République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

NITA Hadjer

Titre :

Statistiques robustes

Membres du Comité d'Examen :

Dr.	Djabrane Yahia	UMKB	Président
Pr.	Meraghni Djamel	UMKB	Encadreur
Dr	Benameur Sana	UMKB	Examinatrisse

Juin 2019

Dédicace

Je dédie ce modeste travail :

A mes Parents qui m'ont aidé et encouragé à poursuivre mes études

A ma Adorable Sœur Dr Sihem

A mes frères Dr Lokmane et Dr Mohamed

A mon fiancé Dr Abdellatif

A ma Famille pour leur soutient

A tous mes Amies surtout G.Sabrina

A tous ceux qui m'ont aidé à finaliser ce mémoire

REMERCIEMENTS

Je remercie tout d'abord «Allah » le tout puissant, de m'avoir donnée la santé, le courage et la patience pour mener à bien ce projet de fin d'études.

Je tiens à remercier le professeur Meraghni Djamel pour son encadrement, sa disponibilité, son suivi, ses conseils et ses critiques constructives.

Je tiens à remercier toute personne ayant contribué de près ou de loin à la réalisation de ce travail.

Pour finir j'aimerais remercier toute ma famille pour leur soutien constant

Une autre fois Je remercie mon encadreur

Pr. Meraghni Djamel

Table des matières

Table des figures

Liste des tableaux

Introduction

En statistique, la robustesse d'un estimateur est sa capacité à ne pas être modifié par une petite perturbation des données où par des paramètres du modèle à estimer. La robustesse est donc un terme qui désigne une constitution résistance et solide. La robustesse implique une insensibilité aux écarts dus à une non-conformité aux hypothèses sous-jacentes à un modèle probabiliste. Le plus souvent, on veut obtenir des méthodes dont les hypothèses d'application ne soient pas trop restrictives. Les méthodes classiques d'analyse statistique nécessitent très souvent une distribution gaussienne des données.

Les méthodes robustes garantiront que le résultat est bon pour une très grande collection de distributions sans pour autant être les meilleures pour une en particulier. L'analyse de variance classique, par exemple, nécessite que les résidus suivent une loi gaussienne avec un même écart-type pour chacun des groupes étudiés ; ce n'est pas une méthode robuste.

Un problème courant en statistique est d'essayer d'expliquer comment une variable d'intérêt Y est reliée à une variable explicative X. La régression est l'outil principal pour répondre à cette question. Cependant, cette estimation vu comme la moyenne conditionnelle de Y sachant X peut être inadaptée dans certaines situations. Par exemple, la présence de données aberrantes peut amener à des résultats non pertinents. La régression robuste a été introduite pour résoudre ce genre de problèmes.

Le terme "robuste" a été introduit par Box (1953), mais c'est Tukey (1960) qui a été le premier à reconnaître l'extrême sensibilité de certaines procédures statistiques classiques aux écarts mineurs par rapport aux hypothèses. Sa prise de conscience que les méthodes

statistiques optimisées pour le modèle gaussien classique sont instables sous de petites perturbations était cruciale pour les développements théoriques ultérieurs lancés par Huber (1964) et Hampel (1968).

En plus d'une introduction et d'une conclusion, ce mémoire est composé de deux chapitres. Les différentes notations et abréviations sont rassemblées, en annexe, avec leurs significations.

Chapitre 1. Estimation robuste : Ce chapitre est consacré à la définition d'un estimateur robuste et à la présentation quelques types d'estimateurs robustes parmi les plus populaires.

Chapitre 2. Régression linéaire et robustesse : Dans ce chapitre, on commence par un rappel sur la régression linéaire classique, puis on étudie la régression linéaire robuste. Les deux cas de régression (simple et multiple) sont considérées.

Enfin, il est utile de noter que les calculs numériques et les représentations graphiques, des exemples d'application traités dans le dernier chapitre, sont réalisés à l'aide du package ade4 du logiciel d'analyse statistique R [10].

Chapitre 1

Estimation robuste

Soit $(X_1, ..., X_n)$ un échantillon, de taille $n \geq 1$, d'une certaine population X. Un estimateur est une statistique, c'est à dire une fonction mesurable des variables aléatoires $X_1, ..., X_n$, qui sont indépendantes et identiquement distribuées (iid), portant le plus d'information possible sur une caractéristique inconnue de X.

1.1 Définition

Dans la littérature, il y a plusieurs formes pour la définition d'un estimateur robuste .

Définition 1.1.1 *Un estimateur est dit robuste s'il est insensible à de petites déviations vis à vis du modèle pour lequel il a été optimisé selon par Marie-Odile [1]. En d'autres termes, s'il garde de bonnes propriétés après une petite perturbation du modèle de départ.et s'il ne prend pas des valeurs aberrantes si le modèle de départ est changé ou fortement perturbé.*

Il existe des grandeurs qui permettent de mesurer la robustesse. A titre d'exemple, on peut citer la fonction d'influence et le point de rupture proposés par [5] et [8], pour lesquelles une description détaillée se trouve dans le travail d'El Asri [3].

Remarque 1.1.1 *Par petites déviations, on veut dire que la majorité des données sont entachées d'une petite erreur ou bien quelques données sont entachées d'une très grande erreur. Cette dernière situation correspond à ce qu'on appelle valeurs aberrantes (outliers en anglais).*

Exemple 1.1.1 *Le poids d'une population de personnes est une variable aléatoire d'espérance inconnue μ. Pour estimer μ, on dispose d'un échantillon de 9 personnes dont les poids en kilogrammes (kg) sont :* 48, 63, 75, 50, 72, 54, 80, 77 *et* 48.
Les moyenne et médiane de ces observations sont toutes les deux égales à 63 kg. Dans un deuxième calcul, le poids de la septième personne est pris, par erreur, égal à 180 kg. La moyenne devient approximativement égale à 74 kg et la médiane reste la même. On constate donc que la moyenne empirique est sensible à l'erreur alors que la médiane ne l'est pas. Ceci signifie que, des deux estimateurs de μ, c'est la médiane qui est robuste.

1.2 Différents types d'estimateurs

1.2.1 L-Estimateurs

Définition 1.2.1 *Soient $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ les statistiques d'ordre d'un échantillon de taille $n \geq 1$. Un L-estimateur T est une combinaison linéaire particulière des $X_{(i)}$, c'est à dire*

$$T = \sum_{i=1}^{n} a_i X_{(i)},$$

où $a_1, a_2, ..., a_n$ sont des nombres réels vérifant $0 \leq a_i \leq 1$ pour $i = 1, ..., n$ et $\sum_{i=1}^{n} a_i = 1$, appelés poids de T.

Exemple 1.2.1 *(Moyenne) La moyenne d'un échantillon de taille $n \geq 1$ est le L-estimateur dont tous les poids sont égaux à $1/n$. Elle donne la même importance à toutes les statistiques d'ordre.*

Exemple 1.2.2 *(Médiane) La médiane empirique est le L-estimateur qui ne fait intervenir que la statistique d'ordre centrale si n est impair et qui donne la moyenne des deux statistiques d'ordre centrales si n est pair.*

$$médiane = \begin{cases} X_{(p)} & si \quad n = 2p+1 \\ \frac{1}{2}X_{(p)} + \frac{1}{2}X_{(p+1)} & si \quad n = 2p \end{cases},$$

où $n \geq 1$ désigne la taille de l'échantillon.

