وزارة التعليم العالي والبحث العلمي

Université 20 Aout 1955 de Skikda

Faculté des Sciences

Département de Mathématiques

جامعة 20 أوت 1955 ، سكيكدة

كلية العلوم

قسم الرياضيات

N^0 : U.S/F.S/D.M/2022/2023.

Faculté des Sciences
Département de Mathématiques

Mémoire

Présenté en vue de l'obtention du diplôme de
Master en Mathématiques

Sur Quelques Modèles Prédictifs De Régression Linéaire Et Non Linéaire Et Applications

Option : Commande optimale et système dynamique

Par :

Larit Houria

Encadré par : Lallouche . Abdallah **M.C.B** U. SKIKDA

Devant le jury :

Président : Bouzettouta . Lamine M.C.A U. SKIKDA
Examinateur: Tilbi . Djahida M.C.B U. SKIKDA

Année : 2022/2023

Dédicace

Je dédie ce mémoire de fin d'études :

A mes adorables parents

A mon professeur et mon modèle dans la vie mon père **«Noureddine»** Pour la volonté que tu m'as donnée, pour m'avoir soutenu et guidé à chaque étape de ma vie et pour tous Les efforts qu'il a fait pour mon éducation et mon formation.

À ma Mère **«Souaad»** : Celle qui m'a donné la vie, le symbole de la force et la tendresse, qui s'est sacrifiée pour mon bonheur et ma réussite et qui n'a pas cesser de prier pour moi.

Je prie le bon Dieu de les bénir, de veiller sur eux, en espérant qu'ils seront toujours fiers de moi.

A ma sœur et mon frère

«Imane» et **«Houssem»**, pour tout ce que vous avez fait et ce que vous feriez pour moi.

A celui que j'aime beaucoup et qui m'a soutenue tout au long de ce projet : mon fiancé **«Ilyes »**

Je prie Dieu, le tout puissant de vous accorder santé, bonheur et succès…

A tous mes amies et mes collègues

Qu'ils trouvent ici le témoignage d'une fidélité et d'une amitié infinie.

A ma famille, mes proche et à ceux qui me donnent l'amour et de la vivacité.

Houria

REMERCIEMENTS

♥ Je remercie tout d'abord **ALLAH** qui m'aide et me donne la santé, la patience et le courage durant ces langues années d'étude et la force pour finir ce travail ♥.

Je tiens à remercier sincèrement mon encadreur

Dr.Lallouche Abdallah , pour m'avoir donné l'opportunité de travailler sur ce projet , pour son grand soutien scientifique et moral, pour les suggestions et les encouragements qu'il m'a apportés durant mon projet.

Mon sincère remerciement aux membres de jury

Dr.Bouzettouta Lamine et **Dr.Tilbi Djahida** qui ont accepté de juger mon travail.

Je remercie vivement tous les enseignants de notre département qui ont toujours donné le meilleur d'eux-mêmes afin de nous assurer une formation de qualité.

Résumé

Ce travail est consacré à l'étude des modèles statistiques linéaires et non linéaires, la régression spécifiquement linéaire et non linéaire, parce que l'étude de ces modèles nous permettent d'estimer et de prédire les valeurs futures basées sur des données complètes pour trouver des solutions appropriées en cas de problème. Pour cela, nous utiliserons deux modèles d'estimation, la régression linéaire (simple et multiple) et non linéaire, comment estimer cet modèles à partir des données complètes, le second concerne appréciées par régression non linéaire, avec des applications expérimentales pour chacun des deux modèles.

Mots clés: Régression , Moindres carrées , Estimation , Prévision .

ABSTRACT

This work is devoted to the study of linear and nonlinear statistical models, specifically linear and nonlinear regression, because the study of these models allow us to estimate and predict future values based on complete data to find appropriate solutions in the event of a problem. For this, we will use two estimation models, linear (single and multiple) and nonlinear regression, how to estimate this model from complete data, the second relates to assessments by nonlinear regression, with experimental applications for each of the two models .

Key words : Regression , Least squares , Estimation , Forecasting .

الملخص

هذا العمل مخصص لدراسة النماذج الإحصائية الخطية وغير الخطية، وتحديداً الانحدار الخطي وغير الخطي، لأن دراسة هذه النماذج تسمح لنا بتقدير وتوقع القيم المستقبلية بناءً على بيانات كاملة لإيجاد الحلول المناسبة في حالة حدوث مشكلة. لهذا، سوف نستخدم نموذجي تقدير، الانحدار الخطي (الفردي والمتعدد) وغير الخطي، وكيفية تقدير هذا النموذج من البيانات الكاملة، والثاني يتعلق بالتقييمات عن طريق الانحدار غير الخطي، مع تطبيقات تجريبية لكل من النموذجين.

الكلمات المتاحية: **الانحدار ، المربعات الصغرى ، تقدير ، التنبؤ .**

TABLE DES MATIÈRES

INTRODUCTION

La régression est une technique statistique employée pour décrire la liaison entre des variables. Les outils statistiques utilisés en analyse de régression permettent d'établir un modèle mathématique de la relation entre les variables et de quantifier l'incertitude associée à la relation qui en résulte. L'objectif le plus fréquent et le plus général est de prédire la valeur d'une certaine variable dite dépendante, connaissant la valeur d'une variable qui lui est associée (ou semble lui être associée), et que l'on nomme généralement variable indépendante ou explicative.

Le concept de régression fut introduit par Galton (1885) dont les recherches portaient sur l'hérédité. Cet auteur fut un des premiers qui décrivit les liens de dépendance entre certaines variables physiologiques. Il Si intéressa à la relation entre la taille des parents et celle de leurs enfants.

Ses résultats démontraient que les enfants dont les parents étaient très grands, s'avéraient en moyenne plus petits que leurs parents et que les enfants dont les parents étaient très petits étaient en moyenne plus grands que leurs parents. Il en déduisit que d'une génération à l'autre la caractéristique de taille au lieu de s'accentuer (de plus en plus grande et de plus en plus petite) "régressait" c'est-à-dire se rapprochait de la moyenne.

On donne aujourd'hui une toute autre interprétation aux résultats de Galton. En fait Galton généralisa à toute la population ce qu'il avait démontré pour un individu (en l'occurence les cas extrêmes, soit les très grands et les très petits). En réalité les parents de grande taille ont des enfants de grande taille, également aussi des enfants plus petits, de sorte que globalement la taille moyenne des enfants d'une famille pourra être plus faible que celle des parents. En somme les grands ne produisent pas tous des grands ou des plus grands. Il en résulte que les tailles extrêmes (dans une famille) demeurent exceptionnelles, tout comme dans la population dont l'échantillon est issu. Ce fait implique que la distribution des tailles et La dispersion autour de La moyenne demeure La

même dans La population et qu'en réalité il n'y a pas "régression" au sens ou l'entendait Galton. Quoique'il en soit le terme régression a survécu mais s'applique plus généralement aux méthodes statistiques qui décrivent et étudient la relation entre des variables.

Par la suite, Legendre (1805) introduisit la méthode d'estimation des coefficients d'un modèle de causalité par les moindres carrés. En parallèle, Gauss (1809) publia un travail sur le mouvement des corps célestes qui incluait un développement de la méthode des moindres carrés. Ainsi, la régression linéaire trouve ses racines dans ces avancées scientifiques Ce travail est composé de quartes chapitres :

Chapitre 1 :consiste à présentée quelques à rappelle qui seront utilisées dans les prochaines chapitres.

Chapitre 2 : une étude la régression linéaire simple et multiple ou on va estimer les paramètres de modèle par la méthode des moindres carrée , les hypothèses de modèle, les testes sur les paramètres du modèle , la prévision.est consacré à étudier la régression linéaire multiple ou on va donnée la forme analytique et la forme matricielle .

Chapitre 3 :est consacré à étude la régression non linéaire et la convertir en régression linéaire par la régression exponentielle et la régression puissance.

Chapitre 4 : est consacré à des applications à savoir la régression linéaire (simple et multipl) et la régression non linéaire .

CHAPITRE 1

RAPPELS PRÉLIMINAIRES

Dans ce chapitre on aborde quelques rappels préliminaire qui seront utilisés dans la suit,à savoir, séries statistique à une seule variable, séries statistique double.

1.1 Rappels statistiques

1.1.1 Séries statistique à une seule variable

Ce type d'analyse porte sur l'étude d'une seule variable. On a besoin de quelques notions de base qui seront à la suite.

Définition 1.1.1. *On appelle (série statistique) la suite des valeurs prises par une variable X sur les unités d'observation. Le nombre d'unités d'observation est noté n. Les valeurs de la variable X sont notées :*

$$x_1, ..., x_i, ..., x_n.$$

Définition 1.1.2. *Individu (**unité statistique**) tout élément sur lequel on peut faire une étude statistique.*

Définition 1.1.3. *Une population est définie comme un ensemble d'unités statistiques de même nature sur lequel porte notre étude statistique*

Définition 1.1.4. *Un échantillon est un sous-ensemble de la population*

Définition 1.1.5. *On appelle variable aléatoire statistique tout application X définie sur Ω , avec Ω un ensemble non vide appelé population , tout élément ω de Ω s'appelle un individu .*

Définition 1.1.6. *Une variable aléatoire continue prend des valeurs dans R ou dans un intervalle de R.*

Remarque 1.1.1. *X est aussi appelée caractère statistique. Le caractère désigne une grandeur, observable sur individu et susceptible de varier et prenant ainsi différent états appelés modalités.*

Définition 1.1.7. *On appelle modalité toute valeur $x_i \in X(\Omega)$ telle que*

$$X(\Omega) = \{x_1, x_2, ..., x_P\},$$

$X(\Omega)$: ensemble des valeurs prises par X ,où P est le nombre de modalités différentes de X

Définition 1.1.8. ***(L'effectif)*** *ou fréquence absolue associée à une valeur d'un caractère est le nombre de fois où cette valeur du caractère a été observée. l'effectif et noté par n_i.*

1.2 Séries statistiques doubles

Il arrive fréquemment d'étudier deux caractères quantitatifs différents X et Y d'un échantillon d'une même population. Pour déterminer s'il existe une relation entre eux (par exemple : la taille et le poids des étudiantes). Pour chaque individu de l'échantillon de la population, on mesure les valeurs de deux caractères X et Y, on obtient alors une liste de couples de nombres (x_i, y_i), k,l $\in \mathbb{N}$ que l'on peut présenter sous forme d'un tableau à deux entrées

X \ Y	y_1	y_2	...	y_j	.	y_l	effictif marginal de X
x_1	n_{11}	n_{12}	...	n_{1j}	.	n_{1l}	n_1
x_2	n_{21}	n_{22}	...	.	.	.	.
x_i	n_{i1}	n_{i2}	...	n_{ij}	.	n_{il}	$n_.$
.	.	.	...	.	.	.	.
x_i	n_{k1}	n_{k2}	...	n_{kj}	.	n_{kl}	$n_{k.}$
effictif marginal de Y	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	.	$n_{.l}$	n

- $x_1, x_2, ..., x_k$ sont les valeurs du caractère X.
- y_1, y_2 , ... , sont les valeurs du caractère Y.
- n_{ij} est l'effectif du couple (x_i, y_i) pour tout $1 \leq i \leq$ k et $1 \leq j \leq$ l.
- n_i est l'effectif marginal de $x_i, n_i = \sum_{j=1}^{1} n_{ij}$.

