UNIVERSITÉ ASSANE SECK DE ZIGUINCHOR

UFR SCIENCES ET TECHNOLOGIES
DÉPARTEMENT DE MATHÉMATIQUES

MÉMOIRE DE MASTER

DOMAINE : SCIENCES ET TECHNOLOGIES
MENTION : MATHÉMATIQUES ET APPLICATIONS
SPÉCIALITÉ : MATHÉMATIQUES APPLIQUÉES
OPTION : STATISTIQUE

Présenté par :

LAMINE MANE

TITRE

Régression Linéaire - Étude Comparative D'Estimateurs ET Application Sur Des Données Médicales

Sous la direction de : Dr. Emmanuel Nicolas CABRAL

Sous la supervision de : Pr. Alassane DIÉDHIOU

Soutenu publiquement le 09 Août 2023 devant le jury ci-après :

Prénoms et Nom	Grade	UFR	Qualité	Établissement
M. Clément MANGA	Professeur Assimilé	ST	Président	UASZ
M. Alassane DIÉDHIOU	Professeur Titulaire	ST	Superviseur	UASZ
M. Emmanuel Nicolas CABRAL	Maître de Conférences Titulaire	ST	Directeur	UASZ
M. Kalilou DIALLO	Maître de Conférences Titulaire	2S	Examinateur	UASZ

Année Universitaire 2021-2022

REMERCIEMENTS

Tout d'abord je tiens à exprimer ma profonde gratitude envers le **Tout-Puissant**, pour m'avoir guidé tout au long de mon parcours scolaire et universitaire et pour m'avoir donné la force et la persévérance nécessaires à l'accomplissement de ce travail de recherche.

Je voudrais également exprimer ma reconnaissance envers mon encadreur, **Dr Emannuel Nicolas CABRAL** et **Dr Kalilou DIALLO** de l'hôpital La Paix de Ziguinchor, pour leur soutien, leur patience et leur expertise. Vos conseils précieux ont été essentiels pour la réussite de mon mémoire de Master. Je tiens à remercier très sincèrement le Professeur **Alassane DIÉDHIOU** pour avoir accepté de superviser ce travail.

Je tiens également à remercier le Professeur **Clément MANGA** et le Docteur **Kalilou DIALLO** pour avoir accepté de participer au jury de la soutenance. Je veux aussi remercier tous les enseignants du département de mathématiques.

J'exprime aussi ma reconnaissance envers mes professeurs de mathématiques du moyen et du secondaire, **M. SECK**, **M. SY**, **M. DIOP**, **M. DIÉDHIOU**, et **M. NDIAYE**; merci d'avoir cultivé en moi l'amour des mathématiques. Vos enseignements ont eu un impact majeur sur ma vie universitaire.

Mes remerciements s'adressent également à mes camarades de Probabilité-statistique, **Fatou DIENG**, **M.Saliou NDIONE** et **Bamba SECK**, et aussi à mes autres camarades du M2, en particulier **Abdoulaye DIOUF**, **Ibrahima TRAORE**, **Thierno DIALLO**, **Marie FAYE**, **Fatou DIÉMÉ**, **Christ Jésus BASSE** et **Abdourahmane DIATTA**. Votre présence et votre soutien ont été précieux tout au long de ce périple, et je suis fier d'avoir partagé cette expérience avec vous. Merci d'avoir été là pour moi, pour les encouragements, les échanges et les moments de camaraderie. Je vous souhaite à tous le meilleur pour vos projets futurs et j'ai la conviction que vous réussirez dans tout ce que vous entreprendrez.

Je suis convaincu que cette expérience universitaire restera gravée dans ma mémoire, et je suis reconnaissant envers tous ceux qui ont contribué à sa réussite.

DÉDICACES

Ce travail est dédié aux personnes de bonne foi qui ont toujours cru en moi et qui m'ont toujours accompagné dans les prières et les bénédictions. Je dédie donc ce mémoire :

- À mon oncle et mentor **Lamine MANE** qui m'a élevé et a cultivé en moi le sens du respect, de la dignité, de l'honneur, de l'intégrité et de la responsabilité. Je vous souhaite une longue vie pleine de bonheur et de réussite.

- À ma bien aimée mère et ma lumière **Yama MANE** qui m'a toujours soutenu et encouragé. Si nous sommes parvenus à ce niveau, c'est grâce à vous. Longue vie à vous maman.

- À ma bien aimée **Fatoumata SY**. Merci d'avoir été là.

- À mes bien aimées tantes **Amy et Mariama MANE** . Merci pour vos conseils, encouragements et votre amour.

- À mes frères et sœurs, **Sophiétou**, **Maimouna**, **Mariama**, **Amy**, **Diénabou SADIO**, **Aliou MANGAL**, **Souleymane**, **Moussa**, **Massar**, **El Hadji DIÉDHIOU** et **Samesidine DIATTA**. Merci pour votre soutien constant et vos encouragements.

- À ma marraine **Madame DIÉDHIOU** et sa famille (mention spéciale au grand frère **Cheikh Wally**). Votre soutien et votre amour ont été une source de motivation pour moi.

- À ma bien aimée tutrice et tante **Jacqueline MANGA** et sa famille (mention spéciale à Tanty Bayo) .

- À **Oustaz Chérif DIBA**, **Martiny DIATTA** et à toute leur famille.

- À **Laye FATY**, **Ismaïl** (mention spéciale à grand **Souleymane**) et toute la team boutique Souleymane. Merci beaucoup grand. Votre soutien et encouragement m'ont toujours permit de tenir bon.

- À mes grandes sœurs, **Fatou bintou** et **Aïssatou**. Merci pour votre soutien inconditionnel et votre amour.

- À mes "bradaframanadamada" **Alioune Badara WADE**, **Malick FAYE**, **Mouhamed Anta GAYE**, **Omar DIOP**, **Diamody KA**, **Saliou DIAW**, **Modou NDOUR**, **Kang-Rang Seth KOUMLA**, **Daouda NDIAYE**, **Mamina AÏDARA**, **Abdou Kounta FATY**, **Nouha COLY**, **Abdoul Khoudoussi DIALLO** et **El Hadji DIALLO**.

- À mes camarades de la TS2, en particulier **Lamine DIÉDHIOU**, **Ibrahima BÂ**, **Sény CAMARA**, **Arouna TRAORÉ** et **Mame Dïarra Camara**, pour leur esprit de fraternité.

Notations et Abréviations

AIC : Akaike Information Criterion.
BIC : Bayesian Information Criterion.
ddl : degrés de liberté.
i.e : c'est-à-dire.
i.i.d : indépendantes et identiquement distribuées.
IC : Intervalle de Confiance.
MCO : Moindres Carrés Ordinaires.
MV : Maximum de Vraisemblance.
rls : régression linéaire simple.
rlm : régression linéaire multiple.
TB : Tuberculose.

Résumé

Notre étude vise à comparer deux méthodes d'estimation en régression linéaire : les Moindres Carrés Ordinaires (MCO) et le Maximum de Vraisemblance (MV). Nous évaluons leurs performances respectives, analysons les facteurs influençant les variables d'intérêt, et appliquons ces méthodes à l'analyse du délai de diagnostic dans la tuberculose pulmonaire. Notre objectif est de fournir une compréhension approfondie de ces méthodes et de déterminer laquelle est la plus appropriée dans différents contextes.

Abstract

Our study aims to compare two estimation methods in linear regression : Ordinary Least Squares (OLS) and Maximum Likelihood (ML). We evaluate their respective performances, analyze the factors influencing the variables of interest, and apply these methods to the analysis of diagnostic delay in pulmonary tuberculosis. The objective of this study is to provide a comprehensive understanding of these methods and determine their suitability in different contexts.

Table des matières

Table des figures

Liste des tableaux

Introduction Générale

L'origine du terme "régression" remonte à Sir Francis Galton en 1885 lorsqu'il étudiait l'hérédité et cherchait à expliquer la taille des fils en fonction de celle des pères. Il observa que lorsque le père était plus grand que la moyenne, son fils avait tendance à être plus petit, et vice versa. Ces observations l'ont conduit à développer sa théorie de la "régression vers la moyenne". D'autre part, l'analyse de la causalité entre plusieurs variables remonte au milieu du XVIIIe siècle avec les travaux de R. Boscovich et Legendre.
En statistique, économétrie et apprentissage automatique, un modèle de régression linéaire cherche à établir une relation linéaire entre une variable expliquée et une ou plusieurs variables explicatives. Le modèle le plus simple est l'ajustement affine, qui tente d'expliquer le comportement d'une variable y en fonction d'une autre variable x à travers une droite. Le modèle de régression linéaire désigne généralement un modèle où l'espérance conditionnelle de y sachant x est une fonction affine des paramètres.
Dans cette étude comparative, notre intérêt se porte sur les estimateurs des Moindres Carrés Ordinaires (MCO) et du Maximum de vraisemblance (MV) dans le contexte de la régression linéaire. Notre objectif est d'examiner en détail ces deux méthodes d'estimation, de mettre en évidence leurs différences, avantages et limites, ainsi que d'évaluer leur performance relative. Nous nous appuierons sur diverses sources pour cette étude (cf. [1], [13], [31], [32]).

Avant d'entamer cette comparaison approfondie, nous commencerons par revisiter certains outils probabilistes et statistiques pertinents. Ces outils nous permettent de mieux comprendre les concepts clés et les principes sous-jacents aux méthodes d'estimation en régression linéaire, d'où le chapitre 1.

Ensuite, au chapitre 2, nous aborderons la régression linéaire simple. Nous examinerons en détail les méthodes d'estimation MCO et MV, en les comparant sur des critères tels que la précision, le biais et d'autres mesures d'évaluation appropriées.

Nous poursuivrons notre analyse au chapitre 3 en étudiant la régression linéaire multiple, où plusieurs variables explicatives sont utilisées pour prédire une variable cible. Nous réexaminerons les méthodes d'estimation MCO et MV dans ce contexte plus complexe et évaluerons leur performance respective.

Afin d'évaluer la robustesse des méthodes d'estimation et de mieux comprendre les mécanismes sous-jacents, nous réaliserons des simulations numériques. Le chapitre 4 présentera ces simulations numériques, qui nous permettront d'explorer différents scénarios et de mieux appréhender les facteurs influençant les estimations des coefficients de régression.

Enfin, au chapitre 5, nous illustrerons l'application pratique de nos analyses en étudiant le délai de diagnostic des patients atteints de la tuberculose (**TB**) pulmonaire. Nous analyserons les facteurs susceptibles d'influencer un diagnostic tardif et déterminerons quelle méthode d'estimation, entre les MCO et le MV, est la plus appropriée pour analyser ces données spécifiques.

Chapitre 1

Rappels d'outils probabilistes et statistiques

Dans ce chapitre, nous essayerons de rappeler quelques propriétés de la théorie des probabilités et statistiques qui nous seront utiles dans les chapitres à venir. Le but n'est absolument pas de fournir les bases de calcul des probabilités et statistiques (bases qui nous sont nécessaires à une bonne compréhension de la statistique inférentielle) ; mais plutôt d'exposer quelques résultats qui faciliteront la lecture et la compréhension du document.

1.1 Outils probabilistes

La théorie des probabilités est l'étude mathématique des phénomènes caractérisés par le hasard et l'incertitude. Le calcul des probabilités a commencé avec Blaise Pascal, Pierre Fermat, Christian Huygens et Jacques Bernoulli par l'analyse des jeux dits de hasard. Nous allons présenter ici quelques définitions utiles pour notre travail.
(Voir [14],[18]).

Définition 1.
Une famille $\mathcal{E}$ de parties d'un ensemble Ω (appelé univers ou ensemble fondamental) est appelé ***tribu*** *ou* ***σ- algèbre*** *si elle vérifie les propriétés suivantes :*

- $\Omega \in \mathcal{E}$;
- Si $(A_n)_n$ est une suite (éventuellement finie) d'éléments de $\mathcal{E}$, alors $\cup_{n\in\mathbb{N}} A_n \in \mathcal{E}$;
- Si A est un élément de $\mathcal{E}$, alors $\overline{A} \in \mathcal{E}$.

Exemple 1.
Le couple $(\Omega;\mathcal{E})$ est un ***espace probabilisable*** *et les éléments de $\mathcal{E}$ sont appelés* ***évènements****.*
Exemple *: On se donne un ensemble E non vide. Les trois familles de sous-ensembles de E suivantes sont des tribus de E :*

- $\mathcal{M} = \mathcal{P}(E)$*, la famille de tous les sous-ensembles de E est appelée* ***tribu triviale****.*
- $\mathcal{M} = \{\emptyset, E\}$ *est la* ***tribu grossière****, c'est la plus petite tribu sur E.*
- *Si on fixe A un sous-ensemble de E alors* $\mathcal{M} = \{\emptyset, E, A, \overline{A}\}$ *est la* ***tribu engendrée par*** *A.*

Définition 2.
On appelle ***tribu borélienne*** *sur $\mathbb{R}$, notée $\mathcal{B}(\mathbb{R})$, la plus petite tribu, au sens de l'inclusion, contenant tous les intervalles de $\mathbb{R}$.*

On peut donc donner maintenant la définition d'un espace probabilisé :

Définition 3.
On appelle ***probabilité*** *sur* $(\Omega;\mathcal{E})$ *(ou mesure de probabilité) une application* $\mathbb{P}$ *de* $\mathcal{E}$ *dans* $[0,1]$ *telle que :*

- $\mathbb{P}(\emptyset)=0$ et $\mathbb{P}(\Omega)=1$.
- *Pour toute suite* $(A_n)_{n\geqslant 0}$ *d'évènements, deux à deux incompatibles :*

$$\mathbb{P}(\cup_{n=0}^{+\infty}A_n)=\sum_{n=0}^{+\infty}\mathbb{P}(A_n) \qquad \textit{appelée la } \sigma\textit{-additivité.}$$

On appelle ***espace probabilisé****, le triplet* $(\Omega,\mathcal{E},\mathbb{P})$.

1.1.1 Hypothèses initiales

Dans tout ce qui suit, on se place sur $(\Omega,\mathcal{E},\mathbb{P})$ un espace de probabilité. À noter que Ω et $\mathcal{E}$ ne seront en général jamais précisés. On appellera X, Y ou Z des variables ou vecteurs aléatoires sur $(\Omega,\mathcal{E},\mathbb{P})$ i.e des fonctions mesurables à valeurs dans $\mathbb{R}$ (ou $\mathbb{R}^n$). On suppose par ailleurs que dans les propriétés suivantes, que les différents moments (espérance et variance) de ces variables existent.

1.1.2 Règles opératoires du calcul de l'espérance et de la variance d'une variable aléatoire (v.a)

L'espérance donne la position moyenne théorique d'une variable aléatoire (v.a). C'est un opérateur linéaire (de l'espace vectoriel constitué par l'ensemble des variables aléatoires dans $\mathbb{R}$). Ainsi, si **a** et **b** sont des constantes réelles ; alors :

$$\mathbb{E}(\mathbf{a}X+\mathbf{b}Y)=\mathbf{a}\mathbb{E}(X)+\mathbf{b}\mathbb{E}(Y).$$

La variance mesure l'écart quadratique (théorique) d'une variable aléatoire à son espérance. Elle se définit à partir de l'espérance de la façon suivante :

$$Var(X)=\mathbb{E}[(X-\mathbb{E}(X)^2]=\mathbb{E}(X^2)-(\mathbb{E}(X))^2.$$

Elle est quadratique et invariante par addition d'une constante :

$$Var(\mathbf{a}X+\mathbf{b})=\mathbf{a}^2Var(X).$$

La variance de la somme de variables aléatoires fait intervenir la covariance entre ces variables :

$$Var(\mathbf{a}X+\mathbf{b}Y)=\mathbf{a}^2Var(X)+\mathbf{b}^2Var(Y)+2\mathbf{ab}Cov(X,Y).$$

On note que la variance d'une variable aléatoire est une constante toujours positive, la covariance de deux variables aléatoires peut être négative. Si les variables X et Y sont non corrélées i.e $Cov(X,Y)=0$, ce qui arrive lorsqu'elles sont indépendantes, on obtient :

$$Var(X+Y)=Var(X)+Var(Y).$$

Si $X_1,\cdots\cdots,X_n$ sont indépendantes d'une même variable aléatoire X, on a :

$$\mathbb{E}\left(X_1+\cdots+X_n\right)=n\mathbb{E}(X) \text{ et } Var(X_1+\cdots+X_n)=nVar(X).$$

En conséquence, la variable aléatoire que constitue la moyenne empirique vérifie :

$$\mathbb{E}(\bar{X})=\mathbb{E}\left[\tfrac{1}{n}(X_1+\cdots+X_n)\right]=\mathbb{E}(X) \text{ et } Var(\bar{X})=\tfrac{1}{n}Var(X).$$

1.1.3 Lois de probabilités de variables aléatoires

1.1.3.1 Lois de probabilités et fonctions de répartition

On dira qu'une variable aléatoire suit une loi de probabilité $\mathbb{P}_X$ notée $X \sim \mathbb{P}_X$ lorsque tout ensemble A borélien de $\mathbb{R}$: $\mathbb{P}_X(A) = \mathbb{P}(X \in A)$. On dira que cette loi admet une densité f_x par rapport à la mesure de Lebesgue lorsque l'on pourra écrire pour tout ensemble A borélien de $\mathbb{R}$:

$$\mathbb{P}_X(A) = \int_A f_x(x)dx.$$

On peut également définir la fonction de répartition $\mathbb{F}_X$ de la variable X, telle que $\forall\ x \in \mathbb{R}$

$$\mathbb{F}_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(]-\infty; x]).$$

Cette fonction $\mathbb{F}_X : \mathbb{R} \longrightarrow [0,1]$ est croissante, continue à droite et a une limite à gauche, telle que :

$$\lim_{x\to-\infty} \mathbb{F}_X = 0 \text{ et } \lim_{x\to-\infty} \mathbb{F}_X = 1.$$

Il y'a une correspondance bijective entre la connaissance de $\mathbb{P}_X$ et celle de $\mathbb{F}_X$.
La fonction de répartition (f.d.r) permet également de déterminer les quantiles qui sont essentielles à la construction d'intervalle de confiance (**I.C**) et tests.

1.1.3.2 Quantiles d'une loi de probabilité

Soit $\alpha \in [0,1]$. Des propriétés de la fonction de répartition , on en déduit qu'il $\exists\ x_\alpha$ tel que :

$$\lim_{x\to x_\alpha} \mathbb{F}_X(x) \leq \alpha \leq \mathbb{F}_X(x_\alpha)\ (\bullet.1).$$

Soit $I_\alpha = \{x_\alpha \in \mathbb{R} \text{ tel que } x_\alpha \text{ vérifie } (\bullet.1)\}$. On appelle **quantile** (ou fractile, ou percentile en anglais) d'ordre α de la loi $\mathbb{P}_X$, noté q_α, le milieu de l'intervalle I_α. Évidemment, lorsque X admet une distribution continue par rapport à la mesure de Lebesgue, $q_\alpha = \mathbb{F}_X^{-1}(\alpha)$, où $\mathbb{F}_X^{-1}$ désigne la fonction réciproque de $\mathbb{F}_X$. Deux cas se présentent et il est important de les connaître :
• pour $\alpha = 0.5$, $q_{0.5}$ est appelé la médiane de $\mathbb{P}_X$;
• pour $\alpha = 0.25$ et $\alpha = 0.75$ (resp.), $q_{0.25}$ et $q_{0.75}$ sont appelés **premier** et **troisième** quartile (resp.) de $\mathbb{P}_X$.

1.1.3.3 Les principales lois utilisées en modèle linéaire

Dans le cadre du modèle linéaire, nous allons principalement utiliser des lois de probabilités possédant une densité par rapport à la mesure de Lebesgue.

1. **Loi Normale (ou gaussienne) de moyenne μ et de variance σ^2 ($\mathcal{N}(\mu, \sigma^2)$)**
   Elle est à valeurs dans $\mathbb{R}$ et de densité par rapport à la mesure de Lebesgue :
   $$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \text{ On a : } \mathbb{E}(X) = \mu \text{ et } Var(X) = \sigma^2.$$
   Ainsi lorsque $\mathbb{E}(X) = 0$ et $Var(X) = 1$, elle est dite **centrée** et **réduite** et on a :
   $$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$
2. **Loi de χ^2 à n degrés de liberté**
   Soit $X_1, \cdots\cdots, X_n$ n variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$, alors $S = X_1^2 + \cdots + X_n^2$ suit une loi de χ^2 à n degrés de liberté, noté $\chi^2(n)$. Cette loi est à valeurs dans $\mathbb{R}$, d'espérance n et de variance $2n$. C'est aussi la Loi de Gamma de paramétra $(n/2, 1/2)$ i.e $X \sim \chi^2(n)$ admet pour densité par rapport à la mesure de Lebesgue :
   $$f_X(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp\left(-\frac{x}{2}\right) \mathbb{1}_{\{x\geqslant 0\}},$$

où la fonction Gamma est telle que : $\Gamma(x) = \int_0^\infty x^{a-1}\exp(-\frac{x}{2})$, pour $\mathbf{a} \geq 0$. En fin, si $X \sim \chi^2(n)$, par définition on dira que $Y = \sigma^2 \times X$ suit une loi $\sigma^2 \times \chi^2(n)$.

3. **Loi de Student à n degrés de liberté**

   La loi de Student à n degrés de liberté, notée T(n) est la loi du quotient :

$$T = \frac{N}{\sqrt{S/n}} \text{ où } N \sim \mathcal{N}(0,1) \text{ et } S \sim \chi^2(n).$$

   N et S étant deux variables aléatoires indépendantes, il est possible de déterminer la densité d'une telle loi par rapport à la mesure de Lebesgue :

$$f_X(x) = \frac{1}{\sqrt{n} \times B(1/2, n/2)} \left(1 + \tfrac{x^2}{n}\right)^{(n+1)/2},$$

   où la fonction **Bêta** est telle que $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ $\forall a, b > 0$.

   Remarque.

   Par la loi des grands nombres, plus n est grand, plus S est proche de son espérance qui vaut n. Le dénominateur est donc plus proche de 1. Il s'en suit que la loi T(n) est d'autant plus proche d'une loi normale que n est grand.

   Un des principaux intérêts de la loi de Student réside dans le fait que si $X_1, \cdots\cdots, X_n$ sont des variables aléatoires indépendantes de loi $\mathcal{N}(m, \sigma^2)$ si l'on considère la moyenne et la variance empirique :

$$\bar{X}_n = \frac{1}{n}(X_1, \cdots\cdots, X_n) \text{ et } \bar{\sigma}_n^2 = \frac{1}{n-1}\left((X_1 - \bar{X}_n)^2 + \cdots\cdots + (X_d - \bar{X}_n)^2\right),$$

   alors $T = \dfrac{\sqrt{n}(\bar{X}_n - m)}{\sqrt{\bar{\sigma}^2}_n}$ suit la loi de Student à $(n-1)$ degré de libertés.

4. **Loi de Fisher à n_1 et n_2 degrés de liberté**

   Soit S_1 et S_2 deux variables aléatoires indépendantes de loi respectives $\chi^2(n_1)$ et $\chi^2(n_2)$. Alors par définition :

$$F = \frac{S_1/n_1}{S_2/n_2} \text{ suit une loi de } \textbf{Fisher} \text{ à } n_1 \text{ et } n_2 \text{ degrés de liberté, notée : } F(n_1, n_2)\ .$$

   Remarque.

   Par les mêmes considérations que précédemment, la loi de F est d'autant plus proches de 1 que les degrés de liberté n_1 et n_2 sont grands. On a les propriétés suivantes :

   (a) Si $F \sim F(n_1, n_2)$, alors la loi de $\frac{n_1}{n2}$ est la loi Bêta de seconde espèce de paramètre (n_1 et n_2) i.e F est à valeurs dans $\mathbb{R}_+$ et admet la densité par rapport à la mesure de Lebesgue.

$$f_X(x) = \frac{n_1^{n_1/2} n_2^{n_2/2}}{B(n_1/2, n_2/2)} \frac{x^{n_1/2-1}}{(n_2 + n_1 - x)^{(n_1+n_2)/2}} \mathbb{1}_{\{x \geqslant 0\}};$$

   (b) Si $F \sim F(n_1, n_2)$, alors $\mathbb{E}(F) = \dfrac{n_2}{n_1 - 1}$ lorsque $n_2 > 0$ et $Var(F) = \dfrac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2-4)(n_2-2)^2}$, lorsque $n_2 > 4$;

   (c) Si $T \sim T(n)$, alors $T^2 \sim F(1, n)$.

1.1.4 Vecteurs Aléatoires

Les vecteurs aléatoires ont des applications dans beaucoup de domaines. Ils permettent de décrire des phénomènes aléatoires qui évoluent dans $\mathbb{R}^n$.

Espérance et matrice de covariance

Comme on se place en général dans la base canonique orthonormale de $\mathbb{R}^n$, si μ_i désigne l'espérance de X_i, l'espérance de $X = (X_1, \cdots\cdots, X_n)$ est le vecteur

$$\mathbb{E}(X) = \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_n \end{pmatrix}.$$

Définition 4.

Si X et Y sont de carré intégrables, on appelle covariance de X et Y le nombre

$$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Le vecteur aléatoire $X \in L^2$ si et seulement si pour tout $i = 1, \cdots\cdots, n; X_i \in L^2$. La matrice de variance-covariance Σ de X est définie par :

$$\Sigma = (\Sigma_{ij})_{1\leq i,j\leq n} \begin{pmatrix} \sigma_1^2 & Cov(X_1, X_2) & \cdots & Cov(X_1, X_n) \\ \vdots & \sigma_2^2 & \vdots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \cdot & \sigma_n^2 \end{pmatrix} = \mathbb{E}(XX^t) - \mu\mu^t.$$

Remarque.

La matrice de variance-covariance Σ est symétrique et définie positive (i.e pour tout $v \in \mathbb{R}^n$, $\langle v, \Sigma v\rangle \geq 0$).

On peut supposer que les variables sont centrées, sinon considérer $Y_i = X\mathbb{E}(X_i)$, on a alors

$$\begin{aligned} \langle v, \Sigma v\rangle &= \sum \Sigma_{ij} v_i v_j \\ &= \sum \mathbb{E}(X_i X_j) v_i v_j \\ &= \mathbb{E}(\sum (v_i X_i v_j X_j)) \\ &= \mathbb{E}[\sum (v_i X_i)(\sum (v_j X_j] \\ &= \mathbb{E}[(\sum (v_i X_i)^2] \geq 0. \end{aligned}$$

Transformations linéaires

Considérons le changement de variables linéaires, $Y = AX$ où A est une matrice de constantes telle que l'opération AX soit possible. Alors on a : $\mu_Y = A\mu_X, \Sigma_Y = A\Sigma_X A^t$.

Théorème 1. *(Voir [15])*

Une condition nécessaire et suffisante pour qu'une matrice Σ symétrique soit la matrice de variance d'un vecteur aléatoire est que Σ soit une matrice positive.

Preuve.

Condition nécessaire est évidente (d'après ce qui précède), alors montrons la réciproque.

Supposons que Σ est une matrice positive. Nous devons montrer qu'il existe un vecteur aléatoire dont la matrice de variance-covariance est Σ.

Considérons la décomposition spectrale de Σ :

$$\Sigma = PDP^t,$$

où P est une matrice orthogonale dont les colonnes sont les vecteurs propres de Σ et D est une matrice diagonale dont les éléments sont les valeurs propres correspondantes.

Définissons un vecteur aléatoire $X = PZ$, où Z est un vecteur aléatoire standard (chaque composante est une variable aléatoire indépendante et identiquement distribuée selon une loi normale standard).

Calculons la matrice de variance-covariance de X :

$$\Sigma_X = E[(PZ)(PZ)^t] = E[PZZ^tP^t] = PIdP^t = \Sigma,$$

où Id est la matrice identité.

Ainsi, nous avons montré qu'en choisissant $X = PZ$, où Z est un vecteur aléatoire standard, la matrice de variance-covariance de X est Σ.

Par conséquent, Σ est la matrice de variance-covariance d'un vecteur aléatoire. □

Proposition 1.

Soit X un vecteur aléatoire de $\mathbb{R}^n$ d'espérance μ et de matrice de covariance Σ régulière (det $\Sigma \neq 0$). Alors

1. *Le Vecteur $Y = \Sigma^{\frac{1}{2}}(X - \mu)$ est un vecteur aléatoire centré réduit à composantes non corrélées.*
2. *La variable $(X\mu)\Sigma^{-1}(X - \mu)^t$ a pour espérance n.*

Démonstration.

1. Le vecteur $Y = \Sigma^{-\frac{1}{2}}(X - \mu)$ est un vecteur aléatoire centré réduit à composantes non corrélées.

Pour montrer que Y est centré, nous devons vérifier que son espérance est nulle. Calculons l'espérance de Y :

$$\mathbb{E}(Y) = \mathbb{E}\left(\Sigma^{-\frac{1}{2}}(X - \mu)\right) = \Sigma^{-\frac{1}{2}}\mathbb{E}(X - \mu) = \Sigma^{-\frac{1}{2}}(\mu - \mu) = \Sigma^{-\frac{1}{2}}0 = 0.$$

Pour montrer que les composantes de Y sont non corrélées, nous devons montrer que la matrice de covariance de Y est diagonale. Calculons la matrice de covariance de Y :

$$\mathrm{Cov}(Y) = \mathrm{Cov}(\Sigma^{-\frac{1}{2}}(X - \mu)) = \Sigma^{-\frac{1}{2}}\mathrm{Cov}(X - \mu)(\Sigma^{-\frac{1}{2}})^t = \Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}} = Id.$$

La matrice de covariance de Y est une matrice identité, ce qui signifie que les composantes de Y sont non corrélées.

Ainsi, nous avons montré que le vecteur Y est un vecteur aléatoire centré réduit à composantes non corrélées.

2. La variable $(X - \mu)\Sigma^{-1}(X - \mu)^t$ a pour espérance n.

Calculons l'espérance de cette variable :

$$\mathbb{E}\left((X - \mu)\Sigma^{-1}(X - \mu)^t\right) = \mathbb{E}\left(\mathrm{tr}((X - \mu)\Sigma^{-1}(X - \mu)^t)\right) = \mathrm{tr}\left(\mathbb{E}((X - \mu)(X - \mu)^t)\right) = \mathrm{tr}\left(\Sigma^{-1}\Sigma\right) = \mathrm{tr}(Id) = n.$$

Ce qui montre que la variable $(X - \mu)\Sigma^{-1}(X - \mu)^t$ a pour espérance n. □

Fonction caractéristique

Définition 5.

On appelle fonction caractéristique du vecteur aléatoire X, la fonction de l'argument vectoriel a défini par :

$$\varphi_X(a) = \mathbb{E}\left[\exp(\langle a, X\rangle)\right] = \mathbb{E}\left[\exp(\textstyle\sum_{i=1}^{n} a_i Xi)\right].$$

Théorème 2. *(Voir [4],[15])*

Les composantes $(X_1, \cdots\cdots, X_n)$ de X sont indépendantes si et seulement si, la fonction caractéristique de X est égale au produit des fonctions caractéristiques de ses composantes :

$$\varphi_X(a) = \prod_{i=1}^{n} \varphi_{X_i}(a_i).$$

1.1.5 Vecteurs aléatoires gaussiens

Soit $X_1, \cdots\cdots, X_n$ des variables aléatoires indépendantes de même loi normale centrée et de variance σ^2 (i.e $X_i \sim \mathcal{N}(0,\sigma^2), i = 1, \cdots\cdots, n$). Soit un vecteur $X \in \mathbb{R}^n$ tel que $X = X_1, \cdots\cdots, X_n$. En raison de l'indépendance, X est un vecteur gaussien admettant une densité f_X (par rapport à la mesure de Lebesgue) qui est le produit des densités de chacune des coordonnées, soit :

$$\begin{aligned} f_X(x_1, \cdots\cdots, x_n) &= f_{X_1}(x_1) \times \times f_{X_2}(x_2) \times \cdots \times \times f_{X_n}(x_n) \\ &= (2\pi\sigma^2)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(x_1 + \cdots\cdots + x_n)\right\} \\ &= (2\pi\sigma^2)^{n/2} \exp\left\{-\frac{|| \ x \ ||^2}{2\sigma^2}\right\}. \end{aligned}$$

On voit que la densité de X ne dépend que de la norme $|| \ X \ ||$: elle est constante sur toutes les sphères centrées en zéro. Cela implique qu'elle est invariante par rotation ou symétrie orthogonale d'axes passant par zéro : elle est invariante par toutes les isométries de $\mathbb{R}^n$; on dira que X suit une loi gaussienne isotrope.
Rappel : Les isométries correspondent à des changements de bases orthonormées (**BON**). Par conséquent on a la première propriété importante :

Propriétés 1.
*Soit X un vecteur aléatoire de $\mathbb{R}^n$ de loi normale isotrope et de variance σ^2 i.e dans une **BON** les coordonnées de X vérifient : $\mathbb{E}(X) = 0$ et$Var(X) = \sigma^2 \cdot Id$. Alors les coordonnées de X dans toute BON sont encore des lois $\mathcal{N}(0,\sigma^2)$ indépendantes.*

Voici maintenant l'un des résultats couramment appelé **Théorème de Cochran**, que nous utilisons le plus fréquemment, et nous en donnons donc une démonstration.

