

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي

جامعة عبد الحميد بن باديس - مستغانم -
كلية العلوم الدقيقة والإعلام الآلي

تصريح شرقي خاص بالالتزام بقواعد النزاهة العلمية
لإنجاز البحث

أنا الممضي أدناه،

الطالب(ة): هدا حديدية رقم التسجيل الجامعي 202422.0480.44244

الحامل لبطاقة التعريف الوطنية رقم 119.88.10.8300.649000 والصادرة بتاريخ: 13/12/2024

عن المصطفى ولايت محسنة

المسجل ب: كلية العلوم الدقيقة والإعلام الآلي / قسم: الرياضيات

شعبة: الرياضيات / التخصص: تحليل دالي

والمكلف بإنجاز مذكرة ماستر بعنوان:

Régression linéaire Multiple et Application

أصرح شرقي أنني ألتزم بمراعاة المعايير العلمية والمنهجية ومعايير الأخلاقيات العلمية والنزاهة الأكاديمية
المطلوبة في إنجاز البحث، وأتحمل المسؤولية الشخصية عن كل المحتوى المتضمن في البحث المذكور أعلاه.

التاريخ: 16/06/2025

إمضاء المعني



MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ ABDELHAMID IBN BADIS - MOSTAGANEM
FACULTÉ DES SCIENCES EXACTES ET DE L'INFORMATIQUE
DÉPARTEMENT DE MATHÉMATIQUES ET INFORMATIQUE
FILÈRE : MATHÉMATIQUES



MÉMOIRE DE FIN D'ÉTUDES

Pour l'Obtention du Diplôme de Master en Mathématiques délivré par

Université de Mostaganem

Spécialité "Analyse fonctionnelle"

présenté par :

Meddah KHADIDJA

Régression Linéaire Multiple et Application

soutenu publiquement le 12 Juin 2025 devant le jury composé de :

Présidente :	H.KERIM	MCB	Université de Mostaganem
Examineur :	M. MOHAMED	MCA	Université de Mostaganem
Encadreur :	O. BOUANANI	MCA	Université de Mostaganem

Année Universitaire : 2024 / 2025

Table des matières

Introduction	6
1 Régression linéaire Simple	8
1.1 Exemples concrets sur le modèle RLS	8
1.1.1 Relation entre le nombre de tests effectués et le nombre de cas confirmés	8
1.1.2 Consommation d'alcool et espérance de vie par pays	9
1.2 Modèle de régression linéaire simple	9
1.2.1 Formulation vectorielle	10
1.3 Estimation des paramètres β_0 et β_1 par approche non inférentielle	11
1.3.1 Moindres Carré Ordinaires	11
1.3.2 Estimateurs du maximum de vraisemblance	17
1.4 Décomposition de la variance et tableau d'ANOVA	17
1.4.1 Le coefficient de détermination	18
1.5 Test de signification	19
1.5.1 Test de signification β_1	19
1.5.2 Test de signification globale du modèle	20
1.6 Intervalles de confiance et intervalle de prévision	20
1.6.1 Intervalles de confiance pour les coefficients de régression .	20
1.6.2 Prévision	20
2 Régression linéaire Multiple	22
2.1 Modèle de régression linéaire multiple	22
2.2 Estimation de paramètre β	23
2.2.1 Estimateurs des Moindres Carrés Ordinaires	23
2.2.2 Estimation de variance du résidu	26
2.2.3 Estimateurs du maximum de vraisemblance	27
2.3 Analyse de la variance et coefficient de détermination	28
2.3.1 Décomposition de la variance et tableau d'ANOVA	28
2.4 Test de signification	29
2.4.1 Test sur les paramètres du modèle	29
2.4.2 Test de signification globale du modèle	29
2.5 Intervalles de confiance et intervalle de prévision	30
2.5.1 Intervalles de confiance pour les coefficients de régression .	30
2.5.2 Prevision	30
3 Application RLM sur le Volume des Marchandises au Port maritime	32
3.0.1 Estimation par intervalle de confiance	41
Conclusion	44
Bibliography	44

Table des figures

1.1	Représentations graphiques des nuages de points	9
1.2	Droite et résidu de la régression linéaire	11
1.3	Représentation de la projection dans l'espace des variables	13
3.1	Graphique des résidus vs valeurs prédites	38
3.2	Corrélogramme des résidus du modèle avec 30 décalages	39
3.3	Normalité des résidus	40
3.4	La variance des erreurs n'est pas constante	41

Remerciement

Je rends grâce à **ALLAH** de m'avoir donné la force, la patience, le courage et la volonté nécessaires pour mener à bien ce travail.

Je tiens à exprimer ma profonde gratitude à M. **O. Bouanani** pour la confiance qu'il m'a accordée en acceptant de diriger ce mémoire, pour sa disponibilité tout au long de son élaboration, ainsi que pour ses précieux conseils et suggestions. Il a su me rendre fière de travailler à ses côtés.

Je suis infiniment reconnaissante à Mme **H. Kerim** d'avoir accepté d'examiner ce travail et de m'avoir fait l'honneur de présider le jury de ce travail.

Je remercie chaleureusement M. **M. Mohamedi**, qui a toujours partagé avec générosité ses connaissances précieuses durant notre formation en master, et qui a accepté d'évaluer mon travail en tant que membre du jury.

Je n'oublie pas tous les enseignants qui ont contribué à ma formation durant mes études de master.

Enfin, je garde une pensée reconnaissante pour toutes les personnes qui, de près ou de loin, ont participé à la réalisation de ce mémoire.

À vous tous, un grand Merci

Dédicace

Je dédie ce modeste travail :

À mes chères parents, pour leurs prières ,leur amour inconditionnel,leurs sacrifices.
Qu'Allah les protège et les récompense.

À mes soeurs et mon frère Hicham.
À mon mari et à mes enfants.

À toute ma famille, de près ou de loin.
À tout mes amies avec qui suivi cursus de master.

À tous les professeurs du département de mathématique .
Merci à tous pour votre présence,votre soutien et vos encouragements.

Khadija

Résumé

Ce mémoire est consacré à l'étude des méthodes de régression linéaire simple et multiple, visant à analyser les relations entre une variable dépendante et une ou plusieurs variables indépendantes. Une application pratique est proposée à l'aide du logiciel statistique **R**.

Dans un premier temps, nous avons étudié la régression linéaire simple, qui permet d'analyser la relation entre deux variables. Ce modèle est particulièrement utile pour comprendre l'influence d'une variable explicative sur une variable à expliquer, et pour effectuer des prévisions.

Dans un second temps, nous avons abordé la **régression linéaire multiple**, qui étend le modèle précédent en considérant plusieurs variables explicatives. Cette approche permet une modélisation plus précise des phénomènes, notamment lorsque plusieurs facteurs influencent simultanément la variable dépendante.

Sur le plan pratique, nous avons appliqué ces méthodes à l'analyse du volume de marchandises au port de Mostaganem, en exploitant des données réelles telles que le nombre de navires accueillis chaque année, leur taille moyenne ainsi qu'un indicateur économique global. L'analyse statistique réalisée avec **R** a révélé une forte corrélation entre les variables explicatives et le volume de marchandises, confirmant ainsi la pertinence du modèle pour la prévision.

Mots-clés : régression linéaire simple, régression linéaire multiple, port de Mostaganem, volume de marchandises, logiciel R, analyse statistique.

Introduction

La régression linéaire est l'une des méthodes statistiques les plus fondamentales et les plus utilisées pour modéliser la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Elle remonte aux travaux pionniers de Stigler [10] et aux contributions de Carl Friedrich Gauss et Adrien-Marie Legendre au début du XIX^e siècle, qui ont introduit la méthode des moindres carrés pour ajuster une droite à un ensemble de données. Cette technique permet d'estimer les paramètres d'un modèle linéaire en minimisant la somme des carrés des résidus, c'est-à-dire les différences entre les valeurs observées et les valeurs prédites.

La régression linéaire simple consiste à modéliser la relation entre deux variables, tandis que la régression linéaire multiple généralise ce concept à plusieurs variables explicatives. Introduite formellement par Yule [12] et développée davantage par Fisher [2], la régression multiple est aujourd'hui un outil essentiel en sciences sociales, en économie, en biologie et en apprentissage automatique. Elle permet non seulement de prédire une variable d'intérêt, mais aussi d'évaluer l'importance relative de plusieurs facteurs explicatifs dans un phénomène donné.

Dans un contexte moderne, la régression linéaire est également la base de nombreuses techniques statistiques avancées, comme les modèles linéaires généralisés, la régression ridge et elle est au cœur de nombreux algorithmes d'apprentissage supervisé.

Ce mémoire, structuré en trois chapitres, a pour objectif d'étudier :

Régression Linéaire Simple

Dans ce chapitre, nous présentons brièvement les formules de la régression linéaire simple, qui permettent d'expliquer les hypothèses nécessaires ainsi que les différents termes du modèle. Nous abordons les notions d'estimation des paramètres du modèle par la méthode des moindres carrés ordinaires, l'estimation de l'intervalle de confiance, les tests de signification des paramètres et la signification globale du modèle. Enfin, une attention particulière est portée aux prévisions.

Régression Linéaire Multiple

Ce chapitre est consacré à la régression multiple, qui constitue une généralisation naturelle de la régression simple. L'objectif est d'expliquer les valeurs prises par la variable endogène Y en fonction de p variables exogènes $X_1, \dots, X_p$. La principale différence réside dans le formalisme, qui fait appel aux écritures matricielles des estimateurs et de leurs variances.

Application

Le dernier chapitre est dédié à l'application du modèle de régression linéaire multiple, traité dans le chapitre 2. L'objectif est d'établir une relation entre le volume des marchandises au port de Mostaganem et les variables suivantes : le nombre de navires arrivant au port chaque année, la taille moyenne des navires (en tonnage), et un indicateur économique global.

Régression linéaire Simple

La régression linéaire appartient aux méthodes d'analyse multivariée utilisées pour traiter des données quantitatives. Il s'agit d'une approche basée sur l'observation ou l'expérimentation, visant principalement à identifier une relation linéaire entre une variable dépendante Y quantitative et une variable indépendante X , également quantitative.

Ce chapitre présente une introduction à la modélisation linéaire en se concentrant sur le modèle fondamental : la régression linéaire simple (RLS). Cette méthode vise à représenter la relation entre deux variables quantitatives par une équation linéaire, permettant ainsi de capturer au mieux leur connexion.

Dans un premier temps, nous explorerons l'estimation des paramètres du modèle à travers deux approches. D'une part, la méthode des moindres carrés ordinaires (MCO), qui constitue une approche classique. D'autre part, l'approche inférentielle. Nous introduirons également les intervalles de confiance comme outil essentiel pour quantifier l'incertitude liée aux estimations.

Par la suite, nous analyserons la signification statistique de chaque paramètre individuel afin de déterminer leur contribution significative au modèle. Enfin, nous évaluerons la pertinence globale du modèle en termes de sa capacité explicative et de sa validité

1.1 Exemples concrets sur le modèle RLS

1.1.1 Relation entre le nombre de tests effectués et le nombre de cas confirmés

Une étude menée par l'Organisation mondiale de la santé (OMS) et d'autres institutions a montré que le nombre de tests effectués dans une population est directement lié au nombre de cas confirmés de COVID-19.

