted R-squared:  0.7527
F-statistic: 305.4 on 1 and 99 DF,  p-value: < 2.2e-16
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.1996$, $\beta_1 = 7.0708$, RSE $= 1.085$ et $R^2 = 0.7552$.
On remarque RSE est augmenter et le R^2 diminue par rapport au cas sans valeur extrême.
La droite de régression est définie par l'équation :

$$\hat{y} = 2.1996 + 7.0708x$$

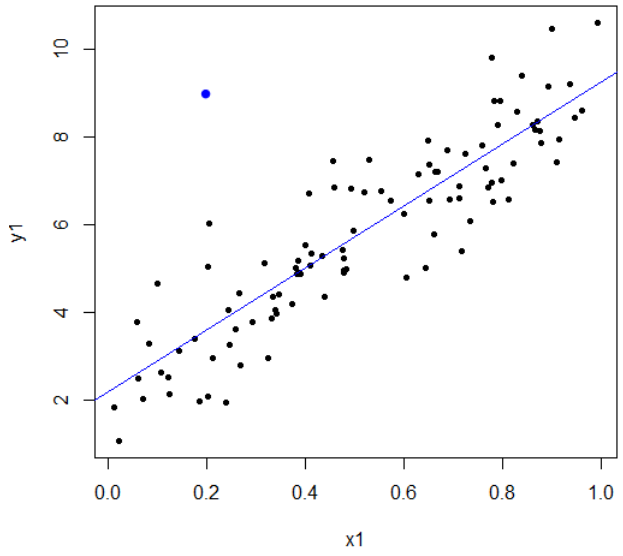


FIGURE 3.9 – Nuage de points avec droite de régression pour 1% de valeurs aberrantes

On remarque que les points sont bien alignées sur la droite de régression donc on a pas un influence par le point blue .

Analyse des résidus

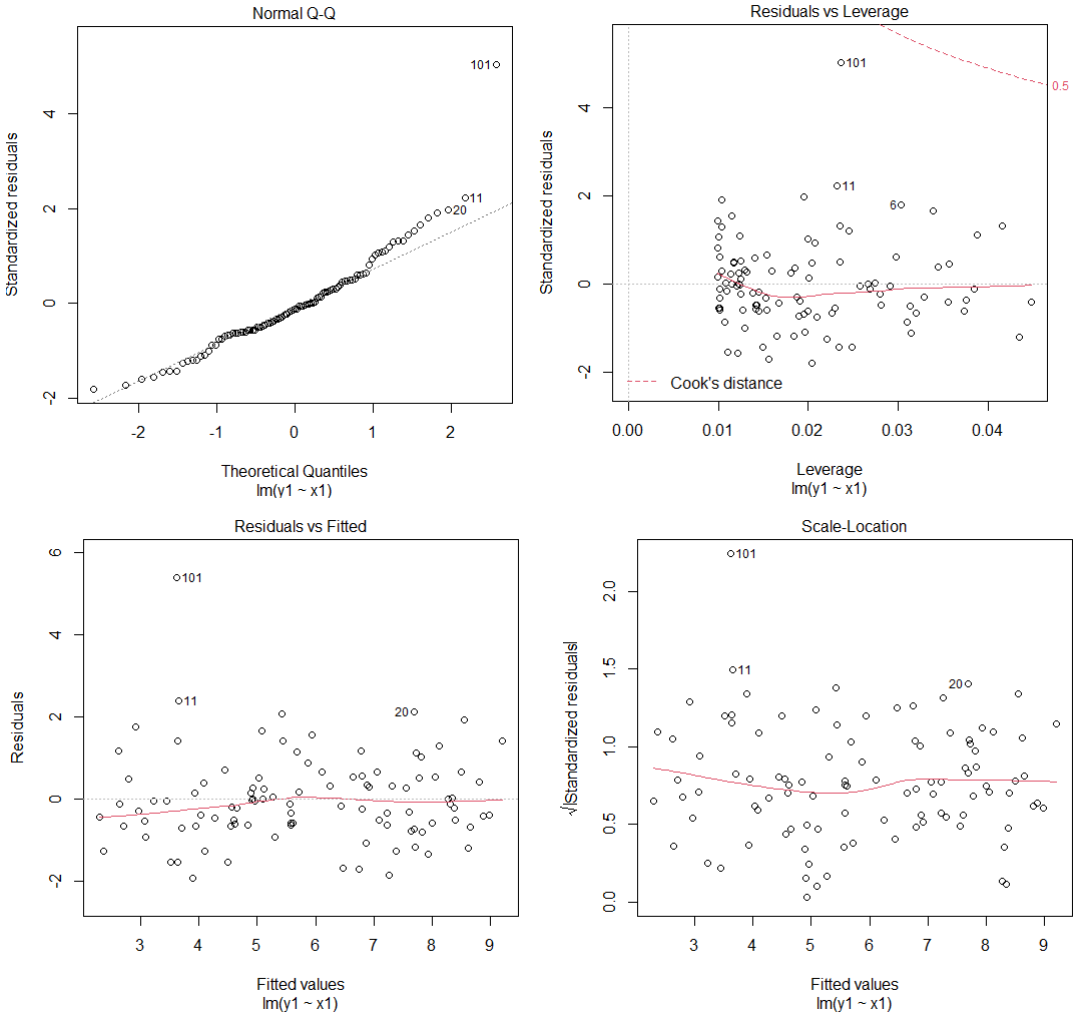


FIGURE 3.10 – Les graphiques des résidus

Le premier graphe Residuals vs Fitted : le graphe nous montre que lorsque les réponses prédites par le modèle (fitted values) augmentent, les résidus restent globalement uniformément distribués de part et d'autre de 0. Cela montre, qu'en moyenne, la droite de régression, est bien adaptée aux données, et donc que l'hypothèse de linéarité est acceptable.

Le deuxième graphe Normal Q-Q : Evaluation de l'hypothèse de normalité des résidus, cette hypothèse peut s'évaluer graphiquement à l'aide d'un QQplot. Si les résidus sont bien distribués le long de la droite figurant

sur le graphe , alors l'hypothèse de normalité on peut la considérer acceptée dans notre cas .

le troisième graphe Residuals vs Leverage : permet d'identifier les points avec une forte influence, la ligne pointillée démarque le seuil de 1 pour la distance de cook dans notre cas aucun points est proche ou dépasse le 1 .donc on peut considérer aucun valeur influent.

le quatrième graphique : Aucun motif discernable est remarquable ce la impliquant que les erreurs sont normalement distribué avec présence de l'homoscédasticitè.

Méthode moindre carré pondérer

Estimation des paramètres

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1954     0.2449   8.966 1.97e-14 ***
data$R        7.0789     0.3988  17.753  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.355 on 99 degrees of freedom
Multiple R-squared:  0.761,     Adjusted R-squared:  0.7585
F-statistic: 315.2 on 1 and 99 DF,  p-value: < 2.2e-16
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.1954$, $\beta_1 = 7.0789$, RSE $= 1.355$ et $R^2 = 0.761$.

Méthode MM-estimation

Estimation des paramètres

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9411     0.2124   9.138 8.29e-15 ***
xl            7.3837     0.3678  20.076  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.8876
Multiple R-squared:  0.8196,    Adjusted R-squared:  0.8178
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 1.9411$, $\beta_1 = 7.3837$, RSE $= 0.8876$ et $R^2 = 0.8196$.
On remarque que le RSE est diminue et R^2 est augmente par rapport au méthode précédente.

Mèthodes t-régression

```
Location model :

Call:
tlm(lform = yl ~ xl, data = data, estDof = TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-1.81170 -0.52187 -0.02203  0.59412  5.57542

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9542     0.2026   9.644 6.54e-16 ***
xl            7.3519     0.3495  21.034  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Scale Model :

Call:
tlm(lform = yl ~ xl, data = data, estDof = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6378 -1.3617 -0.8175  1.1629  6.8639

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4306     0.1801  -2.391   0.0168 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 1.9542$, $\beta_1 = 7.3519$, RSE $= 2.200651$ et $R^2 = 0.75185$.

Méthode de régression quantiles

Estimation des paramètres

```
Coefficients:
            coefficients lower bd upper bd
(Intercept) 0.89456      0.41391  1.29960
xl          7.19387      6.51153  8.27793

Call: rq(formula = yl ~ xl, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.25

Coefficients:
            coefficients lower bd upper bd
(Intercept) 1.53035      1.22632  1.70211
xl          7.15486      6.74342  7.49164

Call: rq(formula = yl ~ xl, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept) 1.92367      1.62892  2.32570
xl          7.36976      6.55833  8.17603

tau: [1] 0.75

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.46484      1.96237  3.32608
xl          7.50194      6.43192  8.58312

Call: rq(formula = yl ~ xl, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.9

Coefficients:
            coefficients lower bd upper bd
(Intercept) 3.73999      3.09963  4.91897
xl          6.76538      5.12007  7.66471
```

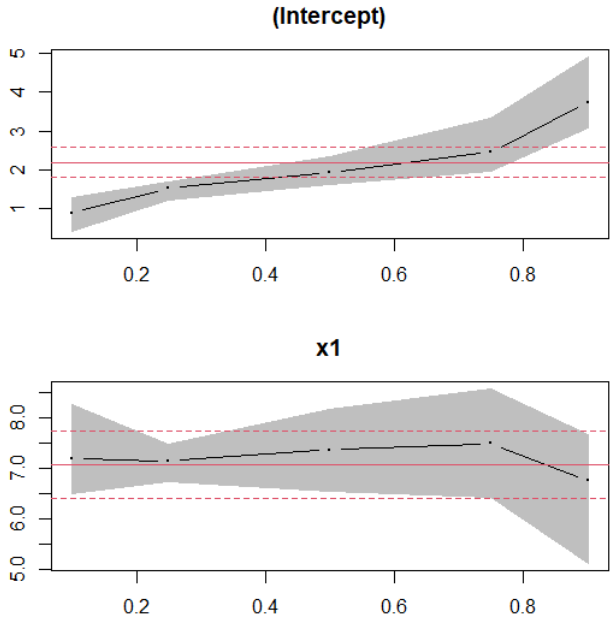


FIGURE 3.11 – Intervalle de confiance pour coefficient de modèle

A partir des résultats d'estimation déterminer sous **R**,et le graphe ci dessus,on remarque que le coefficient constant β_0 avoir une valeur initial 2.2 pour un quantile dont l'ordre est compris entre 0.6 et 0.8.
Pour la pente β_1 proche de sa vrai valeur pour des quantiles dont l'ordre est compris entre 0.8 et 1.

Les résultats obtenue dans ce cas sont résumé dans le tableau suivant :

	MCO	MCP	MM	t-reg	reg-quant
R^2	0.7552	0.761	0.8196	0.7518	0.75131
RSE	1.085	1.355	0.8876	1.0977	1.09894

Dans ce cas on remarque que la MM-estimation qui a une meilleur RSE et R^2.

3.3 Simulation avec 5% valeur aberrante :

3.3.1 Avec variance constante :

moindre carrés ordinaire :

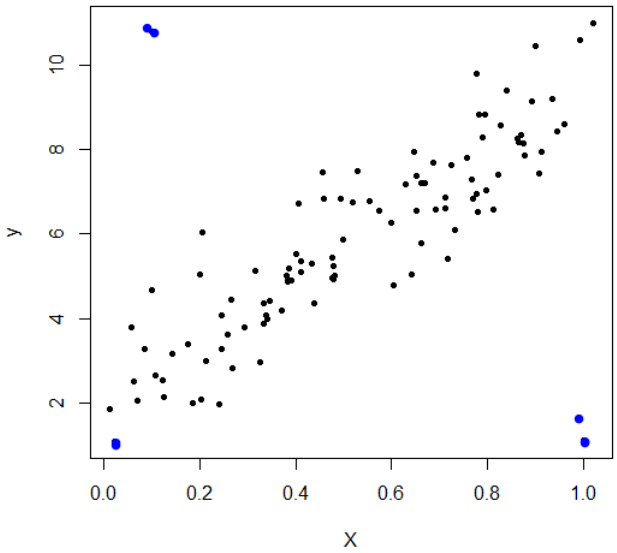


FIGURE 3.12 – Nuage de points des y en fonction des x pour 5% des valeurs aberrantes

Un nuage de points peut également montre si les valeurs extrême sont présentes dans l'ensemble de données, les valeurs extrêmes sont celles qui sont éloignées des autres données de l'ensemble de données, comme les 5 points en blue au figure ci-dessus.

Estimation des paramètres :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.8240     0.3494  10.946  < 2e-16 ***
x1            3.7305     0.5306   7.031 1.95e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.925 on 108 degrees of freedom
Multiple R-squared:  0.314,     Adjusted R-squared:  0.3076
F-statistic: 49.43 on 1 and 108 DF,  p-value: 1.947e-10
```

A partir des résultats d'estimation déterminer sous **R**, nous avons :
$\beta_0 = 3.8240$, $\beta_1 = 3.7305$, RSE $= 1.925$ et $R^2 = 0.314$.

