UNIVERSITE KASDI MERBAH
OUARGLA

Faculté des Mathématiques et Sciences de La Matière

N° d'ordre :
N° de série :

DEPARTEMENT DE MATHEMATIQUES

MASTER

Spécialité : Mathématiques

Option : Probabilité et Statistique

Présenté Par :Hachmine Abdelkader

Théme :

Spécification des modèles linéaires et non-linéaires dans l'inférence statistique (régression linéaire et non-linéaire)

Soutenu publiquement le :30 mai 2017

Devant le jury composé de :

Mr. **Mohamed Agti**	M. A. université de KASDI Merbah - Ouargla	Président
Melle. **Fatima Meddi**	M. C. université de KASDI Merbah - Ouargla	Examinateur
Mr. **Djamel Eddine Chetti**	M. C. université de KASDI Merbah - Ouargla	Rapporteur

Année universitaire 2016/2017

Dédication

A ma chère mère , et mon cher père.

Pour leur amour inestimable,leur confiance,

leur soutien, leurs sacri ces et toutes les valeurs qu'ils ont su m'inculquer.

A toute ma famille ainsi qu'à mes amis.

Remerciements

Avant toute considération,je remercie le Grand Dieu le tout puissant qui, m'a aidé pour achever ce travail.

Je tiens tout a remercier premier lieu mon encadreur Monsieur **Chetti Djamel Eddine** d'avoir accepté de m'encadrer et pour sa continuité à me soutenir et à m'encourager. Je voudrai aussi le remercier pour sa gentillesse, sa disponibilité et du temps consacré à mon travail.

J'ai de la chance parce que Monsieur **Agti Mohamed** ait acceptée d'être président de ce travail ainsi que pour l'attention qu'il a porté mon travail malgré son emploi du temps très chargé.

Je tiens également remercier Melle. **Meddi Fatima** pour avoir accepté de faire partie de mon jury. Je remercie également les membres du département de Mathématique et Informatique de m'avoir permis de travailler dans de bonnes conditions pendant la réalisation de mon travail.

Merci également a tous les enseignants qui m'ont aidé pendant mon cursus, sans oublier leurs conseils précieux.

Je remercie aussi toute personne de prés ou de loin a contribué à la finalisation de ce travail.

Table des matières

Table des figures

Liste des tableaux

Notations et conventions

θ :paramétre de forme.

$Var[X]$: variance mathémaatique de v.a.X

$E[X]$: espérance mathématique ou moyenne du v.a. X

$\Re$: ensemble des valeurs réelles.

i, i, d : indépendantes et identiquement distribuées.

$L(x, \theta)$: fonction de vraisemblance.

$X_1, X_2, ..., X_n$: échantillon de taille n de X

(Ω, F, P) : espace de probabilité.

ε_i : résidus ou erreur.

SCT : somme des carrées totale

SCE : somme des carrées expliquée

SCR : somme des carrés resoluelle

I_c : l'intervalle de confiance

$\nabla f(x)$: La dérivée partielle de f

$ANOVA$: analyse de la variance

$N(0, 1)$: loi normale centre reduite.

INTRODUCTION

Pour certaines observations, il arrive que les valeurs de la variable réponse ou des préviseurs semblent , se comporter différemment de la majorité des observations donc ces observations qui ne suivent pas le même modèle (linéaire ou non linéaire) que la majorité des données ces appelées des valeurs aberrantes (de la régression).

La régression est l'un des méthodes le plus appliquées en statistique pour l'analyse de données quantitatives. Elle est utilisée pour établir une relation entre une variable quantitative et une ou plusieurs autres variables quantitatives, sous la forme d'un modèle. Si on s'intéresse à la relation entre deux variables, on parlera de régression simple en exprimant une variable en fonction de l'autre. Si la relation porte entre une variable et plusieurs variables auters, on parlera de régression multiple. La mise en oeuvre d'une régression impose l'existence d'une relation de cause à effet entre les variables prises en compte dans le modèle Pour pouvoir faire, au minimum, l'estimation ponctuelle des paramétres b_i et σ^2, il est indispensable de répliquer, de manières indépendantes, les observations simultanées des variables x_i et y .