- n_j est l'effectif marginal de $y_j, n_j = \sum_{j=1}^{1} n_{ij}$.
- n est l'effectuf total , n=$\sum_{j=1}^{1}\sum_{j=1}^{1} n_{ij} = \sum_{j=1}^{k} n_i = \sum_{j=1}^{l} n_j$.

Définition 1.2.1. *(**Nuage de points**) Dans un repère orthogonal du plan, l'ensemble des points M_{ij} de cordonnées (x_i , y_i) constitue le nuage de points associé la série statistique double donnée ci-dessus*

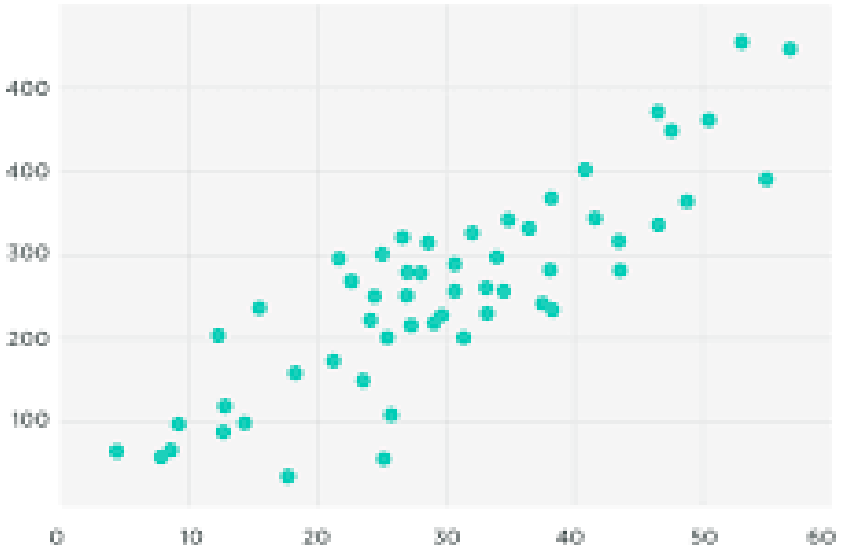

FIGURE 1.1: Nuage de points

Définition 1.2.2. *(**le point moyen**) On appelle le point moyen du nuage de points associé à la série statistique double le point G de coordonnées $(\bar{X}, \bar{Y})$ où x et y sont les moyennes arithmétiques des séries statistiques et respectivement définies comme suit :*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} x_i,$$

et

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{l} y_i,$$

1.2.1 Covariance d'une série statistique double

Définition 1.2.3. *On appelle covariance d'une série statistique double (X , Y) où les caractère X et Y sont quantitatifs le nombre noté Cov(X, Y) ou σ_{xy} défini par :*

$$\sigma_{xy} = Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Théorème 1.2.1. *La covariance peut également s'écrire :*

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{X}\bar{Y}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} x_i y_i\right) - \bar{X}\bar{Y},$$

ou X et Y sont les moyennes arithmétiques des deux séries simples, X et Y respectivement définies précédemment,

Démonstration

$$\sigma_{xy} = Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

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

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

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

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{X}\bar{Y}$$

Définition 1.2.4. *La variance est la somme des carrés des écarts 'a la moyenne divisée par le nombre d'observations :*

$$\sigma_x^2 = VarX = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{X})^2,$$

$$\sigma_y^2 = VarY = \frac{1}{n}\sum_{i=1}^{k}(y_i - \bar{Y})^2,$$

Théorème 1.2.2. *La variance peut aussi s'écrire*

$$VarX = \frac{1}{n}\sum_{i=1}^{k} x_i^2 - \bar{X}^2,$$

$$VarY = \frac{1}{n}\sum_{i=1}^{l} y_i^2 - \bar{Y}^2$$

Définition 1.2.5. *L'écart type de X est :*

$$\sigma_X = \sqrt{VarX}$$

Remarque 1.2.1. *La covariance mesure les dispersions des deux variables X et Y autour de leurs moyennes.*

1. *si Cov(X, Y) est nulle, on dit que les deux variables X et Y se produisent indépendamment.*
2. *si Cov(X, Y) < 0, les deux variables X et Y sont liées négativement $\Longrightarrow$ l'une augmente , l'autre diminue.*
3. *si Cov(X, Y)>0, les deux variables X et Y sont liées positivement $\Longrightarrow$ l'une augmente , l'autre augmente aussi .*

Proposition 1.2.1. *(**propriétés de la covariance**) soient X et Y deux variables statistiques et a,b,a' et b' des constantes réelles, alors*

- *Cov(X, Y) = Var(X).*
- *Cov(X, Y) = Cov(Y ,X).*
- *Cov(aX + b, a'Y + b') = aa'Cov(X, Y).*

Définition 1.2.6. *le Coefficient de corrélation linéaire d'une série statistique à deux variables X et Y est le nombre ρ défini par :*

$$\rho = \frac{Cov(X,Y)}{\sqrt{VarX}\sqrt{VarX}} = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Remarque 1.2.2. *L'importance des paramètres Cov(X,Y) et $\rho(X, Y)$ apparaîtra quand on s'intéressera au lien (ou corrélation) éventuel entre X et Y .*

Propriété 1.2.1. $|\rho(X, Y)| \leq 1$.

1.2.2 Rappels sur les variables aléatoires

Définition 1.2.7. *On appelle variable aléatoire X(en abrégé V.A.X) toute application de* Ω *vers* $\mathbb{R}$.

Ω *: l'ensemble des valeurs possibles de la série statistique.*

Une variable aléatoire continue prend des valeurs dans $\mathbb{R}$ ou dans un intervalle de $\mathbb{R}$.
La probabilité qu'une variable aléatoire continue soit inférieure à une valeur particulière est donnée par sa fonction de répartition.

$$F_X(x) = P(X \leqslant x)$$

où P est une mesure de probabilité, x $\in \mathbb{R}$.
La fonction de répartition d'une variable aléatoire continue est toujours :

- dérivable,
- positive : F(x)$\geq$ 0, pour tout x,
- croissante,
- $\lim_{x \longrightarrow \infty} F(x) = 1$,
- $\lim_{x \longrightarrow -\infty} F(x) = 0$.

On a

$$P(a \leq X \leq b) = F(b) - F(a).$$

La fonction de densité d'une variable aléatoire continue est la dérivée de la fonction de répartition en un point

$$f(x) = \frac{dF(x)}{dx}$$

Une fonction de densité est toujours :

- positive : $f(x) \geq 0$, pour tout x $\in \mathbb{R}$,
- d'aire égale à un : $\int_{-\infty}^{+\infty} f(x)dx = 1$.

On a évidemment la relation :

$$F(b) = \int_{-\infty}^{b} f(x)dx, \tag{1.1}$$

La probabilité que la variable aléatoire soit inférieure à une valeur quelconque vaut :

$$P(X \leq a) = \int_{-\infty}^{a} f(x)dx = F(a).$$

Dans la Figure 1.2, la probabilité Pr[X $\leq a$] est l'aire sous la densité de $-\infty$ à a.

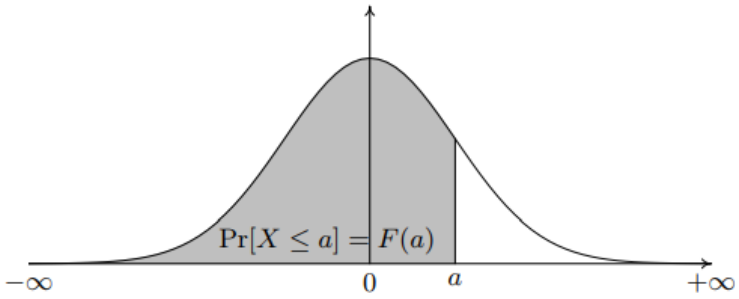

FIGURE 1.2: Probabilité que la variable aléatoire soit inférieure à a

La probabilité que la variable aléatoire prenne une valeur comprise entre a et b vaut

$$P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a).$$

Si la variable aléatoire est continue, la probabilité qu'elle prenne exactement une valeur quelconque est nulle :

$$Pr(X = a) = 0$$

1.2.3 Espérance mathématique et variance

Définition 1.2.8. *L'espérance mathématique d'une V.A.X à densité définie sur* Ω *est donnée par :*

$$\mu = \mathbb{E}(X) = \int_{X(\Omega)} xf(x)d(x). \tag{1.2}$$

X(Ω*) : l'ensemble des valeurs prises par X.*

Définition 1.2.9. *La variance d'une V.A.X existe et finie si* $\mathbb{E}(X^2)$ *existe, alors on a*

$$\sigma^2 = Var(X) = \mathbb{E}(X - \mathbb{E}(X^2)) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2. \tag{1.3}$$

où

$$\mathbb{E}(X^2) = \int_{(\Omega)} x^2 f(x)d(x). \tag{1.4}$$

$\mathbb{E}(X^2)$ *est le moment d'ordre é de la V.A.X.*

Définition 1.2.10. *l'écart type de la variable aléatoire X est :*

$$\sigma_X = \sqrt{VarX}.$$

Remarque 1.2.3. $VarX \geqslant 0, \sigma_X \geqslant 0$.

1.2.4 Loi normale (Gauss)

Définition 1.2.11. *Une variable aléatoire X est dite normale si sa densité vaut*

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}}, x \in \mathbb{R}. \tag{1.5}$$

où $\mu \in \mathbb{R}$ et $\sigma \in \mathbb{R}^+$ sont les paramètres de la distribution. Le paramètre μ est appelé la moyenne et le paramètre σ l'écart-type de la distribution.

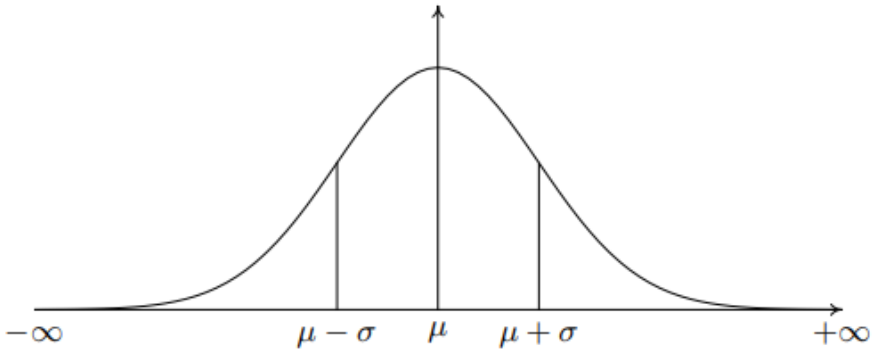

FIGURE 1.3: Fonction de densité d'une variable normale

De manière synthétique, pour noter que X suit une loi normale (ou gaussienne, d'après Carl Friedrich Gauss) de moyenne μ et de variance σ^2 on écrit :

$$X \sim N(\mu, \sigma^2).$$

La loi normale est une des principales distributions de probabilité. Elle a de nombreuses applications en statistique. Sa fonction de densité dessine une courbe dite courbe de Gauss. On peut montrer (sans démonstration) que :

$$E(X) = \mu,$$

et

$$Var(X) = \sigma^2$$

La fonction de répartition vaut

$$f_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx, x \in \mathbb{R}. \tag{1.6}$$

Cas Particulier

La variable aléatoire normale centrée réduite X $\sim$ N(0, 1) est une variable normale, d'espérance nulle, μ, et de variance $\sigma^2 = 1$. Sa fonction de densité vaut

$$f_{0,1}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in \mathbb{R}.$$

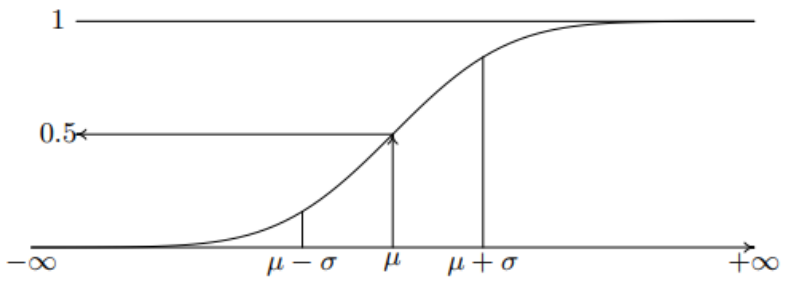

FIGURE 1.4: Fonction de répartition d'une variable normale

CHAPITRE 2

RÉGRESSION LINÉAIRE SIMPLE ET MULTIPLE

Ce chapitre traite la modélisation mathématique par le modèle le plus célèbre dit : régression linéaire (simple et multiple). on explicite les hypothèes fondamentales et les termes du modèle, les notions d'estimation des paramètres du modèle par la méthode de moindres carrés ordinaires, l'estimation par intervalles . On aborde quelques propriétés des estimateurs , la signification des paramètres et la signification globale du modèle .
commençons par la régression linéaire

2.1 Introduction

La régression linéaire est une méthode permettant de réaliser des prédictions ou des estimations. À l'aide d'un algorithme d'apprentissage supervisé, une relation linéaire est déterminée entre une variable dépendante et une ou plusieurs variables explicatives.