Les données proviennent du [Our World in Data](#), une plateforme qui collecte et publie des données ouvertes sur la pandémie de COVID-19.

1.1.2 Consommation d'alcool et espérance de vie par pays

La consommation d'alcool est souvent associée à des effets négatifs sur la santé, ce qui pourrait influencer l'espérance de vie. Cependant, la relation exacte entre ces deux variables peut varier selon les contextes socio-économiques et culturels.

Les données issues du rapport de l'OMS datant de février 2011 sur la consommation d'alcool (en L d'alcool pur par adulte de 15 ans ou plus) en projection pour l'année 2008, et l'espérance de vie en 2009 par pays.

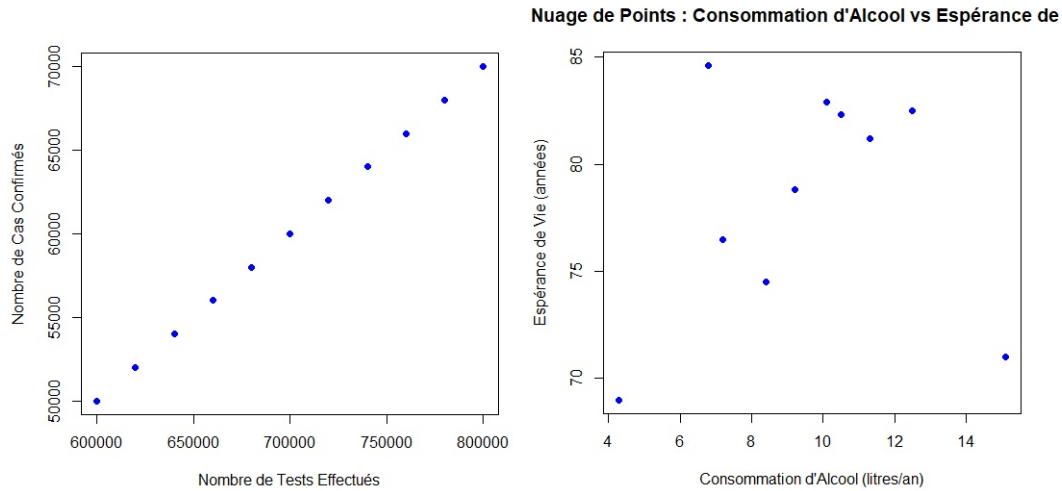


FIGURE 1.1 – Représentations graphiques des nuages de points

1.2 Modèle de régression linéaire simple

L'objectif est d'expliquer les variations d'une variable Y (appelée variable endogène, dépendante ou à prédire) en fonction des fluctuations d'une fonction linéaire d'une variable explicative X (également désignée comme variable exogène, indépendante ou prédictive).

Définition 1.2.1. Le modèle de régression linéaire simple (RLS), ou modèle linéaire simple, est défini par.

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \forall i \in \{1 \dots n\},$$

où

- x_i est une variable déterministe appelée *régressseur* ou *variable explicative*,
- Y_i est une variable aléatoire, appelée *variable à expliquer*,
- β_0 et β_1 sont des paramètres réels inconnus appelés *paramètres de régression* ou *coefficients de régression*,
- les ε_i sont des variables aléatoires indépendantes des x_i , non observées, appelées *erreurs* ou *bruits*

Remarque 1.2.1. Le aléatoire ε joue un rôle fondamental en régression. Il sert à condenser l'ensemble des informations non prises en compte dans la relation

linéaire recherchée entre Y et X . Par exemple, si vous modélisez le salaire (Y) en fonction de l'expérience professionnelle (X), le terme aléatoire pourrait capturer des facteurs tels que l'éducation, le secteur d'activité, ou encore des événements personnels qui influencent le salaire mais ne sont pas explicitement pris en compte dans le modèle.

Les hypothèses associées à ce modèle sont les suivantes :

- **(H₁) Espérance nulle :**

$$\forall i \in \{1 \dots n\}, \quad \mathbb{E}[\varepsilon_i] = 0$$

Cela signifie que le terme aléatoire n'introduit pas de biais en moyenne dans la relation entre X et Y .

- **(H₂) Homoscédasticité :**

$$\forall i \in \{1 \dots n\}, \quad \text{var}(\varepsilon_i) = \sigma^2$$

Cela indique que la dispersion des résidus est uniforme autour de la ligne de régression.

- **(H₃) l'absence de corrélation**

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

- **(H₄) Normalité (dans certains cas) :** Lors de tests statistiques basés sur la distribution normale (tels que les intervalles de confiance ou les tests d'hypothèse), on suppose souvent que :

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Ce modèle est appelé **modèle de régression linéaire simple**.

1.2.1 Formulation vectorielle

$$Y = \beta_0 \mathbf{1} + \beta_1 x + \varepsilon = X\beta + \varepsilon.$$

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Les conditions (H₁) à (H₃) se traduisent par :

- $\mathbb{E}[\varepsilon] = 0$ et $\mathbb{E}[Y] = \beta_0 \mathbf{1} + \beta_1 x = X\beta$.
- $\text{Var}(\varepsilon) = \text{Var}(Y) = \sigma^2 I_n$.

1.3 Estimation des paramètres β_0 et β_1 par approche non inférentielle

1.3.1 Moindres Carrés Ordinaires

Définition 1.3.1. (*Estimateurs des Moindres Carrés Ordinaires*) On appelle estimateurs des Moindres Carrés Ordinaires (notée par MCO) $\hat{\beta}_0$ et $\hat{\beta}_1$ les valeurs qui minimisent la quantité :

$$S(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

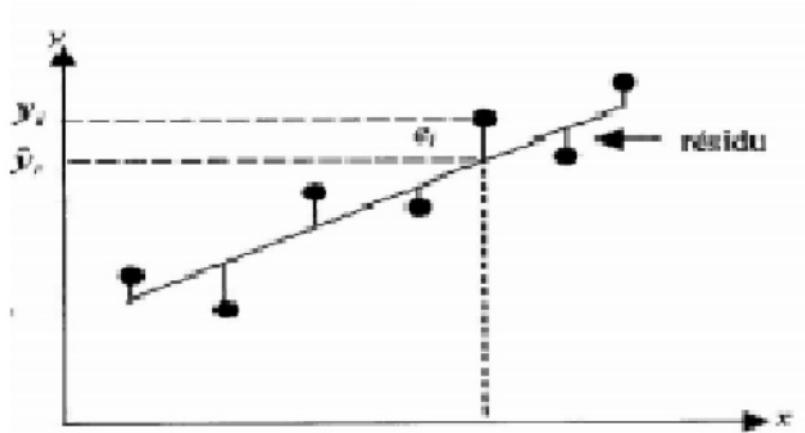


FIGURE 1.2 – Droite et résidu de la régression linéaire

Proposition 1.3.1. (Estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$) Les estimateurs des MCO ont pour expressions :

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

Preuve. On remarque que la fonction $S(\beta_0, \beta_1)$ est strictement convexe, donc qu'elle admet un minimum en un unique point $(\hat{\beta}_0, \hat{\beta}_1)$, lequel est déterminé en annulant les dérivées partielles de S . On obtient les équations suivantes :

$$\begin{cases} \frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

La première équation donne :

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n x_i &= \frac{1}{n} \sum_{i=1}^n y_i \\ \Rightarrow \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

où $\bar{x}$ et $\bar{y}$ sont comme d'habitude les moyennes empiriques des x_i et des y_i . La seconde équation donne :

$$\hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \quad (1.1)$$

En remplaçant $\hat{\beta}_0$ dans l'équation (1.1), on obtient :

$$\begin{aligned} (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i \\ \Rightarrow \bar{y} \sum_{i=1}^n x_i - \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i \\ \Rightarrow \hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \bar{x} \right) &= \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} \\ \Rightarrow \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \bar{x}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

car :

$$\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

et

$$\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i = \sum_{i=1}^n x_i^2 - n \bar{x}^2 = \sum_{i=1}^n (x_i - \bar{x})^2$$

Représentation géométrique

Si nous considérons le problème sous un angle vectoriel, nous disposons de deux vecteurs : le vecteur $\mathbf{X} = [x_1, \dots, x_n]'$ représentant les n observations de la variable explicative, et le vecteur $\mathbf{Y} = [y_1, \dots, y_n]'$ représentant les n observations de la variable à expliquer. Ces deux vecteurs appartiennent au même espace $\mathbb{R}^n$, appelé espace des variables. En ajoutant le vecteur $\mathbf{1} = [1, \dots, 1]'$, on observe que, par hypothèse, tous les x_i ne sont pas égaux, ce qui implique que les vecteurs $\mathbf{1}$ et $\mathbf{X}$ ne sont pas colinéaires. Ils engendrent donc un sous-espace de $\mathbb{R}^n$ de dimension 2, noté $M(\mathbf{X})$. On peut alors projeter orthogonalement le vecteur $\mathbf{Y}$ sur ce sous-espace $M(\mathbf{X})$. Notons provisoirement $\tilde{\mathbf{Y}}$ cette projection. Comme $(\mathbf{1}, \mathbf{X})$ forme une base de $M(\mathbf{X})$, il existe une unique décomposition de la forme $\tilde{\mathbf{Y}} = \tilde{\beta}_0 \mathbf{1} + \tilde{\beta}_1 \mathbf{X}$. Par définition de la projection orthogonale, $\tilde{\mathbf{Y}}$ est l'unique vecteur de $M(\mathbf{X})$ qui minimise la distance euclidienne $\|\mathbf{Y} - \tilde{\mathbf{Y}}\|$, ce qui équivaut à minimiser son carré. Or, par définition de la norme euclidienne, cette quantité correspond à :

$$\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^2 = \sum_{i=1}^n \left(y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_i) \right)^2$$

ce qui nous ramène à la méthode des moindres carrés ordinaires. On en déduit que $\tilde{\beta}_0 = \hat{\beta}_0$, $\tilde{\beta}_1 = \hat{\beta}_1$ et $\tilde{\mathbf{Y}} = \hat{\mathbf{Y}} = [\hat{y}_1, \dots, \hat{y}_n]'$, avec les expressions de $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\mathbf{Y}}$ vues précédemment.

[h] Autrement dit, dans $\mathbb{R}^n$, $\hat{\beta}_0$ et $\hat{\beta}_1$ s'interprètent comme les coordonnées de la projection orthogonale $\hat{\mathbf{Y}}$ de $\mathbf{Y}$ sur le sous-espace de $\mathbb{R}^n$ engendré par $\mathbf{1}$ et $\mathbf{X}$.