Théorème 3. *(Théorème de Cochran. Voir [1],[11])*
Soient E_1et E_2 deux sous-espace vectoriel (s.e.v) de $\mathbb{E} = \mathbb{R}^n$ de dimensions respectives d_1 et d_2 et soit X un vecteur aléatoire de $\mathbb{R}^n$ de loi normale centrée isotrope de variance σ^2. Alors $\mathbf{P}_{E_1}(X)$ et $\mathbf{P}_{E_2}(X)$ sont deux variables aléatoires centrées indépendantes et $|| \ \mathbf{P}_{E_1}(X) \ ||^2$ (resp. $|| \ \mathbf{P}_{E_2}(X) \ ||^2$) est une loi $\sigma^2 \cdot \chi^2(d_1)$ (rep. $\sigma^2 \cdot \chi^2(d_1)$).
Ce théorème se généralise naturellement pour $2 \leq m \leq n$ s.e.v orthogonaux $(E_i)_{1\leq i\leq m}$ de $\mathbb{E} = \mathbb{R}^n$.

Preuve.
Soient $(e_1, \cdots\cdots, e_{d_1})$ et $(e_{d_1+1}, \cdots\cdots, e_{d_1+d_2})$ deux BON de E_1et E_2 respectivement. L'ensemble de ces deux bases peut être complété en :
$(e_1, \cdots\cdots, e_{d_1}, e_{d_1+1} \cdots\cdots, e_{d_1+d_2}, e_{d_1+d_2+1}, \cdots\cdots, e_n)$ pour former un BON de $\mathbb{R}^n$.
Soient $(X_1, \cdots\cdots, X_n)$, les coordonnées de X dans cette base, elles sont indépendantes de loi $\mathcal{N}(0,\sigma^2)$ car le changement de base est orthonormal et on a vu que la distribution de X était conservée par transformation isométrique . Comme :

$$\begin{aligned} \mathbf{P}_{E_1}(X) &= X_1 e_1 + \cdots\cdots + X_{d_1} e_{d_1} = || \ \mathbf{P}_{E_1}(X) \ ||^2 \\ &= \sigma^2 \left(\left(\frac{X_1}{\sigma}\right)^2 + \cdots\cdots + \left(\frac{X_{d_1}}{\sigma}\right)^2 \right); \\ \mathbf{P}_{E_2}(X) &= X_{d_1+1} e_{d_1+1} + \cdots\cdots + X_{d_1+d_2} e_{d_1+d_2} = || \ \mathbf{P}_{E_2}(X) \ ||^2 \\ &= \sigma^2 \left(\left(\frac{X_{d_1}}{\sigma}\right)^2 + \cdots\cdots + \left(\frac{X_{d_1+d_2}}{\sigma}\right)^2 \right). \end{aligned}$$

On remarque bien l'indépendance entre les deux projections et le fait que la loi de $|| \mathbf{P}_{E_1}(X) ||^2$ (resp. $|| \mathbf{P}_{E_2}(X) ||^2$) est une loi de $\sigma^2 \cdot \chi^2(d_1)$ (rep. $\sigma^2 \cdot \chi^2(d_2)$). □

Densité de la loi Multi-normale

Théorème 4.
Si la matrice de variance-covariance de $X \in \mathbb{R}^n$, un vecteur gaussien, Σ, est régulière, alors le vecteur admet une densité par rapport à la mesure de Lebesgue qui s'écrit :

$$f_X(x_1, x_2, \cdots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}(|det\Sigma|)^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}(x-\mu)^t\Sigma(x-\mu)\right).$$

Preuve : Le vecteur $Y = \Sigma^2(X - \mu)$ est gaussien, et les composantes sont centrées réduites et indépendantes. La densité de Y s'écrit alors :

$$g(y) = \prod_{i=1}^{n} g(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\sum_{i=1}^{n} y_i^2).$$

Il suffit ensuite d'appliquer la formule du changement de variables, avec comme jacobien
$detJ = det\Sigma^2$.

1.1.6 Théorèmes Limites

Quelle est la fréquence des piles lorsque l'on jette un grand nombre de fois une pièce de monnaie ?
Comment le sondage peut-il permettre d'estimer le score d'un candidat à une élection ?
C'est à ce genre de questions que répondent les deux théorèmes limites suivants. Le premier appelé loi des Grands Nombres car on a longtemps vu qu'il ressortait des lois de la nature (du même ordre que la gravitation) de plus en plus de la moyenne théorique, ou espérance, lorsque le nombre de données croit. Plus précisément son énoncé est le suivant :

Théorème 5. *(Loi Forte des Grands Nombres : LGN)*
Soient $(Y_1, \cdots\cdots, Y_n)$ n variables aléatoires indépendantes et identiquement distribuées suivant la même loi d'espérance μ (donc $\mathbb{E}(Y) < +\infty$).Soient $\bar{Y}_n = \frac{1}{n}(Y_1 + \cdots\cdots + Y_n)$ la moyenne empirique. Alors quand n est grand

$$\bar{Y}_n \xrightarrow[n\to\infty]{} \mu \; p.s \; .$$

Preuve.
La preuve de la loi forte des grands nombres repose sur deux éléments clés : l'inégalité de Bienaymé-Tchebychev et la convergence presque sûre de la série des variables aléatoires.

Étape 1 : Utilisation de l'inégalité de Bienaymé-Tchebychev

Pour tout $\varepsilon > 0$, nous avons :

$$P\left(\left|\bar{Y}_n - \mu\right| > \varepsilon\right) \leq \frac{\mathrm{Var}(\bar{Y}_n)}{\varepsilon^2}.$$

Étape 2 : Application de la convergence presque sûre

Comme les variables aléatoires $Y_1, \cdots\cdots, Y_n$ sont i.i.d, nous pouvons utiliser la convergence presque sûre de la série des variables aléatoires pour montrer que :

$$\lim_{n\to\infty} \bar{Y}_n = \mu \quad \text{presque sûrement.}$$

Étape 3 : Convergence vers zéro de la probabilité

En utilisant l'étape 2, nous savons que $\lim_{n\to\infty} \bar{Y}_n = \mu$ presque sûrement. Cela signifie que la probabilité que $\bar{Y}_n$ s'écarte de μ de plus de ε devient de plus en plus petite lorsque n tend vers l'infini.

Formellement, pour tout $\varepsilon > 0$:

$$P\left(\lim_{n\to\infty} \left|\bar{Y}_n - \mu\right| > \varepsilon\right) = 0.$$

En combinant les étapes 1 et 3, nous obtenons :

$$P\left(\lim_{n\to\infty} \bar{Y}_n = \mu\right) = 1,$$

ce qui établit la loi forte des grands nombres. □

On note que les hypothèses peuvent être étendues à des variables qui ne sont pas indépendantes entre elles, ou qui n'ont forcément la même variance.
Le théorème 6 suivant précise en quelque sorte la manière dont la moyenne empirique se rapproche de l'espérance :

Théorème 6. *(Théorème central Limite : TCL)*
Soient $(Y_1, \cdots\cdots, Y_n)$ n variables aléatoires indépendantes distribuées suivant la même loi d'espérance μ et de variance σ^2.Soit $\bar{Y}_n = \frac{1}{n}(Y_1 + \cdots\cdots + Y_n)$ la moyenne empirique. Alors quand n est grand,

$$\sqrt{n}\left(\frac{\bar{Y}_n - \mu}{\sigma}\right) = \frac{\bar{Y}_n - \mu}{\sqrt{\sigma^2/n}} \xrightarrow[n\to\infty]{} \mathcal{N}(0,1) \text{ en loi.}$$

Preuve.
La preuve du TCL repose sur l'utilisation de la fonction caractéristique. La fonction caractéristique d'une variable aléatoire X est définie comme $\phi_X(a) = E[e^{iaX}]$, où a est un réel et i est l'unité imaginaire.

Pour chaque variable aléatoire Y_i, nous pouvons définir sa fonction caractéristique $\phi_{Y_i}(a)$. Puisque les variables aléatoires $Y_1, \ldots, Y_n$ sont indépendantes, la fonction caractéristique de leur somme peut être calculée en multipliant les fonctions caractéristiques individuelles :

$$\phi_{\overline{Y}_n}(a) = \prod_{i=1}^{n} \phi_{Y_i}\left(\frac{a}{n}\right).$$

En utilisant l'approximation de Taylor du premier ordre pour la fonction exponentielle, nous pouvons écrire :

$$\phi_{Y_i}\left(\frac{a}{n}\right) = 1 + \frac{ia}{n}\mathbb{E}[Y_i] - \frac{a^2}{2n^2}\mathbb{E}[Y_i^2] + O\left(\frac{a}{n}\right),$$

où $O\left(\frac{a}{n}\right)$ est le terme d'erreur qui tend vers zéro lorsque n tend vers l'infini.
En substituant cette approximation dans la fonction caractéristique de $\overline{Y}_n$, nous obtenons :

$$\phi_{\overline{Y}_n}(a) = \left(1 + \frac{ia}{n}\mu - \frac{a^2}{2n^2}\sigma^2 + O\left(\frac{a}{n}\right)\right)^n.$$

Maintenant, nous prenons la limite lorsque n tend vers l'infini :

$$\lim_{n\to\infty} \phi_{\overline{Y}_n}(a) = \lim_{n\to\infty} \left(1 + \frac{ia}{n}\mu - \frac{a^2}{2n^2}\sigma^2 + O\left(\frac{a}{n}\right)\right)^n.$$

En utilisant la limite classique $e^x = \lim_{n\to\infty}(1 + \frac{x}{n})^n$, nous pouvons réécrire la limite comme suit :

$$\lim_{n\to\infty} \phi_{\overline{Y}_n}(a) = e^{ia\mu - \frac{a^2}{2}\sigma^2}.$$

Cette expression est la fonction caractéristique de la loi normale $\mathcal{N}(\mu, \sigma^2)$. Ainsi, la limite de la fonction caractéristique de $\sqrt{n}(\frac{\overline{Y}_n - \mu}{\sigma})$ est la fonction caractéristique de la loi normale standard $\mathcal{N}(0,1)$.

Par le théorème de continuité des fonctions caractéristiques, cela implique que la variable aléatoire $\sqrt{n}(\frac{\overline{Y}_n - \mu}{\sigma})$ suit la loi normale standard $\mathcal{N}(0,1)$. □

La traduction « intuitive » de ce résultat est que « la moyenne empirique tend à être gaussienne » quand le nombre de données devient grand, $\bar{Y}_n$ a une loi très proche la loi gaussienne de moyenne μ et de variance σ^2/n.

1.2 Outils Statistiques

1.2.1 Échantillonnage et Modèles statistiques

Le problème de l'estimation est l'impossibilité de connaître exactement la valeur d'un paramètre inconnu θ. Ce problème est très général et a des aspects distincts. Les observations par exemple obtenues à partir d'une méthode d'échantillonnage, permettent de construire une estimation de θ. Ainsi chaque observation est la valeur d'une variable aléatoire X dont la loi dépend de θ. Cela revient à doter l'état de la nature inconnue d'un modèle probabiliste. Ce dernier est complété par un modèle d'échantillonnage décrivant la manière dont les observations sont recueillies.

Échantillonnage

Soit X une variable aléatoire réelle (v.a.r). Un **échantillon aléatoire** d'effectifs $n \geq 1$ est un vecteur aléatoire $X = (X_1, \cdots\cdots, X_n)$ à n paramètres qui sont n v.a indépendantes suivant la même loi que X (appelé variable parente).

• **Statistique de l'échantillon** : Toute variable aléatoire T, fonction de l'échantillon aléatoire $(X_1, \cdots\cdots, X_n)$ est appelée statistique de l'échantillon.

Modèles

• **Modèles Statistiques** : Un modèle statistique est définie par la donnée d'une caractéristique d'un vecteur X et d'une famille de loi de probabilités de X notée $(\mathbb{P}_\theta)_{\theta\in\Theta}$.

• **Modèles paramétriques et non paramétriques** : lorsque la famille des lois de probabilité $(\mathbb{P}_\theta)_{\theta\in\Theta}$ peut être indexée par un paramètre θ dont l'ensemble des valeurs possibles, notée Θ, est un ensemble de $\mathbb{R}^n$, le modèle est appelé **modèle paramétrique**. dans le cas contraire, le modèle est appelé **modèle non paramétrique**.

1.2.2 Estimateur et propriétés d'un estimateur

Estimateur et Estimation

• **Estimateur** : Si $(X_1, \cdots\cdots, X_n)$ est un échantillon aléatoire d'effectifs n de loi parente de X, alors on appelle **estimateur** du paramètre θ toute fonction h_n de l'échantillon aléatoire $(X_1, \cdots\cdots, X_n)$ notée $\hat{\theta}_n$: $\hat{\theta}_n = h_n(X_1, \cdots\cdots, X_n)$.

Remarque.

- *À priori l'estimateur $\hat{\theta}_n$ est à valeur dans un ensemble Θ contenant l'ensemble des valeurs possibles du paramètre θ ;*
- *$\hat{\theta}_n$ est une v.a de loi de probabilité qui dépend du paramètre θ ;*
- *$\hat{\theta}_n$ peut être univarié ou multivarié.*

•**Estimation** : Une fois l'échantillon prélevé, nous disposons de n valeurs observées $x_1, \cdots\cdots, x_n$, ce qui nous fournit une valeur $h_n(x_1, \cdots\cdots, x_n)$, qui est une réalisation de $\hat{\theta}_n$ et que nous appelons **estimation**.

Remarque.

- *Nous distinguons de la variable aléatoire $\hat{\theta}_n$ de sa variable observée $\hat{\theta}_n \quad (x_1, \cdots\cdots, x_n)$;*
- *Nous utilisons les notations suivantes :*
  - *i) $(X_1, \cdots\cdots, X_n)$ désigne l'échantillon aléatoire de taille n et les n observations ne sont pas encore à disposition ;*
  - *ii) $(x_1, \cdots\cdots, x_n)$ désigne une réalisation de l'échantillon aléatoire et les n observations sont à disposition.*

∗ *Il faut systématiquement se demander « suis-je entrain de manipuler des v.a ou l'une de ses réalisation ? »*

Propriétes d'un estimateur

Le choix d'un estimateur va reposer sur ses qualités. Le premier défaut concerne la possibilité de comporter un biais.

• Biais d'un estimateur

Le biais de $\hat{\theta}_n$ se définit comme étant l'écart entre la moyenne et la vraie valeur du paramètre θ. IL est noté par

$$b(\hat{\theta}_n, \theta) = \mathbb{E}(\hat{\theta}_n) - \theta \ .$$

• Estimateur sans biais

$\hat{\theta}_n$ est un estimateur sans biais (ou non biaisé) du paramètre θ si

$$b(\hat{\theta}_n, \theta) = 0 \text{ i.e } \mathbb{E}(\hat{\theta}_n) = \theta \ .$$

• Estimateur asymptotiquement sans biais

Un estimateur $\hat{\theta}_n$ est asymptotiquement sans biais pour θ si

$$\lim_{n \to +\infty} b(\hat{\theta}_n, \theta) = 0 \Leftrightarrow \lim_{n \to +\infty} \mathbb{E}(\hat{\theta}_n) = \theta.$$

• Écart (ou Erreur) Quadratique Moyen

Si $\hat{\theta}_n$ est un estimateur de θ, on mesure la précision de $\hat{\theta}_n$ par l'écart quadratique moyen, noté **E.Q.M** :

$$\textbf{E.Q.M}(\hat{\theta}_n) = \mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right] = Var(\hat{\theta}_n) + b(\hat{\theta}_n, \theta)^2 \ .$$

En effet, rappelons nous d'abord que $b(\hat{\theta}_n, \theta) = \mathbb{E}(\hat{\theta}_n) - \theta$ et $\mathbb{E}(\hat{\theta}_n)$ sont des constantes. De plus par la linéarité de l'espérance on a :

$$\begin{aligned}
\textbf{E.Q.M}(\hat{\theta}_n) &= \mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n) + b(\hat{\theta}_n, \theta)\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n, \theta)\right)^2 + 2\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right) b(\hat{\theta}_n, \theta) + b(\hat{\theta}_n, \theta)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right)^2\right] + 2\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right) b(\hat{\theta}_n, \theta) + b(\hat{\theta}_n, \theta)^2 \\
&= Var(\hat{\theta}_n) + 2\left(\mathbb{E}(\hat{\theta}_n) - \mathbb{E}(\hat{\theta}_n)\right) b(\hat{\theta}_n, \theta) + b(\hat{\theta}_n, \theta)^2 \\
&= Var(\hat{\theta}_n) + b(\hat{\theta}_n, \theta)^2.
\end{aligned}$$

Remarque.

Si $\hat{\theta}_n$ est un estimateur sans biais, alors :

$$\boldsymbol{E.Q.M}(\hat{\theta}_n) = Var(\hat{\theta}_n) \ .$$

De plus si $\Theta \subset \mathbb{R}^n$, elle se définit comme suit : $\boldsymbol{E.Q.M}(\hat{\theta}_n) = \mathbb{E}(\|\hat{\theta}_n - \theta\|^2)$.

Propriétés 2.

Entre deux estimateurs de θ, nous choisissons celui dont l'écart quadratique moyen, où le risque est le plus faible.

• Estimateur relativement plus efficace

Un estimateur $\hat{\theta}_n^1$ est **relativement plus efficace** qu'un estimateur $\hat{\theta}_n^2$, s'il est plus précis que le second i.e :

$$\textbf{E.Q.M}(\hat{\theta}_n^1) \leq \textbf{E.Q.M}(\hat{\theta}_n^2) \ .$$

L'efficacité relative de deux estimateurs, $\hat{\theta}_n^1$ et $\hat{\theta}_n^2$ est donnée par :

$$eff(\hat{\theta}_n^1, \hat{\theta}_n^2) = \frac{\textbf{E.Q.M}(\hat{\theta}_n^1)}{\textbf{E.Q.M}(\hat{\theta}_n^1)} \ .$$

• Estimateur sans biais optimal

On appelle **estimateur sans biais optimal** parmi les estimateurs sans biais, un estimateur $\hat{\theta}_n$ préférable à tout autre au sens de la variance i.e l'estimateur le plus efficace parmi tous les estimateurs sans biais.

•Estimateur convergent

Un estimateur $\hat{\theta}_n$ est un estimateur **convergent**, s'il converge en proba vers θ quand $n \longrightarrow +\infty$.

Propriétés 3.

Si un estimateur est sans biais et que sa variance tend vers 0 *quand* $n \longrightarrow \infty$ *alors cet estimateur est convergent.*

Démonstration.

Supposons que l'estimateur $\hat{\theta}_n$ est sans biais, c'est-à-dire $\mathbb{E}(\hat{\theta}_n) = \theta$, où θ est la vraie valeur du paramètre à estimer.

De plus, supposons que la variance de l'estimateur $\hat{\theta}_n$ tende vers zéro lorsque $n \to \infty$, i.e $\lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_n) = 0$.

Maintenant, nous voulons montrer que $\hat{\theta}_n$ converge en probabilité vers θ, i.e $\lim_{n\to\infty} \Pr(|\hat{\theta}_n - \theta| > \varepsilon) = 0$ pour tout $\varepsilon > 0$.

En utilisant l'inégalité de Tchebychev, nous avons :

$$\Pr(|\hat{\theta}_n - \theta| > \varepsilon) \leq \frac{\mathrm{Var}(\hat{\theta}_n)}{\varepsilon^2}.$$

Puisque $\lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_n) = 0$, cela implique que pour tout $\varepsilon > 0$, il existe un entier N tel que pour tout $n \geq N$, $\mathrm{Var}(\hat{\theta}_n) < \varepsilon^2$.

Ainsi, nous avons :

$$\Pr(|\hat{\theta}_n - \theta| > \varepsilon) \leq \frac{\mathrm{Var}(\hat{\theta}_n)}{\varepsilon^2} < \frac{\varepsilon^2}{\varepsilon^2} = 1.$$

Cela signifie que la probabilité que $|\hat{\theta}_n - \theta| > \epsilon$ est inférieure à 1 pour tout $n \geq N$.

Par conséquent, nous avons montré que $\hat{\theta}_n$ converge en probabilité vers θ lorsque $n \to \infty$.

□

Quelques exemples d'estimateurs

• Estimateur de la moyenne

L'estimateur $\hat{\mu}_n$ est égale à :

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Propriétés 4.

Pour un échantillon aléatoire dont la loi de X admet une espérance notée μ*, et une variance notée* σ^2*,* $\hat{\mu}_n$ *est un estimateur sans biais de la moyenne* μ*, i.e* $\mathbb{E}(\hat{\mu}_n) = \mu$ *et la variance de* $\hat{\mu}_n$ *est égale à* $Var(\hat{\mu}_n) = \sigma^2/n$*. De plus* $\hat{\mu}_n$ *est un estimateur convergent de la moyenne* μ*.*

• Estimateur de la variance

l'estimateur S_n^2 est égale à :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n))^2.$$

Propriétés 5.

Pour un échantillon aléatoire dont la loi de X admet une espérance notée μ *et une variance notée* σ^2*,* S_n^2 *est un estimateur biaisé de la variance et le biais* $\mathrm{b}(S_n^2) = \sigma^2/n$*.*

S_n^2 *est donc un estimateur asymptotiquement sans biais .*

•Estimateur corrigé de la variance

L'estimateur corrigé de la variance, noté $S_{n.c}^2$ est égale à :

$$S_{n.c}^2 = \frac{nS_n^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2.$$

Propriétés 6.

Pour un échantillon aléatoire dont la loi de X admet une espérance notée μ et une variance notée σ^2, $S_{n.c}^2$ est un estimateur sans biais de la variance σ^2.

Définition 6.

La fonction quantile d'ordre α, $\alpha \in [0,1]$, est liée à la fonction de répartition inverse généralisée (ou pseudo-inverse) de F notée F^{-1} définie par :

$$q_\alpha = inf_{t\in\mathbb{R}}\left\{t : F(t) > \alpha\right\}.$$

Si F est continue et strictement croissante, alors :

$$q_\alpha = F^{-1}(\alpha).$$

La **médiane** d'une variable aléatoire X est le quantile d'ordre $\frac{1}{2}$. Elle est notée et définie par :

$$M_e = q_{0.5}.$$

La **vraisemblance** (likelihood en anglais) de l'échatillon $(X_1, \cdots\cdots, X_n)$ est la loi de probabilité de ce n-uplet, notée $L(x_1, \cdots\cdots, x_n; \theta)$ et définie par :

$$L(x_1, \cdots\cdots, x_n; \theta) = \prod_{i=1}^{n}\mathbb{P}(X_i = x_i|\theta) \qquad \text{dans le cas discret}$$

$$L(x_1, \cdots\cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \qquad \text{dans le cas continu.}$$

Sous certaines conditions, la quantité d'information de Fisher sur θ fournie par l'échantillon est le réel positif noté $I_n(\theta)$, définie par :

$$I_n(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial lnL(X_1, \cdots\cdots, X_n)}{\partial\theta}\right)^2\right].$$

On appelle score de l'echatillon, noté par $S_n(\theta)$, la dérivée de la log-vraisemblance définie par :

$$S_n(\theta) = \frac{1}{L(X_1, \cdots\cdots, X_n)}\frac{\partial}{\partial\theta}L(X_1, \cdots\cdots, X_n).$$

En posant $l(x,\theta) = lnL(x,\theta)$, on a :

$$S_n(\theta) = \frac{\partial}{\partial\theta}l(X_1, \cdots\cdots, X_n).$$

propriété de l'information de Fisher) (cf [18])

▷ Si le domaine de définition de $f(x,\theta)$ est indépendant de θ alors on a :

$$I_n(\theta) = \mathbb{E}\left(-\frac{\partial}{\partial\theta}lnL(X_1, \cdots\cdots, X_n)\right).$$

▷ Si le domaine de définition de $f(x,\theta)$ est indépendant de θ , chaque observation apporte la même information

$$I_n(\theta) = nI_1(\theta).$$

Ainsi pour les cas p-dimensionnel (i.e $\theta = (\theta_1, \cdots\cdots, \theta_n)$), le score est un vecteur défini par :

$$S_n(\theta) = \frac{\partial}{\partial\theta_1} l\left(X_1, \cdots\cdots, X_n; \theta_1\right), \cdots\cdots, \frac{\partial}{\partial\theta_n} l\left(X_1, \cdots\cdots, X_n; \theta_n\right).$$

Définition 7.

Soit $(X_1, ..., X_n)$ un échantillon extrait d'une variable aléatoire dont la loi dépend d'un paramètre θ inconnu et $\alpha \in]0, 1[$ fixé. On appelle ***intervalle de confiance*** *de paramètre θ, de niveau de confiance $1 - \alpha$, tout intervalle de la forme $[a_n, b_n]$, avec a_n, b_n deux statistiques dépendantes de $(X_1, ..., X_n)$ tel que :*

$$\mathbb{P}(a_n \leq \theta \leq b_n) = 1 - \alpha.$$

Chapitre 2

Régression Linéaire Simple, Comparaison des Méthodes d'estimation

La régression linéaire simple peut être considérée comme une technique statistique permettant de modéliser la relation linéaire entre une variable explicative (notée X) et une variable à expliquer (notée Y). Cette présentation va nous permettre d'exposer la régression linéaire dans un cas simple afin de bien comprendre les enjeux de cette méthode, les problèmes posés et les réponses apportées.

2.1 Modèle et Hypothèses

Définition 8.
*Un modèle de régression linéaire simple, noté **rls** est défini par :*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \forall i \in \{1, \cdots\cdots, n\}, \tag{2.1}$$

où β_0 et β_1 sont les paramètres réels inconnus (les coefficients du modèle, à savoir l'ordonnée à l'origine et la pente) et ε_i est une erreur aléatoire.

Remarque.
• *Le coefficient β_0 est aussi appelé intercept ou constante. C'est la valeur prédite de y lorsque $x = 0$ et le coefficient β_1, la pente est le changement sur y lorsque x change d'une unité.*
• *L'erreur aléatoire ε tient un rôle très important en régression, Il nous permet de résumer toute information non prise en compte dans la relation linéaire que nous cherchons à établir entre Y et X (i.e résumer le rôle des variables explicatives absentes).*

Avant de procéder à l'étude de ce modèle, il est important de donner les différentes hypothèses qui rendent son étude possible. Elles sont les suivantes :
H1 Le modèle est linéaire en x_i (ou en n'importe quelle transformation de x_i) et les valeurs de celui-ci sont observées sans erreurs (x_i non aléatoires).
H2 Les erreurs ε_i sont centrées (i.e en moyenne le modèle est bien spécifié et donc l'erreur moyenne est nulle) et de variance constante (homoscédasticité) :

$$\mathbb{E}(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma_\varepsilon^2, \forall i \in \{1, \cdots\cdots, n\}.$$

H3 Les erreurs relatives à deux observations sont indépendantes :

$$Cov(\varepsilon_i, \varepsilon_j) = 0.$$

H4 Les erreurs ε_i sont indépendantes et identiquement distribuées (i.i.d) suivent une loi normale de moyenne nulle est de variance $\sigma_\varepsilon^2 : \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

2.2 Estimation des paramètres par la Méthode des Moindres Carrés Ordinaires (MCO)

Les paramètres β_0 et β_1

On suppose n couples d'observations $(x_i, y_i), i = 1, \cdots\cdots, n$ ont été réalisées, en substituant dans le modèle linéaire, on obtient :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Rightarrow \varepsilon_i = y_i - \beta_0 - \beta_1 x_i \, .$$

Ces coefficients sont déterminer par la méthode des moindres carrés qui minimise les carrés des erreurs :

$$d(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \|Y - \beta_0 - \beta_1 X\|^2;$$

$$\nabla d(\beta_0, \beta_1) = 0 \Rightarrow \begin{cases} \hat{\beta}_0 & = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 & = \dfrac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2} = \dfrac{S_{xy}}{S_{xx}} \end{cases}$$

avec

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{la moyenne empirique des } x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{la moyenne empirique des } y_i$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \qquad \text{la variance empirique des } x_i$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \qquad \text{la variance empirique des } y_i$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \qquad \text{la variance empirique entre les } x_i \text{ et les } y_i.$$

On voit que c'est un minimum global de $\mathbb{R}^2$ en (β_0, β_1). C'est pourquoi cette méthode s'appelle la méthode des moindres carrés ordinaires (MCO).(Voir [5],[26],[32]).

Preuve.

Les estimateurs peuvent être écrits sous la forme :

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0 \times \beta_1) \in \mathbb{R}\times\mathbb{R}}{\operatorname{argmin}} \; d(\beta_0, \beta_1).$$

De plus la fonction $d(\beta_0, \beta_1)$ est strictement convexe. Si elle admet un point singulier, celui-ci correspond au minimum. En annulant les dérivées partielles, nous obtenons les deux équations ci-dessous appelées équations normales

$$\begin{aligned} \frac{\partial d(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta_0}} = 0 &\Leftrightarrow -2 \sum_{1=n}^{n} \left(y_i - \hat{\beta}_1 x_i - \hat{\beta_0}\right) = 0 \\ &\Leftrightarrow -2 \left[\sum_{1=n}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i - n\hat{\beta_0}\right] = 0. \end{aligned} \tag{2.2}$$

$$\begin{aligned} \frac{\partial d(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta_1}} = 0 &\Leftrightarrow -2 \sum_{1=n}^{n} x_i \left(y_i - \hat{\beta}_1 x_i - \hat{\beta_0}\right) = 0 \\ &\Leftrightarrow -2 \left[\sum_{1=n}^{n} x_i y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \hat{\beta_0} \sum_{i=1}^{n} x_i\right] = 0. \end{aligned} \tag{2.3}$$

Ainsi, si on multiplie l'équation (2.2) par $\frac{1}{n}$, on obtient :

$$\frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1n}^{n} x_i - \hat{\beta}_0 = 0.$$

Alors

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{2.4}$$

En remplaçant l'expression de (2.4), on a :

$$-2 \left(\sum_{i=1}^{n} x_i y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \bar{y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i\right) = 0. \tag{2.5}$$

En multipliant (2.5), par $\frac{n}{n}$ on obtient :

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) - n\bar{x}\bar{y} = 0.$$

D'où

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}.$$

□

Droite de régression

Pour n points ou couples de variables, $(x_i, y_i), i = 1, \cdots\cdots, n$ de $\mathbb{R}^2$, on essaie de trouver l'équation d'une droite qui passe par les n points. Cette équation est $Y = \beta_0 + \beta_1 X$ avec β_0 et β_1 les solutions du système linéaire $X\beta = Y$ où :

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

et au sens des moindres carrés on a :

$$(\hat{\beta}_0, \hat{\beta}_1) = \left(X^t X\right)^{-1} X^t Y.$$

Ainsi, la droite de régression estimée est $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

Remarque.

1. *Le résidus de la régression $\hat{\varepsilon}$ est donné par :*

$$\hat{\varepsilon}_i = y_i - \hat{y}_i.$$

2. *La somme moyenne des résidus est nulle dans une régression avec constante :*

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0.$$

En effet,

$$\begin{aligned}\sum_{i=1}^{n} \hat{\varepsilon}_i &= \sum_{i=1}^{n} y_i - \hat{y}_i \\ &= \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right] \\ &= n\bar{y} - n\hat{\beta}_0 + n\hat{\beta}_1\bar{x} \\ &= n\bar{y} - n(\bar{y} - \hat{\beta}_1\bar{x} - n\hat{\beta}_1\bar{x} = 0.\end{aligned}$$

3. *Da droite de régression avec constante passe forcément par le centre de gravité du nuage des points $(\bar{x}, \bar{y})$. Pour le montrer, il suffit simplement de réaliser la projection pour le point $\bar{x}$:*

$$\begin{aligned}\hat{y}(\bar{x}) &= \hat{\beta_0} + \hat{\beta_1} x_i \\ &= (\bar{y} - \hat{\beta_1}\bar{x}) - \hat{\beta_1} x_i = \bar{y}.\end{aligned}$$

4. *L'estimateur de la variance de l'erreur, note $\hat{\sigma}_\varepsilon^2$ est donné comme suit :*

$$\begin{aligned}\hat{\sigma}_\varepsilon^2 &= \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i}{n-2} \\ &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \frac{SC_{res}}{n-2},\end{aligned}$$

où SC_{res} est la somme carré des résidus que nous verrons un peu plu tard.
De plus $(n-2)$ peut s'expliquer par la règle : nombre d'observations n moins le nombre de paramètre du modèle à estimer.