2.2 Domaines d'applications de régression linéaire simple et multiple

la régression linéaire est très utilisée dans plusieurs domaines soit pour étudier la relation linéaire ou les prévisions. Parmi ces domaines on cite,

- description.
- Sociologie.
- Marketing.

- Finance.
- Géographie.
- Ingénierie ...etc.

commençons par la régression linéaire simple.

2.3 Formulation analytique de la régression linéaire simple

La régression linéaire simple cherche à expliquer ou prédire les valeurs prises par une variable quantitative Y dite (endogène,dépendante ou encore réponce) par une variable quantitative X dite variable explicative (exoène, indépendante ou encore contrôle). On formule analytiquement le modèle comme suit :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \forall i \in \{1, .., n\}. \tag{2.1}$$

Où

- Y_i : Les valeurs de la variable endogène (dépendante,à expliquer).
- X_i : les valeurs de la variable exogène (indépendante, explicative).
- n : nombre d'observations.
- β_0 et β_1 : sont des paramètres réels inconnus (les coefficients du modèle).
- ε_i : est l'erreur aléatoire du modèle.

Remarque 2.3.1.

Le coefficient β_0 est appelé l'ordonnée à l'origine et le coefficient β_1 est appelé la pente .

le terme altéatoire ε permet de résumer toute l'information qui n'est pas prise en compte dans la relation linéaire que l'on cherche à établir entre Y et X .

2.4 Hypothèses fondamentales du modèle

Le modèle de la régression linéaire simple s'appuie sur les hypothèses suivante :

1. Hypothèses sur X et Y : Ce sont des grandeurs numérique mesurées . X est une variable (exogène) dans le modèle, Y est aléatoire é par l'intermédiaire de ε (c-à-d. la seule erreur que l'on a sur Y provient des insuffisances de X à expliquer ses valeurs dans le modèle).
2. E($\varepsilon = 0$) : en moyenne les erreurs s'annulent, le modèle est bien spécifié.
3. Var(ε_i)=σ_ε^2 : la variance de l'erreur est constante : homoscédastisité .
4. COV(x_i,ε_i) =0 l'erreur est indépendante de la variable oxogène.
5. COV(ε_i,ε_j)=0 , indépendance des erreurs : les erreurs relatives à 2 observations sont indépendantes .
6. $\varepsilon_i \equiv$ N(0 , σ_ε) suit une loi normale (Gaussienne) centrée.

2.5 Estimation des paramètres de modèle par la méthode des moindres carrée

on appelle estimateurs des moindres carrées Ordinaires (notée par MCO) $\hat{\beta}_0$ et $\hat{\beta}_1$ les valeurs qui minimisent la somme des carrés des écarts y_i à $\hat{y}_i$ Autrement dit, la droite des moindres carrées minimise la somme des carrés des carrées des distances verticales des points (x_i, y_i) du nuage à la droite ajustée $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1$x.
où $\hat{y}$ est l'estimation de y , $\hat{\beta}_0$ est l'estimateur de β_0 , $\hat{\beta}_1$ est l'estimateur de β_1 et le résidu ε_i Alors

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

et

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

c'est-à-dire minimiser la fonction

$$F(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Proposition 2.5.1. *les estimateurs des MCO $\hat{\beta}_0, \hat{\beta}_1$ ont pour expressions*

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

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

La valeur de la fonction S(β_0 , β_1) est minimum lorsque les dérivées de S par rapport à β_0 et β_1 s'annulent

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = 0.$$

et

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = 0.$$

Les dérivées par rapport à β_0 et β_1 sont

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = 0. \Longleftrightarrow -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Longleftrightarrow [-2\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i - \eta\hat{\beta}_0], \tag{2.2}$$

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = 0 \Longleftrightarrow -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Longleftrightarrow [-2\sum_{i=1}^{n} x_i y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i], \tag{2.3}$$

On pose

X = $\frac{1}{n}\sum_{i=1}^{n} x_i$: La moyenne empirique des x_i

Y = $\frac{1}{n}\sum_{i=1}^{n} y_i$: La moyenne empirique des y_i

$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i - (x_n)^2$: La variance empirique des x_i.

$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_n)^2 = \frac{1}{n}\sum_{i=1}^{n} y_i - (y_n)^2$: La variance empirique des y_i.

$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_n)(y_i - y_n)$: La covariance empirique entre les x_i et les y_i.

En multipliant l'équation (2.1) par $\frac{1}{n}$ on obtient

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

alors

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

en substituants l'équation (2.3)dans (2.2), on a

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

en multipliant (2.4) par $\frac{1}{n}$, on trouve

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_1(\sum_{i=1}^{n} x_i^2 - nx^2) - xy = 0,$$

donc

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - xy}{\sum_{i=1}^{n} x_i^2 - nx^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sum_{i=1}^{n}(x_i - x_n)^2}.$$

Remarque 2.5.1. - *On peut réécrire* $\hat{\beta}_1$ *sous la forme*

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{Cov(X,Y)}{Var(X)}.$$

Définition 2.5.1. *Le résidu de la i observation est* $\hat{\varepsilon}_i$ *telle que*

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

2.5.1 Propriétés des résidus

1. La droite de régression passe forcément par le centre de gravité du nuage de points (X , Y).

Pour le vérifier simplement , on réalise la projection pour le point X

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$= (y - \hat{\beta}_1 x) + \hat{\beta}_1 x_i$$

$$= y.$$

D'où le point (X,Y)appartient à la droite de régression.

2. La moyenne des valeures ajustées est égale à la moyenne des valeures observées y .

Preuve

$$\frac{1}{n}\sum_{i=1}^{n} \hat{y}(i) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \hat{\beta}_0 + \hat{\beta}_1 x.$$

Or , $y = \hat{\beta}_0 + \hat{\beta}_1 x$; car le point (X,Y) appartient à la droite de régression.

3. Les résidus représentent la partie inexpliquée des y_i par la droit des régression.

4. La somme des résidus est nulle

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

Preuve

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$$

$$= ny - n\hat{\beta}_0 - n\hat{\beta}_1 x$$

$$= ny - n(y - \hat{\beta}_1 x) - n\hat{\beta}_1 x = 0$$

5.

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

Preuve

$$\sum_{i=1}^{n} x_i \hat{\varepsilon}_i = \sum_{i=1}^{n} x_i (y_i - yi)$$

$$= \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$= \sum_{i=1}^{n} x_i (y_i - y - \hat{\beta}_1 (xi))(car \hat{\beta}_0 = y - \hat{\beta}_1 x)$$

$$= \sum_{i=1}^{n} x_i (yi - y) - \hat{\beta}_1 \sum_{i=1}^{n} x_i (xi - x).$$

Or

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

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

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

$$= nVarx$$

De plus

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

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

$$= n(\frac{1}{n}\sum_{i=1}^{n} x_i y_i - x_i y_i)$$

$$= nCov(x,y)$$

En considérant (2,3) et (2,4) , l'expression (2,2) devienne

$$\sum_{i=1}^{n} x_i\hat{\varepsilon}_i = \sum_{i=1}^{n} x_i(yi-y) - \hat{\beta}_1\sum_{i=1}^{n} x_i(xi-x) = nCov(x,y) - \hat{\beta}_1 nVarx = 0,$$

car $\hat{\beta}_1 = \dfrac{Cov(x,y)}{Varx}$.

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

Preuve

$$\sum_{i=1}^{n} \hat{y}_i\hat{\varepsilon}_i = \sum_{i=1}^{n}(\hat{\beta}_0 - \hat{\beta}_1 x_i)\hat{\varepsilon}_i = \hat{\beta}_0\sum_{i=1}^{n}\hat{\varepsilon}_i + \hat{\beta}_1\sum_{i=1}^{n}\hat{x}_i\hat{\varepsilon}_i = 0.$$

7. L'estimateur de la variance de l'erreur , noté par $s^2_{\hat{\varepsilon}}$, est donné comme suit

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

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

2.5.2 Propriétés des estimations de moindres carrés

Propriété 2.5.1. *$\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais de $\hat{\beta}_0$ et $\hat{\beta}_1$.*

Preuve

$\hat{\beta}_1$ est un estimateurs ans biais de $\hat{\beta}_1 \Longleftrightarrow E[\hat{\beta}_1] = \beta_1$,

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

telle que

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \\ y = \frac{1}{n}\sum_{i=1}^{n} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \end{cases} \tag{2.6}$$

$$\Longrightarrow y_i - y = \beta_1(x_i - x) + (\varepsilon_i - \varepsilon_i).$$

Alors

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

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

$$= E\Big(\frac{\sum_{i=1}^{n}(x_i - x)\varepsilon_i}{\sum_{i=1}^{n}(x_i - x)^2}\Big) \Longrightarrow E[\hat{\beta}_1] = \beta_1,$$

car l'erreur ε_i aléatoires $E[\varepsilon_i] = 0$, Pour $\hat{\beta}_0$

$$\hat{\beta}_0 = y - \hat{\beta}_1 x \Longleftrightarrow E[\hat{\beta}_0] = E[y - \hat{\beta}_1 x] = E[y] - E[\hat{\beta}_1]x = \beta_0 + \beta_1 x - \beta_1 x = \beta_0.$$

Propriété 2.5.2. *Les variances des estimateurs sont*

$$Var(\hat{\beta}_0) = \sigma^2_{\hat{\beta}_0} = \frac{\sigma^2_\varepsilon \sum_{i=1}^{n}(x_i^2)}{n\sum_{i=1}^{n}(x_i - x)^2}$$

$$= \frac{\sigma^2_\varepsilon \sum_{i=1}^{n}(x_i^2)}{nS_x}.$$

$$Var(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_\varepsilon}{\sum_{i=1}^{n}(x_i - x)^2}$$

$$= \frac{\sigma^2_\varepsilon}{S_x}.$$

Tendis que leurs covariance est

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2_\varepsilon \hat{x}}{\sum_{i=1}^{n}(x_i - x)^2}$$

$$= \frac{\sigma^2_\varepsilon \hat{x}}{S_x}.$$

Preuve

1)