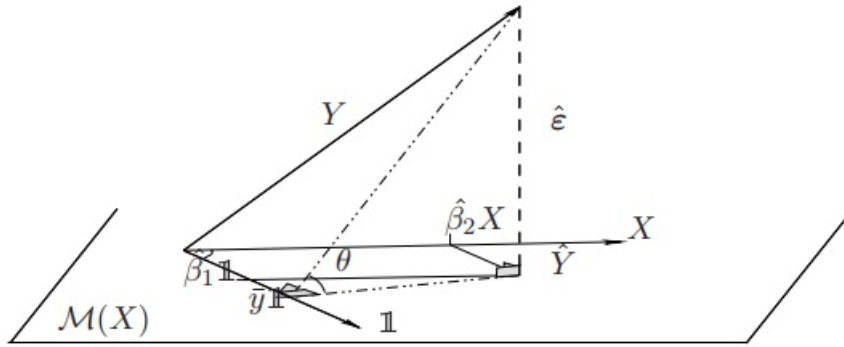


FIGURE 1.3 – Représentation de la projection dans l'espace des variables

- Remarque 1.3.1.** 1. Cette interprétation géométrique peut sembler quelque peu abstraite, mais elle constitue en réalité une approche puissante pour comprendre la régression multiple, comme nous le verrons dans les chapitres suivants.
2. Nous avons supposé que 1 et X ne sont pas colinéaires. En général, ces vecteurs ne sont pas orthogonaux (sauf si $\bar{x} = 0$). Cela implique que $\hat{\beta}_0 1$ n'est pas la projection orthogonale de Y sur 1 (qui est en réalité $\bar{y}1$), et que $\hat{\beta}_1 X$ n'est pas la projection orthogonale de Y sur X (qui est donnée par $\frac{\langle Y, X \rangle}{\|X\|^2} X$).

Propriétés des estimateurs

Théorème 1.3.1. Les estimateurs des MCO vérifient les propriétés suivantes.

1. $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs linéaires en $(y_i)_{i=1, \dots, n}$, sans biais de β_0 et β_1 respectivement
2. $\text{var}(\hat{\beta}_0) = \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2$;
3. $\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$;
4. $\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2$.

Preuve.

$$\frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n w_i y_i, \quad \text{avec} \quad w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

De même

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum_{i=1}^n \left(\frac{1}{n} - \bar{x} w_i \right) y_i$$

donc les estimateurs des MCO sont bien des estimateurs linéaires. par suite, on note au préalable que :

- $\sum_{i=1}^n w_i = \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0,$
- $\sum_{i=1}^n w_i x_i = \sum_{i=1}^n \frac{x_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 1$
- $\sum_{i=1}^n w_i^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}.$

Alors :

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\sum_{i=1}^n w_i Y_i\right] = \sum_{i=1}^n w_i (\beta_0 + \beta_1 x_i) = \beta_1 \sum_{i=1}^n w_i x_i = \beta_1$$

et

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{Y} - \hat{\beta}_1 \bar{x}] = \mathbb{E}[\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}] - \beta_1 \bar{x} = \beta_0.$$

En outre

$$\begin{aligned} \text{var}(\hat{\beta}_0) &= \sum_{i=1}^n \left(\frac{1}{n^2} - \frac{2\bar{x}w_i}{n} + \bar{x}^2 w_i^2 \right) \sigma^2 \\ &= \left(\frac{1}{n} - 0 + \bar{x}^2 \sum_{i=1}^n w_i^2 \right) \sigma^2 \\ &= \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \sigma^2 \\ &= \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2 \end{aligned}$$

et

$$\text{var}(\hat{\beta}_1) = \sum_{i=1}^n w_i^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\begin{aligned} \text{cov}(\hat{\beta}_0, \hat{\beta}_1) &= \text{cov}\left(\sum_{i=1}^n \left(\frac{1}{n} - \bar{x}w_i\right) Y_i, \sum_{i=1}^n w_i Y_i\right) \\ &= \sum_{i=1}^n \sum_{j=1}^n \left(\frac{1}{n} - \bar{x}w_i\right) w_j \text{cov}(Y_i, Y_j) \\ &= \sum_{i=1}^n \left(\frac{1}{n} - \bar{x}w_i\right) w_i \sigma^2 \\ &= -\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2 \end{aligned}$$

Théorème 1.3.2. (Gauss-Markov) Parmi les estimateurs sans biais linéaires en y , les estimateurs $\hat{\beta}_j$ sont des variances minimales.

Preuve. L'estimateur des MCO s'écrit $\hat{\beta}_1 = \sum_{i=1}^n w_i y_i$, avec $w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$. Considérons un autre estimateur $\tilde{\beta}_1$ linéaire en y_i et sans biais, c'est-à-dire :

$$\tilde{\beta}_1 = \sum_{i=1}^n \lambda_i y_i.$$

Montrons que $\sum_{i=1}^n \lambda_i = 0$ et $\sum_{i=1}^n \lambda_i x_i = 1$. L'égalité

$$\mathbb{E}(\tilde{\beta}_1) = \beta_0 \sum_{i=1}^n \lambda_i + \beta_1 \sum_{i=1}^n \lambda_i x_i + \sum_{i=1}^n \lambda_i \mathbb{E}(\varepsilon_i) = \beta_0 \sum_{i=1}^n \lambda_i + \beta_1 \sum_{i=1}^n \lambda_i x_i$$

est vraie pour tout β_1 . L'estimateur $\tilde{\beta}_1$ est sans biais donc $\mathbb{E}(\tilde{\beta}_1) = \beta_1$ pour tout β_1 , c'est-à-dire que $\sum_{i=1}^n \lambda_i = 0$ et $\sum_{i=1}^n \lambda_i x_i = 1$. Montrons que $\text{Var}(\tilde{\beta}_1) \geq \text{Var}(\hat{\beta}_1)$:

$$\text{Var}(\tilde{\beta}_1) = \text{Var}(\tilde{\beta}_1 - \hat{\beta}_1 + \hat{\beta}_1) = \text{Var}(\tilde{\beta}_1 - \hat{\beta}_1) + \text{Var}(\hat{\beta}_1) + 2\text{Cov}(\tilde{\beta}_1 - \hat{\beta}_1, \hat{\beta}_1).$$

Or :

$$\text{Cov}(\tilde{\beta}_1 - \hat{\beta}_1, \hat{\beta}_1) = \text{Cov}(\tilde{\beta}_1, \hat{\beta}_1) - \text{Var}(\hat{\beta}_1) = \sigma^2 \frac{\sum_{i=1}^n \lambda_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0$$

la dernière égalité étant due aux deux relations $\sum_{i=1}^n \lambda_i = 0$ et $\sum_{i=1}^n \lambda_i x_i = 1$. Ainsi :

$$\text{Var}(\tilde{\beta}_1) = \text{Var}(\tilde{\beta}_1 - \hat{\beta}_1) + \text{Var}(\hat{\beta}_1).$$

Une variance est toujours positive, donc :

$$\text{Var}(\tilde{\beta}_1) \geq \text{Var}(\hat{\beta}_1).$$

Le résultat est démontré. On obtiendrait la même chose pour $\hat{\beta}_0$.

Estimation ponctuelle de la variance ,valeur ajustées et résidus

$$V(\hat{\beta}) = \sigma^2 \begin{pmatrix} \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} & \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{pmatrix}$$

Les ε_i ne sont pas observés donc pour pouvoir estimer la variance σ^2 , on introduit les résidus.

Définition 1.3.2. Les résidus sont définis comme les quantité $\hat{\varepsilon} = y_i - \hat{y}_i$, pour tout $i = 1 \dots n$ ou $\hat{y}_i$ représente les valeurs ajustées obtenues par modèle $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Propriétés des résidus

Les $\hat{\varepsilon}_i$ sont des variables aléatoires déterministe, centrés, de somme nulle, c'est à dire $\sum_{i=1}^n \hat{\varepsilon}_i = 0$ (donc indépendance).

Preuve. On a :

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

Théorème 1.3.3. (Estimateur non biaisé de σ^2) La statistique de l'estimateur non biaisé de la variance est donnée par :

$$s^2 = \hat{\sigma}^2 = \frac{1}{(n-2)} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

est un sans biais de σ^2

Preuve. Récrivons les résidus en constatant que $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ et $\beta_0 = \bar{y} - \beta_1 \bar{x} - \bar{\varepsilon}$ ce qui donne :

$$\begin{aligned}\hat{\varepsilon}_i &= \beta_0 + \beta_1 x_i + \varepsilon_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= \bar{y} - \beta_1 \bar{x} - \bar{\varepsilon} + \beta_1 x_i + \varepsilon_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i \\ &= (\beta_1 - \hat{\beta}_1)(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}).\end{aligned}$$

En utilisant l'expression suivante :

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

nous avons :

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

En simplifiant, on obtient :

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

Prenons l'espérance :

$$\mathbb{E} \left[\sum_{i=1}^n \hat{\varepsilon}_i^2 \right] = \mathbb{E} \left[\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 \right] - \sum_{i=1}^n (x_i - \bar{x})^2 \cdot \text{Var}(\hat{\beta}_1) = (n-2)\sigma^2.$$

Approche inférentielle

Dans cette section, nous supposons que les erreurs ε_i suivent une loi gaussienne (Hypothèse H_4). (1.2)

Le modèle de régression simple est devenu un modèle paramétrique, caractérisé par les paramètres $(\beta_0, \beta_1, \sigma^2)$ où :

$$(\beta_0, \beta_1, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+^*.$$

Connaissant la loi des résidus ε_i , on en déduit la loi des observations y_i :

$$\forall i \in \{1, \dots, n\}, \quad y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

De plus, les y_i sont mutuellement indépendants, car les ε_i le sont. Cela permet de calculer la vraisemblance de l'échantillon et de déterminer les estimateurs du maximum de vraisemblance. Cette démarche sera détaillée dans la partie suivante.

1.3.2 Estimateurs du maximum de vraisemblance

La vraisemblance du modèle s'écrit :

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2, y_1, \dots, y_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right)$$

d'où

$$\ln \mathcal{L}(\beta_0, \beta_1, \sigma^2, y_1, \dots, y_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

On a :

σ^2 fixé et $\ln \mathcal{L}(\beta_0, \beta_1, \sigma^2, y_1, \dots, y_n)$ est maximale si $\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ est minimale. Les estimateurs du maximum de vraisemblance de β_0 et β_1 sont donc égaux aux estimateurs des moindres carrés ordinaires.

De plus :

$$\frac{\partial \ln \mathcal{L}}{\partial \sigma^2}(\beta_0, \beta_1, \sigma^2, y_1, \dots, y_n) = 0$$

si et seulement si

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2.$$

L'estimateur du maximum de vraisemblance $\hat{\sigma}^2$ de σ^2 est biaisé

1.4 Décomposition de la variance et tableau d'ANOVA

Théorème 1.4.1. La somme des carrés totale (SCT) est égale à la somme des carrés expliquée (SCE) plus la somme des carrés résiduelle (SCR)

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{SCT}} = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{SCR}} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{SCE}}.$$

où :

- SCT : Somme des carrés totale.
- SCE : Somme des carrés expliquée.
- SCR : Somme des carrés résiduelle.

Preuve. On a :

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

En élevant au carré et en sommant :

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

On a :

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

Or, on sait que $\sum_{i=1}^n \varepsilon_i = 0$, alors $\sum_{i=1}^n (\hat{y}_i - \bar{y}) \varepsilon_i = 0$

Finalement, on obtient : $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$

On produit le tableau d'analyse de variance à partir de la décomposition de la variance

Source de variation	ddl	Somme des carrés	Carrés moyens
Expliquée	1	$SCE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$CME = \frac{SCE}{1}$
Résiduelle	$n - 2$	$SCR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$CMR = \frac{SCR}{n-2}$
Totale	$n - 1$	$SCT = \sum_{i=1}^n (y_i - \bar{y})^2$	

1.4.1 Le coefficient de détermination

Définition 1.4.1. : Le coefficient de détermination R^2 est une fraction de la variabilité totale expliquée par la régression :

$$R^2 = \frac{SCE}{SCT} = \frac{\|\hat{Y} - \bar{y}1\|^2}{\|Y - \bar{y}1\|^2} = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{y}1\|^2} = 1 - \frac{SCR}{SCT}.$$

On voit sur la figure (1.3) que R^2 correspond au cosinus carré de l'angle θ . De façon schématique, on peut différencier les cas suivants :

- Si $R^2 = 1$, le modèle explique tout, l'angle θ vaut zéro et Y est dans $M(X)$, c'est-à-dire que $y_i = \beta_1 + \beta_2 x_i$ pour tout i : les points de l'échantillon sont parfaitement alignés sur la droite des moindres carrés.