On remarque les résultats sont similaire avec celle de cas où on a 1% valeurs aberrantes .
La droite de régression est définie par l'équation :

$$\hat{y} = 3.8240 + 3.7305x$$

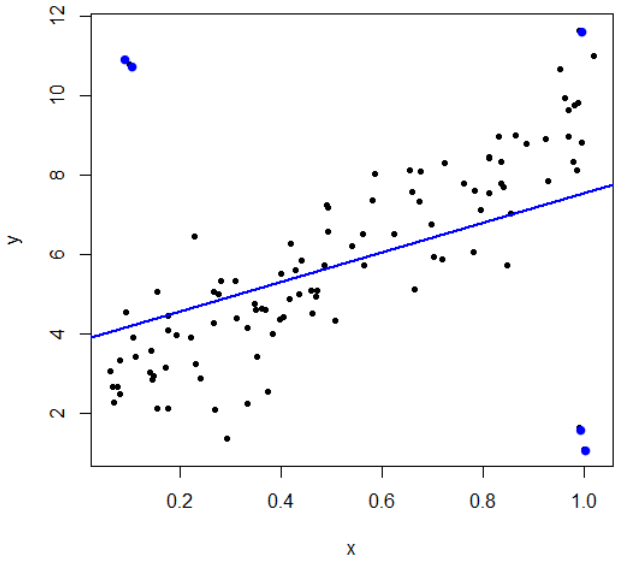


FIGURE 3.13 – Nuage de points avec droite d'ajustement pour 5% de valeurs extrêmes

On remarque que les points ne sont pas bien alignées sur la droite de régression donc on a une influence par les points blue .
Analyse des résidus

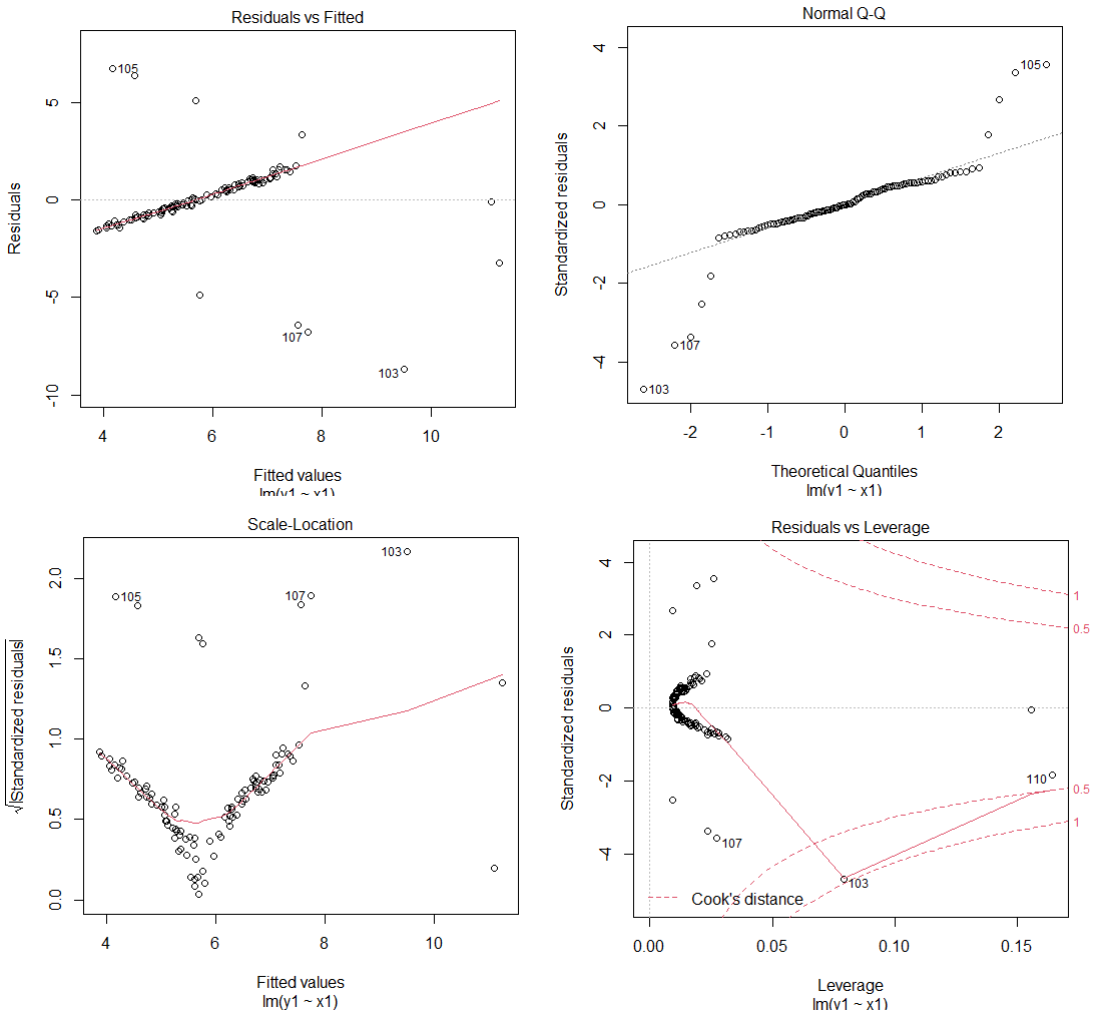


FIGURE 3.14 – Graphiques des résidus

Le premier graphe Residuals vs Fitted : le graphe nous montre qu'en moyenne, la droite de régression, est n'adaptée pas bien aux données, et donc que l'hypothèse de linéarité n'est pas acceptable.

Le deuxième graphe Normal Q-Q : Si les résidus sont bien distribués le long de la droite figurant sur le graphe , alors l'hypothèse de normalité on peut pas la considérer acceptée dans notre cas .

le troisième graphe Residuals vs Leverage : On remarque plusieurs points éloignés des autre, et sont proche de 0.5 et 1 parmi les quelle (103,107,110), donc on peut les considérer comme valeurs influentes.

Le quatrième graphique : Le graphique montre un motif discernable impliquerai une présence de l'hétéroscédasticité.

Test d'hétéroscédasticité :

```
        studentized Breusch-Pagan test

data:  model
BP = 0.0013846, df = 1, p-value = 0.9703
```

On remarque que la p-value =0.9703 est supérieur à 0.05 donc on rejette l'hypothèse de l'hétéroscédasticité.

Estimation par moindres carrés pondéreés

Estimation des paramétres

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.4592     0.3197   7.691 9.04e-12 ***
data$R        6.6821     0.5857  11.409  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.669 on 103 degrees of freedom
Multiple R-squared:  0.5583,    Adjusted R-squared:  0.554
F-statistic: 130.2 on 1 and 103 DF,  p-value: < 2.2e-16
```

A partir des résultats d'estimation déterminer sous **R**, nous avons :
$\beta_0 = 2.4592$, $\beta_1 = 6.6821$, RSE = 1.696 et $R^2 = 0.5583$.
On remarque que les valeurs de la pente et R^2 sont augmentes,tandis que la valeur de RSE diminue.

Estimation par MM-estimation

Estimation des paramètres

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.19526    0.02364   92.87   <2e-16 ***
xl           7.03716    0.04717  149.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.1299
Multiple R-squared:  0.9961,    Adjusted R-squared:  0.9961
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.19526$, $\beta_1 = 7.03716$, RSE $= 0.1299$ et $R^2 = 0.9961$.

Estimation Mèthodes t-régression

```
Location model :

Call:
tlm(lform = yl ~ xl, data = data, estDof = TRUE)

Residuals:
      Min         1Q     Median         3Q        Max
-12.04467   -0.09223   -0.01250    0.07287    8.05675

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.21190    0.01993   111.0   <2e-16 ***
xl           7.01498    0.03027   231.7   <2e-16 ***
Scale Model :

Call:
tlm(lform = yl ~ xl, data = data, estDof = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-4.4620  -2.5888   0.2659   2.3103   3.8630

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.1467     0.2852  -18.04   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Scale parameter taken to be  2 )
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.2119$, $\beta_1 = 7.01498$, RSE $= 0.863286$.

Estimation par régression quantiles

Estimation des paramètres

```
tau: [1] 0.1

Coefficients:
            coefficients lower bd upper bd
(Intercept)  2.65211      2.53536  2.69911
xl           2.70246     -1.46596  4.38648

Call: rq(formula = yl ~ xl, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.25

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.23755      2.18114  2.30016
xl          6.74679      6.31092  6.90370

Call: rq(formula = yl ~ xl, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.25783      2.21784  2.33014
xl          6.89310      6.77460  7.03238

tau: [1] 0.75

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.26732      2.21113  2.35124
xl          7.03810      6.90621  7.14901

Call: rq(formula = yl ~ xl, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.9

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.34285      2.24420  6.83829
xl          7.07812      6.87529  7.28078
```

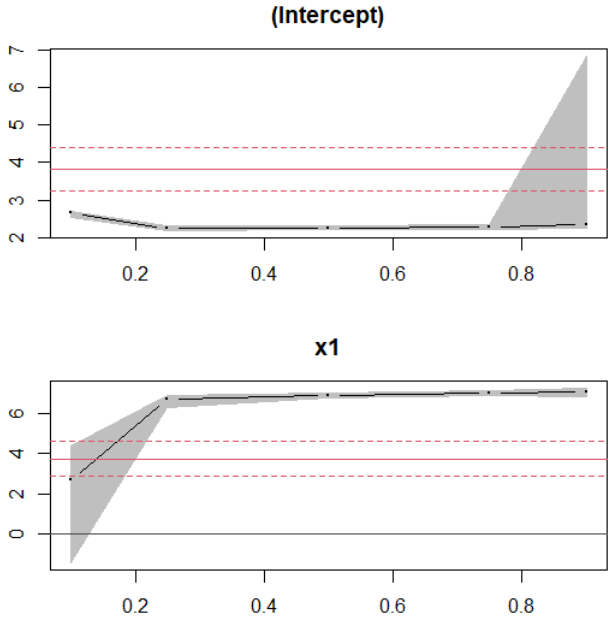


FIGURE 3.15 – Intervalle de confiance pour coefficients de modéle

A partir des résultats d'estimation déterminer sous **R**,et le graphe ci dessus,on remarque que le coefficient constant β_0 avoir une valeur initial 2.2 pour un quantile dont l'ordre est compris entre 0.2 et 0.4. Pour la pente β_1 est proche de sa vraie valeur pour des quantiles dont l'ordre est compris entre 0.8 et 1