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

$$= Var\Big(\frac{\sum_{i=1}^n (x_i - x)\varepsilon_i}{\sum_{i=1}^n (x_i - x)^2}\Big)$$

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

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

$$= \frac{\sigma_\varepsilon^2}{S_x}$$

2)

$$Var(\hat{\beta}_0) = Var(y - \hat{\beta}_1 x)$$

$$= Var\Big(\frac{1}{n}\sum_{i=1}^n (y_i - \hat{\beta}_1 x)\Big)$$

$$= Var\Big(\frac{1}{n}\sum_{i=1}^n (y_i)\Big) + Var(\hat{\beta}_1)x - 2xCov(y, \hat{\beta}_1).$$

Où

$$Cov(y, \hat{\beta}_1 = Cov\Big(\frac{1}{n}\sum_{i=1}^n y_i, \beta_1 + \frac{\sum_{i=1}^n (x_i - x)\varepsilon_i}{\sum_{i=1}^n (x_i - x)^2}\Big)$$

$$= \frac{1}{n} Cov\Big(\sum_{i=1}^n y_i, \beta_1\Big) + \frac{\sum_{i=1}^n (x_i - x)}{n\sum_{i=1}^n (x_i - x)^2}.$$

Théorème 2.5.1. *: **(Gausse Markou)** Les estimateurs sans biais du modèle de la régression linéaire simple sont linéaires et de variance minimales.*

2.6 Interv de cofficiance de $\hat{\beta}_0$ et $\hat{\beta}_1$

Proposition 2.6.1. *1. l'intervalle de confiance pour $\hat{\beta}_0$, noté par IC($\hat{\beta}_0$), est*

$$[\hat{\beta}_0 - t_{n-2}^1 - \frac{\alpha}{2} s_{\hat{\beta}_0}, \hat{\beta}_0 + t_{n-2}^1 - \frac{\alpha}{2} s_{\hat{\beta}_0}],$$

ou $t^1_{n-2} - \frac{\alpha}{2}$ *le quantile de niveau* $1 - \frac{\alpha}{2}$ *d'une loi de student* T_{n-2} *(avec (n-2) degrés de libertés).*

2. L'intervalle de confiance pour $\hat{\beta}_1$ *, noté par IC(*$\hat{\beta}_1$*) , est*

$$[\hat{\beta}_1 - t^1_{n-2} - \frac{\alpha}{2} s_{\hat{\beta}_1}, \hat{\beta}_1 + t^1_{n-2} - \frac{\alpha}{2} s_{\hat{\beta}_1}].$$

2.7 Analyse de la variance et signification de la régression

L'analyse de la variance va nous permettre :

- De quantifier la variation totale dans les observations et de la décompose en deux sources de variations , soit :

  i) une variation attribuable de la régression.

  ii) une variation résiduelle.
- de vérifier , à l'aide d'un tableau d'analyse de la variance (ANOVA) , si la source de la variation attribuable de la régression est significtive.
- De définir un indice qui donne une mesure descriptive de qualité de l'ajustement des points expérimentaux (x_i, y_i) par la droite de régression. On peut écrire que l'écart total $(y_i - y)$ est la somme de deux composantes

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

2.7.1 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 ajoutant puis retranchant by la valeur estimée de y par la droite de régression.

Cette procédure fait apparaître une somme de deux écarts

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

Ainsi l'écart total $(y_i - y)$ peut être vu comme la somme de deux écarts : Un écart entre y_i obsevé et $\hat{y}_i$ la valeur estimée par le modèle . Un écart entre $\hat{y}_i$ la valeur estimée par le modèle et la moyenne y .

On élève les deux membres au carré et on somme sur les observations i

$$\sum_{i=1}^{n}(y_i - y)^2 = \left[\sum_{i=1}^{n}(\hat{y}_i - y) + (y_i - \hat{y}_i)\right]^2$$

$$= \left[\sum_{i=1}^{n}(\hat{y}_i - y) + \hat{\varepsilon}_i\right]^2$$

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

Dans la régression on montre que

$$2\sum_{i=1}^{n}(\hat{y}_i - y)\hat{\varepsilon}_i = 0.$$

Preuve

On a

$$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})\hat{\varepsilon}_i = \sum_{i=1}^{n}(\hat{\beta}_1 x_i + \hat{\beta}_0 - \overline{y})\hat{\varepsilon}_i$$

$$= \hat{\beta}_1 \sum_{i=1}^{n} x_i\hat{\varepsilon}_i + \hat{\beta}_0 \sum_{i=1}^{n}\hat{\varepsilon}_i - \overline{y}\sum_{i=1}^{n}\hat{\varepsilon}_i$$

. On a $\sum_{i=1}^{n}\hat{\varepsilon}_i = 0$, donc on obtient que $\sum_{i=1}^{n}(\hat{y}_i - \overline{y})\hat{\varepsilon}_i = 0$.

On aboutit enfin à l'égalité fondamentale (l'équation d'analyse de la variance)

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

Alors

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

D'où

$$SCT = SCE + SCR.$$

- SCT : $\sum_{i=1}^{n}(y_i - \bar{y})^2$: est la somme des carrés totaux où elle indique la variabilité totale de Y c.à.d. l'information disponible dans les données .
- SCE : $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$: est la somme des carrés expliquées , cette quantité est la variabilité expliquée par le modèle c.à.d. la variation de Y expliquée par X .
- SCR : $\sum_{i=1}^{n}\hat{\varepsilon}_i{}^2$: est la somme des carrés résiduels . Elle écrite la variabilité non-expliquée (résiduelle) par le modèle (l'écart entre y et $\bar{y}$).

Remarque 2.7.1. *Deux situations extrêmes peuvent survenir*

- *Dans des cas , SCR =0 et donc SCT = SCE : les variations de Y sont complètement expliquées par celles de X . On a un modèle parfait , la droite de régression passe exactement par tous les poits du nuage ($\hat{y}_i = y_i$).*
- *Dans le pire des cas , SCE =0 : X n'apporte aucune information sur Y .*

On produire du tableau d'analyse de variance (voir le tableau 2.1) à partir de la décomposition de la variance , comme suit :

Source de variation	*ddl*	*somme des carrées*	*carrées moyens*
Expliquée	*1*	*SCE*	$CME = \frac{SCE}{1}$
Résiduelle	*n-2*	*SCR*	$CMR = \frac{SCR}{n-2}$
Totale	*n-1*	*SCT*	

Tab.2.1- Tableau d'analyse de variance de la régression linéaire simple .

Où ddl : degrés de liberté .

2.7.2 Coefficient de détermination R^2

La propossion de la variation totale dans les observations y_i (autour de la moyenne y) qui est expliquée par la droite de la régression est donnée par le coefficient de détermination , noté R^2 donné par :

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

Théorème 2.7.1.

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

Preuve

d'après l'équation on a

$$\frac{SCT}{SCT} = \frac{SCE}{SCT} + \frac{SCR}{SCT},$$

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

la quantité $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$, est appelée le coefficient de détermination alors

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i{}^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2}.$$

- c'est un indice de la qualité de la justement de la droite de régression aux points expérimentaux .

- **Relation entre les degrés de liberté** L'additivité des degrés de liberté nous donne la relation suivante :

$$\text{dlpourSCT} = \text{dlpourSCE} + \text{dlpourSCR}$$

$$(\text{n -1}) = (\text{n} + 2) + 1 \ .$$

- On présent généralement les quantités nécessaire au calcul de

$$F_c = \frac{CME}{CMR}.$$

- F_c : la valeur du Fischer calculée dans un tableau d'analyse de la variance abrégé par tableau d'ANOVA .

- $CME = \frac{SCE}{1}$: le carré moyenne de la régression .

- $CMR = \frac{SCR}{n-2}$: le carré moyenne résiduel .

Remarque 2.7.2. *- Ce coefficient R^2 qui varie entre 0 et 1 mesure la proportion de variation totale de Y autour de la moyenne expliquée par la régression , c.à.d. prise en compte par le modèle .*

- Plus R^2 se rapproche de valeur 1 , meilleure est l'adéquation du modèle aux données et un R^2 faible (proche de 0) signifie que le modèle a un faible pouvoir explicatif .

2.8 Test de signification du modèle

Ce test permet de connaître l'apport globale de la variable X à la détermination de Y. On veut tester l'hypothèse

$$\begin{cases} H_0 : \beta_1 = 0, \\ H_1 : \beta_1 \neq 0, \end{cases} \tag{2.7}$$

Pour tester cette hypothèse , on a basé sur la statistique de Fisher , notée par F :

$$F_{obs} = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} = \frac{CME}{CMR}. \tag{2.8}$$

cette statistique indique si la variance expliquée est significativement supérieure à la variance résiduelle . Dans ce cas , on peut considérer que l'explication emmenée par la régression traduit une relation qui existe réellement dans la population. Sous H_0 ,SCE est distribué selon un $X^2(1)$ et SCR

selon un $X^2(n-2)$, de fait pour F_{obs} , on a

$$F_{obs} = \frac{\frac{X^2(1)}{1}}{\frac{X^2(n-2)}{n-2}} = f_{1,n-2}^{1-\alpha}. \tag{2.9}$$

Alors , sous H_0, F_{obs} est donc distribué selon une loi de Fisher à (1, n-2) degrés de liberté , où on rejette H_0 si :

$$F_{obs} \geq f_{1,n-2}^{1-\alpha}.$$

Avec $f_{1,n-2}^{1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une loi de Fisher à (1 , n-2), et X^2 est loi du khi-deux.

Remarque 2.8.1. - *On peut réécrire la statistique F en fonction de R^2 cpmme suit :*

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

Dans la plupart des logiciels statistique , on fournit directement la probabilité critique. Elle correspond à la probabilité que la loi de Fisher dépasse la statistique calculée F . ainsi , la régle de décision (rejette H0) au risque devient :

$$p - value < \alpha.$$

passons maintenant à la régression linéaire multiple

2.9 Formulation analytique de la régression linéaire multiple

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

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \forall i \in \{1, ..., n\}, \tag{2.10}$$

où

- $\beta_0, \beta_1, \beta_2, ..., \beta_p$ sont appelés les paramètres ou les coefficients inconnus du modèle que l'on veut estimer à partir des données .
- i = 1,...,n correspond au numéro des observations.

- y_i est la i- ème observation de la variable Y .
- x_{ij} est la i- ème observation de la j-ème variable .
- ε_i est l'erreur du modèle (bruit) . Il représente la déviation entre ce que le modèle prédit et la réalité .
- p : nombre des variables explicatives .

2.10 La forme matricielle du régression linéaire multiple

Ce modèle s'écrit sous la forme des équations comme suit :

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \ldots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \ldots + \beta_p x_{2p} + \varepsilon_2 \\ \quad \vdots \qquad \vdots \qquad \ldots \quad \vdots \qquad \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \ldots + \beta_p x_{np} + \varepsilon_n \end{cases}$$

Qui traduit sous la forme matricielle comme

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

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

où

- Y : est la vecteur à expliquer de taille n x 1 .
- X : est la matrice , de taille n x (p+1) : qui contient l'ensemble des observations sur les exogènes , avec un première colonne formée par la valeur 1 indiquant que l'on intègre la constante β_0 dans l'équation où p étant le nombre de variables explicatives réelles .
- ε : est le vecteur des erreurs de taille n x 1 .