Nous supposerons donc par la suite que n observations indépendantes sont réalisées et nous écrirons le modèle, pour la i-i'eme observation $(i = 1, ..., n)$ sous la forme :

$$Y_i = \sum_{i=0}^{n} b_i X_j^i + a_i$$

Les valeurs observèes des variables seront notées par des minuscules, de sorte qu'on écrira :

$$y_i = \sum_{i=0}^{n} b_i x_j^i + a_i$$

Les modèles linéaire et les modèles non linéaire (définition,théorem .....) sont présontés dans le chapitre 1.

Dans le chapitre 2 ,on à introduit le modèle statistique de la régression linéaire (simple et multiple) et régression non linéaire , qui permet de rendre compte une variable quantitative (variable dite à expliquer) comme fonction afine d'une seule autre variable quantitative (variable dite explicative), à quoi s'ajoute un terme résiduel correspondant aux inffuences d'autres facteurs explicatifs possibles non considérés.

Des application sur la régression linéaire et non linéaire sont exposées dans chapitre 3 , d'étudier les application de régression sur R.

Chapitre 1

Généraliter sur les modèles linéaire et les modèles non linéaire

1.1 Modèles mathématiques

Une grande partie des mathématiques appliquées consiste à faire de la modélisation dans diverses disciplines ,on peut distinguer :
la modélisation déterministe où on ne prend pas en compte des variations aléatoires par le biais d'outils (EDO, EDP, filtrage,...) [10], la modélisation stochastique (qui prend en compte ces variations aléatoires en essayant de leur associer des lois de probabilité)

- modélisation stochastique :

a pour but essentiel de préciser des classes de lois de probabilité rendant compte des variations aléatoires des phénomènes d'intérêt, variations dues à des causes soit inconnues, soit impossible à mesurer (cachées, trop nombreuses,...)

1.2 Modèles Statistique

definition 1.3.1 : On appelle modèle statistique, la donnée d'un espace des observations E, d'une tribu ε d'événements sur E et d'une famille de probabilités p sur l'espace probabilisable (E, ε). On le note(E, ε, ρ) ou, quand il n'y a pas de risque de confusion, plus simplement P.

definition 1.3.2 : Soient $X_1, ..., X_n$ des variables aléatoires défnies sur (Ω, A, φ) dans un espace mesurable (E_n, ε_n) . On suppose qu'on observe $x_1, ..., x_n$ qui sont telles que pour tout i, x_i est une réalisation (un tirage) de la variable aléatoire X_i. Supposons que la distribution jointe de $X = X_1, ..., X_n$ est inconnue mais qu'elle appartient à une famille particulière de distributions. Le couple formé par l'espace d'observation E_n et cette famille de distributions est appelé modèle statistique. On note(E_n, φ_n) .

Variables qualitatives :

Appelèe également variable catégorisée, une variable qualitative est une variable dont les modalitès ne peuvent pas ètre « mesurées » sur une èchelle spécifique. C'est le cas par exemple de la couleur des cheveux ou du degrè d'appréciation d'un certain objet dans le cadre d'un questionnaire avec jugement de prèférence. On distinguera les variables nominales des variables ordinales, qui peuvent ètre « ordonnées » ou recodées sur une èchelle arbitraire. C'est le cas par exemple d'une variable du type « niveau d'expertise » avec les modalités, ou niveaux, « faible », « intermédiaire » et « avancè ».[5]

Variables quantitatives :

Les variables quantitatives, ou numériques, possédent quant à elles une « métrique », c'est à-dire qu'elles peuvent ètre représentées sur une èchelle spécifique de mesure. On distinguera les variables d'intervalle, qui supportent des transformations linéaires (de type $y = ax$), des variables dites de rapport, supportant les transformations afines (de type $y = ax + b$). Dans ce dernier cas, il existe une origine, ou un zéro, qui a un sens. Des

exemples de telles variables sont : la température (intervalle), la taille ou un temps de présentation (rapport), etc.[5]