Les résultats obtenue dans ce cas sont résumé dans le tableau suivant :

	MCO	MCP	MM	t-reg	reg-quant
R^2	0.314	0.5583	0.9961	0.4169	0.4134
RSE	1.925	1.669	0.1299	0.8632	1.8674

3.4 Simulation avec 10% valeur aberrante :

3.4.1 Avec variance constante :

moindre carrés ordinaire :

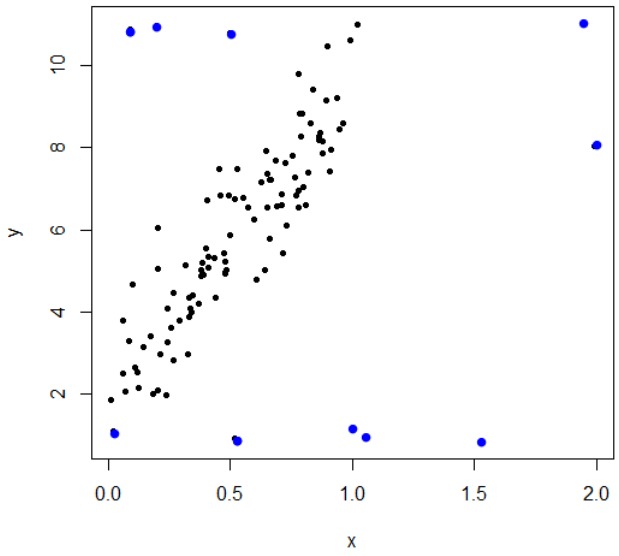


FIGURE 3.16 – Nuage de points des y en fonction des x pour 10 % valeurs aberrantes

Un nuage de points peut également montre si les valeurs extrême sont présentes dans l'ensemble de données, les valeurs extrêmes sont celles qui sont éloignées des autres données de l'ensemble de données, comme les 10 points en blue au figure ci-dessus.

Estimation des paramètres :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7103     0.3895   9.525 5.62e-16 ***
xl            3.8809     0.5916   6.560 1.92e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.147 on 108 degrees of freedom
Multiple R-squared:  0.2849,    Adjusted R-squared:  0.2783
F-statistic: 43.03 on 1 and 108 DF,  p-value: 1.918e-09
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 3.7103$, $\beta_1 = 3.8809$, RSE $= 2.147$ et $R^2 = 0.2849$.
La droite de régression est définie par l'équation :

$$\hat{y} = 3.7103 + 3.8809x$$

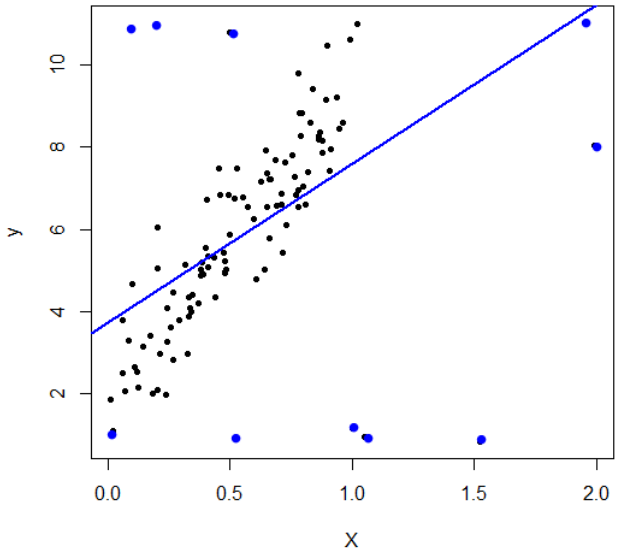


FIGURE 3.17 – Nuage de points avec la droite d'ajustement pour 10% valeurs aberrantes

On remarque que la droite d'ajustement ne passe pas par le centre de nuage donc on a une influence par les points blue.
Analyse des résidus

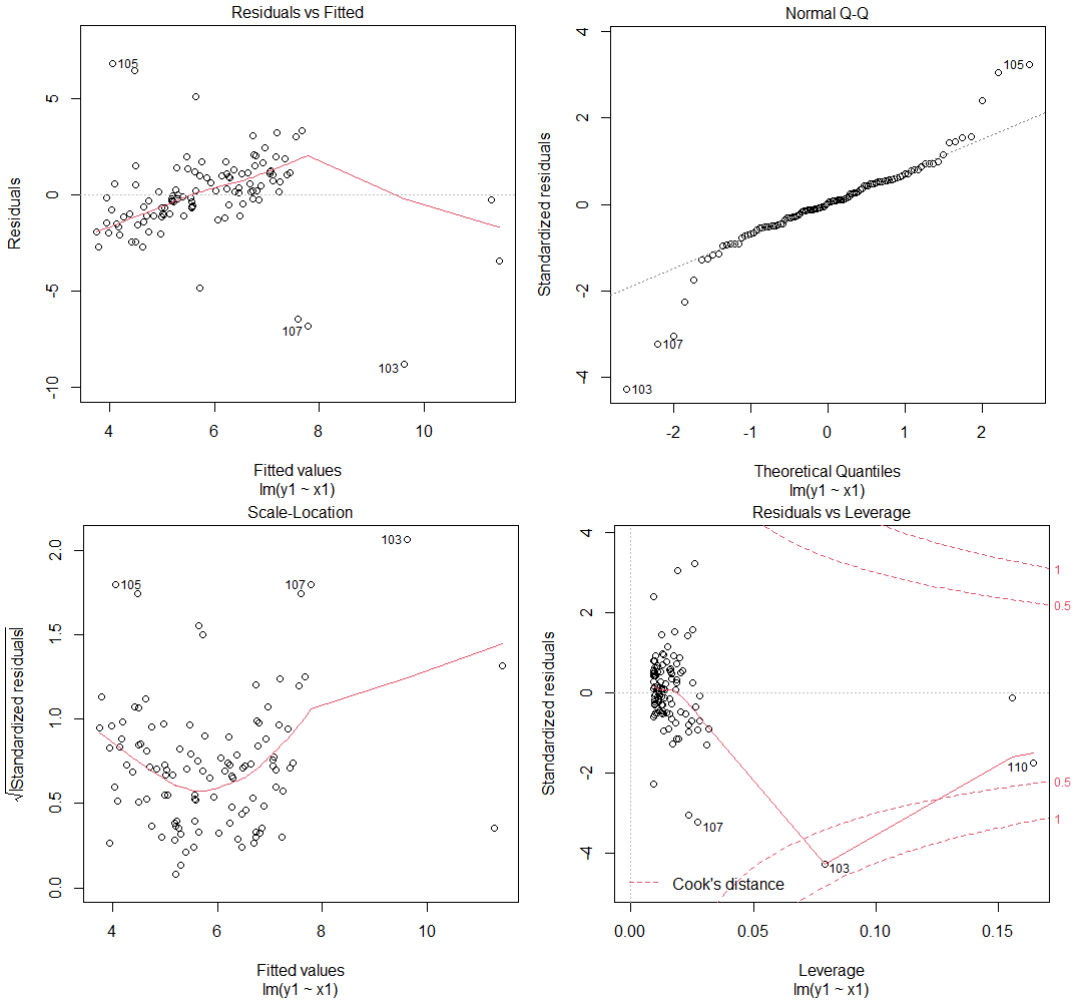


FIGURE 3.18 – Graphique des résidus

Le premier graphe Residuals vs Fitted : le graphe nous montre que la droite de régression, est n'adapte pas aux données, et donc que l'hypothèse de linéarité n'est pas acceptable.

Le deuxième graphe Normal Q-Q : les résidus sont bien distribués le long de la droite figurant sur le graphe, alors l'hypothèse de normalité on peut pas la considérer acceptée dans notre cas.

le troisième graphe Residuals vs Leverage : On remarque plusieurs points éloignés des autre, et sont proche de 0.5 et 1 parmi les quelle (103,107,110), donc on peut les considérer comme valeurs influentes.

Le quatrième graphique : Le graphique montre un motif discernable impliquerai une présence de l'hétéroscédasticité.

Test d'hétéroscédasticité

```
data:  model
BP = 4.9737, df = 1, p-value = 0.02574
```

On remarque que la p-value =0.02574 est inférieur à 0.05 donc on accepte l'hypothèse de l'hétéroscédasticité.

Estimation par moindres carrés pondéreés

Estimation des paramétres

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1650     0.3487   9.076 5.86e-15 ***
data$R        4.9019     0.6446   7.604 1.12e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.474 on 108 degrees of freedom
Multiple R-squared:  0.3487,    Adjusted R-squared:  0.3427
F-statistic: 57.82 on 1 and 108 DF,  p-value: 1.122e-11
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 3.165$, $\beta_1 = 4.9019$, RSE $= 1.474$ et $R^2 = 0.3487$.

Estimation par MM-estimation

Estimation des paramètres

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9199     0.2130   9.013 8.14e-15 ***
xl            7.4740     0.3722  20.083  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 1.013
Multiple R-squared:  0.8161,    Adjusted R-squared:  0.8144
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 1.9199$, $\beta_1 = 7.4740$, RSE $= 1.013$ et $R^2 = 0.8161$.
On remarque que RSE diminue et R^2 augmente.

Estimation Mèthodes t-régression

```
Location model :

Call:
tlm(lform = y1 ~ x1, data = data, estDof = TRUE)

Residuals:
      Min         1Q     Median         3Q        Max
-12.13671   -0.51727   -0.03202    0.66865    8.28549

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9630     0.1769   11.10   <2e-16 ***
x1            7.2393     0.2687   26.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Scale Model :

Call:
tlm(lform = y1 ~ x1, data = data, estDof = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-3.0072  -2.1260  -0.5271   2.2151   4.4513

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6401     0.2340  -2.735  0.00623 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Scale parameter taken to be  2 )


Est. degrees of freedom parameter:  1.491074
Standard error for d.o.f:  0.3046851
No. of iterations of model : 16 in 0.07
Heteroscedastic t Likelihood : -201.7555
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 1.963$, $\beta_1 = 7.2393$, RSE $= 0.346$.

Estimation par régression quantiles

Estimation des paramètres

```
Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.1

Coefficients:
            coefficients lower bd upper bd
(Intercept)  1.81959      1.57529  2.08038
x1           3.12081     -1.02938  4.87414

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.25

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.02641      1.73399  2.34576
x1          5.79021      4.36828  6.82489

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)
Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.75

Coefficients:
            coefficients lower bd upper bd
(Intercept) 3.02248      2.10149  4.38256
x1          6.66555      4.74279  8.52223