Remarque 2.10.1. *1. 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. 2. Les coefficients β_j , (j= 1,...,p) représente le changement subi par $E(y_i)$ correspondant à un changement unitaire dans la valeur de*

la j-iéme variable explicative , lorsque les autres variables explicatives demeurent inchangées .

2.11 Hypothèses fondamentales du modèle

Le modèle de régression linéaire s'appuie sur des hypothèses suivantes :

1. les valeurs x_{ij} sont observées sans erreurs .

2. $E(\varepsilon_i) = 0, \forall i \in \{1, ..., n\}$ espérance nulle . Les erreurs sont centrées (le modèle est bien spécifié en moyenne).

3. $Var(\varepsilon_i) = \sigma^2, \forall i \in \{1, ..., n\}$ (la variance de l'erreur est constante).

4. $Cov(\varepsilon_i, \varepsilon_j) = 0 sii \neq j$ (pas d'auto-corrélation des erreurs).

5. $Cov(x_{ij}, \varepsilon_j) = 0 sii \neq j$ (Les erreurs sont linéairement indépendantes des variables exogènes).

6. $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ les erreurs suivant une loi normale multidimensionnelle .

2.12 Cas général

Il s'agit d'estimer les p paramètres inconnus $\beta_0, \beta_1, \beta_2, ..., \beta_p$ par $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ .En utilisons la méthode des moindres carrées , on chercher à minimiser la quantité

$$\sum_{i=1}^{n} \hat{\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \hat{\beta}_p x_p)^2.$$

c'est-à-dire à minimiser la fonction objectif

$$z = f(\beta_0, \beta_1, \beta_2, ..., \beta_{p-1}) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \hat{\beta_{p-1}} x_{p-1})^2.$$

On obtenu en posant

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

La méthode des moindres carrés consiste donc à minimiser la quantité

$$\sum_{i=1}^{n\hat{\varepsilon}_i^2} \hat{\varepsilon}' \hat{\varepsilon} = (y - \hat{y})'(y - \hat{y}) = (y - X\hat{\beta})'(y - X\hat{\beta}).$$

en développant cette expression , on obtient

$$\hat{\varepsilon}' \hat{\varepsilon} = y'y - \hat{\beta}' X' y - y' X \hat{\beta} + \hat{\beta}' X' X \hat{\beta}$$

$$= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}.$$

Dans ce développement , on a utilisé le fait que $\hat{\beta}'X'y$ est un scalaire , ce qui implique que

$$\hat{\beta}'y = (\hat{\beta}'X'y)' = y'X\hat{\beta},$$

on peut montrer en outre que la dérivée partielle de la fonction objectif $\hat{\varepsilon}'\hat{\varepsilon}$ par rapport à $\hat{\beta}$ est donnée par

$$-2X'y + 2X'X\hat{\beta},$$

l'estimateur $\hat{\beta}$ que l'on cherche est donc la valeur qui annule cette expression . On a ainsi une seule équation normale sous la forme matricielle

$$(X'X)\hat{\beta} = X'y,$$

et on obtient l'estimateur des moindres carrés défini par

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

Remarque 2.12.1. ***cas particulier*** *: régression linéaire simple*

Notons que dans le cas d'un modèle de régression simple , c'est-à-dire le modèle (3.1) avec p = 2

On a

$$X'X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix},$$

et donc

$$(X'X)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (n\sum_{i=1}^n x_i^2)^2} \begin{pmatrix} \sum_{i=1}^n x_i & -\sum_{i=1}^n x_i^2 \\ -\sum_{i=1}^n x_i & n \end{pmatrix},$$

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

et finalement

$$\begin{pmatrix} \sum_{i=1}^n x_i \dfrac{\bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \dfrac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{pmatrix},$$

Ce qui correspond aux estimateurs des moindres carrés .

2.13 Propriétés des moindres carrés

Lorsque on utilise la méthode des moindres carrés , on a toujours la propriété suivante

$$\sum_{i=1}^{n} \hat{y}^2 = \sum_{i=1}^{n} \hat{y} y_i,$$

ou autrement dit sous forme matricielle :

$$\hat{y}'\hat{y} = \hat{y}'y.$$

Preuve

On a

$$\hat{y}'\hat{y} = (X\hat{\beta})'X\hat{\beta}$$

$$= \hat{\beta}X'X\hat{\beta}$$

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

$$= (X\hat{\beta})'y$$

$$= \hat{y}'y.$$

Il s'ensuite que

$$\hat{\varepsilon}'\hat{\varepsilon} = (y - \hat{y})'(y - \hat{y})$$

$$= (y' - \hat{y}')(y - \hat{y})$$

$$= y'y - y'\hat{y} - \hat{y}'y + \hat{y}'\hat{y}$$

$$= y'y - \hat{y}\hat{y}.$$

Autrement dit , la somme des carrés des résidus peut s'exprimer par

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

Par ailleurs , lorsque l'on considères un modèle avec une constante β_0 comme (3.1), on peut montrer

que l'une des équations normales égal à

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x_1} - \hat{\beta}_2\bar{x_2} - ... - \hat{\beta}_P\bar{x_P}.$$

Il s'ensuite que l'hyperplan des moindres carrés contient le point $(\bar{x_1}, \bar{x_2}, ..., \bar{x_P}, \bar{y})$. On a également

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i.$$

Preuve

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n}(y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_p x_{ip})$$

$$= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + ... + \hat{\beta}_P \sum_{i=1}^{n} x_{iP}$$

$$= n(\bar{y} - \hat{\beta}_1\bar{x_1} - \hat{\beta}_2\bar{x_2} - ... - \hat{\beta}_P\bar{x_P})$$

$$= \hat{\beta}_1 n\bar{x_1} + \hat{\beta}_2 n\bar{x_2} + ... + \hat{\beta}_P n\bar{x_P}$$

$$= \sum_{i=1}^{n} y_i.$$

Il s'ensuite que la somme (et donc la moyenne) des résidus est toujours nulle . En effet

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

2.14 Quelques propriétés des estimateurs

- Estimateur sans biais

Théorème 2.14.1. *(Propriétés des estimateurs MCO) L'estimateur $\hat{\beta}$ des moindres carrés ordinaire est sans biais , c.à.d $E(\hat{\beta}) = \beta$ et sa matrice de variance covariance , notée par varcov($\hat{\beta}$) ou par $s^2(\hat{\beta})$, est*

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

- Estimateur convergent

Théorème 2.14.2. *(Gauss-Markov) L'estimateur $\hat{a} = (X'X)^{-1}X'Y$ des moindres carrés est qualifié de BLUE (Best Linear Unbiaised Estimator), car il s'agit du meilleur estimateur linéaire sans biais (au sens qu'il fournit les variances les plus faibles pour les estimateurs).*

2.15 L'intervalle de confiance

Proposition 2.15.1. *(Intervalle et régions de confiance)*

- Pour tout $j \in (1, ..., p)$, un intervalle de confiance de niveau $(1-\alpha)$ pour β_j est

$$[\hat{\beta}_j - t_{n-p-1}^{1-\frac{\alpha}{2}} s_{\hat{\beta}_j}, \hat{\beta}_j + t_{n-p-1}^{1-\frac{\alpha}{2}} s_{\hat{\beta}_j}],$$

où $t_{n-p-1}^{1-\frac{\alpha}{2}}$ est le quantile de niveau $1-\frac{\alpha}{2}$ d'une loi de student τ_{n-p-1}.

2.16 Analyse de la variance et coefficient de détermination

2.16.1 Décomposition de la variance et tableau d'ANOVA

Cette décomposition effectuée de façon similaire à celle effectuée de la régression linéaire simple

.

On peut facilement vérifier que

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

On peut obtenir la décomposition

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

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

$$SCT = SCE + SCR.$$

Où

- SCT : $\sum_{i=1}^{n}(y_i - \bar{y})^2$: désigne la somme des carrés totaux (centrés).

- SCE : $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$: la somme des carrés expliqué (centrés).

- SCR : $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: la somme des carrés des résidus.

Le tableau " d'analyse de la variance" se présenté sous la forme suivante (voir le tableau 3.1)

Source de variation	ddl	somme des carrées	carrées moyens
Expliquée	p	SCE	CME = $\frac{SCE}{P}$
Résiduelle	n-p-1	SCR	CMR = $\frac{SCR}{n-p-1}$
Totale	n-1	SCT	

Tab.2.1- Tableau d'analyse de variance de la régression linéaire multiple .

2.16.2 Coefficient de détermination R^2

Le R^2 est la proportion de variance expliquée par la régression . Pour calculer le R^2, on utilise également les expressions

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

$$= 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon_i}^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

$$= 1 - \frac{SCR}{SCT}.$$

Remarque 2.16.1. *Ce coefficient R est compris entre 0 et 1 : plus il est proche de 1 est plus grand est la part expliquée, autrement dit meilleure est la régression inversement un coefficient R proche de 0 indique que la quantité SCR est élevée .Le R ne permet de comparer que des modèles ayant le même nombre de variables explicatives, le même nombre d'observation et la même forme (on ne peut pas comparer un modèle simple avec un modèle en log).*

2.17 Test de signification du modèle

L'objectif du test global de Fischer est d'étudier la liaison globale entre Y et les variables explicatives $X_j (j = 1, ..., p)$.

On considère les hypothèses

$$\begin{cases} H_0 : \beta_1 = ... = \beta_p = 0, \\ H_1 : \exists \beta_j \neq 0, (j = 1, ..., p). \end{cases} \tag{2.11}$$

Pour tester l'hypothèse , on a vu que l'on peut utiliser la statistique de Fischer F

$$F_{obs} = \frac{\frac{SCE}{p}}{\frac{SCR}{n-p-1}} = \frac{CMR}{CME},$$

où F_{obs} suit une loi de Fisher avec p et (n-p-1) degrés de liberté .

On rejette H_0 si

$$F_{obs} \geq f_{p,n-p-1}^{1-\alpha},$$

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

CHAPITRE 3

RÉGRESSION NON LINÉAIRE

3.1 Introduction

La régression non linéaire est une méthode permettant de déterminer un modèle non linéaire de relation entre deux variables. La régression non linéaire a pour but d'ajuster un modèle non linéaire pour un ensemble de valeurs afin de déterminer la courbe qui se rapproche le plus de celle des données de Y en fonction de X.

3.2 Formulation analytique de la Régression non linéaire

Le modèle de régression non linéaire s'écrit :

$$y_i = f(x_i, \beta) + \varepsilon_i, i = 1, ..., n \tag{3.1}$$

- β :Représente un vecteur à k composantes de paramètre généralement inconnu.
- y_i : Représente l'observation i de la variable dépendante.
- Les ε_i Sont indépendants entre eux.
- $f(x_i, \beta)$ est la fonction de régression non linéaire . elle dépend d'une variable réelle x et de paramètres β .

3.3 Modèles non linéaires, mais linéarisables

Il s'agit des modèles dont la spécification n'est pas linéaire (NL dans les variables), mais ses paramètres sont linéaires. Ci-dessous, nous présentons quelques types de ces modèles .

- La forme exponentielle.
- La forme polynomiales ou inverses.
- La forme Logarithmique .
- La forme puissance .