- Si $R^2 = 0$, cela signifie que $\sum(\hat{y}_i - \bar{y})^2 = 0$, donc $\hat{y}_i = \bar{y}$ pour tout i . Le modèle de régression linéaire est inadapte puisqu'il ne modélise rien de mieux que la moyenne.
- Si R^2 est proche de zéro, cela signifie que Y est quasiment dans l'orthogonal de $M(X)$. Le modèle de régression linéaire est alors inadapte, car la variable x n'explique pas bien la variable réponse y (du moins pas de façon affine).

De façon générale, l'interprétation est la suivante : le modèle de régression linéaire permet d'expliquer $100 \times R^2\%$ de la variance totale des données.

1.5 Test de signification

1.5.1 Test de signification β_1

On souhaite de tester :

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

Sous les hypothèses (H_1) et (H_0) (1.2), on a des propriétés :

1. $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$
2. $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2)$
3. $\hat{\beta}_1$ et S^2 sont indépendants.

On déduit alors des propriétés (1) et (2) que :

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathcal{N}(0, 1)$$

D'après l'hypothèse H_0 on a :

La statistique de test :

$$T_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \sim \mathcal{T}(n-2)$$

où :

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

On rejette H_0 si l'observation de la statistique de test, notée $T_{\hat{\beta}_1}$, est telle que :

$$|T_{\hat{\beta}_1}| \geq t_{n-2}^{1-\frac{\alpha}{2}}$$

où $t_{n-2}^{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ d'une loi Student à $(n-2)$ ddl.

Remarque 1.5.1. Si $n \geq 30$, loi de student tend vers la loi normale.

1.5.2 Test de signification globale du modèle

L'objectif de ce test est de vérifier si la variable X a un effet significatif sur Y ou non. On veut tester l'hypothèse :

$$\begin{cases} H_0 : \beta_1 = 0 & (\text{absence de relation}) \\ H_1 : \beta_1 \neq 0 & (\text{relation significative}) \end{cases}$$

Pour tester cette hypothèse, on se base sur la statistique de Fisher, notée par F , qui est donnée par la formule :

$$F = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}}$$

Cette statistique permet de déterminer si la variance expliquée dépasse significativement la variance résiduelle. Si c'est le cas, on peut estimer que la régression révèle une relation authentique existant dans la population. Sous H_0 , SCE est distribué selon une loi $\chi^2(1)$ et SCR selon une loi $\chi^2(n-2)$. De fait, pour F , on a :

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

Alors, sous H_0 , F est distribué selon une loi de Fisher à $(1, n-2)$ degrés de liberté, et on rejette H_0 si :

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

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

1.6 Intervalles de confiance et intervalle de prévision

1.6.1 Intervalles de confiance pour les coefficients de régression

Théorème 1.6.1. : Soit $\alpha \in]0, 1[$. On note $t_{n-2}(u)$ le u -quantile de la loi $\mathcal{T}(n-2)$. Un intervalle de confiance de niveau de confiance $(1-\alpha)$ pour β_0 est donné par :

$$\left[\hat{\beta}_0 - t_{n-2}^{1-\frac{\alpha}{2}} \sqrt{\frac{s^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{\beta}_0 + t_{n-2}^{1-\frac{\alpha}{2}} \sqrt{\frac{s^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}} \right]$$

Un intervalle de confiance de niveau de confiance $(1-\alpha)$ pour β_1 est donné par :

$$\left[\hat{\beta}_1 - t_{n-2}^{1-\frac{\alpha}{2}} \sqrt{\frac{s^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{\beta}_1 + t_{n-2}^{1-\frac{\alpha}{2}} \sqrt{\frac{s^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}} \right]$$

1.6.2 Prévision

L'un des objectifs de la régression est de réaliser des prévisions, c'est-à-dire de prédire la variable à expliquer y pour une nouvelle valeur de la variable explicative x .

Soit x_{n+1} une nouvelle valeur pour laquelle on souhaite prédire y_{n+1} . Le modèle reste le même :

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

avec les hypothèses suivantes :

- $\mathbb{E}[\varepsilon_{n+1}] = 0$,
- $\text{Var}(\varepsilon_{n+1}) = \sigma_\varepsilon^2$,
- $\text{Cov}(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, \dots, n$.

Il est naturel de prédire la valeur correspondante y_{n+1} à l'aide du modèle ajusté :

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

Dans le cadre de la prévision avec des erreurs gaussiennes, les résultats concernant l'espérance et la variance restent valables. De plus, comme $\hat{y}_{n+1}$ est une combinaison linéaire de $\hat{\beta}_0$ et $\hat{\beta}_1$, et en tenant compte de ε_{n+1} , on peut déterminer la loi de l'erreur de prévision :

$$\hat{\varepsilon}_{n+1} = y_{n+1} - \hat{y}_{n+1} \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right).$$

Cette formulation permet de quantifier l'incertitude associée à la prévision.

On ne connaît pas σ_ε^2 , on l'estime donc par s_ε^2 .

Comme $(y_{n+1} - \hat{y}_{n+1})$ et $s_\varepsilon^2(n-2)/\sigma_\varepsilon^2$ sont indépendants, on peut énoncer un résultat donnant des intervalles de confiance pour y_{n+1} .

Avec les notations et hypothèses précédentes, on a :

$$\frac{y_{n+1} - \hat{y}_{n+1}}{s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathcal{T}_{n-2}$$

où $\mathcal{T}_{n-2}$ désigne la loi de Student à $n-2$ degrés de liberté. On en déduit l'intervalle de confiance suivant pour y_{n+1} :

$$\left[\hat{y}_{n+1} - t_{1-\alpha/2}^{n-2} \cdot s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad \hat{y}_{n+1} + t_{1-\alpha/2}^{n-2} \cdot s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]$$

où $t_{1-\alpha/2}^{n-2}$ est le quantile de la loi de Student à $n-2$ degrés de liberté pour un niveau de confiance $1-\alpha$

Régression linéaire Multiple

Dans ce chapitre nous généralisons et étendons les résultats précédents au cas plus intéressant où l'on cherche à expliquer une variable Y par un ensemble de variables X . De façon à simplifier la notation, on utilisera la notation matricielle. Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en oeuvre pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple.

2.1 Modèle de régression linéaire multiple

Écriture du modèle

Le modèle de régression linéaire multiple est une généralisation du modèle de régression simple lorsque les variables explicatives sont en nombre quelconque. Nous supposons donc que les données collectées suivent le modèle suivant :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n$$

où :

- les x_{ij} sont des nombres connus, non aléatoire.
- les paramètres β_j du modèle sont inconnus, mais non aléatoires ;
- les ε_i sont des variables aléatoires inconnues.

Définition 2.1.1. (Écriture matricielle) Un modèle de régression linéaire est défini par une équation de la forme :

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

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{n,p} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

où :

- Y est un vecteur aléatoire de dimension n ,
- X est une matrice de taille $n \times (p + 1)$ connue, appelée matrice du plan d'expérience,
- β est le vecteur de dimension $p + 1$ des paramètres inconnus du modèle
- ε est le vecteur de dimension n des erreurs.

Les hypothèses concernant le modèle sont

$$(\mathcal{C}) \begin{cases} (\mathcal{C}_1) : \mathbb{E}[\varepsilon] = 0, \text{Var}(Y) = \sigma^2 I_n, \\ (\mathcal{C}_2) : \text{rg}(X) = p + 1, \\ (\mathcal{C}_3) : \varepsilon \text{ est un vecteur gaussien.} \end{cases}$$

L'hypothèse $(\mathcal{C}_1)$ est signifie que les erreurs sont centrées, de même variance (homoscédasticité) et non corrélées entre elles. L'hypothèse $(\mathcal{C}_2)$ assure que la matrice X est bien inversible.

2.2 Estimation de paramètre β

Soit $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)$ le vecteur de dimension $n \times 1$ des résidus défini par :

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

où $\hat{Y} = X\hat{\beta}$ représente les valeurs estimées par le modèle, on les appelle aussi le vecteur des valeurs ajustées. La somme des carrés des résidus est donnée par :

$$\begin{aligned} SCR &= \sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\varepsilon}'\hat{\varepsilon} \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \end{aligned}$$

2.2.1 Estimateurs des Moindres Carrés Ordinaires

Comme pour la régression linéaire simple, nous allons considérer ici une fonction de coût quadratique, d'où la dénomination de Moindres Carrés Ordinaires (MCO).

Définition 2.2.1. L'estimateur des moindres carrés $\hat{\beta}$ est défini comme suit :

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 = \arg \min_{\beta \in \mathbb{R}^{p+1}} \|Y - X\beta\|^2.$$

Proposition 2.2.1. (Expression de $\hat{\beta}$) L'estimateur $\hat{\beta}$ des Moindres Carrés Ordinaires a pour expression :

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

Preuve. Nous cherchons β qui minimise la fonction $S(\beta)$:

$$\hat{\beta} = \arg \min_{(\beta_0, \beta_1, \dots, \beta_p)} S(\beta) = \arg \min_{(\beta_0, \beta_1, \dots, \beta_p)} \|Y - X\beta\|^2$$

On a :

$$\begin{aligned} S(\beta) &= (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta \\ &= \|Y\|^2 - 2Y'X\beta + \beta'X'X\beta \end{aligned}$$

Puisque $S(\beta)$ est une fonction quadratique en β , avec $X'X$ symétrique défini positive, le problème admet une solution unique $\hat{\beta}$. Calculons la dérivée :

$$\frac{\partial S}{\partial \beta} = -2X'Y + 2X'X\beta$$

On a :

$$\frac{\partial S}{\partial \beta} = 0 \Leftrightarrow -2X'Y + 2X'X\beta = 0$$

$$\Leftrightarrow X'X\beta = X'Y$$

Sous l'hypothèse (C) (2.1), nous avons :

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

Théorème 2.2.1. : (Propriétés des estimateurs MCO) L'estimateur $\hat{\beta}$ des moindres carrés ordinaires est sans biais, c'est-à-dire $E(\hat{\beta}) = \beta$, et sa matrice de variance-covariance, notée par $\text{varcov}(\hat{\beta})$:

$$\text{varcov}(\hat{\beta}) = \sigma_\varepsilon^2 (X'X)^{-1}.$$

Remarque 2.2.1. La matrice de variance-covariance $\text{varcov}(\hat{\beta})$ des coefficients, de dimension $(p+1, p+1)$, est donnée par :

$$\text{varcov}(\hat{\beta}) = \begin{pmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \dots & \text{cov}(\hat{\beta}_0, \hat{\beta}_p) \\ \dots & \text{var}(\hat{\beta}_1) & \dots & \text{cov}(\hat{\beta}_1, \hat{\beta}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \dots & \dots & \dots & \text{var}(\hat{\beta}_p) \end{pmatrix}$$

Cette matrice est symétrique, et sur sa diagonale principale, on observe les variances des coefficients estimés $\text{var}(\hat{\beta}_0), \dots, \text{var}(\hat{\beta}_p)$.