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.9

Coefficients:
            coefficients lower bd upper bd
(Intercept) 4.20110      3.49367  7.77183
x1          6.21607      1.90855  7.47597
```

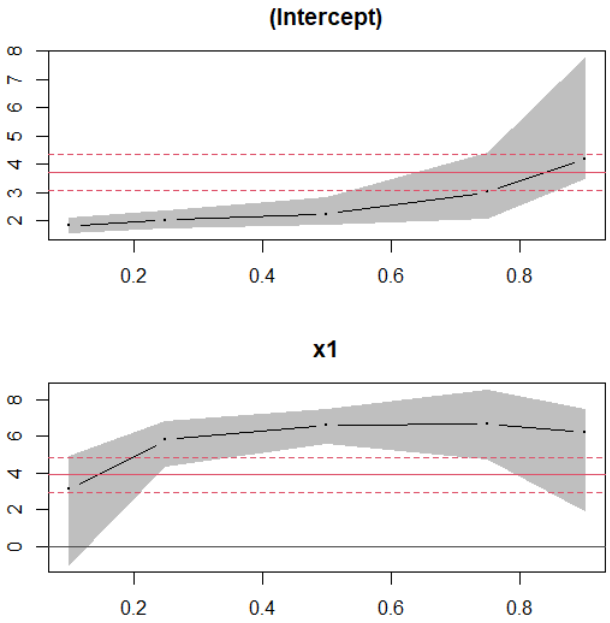


FIGURE 3.19 – Intervalle de confiance pour coefficient de modèle

A partir des résultats d'estimation déterminer sous **R**,et le graphe ci dessus,on remarque que le coefficient constant β_0 avoir une valeur initial 2.2 pour un quantile dont l'ordre est compris entre 0.2 et 0.4. Pour la pente β_1 est proche de sa vraie valeur pour des quantiles dont l'ordre est compris entre 0.4 et 0.8

Les résultats obtenue dans ce cas sont résumé dans le tableau suivant :

	MCO	MCP	MM	t-reg	reg-quant
R^2	0.2849	0.3487	0.8161	0.0687	0.0755
RSE	2.147	1.474	1.013	0.346	2.5623

3.5 Simulation avec 5% valeur aberrante :

3.5.1 Avec variance constante :

Moindre carrés ordinaire :

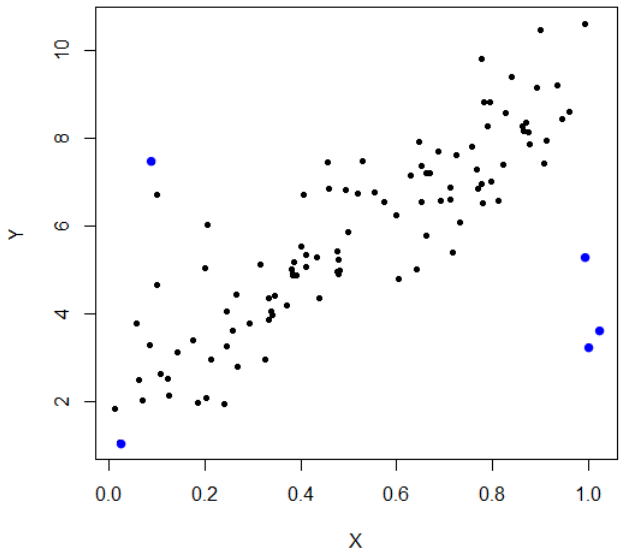


FIGURE 3.20 – Nuage de point des y en fonction des x

Un nuage de points peut également montre si les valeurs extrême sont présentes dans l'ensemble de données, les valeurs extrêmes sont celles qui sont éloignées des autres données de l'ensemble de données, comme les 5 points en blue au figure ci-dessus.

Estimation des paramètres :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6861     0.2866   9.371 1.85e-15 ***
xl            5.9122     0.4832  12.235  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.38 on 103 degrees of freedom
Multiple R-squared:  0.5924,    Adjusted R-squared:  0.5884
F-statistic: 149.7 on 1 and 103 DF,  p-value: < 2.2e-16
```

A partir des résultats d'estimation déterminer sous **R**, nous avons :

$\beta_0 = 2.6861$, $\beta_1 = 5.9122$, RSE $= 1.38$ et $R^2 = 0.5924$.

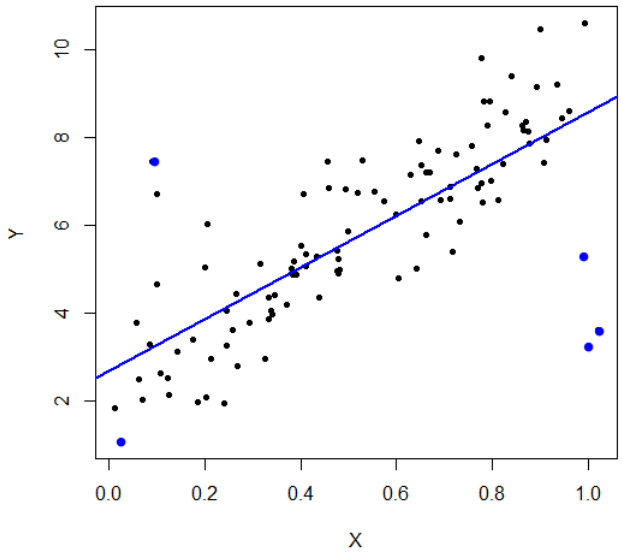


FIGURE 3.21 – Nuage de point avec droite d'ajustement

On remarque que la droite ne passe par le centre de nuage de points a cause de l'influence des valeurs aberrants en bleu. **Analyse des résidus**

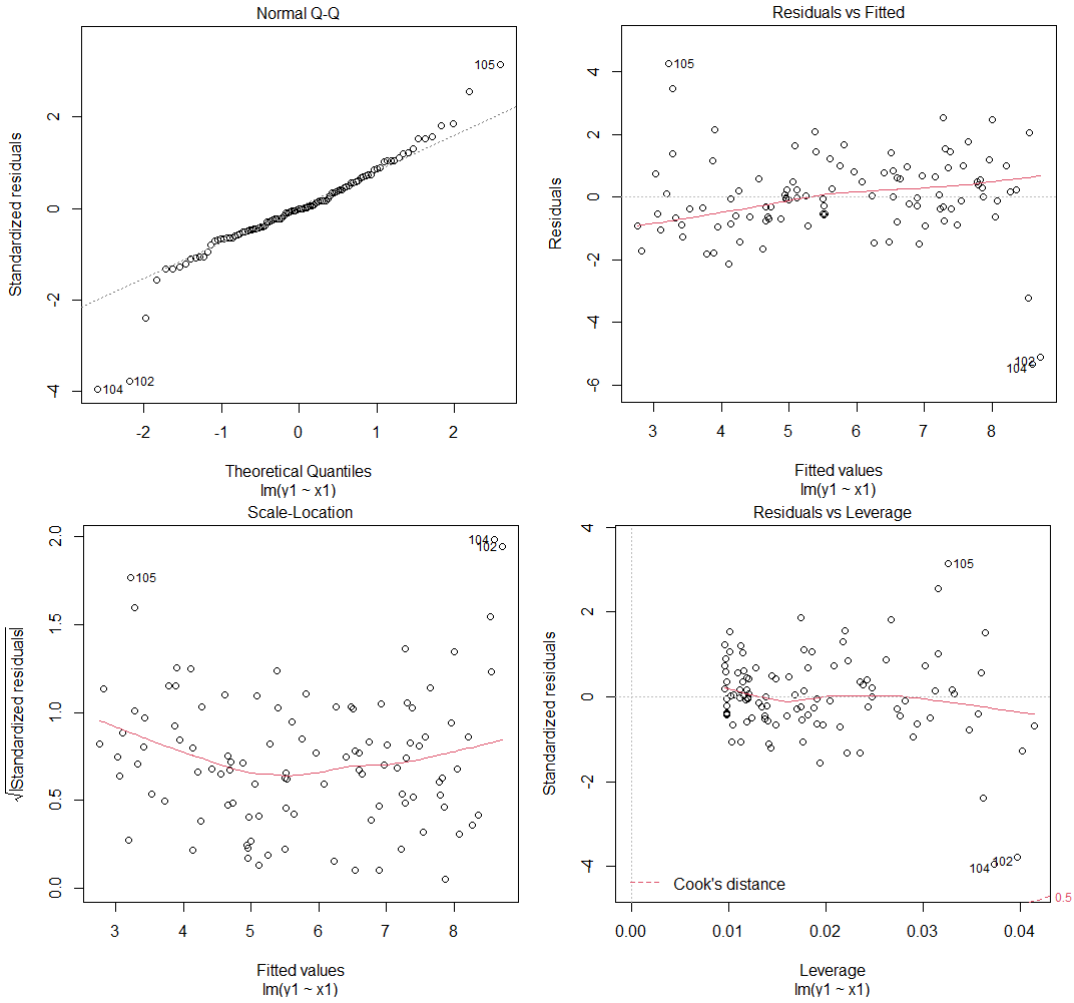


FIGURE 3.22 – Graphiques des résidus

Le premier graphe Residuals vs Fitted : le graphe nous montre que lorsque les réponses prédites par le modèle (fitted values) augmentent, les résidus restent globalement uniformément distribués de part et d'autre de 0. Cela montre, qu'en moyenne, la droite de régression, est bien adaptée aux données, et donc que l'hypothèse de linéarité est acceptable.

Le deuxième graphe Normal Q-Q : Evaluation de l'hypothèse de normalité des résidus, cette hypothèse peut s'évaluer graphiquement à l'aide d'un QQplot. Si les résidus sont bien distribués le long de la droite figurant sur le graphe , alors l'hypothèse de normalité on peut la considérer acceptée dans notre cas .

le troisième graphe Residuals vs Leverage : permet d'identifier les points avec une forte influence, la ligne pointillée démarque le seuil de 1 pour la distance de cook dans notre cas aucun points est proche ou dépasse le 1 .donc on peut considérer aucun valeur influent.

le quatrième graphique : Aucun motif discernable est remarquable ce la impliquant que les erreurs sont normalement distribué avec présence de l'homoscédasticitè.

Méthode moindre carré pondérer

Estimation des paramètres

Estimation des paramètres :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6398     0.2659   9.928   <2e-16 ***
data$R        6.0022     0.4755  12.623   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.425 on 103 degrees of freedom
Multiple R-squared:  0.6074,    Adjusted R-squared:  0.6036
F-statistic: 159.3 on 1 and 103 DF,  p-value: < 2.2e-16
```

A partir des résultats d'estimation obtenue par **R**, nous avons : $\beta_0 = 2.6398$,$\beta_1 = 6.002$,RSE $= 1.425$ et $R^2 = 0.607$

Estimation par MM-estimation

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9618     0.2254   8.702 5.62e-14 ***
xl            7.3524     0.3872  18.989  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.9414
Multiple R-squared:  0.8124,    Adjusted R-squared:  0.8106
```

A partir des résultats d'estimation obtenue par **R**, nous avons : $\beta_0 = 1.9618$,$\beta_1 = 7.3524$,RSE $= 0.9414$ et $R^2 = 0.8124$

Estimation Mèthodes t-régression

```
Location model :

Call:
tlm(lform = y1 ~ x1, data = data, estDof = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-5.9294  -0.5124  -0.0424   0.6476   4.8290

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9973     0.2022   9.879   <2e-16 ***
x1            7.1985     0.3409  21.119   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Scale parameter(s) as estimated below)


Scale Model :

Call:
tlm(lform = y1 ~ x1, data = data, estDof = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-2.2170  -1.7424  -0.9159   1.7117   5.1544

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.5104     0.2055  -2.483    0.013 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Scale parameter taken to be  2 )


Est. degrees of freedom parameter:  2.464284
Standard error for d.o.f:  0.7002051
No. of iterations of model : 9 in 0.03
Heteroscedastic t Likelihood : -167.9015
```

A partir des résultats d'estimation obtenue par **R**, nous avons : $\beta_0 = 1.9973$,$\beta_1 = 7.1985$,RSE $= 0.7002$