3.3.1 Régression exponentielle

La régression exponentielle désigne le processus consistant à trouver une équation pour la courbe exponentielle qui correspond le mieux à un ensemble de données. La régression exponentielle est très similaire à la régression linéaire, qui consiste à essayer de trouver une équation pour la ligne (droite) qui correspond le mieux à un ensemble de données . L'équation exponentielle (Forme non linéaire) est :

$$Y = B \exp^{ax} \tag{3.2}$$

Chaque fonction non linéaire nous pouvons entrer la réciproque(la fonction inverse)pour avoir la forme linéaire . Par exemple la réciproque de la fonction exponentielle c'est la fonction logarithme donc on improvisant logarithme de 2 côté, c'est-à-dire :

$$\ln Y = \ln B \exp^{ax} = \ln B + \ln \exp^{ax},$$

Finalement , on a trouvé

$$\ln Y = \ln B + ax \tag{3.3}$$

Linéarisation de le forme exponentielle

$1^{ére}$ **méthode :** on prend

$$(*) \begin{cases} y = \ln Y \\ b = lnB \end{cases}$$

On obtient

$$y = ax + b.$$

Donc , on écrit une relation linéaire entre deux autres variables en fonction des variables initiales , C'est-à-dire Création d'une relation exponentielle entre (X) et $(\ln Y)$.

Ensuite, vous choisissez une régression linéaire (par la méthode des moindres carrées)

Exemple 3.3.1. *Les deux variables sont X et $\ln Y = y$,et les données de ces deux variables présente dans le tableau suivantes :*

X	*1*	*2*	*13*	*19*	*25*	*31*
Y	*1240*	*1190*	*1160*	*1135*	*1094*	*1065*

Création d'une relation linaire entre (X) et $(\ln Y)$, Les deux variables sont X et $\ln Y = y$ on à calcule $\ln Y$,et les données de ces deux variables présente dans le tableau suivantes :

X	*1*	*2*	*13*	*19*	*25*	*31*
$Y = \ln Y$	*7.123*	*7.082*	*7.056*	*7.034*	*7.085*	*6.971*

et voila

$$y = aX + b.$$

on trouver

- $B \longrightarrow a = -4.445552x10^{-3}$.
- $A \longrightarrow b = 7.111334964$.
- $r = -0.974911601$.

Et voila , l'équation de la relation linéaire est :

$$y = -4.445552 * 10^{-3}X + 7.111334964. \tag{3.4}$$

en introduisant l'exponentiel des deux cotés de l'équation , on obtient :

$$e^{y} = e^{-4.445552\times10^{-3}X+7.111334964}$$

$$= e^{-4.445552*10^{-3}X} \times e^{7.111334964}.$$

Alors :

$$Y = Be^{aX} = 1225.78e^{-4.45\times10^{-3}X}. \tag{3.5}$$

$2^{éme}$ **méthode :**

Maintenant , on utiliser un méthode plus facile.

Directement vous choisissez le modèle : régression exponentielle

$$Y = Be^{aX}.$$

Avec la calculatrice :

$$Y = 1225.78238e^{-4.445555\times10^{-3}X}.$$

Remarque 3.3.1. *La valeur donnée par la calculatrice du coefficient de corrélation r est la valeur de la corrélation linéaire C'est-à-dire :*

$$r = cor(X, \ln Y) = -0.9749.$$

3.3.2 Régression Puissance

l'équation Puissance (Forme non linéaire) est :

$$Y = BX^a.$$

ET l'équation Puissance (forme linéaire transformation logarithmique) :

$$\ln(y) = \ln(BX^a)$$

$$\ln(Y) = \ln(B) + a\ln(X).$$

Linéarisation da le forme puissance

$1^{ére}$ **méthode :** On prend

$$(**)\begin{cases} y = \ln(Y) \\ x = \ln(X) \\ b = ln(B) \end{cases}$$

On obtient

$$y = ax + b$$

Donc , on écrit une relation linéaire entre deux autres variables en fonction des variables initiales , C'est-à-dire Création d'une relation linéaire entre $(\ln X)$ et $(\ln Y)$.

Ensuite, vous choisissez une régression linéaire (par la méthode des moindres carrées)

Exemple 3.3.2. *Les deux variables sont* $\ln X$ *et* $\ln Y$ *,et les données de ces deux variables présente dans le tableau suivantes :*

X	*100*	*200*	*130*	*190*	*250*	*310*
Y	*1240*	*1190*	*1160*	*1135*	*1094*	*1065*

Création d'une relation linaire entre $(\ln X)$ *et* $(\ln Y)$*, Les deux variables sont* X *et* $\ln Y = y$ *,et les données de ces deux variables présente dans le tableau suivantes :*

$\ln(X)$	*2.30*	*5.289*	*4.868*	*5.247*	*5.521*	*5.737*
$y = \ln Y$	*7.123*	*7.082*	*7.056*	*7.034*	*7.085*	*6.971*

et voila

$$y = aX + b.$$

on trouver

- $A \longrightarrow b = 7.65708.$
- $B \longrightarrow a = -0.11763.$
- $r = -0.882249$.

Et voila , l'équation de la relation linéaire est :

$$y = -0.11763 \times 10^{-3}x + 7.65708. \tag{3.6}$$

en introduisant l'exponentiel des deux cotés de l'équation , on obtient :

$$e^y = e^{-0.11763\times 10^{-3}x + 7.65708}$$

$$= e^{-0.11763\times 10^{-3}x} \times e^{7.65708}$$

$$e^{\ln(Y)} = e^{-0.11763\ln(X)} \times e^{7.65708}$$

$$= e^{\ln(X) - 0.11763} \times e^{7.65708}$$

Ajustement puissance

$$Y = e^{\ln(X) - 0.11763} \times e^{7.65708}$$

finalement , on a :

$$Y = 2115.570934 \times X^{-0.11736}.$$

$2^{éme}$ **méthode :**

Maintenant , on utiliser un méthode plus facile.

Directement vous choisissez le modèle : régression puissance

$$Y = BX^a..$$

Avec la calculatrice :

$$Y = 2115.58 \times X^{-0.1176}.$$

Remarque 3.3.2. *La valeur donnée par la calculatrice du coefficient de corrélation r est la valeur de la corrélation linéaire C'est-à-dire :*

$$r = cor(\ln(X), \ln(Y) = -0.882249.$$

3.4 Choix du modèle de régression valable

3.4.1 Choix entre la régression linéaire, exponentielle et puissance

1. Vous calculez les trois coefficient de corrélation .
2. Vous choisissez le coefficient le plus élevé en **valeur absolue**

$$r_1 = cor(X, Y) \text{ , } r_2 = cor(X, \ln(Y)) \text{ et } r_3 = cor(\ln(X), \ln(Y)).$$

3. Le modèle associé à ce coefficient est le plus valable.

– Ou bien, directement vous choisissez le coefficient de détermination le plus élevé .

– Le modèle associe à ce coefficient de détermination est le plus convenable .

– R^2 le carré du coefficient de corrélation.

3.4.2 Exemple

volume (mol)	1	2	3	4	5	6	7	8	9	10
concentration (mol/L)	6.53	4.12	2.05	1.5	1	0.8	0.6	0.5	0.23	0.01

Le modèle linéaire

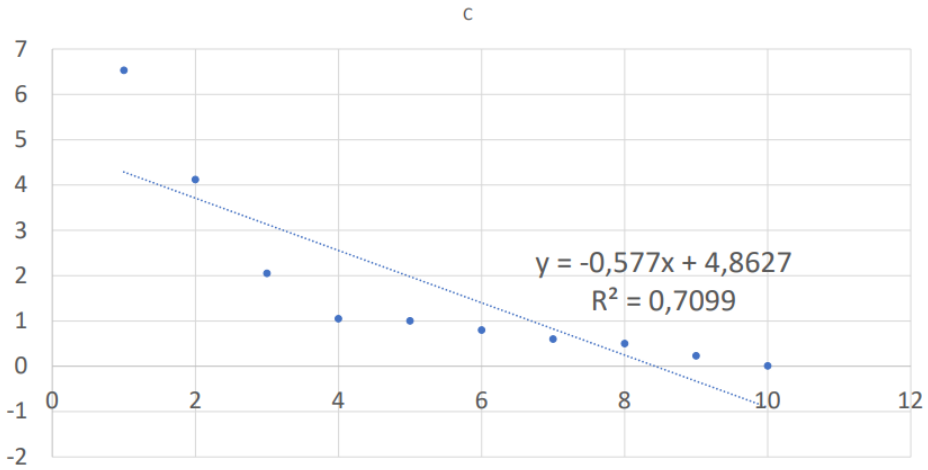

FIGURE 3.1: Nuage de points de modèle linéaire

Le modèle exponentiel

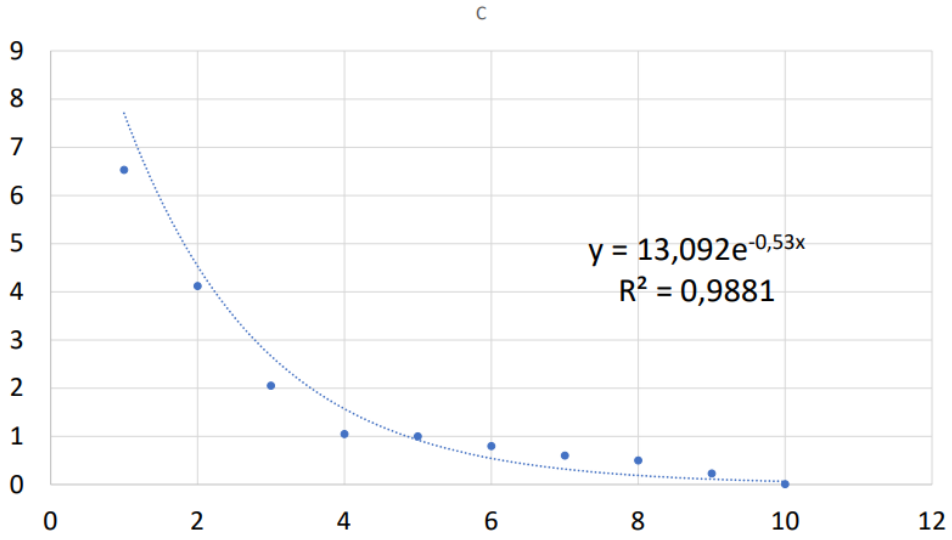

FIGURE 3.2: Nuage de points de modèle exponentiel

Le modèle puissance

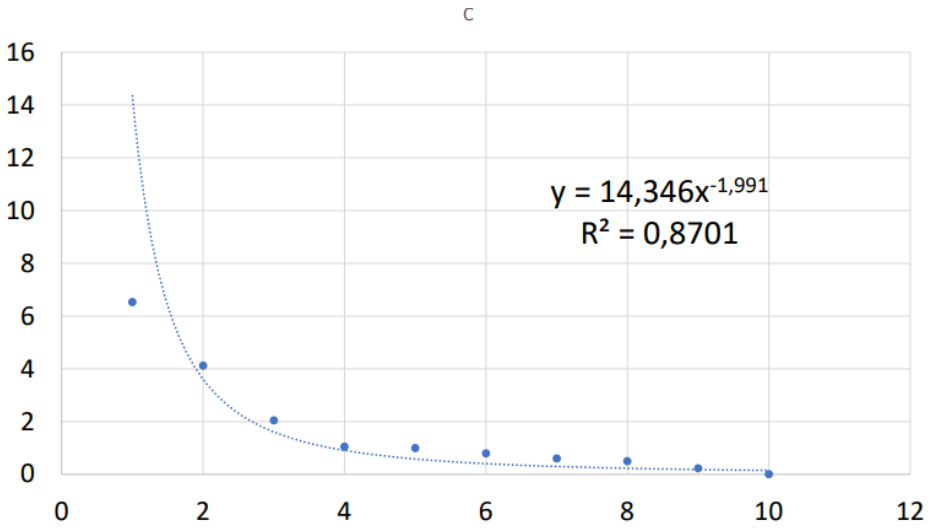

FIGURE 3.3: Nuage de points de modèle puissance

Calcul par SPSS

Modèle	Régression	Coefficient de Détermination
Linéaire	C = -0.577v + 4.8627	0.7099
Exponentiel	$C = 13.092e^{-0.53V}$	0.9881
Puissance	$Ph = 14.346X^{-1.991}$	0.8701

Le coefficient de détermination le plus élevé est 0,9881 .