1.2.1 Modèle paramétrique

Un modèle paramétrique est une famille de lois $M = \{P_\theta : \theta \in \Theta\}$indexée par un ensemble fini, disons P, de paramètres : θ
le modèle est indéxé par un nombres ou un vecteur réel. p est la dimension du modèle

modéle de Bernoulli :$P = B(p), p \in [0; 1]$ pour l'observation des notes de qualité de n piéces mécaniques (0 si piéce correcte, 1 si piéce défectueuse).

1.2.2 Modèles Paramétriques classiques

Le modéle linéaire (gaussien) de base : à la fois le plus simple, le plus ancien et le plus connu des modéles statistiques, il englobe essentiellement la régression linéaire, l'analyse de variance et l'analyse de covariance. Dans ce modéle, les variables explicatives (régresseurs ou facteurs) ne sont pas aléatoires Pour pouvoir ètre exploitè pleinement, ce modéle nécessite l'hypothése de normalitè des erreurs, donc de la variable à expliquer. [3]

Les modéles non linéaires : De façon très gènèrale,il s'agit de modèles permettant d'expliquer la variable rèponse (alèatoire) au moyen des variables explicatives (non alèatoires dans les modèles usuels), à travers une fonction quelconque, inconnue (on est donc en dehors du cadre du modèle linéaire généralisé). Cette classe de modéles est très vaste et relève, en général, de la statistique non paramétrique. Citons, à titre d'exemple, la régression non paramétrique, les GAM (Generalized Additive Models) et les réseaux de neurones.[4]

Les modéles mixtes : On étend les modèles précédents au cas où les variables explicatives sont elles aussi aléatoires avec spécification de leurs lois (effets aléatoires) : Ceci permet d'expliquer une plus grande variabilité des données.

Les modéles pour séries chronologiques : Les séries chronologiques sont les observations, au cours du temps, d'une certaine grandeur représentant un phénoméne économique, social ou autre. Si données répètées et séries chronologiques ont en commun de rendre compte de l'évolution au cours du temps d'un phénoméne donnè, on notera que ces deux types de données ne sont pas réellement de mème nature (dans une sèrie chronologique, ce sont rarement des personnes ou des animaux que l'on observe). Pour les sèries chronologiques, on utilise des modéles spécifiques : modéles AR , MA, ARMA, ARIMA ...[5]

1.3 Modèles linéaires

definition : On appelle modèle linéaire un modèle statistique qui peut s'écrire sous la forme :

$$y = \sum_{i=0}^{n} b_i x^i + a$$

Dans la défnition ci-dessus, les éléments intervenant ont les caractéristiques suivantes :

- y : est une variable aléatoire réelle (v.a.r.) que l'on observe et que l'on souhaite expliquer, ou prédire (ou les deux à la fois), on l'appelle variable à expliquer, ou variable réponse .

- Chaque variable x_i est une variable réelle , non aléatoire dans le modèle de base, également observée ; l'ensemble des x_i est censé expliquer y , en être la cause (au moins partiellement) ; les variables X_j sont appelées variables explicatives.

Les $b_i \quad i = 1, ..., n$ sont des coeficients, des paramétres, non observés ; on devra donc les estimer au moyen de techniques statistiques appropriées. ε_i est le terme d'erreur du modèle, c'est une v.a.r. non observée pour laquelle on fait systèmatiquement les hypothèses suivantes :