Preuve. Puisque $\hat{\beta} = (X'X)^{-1}X'Y$, il s'agit bien d'un estimateur linéaire (en y). De plus $E(\hat{\beta}) = \beta$

On a :

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'[X\beta + \varepsilon] \\ &= (X'X)^{-1}(X'X)\beta + (X'X)^{-1}X'\varepsilon \\ &= \beta + (X'X)^{-1}X'\varepsilon \end{aligned}$$

Alors, l'espérance mathématique de $\hat{\beta}$ est :

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= \mathbb{E}[\beta + (X'X)^{-1}X'\varepsilon] \\ &= \beta + \mathbb{E}[(X'X)^{-1}X'\varepsilon] \\ &= \beta + (X'X)^{-1}X'\mathbb{E}[\varepsilon], \quad \text{car } X \text{ est non aléatoire.}\end{aligned}$$

Et sous l'hypothèse que $\mathbb{E}(\varepsilon) = 0$, il vient que :

$$\mathbb{E}(\hat{\beta}) = \beta. \quad (\text{donc } \hat{\beta} \text{ est sans biais})$$

Montrons que :

$$\text{Varcov}(\hat{\beta}) = \sigma_\varepsilon^2 (X'X)^{-1}$$

On procède de même, on a $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$ donc :

$$\begin{aligned}\text{Varcov}(\hat{\beta}) &= \text{Var}(\beta + (X'X)^{-1}X'\varepsilon) \\ &= (X'X)^{-1}X'\text{Var}(\varepsilon)X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma_\varepsilon^2 I_n X(X'X)^{-1} \\ &= \sigma_\varepsilon^2 (X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma_\varepsilon^2 (X'X)^{-1}\end{aligned}$$

On trouve :

$$\text{Varcov}(\hat{\beta}) = \sigma_\varepsilon^2 (X'X)^{-1}.$$

Théorème 2.2.2. (Gauss-Markov) L'estimateur $\hat{\beta}$ des moindres carrés ordinaires est l'unique estimateur linéaire sans biais de variance minimale parmi les estimateurs linéaires sans biais de β .

Preuve.

Nous allons montrer que, pour tout estimateur $\tilde{\beta}$ de β linéaire et sans biais, $\text{Var}(\tilde{\beta}) \geq \text{Var}(\hat{\beta})$, où l'inégalité entre matrices de variance-covariance est à comprendre au sens précisé ci-dessus. Rappelons la formule générale pour la matrice de covariance de la somme de deux vecteurs aléatoires X et Y :

$$\text{Var}(U + V) = \text{Var}(U) + \text{Var}(V) + \text{Cov}(U, V) + \text{Cov}(V, U),$$

où $\text{Cov}(U, V) = \mathbb{E}[UV'] - \mathbb{E}[U]\mathbb{E}[V]' = \text{Cov}(V, U)'$. Décomposons ainsi la variance de $\tilde{\beta}$:

$$\text{Var}(\tilde{\beta}) = \text{Var}(\tilde{\beta} - \hat{\beta} + \hat{\beta}) = \text{Var}(\tilde{\beta} - \hat{\beta}) + \text{Var}(\hat{\beta}) + \text{Cov}(\tilde{\beta} - \hat{\beta}, \hat{\beta}) + \text{Cov}(\hat{\beta}, \tilde{\beta} - \hat{\beta}).$$

Les variances étant semi-définies positives, si nous montrons que $\text{Cov}(\tilde{\beta} - \hat{\beta}, \hat{\beta}) = 0$, nous aurons fini la démonstration. Puisque $\tilde{\beta}$ est linéaire, $\tilde{\beta} = AY$. De plus, nous savons qu'il est sans biais, c'est-à-dire $\mathbb{E}[\tilde{\beta}] = \beta$ pour tout β , donc $AX = I$. La covariance devient :

$$\begin{aligned}\text{Cov}(\tilde{\beta} - \hat{\beta}, \hat{\beta}) &= \text{Cov}(AY, (X'X)^{-1}X'Y) - \text{Var}(\hat{\beta}) \\ &= \sigma^2 AX(X'X)^{-1} - \sigma^2 (X'X)^{-1} = 0.\end{aligned}$$

2.2.2 Estimation de variance du résidu

Lemme 2.2.1. Soit un vecteur Z composé de n variables aléatoires d'espérances nulles, et tel que $\text{var}(Z) = \sigma_\varepsilon^2 I_n$, et A une matrice symétrique non aléatoire, alors

$$\mathbb{E} [Z'AZ] = \sigma_\varepsilon^2 \text{tr}(A)$$

où $\text{tr}(A)$ est la trace de la matrice A .

Théorème 2.2.3. Soit $\hat{\varepsilon} = Y - X\hat{\beta}$, alors

$$\mathbb{E} [\hat{\varepsilon}'\hat{\varepsilon}] = (n - (p + 1))\sigma_\varepsilon^2$$

Preuve. Soit

$$\begin{aligned} \hat{Y} &= X\hat{\beta} \\ &= \left[X(X'X)^{-1}X' \right] Y = HY \quad \text{avec} \quad H = X(X'X)^{-1}X' \end{aligned}$$

La différence entre les valeurs observées Y et les valeurs ajustées $\hat{Y}$ porte le nom vecteurs des résidus, noté par $\hat{\varepsilon}$ est donné par la relation :

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

où I_n est la matrice unité de dimension (n, n) . La matrice H est appelée la matrice chapeau de taille $n \times n$. Elle vérifie les deux propriétés :

$$\begin{cases} \text{Symétrique :} & H' = H \\ \text{Idempotente :} & H^2 = H \end{cases}$$

La matrice $I_n - H$ a les mêmes propriétés :

$$\begin{cases} (I_n - H)' = I_n - H \\ (I_n - H)^2 = I_n - H \end{cases} \quad (2.1)$$

$$\hat{\varepsilon} = Y - \hat{Y} = Y - HY = (I_n - H)(X\beta + \varepsilon) = (I_n - H)X\beta + (I_n - H)\varepsilon$$

où $HX = (X(X'X)^{-1}X')X = X \Rightarrow (I - H)X = 0$
ce qui donne

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

On obtient :

$$\begin{aligned} \hat{\varepsilon}'\hat{\varepsilon} &= ((I_n - H)\varepsilon)'(I_n - H)\varepsilon \\ &= \varepsilon'(I_n - H)'(I_n - H)\varepsilon \\ &= \varepsilon'(I_n - H)\varepsilon, \quad (\text{d'après l'Équation 2.1}) \\ &= \varepsilon'I_n\varepsilon - \varepsilon'H\varepsilon \end{aligned}$$

Donc, par le lemme, on obtient :

$$\mathbb{E} [\hat{\varepsilon}'\hat{\varepsilon}] = \mathbb{E} [\varepsilon'I_n\varepsilon - \varepsilon'H\varepsilon] = \mathbb{E} [\varepsilon'I_n\varepsilon] - \mathbb{E} [\varepsilon'H\varepsilon]$$

$$= \sigma_\varepsilon^2 \text{tr}(I_n) - \sigma_\varepsilon^2 \text{tr}(H) = [\text{tr}(I_n) - \text{tr}(H)] \sigma_\varepsilon^2$$

où $\text{tr}(I_n) = n$ et

$$\begin{aligned} \text{tr}(H) &= \text{tr}(X(X'X)^{-1}X') \\ &= \text{tr}(X'X(X'X)^{-1}) \quad \text{car } \text{tr}(AB) = \text{tr}(BA) \\ &= \text{tr}(I_{p+1}) \quad \text{car } X'X \text{ est une matrice carrée de taille } (p+1) \\ &= p+1 \end{aligned}$$

Donc

$$\mathbb{E} [\hat{\varepsilon}'\hat{\varepsilon}] = (n - (p+1))\sigma_\varepsilon^2.$$

D'après le théorème (2.2.3), on peut construire l'estimateur sans biais pour σ_ε^2 qui est :

$$s_\varepsilon^2 = \hat{\sigma}_\varepsilon^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n - (p+1)} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n - (p+1)} = \frac{SCR}{n - (p+1)}$$

2.2.3 Estimateurs du maximum de vraisemblance

La vraisemblance du modèle s'écrit :

$$\begin{aligned} \mathcal{L}(\beta, \sigma^2, Y_1, \dots, Y_n) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left(-\frac{1}{2\sigma^2} \|Y - X\beta\|^2 \right) \end{aligned}$$

d'où

$$\ln \mathcal{L}(\beta, \sigma^2, Y_1, \dots, Y_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \|Y - X\beta\|^2.$$

On cherche les estimateurs $\hat{\beta}_{mv}$ et $\hat{\sigma}_{mv}^2$ qui maximisent cette ln-vraisemblance. Il est clair qu'il faut minimiser la quantité $\|Y - X\beta\|^2$, ce qui est justement le principe des moindres carrés ordinaires, donc :

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

et :

$$\frac{\partial \ln \mathcal{L}(Y, \hat{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \|Y - X\hat{\beta}\|^2$$

d'où il vient :

$$\hat{\sigma}_{mv}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n} = \frac{\|\hat{\varepsilon}\|^2}{n}$$

On remarque que l'estimateur du maximum de vraisemblance de β est égal à l'estimateur des MCO $\hat{\beta}$ déterminé dans le chapitre précédent. Il est donc linéaire, sans biais de variance minimale parmi les estimateurs linéaires sans biais (théorème de Gauss-Markov), et l'estimateur $\hat{\sigma}_{mv}^2$ du maximum de vraisemblance est biaisé, mais d'autant moins que le nombre de variables explicatives est petit devant le nombre n d'observations. Dans la suite, nous continuerons à considérer l'estimateur $\hat{\sigma}^2$ des moindres carrés vu au chapitre précédent et nous conserverons aussi la notation adoptée pour les résidus $\hat{\varepsilon}_i$, de sorte que :

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

2.3 Analyse de la variance et coefficient de détermination

2.3.1 Décomposition de la variance et tableau d'ANOVA

On peut aisément vérifier que, comme dans le modèle de régression simple, l'écart entre y_i et la moyenne des y_i en ajoutant puis retranchant $\hat{y}_i$ la valeur estimée de y par la droite de régression. Cette procédure fait apparaître une somme de deux écarts :

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

On peut obtenir la décomposition :

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$SCT = SCE + SCR$$

$$= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \varepsilon_i^2$$

Où :

- **SCT** : désigne la somme des carrés totaux .
- **SCE** : la somme des carrés expliqués .
- **SCR** : la somme des carrés des résidus.

Le tableau "d'analyse de la variance" se présente sous la forme suivante (voir le tableau).