Exemple 1.2.3 *(Moyenne $\alpha-$censurée) Soit α un réel vérfiant $0 \leq \alpha < 0.5$. La moyenne $\alpha-$censurée, notée $T(\alpha)$, est un L-estimateur de poids*

$$a_i = \begin{cases} 0 & si \quad i \leq g \ ou \ i \geq n-g+1 \\ \frac{1-r}{n(1-2\alpha)} & si \quad i = g+1 \ ou \ i = n-g \\ \frac{1}{n(1-2\alpha)} & si \quad g+2 \leq i \leq n-g-1 \end{cases},$$

où $g = [\alpha n]$ et $r = \alpha n - g$.

Le principe général de cet estimateur est d'éliminer une proportion α des plus petites valeurs et une proportion α des plus grandes valeurs, puis de calculer la moyenne de l'ensemble des valeurs restantes. Si on choisit $n = 18$ et $\alpha = 0.4$. On obtient $\alpha n = 7.2$ et par suite $g = 7$. Après avoir ordonné les observations, on élimine les 7 premières et les 7 dernières. Les valeurs restantes interviennent dans le calcul de la moyenne comme suit :

$$T(0.4) = \frac{0.8}{3.6}X_{(8)} + \frac{1}{3.6}\sum_{i=9}^{10} X_{(i)} + \frac{0.8}{3.6}X_{(11)}.$$

La moyenne 0.25-censurée est appelée mimoyenne car son calcul ne fait intervenir que la moitié centrale des observations triées. Si on prend $n = 20$ et $\alpha = 0.25$. On a $\alpha n = 5$, donc on enlève la moitié de cet échantillon. On obtient

$$T(0.25) = \frac{1}{20}\sum_{i=6}^{15} X_{(i)}.$$

1.2.2 M-estimateurs

Les M-estimateurs, introduits en 1964 par Huber [7], représentent une généralisation de l'estimation par maximum de vraisemblance.

Définition 1.2.2 *Un M-estimateur $\hat{\theta}$ du paramètre θ est défini par*

$$\hat{\theta} := \arg\min_{\theta} \sum_{i=1}^{n} \rho(x_i, \theta),$$

où ρ est une fonction donnée dans la définition 1.2.4.

La recherche d'un M-estimateur revient donc à minimiser la quantité $\sum_{i=1}^{n} \rho(x_i, \theta)$. La fonction ρ est obtenue par l'intégration d'une fonction ψ qu'on présente dans la définition 1.2.3.

Définition 1.2.3 *Une $\psi-$fonction est une fonction définie et continue par morceaux sur* $\mathbb{R}$, *telle que :*

1. *ψ est impaire.*
2. *$\psi(x) \leq 0$ pour $x \leq 0$ et $\psi(x) > 0$ pour $0 < x < x_r$ où $x_r = \sup\{x : \psi(x) > 0\}$*
3. *$\psi'(0) = 1$.*

Définition 1.2.4 *Une $\rho-$fonction, est représentée par l'intégrale d'une $\Psi-$fonction sur $[0, x]$, c'est à dire $\rho(x) := \int_0^x \psi(u)du$, $x > 0$.[9].*

Exemple 1.2.4 *La médiane est un M-estimateur pour la fonction ρ définie par*

$$\rho(x; t) = |x - t| .$$

Le problème, dans ce cas, est que la fonction valeur absolue n'a pas de dérivée en 0. *Cependant, il est raisonnable de prendre*

$$\psi(x; t) = sgn(x; t),$$

où

$$sgn(u) = \begin{cases} 1 & si \quad u > 0, \\ 0 & si \quad u = 0, \\ -1 & si \quad u < 0. \end{cases}$$

Il y a plusieurs solutions au problème de minimisation dans le cas d'un nombre pair d'observations. L'expression

$$\sum_{i=1}^{n} \Psi(x_i; t) = \sum_{i=1}^{n} sgn(x_i; -t),$$

calcule la différence entre le nombre des observations supérieures à t et celui des observations qui lui sont inférieures. Dans le cas où n est impair, la médiane annule cette expression. Si n est pair, toute valeur comprise entre les deux statistiques d'ordre centrales annule cette expression.

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 détriment de leur éfficacité. Les deux fonctions ρ les plus utilisées sont celles proposés par Huber [7] et par Tukey [15]. La première est définie, pour une constante $c > 0$,

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

La seconde, appelée Tukey's biweight, est définie par

$$\rho(x) := \begin{cases} \frac{1}{6}[1 - (1 - x^2)^3], & |x| \leq 1, \\ \frac{1}{6}, & |x| > 1. \end{cases}$$

Dans le tableau 1.1, on rassemble les $\rho-$fonctions et $\psi-$fonctions correspondantes de quelques M-estimateurs parmi les plus connus.

Type	ρ	ψ
médiane	$\|x-t\|$	$sgn(x;t)$
Huber	$\begin{cases} \frac{1}{2}x^2, & \|x\| \le c \\ c(\|x\| - \frac{c}{2}), & \|x\| > c \end{cases}$	$\begin{cases} x & \|x\| \le c \\ c\ sign(x) & \|x\| > c \end{cases}$
Tukey's biweight	$\begin{cases} \frac{1}{6}[1-(1-x^2)^3] & \|x\| \le 1 \\ \frac{1}{6} & \|x\| > 1 \end{cases}$	$\begin{cases} x(1-x^2)^2 & \|x\| \le 1 \\ 0 & \|x\| > 1 \end{cases}$
Andrew	$\begin{cases} \frac{1}{\pi^2}(1-\cos \pi x) & \|x\| \le 1 \\ \frac{2}{\pi^2} & \|x\| > 1 \end{cases}$	$\begin{cases} \frac{1}{\pi}\sin \pi x & \|x\| \le 1 \\ 0 & \|x\| > 1 \end{cases}$

TAB. 1.1 – rho et psi fonctions de quelques M-estimateurs

1.2.3 W-estimateurs

Définition 1.2.5 *Les W-estimateurs, ou Generalized M-estimators, sont des M-estimateurs pondérés. Chaque W-estimateur possède une fonction poids caractéristique, notée $w(.)$, qui représente l'importance de chaque observation dans l'estimation de θ. L'estimation optimale est déterminée en résolvant le système de p équations non-linéaires suivant*

$$\sum_{i=1}^{n} w(X_i)\psi(\epsilon_i/\hat{\sigma})X_{ij} = 0, j = 1, ..., p$$

voir[14].

Définition 1.2.6 *Une autre forme des M-estimateurs est appelée W-estimateurs. On prend T_n un estimateur défini par*

$$\sum_{i=1}^{n} \psi(\frac{x_i - T_n}{cS_n}) = 0.$$

où S_n est une grandeur de dispesion de l'échantillon. On introduit une fonction w telle que $uw(u) = \psi(u)$, alors

$$\sum_{i=1}^{n} \psi(\frac{x_i - T_n}{cS_n})w(\frac{x_i - T_n}{cS_n}) = 0,$$

et après modifications, on a

$$T_n = \frac{\sum_{i=1}^{n} x_i w[(x_i - T_n)/cS_n]}{\sum_{i=1}^{n} w[(x_i - T_n)/cS_n]}. \tag{1.1}$$

Ainsi, T_n est une moyenne pondérée des x_i. On dit que T_n est défini itérativement par l'équation 1.1 comme le W-estimateur basé sur la fonction de poids w.