Propriétés des estimateurs

Sous les hypothèses H_1 et H_2, on peut donner certaines propriétés des estimateurs moindres carrés de $\hat{\beta_0}$ et $\hat{\beta_1}$ (pour plus de détails des calcul ci-dessous voir [5],[26]).

Théorème 7.

1. $\mathbb{E}(\hat{\beta}_0) = \beta_0$ *et* $\mathbb{E}(\hat{\beta}_1) = \beta_1$ *(non biaisés).*

2. $Var(\hat{\beta}_0) = \sigma_\varepsilon^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$, *et* $Var(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{S_{xx}}$.

3. $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma_\varepsilon^2\bar{x}}{S_{xx}}$.

Preuve.

1. Montrons que les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont non biaisés.
   Considérons notre modèle de rls suivant :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{2.6}$$

   Il nous est possible de calculer :

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}(n\beta_0) + \beta_1\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i;$$

$$\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{\varepsilon}. \tag{2.7}$$

   En faisant la différence entre (2.6) et (2.7), on obtient :

$$y_i - \bar{y} = \beta_1(x_i - \bar{x}) + (\epsilon_i - \bar{\varepsilon}).$$

   Et comme nous avons déjà :

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

   par suite on a :

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})\left[\beta_1(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})\right]}{\sum_{i=1}^{n}(x_i - \bar{x}^2)^2} \\ &= \frac{\beta_1\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n}(x_i - \bar{x}^2)^2} \\ &= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n}(x_i - \bar{x}^2)^2}\end{aligned}$$

   et comme $\bar{\varepsilon}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, alors il reste :

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

En faisant intervenir le calcul de l'espérance on obtient :

$$\begin{aligned}\mathbb{E}\left(\hat{\beta}_1\right) &= \mathbb{E}\left(\beta_1\right) + \mathbb{E}\left[\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\varepsilon_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}^2\right)^2}\right] \\ &= \beta_1 + \mathbb{E}\left[\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\varepsilon_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}^2\right)^2}\right] \\ &= \beta_1 + \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}^2\right)^2} \times \mathbb{E}(\varepsilon_i).\end{aligned}$$

Et d'après l'hypothèse que $\mathbb{E}(\varepsilon_i) = 0$, on obtient le résultat : $\mathbb{E}(\hat{\beta}_1) = \beta_1$.
Pour ce qui est de β_0, on part de l'expression $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$.
En utilisant toujours le calcul de l'espérance, on a :

$$\mathbb{E}(\beta_0) = \mathbb{E}(\bar{y}) - \bar{x}\mathbb{E}(\hat{\beta}_1) = \mathbb{E}\left(\beta_0 + \beta_1\bar{x} + \bar{\varepsilon}\right) - \bar{x}\mathbb{E}(\hat{\beta}_1) = \beta_0 + \bar{\varepsilon} + \bar{x}\left(\beta_1 - \beta_1\right).$$

D'où

$$\mathbb{E}(\hat{\beta}_0) = \beta_0.$$

2. Les variances des estimateurs
- Montrons que $Var(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{S_{xx}} = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}^2\right)^2}$.

On a

$$\begin{aligned}Var(\hat{\beta}_1) &= Var\left(\beta_1 + \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\varepsilon_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}^2\right)^2}\right) \\ &= Var\left(\beta_1\right) + Var\left(\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\varepsilon_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}^2\right)^2}\right), \text{réécriture} \\ &= Var\left(\beta_1\right) + Var\left(\sum_{i=1}^{n}\frac{\left(x_i-\bar{x}\right)\varepsilon_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}^2\right)^2}\right), \text{posons } a_i = \frac{\left(x_i-\bar{x}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}^2\right)^2} \\ &= Var\left(\sum_{I=1}^{n} a_i\varepsilon_i\right), \text{vue que } \beta_1 \text{ est une constante} \\ &= \sum_{i=1}^{n} a_i^2 Var\left(\varepsilon_i\right) + 2\sum_{1\le i\le j\le n} a_i a_j Cov(\varepsilon_i,\varepsilon_j), \text{où } Cov(\varepsilon_i,\varepsilon_j) = 0\end{aligned}$$

et alors

$$Var(\hat{\beta}_1) = \sigma_\varepsilon^2\sum_{i)1}^{n} a_i^2, \text{avec } Var(\varepsilon_i) = \sigma_\varepsilon^2.$$

En remplaçant a_i par son expression, on on :

$$\begin{aligned} Var(\hat{\beta}_1) &= \sigma_\varepsilon^2 \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x}^2)^2} \right)^2 \\ &= \sigma_\varepsilon^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} \\ &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sigma_\varepsilon^2}{S_{xx}}. \end{aligned}$$

-Montrons que $Var(\hat{\beta}_0) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$.

On a :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

alors

$$\begin{aligned} Var(\hat{\beta}_0) &= Var\left(\bar{y} - \hat{\beta}_1 \bar{x} \right) \\ &= Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x} Cov(\bar{y}, \hat{\beta}_1); \end{aligned}$$

où $Cov(\bar{y}, \hat{\beta}_1) = 0$
donc

$$\begin{aligned} Var(\hat{\beta}_0) &= Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) \\ &= Var\left(\frac{1}{n} \sum_{i=1}^{n} y_i \right) + \bar{x}^2 Var(\hat{\beta}_1) \\ &= \frac{\sigma_\varepsilon^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ &= \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]. \end{aligned}$$

3. -Montrons que $Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\sigma_\varepsilon^2 \bar{x}}{S_{xx}} = -\frac{\sigma_\varepsilon^2 \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$.

   D'après l'équation (2.7), on a :

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon} \tag{2.8}$$

et

$$\hat{\beta}_0 = \bar{y} - \bar{x} \hat{\beta}_1. \tag{2.9}$$

En substituant (2.8) dans l'équation (2.9), on obtient

$$\begin{aligned}\hat{\beta}_0 &= \beta_0 + \bar{x}\beta_1 + \bar{\varepsilon} - \bar{x}\hat{\beta}_1 \\ &= \beta_0 + \bar{x}\left(\beta_1 - \hat{\beta}_1\right) + \bar{\varepsilon},\end{aligned}$$

donc

$$\hat{\beta}_0 - \beta_0 = -\bar{x}\left(\hat{\beta}_1 - \beta_1\right) + \bar{\varepsilon}.$$

D'autre part

$$\begin{aligned}Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) &= \mathbb{E}\left[\left(\hat{\beta}_0 - \mathbb{E}(\hat{\beta}_0)\right)\left(\hat{\beta}_1 - \mathbb{E}(\hat{\beta}_1)\right)\right] \\ &= \mathbb{E}\left[\left(-\bar{x}\left(\hat{\beta}_1 - \beta_1\right) + \bar{\varepsilon}\right)\left(\hat{\beta}_1 - \mathbb{E}(\hat{\beta}_1)\right)\right] \\ &= \mathbb{E}\left[-\bar{x}\left(\hat{\beta}_1 - \beta_1\right)^2 + \bar{\varepsilon}\left(\hat{\beta}_1 - \mathbb{E}(\hat{\beta}_1)\right)\right] \\ &= -\bar{x}\mathbb{E}\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right] + \bar{\varepsilon}\mathbb{E}\left(\hat{\beta}_1 - \mathbb{E}(\hat{\beta}_1)\right), \quad \text{car } \bar{\varepsilon} = 0 \\ &= -\bar{x}\mathbb{E}\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right] = \sigma_\varepsilon^2 Var(\hat{\beta}_1).\end{aligned}$$

D'où

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\bar{x}\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})} = -\frac{\bar{x}\sigma_\varepsilon^2}{S_{xx}}.$$

□

2.2.1 Lois des estimateurs - intervalles et régions de confiance

Jusqu'ici, nous avons pu calculer les estimateurs $\hat{\beta_0}$ et $\hat{\beta_1}$ de β_0 et β_1, mais aussi déterminer leur qualité (non biaisé).
Cependant, l'étude ne se limite pas qu'à ça, nous souhaitons en général connaître la loi de ces estimateurs afin de calculer les intervalles ou régions de confiance et les tests d'hypothèses, notamment les tests de significativité que nous verrons un peu plu tard (voir [5]).

Proposition 2. *(Lois des estimateurs avec variance connue)*

(i) $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim \mathcal{N}(\beta, \Gamma)$ *où* $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ *et*

$$\begin{aligned}\Gamma &= \sigma_\varepsilon^2 \begin{bmatrix} \sum_{I=1}^{n} x_i^2 / \left(n \sum_{I=1}^{n}(x_i - \bar{x})^2\right) & -\bar{x} / \sum_{I=1}^{n}(x_i - \bar{x})^2 \\ -\bar{x} / \sum_{I=1}^{n}(x_i - \bar{x})^2 & 1 / \sum_{I=1}^{n}(x_i - \bar{x})^2 \end{bmatrix} \\ &= \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{bmatrix}.\end{aligned}$$

(ii) $\dfrac{(n-2)}{\sigma_\varepsilon^2}\hat{\sigma}_\varepsilon^2 \sim \chi^2_{n-2}.$

(ii) $\hat{\beta}$ *et* $\hat{\sigma}_\varepsilon^2$ *sont indépendantes .*

Pour la preuve et plus de détail (voir [4],[1]).

Remarque. *Le problème des propriétés ci-dessus vient de ce qu'elles font intervenir la variance théorique de* σ_ε^2*, généralement inconnue. La façon naturelle de procéder est de la remplacer par son estimateur* $\hat{\sigma}_\varepsilon^2$*.*

Proposition 3. *(lois des estimateurs avec variance inconnue)*
Les lois des estimateurs MCO avec variance σ_ε^2inconnue sont :

(i) $\dfrac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_\varepsilon(\hat{\beta}_0)}$ *et* $\dfrac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_\varepsilon(\hat{\beta}_1)}$ *suivent une loi de Student à* $(n-2)$ *degré de liberté.*

(ii) $\dfrac{1}{2\hat{\sigma}_\varepsilon^2}\left(\hat{\beta}_0 - \beta_0\right)^t \Gamma^{-1}\left(\hat{\beta}_1 - \beta_1\right) \sim \mathcal{F}_{(2,n-2)}$ *où* $\mathcal{F}_{(2,n-2)}$ *est la loi de Fisher à* 2 *ddl au numérateur et* $(n-2)$ *ddl au dénominateur.*

(Pour la preuve cf proposition 8 chapitre 3 et [9],[4].)

Ces dernières propriétés nous permettent de donner des intervalles de confiance **(I.C)** ou des régions de confiance **(R.C)** des estimateurs. En effet, la valeur ponctuelle d'un estimateur est de peu d'intérêt en général et il lui est intéressant de lui associer un intervalle de confiance.

Proposition 4. *(intervalles et régions de confiance)*
Soit un seuil α petit où $\alpha \in [0,1]$, on a :

1. *L'intervalle de confiance pour β_0 , noté par IC(β_0), est :*
$$\left[\hat{\beta}_0 - t_{n-2}(1-\alpha/2)\hat{\sigma}_\varepsilon(\hat{\beta}_0), \hat{\beta}_0 + t_{n-2}(1-\alpha/2)\hat{\sigma}_\varepsilon(\hat{\beta}_0)\right],$$
où $t_{n-2}(1-\alpha/2)$ le quantile de niveau $(1-\alpha/2)$ d'une loi de Student $\mathcal{T}_{n-2}$.

2. *L'intervalle de confiance pour β_1 , noté par IC(β_1), est :*
$$\left[\hat{\beta}_1 - t_{n-2}(1-\alpha/2)\hat{\sigma}_\varepsilon(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2}(1-\alpha/2)\hat{\sigma}_\varepsilon(\hat{\beta}_1)\right].$$

3. *Une région de confiance simultanée pour β_0 et β_1, notée RC(β), au niveau $(1-\alpha)$ est :*
$$\frac{1}{2\hat{\sigma}_\varepsilon^2}\left[(\hat{\beta}_0 - \beta_0) - 2n\bar{x}(\hat{\beta}_0 - \beta_0) + \sum_{I=1}^{n} x_i^2(\hat{\beta}_1 - \beta_1)^2\right] \leq \mathcal{F}_{(2,n-2)}(1-\alpha),$$
où $\mathcal{F}_{(2,n-2)}(1-\alpha)$ le quantile de niveau $(1-\alpha)$ de la loi $\mathcal{F}_{(2,n-2)}$.

4. *Intervalle de confiance pour σ_ε^2 est donné par :*
$$\left[\frac{(n-2)\hat{\sigma}_\varepsilon^2}{c_{n-2}(1-\alpha/2)}, \frac{(n-2)\hat{\sigma}_\varepsilon^2}{c_{n-2}(\alpha/2)}\right],$$
où $c_{n-2}(1-\alpha/2)$ est le fractile ou quantile de niveau $(1-\alpha/2)$ d'une loi de χ^2_{n-2} à $(n-2)$ degré de liberté.

Pour la preuve et plus de détail, voir [4],[9].

2.2.2 Analyse de la Variance, Coefficient de Détermination et de Corrélation

a) **Décomposition de la variance et Tableau d'ANOVA**

En un point d'observation (x_i, y_i) on décompose l'écart entre y_i et la moyenne des y_i en y ajoutant et retranchant $\hat{y}$ la valeur estimée par la droite de régression.

Cette procédure nous donne :

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

À présent, on élève les deux membres au carré, puis on somme sur les observations i, on obtient :

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}\left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right]^2 = \sum_{i=1}^{n}\left[(\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i\right]^2 \\ &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})\hat{\varepsilon}_i + \sum_{i=1}^{n}\hat{\varepsilon}_i^2.\end{aligned}$$

Or, vu que $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$, alors

$$\begin{aligned}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})\hat{\varepsilon}_i &= \sum_{i=1}^{n}(\hat{\beta_1}x_i + \hat{\beta_0} - \bar{y})\hat{\varepsilon}_i \\ &= \hat{\beta_1}\sum_{i=1}^{n} x_i\hat{\varepsilon}_i + \hat{\beta_0}\sum_{i=1}^{n}\hat{\varepsilon}_i - \bar{y}\sum_{i=1}^{n}\hat{\varepsilon}_i \\ &= 0.\end{aligned}$$

Ce qui aboutit à l'égalité fondamentale (équation d'analyse de la variance) :

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SC_{total}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SC_{reg}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SC_{res}}. \tag{2.10}$$

Alors

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SC_{total}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SC_{reg}} + \underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}_{SC_{res}}. \tag{2.11}$$

Où

- SC_{total} : est la somme des carrés totaux indiquant la variabilité de Y (i.e l'information disponible dans la base de données).
- SC_{reg} : est la somme des carrés expliquées , elle est la variabilité expliquée par le modèle (i.e la variation de Y expliquée par X).
- SC_{res} : est la somme des carrés résiduels. Elle décrit la variabilité non expliquée par le modèle.

Remarque. *Deux situations extrêmes peuvent survenir :*

∗ *Dans le meilleur des cas, $SC_{res} = 0$ et donc $SC_{total} = SC_{reg}$: les variations de Y sont expliquées par celle de X. Ainsi, nous obtenons un modèle parfait, par suite la droite de régression passe exactement par tous les points du nuage $(\hat{y}_i - y_i)$.*

∗ *Dans le pire des cas, $SC_{reg} = 0$: X n'apporte aucune information sur Y (i.e X n'influe en aucune manière sur Y) .*

À présent, que nous avons décomposé la variance et établit l'équation d'analyse de la variance, il nous est possible d'établir le tableau d'analyse de la variance ci-dessous :

Source de variation	*Somme des carrés*	*Degrés de liberté*	*Moyenne des carrés*
Expliquée par la régression	SC_{reg}	1	$MC_{reg} = \frac{SC_{reg}}{1}$
Résidus	SC_{res}	$n-2$	$MC_{res} = \frac{SC_{res}}{n-2}$
Total	$SC_{total} = S_{yy}$	$n-1$	

TABLE 2.1 – Tableau d'analyse de la variance pour la régression linéaire simple

b) **Coefficient de détermination**

Le coefficient de détermination R^2 est définie par :

$$R^2 = \frac{SC_{reg}}{SC_{total}} = 1 - \frac{SC_{res}}{SC_{total}}.$$

Il mesure le pourcentage de la variabilité totale SC_{total} qui est expliqué par le modèle.

Remarque.

- *Ce coefficient de détermination R^2 varie entre 0 et 1 ($R^2 \in [0,1]$).*
- *Plus R^2 se rapproche de la valeur 1, meilleur est l'adéquation du modèle sur la base de données et un R^2 faible (proche de 0) signifie que le modèle est inadéquat (i.e un faible pouvoir explicatif).*

c) **Coefficient de corrélation**

Pour rappel ; la corrélation entre deux variables aléatoires X et Y est mesurée par le coefficient

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Définition 9.

Le coefficient de corrélation d'un échantillon est

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

Pour ce qui est de notre exemple, le coefficient de corrélation ρ est estimé ponctuellement par r.

Interprétation du coefficient de corrélation

On peut montrer que $-1 \leq r \leq 1$.

$*$ Si $r = -1$ ou $r = 1$ alors il y a corrélation parfaite entre X et Y et les points (x_i, y_i) sont tous sur la droite de régression.

$*$ Si $r = 0$ alors il n'y a pas de corrélation entre X et Y et les points (x_i, y_i) sont dispersés au hasard.

$*$ Si $0 < r < 1$ alors il y a corrélation positive faible, moyenne ou forte entre X et Y. Dans ce cas, une augmentation de X entraîne une augmentation de Y.

$*$ Si $-1 < r < 0$ alors il y a corrélation négative faible, moyenne ou forte entre X et Y. Dans ce cas, une augmentation de X entraîne une diminution de Y.

Remarque.

- *De plus, il est important de noter que dans le cas d'une régression linéaire, $R^2 = r^2$.*

2.2.3 Tests de significativité du modèle

a) **Test de significativité globale du modèle**

Ce test nous permet de connaître l'apport global de la variable X à la détermination de Y.

On veut tester :

$$\begin{cases} H_0 : \beta_1 & = 0 \\ H_1 : \beta_1 & \neq 0. \end{cases}$$

Le test de cette hypothèse est basé sur la statistique de Fisher, notée par F :

$$F = \frac{\frac{SC_{reg}}{1}}{\frac{SC_{res}}{n-2}} = \frac{MC_{reg}}{MC_{res}}. \tag{2.12}$$

Cette statistique nous indique que si la variance explicative est significativement supérieure à la variance résiduelle, dans ce cas, on peut considérer que l'explication donnée par la régression traduit une relation qui existe réellement dans la population.

Sous H_0, SC_{reg} est distribuée selon un loi de $\chi^2(1)$ et SC_{res} selon une loi de $\chi^2(n-2)$, de fait pour F, on :

$$F = \frac{\frac{\chi^2(1)}{1}}{\frac{\chi^2(n-2)}{n-2}} = f_{(1,n-2)}(1-\alpha). \tag{2.13}$$

Alors, Sous H_0, F est donc distribuée selon une loi de Fisher à $(1, n-2)$ ddl, où on rejette H_0, si :

$$F \geq f_{(1,n-2)}(1-\alpha),$$

avec $f_{(1,n-2)}(1-\alpha)$ le quantile d'ordre $1-\alpha$ d'une loi de Fisher à $(1, n-2)$ ddl.

Remarque. *(i) On peut réécrire la statistique F en fonction de R^2 comme suit :*

$$F = \frac{\frac{R^2}{1}}{\frac{1-R^2}{n-2}} = (n-2)\frac{R^2}{1-R^2}.$$

(ii) Dans la plupart des logiciels statistiques, on fournit directement la probabilité critique (p-value). Elle correspond à la probabilité que la loi de Fisher dépasse la statistique calculée F. Ainsi, la règle de décision (rejeter H_0) au risque α devient :

$$\textit{p-value} \leq \alpha.$$

b) **Test de significativité des paramètres**

Pour le paramètre β_0, l'ordonnée à l'origine

On veut tester l'hypothèse :

$$\begin{cases} H_0 : \beta_0 & = 0 \\ H_1 : \beta_0 & \neq 0. \end{cases}$$

La statistique de test permettant d'effectuer ce test est :

$$T_{\hat{\beta}_0} = \frac{\hat{\beta}_0}{\hat{\sigma}_\varepsilon(\hat{\beta}_0)}.$$

On rejette H_0 au seuil $(1-\alpha)$ si $|T_{\hat{\beta}_0}| \geq t_{n-2}(1-\alpha/2)$ où $t_{n-2}(1-\alpha/2)$ est le quantile d'ordre $(1-\alpha/2)$ de $\mathcal{T}_{n-2}$.

Pour le paramètre β_1, la pente

On veut tester l'hypothèse :

$$\begin{cases} H_0 : \beta_1 & = 0 \\ H_1 : \beta_1 & \neq 0. \end{cases}$$

La statistique de test permettant d'effectuer ce test est :

$$T_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{\hat{\sigma}_\varepsilon(\hat{\beta}_1)}.$$

On rejette H_0 au seuil $(1-\alpha)$ si $|T_{\hat{\beta}_1}| \geq t_{n-2}(1-\alpha/2)$ où $t_{n-2}(1-\alpha/2)$ est le quantile d'ordre $(1-\alpha/2)$ de $\mathcal{T}_{n-2}$.

Remarque. *Si le nombre d'observation $n \geq 30$, la loi de Student tend vers la loi Normale.*

2.2.4 Prévision et Intervalle de prédiction

Un des buts de la régression est de proposer des prévisions pour la variable à expliquer Y. Soit x_{n+1} une nouvelle valeur de la variable explicative X, nous voulons prédire y_{n+1}. Le modèle est toujours le même :

$$y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1},$$

avec $\mathbb{E}(\varepsilon_{n+1}) = 0$, $Var(\varepsilon_{n+1}) = \sigma_\varepsilon^2$ et $Cov(\varepsilon_{n+1}, \varepsilon_i) = 0$, $\forall i \in \{1, \cdots\cdots, n\}$. Il est naturel de prédire la valeur correspondante grâce au modèle ajusté :

$$\hat{y}_{n+1}^p = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}.$$

Nous utilisons la notation $\hat{y}_{n+1}^p$ afin d'insister sur la notion de prévision.

Proposition 5. *(Variance de la prévision $\hat{y}_{n+1}^p$)*
La variance de la valeur prévue de $\hat{y}_{n+1}^p$ vaut :

$$Var(\hat{y}_{n+1}^p) = \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right).$$

Elle nous donne une idée de la stabilité de l'estimation.

Preuve :
Considérons la prédiction $\hat{y}_{n+1}^p$:

$$\begin{aligned} \hat{y}_{n+1}^p &= \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \\ &= \beta_0 + \beta_1 x_{n+1} + (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) x_{n+1} \\ &= \beta_0 + \beta_1 x_{n+1} + (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) x_{n+1} + 0 \cdot \varepsilon_{n+1}, \end{aligned}$$

où ε_{n+1} est l'erreur associée à la prédiction $\hat{y}_{n+1}^p$.

En utilisant les propriétés de l'espérance et de la variance, nous pouvons calculer la variance de $\hat{y}_{n+1}^p$:

$$\begin{aligned} \mathrm{Var}(\hat{y}_{n+1}^p) &= \mathrm{Var}(\hat{\beta}_0) + \mathrm{Var}(\hat{\beta}_1 x_{n+1}) + 2\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1 x_{n+1}) + \mathrm{Var}(0 \cdot \varepsilon_{n+1}) \\ &= \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) + x_{n+1}^2 \frac{\sigma_\varepsilon^2}{S_{xx}} - 2x_{n+1}\bar{x} \frac{\sigma_\varepsilon^2}{S_{xx}} \\ &= \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right). \end{aligned}$$

Ainsi, la variance de la prédiction $\hat{y}_{n+1}$ est égale à $\sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right)$.

Proposition 6. *(Erreur de prévision)*
L'erreur de prévision, définie par $\hat{\varepsilon}_{n+1}^p = y_{n+1} - \hat{y}_{n+1}^p$, satisfait les propriétés suivantes :

$$\mathbb{E}(\hat{\varepsilon}_{n+1}^p) = 0 \; et \; Var(\hat{\varepsilon}_{n+1}^p)) = \sigma_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right).$$

En d'autre terme, $\hat{\varepsilon}_{n+1}^p \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right) \right)$.

Preuve.
Considérons l'erreur de prévision $\hat{\varepsilon}_{n+1}^p$ définie par $\hat{\varepsilon}_{n+1}^p = y_{n+1} - \hat{y}_{n+1}^p$.

En utilisant la définition de $\hat{y}_{n+1}^p$, nous avons :

$$\hat{\varepsilon}_{n+1}^p = y_{n+1} - (\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}).$$

En utilisant les propriétés de l'espérance et de la variance, nous pouvons calculer l'espérance et la variance de

$\hat{\varepsilon}^p_{n+1}$:

$$\begin{aligned}\mathbb{E}(\hat{\varepsilon}^p_{n+1}) &= \mathbb{E}(y_{n+1}) - \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}) \\ &= \mathbb{E}(y_{n+1}) - \mathbb{E}(\hat{\beta}_0) - \mathbb{E}(\hat{\beta}_1 x_{n+1}) \\ &= \beta_0 + \beta_1 x_{n+1} - \beta_0 - \beta_1 x_{n+1} \\ &= 0.\end{aligned}$$

La variance de $\hat{\varepsilon}^p_{n+1}$ est donnée par :

$$\begin{aligned}\mathrm{Var}(\hat{\varepsilon}^p_{n+1}) &= \mathrm{Var}(y_{n+1}) + \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}) + 2\mathrm{Cov}(y_{n+1}, \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}) \\ &= \mathrm{Var}(y_{n+1}) + \mathrm{Var}(\hat{\beta}_0) + \mathrm{Var}(\hat{\beta}_1 x_{n+1}) + 2\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1 x_{n+1}) + 2\mathrm{Cov}(y_{n+1}, \hat{\beta}_0) + 2\mathrm{Cov}(y_{n+1}, \hat{\beta}_1 x_{n+1}) \\ &= \sigma^2_\varepsilon + \sigma^2_\varepsilon \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) + x^2_{n+1}\frac{\sigma^2_\varepsilon}{S_{xx}} + 2(-x_{n+1}\bar{x})\frac{\sigma^2_\varepsilon}{S_{xx}} + 2(0) + 2(0) \\ &= \sigma^2_\varepsilon \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}\right).\end{aligned}$$

Ainsi, on a $\mathbb{E}(\hat{\varepsilon}^p_{n+1}) = 0$ et $\mathrm{Var}(\hat{\varepsilon}^p_{n+1}) = \sigma^2_\varepsilon \left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{S_{xx}}\right)$.

□

Remarque. *La variance augmente lorsque x_{n+1} s'éloigne du centre de gravité du nuage de points. Effectuer une prévision lorsque x_{n+1} est « loin » de $\bar{x}$ est donc périlleuse, la variance de l'erreur de prévision peut être alors très grande .*

On ne connaît pas σ^2_ε, on l'estime donc par $\hat{\sigma}^2_\varepsilon$. De plus comme $(y_{n+1} - \hat{y}^p_{n+1})$ et $\frac{(n-2)}{\sigma^2_\varepsilon}\hat{\sigma}^2_\varepsilon$ sont indépendantes, on peut énoncer des résultats donnant des intervalles de confiance pour y_{n+1}.
Par suite, on a :

$$\frac{y_{n+1} - \hat{y}^p_{n+1}}{\hat{\sigma}_\varepsilon \sqrt{\left[\frac{1}{n} + \frac{(\bar{x} - x_{n+1})^2}{S_{xx}}\right]}} \text{ et } \frac{y_{n+1} - \hat{y}^p_{n+1}}{\hat{\sigma}_\varepsilon \sqrt{\left[1 + \frac{1}{n} + \frac{(\bar{x} - x_{n+1})^2}{S_{xx}}\right]}}$$

suivent une loi de Student à $(n-2)$ degrés de liberté. Ainsi, on obtient :
Intervalle d'estimation, au niveau $1-\alpha$, pour y_{n+1} :

$$\left[\hat{y}^p_{n+1} \pm t_{(n-2)}(1-\alpha/2)\hat{\sigma}_\varepsilon \sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_{n+1})^2}{S_{xx}}\right)}\right]. \tag{2.14}$$

Intervalle de prévision, au niveau $1-\alpha$, pour y_{n+1} :

$$\left[\hat{y}^p_{n+1} \pm t_{(n-2)}(1-\alpha/2)\hat{\sigma}_\varepsilon \sqrt{\left(1 + \frac{1}{n} + \frac{(\bar{x} - x_{n+1})^2}{S_{xx}}\right)}\right]. \tag{2.15}$$

2.3 Estimation des paramètres par Maximum de Vraisemblance (MV)

Il est vrai qu'en régression linéaire, la méthode d'estimation la plus utilisée est celle des moindres carrés ordinaires comme nous l'avons vue précédemment.
Toutefois, il existe d'autres méthodes pour estimer les paramètres des estimateurs linéaires en l'occurrence la méthode du maximum de vraisemblance (MV).
Considérons le modèle rls suivant :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \cdots\cdots, n,$$

avec y_i, linéairement dépendantes du terme d'erreur ε_i une variable aléatoire normalement distribuée de paramètre (moyenne et variance) : $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$ et $Var(y_i) = \sigma_\varepsilon^2$.
En effet, si $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2) \Rightarrow \mathbb{E}(\varepsilon_i) = 0$, alors $\mathbb{E}(y_i) = \mathbb{E}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i$.
Puisque $y_i - \mathbb{E}(y_i) = (\beta_0 + \beta_1 x_i + \varepsilon_i) - (\beta_0 + \beta_1 x_i) = \varepsilon_i$, alors la variance de y_i est :

$$Var(y_i) = \mathbb{E}\left[y_i - \mathbb{E}(y_i)\right]^2,$$

d'où la distribution de y_i : $y_i \sim \mathcal{N}\left([\beta_0 + \beta_1 x_i], \sigma_\varepsilon^2\right)$
- Fonction de densité de probabilité conjointe :
Elle s'écrit comme suit : $f(y_1, y_2, \cdots\cdots, y_n | \beta_0 + \beta_1 x_i, \sigma_\varepsilon^2)$. Si les y_i sont indépendantes, cette fonction peut s'écrire comme suit :

$$f(y_1, y_2, \cdots\cdots, y_n | \beta_0 + \beta_1 x_i, \sigma_\varepsilon^2) = f(y_1 | \beta_0 + \beta_1 x_i, \sigma_\varepsilon^2) \times f(y_2 | \beta_0 + \beta_1 x_i, \sigma_\varepsilon^2) \times \cdots \times f(y_n | \beta_0 + \beta_1 x_i, \sigma_\varepsilon^2).$$

- Fonction de densité de la loi normale générale
De manière générale, cette fonction se présente comme suit :

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\}.$$

À présent que nous avons la densité de la loi normale générale, nous pouvons calculer la vraisemblance et la log-vraisemblance de celle-ci.
- La vraisemblance :
Grâce à l'indépendance des erreurs, les observations sont indépendantes et la vraisemblance s'écrit :

$$\mathcal{L}(\beta_0, \beta_1, \sigma_\varepsilon^2) = \prod_{i=1}^{n} f(y_i) = \left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{n/2} exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}.$$

- La log-vraisemblance :
Lorsqu'on effectue une transformation logarithmique de la fonction de vraisemblance ci-dessus, l'on obtient la fonction dite log-vraisemblance qui servira de base à l'estimation des paramètres β_0 et β_1.
La fonction log-vraisemblance s'écrit :

$$\begin{aligned} log\left[\mathcal{L}(\beta_0, \beta_1, \sigma_\varepsilon^2)\right] &= log\left[\left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{n/2} exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}\right] \\ &= log\left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{n/2} + log\left[exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}\right] \\ &= -\frac{n}{2}log\sigma_\varepsilon^2 - \frac{n}{2}log(2\pi) - \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2. \end{aligned}$$

- Estimation des paramètres β_0,β_1 et σ_ε^2.
Pour estimer les paramètres β_0, β_1 et σ_ε^2 par le maximum de vraisemblance, la démarche consiste à maximiser la fonction log-vraisemblance ci-dessous, ce qui revient à annuler ses dérivées premières par rapport aux arguments