$$E(a) = 0 \quad , \quad Var(a) = \sigma^2$$

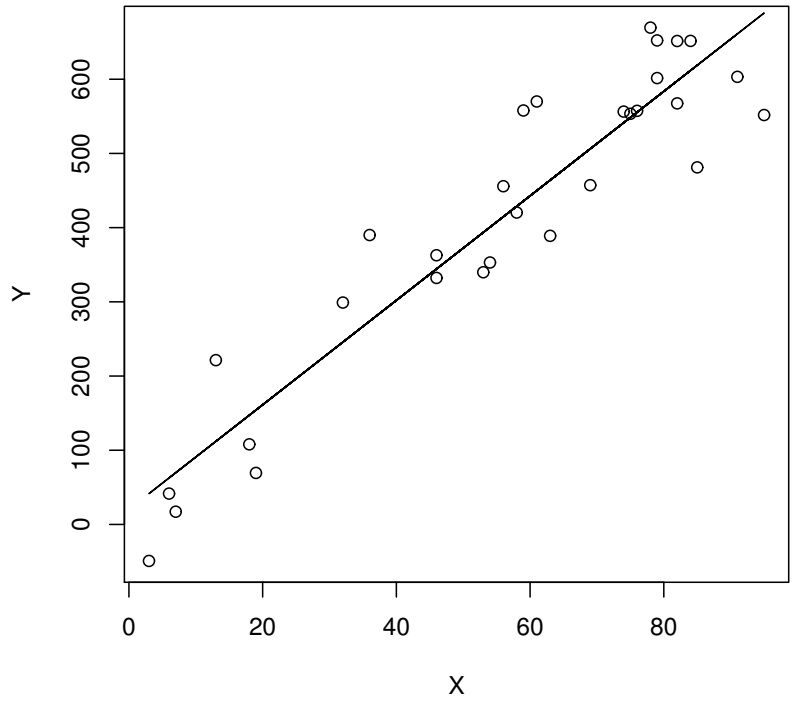

FIGURE 1.1 – nuage de points et droite de modèle linéaire

1.3.1 Notations des modéles linéaires

Pour pouvoir faire, au minimum, l'estimation ponctuelle des paramétres b_i et σ^2, il est indispensable de répliquer, de manières indépendantes, les observations simultanées des variables x_i et y .

Nous supposerons donc par la suite que n observations indépendantes sont réalisées et nous écrirons le modèle, pour la i-i'eme observation $(i = 1, ..., n)$ sous la forme :

$$Y_i = \sum_{i=0}^{n} b_i X_j^i + a_i$$

Les valeurs observèes des variables seront notées par des minuscules, de sorte qu'on écrira :

$$y_i = \sum_{i=0}^{n} b_i x_j^i + a_i$$

Par ailleurs, on notera $y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$ le vecteur aléatoire de R_n correspondant à l'ensemble de l'échantillon des v.a.r. réponses (la notation Y est identique à celle introduite pour une seule v.a.r. réponse, mais cela ne devrait pas entr â iminer de confusion puisqu'on travaillera dorénavant avec un échantillon), $X = X_i^j$ la matrice réelle, $n \times n$.

Exemples basiques

- ***Modèle de régressions linènaire simpele***

un modèle de la régrission linéaire simple est ecrit sous la forme :

$$y_i = ax_i + b + \varepsilon_i$$

généralement on écrit :

$$Y = b + aX + \varepsilon$$

avece :

✓ y : variable à expliquer.

✓ x : variable explicative.

✓ ε_i : résidus ou erreur.

- ***Modèle de régressions linènaire multiple :***

La régressions linènaire multiple est une généraisation de Régressions linènaire simpele son but est d'étudier et modéliser la relation enter une variable expliquée (y) et plusieurs variables explicatives $(x_1, x_2, ..., x_n)$.
le modèle de la (RLM) s'ecrit sous la forme :

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + .. + b_n x_{in} + \varepsilon_i$$

avece :

✓ y_i : i éme observation de y.

✓ $x_{i1}, ..., x_{in}$: les i éme observation de x.

✓ $b_1, ..., b_n$: les parameter du modèle de régressions linènaire multiple .

✓ ε_i : les erreurs de modèle.

- ***Modèle de analyse de la variance :***

Est une téchnique statistique qui sert à étudier l'infuenee d' un plusieurs facteurs qualitatifs sur une variable quantitative. Il y à pluseur cos de l'analyse de la variance commme AV(1) ,AV(2)....