Dans le tableau 1.2, on rassemble les poids w correspondantes de quelques W-estimateurs parmi les plus connus.

Type	Poids
moyenne	1
Biweight	$\begin{cases} (1-u^{2)^2} & \lvert u\rvert \leq 1 \\ 0 & \lvert u\rvert > 1 \end{cases}$
Huber	$\begin{cases} 1 & \lvert u\rvert \leq k \\ \frac{ksgn(u)}{u} & \lvert u\rvert > k \end{cases}$
Andrew	$\begin{cases} \frac{1}{\pi u}\sin u\pi & \lvert u\rvert \leq 1 \\ 0 & \lvert u\rvert > 1 \end{cases}$

TAB. 1.2 – Fonctions poids de quelques W-estimateurs

Chapitre 2

Régression linéaire et robustesse

Un modèle de régression linéaire est un modèle de qui cherche à établir une relation linéaire entre une variable, dite expliquée et une ou plusieurs variables, dites explicatives. Le modèle de régression linéaire est souvent estimé par la méthode des moindres carrés mais il existe aussi de nombreuses autres méthodes pour estimer ce modèle. On peut par exemple estimer le modèle par maximum de vraisemblance. C'est l'une des méthodes les plus connues et les plus appliquées en statistique pour l'annalyse de données quantitatives. Pour une déscription détaillée de la régression linéaire, on peut consulter [2] par exemple.

2.1 Régression linéaire non robuste

2.1.1 Régression linéaire simple

Elle est appliquée quand on s'interesse à la relation entre deux variables. Soient Y la v.a réelle à expliquer et X la variable explicative (déterministe). Le modèle théorique de régression linéaire simple est de la forme

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{2.1}$$

où β_0 et β_1 sont deux nombres réels inconnus, appelés paramètres du modèle (coefficients ou constantes de régression) et ϵ est une v.a d'espérance nulle et de variance finie, appelée bruit ou erreur.

Le but est d'estimer les coefficients de régression. Pour cela, on mesure $n \geq 1$ observations de X et Y notées x_i et y_i; $i = 1, ..., n$, respectivement. On confondra la variable aléatoire Y_i et sa réalisation y_i. On écrit alors

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, ..., n, \tag{2.2}$$

où les ϵ_i sont des v.a iid vérifiant $E(\epsilon_i) = 0$ et $E(\epsilon_i^2) = \sigma^2 < \infty$. Si on pose

- $\underline{Y} := (y_1, ..., y_n)'$: vecteur de dimension n.
- $\underline{X} := \begin{pmatrix} 1 & ... & 1 \\ x_1 & ... & x_n \end{pmatrix}'$: matrice de dimensions $n \times 2$.
- $\beta := (\beta_0, \beta_2)'$: vecteur des coefficients de la régression.
- $\underline{\epsilon} := (\epsilon_1, ..., \epsilon_n)'$: vecteur de dimention n.

Alors le système (2.2) s'écrit sous la forme matricielle $\underline{Y} = \underline{X}\beta + \underline{\epsilon}$.

Estimation des coefficients de régression

Méthode des moindres carrés

Définition 2.1.1 (Estimateurs des MC) *On appelle estimateurs des MC des coefficients β_0 et β_1, les quantités $\hat{\beta}_0$ et $\hat{\beta}_1$ obtenues par minimisation de la quantité*

$$S(\beta_0, \beta_1) := \|\underline{Y} - \underline{X}\beta\|^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

En d'autres termes

$$(\hat{\beta}_0, \hat{\beta}_1) := \arg \min_{(\beta_0, \beta_1) \in \mathbb{R} \times \mathbb{R}} S(\beta_0, \beta_1).$$

Proposition 2.1.1 *Les estimateurs des MC de β_0 et β_1 sont*

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ et } \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}, \tag{2.3}$$

où $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, $S_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$, $S_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$ *et* $S_{xy} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$.

Preuve. Le minimum de la fonction $S(\beta_0, \beta_1)$ est le point où son gradient s'annule. On annule les dérivées partielles et on obtient le système d'équations

$$\begin{cases} \dfrac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \dfrac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0, \end{cases}$$

appelées équations normales. En résolvant ce système, on trouve le résultat. ■

Méthode du maximum de vraisemblance

Dans ce cas, on suppose que l'erreur ϵ est de distribution Gaussienne, c-a-d $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Définition 2.1.2 (Estimateurs du MV) *On appelle estimateurs du MV des coefficients β_0 et β_1, les quantités $\hat{\beta}_0$ et $\hat{\beta}_1$ obtenues par maximisation de la fonction de vraisemblance $L(\beta_0, \beta_1)$ de l'échantillon $(Y_1, ..., Y_n)$. En d'autres termes*

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\max_{\beta_0, \beta_1} L(\beta_0, \beta_1)$$

Proposition 2.1.2 *Les estimateurs du MV de $\hat{\beta}_0$ et $\hat{\beta}_1$ sont les mêmes que ceux des MC.*

Preuve. Puisque $\epsilon \sim \mathcal{N}(0, \sigma^2)$ alors $Y \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$, c-a-d la densité de Y est définie par

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y - \beta_0 - \beta_1 x)^2)$$

Par conséquent, on a

$$L(\beta_0, \beta_1) := \prod_{i=1}^{n} f(y_i) = \left(2\pi\sigma^2\right)^{-n/2} \exp(-\frac{1}{2\sigma^2} S(\beta_0, \beta_1)).$$

Il est clair que maximiser $L(\beta_0, \beta_1)$ équivaut à minimiser $S(\beta_0, \beta_1)$. D'où le résultat. ■

Propriétés des estimateurs

Proposition 2.1.3 *Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$sont sans biais pour β_0 et β_1 respectivement.*

Preuve. *En remplaçant $y_i - \bar{y}$ par $\beta_1(x_i - \bar{x}) + \epsilon_i$ dans* (2.3) , *on obtient*

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\epsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{2.4}$$

Alors

$$E(\hat{\beta}_1) = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(\epsilon_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

c-a-d $E(\hat{\beta}_1) = \beta_1$. D'aure part

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \beta_0 + \beta_1\bar{x} - \hat{\beta}_1\bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x}, \tag{2.5}$$

d'où $E(\hat{\beta}_0) = \beta_0$. ■

Proposition 2.1.4 *On a*

$$V(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2},\ V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \textit{et}\ Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Preuve. Les résultats ci-dessus s'obtiennent de façon directe à partir des formes (2.4) et (2.5) . Pour $V(\hat{\beta}_1)$, on a

$$V(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 V(\epsilon_i)}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Pour $V(\hat{\beta}_0)$, on montre d'abord que $\bar{y}$ et $\hat{\beta}_1$ sont non corrélées. En effet, on a

$$cov(\bar{y}, \hat{\beta}_1) = cov(\frac{\sum_{i=1}^n y_i}{n}, \frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}) = \frac{1}{n\sum_{i=1}^n (x_i - \bar{x})^2} cov\left(\sum_{i=1}^n y_i, \sum_{i=1}^n (x_i - \bar{x})\epsilon_i\right),$$

avec

$$cov\left(\sum_{i=1}^n y_i, \sum_{i=1}^n (x_i - \bar{x})\epsilon_i\right) = \sum_{i=1}^n \sum_{j=1}^n (x_j - \bar{x}) cov(y_i, \epsilon_j).$$