β_0, β_1 et σ_ε^2 comme suit :

$$\begin{cases} \dfrac{\partial}{\partial\beta_0}\left(log\left[\mathcal{L}(\beta_0,\beta_1,\sigma_\varepsilon^2)\right]\right) = -\displaystyle\sum_{I=1}^{n}\left(\frac{y_i-\beta_0-\beta_1x_i}{\sigma_\varepsilon^2}\right)(-1) = 0 & (2.16a) \\ \dfrac{\partial}{\partial\beta_1}\left(log\left[\mathcal{L}(\beta_0,\beta_1,\sigma_\varepsilon^2)\right]\right) = -\displaystyle\sum_{I=1}^{n}\left(\frac{y_i-\beta_0-\beta_1x_i}{\sigma_\varepsilon^2}\right)(-x_i) = 0 & (2.16b) \\ \dfrac{\partial}{\partial\sigma_\varepsilon^2}\left(log\left[\mathcal{L}(\beta_0,\beta_1,\sigma_\varepsilon^2)\right]\right) = \dfrac{\partial}{\partial\sigma_\varepsilon^2}\left(-\dfrac{n}{2}log\sigma_\varepsilon^2\right) - \dfrac{1}{2}\dfrac{\partial}{\partial\sigma_\varepsilon^2}\left(\displaystyle\sum_{I=1}^{n}\left(\frac{y_i-\beta_0-\beta_1x_i}{\sigma_\varepsilon^2}\right)^2\right) = 0. & (2.16c) \end{cases}$$

Posons $L = \mathcal{L}(\beta_0,\beta_1,\sigma^2)$ et égalisons les équations du système à 0, elles s'écrivent alors comme suit :

$$\begin{cases} \dfrac{\partial logL}{\partial\beta_0} = 0 \Rightarrow \displaystyle\sum_{I=1}^{n}\left(y_i-\tilde{\beta}_0-\tilde{\beta}_1x_i\right) = 0 \Rightarrow \sum_{I=1}^{n}y_i = n\tilde{\beta}_0 + \tilde{\beta}_1\sum_{I=1}^{n}x_i & (2.17a) \\ \dfrac{\partial logL}{\partial\beta_1} = 0 \Rightarrow \displaystyle\sum_{I=1}^{n}\left(y_i-\tilde{\beta}_0-\tilde{\beta}_1x_i\right)x_i = 0 \Rightarrow \sum_{I=1}^{n}y_ix_i = \tilde{\beta}_0\sum_{I=1}^{n}x_i + \tilde{\beta}_1\sum_{I=1}^{n}x_i^2 & (2.17b) \\ \dfrac{\partial logL}{\partial\sigma_\varepsilon^2} = 0 \Rightarrow -n + \dfrac{1}{\tilde{\sigma}_\varepsilon^2}\displaystyle\sum_{I=1}^{n}\left(y_i-\tilde{\beta}_0-\tilde{\beta}_1x_i\right)^2 = 0 \Rightarrow \tilde{\sigma}_\varepsilon^2 = \dfrac{1}{n}\sum_{I=1}^{n}\left(y_i-\tilde{\beta}_0-\tilde{\beta}_1x_i\right)^2 & (2.17c) \end{cases}$$

avec $\tilde{\beta}_0$, $\tilde{\beta}_1$ et $\tilde{\sigma}_\varepsilon^2$ les estimateurs du maximum de vraisemblance respectifs des paramètres β_0, β_1 et σ_ε^2.
En effet, de l'expression $\dfrac{\partial}{\partial\sigma_\varepsilon^2}\left(-\dfrac{n}{2}log\sigma_\varepsilon^2\right) - \dfrac{1}{2}\dfrac{\partial}{\partial\sigma_\varepsilon^2}\left(\displaystyle\sum_{I=1}^{n}\left(\frac{y_i-\beta_0-\beta_1x_i}{\sigma_\varepsilon^2}\right)^2\right)$, on a :

$$\frac{\partial}{\partial\sigma_\varepsilon^2}\left(-\frac{n}{2}log\sigma_\varepsilon^2\right) = -\frac{n}{2}\frac{1}{\sigma_\varepsilon^2} = -\frac{n}{2\sigma_\varepsilon^2}$$

et

$$-\frac{1}{2}\frac{\partial}{\partial\sigma_\varepsilon^2}\left(\sum_{I=1}^{n}\left(\frac{y_i-\beta_0-\beta_1x_i}{\sigma_\varepsilon^2}\right)^2\right) = \frac{1}{2}\frac{1}{\sigma_\varepsilon^4}\sum_{I=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2.$$

Donc

$$\begin{aligned}\frac{\partial logL}{\partial\sigma_\varepsilon^2} &= -\frac{n}{2\sigma_\varepsilon^2} + \frac{1}{2}\frac{1}{\sigma_\varepsilon^4}\sum_{I=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2 \\ &= -n + \frac{1}{\sigma_\varepsilon^2}\sum_{I=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2,\end{aligned}$$

d'où l'expression (2.17c) ci-dessus.
Ainsi de l'expression (2.17a), on obtient :

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1\bar{x}.$$

En substituant l'expression de $\tilde{\beta_0}$ dans celle de $\displaystyle\sum_{I=1}^{n}y_ix_i = \tilde{\beta}_0\sum_{I=1}^{n}x_i + \tilde{\beta}_1\sum_{I=1}^{n}x_i^2$, nous obtenons l'estimateur $\tilde{\beta}_1$:

$$\tilde{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}}{\displaystyle\sum_{i=1}^{n}x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}.$$

Pour ce qui est de l'estimateur de la variance $\tilde{\sigma}_\varepsilon^2$, de l'expression (2.17c) ci dessus, nous obtenons :

$$\tilde{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{I=1}^{n}\left(y_i-\tilde{\beta}_0-\tilde{\beta}_1x_i\right)^2 = \frac{1}{n}\sum_{I=1}^{n}\tilde{\varepsilon}_i^2.$$

Le constat est immédiat, on remarque que les estimateurs MCO vue au chapitre précédant des paramètres β_0 et β_1 et ceux du maximum de vraisemblance sont égaux ou les mêmes : $\hat{\beta}_0 = \tilde{\beta}_0$ et $\hat{\beta}_1 = \tilde{\beta}_1$. Par contre l'estimateur du maximum de vraisemblance $\tilde{\sigma}_\varepsilon^2$ de la variance de l'erreur σ_ε^2 est différent de l'estimateur MCO :

$$\tilde{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{I=1}^{n}\tilde{\varepsilon}_i^2 \neq \hat{\sigma}_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

2.4 Comparaison des deux méthodes d'estimation

Comparer deux méthodes d'estimations revient à comparer pour chacune des méthodes les estimateurs trouvés. Pour y arriver nous ferons recours aux critères de comparaison que sont :

$*$ voir si l'estimateur est biaisé ou non ;

$*$ dans le cas ou l'estimateur serait biaisé, vérifié le critère de convergence ;

$*$ comparer les erreurs quadratiques moyennes de chacun des estimateurs (l'éfficacité) ;

$*$ pour finir comparer les valeurs des variances .

2.4.1 Calcul du biais des estimateurs

Nous avons pu voir en régression linéaire simple, que les estimateurs des paramètres β_0 et β_1 sont égaux pour chacune des deux méthodes : $\hat{\beta}_0 = \tilde{\beta}_0$ et $\hat{\beta}_1 = \tilde{\beta}_1$ et que les estimateurs MCO sont sans biais ,alors ces EMV le sont aussi.

Par contre l'estimateur du maximum vraisemblance de la variance $\tilde{\sigma}_\varepsilon^2$ de l'erreur est biaisé, et est différent de l'estimateur MCO car :

$$\mathbb{E}(\tilde{\sigma}_\varepsilon^2) = \frac{1}{n}\mathbb{E}\left(\sum_{I=1}^{n}\tilde{\varepsilon}_i^2\right) \neq \mathbb{E}(\hat{\sigma}_\varepsilon^2) = \frac{1}{(n-2)}\mathbb{E}\left(\sum_{I=1}^{n}\hat{\varepsilon}_i^2\right) = \sigma_\varepsilon^2.$$

On dit que $\tilde{\sigma}_\varepsilon^2$ est biaisé vers le bas (en moyenne, il sous estime la valeur réelle de σ_ε^2).

Cependant il reste convergent :

$$\lim_{n\to\infty}\mathbb{E}(\tilde{\sigma}_\varepsilon^2) = \sigma_\varepsilon^2, \text{ ou lorsque } \quad n \quad \text{ est grand, on a : } \quad \tilde{\sigma}_\varepsilon^2 \simeq \hat{\sigma}_\varepsilon^2.$$

Cela garantit la minimisation du biais avec l'accroissement de la taille de l'échantillon . Autrement dit $\tilde{\sigma}_\varepsilon^2$ est asymptotiquement sans biais.

2.4.2 Calcul de la variance des estimateurs

Tout d'abord on se rappelle que pour les estimateurs MCO $\hat{\beta}_0$ et $\hat{\beta}_1$, on a :

$$Var(\hat{\beta}_0) = \sigma_\varepsilon^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right], \text{ et } Var(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{S_{xx}}.$$

Il en est de même pour ceux de MV :

$$Var(\tilde{\beta}_0) = \sigma_\varepsilon^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right], \text{ et } Var(\tilde{\beta}_1) = \frac{\sigma_\varepsilon^2}{S_{xx}}.$$

Pour ce qui est des estimateurs $\hat{\sigma}_\varepsilon^2$ et $\tilde{\sigma}_\varepsilon^2$, on a :

$$Var(\hat{\sigma}_\varepsilon^2) = \frac{2\sigma_\varepsilon^4}{n-2} \text{ et } Var(\tilde{\sigma}_\varepsilon^2) = \frac{2\sigma_\varepsilon^4}{n}.$$

En effet,

$$\begin{aligned} Var(\hat{\sigma}_\varepsilon^2) &= Var\left(\frac{\sigma_\varepsilon^2}{n-2} \times \frac{\|Y - X\hat{\beta}\|^2}{\sigma_\varepsilon^2}\right) \\ &= \frac{\sigma_\varepsilon^4}{n-2} Var\left(\frac{\|Y - X\hat{\beta}\|^2}{\sigma_\varepsilon^2}\right) \\ &= \frac{\sigma_\varepsilon^4}{(n-2)^2} \times 2(n-2) \qquad \text{car } \frac{\|Y - X\hat{\beta}\|^2}{\sigma_\varepsilon^2} \sim \chi^2_{n-2} \end{aligned}$$

donc, $Var(\hat{\sigma}_\varepsilon^2) = \dfrac{2\sigma_\varepsilon^4}{n-2}$ et la démonstration est la même pour $Var(\tilde{\sigma}_\varepsilon^2)$ vu que $\dfrac{\|Y - X\tilde{\beta}\|^2}{\sigma_\varepsilon^2} \sim \chi^2_n$.

Remarque. *La variance de l'estimateur MCO est plus importante que celle de l'estimateur MV :*

$$Var(\tilde{\sigma}_\varepsilon^2) = \frac{2\sigma_\varepsilon^4}{n} < \frac{2\sigma_\varepsilon^4}{n-2} = Var(\hat{\sigma}_\varepsilon^2);$$

parce que le dénominateur pour l'estimateur MCO est plus petit que celui de l'estimateur MV.
En conséquence, l'estimateur du maximum de vraisemblance sera biaisé, mais plus précis que l'estimateur MCO !

On doit faire un arbitrage entre le biais et la précision de l'estimateur. C'est pourquoi nous développons un critère pour évaluer les propriétés des estimateurs, même s'ils sont biaisés : l'erreur quadratique moyenne.

2.4.3 Calcul de l'erreur quadratique moyenne EQM des estimateurs

On se rappelle que l'erreur quadratique moyenne est définie par :

$$\mathbf{E.Q.M}(\hat{\theta}_n) = \mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right] = Var(\hat{\theta}_n) + b(\hat{\theta}_n, \theta)^2.$$

Pour les estimateurs MCO et MV des paramètres β_0 et β_1.
Avec les estimateurs MCO qui sont non biaisés, nous avons les résultats suivants :

$$\mathbf{E.Q.M}(\hat{\beta}_0) = \mathbf{E.Q.M}(\tilde{\beta}_o) = Var(\tilde{\beta}_0) = \sigma_\varepsilon^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right].$$

et

$$\mathbf{E.Q.M}(\hat{\beta}_o) = \mathbf{E.Q.M}(\tilde{\beta}_1) = Var(\tilde{\beta}_0) = \frac{\sigma_\varepsilon^2}{S_{xx}}.$$

Pour l'estimateur MCO de la variance σ_ε^2, on a :

$$\mathbf{E.Q.M}(\hat{\sigma}_\varepsilon^2) = Var(\hat{\sigma}_\varepsilon^2) + \underbrace{b^2(\hat{\sigma}_\varepsilon^2)}_{0} = \frac{2\sigma_\varepsilon^4}{n-2}.$$

Pour l'estimateur MV de la variance σ_ε^2, on a :

$$\mathbf{E.Q.M}(\tilde{\sigma}_\varepsilon^2) = Var(\tilde{\sigma}_\varepsilon^2) + \underbrace{b^2(\tilde{\sigma}_\varepsilon^2)}_{=-\frac{2\sigma_\varepsilon^2}{n}} = \frac{2\sigma_\varepsilon^4}{n} + \left(-\frac{2\sigma_\varepsilon^2}{n}\right)^2.$$

Celle-ci se réécrit : $\mathbf{E.Q.M}(\tilde{\sigma}_\varepsilon^2) = \left(\dfrac{2}{n} + \dfrac{4}{n^2}\right)\sigma_\varepsilon^4.$

Voici le tableau récapitulatif qui résume tout :

Estimateurs	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}_\varepsilon^2$	$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\sigma}_\varepsilon^2$
biaisé/ non biaisé	non biaisé	non biaisé	non biaisé	non biaisé	non biaisé	biaisé
Variance	$\sigma_\varepsilon^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)$	$\frac{\sigma_\varepsilon^2}{S_{xx}}$	$\frac{2\sigma_\varepsilon^4}{n-2}$	$\sigma_\varepsilon^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)$	$\frac{\sigma_\varepsilon^2}{S_{xx}}$	$\frac{2\sigma_\varepsilon^4}{n}$
E.Q.M	$\sigma_\varepsilon^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)$	$\frac{\sigma_\varepsilon^2}{S_{xx}}$	$\frac{2\sigma_\varepsilon^4}{n-2}$	$\sigma_\varepsilon^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)$	$\frac{\sigma_\varepsilon^2}{S_{xx}}$	$\left(\frac{2}{n}+\frac{4}{n^2}\right)\sigma_\varepsilon^4$

TABLE 2.2 – Tableau comparatif des estimateurs MCO et MV pour la régression linéaire simple

Chapitre 3

Régression linéaire multiple, Comparaison des Méthodes d'Estimation

La régression linéaire multiple est une généralisation de la régression linéaire simple, dans le sens où cette approche permet d'évaluer les relations linéaires entre une variable à expliquer et plusieurs variables explicatives (de type numérique ou catégoriel).

3.1 Modèle théorique

1. **Définition**
   Le modèle de régression linéaire multiple, noté **rlm** ou modèle linéaire multiple, est défini par :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad i \in \{1, \ldots, n\}, \tag{3.1}$$

   où $\beta_0, \beta_1, \ldots, \beta_p$ sont appelés les paramètres ou les coefficients inconnus du modèle que l'on veut estimer à partir des données, et ε_i est l'erreur du modèle (bruit) qui représente la déviation entre ce que le modèle prédit et la réalité.

2. **Écriture matricielle**
   Ce modèle s'écrit sous la forme d'équations comme suit :

$$\begin{cases} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1P} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2P} + \varepsilon_2 \\ & \vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_p x_{nP} + \varepsilon_n \end{cases}$$

   Ou sous la forme matricielle :

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

   De façon équivalente, on écrit :

$$Y = X\beta + \varepsilon. \tag{3.2}$$

où

Y : est le vecteur à expliquer de taille n;

X : est la matrice, de taille $n \times (p+1)$, qui contient l'ensemble des observations sur les variables exogènes, avec une première colonne formée par la valeur 1 indiquant que l'on intègre la constante β_0 dans l'équation, où p est le nombre de variables explicatives réelles ;

ε : est le vecteur des erreurs de taille n.

Remarque. *(i) Le coefficient β_0 est un paramètre appelé intercepte qui représente la moyenne des y_i lorsque la valeur de chaque variable explicative est égale à 0.*

(ii) Les coefficients β_j $\{j = 1, \ldots, p\}$ représentent le changement subi par $E(y_i)$ correspondant à un changement unitaire dans la valeur de la j-ième variable explicative, lorsque les autres variables explicatives demeurent inchangées.

3.2 Hypothèses relatives du modèle rlm

Comme en régression simple, les hypothèses permettent de déterminer les propriétés des estimateurs (biais, convergence) et leurs lois de distributions (pour les estimations par intervalle et les tests d'hypothèses) (voir Bourbonnais [6], Labrousse [23] et Giraud et Chaix [17]).

Il existe principalement deux catégories d'hypothèses :

▷**Hypothèses stochastiques**

H1 Les erreurs sont centrées (le modèle est bien spécifié en moyenne), c'est-à-dire que l'ensemble des déterminants de y qui n'ont pas été retenus dans le modèle a une espérance nulle :

$$E(\varepsilon_i) = 0, \quad \forall i \in \{1, \ldots, n\}.$$

H2 La variance des erreurs est constante, on parle d'homogénéité des variances ou encore d'homoscédasticité :

$$\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2, \quad \forall i \in \{1, \ldots, n\}.$$

H3 Les erreurs relatives à deux termes aléatoires ne sont pas corrélées, on dit qu'il n'y a pas de corrélation sérielle :

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, \quad \forall i \neq j.$$

H4 Les ε_i sont indépendants et identiquement distribués (i.i.d) et suivent une loi normale de moyenne nulle et de variance σ_ε^2, on écrit :

$$\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2), \quad \forall i \in \{1, \ldots, n\}.$$

▷**Hypothèses structurelles**

H1 La matrice (X^tX) (où X^t est la matrice transposée de X) est non singulière de rang p, c'est-à-dire que $\det(X^tX) \neq 0$ et $(X^tX)^{-1}$ existe. Cette hypothèse implique l'absence de colinéarité entre les variables exogènes $(X_1, \ldots, X_p)$, c'est-à-dire que les différents vecteurs X_j sont linéairement indépendants. En cas de multicolinéarité, la méthode des MCO devient défaillante.

H2 $\dfrac{(X^tX)}{n}$ tend vers une matrice finie non singulière lorsque $n \to +\infty$:

$$\lim_{n \to +\infty} \frac{(X^tX)}{n} = A, \quad \text{avec } \det(A) \neq 0.$$

H3 : $n > p+1$, c'est-à-dire que le nombre d'observations est supérieur au nombre de paramètres du modèle p $(j = 0, \ldots, p)$.

3.3 Estimation des paramètres par MCO

Conditionnellement à la connaissance des valeurs des X_j $(j = 1, \ldots, p)$, les paramètres inconnus du modèle : le vecteur $\beta = (\beta_0, \ldots, \beta_p)$ et σ_ε^2, sont estimés par minimisation du critère des moindres carrés ordinaires (MCO).
Le principe des moindres carrés choisit le vecteur $\hat{\beta}$ minimisant la fonction de la somme des carrés des résidus.

•**Estimation de** β

Problème : On cherche à estimer le paramètre $\beta = (\beta_0, \cdots\cdots, \beta_p)$ du modèle de régression et ce de manière optimale. Pour cela, on utilise la méthode des moindres carrés ordinaires (M.C.O). Cette méthode consiste à minimiser la quantité suivante :

$$\underset{\beta_0,\cdots\cdots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots\cdots + \beta_p x_{i,p}\right)\right)^2.$$

Sachant que notre modèle linéaire s'écrit comme suit : $\mathbf{Y} = X\beta + \epsilon$, alors le vecteur des résidus est :

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta}.$$

On remarque que les variables Y et X sont mesurées tandis que l'estimateur $\hat{\beta}$, elle est à déterminer.
En d'autre termes, il s'agit de trouver le vecteur $\hat{\beta}$ qui minimise $\|\varepsilon\|^2 = \varepsilon^t\varepsilon$.

$$\|\varepsilon\|^2 = \hat{\beta} = \underset{\beta_0,\cdots\cdots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{i,j}\right)^2 = \underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y - X\beta\|^2.$$

Calcul : posons

$$d(\beta) = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{i,j}\right)^2,$$

alors

$$\begin{aligned} d(\beta) = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{i,j}\right)^2 &= \|Y - X\beta\|^2 \\ &= (Y - X\beta)^t (Y - X\beta) \\ &= Y^tY - 2\beta^tX^tY + \beta^tX^tX\beta. \end{aligned}$$

Une condition nécessaire d'optimum est que la dérivée première par rapport à β s'annule. Or la dérivée s'écrit comme suit :

$$\frac{\partial d(\beta)}{\partial \beta} = -2X^tY + 2X^tX\beta,$$

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

$$-2X^tY + 2X^tX\hat{\beta} = 0.$$

Ce qui aboutit au résultat :

$$\hat{\beta} = \left(X^tX\right)^{-1} X^tY. \tag{3.3}$$

Pour s'assurer que ce point $\hat{\beta}$ est bien un minimum strict, il faut que la dérivée seconde soit une matrice définie

positive. Or la dérivée seconde s'écrit

$$\frac{\partial^2 d(\beta)}{\partial \beta^2} = 2X^t X.$$

Et donc X est de plein rang, donc $X^t X$ est inversible et n'a pas de valeur propre nulle. La matrice $X^t X$ est donc définie. De plus $\forall w \in \mathbb{R}^{p+1}$, nous avons

$$w^t 2X^t X w = 2\langle Xw, Xw\rangle = 2\|Xw\|^2 \geq 0.$$

$(X^t X)$ est donc bien définie positive et $\hat{\beta}$ est bien un minimum strict.
Et les valeurs ajustées (ou estimées, prédites) de Y ont pour expression :

$$\hat{Y} = X\hat{\beta} = X\left(X^t X\right)^{-1} X^t Y = HY, \tag{3.4}$$

où $H = X\left(X^t X\right)^{-1} X^t$ est appelée « *hat matrix* » (ou matrice chapeau).
Géométriquement, c'est la matrice de projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace vectoriel Vect(X) engendré par les vecteurs colonnes de X.
On note

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta} = (Id - H)Y,$$

le vecteur des résidus ; c'est la projection de Y sur le sous-espace orthogonal Vect(X) dans $\mathbb{R}^n$.

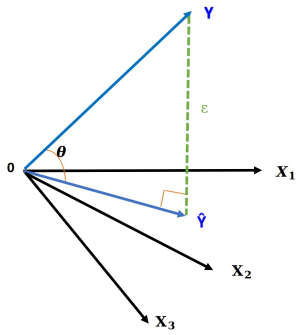

FIGURE 3.1 – Géométriquement, la régression est la projection $\hat{Y}$ de Y sur l'espace vectoriel $Vect\left\{1, X_1, \cdots\cdots, X_p\right\}$; de plus $R^2 = cos^2(\theta)$.

Quelques propriétés statistiques

Le statisticien cherche à vérifier que les estimateurs des MC que nous avons construits admettent de bonnes propriétés au sens statistique. Dans notre cadre de travail, cela peut se résumer en deux parties : l'estimateur des MC est-il sans biais et est-il de variance minimale dans sa classe d'estimateurs ?
Il est aussi important de noter que le passage du modèle de régression de 2 à p variables explicatives ne modifie en rien les propriétés statistiques de l'estimateur MCO. De même l'interprétation de ces propriétés reste inchangée. (Voir [2]).
Pour essayer de répondre aux questions précédentes, nous utiliserons les hypothèses $H1$ et $H2$ à savoir $\mathbb{E}(\varepsilon_i) = 0$ et $Var(\epsilon_i) = \sigma_\varepsilon^2 Id$. Ce qui nous permet de calculer

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}\left(\left(X^t X\right)^{-1} X^t Y\right) = \left(X^t X\right)^{-1} X^t \mathbb{E}(Y) = \left(X^t X\right)^{-1} X^t X\beta = \beta.$$

L'estimateur des M.C.O est donc sans biais. Calculons sa variance

$$Var(\hat{\beta}) = Var\left(\left(X^t X\right)^{-1} X^t Y\right) = \left(X^t X\right)^{-1} X^t Var(Y) X \left(X^t X\right)^{-1} = \sigma_\varepsilon^2 \left(X^t X\right)^{-1}.$$

Propriétés 7.
Les estimateurs des M.C.O $\hat{\beta}_j$, $j = 0, \cdots\cdots, p$ sont des estimateurs sans biais : $\mathbb{E}(\hat{\beta}) = \beta$ et parmi les estimateurs sans biais fonctions linéaires des y_i, ils sont de variance minimum (propriété de Gauss-Markov) ; ils sont donc "BLUE" : best linear unbiaised estimators.

Matrice de variance-covariance des coefficients

La matrice de variance-covariance des coefficients est importante car elle renseigne sur la variance de chaque coefficient estimé, et permet de faire des tests d'hypothèse, notamment de voir si chaque coefficient est significativement différent de zéro. Elle est définie par :

$$Var(\hat{\beta}) = \Sigma = \mathbb{E}\left[(\hat{\beta} - \beta)^t (\hat{\beta} - \beta)\right] = \sigma_\varepsilon^2 \left(X^t X\right)^{-1}.$$

En effet,

- On sait que $Y = X\beta + \varepsilon$ et $\hat{\beta} = \left(X^t X\right)^{-1} X^t Y$.
  En remplaçant Y par son expression on obtient :

$$\begin{aligned} \hat{\beta} = \left(X^t X\right)^{-1} X^t (X\beta + \Leftrightarrow \hat{\beta} &= \left(X^t X\right)^{-1} X^t (X\beta) + \left(X^t X\right)^{-1} X^t \varepsilon \\ &\Leftrightarrow \hat{\beta} = \beta + \left(X^t X\right)^{-1} X^t \varepsilon. \end{aligned}$$

  En utilisant cette nouvelle écriture de $\hat{\beta}$, on obtient que :

$$\begin{aligned} Var(\hat{\beta}) &= Var\left[\left(X^t X\right)^{-1} X^t \varepsilon\right] \\ &= \left(X^t X\right)^{-1} X^t Var[\varepsilon] X \left(X^t X\right)^{-1} \\ &= \left(X^t X\right)^{-1} X^t \sigma_\varepsilon^2 X \left(X^t X\right)^{-1} \\ &= \sigma_\varepsilon^2 \left(X^t X\right)^{-1} X^t X \left(X^t X\right)^{-1} \\ &= \sigma_\varepsilon^2 \left(X^t X\right)^{-1}. \end{aligned}$$

C'est dans cette suite logique que nous énonçons le théorème suivant.

Théorème 8. *(Théorème de Gauss-Markov. Pour plus de détail voir [16])*
L'estimateur des moindres carrés est de variance minimale parmi les estimateurs linéaires sans biais de β.

Preuve.
Supposons que $\tilde{\beta}$ est un autre estimateur linéaire et sans biais de β. Essayons de montrer que $Var(\tilde{\beta}) \geq Var(\hat{\beta})$. Cette inégalité entre matrices est à comprendre au sens suivant : $Var(\tilde{\beta}) \geq Var(\hat{\beta})$ si la matrice $\Delta := Var(\tilde{\beta}) - Var(\hat{\beta})$ est semi-définie positive.
L'estimateur $\tilde{\beta}$ est linéaire donc il s'écrit sous la forme $\tilde{\beta} = AY$, pour une certaine matrice A de dimension $(P+1) \times n$.
De plus, il est sans biais, donc pour tout $\beta \in \mathbb{R}^{p+1}$, on a :

$$\mathbb{E}(\tilde{\beta}) = \mathbb{E}(AY) = AX\beta = \beta, \text{ d'où } AX = Id_{P+1}.$$

Décomposons maintenant la variance de $\tilde{\beta}$ comme suit :

$$\begin{aligned} Var(\tilde{\beta}) &= Var(\tilde{\beta} - \hat{\beta} + \hat{\beta}) \\ &= Var(\tilde{\beta} - \hat{\beta}) + Var(\hat{\beta}) + Cov(\tilde{\beta} - \hat{\beta}, \hat{\beta}) + Cov(\hat{\beta}, \tilde{\beta} - \hat{\beta}). \end{aligned}$$

On a :

$$\begin{aligned} Cov(\tilde{\beta}-\hat{\beta},\hat{\beta}) &= Cov\left(AY,(X^tX)^{-1}X^tY\right) - Var(\tilde{\beta}) \\ &= \sigma_\varepsilon^2 AX(X^tX)^{-1} - \sigma_\varepsilon^2(X^tX)^{-1} \\ &= 0. \qquad \square \end{aligned}$$

Car $AX = Id_{p+1}$. Il s'en suit que $Var(\tilde{\beta}) = Var(\tilde{\beta}-\hat{\beta}) + Var(\hat{\beta})$ et comme les matrices de variance-covariances sont semi-définies positives, on a bien $Var(\tilde{\beta}) \geq Var(\hat{\beta})$. $\square$

•**Estimation de σ_ε^2**

On note : σ_ε^2 la vraie valeur théorique de la variance des résidus et $\hat{\sigma}_\varepsilon^2$ l'estimateur de σ_ε^2.
Un estimateur sans biais de la variance est donné par :

$$\hat{\sigma}_\varepsilon^2 = \frac{\|\hat{\varepsilon}\|^2}{n-p-1} = \frac{\|Y-X\hat{\beta}\|^2}{n-p-1} = \frac{\sum_{i=1}^{n}(\hat{\varepsilon}_i)^2}{n-p-1} = \frac{SC_{res}}{n-p-1}.$$

Lemme 1.
Soit un vecteur Z composé de n variables aléatoires d'espérances nulles tel que $Var(Z) = \sigma_z^2 Id_n$ et A une matrice symétrique non aléatoire, alors

$$\mathbb{E}\left[Z^tAZ\right] = \sigma_z^2 tr(A),$$

avec $tr(A)$, la trace de la matrice A.

Ainsi, grâce au lemme 1, on peut calculer l'espérance de $\hat{\varepsilon}^t\hat{\varepsilon}$.

Théorème 9.
Soit $\hat{\varepsilon} = Y - X\hat{\beta}$, alors

$$\mathbb{E}(SC_{res}) = (n-p-1)\sigma_\varepsilon^2.$$

Preuve.
En utilisant l'équation (3.4) et l'expression du vecteur résidus $\hat{\varepsilon}$, on a :

$$\hat{\varepsilon} = Y - \hat{Y} = Y - HY = (Id_n - H)Y,$$

où Id_n, est la matrice identité de dimension (n,n) et H la matrice chapeau telle définie précédemment de taille $(n \times n)$.
H vérifie les deux propriétés (voir [25]) :

$$\begin{cases} symetrique : H^t &= H \\ idempotente : H^2 &= H. \end{cases} \tag{3.5}$$

Mais aussi la matrice $(Id_n - H)$ a les même propriétés :

$$\begin{cases} (Id_n - H)^t &= (Id_n - H) \\ (Id_n - H)^2 &= (Id_n - H). \end{cases} \tag{3.6}$$

Donc,

$$\begin{aligned}\hat{\varepsilon} &= Y - \hat{Y} = Y - HY \\ &= (Id_n - H)(X\beta + \varepsilon) \\ &= X\beta - HX\beta + \varepsilon + H\varepsilon,\end{aligned}$$

où $HX\beta = \left(X(X^tX)^{-1}X^t\right)X = X$ et $(Id_n - H)X = 0$, ce qui nous donne :

$$\hat{\varepsilon} = (Id_n - H)\varepsilon.$$

On obtient :

$$\begin{aligned}\hat{\varepsilon}^t\hat{\varepsilon} &= \left[((Id_n - H)\varepsilon)^t\left((Id_n - H)\varepsilon\right)\right] \\ &= \varepsilon^t(Id_n - H)^t(Id_n - H)\varepsilon \\ &= \varepsilon^t(Id_n - H)^2\varepsilon, \text{ d'après l'équation (3.5)} \\ &= \varepsilon^t(Id_n - H)\varepsilon = \varepsilon^t Id_n\varepsilon - \varepsilon^t H\varepsilon, \text{ d'après l'équation (3.6)}\end{aligned}$$

Et par le lemme 1, on obtient :

$$\begin{aligned}\mathbb{E}(\hat{\varepsilon}^t\hat{\varepsilon}) &= \left[\varepsilon^t Id_n\varepsilon - \varepsilon^t H\varepsilon\right] \\ &= \mathbb{E}\left[\varepsilon^t Id_n\varepsilon\right] - \mathbb{E}\left[\varepsilon^t H\varepsilon\right] \\ &= \sigma_\varepsilon^2 tr(Id_n) - \sigma_\varepsilon^2 tr(H) = \sigma_\varepsilon^2\left[tr(Id_n) - tr(H)\right],\end{aligned}$$

où $tr(Id_n) = n$ et

$$\begin{aligned}tr(H) &= tr\left(X(X^tX)^{-1}X^t\right) \\ &= tr\left(X^tX(X^tX)^{-1}\right), \text{puisque } tr(AB) = tr(BA) \\ &= tr(Id_n) = p + 1, \text{ vu que la } X^tX \text{ est une matrice carrée de taille } p + 1.\end{aligned}$$

Donc,

$$\mathbb{E}\left[\varepsilon^t\varepsilon\right] = (n - p - 1)\sigma_\varepsilon^2.$$

□

D'après le théorème 9, on peut construire l'estimateur sans biais pour σ_ε^2 qui est :

$$\hat{\sigma}_\varepsilon^2 = \frac{SC_{res}}{n - p - 1} \equiv MC_{res}.$$

3.3.1 Lois des estimateurs - Intervalles et régions de confiance

Comme nous connaissons l'estimateur de β_j , son espérance, ainsi qu'une estimation de sa variance, nous pouvons déterminer la loi de distribution de celui-ci et construire des intervalles de confiance ou tests d'hypothèses sur β.