Estimation par régression quantiles

Estimation des paramètres

```
Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.1

Coefficients:
            coefficients lower bd upper bd
(Intercept) 1.23943      0.65561  1.74004
x1          5.90329      3.62333  7.50573

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.25

Coefficients:
            coefficients lower bd upper bd
(Intercept) 1.65192      1.43037  1.88082
x1          6.82815      6.03273  7.22762

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.75

Coefficients:
            coefficients lower bd upper bd
(Intercept) 3.02248      2.10149  4.38256
x1          6.66555      4.74279  8.52223

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.9

Coefficients:
            coefficients lower bd upper bd
(Intercept) 4.20110      3.49367  7.77183
x1          6.21607      1.90855  7.47597
```

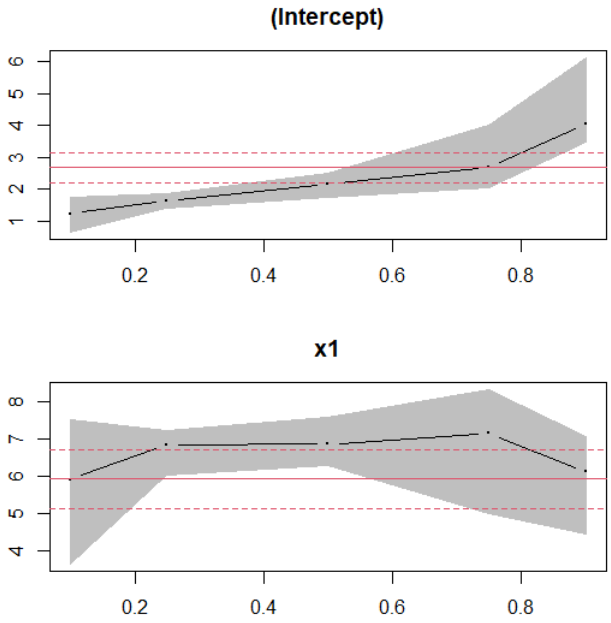


FIGURE 3.23 – Intervalle de confiance pour coefficients de modèle

A partir des résultats d'estimation déterminer sous **R**,et le graphe ci dessus,on remarque que le coefficient constant β_0 avoir une valeur initial 2.2 pour un quantile dont l'ordre est compris entre 0.6 et 0.8.
Pour la pente β_1 proche de sa vraie valeur pour des quantiles dont l'ordre est compris entre 0.2 et 0.4.
Les résultats obtenue dans ce cas sont résumé dans le tableau suivant :

	MCO	MCP	MM	t-reg	reg-quant
R^2	0.592	0.607	0.8124	0.5642	0.432
RSE	1.38	1.425	0.94	0.7002	1.668

3.6 Simulation avec 10% valeur aberrante :

3.6.1 Avec variance constante :

Moindre carrés ordinaire :

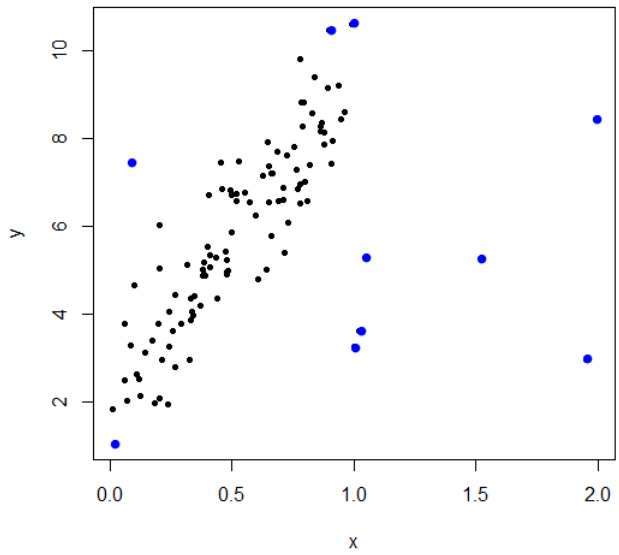


FIGURE 3.24 – Nuage de points des y en fonction des x

Le graphique représente la nuage de points de données contient des valeurs extrême qui sont colorés par blue. **Estimation des paramètres :**

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.6947     0.3127  11.814  < 2e-16 ***
xl            3.6968     0.4750   7.783 4.54e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.723 on 108 degrees of freedom
Multiple R-squared:  0.3594,    Adjusted R-squared:  0.3534
F-statistic: 60.58 on 1 and 108 DF,  p-value: 4.54e-12
```

A partir des résultats d'estimation déterminer sous **R**, nous avons :
$\beta_0 = 3.6947$, $\beta_1 = 3.6968$, RSE $= 1.723$ et $R^2 = 0.3594$.
Donc la droite de régression est définie par l'équation :

$$\hat{y} = 3.6947 + 3.6968x$$

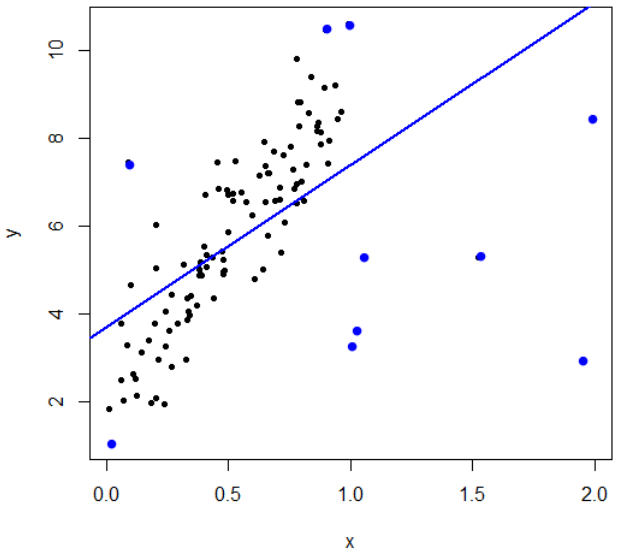


FIGURE 3.25 – Nuage de point avec droite d'ajustement

On remarque que la droite régression ne passe pas par le centre de nuage.

Analyse des résidus

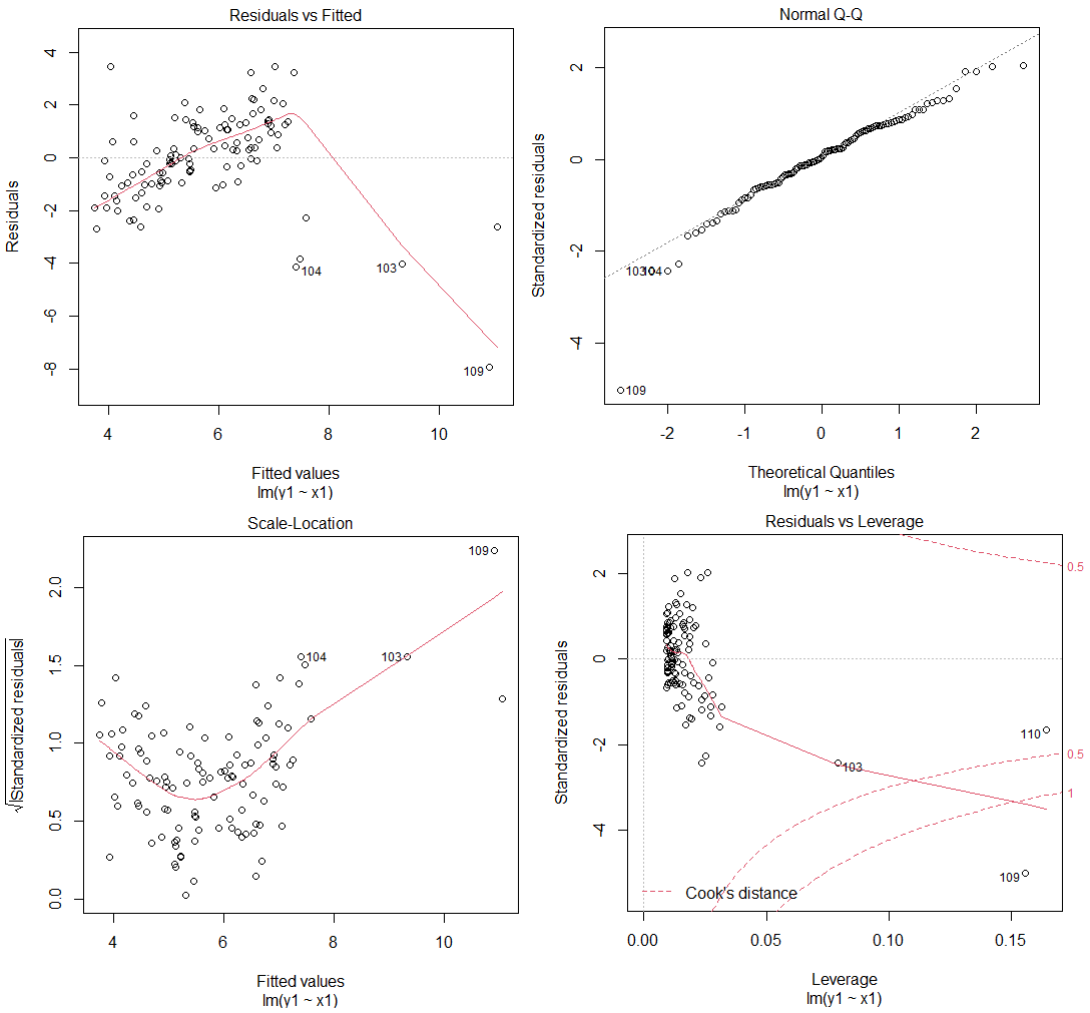


FIGURE 3.26 – Graphiques des résidus

Le premier graphe Residuals vs Fitted : le graphe nous montre que la droite de régression, est n'adapte pas aux données, et donc que l'hypothèse de linéarité n'est pas acceptable.

Le deuxième graphe Normal Q-Q : les résidus sont bien distribués le long de la droite figurant sur le graphe, alors l'hypothèse de normalité on peut pas la considérer acceptée dans notre cas .

le troisième graphe Residuals vs Leverage : On remarque plusieurs points éloignés des autre, et on a un point qui dépasse le seil 1(109), donc on peut les considérer comme valeurs influentes.

Le quatrième graphique : Le graphique montre un motif discernable impliquerai une présence de l'hétéroscédasticité.

Test d'hétéroscédasticité

```
data:  model
BP = 29.547, df = 1, p-value = 5.457e-08
```

On remarque que la p-value = 5.457e-08 est inférieur à 0.05 donc on accepte l'hypothèse de l'hétéroscédasticité.

Méthode moindre carré pondérer

Estimation des paramètres :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.4592     0.3197   7.691 9.04e-12 ***
data$R        6.6821     0.5857  11.409  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.669 on 103 degrees of freedom
Multiple R-squared:  0.5583,    Adjusted R-squared:  0.554
F-statistic: 130.2 on 1 and 103 DF,  p-value: < 2.2e-16
```

A partir des résultat d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.4592$, $\beta_1 = 6.6821$, RSE = 1.669 et $R^2 = 0.5583$

Estimation par MM-estimation

Estimation des paramètres

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.20124    0.02382   92.42   <2e-16 ***
x1           7.02895    0.04726  148.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.1298
Multiple R-squared:  0.9961,    Adjusted R-squared:  0.9961
```

A partir des résultat d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.20124$, $\beta_1 = 7.02895$, RSE $= 0.1298$ et $R^2 =$

Estimation Mèthodes t-régression

```
Location model :

Call:
tlm(lform = y1 ~ x1, data = data, estDof = TRUE)

Residuals:
      Min          1Q      Median          3Q         Max
-12.90744   -0.09081   -0.01147    0.07218    4.62802

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.21527    0.02046   108.3   <2e-16 ***
x1           7.00883    0.03107   225.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Scale Model :

Call:
tlm(lform = y1 ~ x1, data = data, estDof = TRUE)