Or $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, d'où

$$cov(y_i, \epsilon_j) = cov\left(\epsilon_i, \epsilon_j\right) = \begin{cases} \sigma^2 & \text{si} \quad i = j \\ 0 & \text{si} \quad i \neq j \end{cases}.$$

Par conséquent, on a

$$cov\left(\sum_{i=1}^n y_i, \sum_{i=1}^n (x_i - \bar{x})\epsilon_i\right) = \sigma^2 \sum_{i=1}^n (x_i - \bar{x}) = 0,$$

et donc $cov(\bar{y}, \hat{\beta}_1) = 0$. Ainsi

$$V(\hat{\beta}_0) = V\left(\bar{y}\right) + \bar{x}^2 V\left(\hat{\beta}_1\right) = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2 \sum x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Enfin, pour la covariance des deux estimateurs, on a

$$cov(\hat{\beta}_0, \hat{\beta}_1) = cov(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1) = cov(\bar{y}, \hat{\beta}_1) - \bar{x}V(\hat{\beta}_1).$$

En remplaçant $cov(\bar{y}, \hat{\beta}_1)$ et $V(\hat{\beta}_1)$ par leurs valeurs, on obtient le résultat. ■

Un troisième paramètre, à savoir la variance résiduelle σ^2, peut à son tour être estimé. Pour cela, on utilise ce que l'on appelle les résidus.

Définition 2.1.3 ***(Résidus)****Les résidus sont définis par*

$$\hat{\epsilon}_i = y_i - \hat{y}_i,$$

où $\hat{y}_i$ est la valeur ajustée de y_i par le modèle, c-a-d $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Remarque 2.1.1 *Dans un modèle de régression linéaire simple, la somme des résidus est nulle.*

Proposition 2.1.5 *L'estimateur du MV de la variance résiduelle σ^2 est*

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

Preuve. Il suffit de considérer $L(\beta_0, \beta_1)$ comme fonction des trois paramètres β_0, β_1 et σ^2 et chercher son maximum par rapport à σ^2. ■

Remarque 2.1.2 *L'estimateur $\hat{\sigma}^2_{MV}$ est biaisé.*

En effet, on écrit le résidus comme

$$\hat{\epsilon}_i = (\beta_1 - \hat{\beta}_1)(x_i - \bar{x}) + (\epsilon_i - \bar{\epsilon}),\ i = 1, ..., n.$$

La somme des carrés des résidus devient après simplification

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (\epsilon_i - \bar{\epsilon})^2 - (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

L'espérance de cette quantité est

$$E(\sum_{i=1}^{n} \hat{\epsilon}_i^2) = \sum_{i=1}^{n} E(\epsilon_i - \bar{\epsilon})^2 - V(\hat{\beta}_1) \sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-2)\sigma^2. \tag{2.6}$$

Alors,

$$E\left(\hat{\sigma}^2_{MV}\right) = \frac{n-2}{n}\sigma^2 \neq \sigma^2.$$

Le résultat (2.6) permet de construire, de façon directe, un estimateur sans biais pour la variance résiduelle σ^2.

Corollaire 2.1.1 *La statistique*

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2,$$

est un estimateur sans biais de σ^2.

Exemple 2.1.1 *La taille (en centimètres) et le poids (en kilogrammes) de dix étudiants* E_1,...,E_{10} *sont donnés dans le tableau 2.1 et représentés par le nuage de points de la figure 2.1.*

	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
Taille	172	178	175	180	174	171	170	181	174	175
Poids	64	67	63	69	65	62	64	69	65	62

TAB. 2.1 – Tailles et poids de 10 étudiants

Les estimations des deux coefficients de régression β_0 *et* β_1 *sont respectivement* -35.41 *et* 0.57. *Ceci donne une droite d'ajustement d'équation* $y = 0.57x - 35.41$ *(voir la figure 2.1).*

2.1.2 Régression linéaire multiple

Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en œuvre pour l'étude de données multidimensionnelles. C'est une généralisation du modèle de régression simple lorsque les variables explicatives $X_1, ..., X_p$ sont en nombre $p \geq 2$. Etant donné un échantillon $(Y_i, X_{i1}, ..., X_{ip})_{i\in\{1,n\}}$, on cherche à expliquer, avec le plus de précision possible, les valeurs prises par une variable expliquée Y à partir d'une série de variables explicatives $X_1, ..., X_p$. Le modèle théorique de la régression linéaire multiple prend la forme

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon,$$

où $\beta_0, \beta_1, ..., \beta_p$ sont des paramètres inconnus (donc à estimer) et ϵ est v.a d'espérance nulle et de variance σ^2, indépendantes des X_j. Lorsqu'on dispose de n réalisations $(y_i, x_{i1}, ..., x_{ip})_{i\in\{1,n\}}$

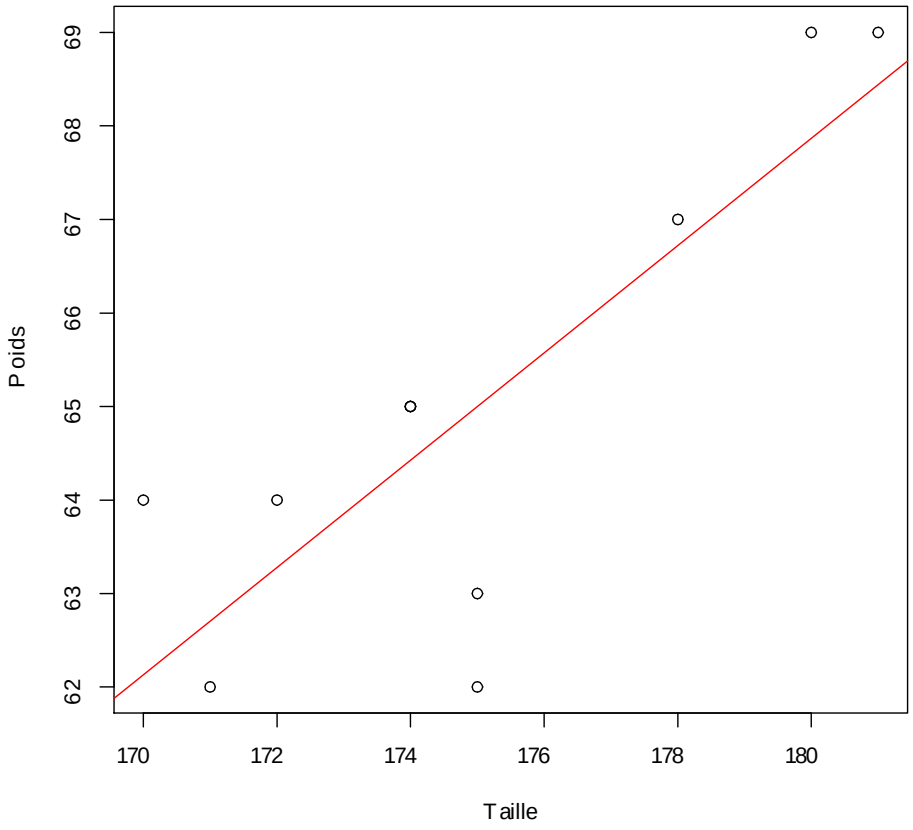

FIG. 2.1 – Nuage de points et droite de régression pour la taille et le poids de 10 étudiants

des v.a $(Y, X_1, ..., X_p)$, le modèle de régression s'écrit

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i, \ \ i = 1, ..., n, \tag{2.7}$$

où les ϵ_i sont des copies non corrélées (entre-elles) de ϵ. En utilisant l'écriture matricielle du système (2.7), on obtient la définition suivante.