Proposition 7. *(Lois des estimateurs avec variance connue)*

(i) *$\hat{\beta}$ est un vecteur gaussien de moyenne β et variance $\sigma_\varepsilon^2\left(X^tX\right)^{-1}$:*

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma_\varepsilon^2\left(X^tX\right)^{-1}).$$

(ii) $\dfrac{(n-p-1)}{\sigma_\varepsilon^2}\hat{\sigma}_\varepsilon^2 \sim \chi^2_{n-p-1}$.

(ii) $\hat{\beta}$ *et* $\hat{\sigma}_\varepsilon^2$ *sont indépendantes.*

Pour la preuve et plus de détail (voir [4],[11]).

Remarque.
Le premier point du précédant résultat, n'assure pas l'obtention des régions de confiance car sur β*, on suppose que* σ_ε^2 *est connue, ce qui n'est pas le cas en général. Toutefois, la proposition suivante pallie ce manquement.*

Proposition 8. *(lois des estimateurs avec variance inconnue)*
Les lois des estimateurs MCO avec variance σ_ε^2 *inconnue sont :*

(i) $\dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_\varepsilon(\hat{\beta}_j)}$ *suit une loi de Student à* $(n-p-1)$ *degrés de liberté.*

(ii) Soit M *une matrice de taille* $m \times (p+1)$ *de rang* m *(nombre maximum de lignes ou de colonnes linéairement indépendantes), alors :*

$$\frac{1}{m\hat{\sigma}_\varepsilon^2}\left(M(\hat{\beta}-\beta)\right)^t \left[M(X^tX)^{-1}M^t\right]^{-1} M\left(\hat{\beta}-\beta\right) \sim \mathcal{F}_{(m,n-p-1)}(1-\alpha)\,,$$

où $\mathcal{F}_{(m,n-p-1)}$ *est le quantile d'ordre* $(1-\alpha)$ *d'une loi de Fisher à* $(m, n-p-1)$ *degrés de liberté (ddl) .*

Preuve. (Pour plus de détail (voir [4],[11])

(i) D'après la **proposition 7**, on a :

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma_\varepsilon^2 \left(X^tX\right)^{-1}).$$

Alors pour tout $j \in \{0,\dots,p\}$, en notant $\left[\left(X^tX\right)^{-1}\right]_{jj}$ la j-ème composante diagonale de $\left(X^tX\right)^{-1}$, on peut écrire :

$$\hat{\beta}_j \sim \mathcal{N}(\beta, \sigma_\varepsilon^2 \left[\left(X^tX\right)^{-1}\right]_{jj}) \qquad et \qquad \hat{\sigma}_\varepsilon(\hat{\beta}_j) = \hat{\sigma}_\varepsilon\sqrt{[(X^tX)^{-1}]_{jj}}.$$

Ici comme $\hat{\beta}_j \sim \mathcal{N}(\beta, \sigma_\varepsilon^2 \left[\left(X^tX\right)^{-1}\right]_{jj})$, on en déduit que $\dfrac{\hat{\beta}_j - \beta_j}{\sigma_\varepsilon} \sim \mathcal{N}(0,1)$. En outre la proposition précédente établit que $\dfrac{(n-p-1)}{\sigma_\varepsilon^2}\hat{\sigma}_\varepsilon^2 \sim \chi^2_{n-p-1}$ et est indépendante de $\hat{\beta}_j$. On a donc par application directe de la définition

$$\frac{(\hat{\beta}_j - \beta_j)/\sigma_\varepsilon}{\sqrt{\dfrac{(n-p-1)}{\sigma_\varepsilon^2}\hat{\sigma}_\varepsilon^2/(n-p-1)}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_\varepsilon(\hat{\beta}_j)} \sim Student(n-p-1).$$

(ii) Dans un premier temps, rappelons une caractérisation de la loi de Fisher. Soient A et B deux variables indépendantes avec $A \sim \chi^2(\nu_1)$ et $B \sim \chi^2(\nu_2)$, alors $F = \frac{\nu_2 A}{\nu_1 B} \sim F(\nu_1,\nu_2)$. On pose alors :

$$A = \frac{(M\hat{\beta} - M\beta)^t (M(X^tX)^{-1}M^t)^{-1}(M\hat{\beta} - M\beta)}{\sigma^2},$$
$$B = \frac{(n-p-1)}{\sigma_\varepsilon^2}\hat{\sigma}_\varepsilon^2.$$

En utilisant le théorème de Cochran, on peut montrer que A et B sont indépendantes avec $A \sim \chi^2(k)$ et $B \sim \chi^2(n-(p+1))$. Par la caractérisation de la loi de Fisher, il s'ensuit

$$\frac{(M\hat{\beta} - M\beta)^t (M(X^tX)^{-1}M^t)^{-1}(M\hat{\beta} - M\beta)}{(n-(m+1))A} \sim F(m, n-(p+1)).$$

□

Certains logiciels et ouvrages donnent les IC pour des paramètres pris séparément, mais ne tiennent pas compte de la dépendance de ces derniers. C'est pour cette raison que nous nous permettons de construire les RC.

Proposition 9. *(intervalles et régions de confiance)*
Soit un seuil α petit où $\alpha \in [0,1]$, on a :

1. *L'intervalle de confiance pour β_j , noté par IC(β_j), est :*
$$\left[\hat{\beta}_j - t_{n-p-1}(1-\alpha/2)\sigma_{\hat{\beta}_j}, \hat{\beta}_j + t_{n-p-1}(1-\alpha/2)\sigma_{\hat{\beta}_j}\right],$$
où $\sigma_{\hat{\beta}_j} = Var(\hat{\beta}_j)$, $t_{n-p-1}(1-\alpha/2)$ le quantile de niveau $(1-\alpha/2)$ d'une loi de Student $\mathcal{T}_{n-p-1}$.

2. *Une région de confiance au niveau $(1-\alpha)$ pour $M\beta$ est :*
$$\frac{1}{m\hat{\sigma}_\varepsilon^2}\left(M(\hat{\beta}-\beta)\right)^t \left[M(X^tX)^{-1}M^t\right]^{-1} M\left(\hat{\beta}-\beta\right) \sim \mathcal{F}_{(m,n-p-1)}(1-\alpha),$$
où $\mathcal{F}_{(m,n-p-1)}$ est le quantile d'ordre $(1-\alpha)$ de la loi de Fisher à $(m, n-p-1)$ (ddl).

3. *L'intervalle de confiance pour σ_ε^2 est donné par :*
$$\left[\frac{(n-p-1)\hat{\sigma}_\varepsilon^2}{c_{n-p-1}(1-\alpha/2)}, \frac{(n-p-1)\hat{\sigma}_\varepsilon^2}{c_{n-p-1}(\alpha/2)}\right],$$
où $c_{n-p-1}(1-\alpha/2)$ est le fractile ou quantile de niveau $(1-\alpha/2)$ d'une loi de χ^2_{n-p-1} à $(n-p-1)$ degré de liberté.

Pour plus de détail (voir [4],[11]).

Preuve.

1. Posons $k = p+1$ avec p le nombre de variables explicatives. Selon les propriétés de $\hat{\beta}$, on peut écrire que $\hat{\beta} \sim \mathcal{N}_k\left(\beta, \sigma_\varepsilon^2(X^tX)^{-1}\right)$, ce qui implique que $\hat{\beta}_j \sim \mathcal{N}\left(\beta, \sigma_\varepsilon^2(X^tX)^{-1}_{jj}\right)$.
La variable aléatoire $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma_\varepsilon^2(X^tX)^{-1}_{jj}}}$ est distribuée selon la loi $\mathcal{N}(0,1)$ et la variable aléatoire $\dfrac{(n-k)}{\sigma_\varepsilon^2}\hat{\sigma}_\varepsilon^2$ est distribuée selon une loi du χ^2_{n-k}.
Ces deux variables aléatoires étant indépendantes, on peut écrire que :
$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma_\varepsilon^2(X^tX)^{-1}_{jj}}} \Big/ \sqrt{\frac{(n-k)\hat{\sigma}_\varepsilon^2}{(n-k)\sigma_\varepsilon^2}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}_\varepsilon^2(X^tX)^{-1}_{jj}}} \sim \text{Student } (n-k).$$
Si on note $t_{n-p-1}(1-\frac{\alpha}{2})$ le $(1-\frac{\alpha}{2})$ quantile de la distribution de Student $(n-k)$, l'intervalle de confiance de $\hat{\beta}_j$ de niveau $1-\alpha$ est alors défini par :
$$IC_{1-\alpha}\left(\hat{\beta}_j\right) = \left[\hat{\beta}_j \pm t_{n-p-1}(1-\tfrac{\alpha}{2})\sqrt{\hat{\sigma}_\varepsilon^2(X^tX)^{-1}_{jj}} \quad \right]$$
2. Pour la région de confiance de $M\beta$, on utilise le fait que $M\hat{\beta} - M\beta \sim \mathcal{N}_m(0, M(X^tX)^{-1}M^t\sigma_\varepsilon^2)$. On peut également réécrire cette expression comme :
$$Q = \frac{1}{m\hat{\sigma}_\varepsilon^2}(M(\hat{\beta}-\beta))^t(M(X^tX)^{-1}M^t)^{-1}M(\hat{\beta}-\beta) \sim \mathcal{F}_{(m,n-p-1)},$$
où $\mathcal{F}_{(m,n-p-1)}$ est le quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à $(m, n-p-1)$ degrés de liberté.
3. En ce qui concerne l'intervalle de confiance pour σ_ε^2, on sait que $\frac{(n-p-1)\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}$ suit une loi du χ^2 à $(n-p-1)$ degrés de liberté. On peut donc définir l'intervalle de confiance de niveau $1-\alpha$ pour σ_ε^2 comme suit :
$$\left[\tfrac{(n-p-1)\hat{\sigma}_\varepsilon^2}{c_{n-p-1}(1-\alpha/2)}, \tfrac{(n-p-1)\hat{\sigma}_\varepsilon^2}{c_{n-p-1}(\alpha/2)}\right],$$
où $c_{n-p-1}(1-\alpha/2)$ est le quantile de niveau $(1-\alpha/2)$ d'une loi du χ^2 à $(n-p-1)$ degrés de liberté.

□

3.3.2 Tests de significativité

a) **Test de significativité globale du modèle**

Ce test global de Fisher nous permet d'étudier la liaison entre Y et les variables explicatives $X_j \quad (j = 0, \ldots, p)$ (Voir [18]) .

On veut tester :

$$\begin{cases} H_0 : & \beta_0 = \beta_1 = \cdots = \beta_p = 0 \\ H_1 : & \exists \beta_j \neq 0, \forall j \in \{j = 0, \cdots\cdots, p\} . \end{cases}$$

Le test de cette hypothèse est basé sur la statistique de Fisher, notée par F :

$$F = \frac{\frac{SC_{reg}}{p}}{\frac{SC_{res}}{n-p-1}} = \frac{MC_{reg}}{MC_{res}}. \tag{3.7}$$

On rejette H_0 si :

$$F \geq f_{p,n-p-1}(1-\alpha),$$

avec $f_{p,n-p-1}(1-\alpha)$ le quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à $(p, n-p-1))$ ddl. (voir [1]).

Remarque.

(i) On peut réécrire la statistique F en fonction de R^2 comme suit :

$$F = \frac{(n-p-1)}{p} \frac{R^2}{1-R^2}.$$

(ii) Dans la plupart des logiciels statistiques, on fournit directement la probabilité critique (p-value). Elle correspond à la probabilité que la loi de Fisher dépasse la statistique calculée F. Ainsi, la règle de décision (rejeter H_0) au risque α devient :

$$p\text{-}value \leq \alpha.$$

b) **Tests de significativité de Student du paramètre du modèle**

Pour le paramètre β_0, l'ordonnée à l'origine

On veut tester l'hypothèse :

$$\begin{cases} H_0 : \beta_j & = 0 \\ H_1 : \beta_j & \neq 0, \end{cases}$$

avec β_j le paramètre associé à la variable explicative X_j.

La statistique de test permettant d'effectuer ce test est :

$$T_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\hat{\sigma}_\varepsilon(\hat{\beta}_j)}.$$

On rejette H_0 au seuil $(1-\alpha)$ si $|T_{\hat{\beta}_j}| \geq t_{n-p-1}(1-\alpha/2)$ où $t_{n-p-1}(1-\alpha/2)$ est le quantile d'ordre $(1-\alpha/2)$ de $\mathcal{T}_{n-p-1}$.

3.3.3 Tableau d'analyse de la variance et coefficient détermination

a) **Tableau d'analyse de la variance**

Bien avant de dresser le tableau d'analyse de la variance pour la régression multiple, nous procédons à la décomposition de la variance. Toutefois, le procédé reste le même que pour le cas simple. Nous avons ainsi l'équation d'ANOVA qui s'écrit comme suit :

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SC_{total}}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SC_{reg}}+\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}_{SC_{res}}.$$

Source de variation	Somme des carrés	Degrés de liberté	Moyenne des carrés
Expliquée par la régression	SC_{reg}	p	$MC_{reg}=\frac{SC_{reg}}{p}$
Résidus	SC_{res}	$n-p-1$	$MC_{res}=\frac{SC_{res}}{n-p-1}$
Total	$SC_{total}=S_{yy}$	$n-1$	

TABLE 3.1 – Tableau d'analyse de la variance pour la régression linéaire multiple

b) **Coefficient de détermination R^2**

Comme dans le cas linéaire simple, le coefficient de détermination est le rapport entre SC_{reg} et SC_{total}. Il représente la proportion de la variance expliquée et est notée par :

$$R^2=\frac{SC_{reg}}{SC_{total}}=1-\frac{SC_{res}}{SC_{total}},$$

$R^2\in[0,1]$.

Effectivement, deux cas peuvent être distingués en fonction de la valeur du coefficient de détermination (R^2). Si R^2 est proche de 1, cela indique une bonne adéquation du modèle de régression aux données. En d'autres termes, la régression est considérée comme meilleure. En revanche, si R^2 est proche de 0, cela signifie que la quantité d'erreurs résiduelles (SC_{res}) est élevée, ce qui suggère que la régression peut être de mauvaise qualité. En résumé, un R^2 proche de 1 est favorable pour une bonne régression, tandis qu'un R^2 proche de 0 suggère une régression insatisfaisante ou peu précise.

Remarque.

Il est aussi important de noter, comme R^2 est le coefficient de détermination, alors la racine carrée de cette quantité qui équivaut à R est le coefficient de corrélation multiple.

b) **Coefficient de détermination ajusté R_a^2**

Il est vrai que R^2 est un indicateur de la qualité d'ajustement des valeurs observées par le modèle mais il a le défaut de ne pas tenir compte du nombre de variables explicatives utilisées dans celui-ci. Dans le cas $n=p+1$, (i.e le nombre de variables explicatives est grand comparativement au nombre d'observations), $R^2=1$. Ou encore, il est géométriquement facile de voir que l'ajout de variables explicatives ne peut que faire croître le coefficient de détermination. Donc un R^2 proche de 1, n'est pas synonyme de bonne qualité de prévision.

Pour pallier à cela, mous définissons donc un R^2 ajusté qui tient compte des degrés de libertés. Ce coefficient est noté par R_a^2, et est défini comme suit :

$$R_a^2=1-\frac{Sc_{res}}{(n-p-1)}\frac{(n-1)}{SC_{total}}=1-\frac{(n-1)}{(n-p-1)}(1-R^2).$$

Remarque.

On a toujours $R_a^2\leq R^2$ et ceci vu que le modèle contient un grand nombre de variables explicatives.

3.3.4 Prévision et Intervalle de prédiction

Rappelons nous que l'un des buts de la régression est de proposer des prévisions pour la variable à expliquer Y lorsque nous avons de nouvelles valeurs de X. Soit une nouvelle valeur $x'_{n+1}=(1,x_{n+1,1},\cdots\cdots,x_{n+1,p})$, nous

voulons prédire y_{n+1} . Or

$$y_{n+1} = x'_{n+1}\beta + \varepsilon_{n+1},$$

avec $\mathbb{E}(\varepsilon_{n+1}) = 0, Var(\varepsilon_{n+1}) = \sigma^2$ et $Cov(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, \cdots\cdots, n$. Nous pouvons prédire la valeur correspondante grâce au modèle ajusté

$$\hat{y}^p_{n+1} = x'_{n+1}\hat{\beta}. \tag{3.8}$$

Intervalle d'estimation, au niveau $1-\alpha$, pour y_{n+1} :

$$\left[\hat{y}^p_{n+1} \pm t_{(n-p-1)}(1-\alpha/2)\hat{\sigma}_\varepsilon\sqrt{x'_{n+1}(X^tX)^{-1}(x'_{n+1})^t}\right]; \tag{3.9}$$

Intervalle de prévision, au niveau $1-\alpha$, pour y_{n+1} :

$$\left[\hat{y}^p_{n+1} \pm t_{(n-p-1)}(1-\alpha/2)\hat{\sigma}_\varepsilon\sqrt{1+x'_{n+1}(X^tX)^{-1}(x'_{n+1})^t}\right]. \tag{3.10}$$

Deux types d'erreurs vont entacher la prévision : la première est due à l'incertitude sur ε_{n+1} et l'autre est due à l'incertitude liée à l'estimation. Calculons la variance de l'erreur de prévision.

$$\begin{aligned} Var(y_{n+1} - \hat{y}^p_{n+1}) &= Var\left(x'_{n+1}\beta + \epsilon_{n+1} - x'_{n+1}\hat{\beta}\right) = \sigma^2 + x'_{n+1}Var(\hat{\beta})x_{n+1} \\ &= \sigma^2\left(1 + x'_{n+1}(X^tX)^{-1}x_{n+1}\right). \end{aligned}$$

Nous retrouvons bien l'incertitude due aux erreurs σ^2 sur laquelle vient s'ajouter l'incertitude d'estimation.

Remarque.
Puisque l'estimateur $\hat{\beta}$ est un estimateur non biaisé de β et l'espérance de ε vaut zéro, les espérances de y_{n+1} et $\hat{y}^p_{n+1}$ sont identiques. La variance de l'erreur de prévision s'écrit :

$$Var\left(y_{n+1} - \hat{y}^p_{n+1}\right) = \mathbb{E}\left[y_{n+1} - \hat{y}^p_{n+1} - \mathbb{E}(y_{n+1}) + \mathbb{E}(\hat{y}^p_{n+1})\right]^2 = \mathbb{E}\left(y_{n+1} - \hat{y}^p_{n+1}\right)^2.$$

Nous voyons donc ici que la variance de l'erreur de prévision est mesurée par l'erreur quadratique moyenne de prévision (EQMP).

3.4 Estimation des paramètres par MV

Soit le modèle de régression multiple suivant :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots\cdots + \beta_p x_{i,p} + \varepsilon_i, \quad \forall i \in \{1, \cdots\cdots, n\}. \tag{3.11}$$

L'écriture matricielle de celui reste la même que lorsque nous utilisons la méthode MCO, où on a :

$$\mathbf{Y} = X\beta + \varepsilon, \tag{3.12}$$

avec

$$\mathbb{E}(y_i) = X\beta \text{ et } Var(y_i) = \sigma^2_\varepsilon.$$

- Densité de probabilité des y_i.
Elle est sous la forme :

$$\begin{aligned} f(y_i|\beta,\sigma_\varepsilon^2) &= \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left(y_i-\beta_j x_{i,j}\right)^2\right\}, \quad \forall j \in \{0,\cdots\cdots,p\} \\ &= \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left(y_i-\sum_{j=0}^{p}\beta_j x_{ij}\right)^2\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left(y_i-\beta x_i\right)^2\right\}. \end{aligned}$$

Maintenant que nous avons la densité, nous pouvons calculer la vraisemblance et la log-vraisemblance.
- La vraisemblance :

$$\begin{aligned} \mathcal{L}(Y|\beta,\sigma_\varepsilon^2) = \prod_{I=1}^{n} f_Y(y_i|\beta,\sigma_\varepsilon^2) &= \left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{n/2} exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i-\sum_{j=0}^{p}\beta_j x_{i,j})^2\right\} \\ &= \left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{n/2} exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\|Y-X\beta\|^2\right\} \\ &= \left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{n/2} exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(Y-X\beta)^t(Y-X\beta)\right\}. \end{aligned}$$

Ainsi la fonction log-vraisemblance s'écrit :

$$log\left[\mathcal{L}(Y|\beta,\sigma_\varepsilon^2)\right] = log\left[\left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{n/2} exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(Y-X\beta)^t(Y-X\beta)\right\}\right].$$

Or

$$log\left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{n/2} = -\frac{n}{2}log(\pi) - \frac{n}{2}log(2\sigma_\varepsilon^2)$$

et

$$log\left(exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(Y-X\beta)^t(Y-X\beta)\right\}\right) = -\frac{1}{2\sigma_\varepsilon^2}(Y-X\beta)^t(Y-X\beta).$$

Donc

$$log\left[\mathcal{L}(Y|\beta,\sigma_\varepsilon^2)\right] = -\frac{n}{2}log(\pi) - \frac{n}{2}log(2\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}(Y-X\beta)^t(Y-X\beta) \tag{3.13}$$

$$= -\frac{n}{2}log(\pi) - \frac{n}{2}log(2\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}(Y^tY - 2X^tY\beta + X^tX\beta^2). \tag{3.14}$$

Estimation des paramètres β et σ_ε^2
Estimer les paramètres β et σ_ε^2 par MV, consiste à maximiser la fonction log-vraisemblance ci-dessus (3.14). Ce qui revient annuler les dérivées premières par rapport aux arguments β et σ_ε^2 comme suit :
posons $L = \mathcal{L}(Y|\beta,\sigma_\varepsilon^2)$

$$\begin{cases} \dfrac{\partial logL}{\partial\beta} = -\dfrac{1}{2\sigma_\varepsilon^2}(-2X^tY + 2X^tX\beta) = 0; & \text{(3.15a)} \\ \dfrac{\partial logL}{\partial\sigma_\varepsilon^2} = \dfrac{\partial}{\partial\sigma_\varepsilon^2}\left(-\dfrac{n}{2}log(2\sigma_\varepsilon^2)\right) + \dfrac{\partial}{\partial\sigma_\varepsilon^2}\left(-\dfrac{1}{2\sigma_\varepsilon^2}(Y^tY - 2X^tY\beta + X^tX\beta^2)\right) = 0. & \text{(3.15b)} \end{cases}$$

Ainsi de l'équation (3.15a) on a :

$$\tilde{\beta} = (X^tX)^{-1}X^tY. \tag{3.16}$$

À partir de l'équation (3.15b), nous obtenons :

$$\frac{\partial logL}{\partial \sigma_\varepsilon^2} = 0 \Rightarrow -\frac{n}{2\sigma_\varepsilon^2} + \frac{1}{2\sigma_\varepsilon^4}(Y - X\beta)^t(Y - X\beta) = 0.$$

Ce qui, si on remplace β par $\tilde{\beta}$,

$$\tilde{\sigma}_\varepsilon^2 = \frac{(Y - X\tilde{\beta})^t(Y - X\tilde{\beta})}{n}.$$

Donc

$$\tilde{\sigma}_\varepsilon^2 = \frac{\tilde{\varepsilon}^t\tilde{\varepsilon}}{n} = \frac{1}{n}\sum_{I=1}^{n}\tilde{\varepsilon}_i^2. \tag{3.17}$$

D'après les équations (3.16) et (3.17), on note que : l'estimateurs MCO de β et celui du maximum de vraisemblance sont égaux : $\hat{\beta} = \tilde{\beta}$. Par contre l'EMV $\tilde{\sigma}_\varepsilon^2$ de la variance de l'erreur et celui du MCO sont différent : $\tilde{\sigma}_\varepsilon^2 \neq \hat{\sigma}_\varepsilon^2$.

3.5 Comparaison des deux méthodes d'estimation

Les résultats restent les mêmes. En effet, sachant que pour le paramètre β, peu importe ces deux méthodes, nous avons vu que celui-ci est sans biais. Donc on a :

$$Var(\hat{\beta}) = Var(\tilde{\beta}) = \sigma^2\left(X^tX\right)^{-1} \text{ et } \mathbf{E.Q.M}(\tilde{\beta}) = \mathbf{E.Q.M}(\hat{\beta}) = \sigma^2\left(X^tX\right)^{-1}.$$

Pour ce qui est des estimateurs $\hat{\sigma}_\varepsilon^2$ et $\tilde{\sigma}_\varepsilon^2$ on obtient des résultats différents .
Pour l'estimateur MCO de la variance σ_ε^2, on a :

$$\hat{\sigma}_\varepsilon^2 = \frac{\hat{\varepsilon}^t\hat{\varepsilon}}{n-p-1} \rightarrow \mathbb{E}(\hat{\sigma}_\varepsilon^2) = \sigma_\varepsilon^2 \text{ et } Var(\hat{\sigma}_\varepsilon^2) = \frac{2\sigma_\varepsilon^4}{n-p-1}$$

et une erreur quadratique moyenne :

$$\mathbf{E.Q.M}(\hat{\sigma}_\varepsilon^2) = Var(\hat{\sigma}_\varepsilon^2) + \underbrace{b^2(\hat{\sigma}_\varepsilon^2)}_{0} = \frac{2\sigma_\varepsilon^4}{n-p-1}.$$

Pour l'estimateur MV de la variance σ_ε^2, on a :

$$\tilde{\sigma}_\varepsilon^2 = \frac{\tilde{\varepsilon}^t\tilde{\varepsilon}}{n-p-1} \rightarrow \mathbb{E}(\tilde{\sigma}_\varepsilon^2) = \frac{n-p-1}{n}\sigma_\varepsilon^2 \text{ et } Var(\tilde{\sigma}_\varepsilon^2) = \frac{2\sigma_\varepsilon^4}{n}$$

et une erreur quadratique moyenne :

$$\mathbf{E.Q.M}(\tilde{\sigma}_\varepsilon^2) = Var(\tilde{\sigma}_\varepsilon^2) + \underbrace{b^2(\tilde{\sigma}_\varepsilon^2)}_{-\frac{(p+1)}{n}\sigma_\varepsilon^2} = \frac{2\sigma_\varepsilon^4}{n} + \left(-\frac{(p+1)\sigma_\varepsilon^2}{n}\right)^2.$$

Celle-ci se réécrit : $\mathbf{E.Q.M}(\tilde{\sigma}_\varepsilon^2) = \left(\frac{2}{n} + \frac{(p+1)^2}{n^2}\right)\sigma_\varepsilon^4.$

Remarque.
*On ne peut pas savoir laquelle des deux **E.Q.M** est toujours plus grande que l'autre, avant de les avoir calculées, sauf dans les cas où $p+1 \leq 2$ (exemple du modèle simple) où $\mathbf{E.Q.M}(\tilde{\sigma}_\varepsilon^2) < \mathbf{E.Q.M}(\hat{\sigma}_\varepsilon^2)$.*
Cependant on peut montrer que $\mathbf{E.Q.M}(\hat{\sigma}_\varepsilon^2) < \mathbf{E.Q.M}(\tilde{\sigma}_\varepsilon^2)$ si $n > (p+1)^2/(p-1)$.

Ainsi, on a le tableau récapitulatif suivant :

Estimateurs	$\hat{\beta}$	$\hat{\sigma}_\varepsilon^2$	$\tilde{\beta}$	$\tilde{\sigma}_\varepsilon^2$
biaisé/ non biaisé	non biaisé	non biaisé	non biaisé	biaisé
Variance	$\sigma_\varepsilon^2 \left(X^t X\right)^{-1}$	$\frac{2\sigma_\varepsilon^4}{n-p-1}$	$\sigma_\varepsilon^2 \left(X^t X\right)^{-1}$	$\frac{2\sigma_\varepsilon^4}{n}$
E.Q.M	$\sigma_\varepsilon^2 \left(X^t X\right)^{-1}$	$\frac{2\sigma_\varepsilon^4}{n-p-1}$	$\sigma_\varepsilon^2 \left(X^t X\right)^{-1}$	$\left(\frac{2}{n} + \frac{(p+1)^2}{n^2}\right) \sigma_\varepsilon^4$

TABLE 3.2 – Tableau comparatif des estimateurs MCO et MV pour la régression linéaire multiple

Chapitre 4

Simulations numériques

Dans les chapitres précédents, nous avons exploré les outils probabilistes et statistiques pour l'étude de la régression linéaire, qu'elle soit simple ou multiple. Nous avons examiné la détermination des estimateurs et leur comparaison selon les méthodes MCO ou MV. Maintenant, nous aborderons les simulations.
Nous commencerons par introduire les concepts clés de la régression linéaire simple dans la première partie, en considérant une variable à expliquer quantitative Y et une variable explicative X, qui peut être quantitative ou qualitative. Dans la deuxième partie, nous présenterons le modèle de régression linéaire multiple pour étudier la relation entre une variable dépendante quantitative Y et plusieurs variables explicatives $X_1, X_2, \ldots, X_n$, pouvant être quantitatives ou qualitatives.

4.1 Régression linéaire simple

Dans cette section, nous utiliserons l'exemple 1 qui concerne la pression artérielle comme base de données.

4.1.1 Exemple 1 : Pression artérielle systolique

Pour le contexte (voir [9])
La pression artérielle systolique est la pression maximale du sang dans les artères au moment de la contraction du cœur. Celle-ci a été mesurée pour 29 individus de différents âges. Ainsi, pour chacun d'entre eux, on dispose :
– de leur pression systolique en mmHg (variable Y),
– de leur âge en années (variable $X1$).
On souhaite expliquer Y à partir de $X1$. Pour ce faire on utilise le logiciel R et ses commandes.
- Affichage des 6 premières valeurs de la base.

X1	39	45	47	65	46	67
Y	144	138	145	162	142	170

- Analyse du nuage des points et droite de régression
On trace le nuage des points $\{(x_i, y_i), i \in \{1, \cdots\cdots, n\}\}$

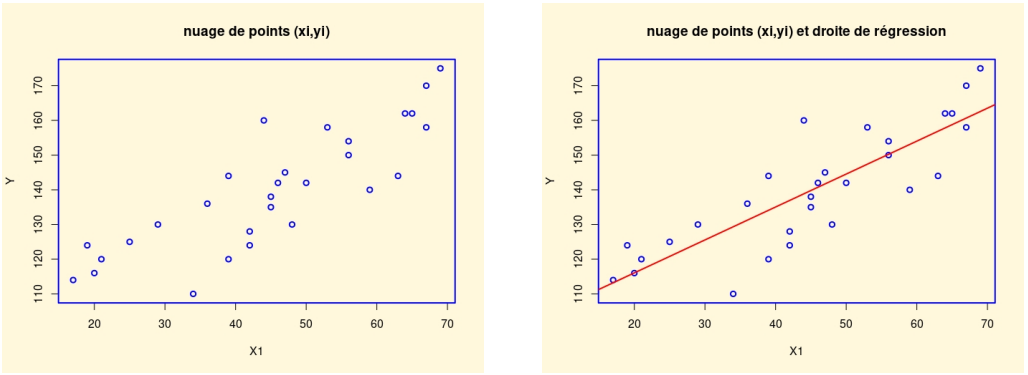

FIGURE 4.1 – Nuage des points (x_i, y_i) et droite de régression $y = 97.0771 + 0.9493x$.