Source de variation	ddl	Somme des carrés	Carrés moyens
Expliquée	p	$SCE = \sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2$	$CME = \frac{SCE}{p}$
Résiduelle	$n - (p + 1)$	$SCR = \sum_{i=1}^n (y_i - \hat{y})^2$	$CMR = \frac{SCR}{n - (p + 1)}$
Totale	$n - 1$	$SCT = \sum_{i=1}^n (y_i - \bar{y})^2$	

Définition 2.3.1. Le coefficient de détermination R^2 est défini par :

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n \varepsilon_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{SCR}{SCT}.$$

Le coefficient R^2 varie entre 0 et 1 : plus il est proche de 1, plus la part de la variance expliquée par le modèle est importante, ce qui traduit une bonne qualité de la régression. À l'inverse, un coefficient R^2 proche de 0 indique que la somme des carrés des résidus (SCR) est élevée, suggérant une mauvaise capacité explicative du modèle.

2.4 Test de signification

2.4.1 Test sur les paramètres du modèle

Soit $j \in \{0, \dots, p\}$. Le test de student permet d'évaluer l'influence de X_j sur Y . On considère les hypothèses suivantes :

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

où β_j est le paramètre associé à la variable explicative X_j . L'hypothèse H_0 de nullité d'un paramètre du modèle peut être testée au moyen de la statistique de student suivante :

$$T_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\hat{s}_{\beta_j}} \sim \mathcal{T}(n - (p + 1))$$

à comparer avec $t_{1-\alpha/2}^{n-p-1}$, où $t_{1-\alpha/2}^{n-p-1}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ d'une loi de student à $n - (p + 1)$ degrés de liberté (ddl).

- Si $|T_{\hat{\beta}_j}| \geq t_{1-\alpha/2}^{n-p-1}$, on rejette H_0 .
- Si $|T_{\hat{\beta}_j}| < t_{1-\alpha/2}^{n-p-1}$, on accepte H_0 .

2.4.2 Test de signification globale du modèle

L'objectif du test global de Fisher est d'étudier la liaison globale entre Y et les variables explicatives X_j (avec $j = 1, \dots, p$). On considère les hypothèses suivantes :

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \dots = \beta_p = 0 \\ H_1 : \exists j \in \{1, \dots, p\}, \text{ tel que } \beta_j \neq 0 \end{cases}$$

Pour tester cette hypothèse, on utilise la statistique de Fisher F , définie par :

$$F = \frac{\frac{\text{SCE}}{p}}{\frac{\text{SCR}}{(n-(p+1))}}$$

Cette statistique permet de déterminer si la variance expliquée dépasse significativement la variance résiduelle. Si c'est le cas, on peut estimer que la régression révèle une relation authentique existant dans la population. Sous H_0 , SCE est distribué selon une loi $\chi^2(p)$ et SCR selon une loi $\chi^2(n - (p + 1))$. De fait, pour F , on a :

$$\frac{\frac{\chi^2(1)}{p}}{\frac{\chi^2(n-2)}{n-(p+1)}} = f_{p, n-(p+1)}^{1-\alpha}$$

où F suit une loi de Fisher avec p et $n - (p + 1)$ degrés de liberté. On rejette l'hypothèse nulle H_0 si :

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

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

2.5 Intervalles de confiance et intervalle de prévision

Corollaire 2.5.1. Sous les conditions (C) (voir 2.1), pour tout $j = 0 \dots p$,

$$\frac{\hat{\beta}_j - \beta_j}{s \sqrt{[(X'X)^{-1}]_{j+1,j+1}}} \sim \mathcal{T}(n - (p + 1)).$$

2.5.1 Intervalles de confiance pour les coefficients de régression

On peut, à partir du Corollaire (2.5.1), construire un intervalle de confiance bilatéral pour le coefficient β_j :

Proposition 2.5.1. Soit $\alpha \in]0, 1[$. On note $t_{n-p-1}(u)$ le u-quantile de la loi $\mathcal{T}(n - (p + 1))$.

Un intervalle de confiance de niveau de confiance $(1 - \alpha)$ pour β_j ($j = 0 \dots p$) est donné par :

$$I_j = \left[\hat{\beta}_j - t_{n-p-1}^{1-\frac{\alpha}{2}} \sqrt{s^2[(X'X)^{-1}]_{j+1,j+1}}, \hat{\beta}_j + t_{n-p-1}^{1-\frac{\alpha}{2}} \sqrt{s^2[(X'X)^{-1}]_{j+1,j+1}} \right].$$

2.5.2 Prevision

Comme dans le cadre du modèle de régression linéaire multiple, les buts de la régression est de proposer des prédictions pour la variable à expliquer Y lorsque on a de nouvelles valeurs $x_{n+1} = (1, x_{n+1,1}, \dots, x_{n+1,p})'$. Soit donc $x_{n+1} = (1, x_{n+1,1}, \dots, x_{n+1,p})'$ une nouvelle valeur pour laquelle, on veut prédire y_{n+1} qui est définie par :

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

avec $\varepsilon_{n+1} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ indépendant des $(\varepsilon_i)_{1 \leq i \leq n}$.

À partir des n observations précédentes, on a pu calculer un estimateur $\hat{\beta}$ de β . Donc, il est naturel de prédire la valeur correspondante via le modèle ajusté $\hat{y}_{n+1}$ définie par :

$$\hat{y}_{n+1} = x'_{n+1}\hat{\beta}.$$

Pour quantifier l'erreur de prévision $(y_{n+1} - \hat{y}_{n+1})$, on utilise la décomposition :

$$y_{n+1} - \hat{y}_{n+1} = x'_{n+1}(\beta - \hat{\beta}) + \varepsilon_{n+1}$$

$(y_{n+1} - \hat{y}_{n+1})$ est une variable gaussienne, dont moyenne 0 et variance $\sigma_\varepsilon^2(1 + x'_{n+1}(X'X)^{-1}x_{n+1})$, ce qui donne :

$$y_{n+1} - \hat{y}_{n+1} \sim \mathcal{N}\left(0, \sigma_\varepsilon^2(1 + x'_{n+1}(X'X)^{-1}x_{n+1})\right)$$

Comme dans le cas de la régression linéaire simple, on obtient que :

$$\frac{y_{n+1} - \hat{y}_{n+1}}{s_\varepsilon \sqrt{x'_{n+1}(X'X)^{-1}x_{n+1}}} \sim \mathcal{T}_{(n-(p+1))}$$

ce qui permet de construire l'intervalle de confiance pour y_{n+1} :

$$\left[\hat{y}_{n+1} - t_{(n-p-1)}^{1-\alpha/2} s_\varepsilon \sqrt{x'_{n+1}(X'X)^{-1}x_{n+1}}, \hat{y}_{n+1} + t_{(n-p-1)}^{1-\alpha/2} s_\varepsilon \sqrt{x'_{n+1}(X'X)^{-1}x_{n+1}} \right].$$

Application RLM sur le Volume des Marchandises au Port maritime

La régression linéaire constitue une méthode statistique couramment employée pour modéliser une relation linéaire entre une variable dépendante, dans ce cas les volumes de marchandises, et une ou plusieurs variables indépendantes, telles que les saisons, les indicateurs économiques ou d'autres facteurs pertinents. Afin d'atteindre notre objectif, nous mettrons en œuvre les étapes ci-dessous :

1. **Collecte des Données** : Rassembler des données historiques sur les volumes de marchandises traités par le port ainsi que des informations contextuelles pertinentes.
2. **Préparation des Données** : Nettoyer et transformer les données pour éliminer les erreurs, gérer les valeurs manquantes et normaliser les variables.
3. **Modélisation** : Construire un modèle de régression linéaire en identifiant les variables explicatives les plus pertinentes.
4. **Validation** : Évaluer la performance du modèle à l'aide de métriques appropriées (R^2 , erreur quadratique moyenne, etc.) et tester sa capacité à généraliser sur des données inconnues.
5. **Interprétation** : Analyser les résultats obtenus, interpréter les coefficients du modèle et formuler des recommandations pour son utilisation pratique.

Le volume des marchandises transitant par un port est un indicateur clé de son activité économique et de son efficacité logistique. Il est influencé par plusieurs facteurs, tels que les conditions météorologiques, la capacité des infrastructures portuaires, la demande du marché, les politiques douanières et les fluctuations économiques globales.

Dans cette section, nous examinerons comment la RLM peut être appliquée à l'étude du volume des marchandises au port de Mostaganem, en sélectionnant les variables explicatives pertinentes, en estimant les coefficients du modèle et en interprétant les résultats pour améliorer la gestion portuaire et optimiser la logistique.

Tableau des marchandises de port

- **Y : Volume de marchandises** : Quantité totale de marchandises traitées (en tonnes).
- **X1** : Nombre de navires arrivant au port chaque année.
- **X2** : Taille moyenne des navires (en tonnage).
- **X3** : Indicateur économique global (taux de change : moyen annuel de la devise en EURO).

Nous cherchons à expliquer une variable quantitative Y (volume des marchandises) en fonction de trois autres variables : X_1 , X_2 et X_3 . Pour ce faire, nous disposons de 20 observations recueillies sur la période de 2005 à 2024. les données de 2020 à 2021, issues des rapports annuels de l'administration portuaire de Mostaganem. En raison du manque de données précises disponibles au sein des services statistiques du port de Mostaganem, il a été nécessaire de simuler les données de manière aléatoire afin de compléter les informations manquantes. Cette approche vise à fournir une estimation approximative des activités portuaires, telles que le volume des marchandises, le nombre de navires, la taille moyenne des navires, ainsi que les indicateurs économiques associés. Bien que ces données simulées puissent offrir une approximation utile pour certaines analyses, elles ne doivent pas être considérées comme des informations officielles ou totalement fiables. Il est essentiel de continuer à améliorer la collecte de données et la gestion statistique au sein du port pour obtenir des informations plus précises et cohérentes

Années	Volume de marchandise	Nombre de navires	Taille moyenne des navires	Indicateur économique globale (EURO)
2005	800000	450	4500000	90.0
2006	828571	461	4642857	93.6
2007	857143	471	4785714	97.1
2008	885714	482	4928571	100.7
2009	914286	493	5071429	104.3
2010	942857	504	5214286	107.9
2011	971429	514	5357143	111.4
2012	1000000	525	5500000	115.0
2013	1028571	536	5642857	118.6
2014	1057143	546	5785714	122.1
2015	1085714	557	5928571	125.7
2016	1114286	568	6071429	129.3
2017	1142857	579	6214286	132.9
2018	1171429	589	6357143	136.4
2019	1200000	600	6500000	140.0
2020	857798	291	2260000	144.9
2021	900000	320	2500000	142.0
2022	1033333	380	3600000	146.3
2023	1166667	440	4700000	150.7
2024	1300000	500	5800000	155.0

On peut réécrire ce modèle sous la forme matricielle suivante :

$$Y = X\beta + \varepsilon \quad (3.1)$$

Notre objectif ici est de déterminer :

1. **Coefficients** : Les coefficients estimés pour chaque variable explicative (*Nombre de navires* , *Taille moyenne des navires*) et *Indicateur économique globale*. Ces coefficients mesurent l'impact de chaque variable sur le *volume de marchandises*.
2. **R² (R-carré)** : Cette valeur représente la proportion de la variance de la variable dépendante (*Volume de marchandises*) expliquée par les variables indépendantes. Plus elle est proche de 1, meilleur est l'ajustement du modèle.
3. **p-values** : Les p-values associées à chaque coefficient testent leur significativité statistique. Une p-value inférieure à 0.05 suggère que la variable a un effet significatif (comme le test de Fisher pour la significativité globale ou le test de Student pour les coefficients) .