Définition 2.1.4 *Un modèle de régression linéaire multiple est défini par*

$$\underline{Y} = \underline{X}\beta + \underline{\epsilon},$$

où

- $\underline{Y} := (y_1, ..., y_n)'$: vecteur de dimension n.
- $\underline{X} := \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix}$: matrice de dimension $n \times (p+1)$.
- $\beta := (\beta_0, ..., \beta_p)'$: vecteur des coefficients de la régression de dimension $(p+1)$.
- $\underline{\epsilon} := (\epsilon_1, ..., \epsilon_n)'$: vecteur aléatoire centré, des erreurs non corrélées, de dimension n.

Estimation des coefficients de régression

Méthode des moindres carrés

La méthode des MC consiste à minimiser l'erreur quadratique moyenne relative aux terme d'erreur du modèle.

Définition 2.1.5 (Estimateur des MC) *L'estimateur des MC $\hat{\beta}$ de β est défini par*

$$\hat{\beta} := \arg\min \|\underline{\epsilon}\|^2 = \arg\min \underline{\epsilon}'\underline{\epsilon} = \arg \min_{\beta \in \mathbb{R}^{p+1}} (\underline{Y} - \underline{X}\beta)'(\underline{Y} - \underline{X}\beta).$$

En d'autres termes, si on pose $S(\beta) := (\underline{Y} - \underline{X}\beta)'(\underline{Y} - \underline{X}\beta) = \sum_{i=1}^{n}(y_i - \sum_{j=0}^{p}\beta_j x_{ij})^2$, alors

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^{p+1}} S(\beta).$$

Proposition 2.1.6 *L'estimateur des MC $\hat{\beta}$ de β vaut*

$$\hat{\beta} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y}.$$

Preuve. Une condition nécessaire d'optimum est que la dérivée première de

$$S(\beta) = \underline{Y}'\underline{Y} + \beta'\underline{X}'\underline{X}\beta - 2\underline{Y}'\underline{X}\beta,$$

par rapport à β s'annule. Or la dérivée s'écrit comme suit

$$\frac{\partial S(\beta)}{\partial \beta} = 2\underline{X}'\underline{X}\beta - 2\underline{X}'\underline{Y},$$

d'où, s'il existe, l'optimum, noté $\hat{\beta}$, vérifie l'équation

$$2\underline{X}'\underline{X}\beta - 2\underline{X}'\underline{Y} = 0,$$

dont la solution est $\hat{\beta} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y}$. ■

Méthode du maximum de vraisemblance

On suppose maintenant que les erreurs ϵ_i sont de distribution normale de variance commune σ^2, c-a-d $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, pour $i = 1, ..., n$ et indépendants deux à deux. Autrement dit, le vecteur des résidus est multinormal d'espérance le vecteur nul 0_n de $\mathbb{R}^n$ et de matrice de covariance $\Sigma := \sigma^2 \mathbb{I}_n$, c-a-d $\underline{\epsilon} \sim \mathcal{N}(0_n, \Sigma)$.

Définition 2.1.6 (Estimateur du MV) *On appelle estimateur du MV de β, le vecteur $\hat{\beta}$ obtenu par maximisation de la fonction de vraisemblance $L(\beta)$ de l'échantillon $(Y_1, ..., Y_n)$. En d'autres termes*

$$\hat{\beta} := \arg \max_{\beta \in \mathbb{R}^{p+1}} L(\beta).$$

Proposition 2.1.7 *L'estimateur du MV de β est exactement le même que celui des MC.*

Preuve. La fonction de vraisemblace est donnée par

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-\epsilon_i^2}{2\sigma^2}\right\} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2\right\} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{\underline{\epsilon}'\underline{\epsilon}}{2\sigma^2}\right\} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{(\underline{Y}-\underline{X}\beta)'(\underline{Y}-\underline{X}\beta)}{2\sigma^2}\right\}. \end{aligned}$$

Ainsi, il est clair que maximiser $L(\beta)$ par rapport à β est équivalent à minimiser $S(\beta)$ par rapport à β, d'où le résultat. ■

Proposition 2.1.8 *L'estimateur $\hat{\beta}$ est un estimateur sans biais de β et sa variance vaut $V(\hat{\beta}) = \sigma^2(\underline{X}'\underline{X})^{-1}$.*

Preuve. En écrivant $\underline{Y} = \underline{X}\beta + \underline{\epsilon}$, on a $\hat{\beta} = \beta + (\underline{X}'\underline{X})^{-1}X'\underline{\epsilon}$. Donc

$$E(\hat{\beta}) = \beta + (\underline{X}'\underline{X})^{-1}\underline{X}'E(\underline{\epsilon}) = \beta,$$

et

$$\begin{aligned} V(\hat{\beta}) &= V((\underline{X}'\underline{X})^{-1}X'\underline{\epsilon}) = (\underline{X}'\underline{X})^{-1}\underline{X}'V(\underline{\epsilon})\underline{X}(\underline{X}'\underline{X})^{-1} \\ &= (\underline{X}'\underline{X})^{-1}\underline{X}'\Sigma\underline{X}(\underline{X}'\underline{X})^{-1}. \end{aligned}$$

On termine le calcul en remplaçant la matrice Σ par sa valeur $\sigma^2\mathbb{I}_n$. ■

Définition 2.1.7 (Résidu) *Le résidu $\hat{\epsilon}$ est défini par la différence entre la valeur théorique Y et sa valeur ajustée $\hat{Y} := X\hat{\beta}$.*

$$\hat{\epsilon} := Y - \hat{Y}.$$

Proposition 2.1.9 *L'estimateur du MV de la variance résiduelle σ^2 est*

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{1}{n}\left\|\hat{\epsilon}\right\|^2.$$

Preuve. *Il suffit de considérer $L(\beta)$ comme fonction des paramètres $\beta_i, i = 1, ..., n$ et σ^2 et chercher son maximum par rapport à σ^2.* ■

Remarque 2.1.3 *L'estimateur $\hat{\sigma}^2_{MV}$ est biaisé.*
En effet, on calcule $E(\|\hat{\epsilon}\|^2)$, puisque c'est un scalaire, il est égal à sa trace d'aprés Ruse de sioux ce qui donne

$$E(\|\hat{\epsilon}\|^2) = E(Tr(\|\hat{\epsilon}\|^2)) = E(Tr((\hat{\epsilon}'\hat{\epsilon})),$$

et puisque pour toute matrice A, on a $Tr(AA') = Tr(A'A)$, alors

$$E(\|\hat{\epsilon}\|^2) = E(Tr(\hat{\epsilon}\hat{\epsilon}')) = Tr(E(\hat{\epsilon}\hat{\epsilon}')) = Tr(V(\hat{\epsilon})) = Tr(\sigma^2 P_{X^\perp}).$$

Et comme $P_{X^\perp}$ est la matrice de la projection orthogonale sur un espace de dimension$(n-p)$, on a bien :

$$E(\|\hat{\epsilon}\|^2) = (n-p)\sigma^2 \tag{2.8}$$

donc

$$E\left(\hat{\sigma}^2_{MV}\right) = \frac{n-p}{n}\sigma^2 \neq \sigma^2.$$

Le résultat (2.8)permet de construire, de façon directe, un estimateur sans biais pour la variance résiduelle σ^2.