On constate qu'une liaison linéaire entre Y et X1 est envisageable.
En utilisant les estimateurs ponctuels de β_0 et β_1 l'équation de la droite de régression est :

$$y = 97.0771 + 0.9493x.$$

-Modélisation
Une première approche est de considérer le modèle de *rls* :

$$Y = \beta_0 + \beta_1 X1 + \varepsilon,$$

où β_0 et β_1 sont deux coefficients inconnus, et $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ avec σ inconnu.
La présence de β_0 est justifiée car même un très jeune individu peut avoir une pression artérielle systolique élevée.
Objectif : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

4.1.1.1 Calcul des estimateurs MCO

À l'aide des commandes de R, on a :

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	97.0771	5.5276	17.56	0.0000	$***$
X1	0.9493	0.1161	8.17	0.0000	$***$

-Estimation ponctuelle de β_0 et β_1 et leurs écart-types :

$\hat{\beta}_0$	$\hat{\beta}_1$
97.0771	0.9493

$\hat{\sigma}_\varepsilon(\hat{\beta}_0)$	$\hat{\sigma}_\varepsilon(\hat{\beta}_1)$
5.5276	0.1161

– t_{obs} :

H_1	$\beta_0 \neq 0$	$\beta_1 \neq 0$
t_{obs}	17.56	8.17

-Test de Student pour β_1 : influence se $X1$ sur Y : p-value $< 0;001$, $***$: hautement significative.
- $R^2 = 0.7122$ et $\bar{R}^2 = 0.7015$: cela est satisfait.
-Test de Fisher : p-value $= 8.876e - 09 < 0.001$, $***$: l'utilisation du modèle de rls est pertinente.
La valeur prédite de Y quand $X1 = 8$ (par exemple) est donnée par :
Cela renvoie : 104.6717
Ainsi, la pression artérielle systolique moyenne d'un enfant de 8 ans est de 104.6717 mmHg.
On peut aussi s'intéresser :

- aux intervalles de confiance (I.C) pour β_0 et β_1 au niveau 95%.
On obtient :

i_{β_0}	i_{β_1}
$[85.7354850, 108.418684]$	$[0.7110137, 1.187631]$

- à l'intervalle de prédiction pour la valeur moyenne de Y quand $X1 = 8$ (par exemple), on a :

fit	lwr	upr
104.6717	95.11582	114.2275

.

La première valeur est celle de la valeur prédite de Y quand $X1 = 8$, les deux autres correspondent aux bornes inférieures et supérieures des intervalles de confiance de cette valeur.

- **Analyse des résidus**

À présent nous essayons de valider les hypothèses en faisant une analyse graphique des résidus, indépendances de ε et $X1$, indépendances des ε_I, égalités des variances des ε_i.

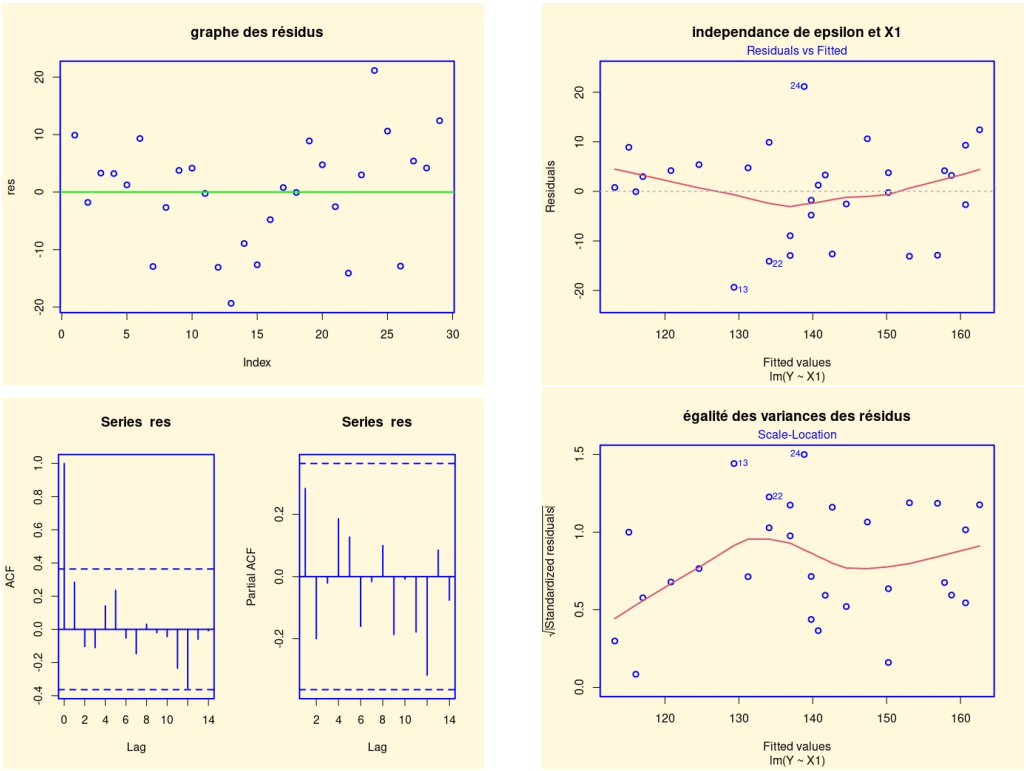

FIGURE 4.2 – Graphes d'analyse des résidus

Analyse graphique des résidus : le graphe des résidus nous montre une relative symétrie des résidus par rapport à l'axe des abscisses et pas de structures évidentes. Cela est encourageant pour la validation des hypothèses standards du modèle rls .

Indépendance de ϵ et $X1$: on trace le nuage de points (résidus, prédictions en $x_{1,i}$).

On constatera que le nuage de points obtenu n'est pas ajustable par une "ligne" et la moyenne des valeurs de la ligne rouge est quasi nulle ; on admet que ε et $X1$ sont indépendantes.

Indépendances de $\varepsilon_1, \cdots\cdots, \varepsilon_n$: les observations de $(Y, X1)$ portent sur des individus tous différents, il doit donc y avoir indépendance de $\varepsilon_1, \cdots\cdots, \varepsilon_n$. On vérifie cela avec les graphiques acf et $pacf$. On ne constatera aucune structure particulière et peu de bâtons dépassent les bornes limites ; on admettra l'indépendance de $\epsilon_1, \cdots\cdots, \epsilon_n$.
(voir figure ci-dessous).

Égalité des variances de $\varepsilon_1, \cdots\cdots, \varepsilon_n$. : on ne constate pas de structure particulière, ce qui traduit une égalité des

variances.
On étudie celle-ci avec le test de Breusch-Pagan :

cela donne : p-value $= 0.7636$. Comme p-value > 0.05, on admet l'égalité des variances.
- **Normalité de $\epsilon_1, \cdots\cdots, \epsilon_n$ et calcul des distances de Cook**
On trace le QQ associé :

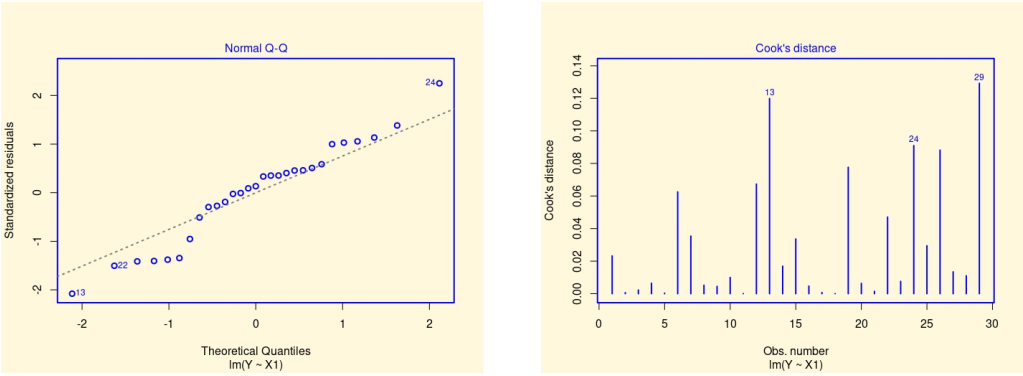

FIGURE 4.3 – Normalité de $\epsilon_1, \cdots\cdots, \epsilon_n$ et calcul des distances de Cook

On constate que les points sont à peu près alignés, ce qui traduit la normalité de$\varepsilon_1, \cdots\cdots, \varepsilon_n$.
De plus nous remarquons aussi qu'aucune valeur ne dépasse 1, il n'y a pas de valeur anormale a priori.

On peut vérifier cela avec le test de Shapiro-Wilk :
Cela renvoie : p-value $= 0.38$. Comme p-value > 0.05, on admet la normalité de $\varepsilon_1, \cdots\cdots, \varepsilon_n$.

On peut compléter notre étude en essayant de détecter les valeurs anormales, en calculant le AIC (**Akaike Information Criterion**) et BIC (**Bayesian Information Criterion**) du modèle.
AIC et BIC
En complément de $\bar{R}^2$, calculons le AIC et le BIC du modèle.
Cela renvoie respectivement : 217.1864 et 221.2883.
Conclusion
L'étude statistique mise en œuvre montre que le modèle de rls est adapté au problème ; les hypothèses permettant la validation des principaux résultats d'estimation sont vérifiées.

4.1.1.2 Calcul des estimateurs par la méthode MV

On se rappel que le modèle rls s'écrit :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \cdots\cdots, n.$$

Par suit, la fonction log-vraisemblance correspondante est :

$$log\left[\mathcal{L}(\beta_0, \beta_1, \sigma_\varepsilon^2)\right] = -\frac{n}{2} log\sigma_\varepsilon^2 - \tfrac{n}{2} log(2\pi) - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Et comme estimer les paramètres β_0, β_1, et σ_ε^2 par MV, consiste à maximiser la fonction log-vraisemblance ci-dessus, alors ceci revient à annuler les dérivées premières par rapport aux arguments β_0, ,β_1, et σ_ε^2 .

On a le résultat tant attendu qui est :

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x1}, \tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x1_i y_i - n\bar{x1}\bar{x1}}{\sum_{i=1}^{n} x1_i^2 - n\bar{x1}^2} = \frac{S_{xy}}{S_{xx}} \text{ et } \tilde{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{I=1}^{n}\left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x1_i\right)^2 = \frac{1}{n}\sum_{I=1}^{n}\tilde{\varepsilon}_i^2.$$

Les résultats de simulations sont :
les estimateurs et écart-types de nos estimateurs après compilation ci-dessous.

	Estimate	Std.Error	t value	Pr(>|t|)	
b0	97.03791	5.33068	18.2037	$< 2.2 \times 10^{-16}$	$***$
b1	0.95010	0.11201	8.4824	$< 2.2 \times 10^{-16}$	$***$
sigma	9.22271	1.21002	7.6219	2.499×10^{-14}	$***$

$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\sigma}_\varepsilon$
97.03791	0.95010	9.22271
$\tilde{\sigma}_\varepsilon(\tilde{\beta}_0)$	$\tilde{\sigma}_\varepsilon(\tilde{\beta}_1)$	$\tilde{\sigma}_\varepsilon(\tilde{\sigma}_\varepsilon)$
5.33068	0.11201	1.21002

4.1.2 Comparaison des estimateurs suivant chacune des deux méthodes d'estimations

Comme nous l'avons pu constater théoriquement au chapitre 2, les estimateurs $\hat{\beta}$ et $\tilde{\beta}$ sont égaux et que la différence semble être observée avec les estimateurs $\hat{\sigma}_\varepsilon$ et $\tilde{\sigma}_\varepsilon$.
À présent, calculons la variance, le biais et l'erreur quadratique moyenne de chacun des estimateurs. Nous obtenons :

Estimateurs	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}_\varepsilon^2$	$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\sigma}_\varepsilon^2$
Biais	2.93565×10^{-16}	2.2045×10^{-16}	4.263256×10^{-14}	1.573422×10^{-05}	2.253733×10^{-05}	6.298751
Variance	30.55383	0.01348955	619.5877	30.55383	0.01348955	500.0621
E.Q.M	30.55383	0.01348955	619.5877	30.55383	0.01348955	539.7363

TABLE 4.1 – Tableau comparatif des estimateurs MCO et MV pour la régression linéaire simple

Le tableau 4.1 montre clairement les différences théoriques. Les estimateurs $\hat{\beta}_j$ et $\tilde{\beta}_j$, où j est soit 0 ou 1, sont presque identiques. En revanche, les estimateurs $\hat{\sigma}_\varepsilon^2$ et $\tilde{\sigma}_\varepsilon^2$ ont des variances et des erreurs quadratiques moyennes différentes. L'estimateur de la variance pour l'erreur aléatoire du modèle est plus petit avec la méthode MV qu'avec la méthode MCO, ce qui signifie que $\mathbf{E.Q.M}(\tilde{\sigma}_\varepsilon^2) < \mathbf{E.Q.M}(\hat{\sigma}_\varepsilon^2)$.

L'efficacité relative des deux estimateurs, $\hat{\sigma}_\varepsilon^2$ et $\hat{\sigma}_\varepsilon^2$, est donc de 0.87, calculée comme le rapport des erreurs quadratiques moyennes :

$$eff(\hat{\sigma}_\varepsilon, \tilde{\sigma}_\varepsilon) = \frac{\mathbf{E.Q.M}(\tilde{\sigma}_\varepsilon)}{\mathbf{E.Q.M}(\hat{\sigma}_\varepsilon)} = 0.87.$$

4.2 Régression linéaire multiple

Ici, nous utiliserons comme base de données l'exemple 2, à savoir la vente d'horloges anciennes.

4.2.1 Exemple 2 vente d'horloges anciennes

Pour le contexte (voir[9],[10])
Dans une vente aux enchères, 32 horloges anciennes différentes ont trouvé preneur. Pour chacune d'entre elles, on dispose :
– du prix de vente en pounds sterling (variable Y), (1 pound = 1.2553 euros),
– de l'âge de l'horloge en années (variable $X1$),
– du nombre de personnes qui ont fait une offre sur celle-ci (variable $X2$).

On souhaite expliquer Y à partir de $X1$ et $X2$. Pour ce faire on utilise R et ses commandes.
-Affichage de l'entête de la base.

	X1	**X2**	**Y**
1	127	13	1235
2	115	12	1080
3	127	7	845
4	150	9	1522
5	156	6	1047
6	182	11	1979

-Modélisation
Une première approche est de considérer le modèle de *rlm* :

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \varepsilon,$$

où β_0 , β_1 et β_2 sont 3 coefficients inconnus, et $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ avec σ inconnu.
La présence de β_0 est justifiée car même une horloge récente avec une offre peut avoir un prix non négligeable.
Objectif : Estimer les paramètres inconnus à partir des données et étudier la qualité du modèle.

4.2.1.1 Calcul des estimateurs par MCO

Par le biais du logiciel R, nous avons les sorties suivantes :

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	-1336.7221	173.3561	-7.71	0.0000	$***$
X1	12.7362	0.9024	14.11	0.0000	$***$
X2	85.8151	8.7058	9.86	0.0000	$***$

-Estimation ponctuelle de β_0, β_1 et β_2 et leurs écart-types

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
-1336.7221	12.7362	85.8151

$\hat{\sigma}(\hat{\beta}_0)$	$\hat{\sigma}(\hat{\beta}_1)$	$\hat{\sigma}(\hat{\beta}_1)$
173.3561	0.9024	8.705

– t_{obs} et degrés de significativité :

H_1	$\beta_0 \neq 0$	$\beta_1 \neq 0$	$\beta_2 \neq 0$
t_{obs}	-7.71	14.11	9.86

H_1	$\beta_0 \neq 0$	$\beta_1 \neq 0$	$\beta_2 \neq 0$
degré	$***$	$***$	$***$

- $R^2 = 0.8927$ et $\bar{R}^2 = 0.8853$: cela est tout à fait correct.
-Test de Fisher : p-value < 0.001, $***$: l'utilisation du modèle de rlm est pertinente.
La valeur prédite de Y quand $X1 = 157$ et $X2 = 11$ (par exemple) vaut : 1606.828.
Par conséquent, une horloge qui a 157 ans et sur laquelle 11 personnes ont fait une offre sera vendue, en moyenne, 1606.828 pounds.
On peut aussi s'intéresser :
– aux intervalles de confiance pour β_0 , β_1 et β_2 au niveau 99% (par exemple).
Ainsi, nous obtenons :

i_{β_0}	i_{β_1}	i_{β_2}
$[-1814.55843, -858.88567]$	$[10.24889, 15.22351]$	$[61.81871, 109.81156]$

-à l'intervalle de prévision pour la valeur moyenne de Y quand $X1 = 157$ et $X2 = 11$ (par exemple), nous obtenons :

fit	lwr	upr
1606.828	1545.249	1668.407

La première valeur est celle de la valeur prédite de Y quand $X1 = 157$ et $X2 = 11$, les deux autres correspondent aux bornes inférieures et supérieures des intervalles de confiance de cette valeur. Ainsi, pour ($X1 = 157$,$X2 = 11$) on a :

$$i_{y_x} = [1545.249, 1668.407].$$

Validation des hypothèses

Analyse des nuages de points

On trace les nuages de points des variables par pairs :

On remarquera qu'une liaison linéaire entre Y et $X1$ est effectivement envisageable. C'est un peu moins clair entre Y et $X2$. (voir figure ci-dessous)

Cette analyse amène une vague idée de modélisation ; des méthodes plus rigoureuses seront présentées dans des études futures.

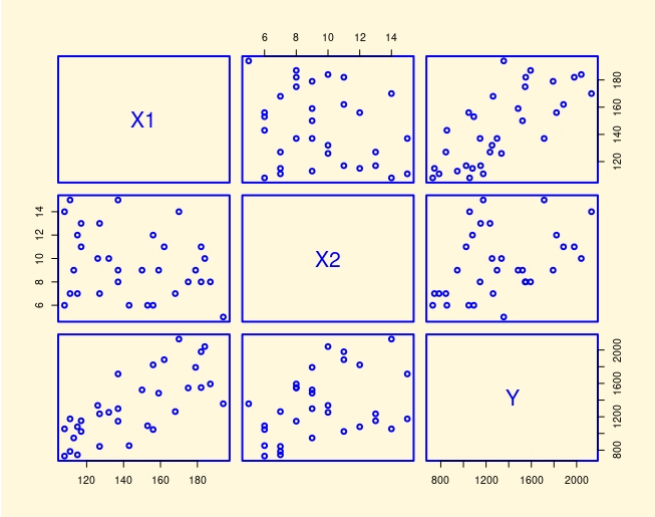

FIGURE 4.4 – Nuages de points des variables par pairs

-Analyse des résidus

Comme avec le modèle rls, nous essayons de valider les hypothèses en en faisant une analyse graphique des résidus, indépendances de ε et $(X1, X2)$, indépendances des ε_I, égalités des variances des ϵ_i et la normalité de ceux-ci

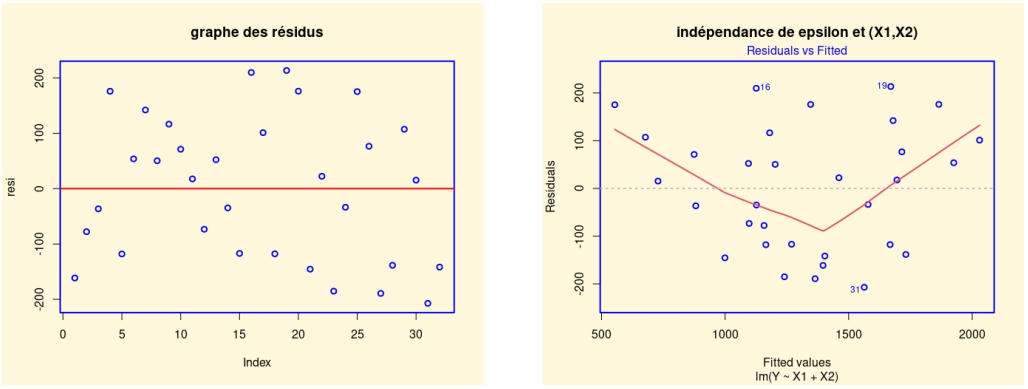

FIGURE 4.5 – Graphe d'analyse des résidus, indépendances de ϵ et $(X1, X2)$ et indépendances des ϵ_I

Graphe des résidus : on constate une relative symétrie des résidus par rapport à l'axe des abscisses et pas de structure évidente. Cela est encourageant pour la validation des hypothèses standards du modèle de *rlm*.
-Indépendance de ε **et** $X1$,$X2$: une fois tracer le nuage de points (résidus, prédictions en $(x_{1,i},x_{2,i})$), on constate que le nuage de points obtenu n'est pas ajustable par une "ligne" et la moyenne des valeurs de la ligne rouge est quasi nulle ; on admet que ϵ et $X1$,$X2$ sont indépendantes.
-Indépendances de $\varepsilon_1, \cdots\cdots, \varepsilon_n$: les observations de $(Y, X1, X2)$ portent sur des horloges toutes différentes, il doit donc y avoir indépendance de $\varepsilon_1, \cdots\cdots, \varepsilon_n$. On vérifie cela avec les graphiques acf et $pacf$ ci dessus .

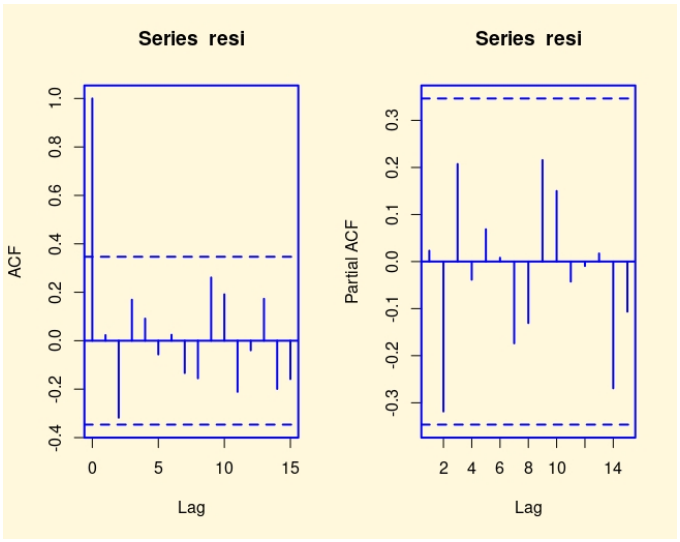

FIGURE 4.6 – Graphe de l'indépendance des ε_i

On ne constate aucune structure particulière (et pas de bâtons qui dépassent les bornes limites, à part le premier, ce qui est normal) ; on admettra l'indépendance de $\epsilon_1, \cdots\cdots, \epsilon_n$.
Égalité des variances de $\varepsilon_1, \cdots\cdots, \varepsilon_n$**. et leur normalité**
Une indication graphique sur l'égalité des variances de $\varepsilon_1, \cdots\cdots, \varepsilon_n$ est donnée par :

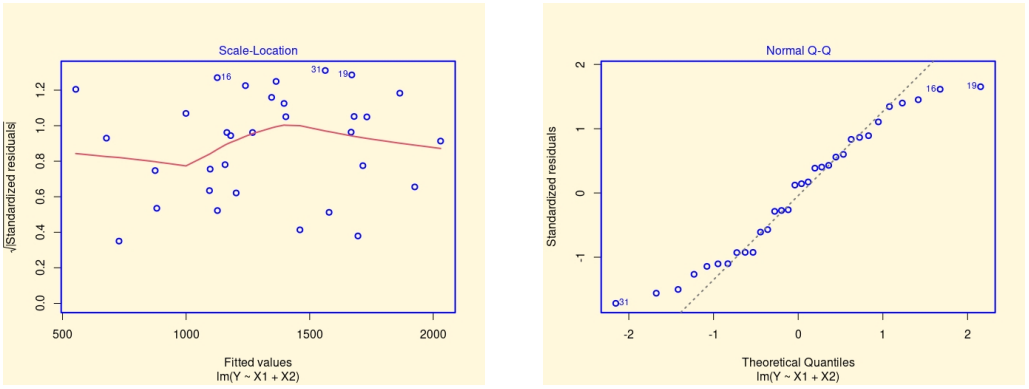

FIGURE 4.7 – Graphes d'analyse des résidus

On ne constate pas de structure particulière, ce qui traduit une égalité des variances.
On étudie celle-ci avec le test de Breusch-Pagan :

cela renvie : p-value $= 0.8038$. Comme p-value > 0.05, on admet l'égalité des variances.

-Normalité de $\varepsilon_1, \cdots\cdots, \varepsilon_n$

On trace le QQ associé.

On constate que les points sont à peu près alignés, ce qui traduit la normalité de $\varepsilon_1, \cdots\cdots, \varepsilon_n$.

On peut vérifier cela avec le test de Shapiro-Wilk.

Cela renvoie : $p - valeur = 0.1215$. Comme $p - valeur > 0.05$, on admet la normalité de $\varepsilon_1, \cdots\cdots, \varepsilon_n$.

Compléments

-Étude de la multicolinéarité

On calcule le carré du coefficient de corrélation entre $X1$ et $X2$.

Cela renvoie : 0.06438861, lequel est éloigné de $R^2 = 0.8811$. Donc, par la règle de Klein, il n'y a pas de lien linéaire entre $X1$ et $X2$.

On peut obtenir la même conclusion avec les `vif` :

V_1	V_2
1.06882	1.06882

Comme les vif sont inférieurs à 5, il n'y a pas de lien linéaire entre $X1$ et $X2$.

-Détection des valeurs anormales

On étudie les distances de Cook des observations :

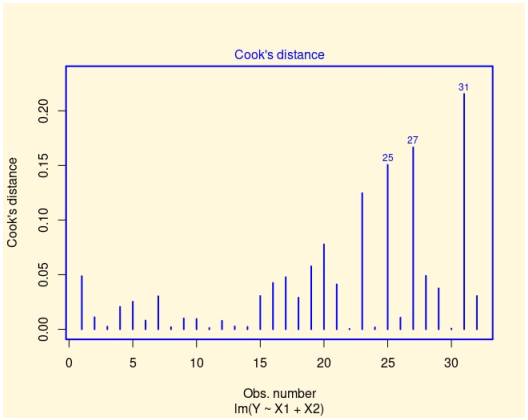

FIGURE 4.8 – Graphe des distances de Cook

Aucune d'entre elles ne dépasse 1, il n'y a pas de valeur anormale a priori.

AIC et BIC

En complément de $\bar{R}^2$, calculons le AIC et le BIC du modèle.

Cela renvoie : 408.71 et 414.5729 comme résultats.

Conclusion

L'étude statistique mise en œuvre montre que le modèle de rlm est adapté au problème ; les hypothèses permettant la validation des principaux résultats d'estimation sont vérifiées.

4.2.2 Calcul des estimateurs par la méthode MV

Comme nous avons estimé les paramètres $\beta = (\beta_0, \beta_1, \beta_2)$ et σ_ε par MCO précédemment, maintenant nous utilisons la méthode (MV) pour déterminer les estimateurs associés. De plus, sachant que X est la matrice composée des valeurs des variables $X1$ et $X2$, alors on a :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i, i = 1, \cdots\cdots, n.$$

Par suit, la fonction log-vraisemblance correspondant est :

$$log\left[\mathcal{L}(Y|\beta, \sigma_\varepsilon^2)\right] = log\left[\left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{n/2} exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(Y - X\beta)^t(Y - X\beta)\right\}\right].$$

ET on obtient le système d'équations ci-dessous.

$$\begin{cases} \dfrac{\partial logL}{\partial \beta} = -\dfrac{1}{2\sigma_\varepsilon^2}(-2X^tY + 2X^tX\beta) = 0; & (4.1a) \\ \dfrac{\partial logL}{\partial \sigma_\varepsilon^2} = \dfrac{\partial}{\partial \sigma_\varepsilon^2}\left(-\dfrac{n}{2}log(2\sigma_\varepsilon^2)\right) + \dfrac{\partial}{\partial \sigma_\varepsilon^2}\left(-\dfrac{1}{2\sigma_\varepsilon^2}(Y^tY - 2X^tY\beta + X^tX\beta^2)\right) = 0. & (4.1b) \end{cases}$$

Ainsi, après calcul, nous obtenons :

$$\tilde{\beta} = (X^tX)^{-1}X^tY \text{ et } \tilde{\sigma}_\varepsilon^2 = \frac{(Y - X\tilde{\beta})^t(Y - X\tilde{\beta})}{n} = \frac{1}{n}\sum_{I=1}^{n}\tilde{\varepsilon}_i^2.$$

À l'aide du logiciel R, nous obtenons les résultats comme suit :

	Estimate	Std.Error	z value	Pr(>|t|)	
(Intercept)	-1336.72209	165.02988	-8.0999	5.501×10^{-16}	$***$
beta.tild1	12.73619	0.85904	14.8261	$< 2.2 \times 10^{-16}$	$***$
beta.tild2	85.81531	8.28762	10.3546	$< 2.2 \times 10^{-16}$	$***$
sigma	126.74200	15.84272	8.0000	1.244×10^{-15}	$***$
—— -2 log L : 400.71					

$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\sigma}_\varepsilon$
-1336.72209	12.73619	85.81531	126.74200
$\tilde{\sigma}_\varepsilon(\tilde{\beta}_0)$	$\tilde{\sigma}_\varepsilon(\tilde{\beta}_1)$	$\tilde{\sigma}_\varepsilon(\tilde{\beta}_2)$	$\tilde{\sigma}_\varepsilon(\tilde{\sigma}_\varepsilon)$
165.02988	0.85904	8.28762	15.84272

4.2.3 Comparaison des estimateurs suivant chacune des deux méthodes d'estimations

À présent, le calcul de la variance, du biais et de l'erreur quadratique moyenne de chacun des estimateurs nous permet de dresser le tableau ci-dessus.

Estimateurs	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}_\varepsilon^2$	$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\sigma}_\varepsilon^2$
Biais	4.216071×10^{-15}	4.134982×10^{-15}	6.280325×10^{-14}	3.637979×10^{-11}	7.551071×10^{-06}	8.594361×10^{-06}	0.0002119444	1652.077
Variances	30052.35	0.8142905	75.7902	20945817	27234.86	0.7379487	68.68468	19615177
EQM	30052.35	0.8142905	75.7902	20945817	27234.86	0.7379487	68.68468	22344534

TABLE 4.2 – Tableau comparatif des estimateurs MCO et MV pour la régression linéaire multiple

Le tableau 4.2 ci-dessus met en évidence les différences théoriques. On peut clairement voir que les estimateurs $\hat{\beta}_j$ et $\tilde{\beta}_j$ (où j est 0, 1, 2 ou 3) sont presque identiques. En revanche, les estimateurs $\hat{\sigma}_\varepsilon^2$ et $\tilde{\sigma}_\varepsilon^2$ ont des variances et des erreurs quadratiques moyennes différentes. L'estimateur de la variance pour l'erreur aléatoire du modèle est plus petit avec la méthode MV qu'avec la méthode MCO, ce qui est normal car cela découle des résultats théoriques lorsque la taille de la base de données n est plus grande que $k/(n-k)$, avec $k = p + 1$.

L'efficacité relative des deux estimateurs, $\hat{\sigma}_\varepsilon^2$ et $\hat{\sigma}_\varepsilon^2$, est donc de 0.94, calculée comme le rapport des erreurs quadratiques moyennes :

$$eff(\hat{\sigma}_\varepsilon, \tilde{\sigma}_\varepsilon) = \frac{\textbf{E.Q.M}(\hat{\sigma}_\varepsilon)}{\textbf{E.Q.M}(\tilde{\sigma}_\varepsilon)} = 0.94.$$

Chapitre 5

Application sur le délai de diagnostic des patients atteints de la tuberculose (TB) pulmonaire

La tuberculose pulmonaire est une maladie infectieuse et contagieuse provoquée par la bactérie Mycobacterium tuberculosis. Elle est largement répandue à l'échelle mondiale, affectant près d'un tiers de la population.
La tuberculose pulmonaire constitue une priorité majeure en termes de santé publique, étant la principale cause de mortalité liée aux maladies infectieuses, notamment dans les pays en développement. Malheureusement, son traitement devient de plus en plus complexe en raison de sa réapparition chez les personnes atteintes du VIH et de l'émergence régulière de souches résistantes aux médicaments.
Au Sénégal, malgré les efforts actifs déployés dans la lutte contre la tuberculose, les statistiques restent préoccupantes. D'après la coordinatrice du programme de lutte contre la tuberculose (PNT), le Dr Yacine Mar Diop, l'organisation mondiale de la santé (OMS) a estimé en 2022 que le nombre de cas s'élevait à 113 pour 100 000 personnes, soit une projection de 19 463 nouveaux cas.
Il est donc crucial de prendre des mesures préventives pour empêcher la transmission de la maladie à des personnes non infectées.
C'est dans cette logique que des médecins et chercheurs du monde entier collaborent pour trouver des traitements plus efficaces, mais aussi pour chercher les facteurs influençant son diagnostic tardif afin de mieux lutter contre et limiter sa transmission (pour plus de détails voir [2], [19], [25], [27]).