Le modèle valable c'est le modèle exponentiel .

CHAPITRE 4

APPLICATIONS

dans ce chapitre nous abordons quelques applications sur la régression linéaire simple , multiple et non linéaire à l'aide du logiciel SPSS .

4.1 Logiciel utilisé

Au cours des applications, on utilise le logiciel SPSS.

Logiciel SPSS

SPSS (signifie statistical package for the social sciences) son objectif est d'offrir un logiciel permettant de réaliser la totalité des analyses statistiques Utilisées en sciences humaines , mais aussi en économie, biologie ,médecine ,ingénierie ... ect . c'est un logiciel très complet. il existe bien d'autres logiciels comme s- plus , R ou SAS qui permettent d'atteindre les mêmes buts, c'est-à-dire faire des analyses statistiques . ression linéaire simple.

4.2 Régression simple

Commençons par une application en santé publique .

4.2.1 Application 1

Les données de cette application sont des prélèvements des patients hémodialyses au niveau de service de l'hémodialyse d'établissement public hospitalier « Saad Guermach » de la wilaya de Skikda, on prendre la calcémie en fonction de l'albuminémie. Les résultats sont présentées dans le

tableau suivant :

Essai numéro	Ca mesuré (Y_i)	ALB (X_i)
1	69	41
2	78	28
3	90	43
4	77	41
5	84	30
6	111	53
7	97	45
8	89	29
9	101	45
10	91	44
11	66	26
12	95	39
13	80	35
14	93	45
15	81	29
16	79	25
17	92	42
18	98	46
19	115	56.7
20	102	46

tab.1 :résultats de dosage de la calcémie et l'albuminémie pour patients hémodialysés.

On a

- la Variable dépendante : Ca mesuré (Y_i).
- la variable explicative : ALB (X_i)

A l'aide de SPSS , on a les corrélations suivantes :

	X_1	Y
X_1	1	0,788
Y	0,788	1

Passons maintenant a l'équation de l'hyperplan de régression .

D'après le SPSS , le tableau ci dessous donne le coefficients de la régression qui sont :

$$\hat{\beta}_0 = 45{,}225 \ , \ \hat{\beta}_1 = 1{,}119.$$

Modèle	$\hat{\beta}$
constante	45,225
X_1	1,119

ainsi L'équation de la régression de Y en fonction de X_1 est :

$$\hat{Y} = 45,225 + 1,119X_1.$$

L'intervalles de confiance pour les paramètres (quatre vingt quinze pour cent)

-$\hat{\beta}_0 \in$ [39.799 , 50.651].

- $\hat{\beta}_1 \in$ [-4.312 , 6.55]

Tableau dANOVA D'après le SPSS ,le tableau dANOVA est le suivant :

Modèle	somme des carrés	ddl	carré moyen	F	sig
Expliquée	1952,984	1	1952,984	29,496	0,000
Résiduelle	1191,816	18	66,212		
Total	314,800	19			

Alors

$$\text{SCE} = 1952{,}984 \ .$$

$$\text{SCR} = 1191{,}816 \ .$$

$$\text{SCT} = 3144{,}800 \ .$$

Le coefficient de détermination R^2 : mesure la proportion de variabilité expliquée par l'une variable explicative , est calculé par le coefficient de détermination que données par :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}.$$

R	R^2
0,788	0,621

$R^2 = 62{,}1$ Ce qui montre un ajustement très fort. ce coefficient ne prendre pas en compte le nombre de variables explicatives .

Terminons par la signification globale du modèle de régression

La liaison globale entre Y est lesX_i est-elle significative ?

Au seuil de signification = 0,05 , d'après le SPSS , f_{obs} =29,496 est supérieure à la valeur critique $f_{1,18}^{0.95} = 4{,}41$ qui est la valeur tabulaire de fisher . ceci est confirmé par la valeurs de tableau ANOVA.

Ce résultat montre que l'albuminémie contribue à expliquer la calcémie .

Passons maintenant à une application en chimique.

4.2.2 Application 2

L'analyse de la température de fonctionnement d'un procédé chimique sur le rendement du produit a donné les valeurs suivantes pour la température X_i et le rendement correspondant Y_i . Les résultats sont présentées dans le tableau suivant :

Essai numéro	la température(Y_i)	le rendement(X_i)
1	45	100
2	51	110
3	54	120
4	61	130
5	66	140
6	70	150
7	74	160
8	78	170
9	85	180
10	89	190

tab.1 :résultats de L'analyse de la température de fonctionnement d'un procédé chimique sur le rendement du produit.

On a

- la Variable dépendante : la température (Y_i).
- la variable explicative : le rendement (X_i)

A l'aide de SPSS , on a les corrélations suivantes :

	X_1	Y
X_1	1	0,998
Y	0,998	1

Passons maintenant a l'équation de l'hyperplan de régression .

D'après le SPSS , le tableau ci dessous donne le coefficients de la régression qui sont :

$$\hat{\beta}_0 = \text{-}2.739 \ , \ \hat{\beta}_1 = 0.483.$$

Modèle	$\hat{\beta}$
constante	-2.739
X_1	0.483

ainsi L'équation de la régression de Y en fonction de X_1 est :

$$\hat{Y} = -2.739 + 0.483X_1.$$

L'intervalles de confiance pour les paramètres (quatre vingt quinze pour cent)

-$\hat{\beta}_0 \in$ [39.799 , 50.651].

- $\hat{\beta}_1 \in$ [-4.312 , 6.55]

Tableau dANOVA D'après le SPSS ,le tableau dANOVA est le suivant :

Modèle	somme des carrés	ddl	carré moyen	F	sig
Expliquée	1924,876	1	1924,876	2131.574	0,000
Résiduelle	7.224	8	0.903		
Total	1932.100	9			

Alors

$$\text{SCE} = 1924{,}876 \ .$$

$$\text{SCR} = 7.224 \ .$$

$$\text{SCT} = 1932.100 \ .$$

Le coefficient de détermination R^2 : mesure la proportion de variabilité expliquée par l'une variable explicative , est calculé par le coefficient de détermination que données par :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}.$$

R	R^2
0,998	0,996

$R^2 = 99.6$ Ce qui montre un ajustement très fort. ce coefficient ne prendre pas en compte le nombre de variables explicatives .

Terminons par la signification globale du modèle de régression

La liaison globale entre Y est lesX_i est-elle significative ?

Au seuil de signification = 0,05 , d'après le SPSS , f_{obs} =2131.574 est supérieure à la valeur critique $f_{1,8}^{0.95} = 5.32$ qui est la valeur tabulaire de fisher . ceci est confirmé par la valeurs de tableau ANOVA. Ce résultat montre que la température de fonctionnement d'un procédé chimique contribue à expliquer le rendement du produit .

4.3 Régression linéaire multiple

Commençant par une application en la consommation de textile .

4.3.1 Application 3

Pour illustrer la mise en oeuvre du test de Durbin-Watson, nous reprenons un exemple extrait de l'ouvrage de Theil . L'objectif est de prédire la consommation de textile à partir du revenu par tête des personnes et du prix. Nous disposons d'observations sur 17 années à partir de 1923. Les résultats sont présentés dans le tableau suivant :

Essai numéro	conso (Y_i)	revenu (X_{i1})	prix (X_{i2})
1	99.2	96.7	101
2	99	98.1	100.1
3	100	100	100
4	111.6	104.9	90.5
5	122.2	104.9	86.5
6	117.6	109.5	89.7
7	121.1	110.8	90.6
8	136	112.3	82.8
9	154.2	109.3	70.1
10	153.6	105.3	65.4
11	158.5	101.7	61.3
12	140.6	95.4	62.5
13	136.2	96.4	63.6
14	168	97.6	52.6
15	154.3	102.4	59.7
16	149	101.6	59.5
17	165.5	103.8	61.3

tab.2 : Données de Theil sur le textile.

On a

- la Variable dépendante :conso (Y_i).
- les variables explicatives :revenu (X_{i1}) et prix (X_{i2}).

A l'aide de SPSS , on a les corrélations suivantes :

	X_1	X_2	Y
X_1	1	0,062	-0,947
X_2	0,062	1	0,179
Y	-0,947	0,179	1

D'après le SPSS , le tableau ci dessous donne le coefficients de la régression qui sont :

$$\hat{\beta}_0 = 130.764 \ , \ \hat{\beta}_1 = 1.061 \ , \ \hat{\beta}_2 = -1.383.$$

Modèle	$\hat{\beta}$
constante	130.764
X_1	1.061
X_2	-1.383

ainsi L'équation de la régression de Y en fonction de X_1 et X_2 est :

$$\hat{Y} = 130.764 + 1.061X_1 - 1.383X_2.$$

L'intervalles de confiance pour les paramètres au seuil (quatre vingt quinze pour cent)

-$\hat{\beta}_0 \in$ [125.946 , 135.582].

- $\hat{\beta}_1 \in$ [-2.912 , 5.034]

- $\hat{\beta}_2 \in$ [-17.855 , 15.089]

Tableau dANOVA : D'après le SPSS ,le tableau dANOVA est le suivant :

Modèle	somme des carrés	ddl	carré moyen	F	sig
Expliquée	8459.486	2	4229.743	136.204	0,000
Résiduelle	434.764	14	31.055		
Total	8894.249	16			

Alors

$$\text{SCE} = 8459.486 .$$

$$\text{SCR} = 434.764 .$$

$$\text{SCT} = 8894.249 .$$

Le coefficient de détermination R^2 : mesure la proportion de variabilité expliquée par les 02 variables explicatives , est calculé par le coefficient de détermination R^2 que données par :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}.$$

R	R^2
0.975	0.951

$R^2 = 95{,}1$ Ce qui montre un ajustement très fort. ce coefficient ne prendre pas en compte le nombre de variables explicatives .

Terminons par la signification globale du modèle de régression

La liaison globale entre Y est lesX_i est-elle significative ?

Au seuil de signification = 0,05 , d'après le SPSS , f_{obs} =136.204 est supérieure à la valeur critique $f_{2,14}^{0.95} = 3.74$ qui est la valeur tabulaire de fisher . ceci est confirmé par la valeurs de tableau ANOVA. Ce résultat montre qu'au moins une des variables contribue à expliquer la consommation de textile Y. Mais, il est global et ne nous indique pas si plusieurs variables qui contribuent et lesquelles.

Passons maintenant à une application en Consommation des véhicules.