Corollaire 2.1.2 *La statistique*

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\epsilon}_i^2,$$

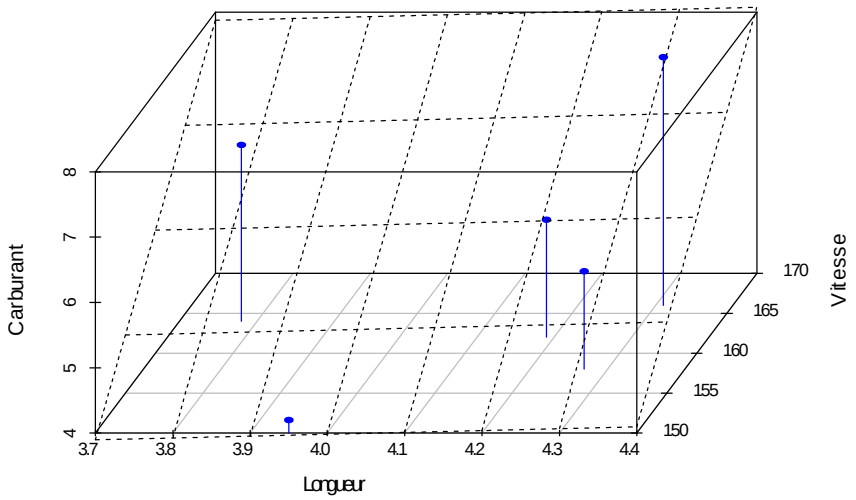

FIG. 2.2 – Nuage de points, dans l'espace, représentant les longueurs, vitesses de pointe et consommations de carburant de 5 voitures. Les lignes en pointillés délimitent le plan de régression multiple.

est un estimateur sans biais de σ^2.

Exemple 2.1.2 *Les longueurs (L), vitesses de pointe (V) et consommations de carburant (C) de* 5 *modèles de voitures sont résumées dans le tableau 2.2 et représentés par le nuage de points de la figure 2.2.*

	L	**V**	**C**
M_1	4.19	162	5.8
M_2	4.31	166	7.8
M_3	4.27	158	5.5
M_4	3.78	164	6.7
M_5	3.95	150	4.2

TAB. 2.2 – Longueurs,vitesses de pointe et consommations de carburant de 5 voitures

Les estimations des deux coefficients de régression β_0, β_1 *et* β_2 *sont respectivement* -26.99, 0.28 *et* 0.20. *Ceci donne un plan d'ajustement d'équation* $z = -26.99 + 0.28x + 0.2y$ *(voir la figure 2.2).*

2.2 Régression linéaire robuste

Bien que les méthodes de régression linéaire soient un outil très répandu parmi les analystes de données, il est très difficile de travailler avec des données présentant une relation linéaire. Pour cette raison, on doit prendre en considération des versions robustes de la méthode de régression linéaire classique.

2.2.1 Régression robuste simple

On part d'une série d'observations $\{(x_i, y_i); i = 1, ..., n\}$, pour faire une régression linéaire avec le modèle (2.1) . La régression linéaire robuste simple est un processus itératif qui aboutit à une approximation de la variable expliquée. Si on désigne par $\beta_0^{(k)}$ et $\beta_1^{(k)}$ les estimations des coefficients β_0 et β_1 obtenues à l'itération numéro k, alors les résidus correspondants sont

$$\epsilon_i^{(k)} = y_i - (\beta_0^{(k)} + \beta_1^{(k)} x_i),\ i = 1, ..., n. \tag{2.9}$$

On peut résumer cette régression en les étapes suivantes :

- On commence par faire une régression des moindres carrées (ordinaires).
- On calcule des poids w_i à partir des résidus obtenus par la régression.
- On fait une régression pondérée des moindres carrés, c-a-d on cherche les valeurs de $\beta_0^{(k)}$ et $\beta_1^{(k)}$ qui minimisent
$$\sum_{i=1}^{n} w_i \epsilon_i^2. \tag{2.10}$$
- On change les poids en fonction des résidus obtenus et on recommence la régression pondérée.

Pour une déscription détaillée de la régression robuste simple, on peut consulter [12] par exemple.

Remarque 2.2.1 *Les W-estimateurs présentés plus haut serviront dans le calcul des poids à chaque itération.*

Il y'a différentes méthodes consernant la régression robuste simple. On va considerer deux méthodes (moindres valeurs absolues et Régression Biweight)

Régression des moindres valeurs absolues

Pour cette méthode on prend, dans (2.10) , comme poids

$$w_i = \begin{cases} \frac{1}{|\epsilon_i|} & \epsilon_i \neq 0 \\ 0 & \epsilon_i = 0 \end{cases}, \ i = 1, ..., n.$$

Avec ce choix l'expression à minimiser devient

$$\sum_{i=1}^{n} \frac{1}{|\epsilon_i|} \epsilon_i^2 = \sum_{i=1}^{n} |\epsilon_i| ,$$

d'où l'appellation "moindres valeurs absolues". On va donc minimiser la somme des valeurs absolues des résidus au lieu de la somme de leur des carrés, comme dans le cas de la régression des moindres carrés (régression non robuste). Ce qui a pour effet de diminuer l'importance des résidus très grands en valeur absolue.

Régression Biweight

L'algorithme de cette méthode est le suivant :

1. A la première itération, on a $k = 1$ et $w_i = 1$ pour $i = 1, ..., n$.

2. On fait une régression linéaire des moindres carrés des valeurs y_i sur les x_i avec les poids w_i. On obtient ainsi des estimations $\beta_0^{(k)}$ et $\beta_1^{(k)}$ des paramêtres.

3. On calcule les résidus selon la formule (2.9) . et les valeurs

$$u_i = \frac{\epsilon_i}{cS_k},$$

où S_k est une grandeur de dispesion de l'échantillon, souvent prise égale à la médiane des $|\epsilon_i|$, et $c = 6$ à 9.

4. On calcule les nouveaux poids

$$w_i = \begin{cases} (1-u_i^2)^2 & |u_i| \leq 1 \\ 0 & u_i > 1 \end{cases}.$$

5. On pose $k = k+1$ et on retourne à l'étape 2.

On arrête cet algorithme quand les valeurs des paramêtres ne diffèrent pas trop d'une itération à l'autre.

2.2.2 Régression robuste multiple

Il y'a différentes méthodes consernant la régression robuste multiple. On va choisir deux méthodes (M-estimation et LTS)

M-estimation des paramètres du modèle

La méthode générale la plus courante de régression robuste est l'estimation M, introduite par [7]. Pour lesquelles une déscription détaillée se trouve dans le travail [4].