Notre analyse doit également inclure des **tests d'hypothèses sur les erreurs** (ou résidus) du modèle, afin de vérifier les conditions d'application de la régression linéaire. Les principales vérifications sont :

1. **Linéarité du modèle**

- *Visualisation par résidus vs. valeurs prédites*

2. **Normalité des erreurs** :

- *Visualisation par Q-Q plot*

3. **Homoscédasticité** (variance constante des erreurs) :

- *Test de Breusch-Pagan* ou *test de White*
- Hypothèses :

$$\begin{cases} H_0 : \text{La variance des erreurs est constante} \\ H_1 : \text{La variance des erreurs n'est pas constante} \end{cases}$$

4. **Indépendance des erreurs** (absence d'autocorrélation) :

- *Visualisation Graphique par la fonction d'autocorrélation (ACF)*

Dans la suite de cette analyse, nous utilisons le logiciel R pour visualiser et vérifier nos objectifs statistiques.

Estimateur des coefficients du modèle (3.1)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 \quad (3.2)$$

```
# code R
donnees <- data.frame(
  Annee = 2005:2024,
  Y = c(800000, 828571, 857143, 885714, 914286, 942857, 971429, 1000000,
        1028571, 1057143, 1085714, 1114286, 1142857, 1171429, 1200000,
        857798, 900000, 1033333, 1166667, 1300000),
  X1 = c(450, 461, 471, 482, 493, 504, 514, 525, 536, 546, 557, 568, 579,
        589, 600, 291, 320, 380, 440, 500),
  X2 = c(4500000, 4642857, 4785714, 4928571, 5071429, 5214286, 5357143,
        5500000, 5642857, 5785714, 5928571, 6071429, 6214286, 6357143,
        6500000, 2260000, 2500000, 3600000, 4700000, 5800000),
  X3 = c(90.0, 93.6, 97.1, 100.7, 104.3, 107.9, 111.4, 115.0, 118.6,
        122.1, 125.7, 129.3, 132.9, 136.4, 140.0, 144.9, 142.0, 146.3,
        150.7, 155.0)
)
modele <- lm(Y ~ X1 + X2 + X3, data = donnemosta)
summary(modele)
```

Résultats de la régression multiple

Distribution des résidus

Min	1Q	Median	3Q	Max
-33955	-1364	-523	552	22385

Coefficients estimés

TABLE 3.1 – Coefficients de régression avec tests de significativité

Variable	Estimation	Erreur type	t-value	p-value
Constante	1.857×10^5	4.101×10^4	4.529	0.000343
X1	-1.607×10^3	2.224×10^2	-7.225	2.03×10^{-6}
X2	1.961×10^{-1}	1.590×10^{-2}	12.338	1.37×10^{-9}
X3	5.040×10^3	1.367×10^2	36.871	$< 2 \times 10^{-16}$

*** Significatif au niveau 0.001

Qualité globale du modèle

- Erreur standard des résidus : 11 000 (16 degrés de liberté)
- R^2 multiple : 0.9948
- R^2 ajusté : 0.9939
- Statistique F : 1 026 (3 et 16 DL)
- p-value : $< 2.2 \times 10^{-16}$

Équation du modèle estimé

L'équation estimée s'écrit :

$$\hat{Y} = 185700 - 1607X_1 + 0.1961X_2 + 5040X_3 \quad (3.3)$$

avec :

- Toutes les variables significatives à $p < 0.001$
- Une excellente qualité d'ajustement ($R^2 > 99\%$)

Interprétation des résultats de régression

- **Intercept** : L'intercept est la valeur prédite du volume de marchandise lorsque le nombre de navires et la taille moyenne des navires sont nuls. Dans cet exemple, il est positif, cete Valeur théorique (peu réaliste en pratique, car aucun port n'a zéro navire ou zéro tonnage).
- **Nombre_navires**

Toutes choses égales par ailleurs, une augmentation de 1 navire est associée à une diminution de 1607 tonne du volume de marchandises.

Problème identifié

Ce résultat est **contre-intuitif** car on s'attendrait à une relation positive :

$$\text{Navires (X1)} \uparrow \Rightarrow \text{Marchandises (Y)} \uparrow$$

Hypothèses explicatives

TABLE 3.2 – Causes possibles de la relation négative

Hypothèse	Mécanisme explicatif
Colinéarité	La variable X1 (nombre de navires) pourrait être corrélée avec X2 (tonnage total), où les gros navires réduisent le besoin de nombreux petits navires : $X1 \downarrow \text{ et } X2 \uparrow \Rightarrow Y \uparrow$
Saturation portuaire	Un excès de navires entraîne des inefficacités opérationnelles : <ul style="list-style-type: none"> • Congestion des terminaux • Allongement des temps d'attente

- **Taille_moyenne_navires** : Le coefficient est également positif, ce qui suggère qu'une augmentation de la taille moyenne des navires est associée à une augmentation du volume de marchandise. Les gros navires sont plus efficaces pour transporter des marchandises (économie d'échelle).
- **Indicateur économique globale** : Le taux de change a un impact positif sur le volume, une augmentation d'1 point de l'indice économique (X3) est associée à une augmentation de 5,040 tonnes du volume de marchandises.

- **R-squared** : Indique que 99% de la variabilité du volume de marchandise est expliquée par les variables indépendantes.
- **p-values** : Les p -values pour les deux variables indépendantes sont inférieures à 0.001, ce qui indique qu'elles sont statistiquement significatives.

Hypothèses relatives au modèle de la RLM

Résidus

Pour tout $i \in \{1, \dots, 30\}$, on appelle i -ème résidu la réalisation ε_i définie par :

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

Ces résidus permettent de vérifier les hypothèses de base du modèle de **régression linéaire multiple** (RLM).

Sous **R**, on les calcule avec la commande :

```
residuals(modele)
```

Index	Résidu	Index	Résidu
1	1077.1476	11	-1424.4931
2	1157.3900	12	-1343.4467
3	136.0095	13	-1263.2043
4	216.2519	14	-2284.5849
5	297.2982	15	-2204.3424
6	377.5407	16	-33954.9945
7	-643.8399	17	22384.7177
8	-563.5974	18	14697.4073
9	-483.3550	19	6507.0705
10	-1504.7355	20	-1180.2399

TABLE 3.3 – Résidus du modèle

Linéarité du modèle

Pour vérifier l'hypothèse de linéarité dans un modèle de régression linéaire, nous utilisons un graphique des résidus vs les valeurs prédites. Ce graphique permet de visualiser si les résidus (les différences entre les valeurs observées et les valeurs prédites) sont distribués de manière aléatoire autour de zéro, ce qui est une indication que la relation entre les variables est linéaire.

```
plot(modele, which = 1)
```

où

```
# Extraire les résidus
residus <- residuals(modele)
```

```
# Extraire les valeurs predites
valeurs_predites <- fitted(modele)
\end{lstlisting}

# Creer le graphique
plot(valeurs_predites, residus,
     xlab = "Valeurs predites",
     ylab = "Residus",
     main = "Graphique des residus vs valeurs predites",
     pch = 16, col = "blue")

# Ajouter une ligne horizontale a y = 0 pour reference
abline(h = 0, col = "red", lty = 2)
```

Le résultat de cette commande est présenté dans la Figure

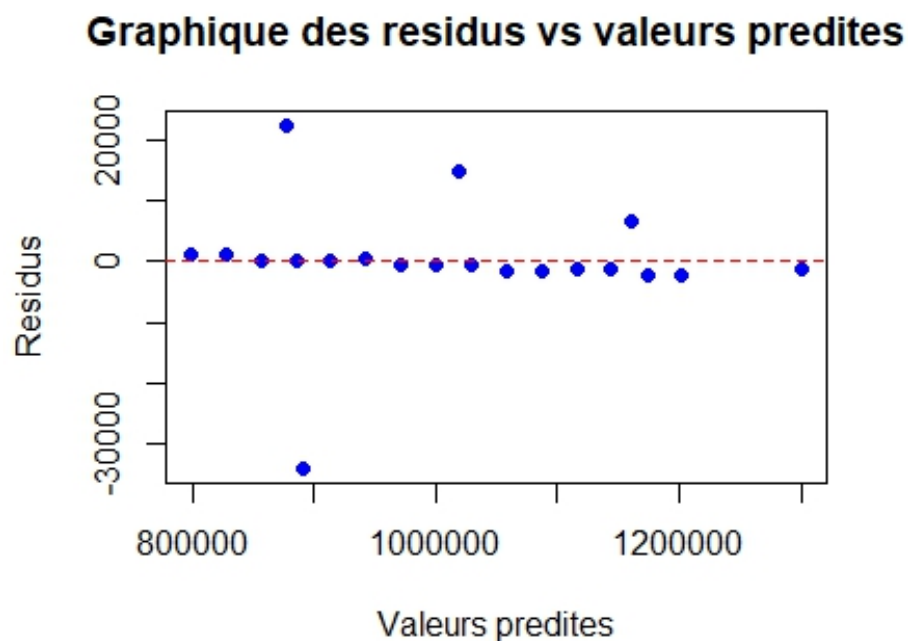


FIGURE 3.1 – Graphique des residus vs valeurs predites

Indépendance de $\varepsilon_1, \dots, \varepsilon_n$

Pour étudier l'indépendance de $\varepsilon_1, \dots, \varepsilon_n$, la première approche consiste à tracer le **corrélogramme**¹.

1. Graphique représentant les coefficients d'autocorrélation en fonction des décalages temporels.

Celui-ci représente les estimations ponctuelles de la fonction d'autocorrélation :

$$\hat{\rho}(h) = \frac{\sum_{t=1}^{n-h} (\varepsilon_t - \bar{\varepsilon})(\varepsilon_{t+h} - \bar{\varepsilon})}{\sum_{t=1}^n (\varepsilon_t - \bar{\varepsilon})^2}$$

où :

- h est le décalage (lag)
- $\bar{\varepsilon}$ est la moyenne des résidus
- n est le nombre d'observations

Sous R, la commande suivante génère le corrélogramme pour 30 décalages :

```
acf(residuals(modele), lag = 30)
```

Le résultat de cette commande est présenté dans la Figure

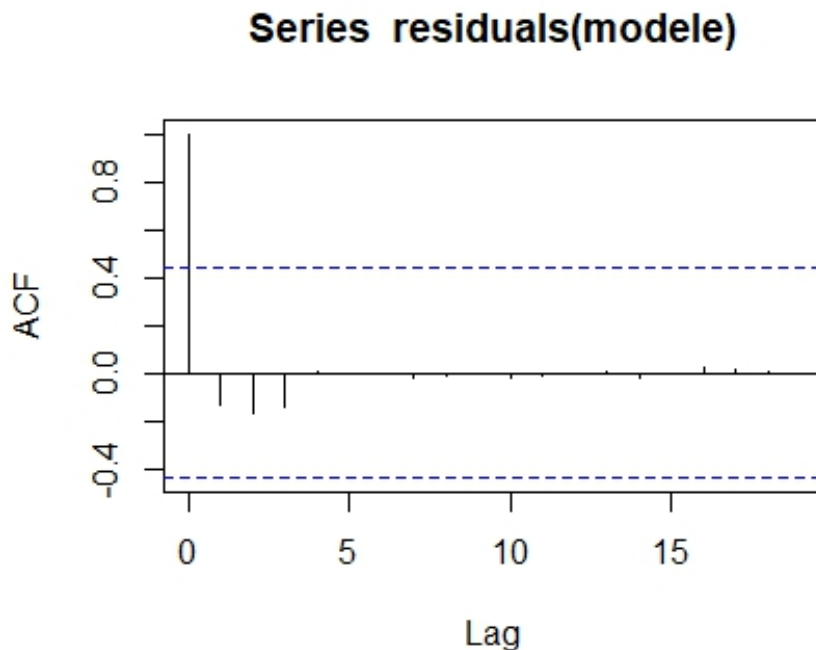


FIGURE 3.2 – Corrélogramme des résidus du modèle avec 30 décalages

D'après cette figure (3.2) on observe que $\varepsilon_1, \dots, \varepsilon_n$ sont indépendantes