Residuals:
     Min        1Q    Median        3Q       Max
-4.04222  -2.58909   0.05464   2.20202   3.98580

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.0620     0.2711  -18.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Scale parameter taken to be  2 )


Est. degrees of freedom parameter:  0.9861215
Standard error for d.o.f:  0.1645651
No. of iterations of model : 30 in 0.23
Heteroscedastic t Likelihood : -1.773079
```

A partir des résultat d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.21527$, $\beta_1 = 7.0088$, RSE $= 0.1645$

Estimation par régression quantiles

Estimation des paramètres

```
Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.1

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.28170      1.82001  2.34110
x1          1.98754      1.18285  3.33678

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.25

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.15986      1.76927  2.45591
x1          5.46318      2.96715  6.67004

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.75

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.96367      2.16607  4.17982
x1          6.69171      4.75235  8.15302

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.9

Coefficients:
            coefficients lower bd upper bd
(Intercept) 4.05995      2.97595  5.64594
x1          6.06839      4.35999  7.29177
```

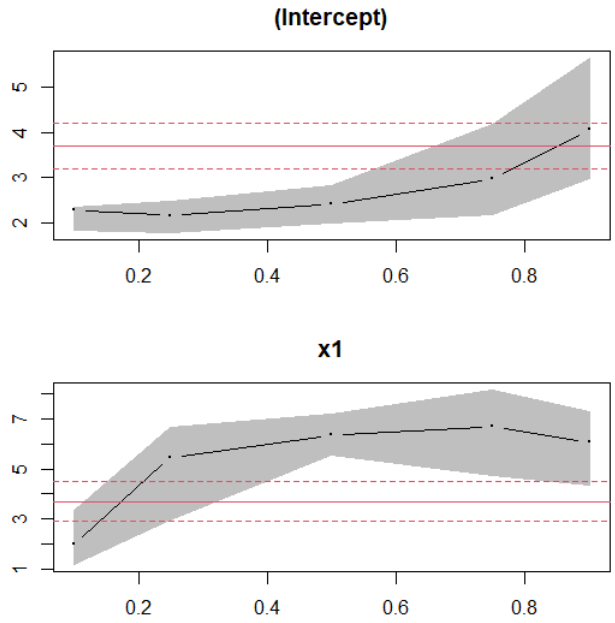


FIGURE 3.27 – Intervalle de confiance pour coefficients de modèle

A partir des résultats d'estimation déterminer sous R, et le graphe ci dessus, on remarque que le coefficient constant β_0 avoir une valeur initial 2.2 pour un quantile dont l'ordre est compris entre 0.2 et 0.4.
Pour la pente β_1 proche de sa vraie valeur pour des quantiles dont l'ordre est compris entre 0.6 et 0.8.
Les résultats obtenue dans ce cas sont résumé dans le tableau suivant :

	MCO	MCP	MM	t-reg	reg-quant
R^2	0.35	0.5583	0.9961	0.039	0.119
RSE	1.72	1.669	0.1298	0.1645	2.2115

3.7 Simulation avec 5% valeur aberrante :

3.7.1 Avec variance constante :

moindre carrés ordinaire :

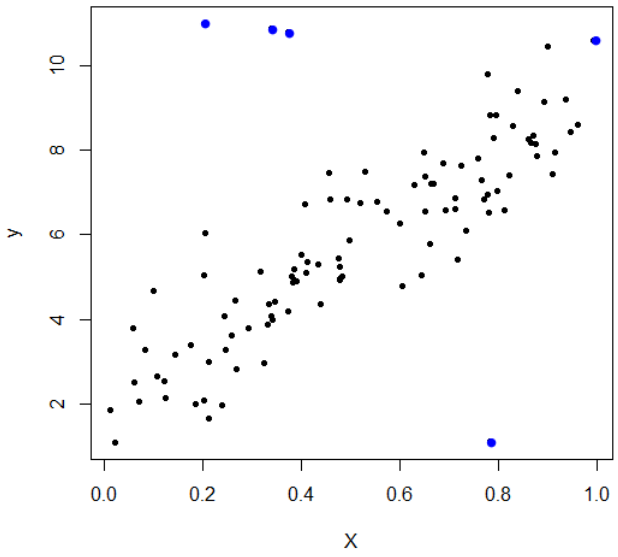


FIGURE 3.28 – Nuage de points des y en fonction des x

Un nuage de points peut également montre si les valeurs extrême sont présentes dans l'ensemble de données, les valeurs extrêmes sont celles qui sont éloignées des autres données de l'ensemble de données, comme les points en blue au figure précédant.

Estimation des paramètres :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5092     0.3386    7.41 3.63e-11 ***
xl            6.5692     0.5877   11.18  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.599 on 103 degrees of freedom
Multiple R-squared:  0.5481,    Adjusted R-squared:  0.5437
F-statistic: 124.9 on 1 and 103 DF,  p-value: < 2.2e-16
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.5092$, $\beta_1 = 6.5692$, RSE $= 1.599$ et $R^2 = 0.5481$.

La droite de régression est définie par l'équation :

$$\hat{y} = 2.5092 + 6.5692x$$

On remarque que la droite ne passe par le centre de nuage de points a cause de l'influence des valeurs aberrants en bleu.

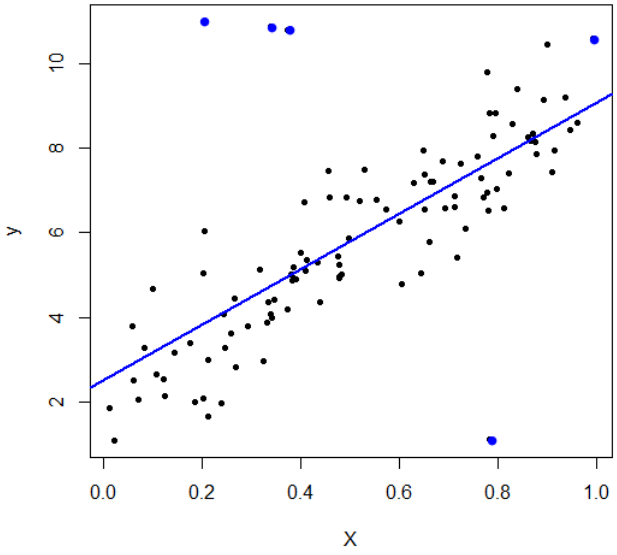


FIGURE 3.29 – Nuage de points avec la droite d'ajustement

Analyse des Résidus

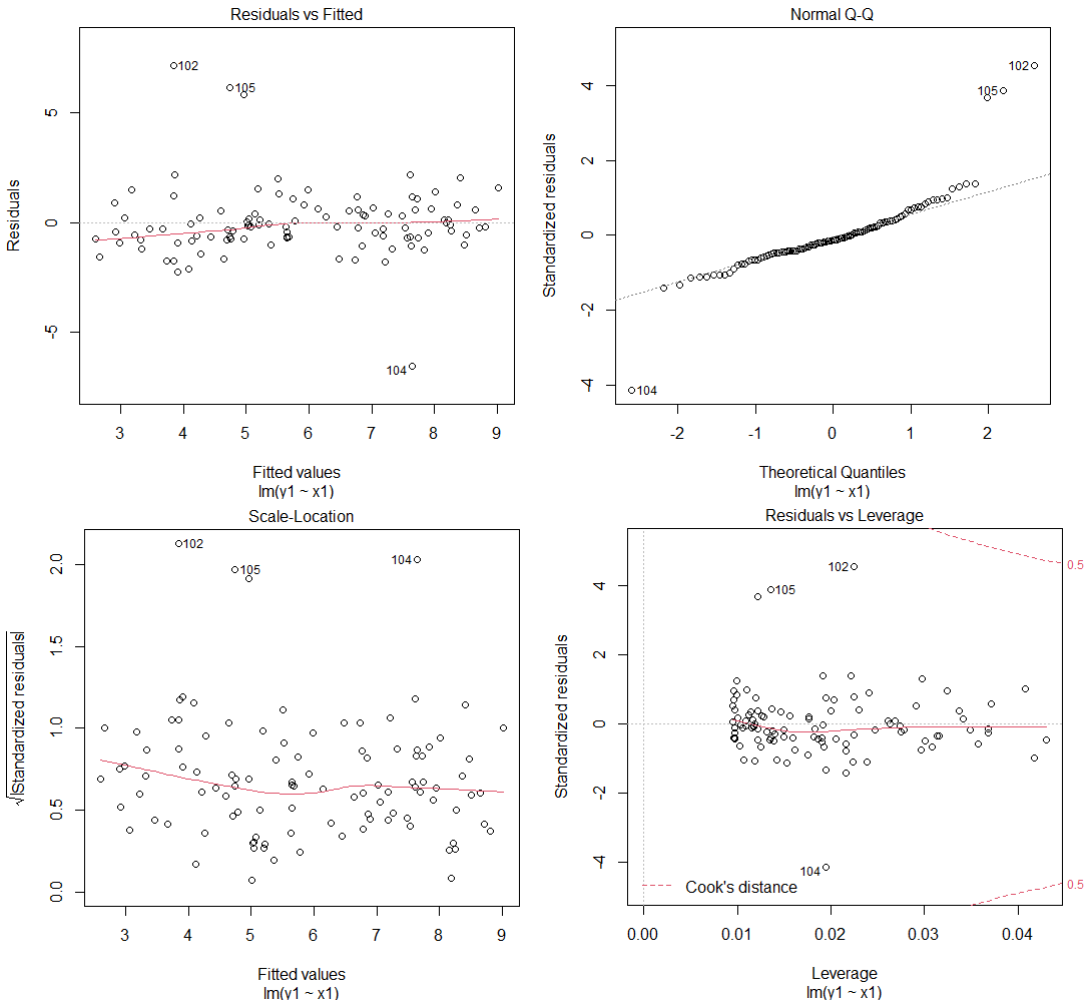


FIGURE 3.30 – Graphiques des résidus

Le premier graphe Residuals vs Fitted : le graphe nous montre que lorsque les réponses prédites par le modèle (fitted values) augmentent, les résidus restent globalement uniformément distribués de part et d'autre de 0. Cela montre, qu'en moyenne, la droite de régression, est bien adaptée aux données, et donc que l'hypothèse de linéarité est acceptable.

Le deuxième graphe Normal Q-Q : Evaluation de l'hypothèse de normalité des résidus, cette hypothèse peut s'évaluer graphiquement à l'aide d'un QQplot. Si les résidus sont bien distribués le long de la droite figurant sur le graphe , alors l'hypothèse de normalité on peut la considérer acceptée dans notre cas .

le troisième graphe Residuals vs Leverage : permet d'identifier les points avec une forte influence, la ligne pointillée démarque le seuil de 1 pour la distance de cook dans notre cas aucun points est proche ou dépasse le 1. donc on peut considérer aucun valeur influent.

le quatrième graphique : Aucun motif discernable est remarquable ce la impliquant que les erreurs sont normalement distribué avec présence de l'homoscédasticitè.

Méthode moindre carré pondérer

Estimation des paramètres :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.4592     0.3197   7.691 9.04e-12 ***
data$R        6.6821     0.5857  11.409  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.669 on 103 degrees of freedom
Multiple R-squared:  0.5583,    Adjusted R-squared:  0.554
F-statistic: 130.2 on 1 and 103 DF,  p-value: < 2.2e-16
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.4592$, $\beta_1 = 6.6821$, RSE $= 1.669$ et $R^2 = 0.5583$.

Estimation par MM-estimation

Estimation des paramètres

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8966     0.2159   8.784 3.69e-14 ***
xl            7.4476     0.3706  20.097  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.9414
Multiple R-squared:  0.815,     Adjusted R-squared:  0.8132
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 1.8966$, $\beta_1 = 7.4476$, RSE $= 0.9414$ et $R^2 = 0.815$.

Estimation Mèthodes t-régression