On a le modèle linéaire multiple(2.7)

$$y_i = x_i'\beta + \epsilon_i$$

pour la nième observation. Le modèle ajusté est

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_p x_{ip} + \hat{\epsilon}_i, \ i = 1, ..., n, \\ &= x_i'\hat{\beta} + \hat{\epsilon}_i \end{aligned}$$

L'estimateur général M minimise la fonction objectif

$$\sum_{i=1}^{n} \rho(\hat{\epsilon}_i) = \sum_{i=1}^{n} \rho(y_i - x_i'\hat{\beta})$$

où la fonction ρ donne la contribution de chaque résidu à la fonction objectif. Un ρ raisonnable devrait avoir les propriétés suivantes :

- $\rho(\hat{\epsilon}) \geq 0$
- $\rho(0) = 0$
- $\rho(\hat{\epsilon}) = \rho(-\hat{\epsilon})$
- $\rho(\hat{\epsilon}_i) \geq \rho(\hat{\epsilon}'_i)$ si $|\hat{\epsilon}_i| > |\hat{\epsilon}'_i|$

On dérive la fonction objectif par rapport aux coefficients $\hat{\beta}$, et en mettant les dérivées partielles à 0, produit un système de $k+1$ équations d'estimation pour les coefficients

$$\sum_{i=1}^{n} \psi(y_i - x'_i\hat{\beta})x'_i = 0,$$

On défini la fonction de pondération $w(\hat{\epsilon}) = \psi(\hat{\epsilon})/\hat{\epsilon}$ et on note $w_i = w(\hat{\epsilon}_i)$. Ensuite, les équations d'estimation peuvent être écrites comme

$$\sum_{i=1}^{n} w_i(y_i - x'_i\hat{\beta})x'_i = 0,$$

La résolution des équations d'estimation est un problème de moindres carrés pondéré, on minimise $\sum_{i=1}^{n} w_i^2\hat{\epsilon}_i^2$. Les poids, cependant, dépendent des résidus, les résidus dépendent des coefficients estimés et des estimations les coefficients dépendent des poids. Une solution itérative (appelée moindres carrés itérativement repondérés, IRLS) est donc nécessaire :

1. On sélectionne les estimations initiales $\hat{\beta}^{(0)}$, comme les estimations des moindres carrés.
2. À chaque itération k, on calcule les résidus $\hat{\epsilon}_i^{(k-1)}$ et les poids associés $w_i^{(k-1)} = w[\hat{\epsilon}_i^{(k-1)}]$ à partir de l'itération précédente.
3. on résoudre de nouvelles estimations de moindres carrés pondérés.

$$\hat{\beta}(t) = [X'W^{(k-1)}X]^{-1}X'W^{(k-1)}Y,$$

où X est la matrice du modèle, avec x_i' as its ith row, et $W^{(k-1)} = diag\{w_i^{(k-1)}\}$ est la matrice de poids actuelle.

Les étapes (2) et (3) sont répétées jusqu'à ce que les coefficients estimés convergent, et le processus se poursuit jusqu'à ce qu'il converge. La matrice de covariance asymptotique de $\hat{\beta}$ est

$$v(\hat{\beta}) = \frac{E(\psi^2)}{[E(\psi')]^2}(X'X)^{-1},$$

En utilisant $\sum_{i=1}^{n} \psi(\hat{\epsilon}_i^2)$ pour estimer $E(\psi^2)$, et $[\sum_{i=1}^{n} \psi'(\hat{\epsilon}_i)/n]^2$ pour estimer $[E(\psi')]^2$ produit la matrice de covariance asymptotique estimée, $\hat{v}(\hat{\beta})$ (qui n'est pas fiable dans les petits échantillons).

Les étapes de cette régression sont détaillé sur [13].

Dans la pondération de Huber, les observations avec de petits résidus ont un poids de 1 et plus le résidu est grand, plus le poids est petit. Ceci est défini par la fonction de poids

$$w(x) = \begin{cases} 1 & |x| \leq c \\ \frac{c}{|x|} & |x| > c \end{cases}.$$

Avec la pondération Tukey's biweight, tous les cas avec un résidu non nul sont sous-pondérés au moins un peu

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

Least trimmed squares (LTS)

La somme des carrés la moins réduite (LTS), est une méthode statistique robuste qui adapte une fonction à un ensemble de données sans être indûment affectée par la présence de valeurs aberrantes. C'est l'une des nombreuses méthodes de régression robuste.

Au lieu de la méthode des moindres carrés standard, qui minimise la somme des résidus au carré sur n points, la méthode LTS tente de minimiser la somme des résidus au carré

sur un sous-ensemble, k, de ces points. Les $n-k$ points inutilisés n'ont aucune influence sur l'ajustement.

Dans un problème classique des moindres carrés, les valeurs de paramètre estimées β sont définies comme celles qui minimisent la fonction objectif $S(\beta)$ des résidus au carré

$$S(\beta) = \sum_{i=1}^{n} \epsilon_i(\beta)^2,$$

où les résidus sont définis comme les différences entre les valeurs des variables dépendantes (observations) et les valeurs du modèle $\epsilon_i(\beta) = y_i - f(x_i, \beta)$, et où n est le nombre total de points de données. Pour une analyse des carrés les moins nets, cette fonction objectif est remplacée par une fonction construite de la manière suivante. Pour une valeur fixe de β, on prend $\epsilon_j(\beta)$ désignant l'ensemble des valeurs absolues ordonnées des résidus (en ordre croissant de valeur absolue). Dans cette notation, la fonction de somme standard de carrés est

$$S(\beta) = \sum_{j=1}^{n} \epsilon_j(\beta)^2,$$

tandis que la fonction objectif pour LTS est

$$S_k(\beta) = \sum_{j=1}^{k} r_j(\beta)^2.$$

voir([6]).

2.2.3 Exemple

A titre d'exemple, on utilise la base de données Duncan du package (car) du logiciel R [10]. Cette base de données contient la relation entre le niveau d'éducation et le revenu. On essaye de créer un barplot de Cook pour voir les valeurs qui ont trop d'influence voir la figure (2.3).

On peut voir dans la figure 2.3 trois valeurs aberrantes qui influent sur la méthode de

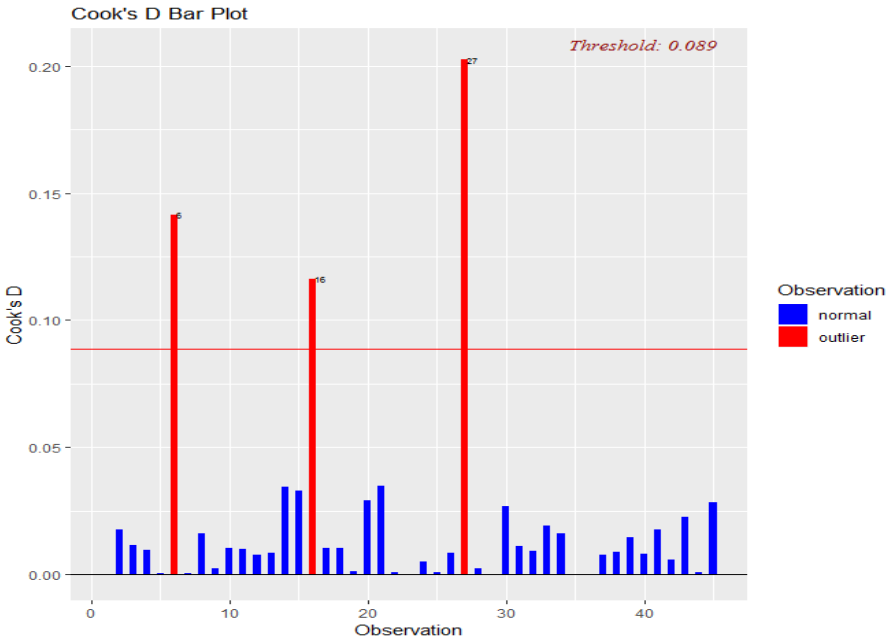

FIG. 2.3 – Diagramme en bâtons de Cook de la régression pour du revenu en fonction du niveau d'éducation