5.1 Présentation des données : « délai de diagnostic»

5.1.1 Présentation

Il s'agit d'une enquête menée à l'hôpital La Paix de Ziguinchor (Médecine Interne) entre 2018 et 2023 pour étudier les facteurs qui contribuent à un diagnostic tardif de la tuberculose chez les patients. Nous avons examiné les données de 141 patients qui ont été consultés ou traités dans ce service de santé ; à partir desquelles nous cherchons à identifier les facteurs qui peuvent influencer le délai de diagnostic, (qui est la différence entre la date du diagnostic et la date approximative des premiers symptômes). Nous considérons qu'un délai de diagnostic supérieur à 31 jours (1 mois) est très élevé.

5.1.2 Description de la population d'étude

Nous analysons les variables différemment selon leur nature : quantitative ou qualitative. Les variables quantitatives sont résumées sous forme d'indicateurs (moyenne, écart-type, ...), dans le tableau 5.1, et sont présentées graphiquement sous forme d'histogramme et de boîtes à moustache ou box-plot, sauf pour les variables âge et délai de traitement qui sont en barplot car elles sont regroupées en classe d'âge (resp. délai de traitement) figures 5.1 et 5.2.

Variables	n	Moyenne	Ecart-type	Médiane	Minimum	Maximum
AGEP (ans)	141	40	16.55508	39	16	86
DUAN.TB (jrs)	141	9	4.75811	10	0	20
DPHT (jrs)	141	12	8.147413	12.0	0	30
RVM (FCFA)	141	34660	17865.7	35000	000	75000
*NBP_*CH	141	2	0.8571175	2	0	3
DELAI.DIAG (jrs)	141	36	9.227865	37	15	58
DELAI.TRAIT (jrs)	141	2	1.677976	1	0	6

TABLE 5.1 – Tableau descriptif des variables quantitatives

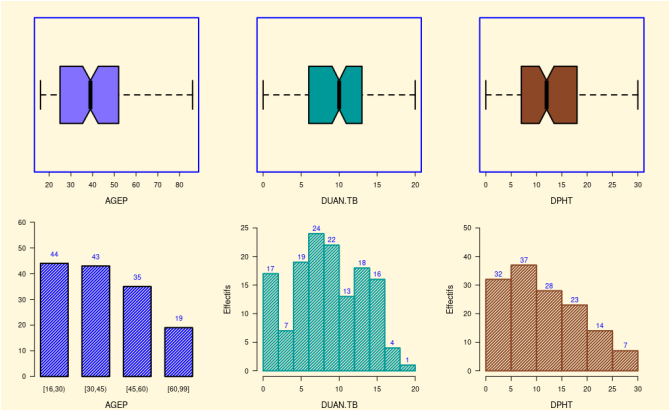

FIGURE 5.1 – Graphes descriptifs 1

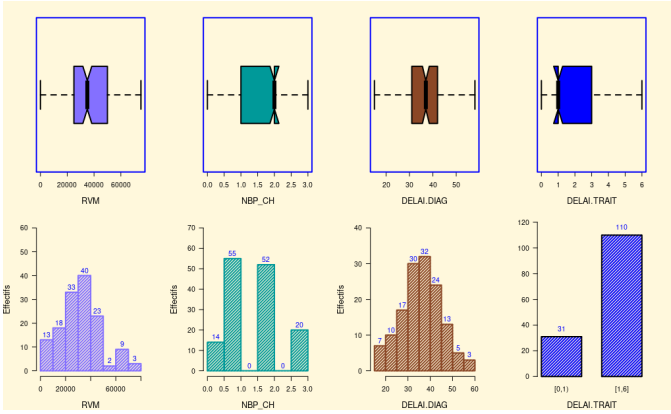

FIGURE 5.2 – Graphes descriptifs 2

Les figures 5.1 et 5.2, représentent la distribution des variables quantitatives : l'âge, l'usage des antibiotiques non TB, l'usage de la phytothérapie, le revenu mensuel, le nombre de personnes avec qui le patient partage la chambre, le délai de diagnostic et le délai de traitement.
Pour les variables qualitatives, on résume les données sous forme de tableau de fréquences (tableaux 5.2 à 5.5) et on les présente graphiquement par des diagrammes en bâtons (Voir Fig. 5.3 à 5.6).

Variables	Modalités	Effectif	Fréquence (%)
SEXE	Masculin(1)	92	65.25
	Féminin (2)	49	34.75
SM	Marié(1)	68	48.2
	Célibataire (2)	55	39.0
	Divorcé (3)	6	4.3
	Veuf(4)	12	8.5
S.GEO	Près(1)	73	51.8
	Loin(2)	68	48.2
CS	Oui(1)	9	6.4
	Non (2)	132	93.6
SECTEURS	Formel(1)	42	29.8
	Informel(2)	99	70.2
NIV1	Analphabète (1)	49	34.8
	Primaire (2)	35	24.8
	Coll/Second(3)	44	31.2
	Universitaire (4)	13	9.2

TABLE 5.2 – Tableau descriptif 1

Variables	Modalités	Effectif	Fréquence (%)
NIV2	Bas(1)	84	59.6
	Élevé(2)	57	40.4
FV	Oui(1)	110	78.
	Non(2)	31	22.0
AMG	Oui(1)	84	59.6
	Non(2)	58	40.4
ASTH	Oui(1)	64	45.4
	Non(2)	77	54.6
TOUX	Oui(1)	127	90.0
	Non(2)	14	10.0S
HPTY	Oui(1)	11	7.8
	Non(2)	130	92.2

TABLE 5.3 – Tableau descriptif 2

Variables	Modalités	Effectif	Fréquence (%)
DTH	Oui(1)	64	45.4
	Non(2)	77	54.6
VAC.BCG	Oui(1)	17	12.0
	Non(2)	107	76.0
	Ignore(3)	17	12.0
NC	Oui(1)	63	44.7
	Non(2)	78	55.3
OC	Familiale(1)	54	38.3
	Milieu Prof(2)	73	51.8
	Voisinage(3)	13	9.2
	Autre(4)	1	0.7
VIH	Oui(1)	19	13.5
	Non(2)	122	86.5
DIABETE	Oui(1)	17	12.1
	Non(2)	124	87.9

TABLE 5.4 – Tableau descriptif 3

Variables	Modalités	Effectif	Fréquence (%)
HTA	Oui(1)	6	4.3
	Non(2)	135	95.7
CANCERS	Oui(1)	1	0.7
	Non(2)	140	99.3
TABAC	Oui(1)	36	25.5
	Non(2)	105	74.5
ALCOOL	Oui(1)	19	13.5
	Non(2)	122	86.5
TXM	Oui(1)	1	0.7
	Non(2)	140	99.3
*ITI_*THERAP1	S.San(1)	20	14.2
	G.tradi(2)	52	36.9
	Ignore	69	48.9

TABLE 5.5 – Tableau descriptif 4

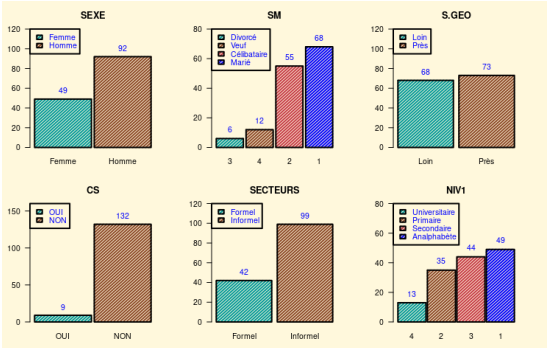

FIGURE 5.3 – Graphes descriptifs 3

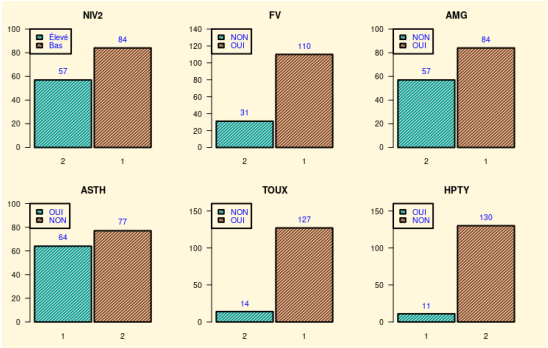

FIGURE 5.4 – Graphes descriptifs 4

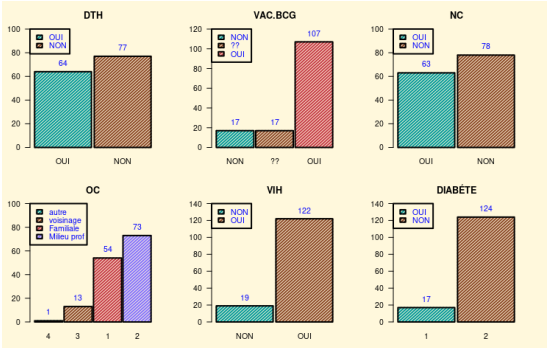

FIGURE 5.5 – Graphes descriptifs 5

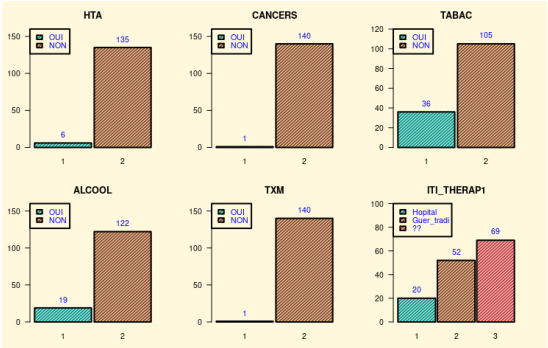

FIGURE 5.6 – Graphes descriptifs 6

Les figures 5.3, 5.4, 5.5 et 5.6 montrent comment les variables qualitatives sont réparties chez nos patients. Ces variables incluent des informations socio-démographiques telles que le sexe et la situation géographique, des symptômes tels que la fièvre, des antécédents cliniques tels que le diabète et les cancers, ainsi que des habitudes de vie telles que la consommation du tabac et d'alcool.

5.1.3 Détermination de la densité de la variable d'étude

Dans cette étude, nous cherchons à comprendre la variation du délai de diagnostic de la TB pulmonaire entre les différents patients en utilisant des modèles linéaires simples et multiples. Afin de déterminer la distribution ou la loi de probabilité de cette variable, il est essentiel de tracer l'histogramme du délai de diagnostic et de générer un graphe QQ-plot. Ensuite, nous procéderons à un test de normalité pour confirmer si la distribution des données suit une distribution normale (voir figure 5.7). Cette étape est cruciale pour nous assurer que les modèles linéaires utilisés pour expliquer la variation du délai de diagnostic sont appropriés et donnent des résultats précis.

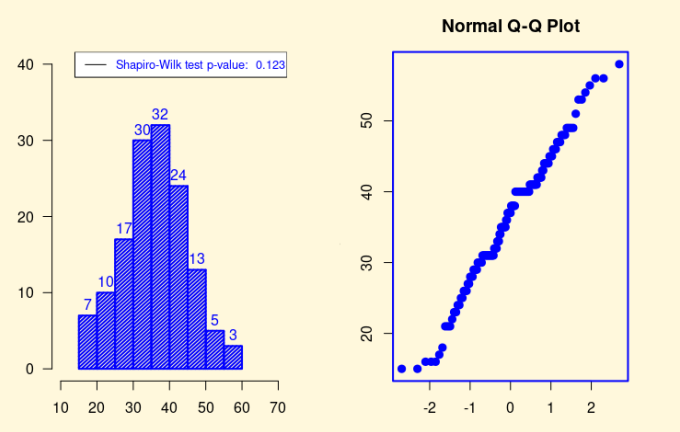

FIGURE 5.7 – Histogramme et graphe QQ-plot de la variable DELAI.DIAG

Nous observons que notre variable d'étude suit une distribution normale. Cette constatation est appuyée par le test de Shapiro-Wilk, qui a donné une valeur de la $p - value$ qui vaut 12,3%, confirmant ainsi notre observation.

Dans ce jeu de données **délai diagnostic**, il s'agit d'expliquer la variabilité du délai de diagnostic de la maladie (tuberculose pulmonaire) en fonction des caractéristiques du patient : ses antécédents et pathologies sous jacentes, son niveau socio-économique et d'étude pour ne citer que ceux là. La variable à expliquer est le délai de diagnostic (variable quantitative **DELAI.DIAG**, exprimée en jours) et les facteurs étudiés (variables explicatives).

5.2 Régression linéaire simple

5.2.1 Analyse du nuage de points $DELAI.DIAG = f(DPHT)$

On trace le nuage de points $\{(x_i, y_i), i \in \{1, \cdots\cdots, 141\}\}$

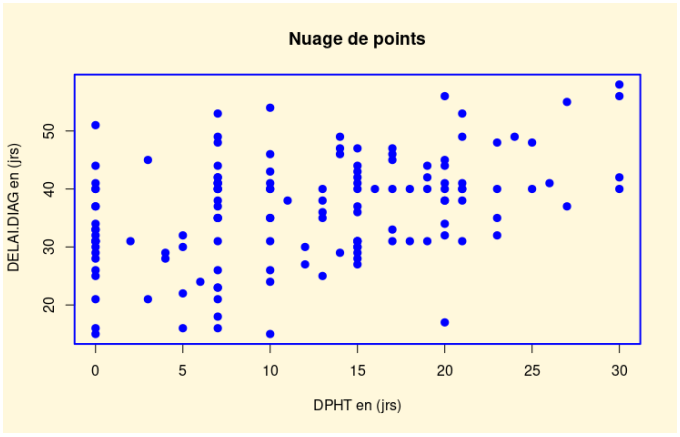

FIGURE 5.8 – Nuage de points du délai de diagnostic la durée de l'utilisation de la phytothérapie.

Nous remarquons bel et bien, qu'il existe un lien entre la durée de l'usage de la phytothérapie et le délai de diagnostic de la maladie.

5.2.2 Modèle de régression linéaire simple

On se donne pour objectif d'expliquer le délai de diagnostic de la TB pulmonaire, noté $Y = DELAI.DIAG$ par une variable explicative quantitative la durée de l'usage de la phytothérapie, noté $X = DPHT$.
Ainsi on a le modèle de régression linéaire simple (rls) qui s'écrit sous la forme :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \qquad i \in \{1, \cdots\cdots, 141\},$$

où ε représente l'erreur tel que $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ et β_0, β_1 et σ_ε^2 les inconnus de la régression.

5.2.3 Hypothèses relatives aux modèles de la rls

a-) **Résidus**
Pour tout $i \in \{1, \cdots\cdots, 141\}$, on appelle i-ème résidus la réalisation $\hat{\varepsilon}_i$ de :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i.$$

Ces résidus vont nous permettre de valider ou non les hypothèses initiales du modèle de la régression linéaire simple (rls).

b-) **Hypothèse de la linéarité du modèle**
En tentant de déterminer si le modèle de régression est adéquat, nous cherchons à savoir si la relation entre Y et X est linéaire. Pour ce faire, nous examinons un graphique qui représente les résidus en fonction des valeurs ajustées. Si le tracé rouge sur le graphique est approximativement horizontal, cela signifie qu'il y a effectivement une relation linéaire entre Y et X.

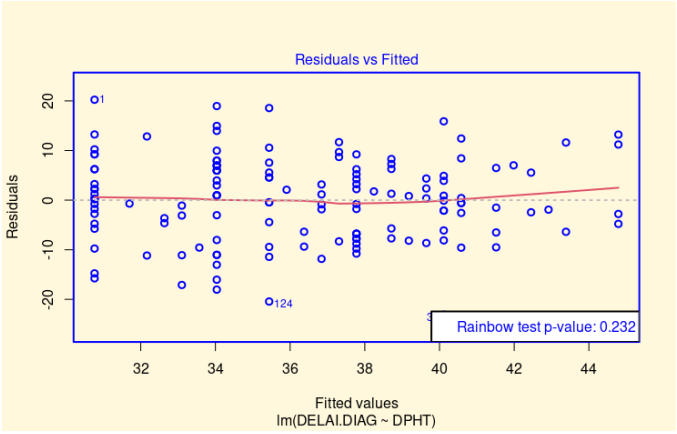

FIGURE 5.9 – Graphe des résidus en fonction des valeurs ajustées

Ici, le tracé rouge est approximativement horizontal. La régression linéaire semble donc être adaptée.
De plus, la p-value du test de **Rainbow** est supérieure à 0.05. Donc l'hypothèse de linéarité est vérifiée.

c-) **Indépendances de** $\varepsilon_1, \cdots\cdots, \varepsilon_{141}$

Les observations de (Y, X) portent sur des individus tous différents, il est donc normal que $\varepsilon_1, \cdots\cdots, \varepsilon_{141}$ soient indépendantes. On examine cela avec le graphique acf et le test de Ljung-Box.

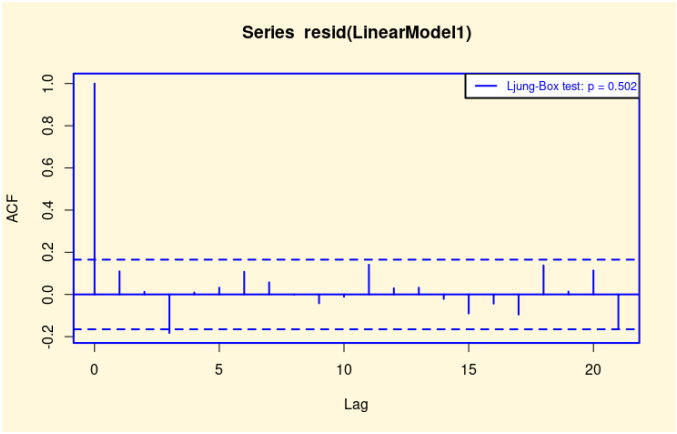

FIGURE 5.10 – Graphe acf et le test de Lijung-Box des résidus

On ne constate aucune structure particulière et peu de bâtons dépassent les bornes limites, et le test de Lijung-Box donne une p-value supérieure à 5%, confirmant l'indépendance des erreurs.

d-) **Test d'hétéroscédasticité (égalité des variances des erreurs)**

Pour verifier l'hypothèse d'hétéroscédasticité ou égalité des variances de $\varepsilon_1, \cdots\cdots, \varepsilon_{141}$, nous analysons le graphe ci-dessous et la valeur de la p-value du test de **Goldfeld-Quandt**.

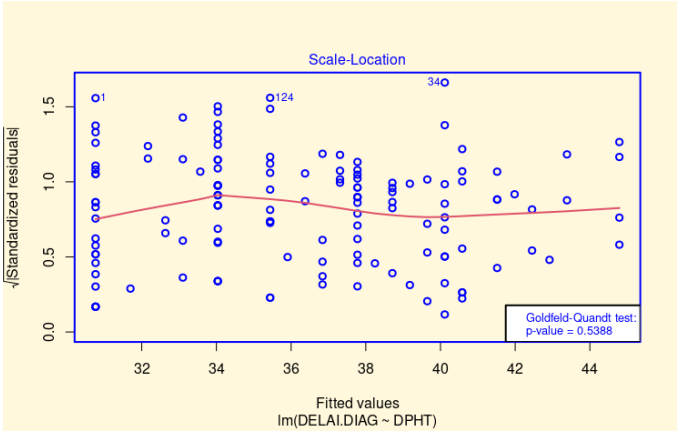

FIGURE 5.11 – Graphe d'égalité des variances

On ne constate pas de structure particulière au niveau de la figure et la p-value du test de **Goldfeld-Quandt** ($= 0.5388 > \alpha = 0.05$), ce qui traduit une égalité des variances.

e-) **Test de normalité des erreurs**

On admet que $\varepsilon_1, \cdots\cdots, \varepsilon_{141}$ soient indépendantes et $\sigma^2_\varepsilon(\varepsilon_1) = \cdots = \sigma^2_\varepsilon(\varepsilon_{141})$.

Pour étudier la normalité de $\varepsilon_1, \cdots\cdots, \varepsilon_n$, on trace le nuage de points **QQ-plot** associé (ou diagramme Quantile-Quantile). Si le nuage de points est bien ajusté par la droite $y = x$, on admet la normalité de $\varepsilon_1, \cdots\cdots, \varepsilon_{141}$. Ceci est aussi justifié en effectuant le test de Shapiro-Wilk où si la p-value associé est supérieure au seuil $\alpha = 0.05$, on dit que la normalité est vérifiée.

À ce diagramme Quantile-Quantile, on peut ajouter le trace de l'histogramme et de la densité des $\varepsilon_1, \cdots\cdots, \varepsilon_n$.

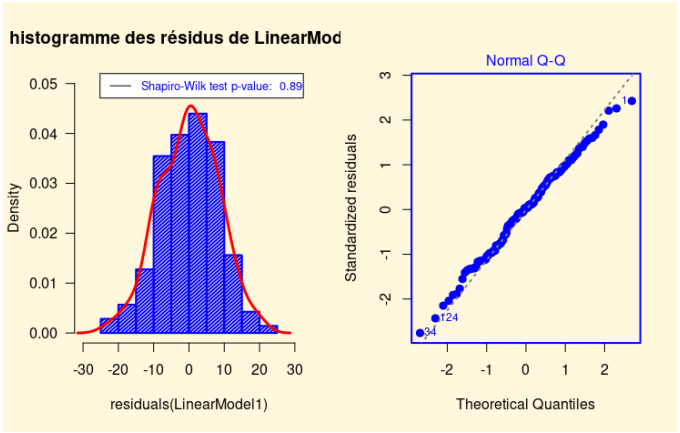

FIGURE 5.12 – Graphes de normalité des résidus

La figure ci-dessus nous montre que le nuage de points est bien ajusté par la droite $y = x$ et l'histogramme des résidus associé à la densité de celles-ci montre que les résidus suivent une loi normale centré de variance σ^2. On remarque une p-value $= 0.89 > \alpha = 0.05$, alors on admet la normalité de $\varepsilon_1, \cdots\cdots, \varepsilon_{141}$.

5.2.4 Estimations des paramètres β_0,β_1 et σ^2_ε par MCO

Maintenant que les hypothèses de validation du modèle $DELAI.DIAG_i = \beta_0 + \beta_1 DPHT_i + \varepsilon_i$ sont vérifiées, nous pouvons passer à l'estimation de ces paramètres inconnus.

Nous allons estimer les paramètres et obtenir :

$$\hat{delai.diag_i} = \hat{\beta}_0 + \hat{\beta}_1 dpht_i,$$

où $delai.diag_i$ et $dpht_i$ respectivement les valeurs observées des variables aléatoires DELAI.DIAG et DPHT.
Il s'agit de calculer le vecteur des estimateurs $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ définie par l'égalité suivante :

$$\hat{\beta} = \left(X^t X\right)^{-1} X^t Y. \tag{5.1}$$

On obtient :

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	30.7638	1.2559	24.495	$< 2 \times 10^{-16}$	$***$
DPHT	0.4675	0.0875	5.343	3.64×10^{-07}	$***$
Residual sd error :	8.435	on 139	degrees	of	freedom

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}_\varepsilon$
30.7638	0.4675	8.435

Par suit, la droite de régression estimée du modèle est donnée par :

$$\hat{delai.diag_i} = 30.7638 + 0.4675 \times dpht_i.$$

Le signe du coefficient de régression nous indique la direction de la relation entre les variables. Dans cette équation, nous remarquons que le coefficient de régression estimé β_1 associé à la variable **l'usage de la phytothérapie** (DPHT) est positif. Cela signifie que plus l'usage de la phytothérapie augmente, plus le délai de diagnostic de la maladie augmente également.

Maintenant que nous avons identifié la direction de la relation, nous pouvons représenter graphiquement la droite de régression sur le nuage de points.

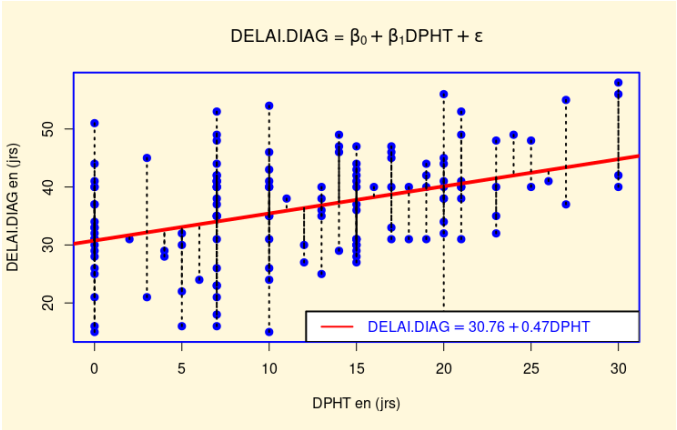

FIGURE 5.13 – Représentation de la droite de régression des moindres carrés sur le nuage de points du délai de diagnostic sur l'usage de la phytothérapie (en jrs).

5.2.4.1 Évaluation

a-) **Estimation de la matrice de variance-covariance de $\hat{\beta}$**

Avec le logiciel d'étude R, on obtient le résultat suivant :

	(Intercept)	DPHT
(Intercept)	1.57736668	−0.090627044
DPHT	−0.09062704	0.007656329.

Les écart-types $\hat{\sigma}_\varepsilon\left(\hat{\beta}\right)$ des estimateurs sont alors donnés par les racines carrés des éléments de la diagonale de cette matrice de variance-covariance.

$$\hat{\sigma}_\varepsilon\left(\hat{\beta}\right) = \begin{bmatrix} \hat{\sigma}_\varepsilon\left(\hat{\beta_0}\right) \\ \hat{\sigma}_\varepsilon\left(\hat{\beta_1}\right) \end{bmatrix} = \begin{bmatrix} 1.2559 \\ 0.0875 \end{bmatrix}.$$

b-) **Intervalles de confiances de β_j**

L'intervalle de confiance pour estimer β_j $(j = 0, 1)$, au niveau de confiance $(1 - \alpha)$, est donnée par :

$$\left[\beta_j - t_{n-2}^{1-\alpha/2}\hat{\sigma}_\varepsilon\left(\hat{\beta}_j\right); \beta_j + t_{n-2}^{1-\alpha/2}\hat{\sigma}_\varepsilon\left(\hat{\beta}_j\right)\right].$$

Avec la commande *confint* appliqué au *modele1*, on obtient :

$$\begin{aligned}\beta_0 &\in [28.2806013 \quad ; \quad 33.2470052]\\ \beta_1 &\in [0.2945238 \quad ; \quad 0.6405317].\end{aligned}$$

5.2.4.2 Évaluation globale de la régression rls

a-) **Tableau d'analyse de la variance**

L'évaluation globale de la pertinence du modèle de prédiction, s'appuie sur l'équation fondamentale d'analyse de la variance qui est donnée par :

$$\underbrace{\sum_{i=1}^{141}(y_i - \bar{y})^2}_{SC_{total}} = \underbrace{\sum_{i=1}^{141}(\hat{y}_i - \bar{y})^2}_{SC_{reg}} + \underbrace{\sum_{i=1}^{141}\hat{\varepsilon}_i^2}_{SC_{res}}.$$

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés	F_{obs}
Expliquée par la régression	1	2031.338	2031.33827	28.54921
Résidus	139	9890.151	71.15217	
Total	140	11921.489		

TABLE 5.6 – Tableau d'analyse de la variance pour la régression linéaire simple (*LinearModel11*)

b-) **Coefficient de détermination et coefficient de détermination ajusté**

On a après calcul :

$$R^2 = 1 - \frac{SC_{res}}{SC_{total}} = 17.04\% \text{ et } R_a^2 = 1 - \frac{(141-1)}{(141-2)}(1 - R^2) = 16.44\%.$$

5.2.4.3 Prévision

Pour créer une prévision d'une nouvelle observation, il suffit de créer un *data.frame* contenant exactement le même nom de colonne que les données initiales.

Ici, la valeur prédite moyenne Y et l'intervalle de prévision pour y_p au niveau $\alpha = 95\%$ pour $DPHT = 56$, donne :

```
fit          lwr       upr
56.94535   49.17694  64.71377,
```

où *fit* : la valeur prédite estimée de Y. *lwr* et *upr* : les bornes inférieures et supérieures de l'intervalle de confiance.

Ainsi il nous est possible pour un nombre quelconque de valeurs à prédire (par exemple 50) de modéliser les tracés d'intervalles de confiances et de prévisions sur le modèle.

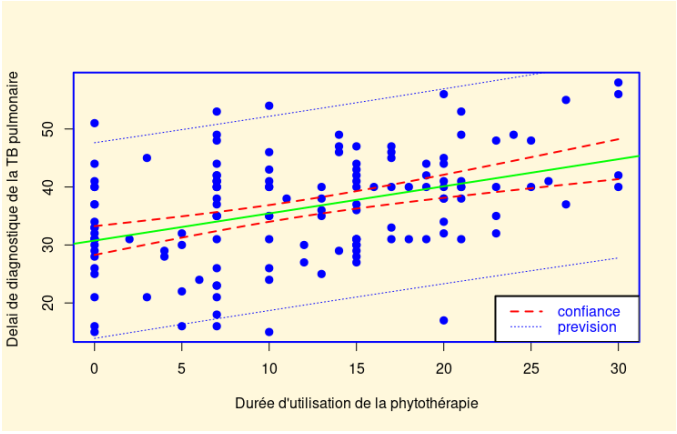

FIGURE 5.14 – Visualisation de l'intervalle de confiance et de l'intervalle de prévision.

5.2.5 Estimation des paramètres β_0,β_1 et σ_ε^2 par Maximum de Vraisemblance (MV)

Nous avons estimé les paramètres β_1, β_0 et σ_ε par MCO précédemment, maintenant nous utilisons la méthode (MV) pour déterminer les estimateurs associés.
On a toujours notre modèle rls suivant :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Leftrightarrow delai.diag_i = \beta_0 + \beta_1 dpht_i + \varepsilon_i, i = 1, \cdots\cdots, 141.$$

Par conséquent, la fonction log-vraisemblance correspondante est :

$$log\left[\mathcal{L}(\beta_0, \beta_1, \sigma_\varepsilon^2)\right] = -\frac{141}{2} log\sigma_\varepsilon^2 - \tfrac{141}{2} log(2\pi) - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{141} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Et comme estimer les paramètres β_0, β_1, et σ_ε^2 par MV, consiste à maximiser la fonction log-vraisemblance ci-dessus, alors ceci revient à annuler les dérivées premières par rapport aux arguments β_0, ,β_1, et σ_ε^2 .