```
Location model :

Call:
tlm(lform = y1 ~ x1, data = data, estDof = TRUE)

Residuals:
      Min         1Q     Median         3Q        Max
-12.90744   -0.09081   -0.01147    0.07218    4.62802

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.21527    0.02046   108.3   <2e-16 ***
x1           7.00883    0.03107   225.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Scale Model :

Call:
tlm(lform = y1 ~ x1, data = data, estDof = TRUE)

Residuals:
     Min       1Q    Median        3Q       Max
-4.04222  -2.58909   0.05464   2.20202   3.98580

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.0620     0.2711  -18.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Scale parameter taken to be  2 )


Est. degrees of freedom parameter:  0.9861215
Standard error for d.o.f:  0.1645651
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.21527$, $\beta_1 = 7.0083$, RSE $= 0.1645$

Estimation par régression quantiles

Estimation des paramètres

```
Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.1

Coefficients:
            coefficients lower bd upper bd
(Intercept) 0.57476      0.10246  1.10554
x1          7.54600      6.62979  8.29968

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.25

Coefficients:
            coefficients lower bd upper bd
(Intercept) 1.51564      1.05391  1.70361
x1          7.13398      6.67782  7.53949

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)
Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.05975      1.64532  2.35498
x1          7.08410      6.51724  8.14403

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.75

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.68251      1.98254  3.78565
x1          7.13862      6.02618  8.64008

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.9

Coefficients:
            coefficients lower bd upper bd
(Intercept) 4.02431      3.49276  6.25745
x1          6.42667      4.63342  7.82484
```

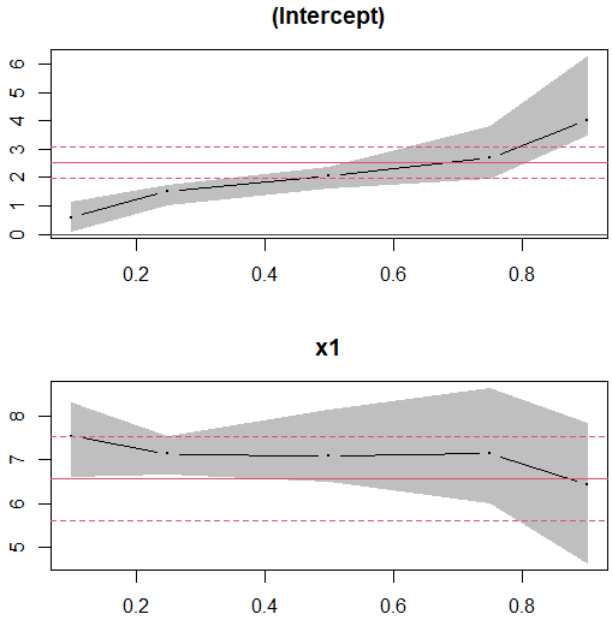


FIGURE 3.31 – Intervalle de confiance pour coefficient de modèle

A partir des résultats d'estimation déterminer sous **R**, et le graphe ci dessus, on remarque que le coefficient constant β_0 avoir une valeur initial 2.2 pour un quantile dont l'ordre est compris entre 0.4 et 0.6.
Pour la pente β_1 proche de sa vraie valeur pour des quantiles dont l'ordre est compris entre 0.2 et 0.6.
Les résultats obtenue dans ce cas sont résumé dans le tableau suivant :

	MCO	MCP	MM	t-reg	reg-quant
R^2	0.548	0.5583	0.815	0.5448	0.5385
RSE	1.599	1.669	0.9414	0.1615	1.6566

3.8 Simulation avec 10% valeur aberrante :

3.8.1 Avec variance non constante :

moindre carrés ordinaire :

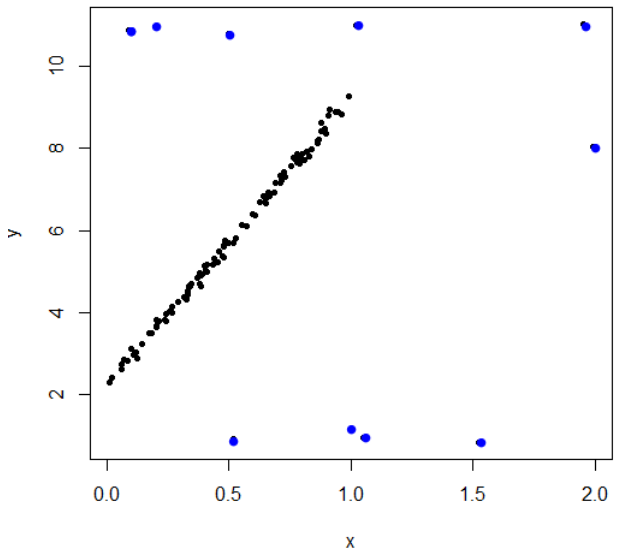


FIGURE 3.32 – Nuage de points des y en fonction des x

On remarque que les points de nuage forment une droite donc le modèle proposé est bon.
Estimation des paramètres :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.8240     0.3494  10.946  < 2e-16 ***
xl            3.7305     0.5306   7.031 1.95e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.925 on 108 degrees of freedom
Multiple R-squared:  0.314,     Adjusted R-squared:  0.3076
F-statistic: 49.43 on 1 and 108 DF,  p-value: 1.947e-10
```

A partir des résultats d'estimation déterminer sous **R**, nous avons :
$\beta_0 = 3.8240$, $\beta_1 = 3.735$, RSE $= 1.925$ et $R^2 = 0.314$.
La droite de régression est définie par l'équation :

$$\hat{y} = 3.8240 + 3.735x$$

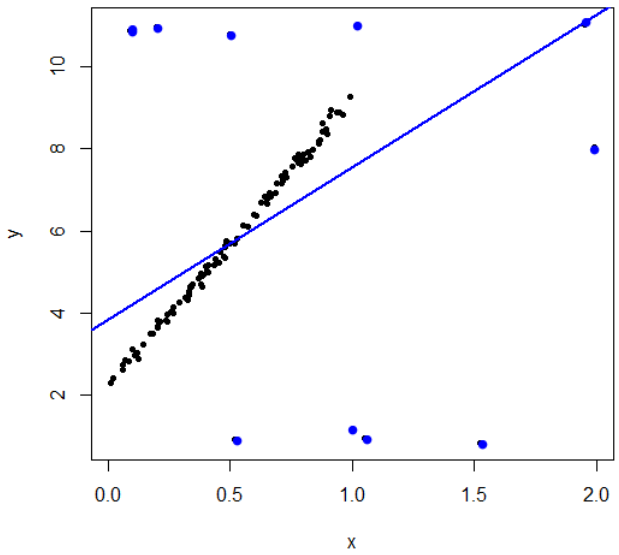


FIGURE 3.33 – Nuage de points avec droite d'ajustement
On remarque que la droite ne passe par le centre de nuage de points a cause de l'influence des valeurs aberrants en bleu.

Les Résidus

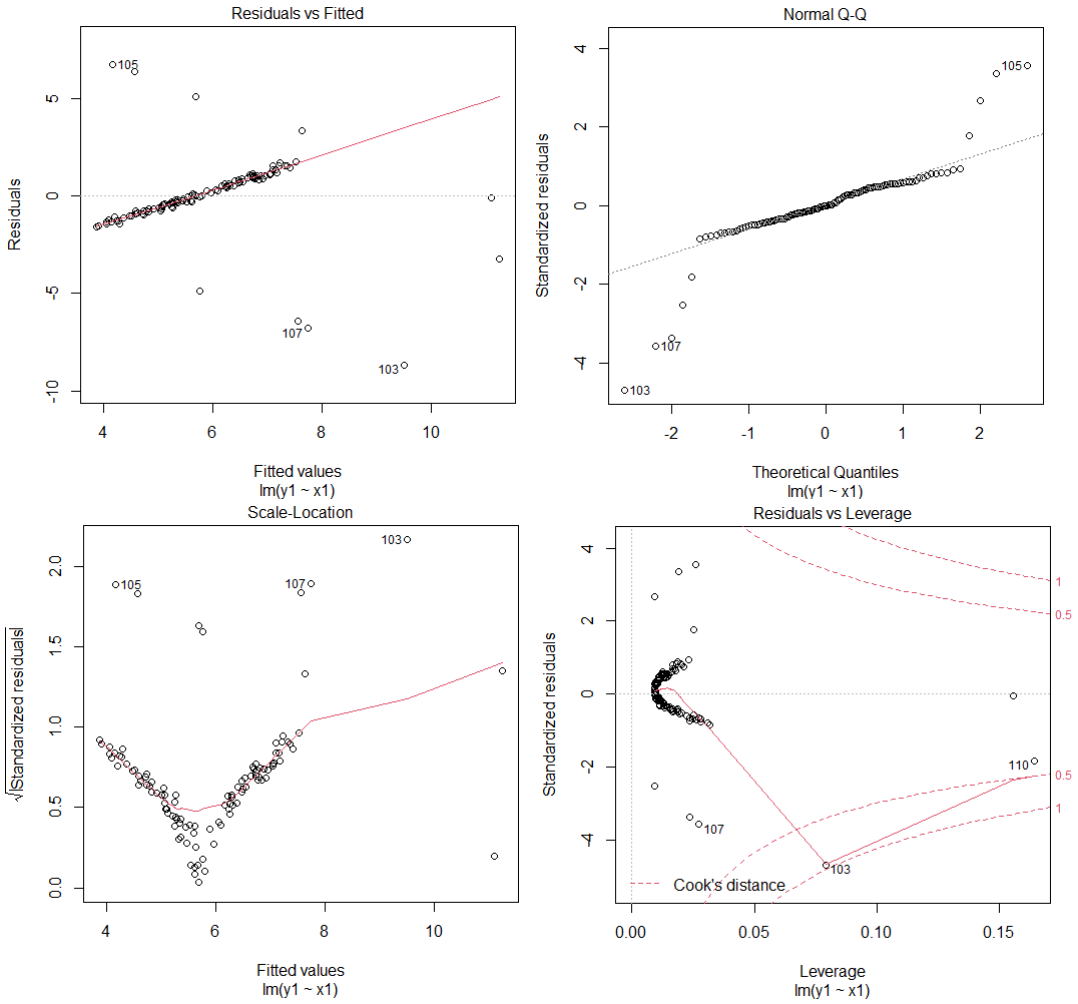


FIGURE 3.34 – Graphiques des résidus

Le premier graphe Residuals vs Fitted : le graphe nous montre que la droite de régression, est n'adapte pas aux données, et donc que l'hypothèse de linéarité n'est pas acceptable.

Le deuxième graphe Normal Q-Q : les résidus sont bien distribués le long de la droite figurant sur le graphe, alors l'hypothèse de normalité on peut pas la considérer acceptée dans notre cas .

le troisième graphe Residuals vs Leverage : On remarque plusieurs points éloignés des autre, et on a un point qui dépasse le seuil 1(109), donc on peut les considérer comme valeurs influentes.

Le quatrième graphique : Le graphique montre un motif discernable impliquerai une présence de l'hétéroscédasticité.

Test d'hétéroscédasticité