4.3.2 Application 4

On veut expliquer la consommation en L/100km de véhicules à partir de p = 4 variables : le prix, la cylindrée, la puissance et le poids . Nous disposons de n = 20 observations. Les résultats sont présentés dans le tableau suivant :

Essai numéro	consommation (Y_i)	prix (X_{i1})	puissance (X_{i2})	cylindrée (X_{i3})	poids (X_{i4})
1	5.7	11600	32	846	650
2	5.8	12490	39	993	790
3	6.1	10450	29	899	730
4	6.5	17140	44	1390	955
5	6.8	14825	33	1195	895
6	6.8	13730	32	658	740
7	7.1	19490	55	1331	1010
8	21.3	285000	325	5474	1690
9	18.7	183900	300	5987	2250
10	14.5	92500	209	2789	1485
11	7.4	25000	74	1597	1080
12	9	22350	74	1761	1100
13	11.7	36600	101	2165	1500
14	9.5	22500	85	1983	1075
15	9.5	31580	85	1984	1155
16	8.8	28750	89	1998	1140
17	9.3	22600	65	1580	1080
18	8.6	20300	54	1390	1110
19	7.7	19900	66	1396	1140
20	10.8	39800	106	2435	1370

tab.3 : Tableau de données CONSO - Consommation des véhicules.

On a

– la Variable dépendante :consommation (Y_i).

– les variables explicatives :prix (X_{i1}) , puissance (X_{i2}) ,cylindrée (X_{i3}) et poids (X_{i4}) .

A l'aide de SPSS , on a les corrélations suivantes :

	X_1	X_2	X_3	X_4	Y
X_1	1	0,939	0,981	0,960	0,903
X_2	0,939	1	0,954	0,929	0,768
X_3	0,981	0,954	1	0,972	0,893
X_4	0,960	0,929	0,972	1	0,927
Y	0,903	0,768	0,893	0,927	1

Passons maintenant à l'équation de l'hyperplan de régression

D'après le SPSS , le tableau ci dessous donne le coefficients de la régression qui sont :

$\hat{\beta}_0$=2.630 , $\hat{\beta}_1$= 2.584E-005 , $\hat{\beta}_2$= 0.027, $\hat{\beta}_3$= -0.001 , $\hat{\beta}_4$= 0.005.

Modèle	$\hat{\beta}$
constante	2.630
X_1	2.584E-005
X_2	0.027
X_3	-0.001
X_4	0.005

ainsi L'équation de la régression de Y en fonction de X_1 , X_2 , X_3 et X_4 est :

$$\hat{Y} = 2.630 + 2.584(E-005)X_1 + 0.027X_2 - 0.001X_3 + 0.005X_4.$$

L'intervalles de confiance pour les paramètres au seuil (quatre vingt quinze pour cent)

-$\hat{\beta}_0 \in$ [0.173 , 5.087].

- $\hat{\beta}_1 \in$ [0.678 , 4.549]

- $\hat{\beta}_2 \in$ [-2.066 , 2.12]

-$\hat{\beta}_3 \in$ [-1.248 , 1.246]

- $\hat{\beta}_4 \in$ [-2.296 , 2.306]

Tableau dANOVA : D'après le SPSS ,le tableau dANOVA est le suivant :

Modèle	somme des carrés	ddl	carré moyen	F	sig
Expliquée	324.467	4	81.117	133.923	0,000
Résiduelle	9.085	15	0.606		
Total	333.552	19			

Alors

$$\text{SCE} = 324.467 \ .$$

$$\text{SCR} = 9.085 \ .$$

$$\text{SCT} = 333.552 \ .$$

Le coefficient de détermination R^2 : mesure la proportion de variabilité expliquée par les 04 variables explicatives , est calculé par le coefficient de détermination R^2 que données par :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}.$$

R	R^2
0.986	0.973

$R^2 = 97{,}3$ Ce qui montre un ajustement très fort. ce coefficient ne prendre pas en compte le nombre de variables explicatives .

Terminons par la signification globale du modèle de régression

La liaison globale entre Y est lesX_i est-elle significative ?

Au seuil de signification = 0,05 , d'après le SPSS , f_{obs} =133.923 est supérieure à la valeur critique $f_{4,15}^{0.95} = 3.07$ qui est la valeur tabulaire de fisher . ceci est confirmé par la valeurs de tableau ANOVA. Ce résultat montre qu'au moins une des variables contribue à expliquer la Consommation des véhicules Y. Mais, il est global et ne nous indique pas si plusieurs variables qui contribuent et lesquelles.

4.4 Régression non linéaire

Commençant par une application en mesures de poids (variable x) et taille (variable y) .

4.4.1 Application5

Soient les 2 mesures de poids (variable x) et taille (variable y) relevées sur un échantillon de 30 objets. on a la relation entre X et Y est écrit par :

$$Y = bX^a$$

donc en linéairement de cet modèle est : $\ln Y = \ln b + a \ln X$ alors par changement des variables on pose :

$$Y' = \ln Y; X' = \ln X; b' = \ln b$$

Alors :

$$Y' = aX' + b'$$

donc l'équation de régression est :

$$\hat{Y_i'} = \hat{a}X_i' + \hat{b}$$

$$\bar{Y'} = \hat{a}\bar{X'} + \hat{b}$$

avec :

$$\hat{a} = \frac{\sum_{i=1}^{30} X_i'Y_i' - n\bar{X'}\bar{Y'}}{\sum X_i'^2 - n\bar{X'}^2}$$

pour calcul $\hat{a}$ et $\hat{b}$ en calcul X_i et Y_i les résulta donne le tableau suivant :

	Poids (X)	Taille (Y)	$X' = \ln X$	$Y' = \ln Y$	X'.Y'
1	46	152	3.8286	5.0238	19.2346
2	78	158	4.3567	5.0626	22.05462
3	85	160	4.4426	5.0752	22.5472
4	85	162	4.4426	50876	22.6024
5	85	158	4.4426	5.0626	22.4913
6	85	159	4.4426	5.0689	22.5194
7	85	161	4.4426	5.0814	22.5749
8	95	165	4.5539	5.1059	23.2519
9	95	166	4.5539	5.1119	23.2794
10	100	168	4.6052	5.1239	23.5967
11	100	163	4.6052	5.0998	23.4576
12	100	164	4.6052	5.1239	23.4858
13	103	168	4.6347	5.1119	23.7482
14	105	166	4.6539	5.1239	23.7909
15	105	168	4.6539	5.1119	23.8467
16	115	166	4.7449	5.1239	24.2560
17	112	168	4.7185	5.1358	24.1774
18	115	170	4.7449	5.0875	24.3690
19	115	162	4.7449	5.0876	24.1403
20	130	165	4.8675	5.1059	24.8533
21	135	167	4.9053	5.1179	25.1052
22	150	172	5.0106	5.1475	25.7922
23	60	170	4.0943	5.1358	21.0277
24	65	172	4.1744	5.1475	21.4876
25	70	168	4.2485	5.1239	21.7691
26	80	184	4.3820	5.2149	22.8519
27	61	158	4.1109	5.0626	20.8117
28	58	160	4.0604	5.07517	20.6074
29	66	164	4.1897	5.0998	21.3667
30	58	160	4.0604	5.07517	20.6074
somme	2742	4944	134.3219	153.1234	685.7066

donc

$$\hat{a} = \frac{\sum_{i=1}^{30} X_i' Y_i' - n\bar{X}'\bar{Y}'}{\sum X_i'^2 - n\bar{X}'^2}$$

$$= \frac{0.1124}{2.345576}$$

Alors :

$$\hat{a} = 0.04792$$

Et on à

$$\hat{b'} = \bar{y'} - \hat{a}\bar{x'} = 5.104112 - 4.477398(0.04792) = 4.88957.$$

Alors :

$$\hat{b'} = \ln(\hat{b}) \Rightarrow = \exp(\hat{b'}) = \exp(4.88957) = 132.8964$$

Alors :

$\hat{b} = 132.8964$ et $\hat{a} = 0.04792$ et la relation enter X et Y est :

$$\hat{y} = \hat{b}x^{\hat{a}} = 132.8964X^{0.04792}.$$

- **Les hypothèses relatives à ce modèle :**

  H_0 : il y à une relation dans la poids (X) et taille (Y)de le homme.

  H_0 : il n y pas une relation dans la poids (X) et taille (Y)de le homme.

- **Teste de signification du modèle :**

  $T_{cal} = 3.813533$ car :

  $T_{cal} = \frac{\hat{a}}{\sqrt{var(\hat{a})}} = \frac{0.04792}{\sqrt{4.889}} = 3.813533$

  et $t_{n-2}^{1-\frac{\alpha}{2}}$ alors

  $T_{cal} \geq t_{n-2}^{1-\frac{\alpha}{2}} = 2.0423$

alors on rejette H_0 donc il n y pas une relation dans la poids et taille de le homme

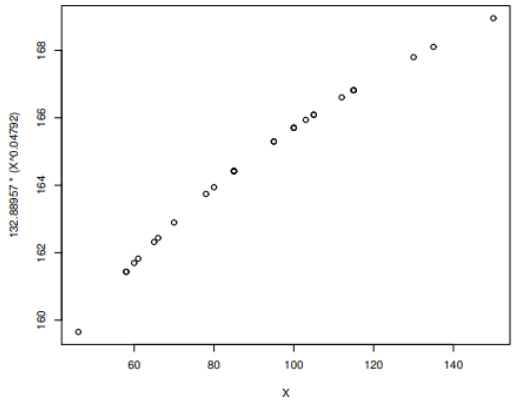

FIGURE 4.1: nuage de points de modéle non linéaire de cet application

CONCLUSION

la méthode de régression est une des solutions qui existe pour observer et expliquer les liens entre une variable quantitative dépendante et plusieurs variables quantitatives indépendantes et de faire des prévisions sur la problématique à étudier . la régression reste toujours une méthode vitale et est utilisé dans divres domaines à savoir : finance , marketing , économie , ingénierie , biologie , ...

BIBLIOGRAPHIE

[1] Cornillon, P.A. Matzner-Lober, E. *Regression theorie et applications (pp. 302-p).Springer,(2007).*

[2] citek Dodge, Y.*Rousson, V Analyse de régression appliquée, Dunod, 2i eme édition, (2004)* .

[3] Faraway, J. , Practical Regression and ANOVA using R, July 2002.

[4] Fourcassie V. E Jost, C. *Introduction aux modeles hésites généraux (General linear model - GLM) Cours Modules Statistiques Master 2 NCC. (2012).*

[5] Francis Galton, *Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute 15 : 246-63, (1886).*

[6] Genest, C. , Modèle de régression linéaire multiple,.

[7] Gerald Baill Argeon statistique, *les étutions SMG, Québec, CANADA, (1995).*

[8] Mme Medouer Nawel . *Cours de Biostatistique, (2020)*

[9] Montgomery, D.C., *E Peck, E.A. Introduction to linear regression analysis, 2nd edition, ed. 1 Wiley E Sons, p 527, (1992).*

[10] Ricco Rakotomalala. *Modèle de régression linéaire Pratique de la Régression Linéaire Multiple - Diagnostic et sélection de variables,May-2015.*

[11] Saporta, G.*Probabilités, analyse des données et statistiques (2 ème édition), Technip, Paris, (2010).*

[12] Theil, H . , Principles of Econometrics, Wiley, (1971). Page 102.

[13] Yadolah Dodge . *analyse de régression appliquée, Dunod, PARIS, (1999)*

[14] Yves Tillé . *Résumé du Cours de Statistique Descriptive, (2010)*