D'où on a :

$$\begin{cases} \sum_{I=1}^{141} delai.diag_i = n\tilde{\beta}_0 + \tilde{\beta}_1 \sum_{I=1}^{141} dpht_i & (5.2a) \\ \sum_{I=1}^{141} delai.diag_i x_i = \tilde{\beta}_0 \sum_{I=1}^{141} dpht_i + \tilde{\beta}_1 \sum_{I=1}^{141} dpht_i^2 & (5.2b) \\ \tilde{\sigma}_\varepsilon^2 = \frac{1}{141} \sum_{I=1}^{141} \left(delai.diag_i - \tilde{\beta}_0 - \tilde{\beta}_1 dpht_i\right)^2 . & (5.2c) \end{cases}$$

Le résultat tant attendu devient :

$$\tilde{\beta}_0 = \bar{delai.diag} - \tilde{\beta}_1 \bar{dpht}, \tilde{\beta}_1 = \frac{\sum_{i=1}^{141} dpht_i delai.diag_i - 141 \bar{dpht} \bar{delai.diagt}}{\sum_{i=1}^{141} dpht_i^2 - 141 \bar{dpht}^2} = \frac{S_{xy}}{S_{xx}} \text{ et}$$

$$\tilde{\sigma}_\varepsilon^2 = \frac{1}{141} \sum_{I=1}^{141} \left(delai.diag_i - \tilde{\beta}_0 - \tilde{\beta}_1 dphtt_i\right)^2 = \frac{1}{141} \sum_{I=1}^{142} \tilde{\varepsilon}_i^2.$$

On obtient :

	Estimate	Std.Error	t value	Pr(>|t|)	
b0	30.763807	1.246994	24.6704	$< 2.2 \times 10^{-16}$	***
b1	0.467528	0.086878	5.3814	7.389×10^{-08}	***
sigma	8.375139	0.498732	16.7928	$< 2.2 \times 10^{-16}$	***

$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\sigma}_\varepsilon$
30.763807	0.467528	8.375139
$\tilde{\sigma}_\varepsilon(\tilde{\beta}_0)$	$\tilde{\sigma}_\varepsilon(\tilde{\beta}_1)$	$\tilde{\sigma}_\varepsilon(\tilde{\sigma}_\varepsilon)$
1.246994	0.086878	0.498732

5.2.6 Comparaison des estimateurs MCO et MV

Afin de comparer nos estimateurs issus des deux méthodes d'estimation, nous dressons le tableau comparatif suivant :

Estimateurs	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}_\varepsilon^2$	$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\sigma}_\varepsilon^2$
Biais	3.576376×10^{-16}	1.819817×10^{-16}	0.000	2.335075×10^{-08}	3.937143×10^{-08}	1.005546
Variance	1.577367	0.007656329	72.84361	1.55501	0.007547811	69.78916
E.Q.M	1.577367	0.007656329	72.84361	1.55501	0.007547811	70.80028

TABLE 5.7 – Tableau comparatif des estimateurs MCO et MV pour la régression linéaire simple

Le tableau 5.7 montre nettement que les estimateurs $\hat{\beta}_j$ et $\tilde{\beta}_j$, $j \in \{0,1\}$ sont quasi identiques, alors que les estimateurs $\hat{\sigma}_\varepsilon^2$ et $\tilde{\sigma}_\varepsilon^2$ ont respectivement des variances et **E.Q.M** différent avec l'estimateur de la variance pour l'erreur aléatoire du modèle plus petit avec la méthode MV qu'avec la méthode MCO, aussi $\textbf{E.Q.M}(\tilde{\sigma}_\varepsilon^2) < \textbf{E.Q.M}(\hat{\sigma}_\varepsilon^2)$.
L'efficacité relative de deux estimateurs, $\hat{\sigma}_\varepsilon^2$ et $\tilde{\sigma}_\varepsilon^2$ est donc :

$$eff(\hat{\sigma}_\varepsilon^2, \tilde{\sigma}_\varepsilon^2) = \frac{\textbf{E.Q.M}(\tilde{\sigma}_\varepsilon^2)}{\textbf{E.Q.M}(\hat{\sigma}_\varepsilon^2)} = 0.97\,.$$

5.3 Régression linéaire multiple

Dans cette partie ou section, nous allons déterminer les variables explicatives qui influencent notre variable réponse, le délai de diagnostic (**DELAI.DIAG**), plutôt que de discuter de cette variable elle-même. Étant donné que nous disposons de 30 variables explicatives, il n'est pas garanti que toutes ces variables contribuent à expliquer notre variable réponse. Nous utiliserons une méthode bien connue pour sélectionner les variables qui ont réellement un impact sur notre variable dépendante.

5.3.1 Sélection des variables explicatives

Il existe plusieurs méthodes de sélection de variables en régression linéaire multiple, telles que la méthode basée sur le coefficient de détermination R^2, celle basée sur le coefficient de détermination ajusté R_a^2, l'utilisation des critères d'information d'Akaike et de Bayes, ainsi que la méthode du Cp de Mallows.
Nous nous permettons d'utiliser une méthode exaustive basée su le critère d'information de Bayes (**BIC**). La méthode de sélection des variables par (BIC) est une approche utilisée pour sélectionner les variables les plus pertinentes dans une régression linéaire multiple. Le BIC est un critère statistique qui permet d'évaluer la qualité de différents modèles statistiques en prenant en compte à la fois l'ajustement aux données et la complexité du modèle. Il est défini comme :

$$BIC = -2 * ln(L) + k * ln(n),$$

où ln(L) est le logarithme de la fonction de vraisemblance maximale du modèle, k est le nombre de paramètres dans le modèle et n est la taille de l'échantillon.

Le processus de sélection des variables par BIC consiste à ajuster plusieurs modèles de régression linéaire multiple en utilisant différentes combinaisons de variables explicatives et en évaluant chaque modèle en fonction de son BIC. Le modèle qui a le BIC le plus petit est considéré comme le meilleur modèle.
On obtient ainsi ce graphe ci-dessous illustrant cette sélection :

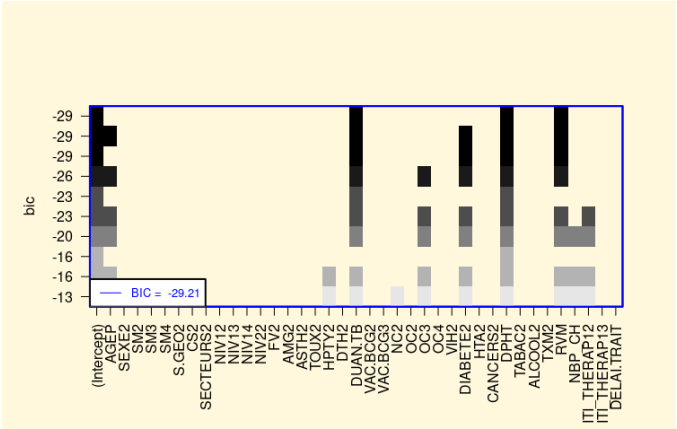

FIGURE 5.15 – Graphe de sélection des variables et la valeur du BIC obtenu.

5.3.2 Modèle de régression linéaire multiple

De ce qui précède, on a le modèle de régression multiple associé qui est :

$$DELAI.DIAG_i = \beta_0 + \beta_1 DPHT_i + \beta_2 DUAN.TB_i + \beta_3 RVM_i + \varepsilon_i, i = 1, \cdots\cdots, 141,$$

avec toujours $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ et β_0, β_1, β_2, β_3 et σ_ε^2 les inconnus de la régression.

5.3.3 Validation des Hypothèses

Résidus

Pour tout $i \in \{1, \cdots\cdots, 141\}$, on a :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i.$$

Analyse graphiques groupées

Ici, nous essayons de tracer les graphiques de manière regroupée puis de les interpréter pour valider ou non le modèle.

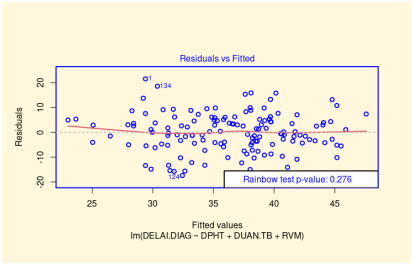

FIGURE 5.16 – Graphe résidus vs valeurs ajustées

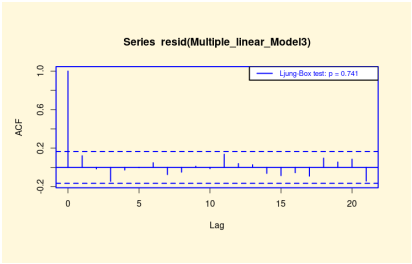

FIGURE 5.17 – Graphe acf des résidus

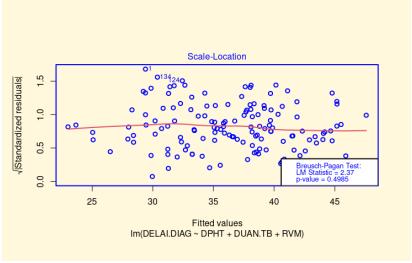

FIGURE 5.18 – Graphe égalité des variances

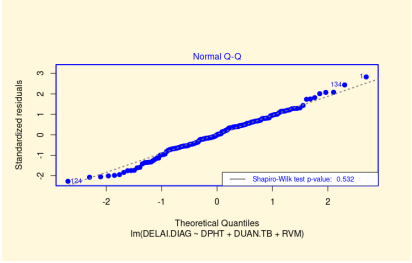

FIGURE 5.19 – Graphe QQ-plot pour la normalité des résidus

Les résultats graphiques obtenus sont satisfaisants et permettent d'observer plusieurs éléments importants.
Tout d'abord, le graphique 5.16 présente un tracé rouge approximativement horizontal et la p-value du test de Rainbow est supérieure à 0.05, ce qui montre que la régression linéaire semble être adaptée.
Le graphique 5.17 ne présente aucune structure particulière, et aucune des valeurs ne dépassent les bornes limites, ce qui suggère l'indépendance des erreurs $\varepsilon_1, \ldots, \varepsilon_{141}$.
Le graphique 5.18 ne montre pas de structure particulière et le test de Breush-Pagan indique une $p - value$ supérieure à 0.05, ce qui permet d'admettre l'égalité des variances de $\varepsilon_1, \ldots, \varepsilon_{141}$.
Enfin, le graphique 5.19 montre un alignement des points sur la diagonale et le test de Shapiro-Wilk indique une p-value supérieure à 0.05, ce qui permet d'admettre la normalité de $\varepsilon_1, \ldots, \varepsilon_{141}$.
Complément : Calcul du facteur d'inflation de variance (VIF)
Le calcul du facteur d'inflation de variance du modèle par le logiciel R, nous donne les résultats ci-dessous :

DPHT	DUAN. TB	RVM
1.090100	1.094787	1.009090 .

Comme toutes les valeurs sont inférieures à 5, alors il n'existe pas de multicolinéarité entre les variables explicatives.

5.3.4 Estimations des paramètres β_0,β_1,β_2,β_3 et σ_ε^2 par MCO

Comme dans le cas simple, il s'agit de calculer le vecteur des estimateurs $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ défini par l'égalité suivante :

$$\hat{\beta} = \left(X^t X\right)^{-1} X^t Y. \tag{5.3}$$

On obtient :

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_\varepsilon$
3.244e+01	4.031e-01	4.601e-01	-1.452e-04	7.736

L'équation de l'hyperplan des moindres carrés est donnée par :

$$\hat{delai.diag_i} = 32.44 + 0.4031 \times dpht_i + 0.4601 \times duan.tb_i - 0.0001452 \times rvm_i.$$

Le signe du coefficient nous indique le sens de la relation. D'après cette équation, on remarque que les coefficients de régression estimés ($\hat{\beta}_1$ et $\hat{\beta}_2$), associés aux variables l'usage de la phytothérapie et des antibiotiques non TB, sont positifs. Cela signifie que l'augmentation des durées d'utilisation de la phytothérapie et des antibiotiques non TB influent positivement sur le délai de diagnostic. Mais une augmentation du revenu mensuel des patients a un impact négatif sur le délai de diagnostic car son coefficient $\hat{\beta}_3$ est négatif.

5.3.4.1 Évaluation

a-) **Estimation de la matrice de variance-covariance de $\hat{\beta}$**
La matrice de variance-covariance définie par :

$$varcov\left(\hat{\beta}\right) = \sigma_\varepsilon^2 \left(X^t X\right)^{-1},$$

est

	(Intercept)	DPHT	DUAN.TB	RVM
(Intercept)	$4.095481 \times 10^{+00}$	$-4.602835 \times 10^{-02}$	$-1.602752 \times 10^{-01}$	$-4.884315 \times 10^{-05}$
DPHT	$-4.602835 \times 10^{-02}$	7.019013×10^{-03}	$-3.438367 \times 10^{-03}$	$-1.819018 \times 10^{-07}$
DUAN.TB	$-1.602752 \times 10^{-01}$	$-3.438367 \times 10^{-03}$	2.066855×10^{-02}	4.653972×10^{-07}
RVM	$-4.884315 \times 10^{-05}$	$-1.819018 \times 10^{-07}$	4.653972×10^{-07}	1.351261×10^{-09}

Les écart-types $\hat{\sigma}_\varepsilon(\hat{\beta}_j)$ des estimateurs $\hat{\beta}_j$ $(j = 0, 1, 2, 3)$ sont alors donnés par les racines des éléments diagonaux de la matrice de variance-covariance. Ainsi, on a :

$$\hat{\sigma}_\varepsilon\left(\hat{\beta}\right) = \begin{bmatrix} \hat{\sigma}_\varepsilon\left(\hat{\beta}_0\right) \\ \hat{\sigma}_\varepsilon\left(\hat{\beta}_1\right) \\ \hat{\sigma}_\varepsilon\left(\hat{\beta}_2\right) \\ \hat{\sigma}_\varepsilon\left(\hat{\beta}_3\right) \end{bmatrix} = \begin{bmatrix} 2.024 \times 10^{+00} \\ 8.378 \times 10^{-02} \\ 1.438 \times 10^{-01} \\ 3.676 \times 10^{-05} \end{bmatrix}.$$

b-) **Intervalles de confiance de β_j**

Avec la commande *confint* appliquée à notre modèle *Multiple_Linear_Model3*, on obtient :

$$\begin{aligned} \beta_0 &\in [28.4417305604 \quad ; \quad 36.4453022837] \\ \beta_1 &\in [0.2374459417 \quad ; \quad 0.5687825333] \\ \beta_2 &\in [0.1757660753 \quad ; \quad 0.7443392424] \\ \beta_3 &\in [-0.0002178623 \quad ; \quad -0.0000724835]. \end{aligned}$$

5.3.4.2 Évaluation globale de la régression rlm

a-) **Tableau d'analyse de la variance**

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrés	F_{obs}
Expliquée par la régression	3	3723.677	1241.22561	20.74308
Résidus	137	8197.813	59.83805	
Total	140	11921.489		

TABLE 5.8 – Tableau d'analyse de la variance pour la régression linéaire multiple (*LinearModel11*)

b-) **Coefficient de détermination et coefficient de détermination ajusté**

On a après calcul :

$$R^2 = 1 - \frac{SC_{res}}{SC_{total}} = 31.23\% \text{ et } R_a^2 = 1 - \frac{(141-1)}{(141-4)}(1-R^2) = 29.73\%.$$

Il n'existe pas de règle précise pour évaluer la qualité d'un coefficient de détermination (R^2). Un R^2 élevé (supérieur à 85%) peut cacher des problèmes et donner des résultats incorrects. Atteindre des valeurs considérées comme "satisfaisantes" peut être difficile en raison des limites des données disponibles. Parfois, un R^2 de seulement 40% ou 30% peut être acceptable, sous réserve de la validation des hypothèses et des tests du modèle. Dans des domaines comme la psychologie, des valeurs de R^2 inférieures à 50% sont généralement attendues, en raison de la complexité de la prédiction du comportement humain. Plus de détail (voir [6], [12], [22], [28]).

5.3.4.3 Test de significativité du modèle

a-) **Test global de Fisher**

Il s'agit surtout de répondre à la question suivante :

Est ce que la liaison globale entre $DELAI.DIAG$ et les variables $DPHT$, $DUANT.TB$ et RVM est -elle significative ?

Autrement dit tester l'hypothèse :

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1 : \exists j \in \{1,2,3\} \text{ tels que } \beta_j \neq 0 \end{cases}.$$

On calcule la statistique de test :

$$F = \frac{\frac{R^2}{3}}{\frac{1-R^2}{141-4}} = \frac{(141-4)}{3}\frac{R^2}{1-R^2} = 20.73826.$$

Règle de décision

Comme la statistique $F = 20.73826$ est supérieure à la valeur critique $f^{0.95}_{(1,137)} = 2.670687$ (valeur théorique), on en conclut que le test est significatif et on rejette H_0 au seuil de significativité $\alpha = 0.05$. Ce qui montre que les variables $X1 = DPHT$, $X2 = DUAN.TB$ et $X3 = RVM$ contribuent à expliquer le délai de diagnostic de la TB pulmonaire.

b-) **Test de Student sur le paramètre β_j**

Il s'agit plutôt de répondre à la question suivante :

L'apport marginal de la variable DPHTou celle de DUAN.TB ou encore celle de RVM est-il significatif ?

En d'autre terme pour $j \in \{0, 1, 2, 3\}$, on considère les hypothèses :

$$\begin{cases} H_0 : \beta_j = 0 \\ \qquad\quad vs \\ H_1 : \beta_j \neq 0. \end{cases}$$

Grâce à la commande la commande *summary* du logiciel R, on a :

H_1	$\beta_0 \neq 0$	$\beta_1 \neq 0$	$\beta_2 \neq 0$	$\beta_3 \neq 0$
t_{obs}	16.032	4.812	3.200	-3.949

Règle de décision

Pour $j \in \{0, 1, 2, 3\}$, comme la valeur critique donnée par $t_{(0.975,141)} = 1.976931$, H_0 est rejeté au seuil de significativité $\alpha = 5\%$.

Donc, cela implique que les variables : l'usage de la phytothérapie, des antibiotiques non TB et le revenu mensuel faible des patients ont une influence très significative (importante) sur le délai de diagnostique des patients atteints de la TB pulmonaire.

5.3.4.4 Prévision

En cherchant à prédire le délai de diagnostic pour deux nouveaux patients qui ne font pas partie de notre base de données, mais dont nous disposons des informations (voir ci-dessous), nous obtenons :

Patient	**DPHT**	**DUAN.TB**	**RVM**	**DELAI.DIAG**	Intervalle de Prévision
1	10	10	150000	19.29925	[1.780345 ; 36.81816]
2	10	10	35000	35.99413	[20.635812 ; 51.35246]

5.3.5 Estimation des paramètres β et σ^2_ε par Maximum de Vraisemblance (MV)

Comme nous avons estimé les paramètres $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ et σ_ε par MCO précédemment, maintenant nous utilisons la méthode (MV) pour déterminer les estimateurs associés. De plus, sachant que $Y = DELAI.DIAG$ et X la matrice composée des valeurs des variables $DPHT$,$DUAN.TB$ et RVM, alors on a :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i \Leftrightarrow delai.diag_i = \beta_0 + \beta_1 dpht_i + \beta_2 duan.tb_i + \beta_3 rvm_i + \varepsilon_i, i = 1, \cdots\cdots, 141.$$

Par conséquent, la fonction log-vraisemblance correspondant est :

$$\begin{cases} \dfrac{\partial logL}{\partial \beta} = -\dfrac{1}{2\sigma^2_\varepsilon}(-2X^tY + 2X^tX\beta) = 0 & (5.4a) \\ \dfrac{\partial logL}{\partial \sigma^2_\varepsilon} = \dfrac{\partial}{\partial \sigma^2_\varepsilon}\left(-\dfrac{n}{2} log(2\sigma^2_\varepsilon)\right) + \dfrac{\partial}{\partial \sigma^2_\varepsilon}\left(-\dfrac{1}{2\sigma^2_\varepsilon}(Y^tY - 2X^tY\beta + X^tX\beta^2)\right) = 0 & (5.4b) \end{cases}$$

Ainsi, après calcul, nous obtenons :

$$\tilde{\beta} = (X^tX)^{-1}X^tY \text{ et } \tilde{\sigma}^2_\varepsilon = \frac{(Y - X\tilde{\beta})^t(Y - X\tilde{\beta})}{n} = \frac{1}{n}\sum_{I=1}^{n} \tilde{\varepsilon}^2_i.$$

Par le biais du logiciel R, nous obtenons les résultats suivants :

	Estimate	Std.Error	z value	Pr(>|t|)	
(Intercept)	$3.2442 \times 10^{+01}$	4.2205×10^{-03}	$7.6868 \times 10^{+03}$	$< 2.2 \times 10^{-16}$	$***$
beta.tilde1	4.0313×10^{-01}	7.9499×10^{-02}	$5.0709 \times 10^{+00}$	3.960×10^{-07}	$***$
beta.tilde2	4.6011×10^{-01}	1.1811×10^{-01}	$3.8957 \times 10^{+00}$	9.792×10^{-05}	$***$
beta.tilde3	-1.4516×10^{-04}	2.7358×10^{-05}	$-5.3058 \times 10^{+00}$	1.122×10^{-07}	$***$
sigma	$7.6250 \times 10^{+00}$	9.3901×10^{-09}	$8.1203 \times 10^{+08}$	$< 2.2 \times 10^{-16}$	$***$
—— -2 log L : 973.0043					

$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$	$\tilde{\sigma}_\varepsilon$
$3.2442 \times 10^{+01}$	4.0313×10^{-01}	4.6011×10^{-01}	-1.4516×10^{-04}	$7.6250 \times 10^{+00}$
$\tilde{\sigma}_\varepsilon(\tilde{\beta}_0)$	$\tilde{\sigma}_\varepsilon(\tilde{\beta}_1)$	$\tilde{\sigma}_\varepsilon(\tilde{\beta}_2)$	$\tilde{\sigma}_\varepsilon(\tilde{\beta}_3)$	$\tilde{\sigma}_\varepsilon(\tilde{\sigma}_\varepsilon)$
4.2205×10^{-03}	7.9499×10^{-02}	1.1811×10^{-01}	2.7358×10^{-05}	9.3901×10^{-09}

5.3.6 Comparaison des estimateurs MCO et MV

Comme nous l'avons pu constater théoriquement au chapitre 3, les estimateurs $\hat{\beta}$ et $\tilde{\beta}$ sont égaux et que la différence semble être observée avec les estimateurs $\hat{\sigma}^2_\varepsilon$ et $\tilde{\sigma}^2_\varepsilon$.
À présent, calculons la variance, le biais et l'erreur quadratique moyenne de chacun des estimateurs.

Estimateurs	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}^2_\varepsilon$	$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$	$\tilde{\sigma}^2_\varepsilon$
Biais	4.278×10^{-18}	7.212×10^{-18}	1.015×10^{-16}	2.513×10^{-20}	2.131628×10^{-14}	1.4949×10^{-4}	3.0809×10^{-05}	5.2228×10^{-06}	4.3818×10^{-08}	1.705188
Variances	4.095481	7.019×10^{-03}	2.067×10^{-02}	1.351×10^{-09}	52.27142	3.979296	0.006813	2.008×10^{-02}	7.4845×10^{-10}	50.80152
E.Q.M	4.095481	0.007019	2.067×10^{-02}	1.351×10^{-09}	52.27142	3.979296	0.006813	2.008×10^{-02}	7.4845×10^{-10}	53.70919

TABLE 5.9 – Tableau comparatif des estimateurs MCO et MV pour la régression linéaire multiple

Le tableau 5.9 met en évidence les différences théoriques. On peut clairement voir que les estimateurs $\hat{\beta}_j$ et $\tilde{\beta}_j$ (où j est 0, 1, 2 ou 3) sont presque identiques. En revanche, les estimateurs $\hat{\sigma}^2_\varepsilon$ et $\tilde{\sigma}^2_\varepsilon$ ont des variances et des erreurs quadratiques moyennes différentes. L'estimateur de la variance pour l'erreur aléatoire du modèle est plus petit avec la méthode MV qu'avec la méthode MCO, ce qui est normal car cela découle des résultats théoriques lorsque la taille de la base de données n est plus grande que $k/(n-k)$, avec $k = p + 1$.

L'efficacité relative des deux estimateurs, $\hat{\sigma}^2_\varepsilon$ et $\tilde{\sigma}^2_\varepsilon$, est donc de 0.96, calculée comme le rapport des erreurs quadratiques moyennes :

$$eff(\hat{\sigma}^2_\varepsilon, \tilde{\sigma}^2_\varepsilon) = \frac{\textbf{E.Q.M}(\hat{\sigma}^2_\varepsilon)}{\textbf{E.Q.M}(\tilde{\sigma}^2_\varepsilon)} = 0.96.$$

Conclusion et Perspectives

L'étude sur le délai tardif de diagnostic a révélé les caractéristiques principales de nos patients : ils sont majoritairement jeunes, de sexe masculin, peu ou pas scolarisés, et la plupart ont un faible revenu. Les signes cliniques les plus fréquents sont la fièvre (78%), la toux (90%) et l'amaigrissement (59, 6%). L'analyse multivariée a identifié trois facteurs associés à ce délai tardif de diagnostic de la tuberculose pulmonaire : l'utilisation de la phytothérapie, l'utilisation d'antibiotiques non spécifiques à la tuberculose et le faible revenu des patients, allant dans la même optique que les résultats obtenus par des spécialistes et chercheurs de la santé bien qu'ils ne soient pas du Sénégal (voir [2], [19], [25] et [27]).
Prendre en compte ces facteurs contribuera à réduire considérablement ce délai, ce qui entraînera une diminution de la transmission de la maladie au sein de la communauté.

En ce qui concerne la régression linéaire simple, la méthode du Maximum de Vraisemblance (MV) présente une Erreur Quadratique Moyenne (E.Q.M) plus faible que celle des Moindres Carrés Ordinaires (MCO) et une variance plus petite. Cependant, en régression multiple, la variance de l'estimateur MV reste généralement plus faible, mais son E.Q.M devient supérieur à celle des MCO.

Il est important de prendre en considération les caractéristiques des deux méthodes de régression. Les MCO sont plus simples à calculer et à interpréter, et ils sont robustes face aux violations des hypothèses. En revanche, la méthode MV offre une approche plus probabiliste et peut être plus appropriée lorsque les hypothèses du modèle sont respectées.

Le choix de la méthode d'estimation dépendra donc du contexte spécifique de l'étude, des hypothèses du modèle et des objectifs de l'analyse. Il est recommandé d'évaluer attentivement les performances des deux méthodes, ainsi que leurs avantages et leurs limitations, avant de prendre une décision finale.

Pour récapituler, la méthode du Maximum de Vraisemblance offre une meilleure précision en termes de variance lorsqu'il s'agit de la régression linéaire simple. Cependant, lorsqu'il s'agit de la régression multiple, son Erreur Quadratique Moyenne peut être plus élevée que celle des Moindres Carrés Ordinaires. Les Moindres Carrés Ordinaires, quant à eux, sont plus simples et robustes. Le choix de la méthode dépendra de plusieurs facteurs, et il est crucial de les évaluer en fonction des besoins spécifiques de chaque étude.

Dans notre étude axée sur le délai tardif de diagnostic (délai patient), nous avons constaté que la majorité des patients subissent un délai de traitement dépassant les 24 heures, voire certains commencent leur traitement six jours (près d'une semaine) après le diagnostic. Cela constitue un problème majeur qui nous pousse à envisager une autre étude visant à identifier les facteurs explicatifs de ce délai tardif de traitement en utilisant un modèle de régression logistique.

En ce qui concerne l'étude comparative des méthodes d'estimation, il serait également intéressant d'examiner le calcul des estimateurs par la méthode d'inférence bayésienne et de les comparer à nos estimateurs issus des deux méthodes développées dans ce travail.

Bibliographie

[1] Azais Jean-Marc, Bardet Jean Marc.(2006) *Le Modèle Linéaire par exemple. Régression,analyse de la variance et plan d'expériences illustrés avec R, SAS et Splus*.

[2] Belkina T. V., Khojiev D. S., Tillyashaykhov M. N., Tigay, Z. N., Kudenov M. U., Tebbens J. D., & Vlcek J. (2014). Delay in the diagnosis and treatment of pulmonary tuberculosis in Uzbekistan : a cross-sectional study. *BMC Infectious Diseases*, 14, 624. (http ://www.biomedcentral.com/1471-2334/14/624).

[3] Bertrand Fréderic, Maumy-Bertrand Myriam. (2017). *Régression Linéaire multiple*. Cours Master1, Université de Strasbourg France.

[4] Bonneu M., Leconte E.. *Polycopié de cours de statistique et économétrie : Modèle Linéaire* Université des Sciences Sociales. Place Anatole France. 31042 .Toulouse.

[5] Bourbonnais Régis. (2015). *Économétrie Cours et exercices corrigés*,Dunod, 9^{ime} édition.

[6] Brownlee J. (2021). How to Interpret Linear Regression Coefficients. Machine Learning Mastery.(https ://machinelearningmastery.com/interpret-linear-regression-coefficients/)

[7] Dr. Cabral Emanuel Nicolas. *Statistiques Inférentielles* (2020-2021). Cours de Master1 Mathématiques et Applications à l'Université Assane Seck de Ziguinchor.

[8] Chatterjee S., & Hadi A. S. (2006). *Regression Analysis by Example*. Wiley.

[9] Chesneau Christophe. (2020). *Études :modèles de régression*. Cours Master2, Université de Caen Basse-Normandie (http : // www.math.unicaen.fr / chesneau/).

[10] Chesneau Christophe. (2017). *Modèles de Régression*. Cours Master2, Université de Caen Basse-Normandie (http : // www.math.unicaen.fr / chesneau/).

[11] Chesneau Christophe. (2017). *Sur l'Estimateur des Moindres Carrés ordinaires(emco)*. Université de Caen (http ://www.math.unicaen.fr/ chesneau/).

[12] Colas J. (2020). Analyse de la régression : Comment interpréter le R-carré et évaluer l'adéquation de l'ajustement ? Blog Minitab.(https ://blog.minitab.com/fr/analyse-de-la-regression-comment-interpreter-le-r-carre-et-evaluer-ladequation-de-lajustement).

[13] Cornillon Pierre-André, Matzner-Lober Eric. (2011). *Régression avec R*. Springer.

[14] De Scheemaekere Xavier. (2015). *Fondements philosophiques du concept de probabilité*.EME Édition.

[15] Pr. Diédhiou Alassane. (2021-2022). *Modèles Aléatoires* , Cours de Master1 Mathématiques et Applications à l'Université Assane Seck de Ziguinchor.

[16] Dupuy, Jean-François. (2018). *Méthodes statistiques pour l'analyse de données de comptage surdispersées*. volume 4, ISTE Group.

[17] Fredon Daniel, Maumy-Bertrand Myriam, Bertrand Frédéric. (2018). *Mathématiques Statistiques et probabilités en 30 fiches*. 1ère édition. pages :125-129.

[18] Fromont Renoir Magalie . *Modèles de régression linéaire*. Master Statistique Appliquée Mention Statistique pour l'Entreprise. Université de Rennes.

[19] Fuge T.G., Bawore S.G., Solomon D.W. et al. (2018). Patient delay in seeking tuberculosis diagnosis and associated factors in Hadiya Zone, Southern Ethiopia. *BMC Res Notes* 11, 115. (https ://doi.org/10.1186/s13104-018-3215-y).

[20] James G., Witten D., Hastie T., & Tibshirani R. (2013). *An Introduction to Statistical Learning*. Springer.

[21] Kibala Kuma Jonas. (2019). *Estimation par la méthode du Maximum de Vraisemblance : Éléments de Théorie et pratiques sur Logiciel*. Licence. Congo-Kinshasa. ffcel-02189969f.(https ://hal.science/cel-02189969/document).

[22] Kutner M. H., Nachtsheim C. J., Neter J., & Li W. (2004). *Applied Linear Statistical Models*. McGraw-Hill/Irwin.

[23] Lalanne Christophe Mesbah Mounir. (2016). *Biostatistique et analyse informatique des données de santé avec R*. Collection : Bioingénierie médicale.

[24] Le Digabel S.. (2017). *Régression linéaire Simple*. Cours MTH2302D, École Polytechnique de Montréal.

[25] Mahato R.K., Laohasiriwong W., Vaeteewootacharn K., Koju R., & Bhattarai R. (2015). Major Delays in the Diagnosis and Management of Tuberculosis Patients in Nepal. *Journal of Clinical and Diagnostic Research*, 9(11), LC05-LC08. DOI : 10.7860/JCDR/2015/16307.6633.

[26] Matzner-Lober Eric. 2006. *Régression théorie et applications*. Collection : Statistique et probabilités appliquées (French Edition). Springer.

[27] Mfinanga, S. G., Mutayoba, B. K., Kahwa, A., Kimaro, G., Mtandu, R., Ngadaya, E., Egwaga, S., & Kitua, A. Y. (2013). The magnitude and factors associated with delays in management of smear positive tuberculosis in Dar es Salaam, Tanzania. *BMC Health Services Research*, *13*(1), 1-9.

[28] Montgomery D. C., Peck E. A., & Vining G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley.

[29] Mulkay Benoît. (2018-2019). *Économétrie (M1) chap 5 Estimateur du Maximum de Vraisemblance* Université de Montpellier.

[30] Polomé Philippe. (2020-2021). *Programmation dans R-ch.2. Régression Linéaire et Exensions*. Cours M2 CEE, Université Lumière Lyon 2.

[31] Rakotomalala, R. (11/03/2016). *Analyse de corrélation, étude des dépendances - Variables quantitatives*, (http ://eric.univ-lyon2.fr/ ricco/publications.html).

[32] Rakotomalala, R. (07/06/2011). *Économétrie - La régression linéaire simple et multiple*. (http :// eric.univ-lyon2.fr/ ricco/publications.html).

[33] Stekhoven D.J et Biihlmann P. (2011). *MissForest - nonparametric missing value imputation for mixed-type data*, Bioinformatics Advance Acess.

[34] Otwombe K. N., Variava E., Holmes C. B., Chaisson R. E., & Martinson N. (2013). Predictors of delay in the diagnosis and treatment of suspected tuberculosis in HIV co-infected patients in South Africa. *International Journal of Tuberculosis and Lung Disease*, 17(9), 1199-1205.doi :(http ://dx.doi.org/10.5588/ijtld.12.0891).