Hypothèse de normalité

Pour vérifier l'hypothèse de normalité des résidus dans un modèle de régression linéaire, nous utilisons un QQ-plot (quantile-quantile plot). Ce graphique compare la distribution des résidus à une distribution normale. Si les points suivent approximativement une ligne droite, cela suggère que les résidus sont normalement distribués

```
plot(modele, which = 2)
```

où

```
Créer le QQ-plot
qqnorm(residus, main = "QQ-plot des résidus", pch = 16, col = "blue")

# Ajouter une ligne de référence
qqline(residus, col = "red", lty = 2)
```

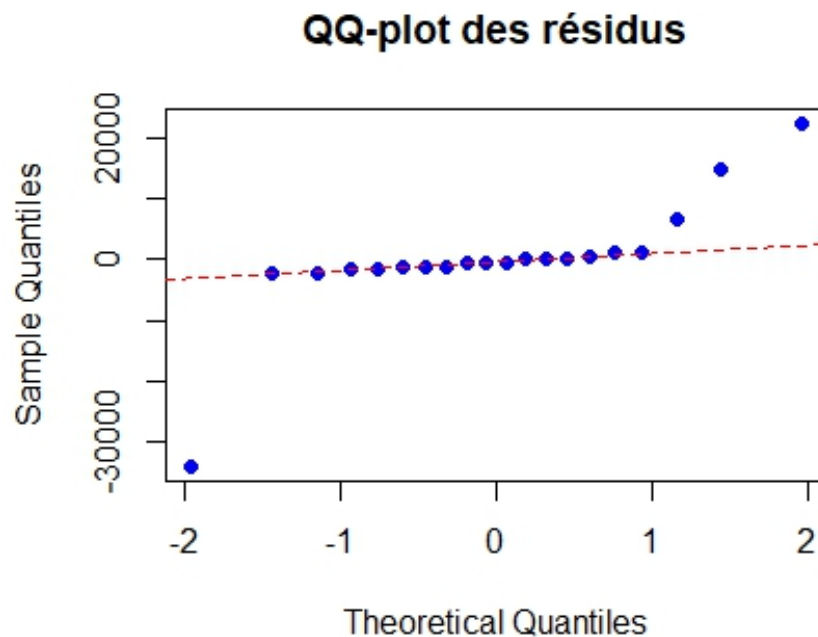


FIGURE 3.3 – Normalité des résidus

Les écarts-types $s(\beta_j)$ des estimateurs β_j ($j = 0, 1, 2, 3$) sont alors donnés par les racines carrées des éléments diagonaux de la matrice de variance-covariance :

$$s(\beta_j) = \sqrt{\text{diag}(\mathbb{V}(\beta))_j}$$

Sous R, on a :

```
> hatsigmabetas <- sqrt(diag(vcov(modele)))
(Intercept)          X1          X2          X3
4.100513e+04 2.223871e+02 1.589625e-02 1.367006e+02
```

Homoscédasticité (variance constante des erreurs)

```
install.packages("lmtest")
library(lmtest)
bptest(modele)
studentized Breusch-Pagan test
```

```
data: modele
BP = 15.484, df = 3, p-value = 0.001447
```

Interprétation : On remarque que $p \text{ value} = 0.001447 < 0.05$, nous décidons de garder l'hypothèse H_1 . La variance des erreurs n'est pas constante

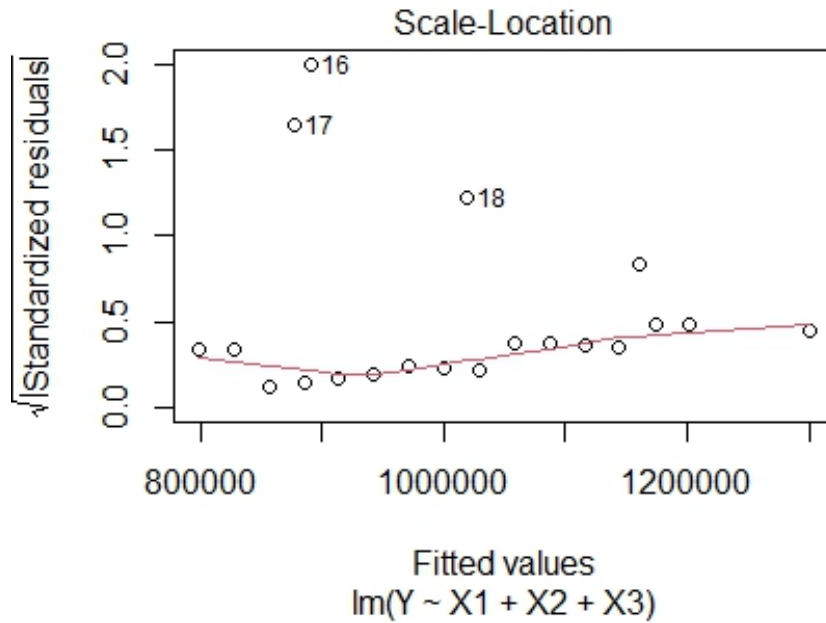


FIGURE 3.4 – La variance des erreurs n'est pas constante

3.0.1 Estimation par intervalle de confiance

L'intervalle de confiance pour estimer β_j ($j = 0, 1, 2, 3$), au niveau de confiance $100(1 - \alpha)\%$, est donné par :

$$\left[\beta_j - t_{1-\alpha/2}^{n-2} \cdot s(\beta_j), \beta_j + t_{1-\alpha/2}^{n-2} \cdot s(\beta_j) \right]$$

On calcule les intervalles de confiance au niveau 95% pour les 4 paramètres $\beta_0, \beta_1, \beta_2$ et β_3 .

Sous R, pour trouver les intervalles de confiance pour β_j ($j = 0, 1, 2, 3$), on écrit la commande suivante :

```
> confint(reg, conf.level = 0.95)
confint(modele, conf.level = 0.95)
              2.5 %      97.5 %
(Intercept) 98766.5002314 2.726205e+05
X1           -2078.0888886 -1.135210e+03
X2              0.1624343 2.298314e-01
X3            4750.4715362 5.330056e+03
```

$$\begin{aligned}
 IC_{0.95}(\beta_0) &= \left[b_0 \pm t_{0.975}^{n-2} \cdot s(b_0) \right] \\
 IC_{0.95}(\beta_1) &= \left[b_1 \pm t_{0.975}^{n-2} \cdot s(b_1) \right] \\
 &\vdots \\
 IC_{0.95}(\beta_3) &= \left[b_3 \pm t_{0.975}^{n-2} \cdot s(b_3) \right]
 \end{aligned}$$

Table d'ANOVA

```

custom_anova <- data.frame(
  "Source de variation" = c("Régression (expliquée)", "Résiduelle", "Totale"),
  "Somme des carrés" = c(
    sum((fitted(modele) - mean(donnees$Y))^2), # SCE
    sum(residuals(modele)^2),                 # SCR
    sum((donnees$Y - mean(donnees$Y))^2) # SCT
  ),
  "ddl" = c(
    length(coef(modele)) - 1, # ddl regression (p)
    df.residual(modele),      # ddl résiduelle (n-p-1)
    length(donnees$Y) - 1     # ddl totale (n-1)
  )
)

# Calculer les carrés moyens et F
custom_anova$"Carré moyen" <- with(custom_anova, Somme.des.carrés/ddl)
custom_anova$"F" <- c(
  custom_anova$"Carré moyen"[1]/custom_anova$"Carré moyen"[2],
  NA,
  NA
)

# Afficher le tableau
knitr::kable(custom_anova, digits = 4,
              caption = "Tableau ANOVA complet pour la régression multiple")
Table: Tableau ANOVA complet pour la régression multiple

|Source.de.variation| Somme.des.carrés| ddl| Carré moyen| F|
|:-----|:-----|:---|:-----|:---|
|Régression (expliquée)| 372423753149| 3| 124141251050| 1026.332|
|Résiduelle| 1935300137| 16| 120956259| NA|
|Totale| 374359053286| 19| 19703108068| NA|

# Créer de nouvelles données pour la prédiction
nouvelles_donnees <- data.frame(
  Nombre_navires = c(500, 600),
  Taille_moyenne_navires = c(2000000, 3000000),

```

CHAPITRE 3. APPLICATION RLM SUR LE VOLUME DES MARCHANDISES AU PORT MARITIME

```
Indicateur_economique_globale = c(1000, 1500)
)

# Faire des prédictions
predictions <- predict(modele, nouvelles_donnees)
print(predictions)
```

Conclusion

La régression linéaire fait partie des méthodes d'analyse multivariée appliquées aux données quantitatives. Il s'agit d'une approche exploratoire, fondée sur des observations ou des expérimentations, dont l'objectif principal est d'étudier la relation linéaire entre une variable quantitative Y et une ou plusieurs variables explicatives X , également quantitatives.

Dans ce mémoire, nous avons présenté les hypothèses fondamentales du modèle, ses composantes, ainsi que les principales méthodes d'estimation des paramètres. Parmi les nombreuses approches possibles, nous avons retenu deux méthodes d'estimation : la méthode des moindres carrés ordinaires et celle du maximum de vraisemblance. Nous avons également abordé les tests de signification des paramètres ainsi que l'estimation des intervalles de confiance.

Bibliographie

- [1] Faraway, J. J. (2016). *Linear Models with R* (2nd ed.). CRC Press.
- [2] Fisher, R. A. (1922). The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85(4), 597–612.
- [3] Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models* (3rd ed.). SAGE Publications.
- [4] Freedman, D. A. (2009). *Statistical Models : Theory and Practice*. Cambridge University Press.
- [5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning : with Applications in R* (2nd ed.). Springer.
- [6] Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2013). *Applied Regression Analysis and Other Multivariable Methods* (5th ed.). Cengage Learning.
- [7] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). McGraw-Hill Education.
- [8] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley.
- [9] Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2001). *Applied Regression Analysis : A Research Tool* (2nd ed.). Springer.
- [10] Stigler, S. M. (1986). *The History of Statistics : The Measurement of Uncertainty before 1900*. Harvard University Press.
- [11] Weisberg, S. (2014). *Applied Linear Regression* (4th ed.). Wiley.
- [12] Yule, G. U. (1897). On the theory of correlation. *Journal of the Royal Statistical Society*, 60(4), 812–854.