```
        studentized Breusch-Pagan test

data:  model
BP = 5.5339, df = 1, p-value = 0.01865
```

On remarque que la p-value = 0.01865 est inférieur à 0.05, donc on accepte l'hypothèse d'hétéroscédasticité

Méthode moindre carré pondérer

Estimation des paramètres :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.4592     0.3197   7.691 9.04e-12 ***
data$R        6.6821     0.5857  11.409  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.669 on 103 degrees of freedom
Multiple R-squared:  0.5583,    Adjusted R-squared:  0.554
F-statistic: 130.2 on 1 and 103 DF,  p-value: < 2.2e-16
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.4592$, $\beta_1 = 6.6821$, RSE = 1.669 et $R^2 = 0.5583$.

Estimation par MM-estimation

Estimation des paramètres

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.19526    0.02364   92.87   <2e-16 ***
xl           7.03716    0.04717  149.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.1299
Multiple R-squared:  0.9961,    Adjusted R-squared:  0.9961
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.19526$, $\beta_1 = 7.03716$, RSE $= 0.1299$ et $R^2 = 0.9961$.

Estimation Mèthodes t-régression

```
Call:
tlm(lform = y1 ~ x1, data = data, estDof = TRUE)

Residuals:
      Min         1Q     Median         3Q        Max
-12.90744   -0.09081   -0.01147    0.07218    4.62802

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.21527    0.02046   108.3   <2e-16 ***
x1           7.00883    0.03107   225.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Scale parameter(s) as estimated below)

Scale Model :

Call:
tlm(lform = y1 ~ x1, data = data, estDof = TRUE)

Residuals:
     Min       1Q    Median        3Q       Max
-4.04222  -2.58909   0.05464   2.20202   3.98580

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.0620     0.2711  -18.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Scale parameter taken to be  2 )


Est. degrees of freedom parameter:  0.9861215
Standard error for d.o.f:  0.1645651
No. of iterations of model : 30 in 0.23
Heteroscedastic t Likelihood : -1.773079
```

A partir des résultats d'estimation déterminer sous **R**, nous avons : $\beta_0 = 2.21527$, $\beta_1 = 7.00883$, RSE $= 0.1645$ et $R^2 = 0.9959$.

Estimation par régression quantiles

Estimation des paramètres

```
Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.1

Coefficients:
            coefficients lower bd upper bd
(Intercept)  2.65211      2.53536  2.69911
x1           2.70246     -1.46596  4.38648

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.25

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.23755      2.18114  2.30016
x1          6.74679      6.31092  6.90370
Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.25783      2.21784  2.33014
x1          6.89310      6.77460  7.03238

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.75

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.26732      2.21113  2.35124
x1          7.03810      6.90621  7.14901

Call: rq(formula = y1 ~ x1, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = dat)

tau: [1] 0.9

Coefficients:
            coefficients lower bd upper bd
(Intercept) 2.34285      2.24420  6.83829
x1          7.07812      6.87529  7.28078
```

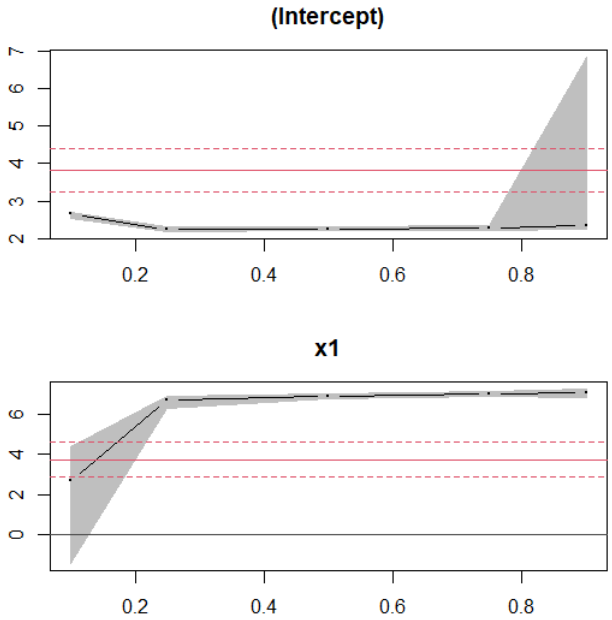


FIGURE 3.35 – Intervalle de confiance pour coefficient de modèle

A partir des résultats d'estimation déterminer sous **R**, et le graphe ci dessus, on remarque que le coefficient constant β_0 avoir une valeur initial 2.2 pour un quantile dont l'ordre est compris entre 0.2 et 0.8.
Pour la pente β_1 proche de sa vraie valeur pour des quantiles dont l'ordre est compris entre 0.2 et 0.8.

Les résultats obtenue dans ce cas sont résumé dans le tableau suivant :

	MCO	MCP	MM	t-reg	reg-quant
R^2	0.314	0.5583	0.9961	0.9959	0.9951
RSE	1.925	1.669	0.1299	0.1645	0.1324

La comparaison :

Pour comparer les différentes méthodes nous avons basé sur le R^2 et RSE. D'après les résultats obtenue a partir de simulation d'un échantillon de 100 variable.

$1^{ére}$ **cas :**

MM		reg quantile	
R^2	RSE	R^2	RSE
0.9961	0.1299	0.9060	0.1194

$2^{éme}$ **cas :**

MM	
R^2	RSE
0.8196	0.8876

$3^{éme}$ **cas :**

MM	
R^2	RSE
0.996	0.1299

$4^{éme}$ **cas :**

MM		t-régression	
R^2	RSE	R^2	RSE
0.8161	1.013	0.0687	0.346

$5^{éme}$ **cas :**

MM		t-régression	
R^2	RSE	R^2	RSE
0.8124	0.94	0.5642	0.7002

$6^{éme}$ **cas :**

MM	
R^2	RSE
0.9961	0.1298

$7^{éme}$ **cas :**

MM		t-régression	
R^2	RSE	R^2	RSE
0.815	0.9414	0.5448	0.1615

$8^{éme}$ **cas :**

MM	
R^2	RSE
0.9961	0.1299

on a conclus que la MM-estimation s'est révélée être la meilleur méthode robuste parmi les quatre autre pour traiter le problème des valeurs extrêmes.

Conclusion générale

Pour conclure, la présence de données aberrantes dans une expérience peut grandement affecter les résultats. Dans ce mémoire nous avons présenté différentes méthodes d'estimation qui permettent d'obtenir des résultats qui ne seront peu affectés par la présence de données aberrantes.

Pour une régression linéaire, l'influence d'une observation augmente si son résidu est grand (valeur extrême de y) ou si elle a un grand effet de levier (valeur extrême de x). La distance de Cook mesure l'effet combiné de ces deux facteurs.

La régression robuste basée sur les MM-estimateurs (fonction lmrob du package robustbase) produit des estimés presque aussi précis que la régression linéaire si les suppositions de celle-ci sont respectées, tout en étant beaucoup moins sensibles à la présence de quelques valeurs extrêmes.

Dans les statistiques robustes, la régression robuste est une forme d'analyse de régression conçue pour surmonter certaines limites des méthodes paramétriques et non paramétriques traditionnelles. L'analyse de régression cherche à trouver la relation entre une ou plusieurs variables indépendantes et une variable dépendante. Certaines méthodes de régression largement utilisées, telles que les moindres carrés ordinaires, ont des propriétés favorables si leurs hypothèses sous-jacentes sont vraies, mais peuvent donner des résultats trompeurs si ces hypothèses ne sont pas vraies ; on dit donc que les moindres carrés ordinaires ne sont pas robustes aux violations de ses hypothèses. Les méthodes de régression robustes sont conçues pour ne pas être trop affectées par les violations des hypothèses par le processus de génération de données sous-jacent.

En particulier, les estimations des moindres carrés pour les modèles de régression sont très sensibles aux valeurs aberrantes. Bien qu'il n'y ait pas de définition précise d'une valeur aberrante, les valeurs aberrantes sont des observations qui ne suivent pas le modèle des autres observations. Ce n'est normalement pas un problème si la valeur aberrante est simplement une observation extrême tirée de la queue d'une distribution normale, mais si la valeur aberrante résulte d'une erreur de mesure non normale ou d'une autre violation des hypothèses standard des moindres carrés ordinaires, cela compromet la validité. des résultats de la régression si une technique de régression non robuste est utilisée Ce mémoire sera principalement consacré à l'étude des méthodes de régression robuste et la dernière partie de celui ci sera dédiée à l'application de ces méthodes aux données obtenues par la simulation d'un modèle de régression linéaire simple sous logiciel R.

On a trouver que la MM-estimation est la plus robuste parmi les cinq méthode qu'on a traité dans ce mémoire.

Bibliographie

[1] Brigitte Dormont, Introduction à l'économétrie, Paris, Montchrestien, 2007, 2e éd., 518 p. (ISBN 978-2-7076-1398-1)

[2] Cade, B.S. et Noon, B.R. (2003) A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment 1 :412-420.

[3] Cowell, F. A., & Flachaire, E. (2007). Income distribution and inequality measurement : The problem of extreme values.Journal of Econometrics, 141(2), 1044-1072.

[4] Cullagh et Nelder, Introduction au modèle de régression linéaire simple, université de toulouse, 1992.

[5] Cullagh et Nelder, Introduction au modèle de régression linéaire multiple, université de toulouse, 1992.

[6] Donoho.D.L and Huber.P.J The notion of breakdown point. In A Festschrift for Erich Lehmann (P.Bickel and K.Dokum and J.L. Hodges,Jr.),Wadsworth,Belmont,1983.

[7] Gumbel.E.J Statistics of extremes. Columbia University Press, 1985

[8] Fox, J.(2002) Robust Regression. Appendix to An R and S-PLUS Companion to Applied Regression. Sage Publications, Thousands Oaks, USA

[9] Picard. F, Première notion de statistique régression linéaire, UMR CNRS-5558, université lyonl.

[10] Droesbeke.J ,Saporta.G, and Thomas-Agnan.C,o Méthodes robustes en statistique.Édition Technip, 2015.

[11] Droesbeke.J ,Lejeune.M , and Saportra.G , Modèle statistiques pour données qualitatives. Number 291. Edition TECHNIP, 2005.

[12] Jureckova.J, Sen.P.K , and Picek.J ,Methodology in Robust and Nonparametrics. CRC Press, 2013.

[13] Marek Banas and Marcin Ligas. Empirical tests of performance of some M-estmators. Polish Academy of Science, 2014.

[14] Huber.P.J , Robust Estimation of a Location Parameter, volume 35. The Annals of Mathematical Statistics, 1964.

[15] Huber.P.J , Robust Statistics. Wiley,1981.

[16] Rousseeuw.P.J and Leroy. Robust.A.M ,Regression and Outlier Detection. Hoboken : Wiley, 1987.

[17] Andersen.R .Modern Methods for Robust Regression. SAGE Publications,2008.

[18] Maronna.R.A, Martin.D,and Yohai.V.J ,Robust Statistics : Theory and Methods. Wiley, 2006.

[19] Koenker.R and Bassett.G ,Regression quantiles. Econometrica, 46(1) :33-50, 1978.

[20] Koenker.R, Quantile Regression. Cambridge University Press,2005.

[21] Portnoy.S ,Asymptotic behavior of number of number of regression quantile breakpoints. SIAM Journal on Scientific and Statistical Computing, 12(4) :867-883, 1992.

[22] Victor J. Yohai. High breakdown-point ans high efficiency robust estimates for regression. The Annals of Statistics, 15(20) :642-656, 1987.

[23] D'Haultfoeuille.X and Givod. P ,La régression quantile en pratique.Economie et statistique,2014.

[24] Susanti.Y, Pratiwi.H,Sulistijowati.S.H., and Liana.T ,Estimation.M, Estimation.S, and M.M , Estimation In Robust Regression. International Journal of Pure and Applied Mathematics, pages 349-360,2013.