régression linéaire. Il est clair qu'on doit essayer de résoudre ce problème avec une version robuste de la méthode de régression. Les supprimer n'est pas une option car il ne s'agit pas d'erreurs de saisie, mais de données réelles représentant des personnes ayant un revenu inhabituellement élevé compte tenu de leur niveau d'instruction. On applique les méthodes de régression robuste qu'on a étudié ci-dessus (Huber,Tukey's biweight et LTS). Les résultats obtenus sont illustrés dans la figure (2.4), où l'on note que la que la méthode la plus robuste est la méthode LTS. Ceci est confirmé par le calcul des coefficients de détermination et leur comparaison. En effet, pour cette méthode on trouve $R^2 = 0.75$, qui est le plus élevé parmi les quatre coefficients.

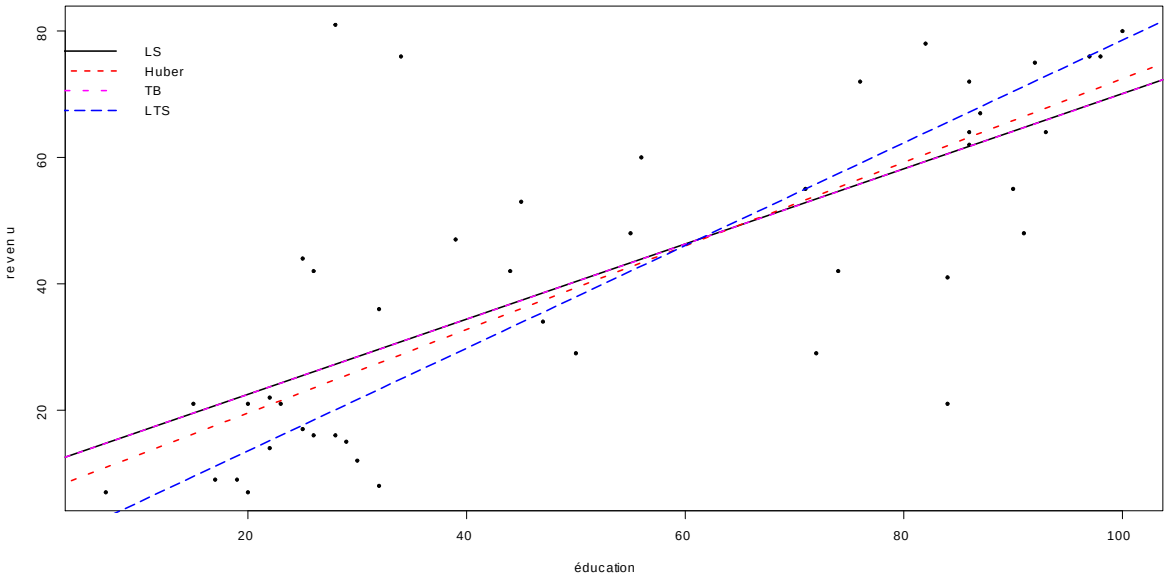

FIG. 2.4 – Régression simple et robuste pour du revenu en fonction du niveau d'éducation

Conclusion

Dans ce mémoire, on a étudié la notion de robustesse en discutant quelques estimateurs robustes (L-, M- et W-estimateurs). On s'est intéressé à la régression robuste après un rappel sur la régression non robuste.

La régression robuste est une méthode très fréquente lorsqu'il y a présence de données aberrantes et aussi lorsque certaines hypothèses de la régression sont violées. Elle permet la détection de ces valeurs aberrantes et donne une bonne estimation des coefficients de régression.

En conclusion, les méthodes statistiques robustes proposent des alternatives meilleures que celles classiques habituellement utilisées.

Bibliographie

[1] Berger, M.O. (2010). Introduction à l'estimation Robuste. Notes de cours.

[2] Cornillon, P.A. ,Matzner-Lober, E. (2007). Régression théorie et application .Springer.

[3] El Asri, M. (2014). Etude des M-estimateurs et leurs versions pondérées pour des données clusterisées. Thèse de Doctorat dirigée par Blanke, D. Université d'Avignon.

[4] Fox, J, Weisberg, S. (2012). Robust Regression in R. Site web : https ://www.researchgate.net/profile/David_Booth14/post/What_robust_statistical_measure_is_used_to_analyse_data_on_small_number_of_participants_for_a_study/attachment/59d62a4979197b8077988b0d/AS%3A338374104764421%401457686081692/download/Appendix-Robust-Regression.pdf

[5] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986). Robust statistics. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

[6] https ://en.wikipedia.org/wiki/Least_trimmed_squares.

[7] Huber, P.J. (1964). Robust estimation of location parameter. The annals of mathematical statistics, 35(1) : 73-101.

[8] Huber, P.J. (1977). Robust methods of estimation of regression coefficients. Statistics : A Journal of Theoretical and Applied Statistics, 8(1) : 41-53.

[9] Huber, P.J. (1981). Robust statistics. Wiley.

[10] Ihaka, R., Gentleman, R. (1996). R : A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5 : 299-314.

[11] Maronna, R.A., Martin, R.D., Yohai, V.J. (2006). Robust Statistics Theory and Methods. Wiley.

[12] Rapacchi, B.(1994). Une introduction à la notion de robustesse. Centre Interuniversitaire de Calcul de Grenoble, disponible en ligne sur www.unige.ch/ses/sococ/eda/bernard/robuste. pdf.

[13] Simard, J. (2018). Méthodes de régression robustes. Mémoire de l'obtention du grade de maître sciences (M.Sc), dirigé par Bouezmarni, M. Université de Sherbrooke.

[14] Souci, S.(2018). Estimation robuste de la régression. Mémoire présenté en vue de l'obtention du diplôme de Master Académique, dirigé par Dr. F. Benziadi, Université Dr Moulay Tahar - Saïda.

[15] Tukey, J.W. (1960). A survey of sampling from contaminated distributions. Contributions to probability and statistics, 2, 448-485.

Annexe : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous.

Notation		Signification
$(\Omega, \mathcal{F}, \mathbb{P})$	:	Espace probabilisé
$v.a$	:	Variable aléatoire
$[]$	:	partie entière
$\mathbb{E}$	:	Espérance mathématique
V	:	Variance ou l'ecart type
Cov	:	Covariance
$N(0;1)$	:	Loi normale centré réduite
$N(\mu;\sigma^2)$	:	Loi normale de moyenne μ et variance σ
iid	:	indépendantes identiquement distribuées
MC	:	moindre carré
cte	:	constant
$c-a-d$	:	c'est-à-dire
Tr	:	trace