1.4 Modèles non linéaire

Un modèles non linéaire puet être dans pluseur cos devient linéaire par linter médiaire des changements des variables par exemple on a les formes non linéaire suivantes :
$y = bx^a \quad , \quad y = \ln(ax + b) \quad , \quad y = \dfrac{1}{b + ax} \quad , \quad ba^x = \ln(y)$

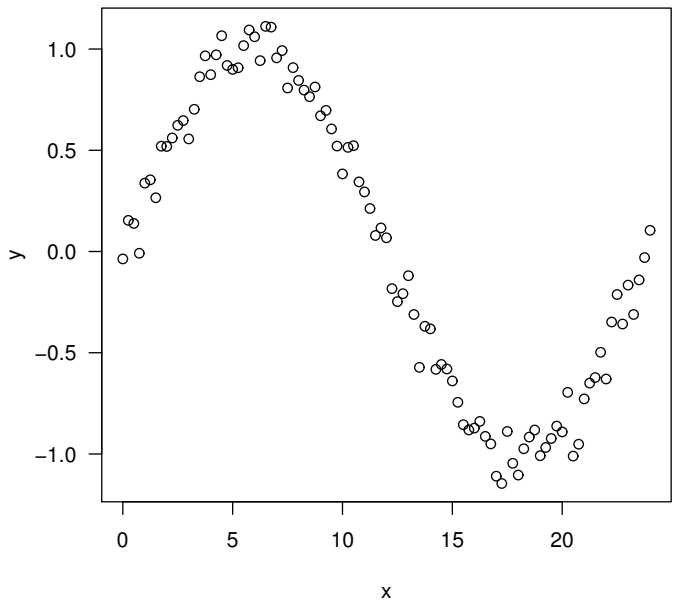

Figure 1.2 – nuage de point de modéle non linéaire

1.5 Estimation des paramétres

definition : On appelle estimateur de $f(\theta)$ toute statistique z (construite à partir de l'échantillon x) à valeurs dans l'image $f(\theta)$ de θ par f. Dans toute la suite, on supposera que Z est de carré intégrable au sens ou $\forall\theta \in \Theta, , E_\theta(|Z|^2) < \theta$.

1.5.1 Méthodes d'estimation

Méthode de moindre carré :

La méthode des moindres carrés consiste à estimer en minimisant la somme des carrés des résidus (SSR), telle que :

$$\wp(\hat{\theta}) = \min \sum_{i=1}^{n} (\hat{e_i})^2 = \min \sum_{i=1}^{n} (Y_i - \hat{Y_i})^2$$

Le critère des moindres carrés peut s'écrire aussi de la façon suivante :

$$\|\hat{e}\|^2 = \|Y - X\hat{\theta}\|^2 = \inf_{\theta \in \Re^n} \|Y - X\theta\|^2$$

Cette méthode d'estimation ne nécessite pas que l'on pose l'hypothèse de normalité des résidus.

Méthode de Maximum de Vraisemblance :

L'estimation par maximum de vraisemblance est basée sur la vraisemblance du modèle linéaire

$$L(\theta, y) = \prod_{i=1}^{n} f(y_i; \theta)$$

où $f(y_i; \theta)$ est la densité de la loi Normale sur y.
Pour obtenir l'estimateur $\hat{\theta}$ du maximum de vraisemblance, on maximise sa log-vraisemblance selon θ en résolvant le système d'équations du maximum de vraisemblance .

$$\frac{\partial \ln L(\theta_1,...,\theta_n,y)}{\partial \theta_i} = 0 \text{ pour } i = 1, ..., n$$

dont $\hat{\theta}$ est solution, sous réserve que la condition de seconde ordre soit vérifiée. On pourra également obtenir l'estimateur du MV de σ^2 en maximisant la log-vraisemblance selon σ^2.

Remarque 1.5.1 *Les estimateurs du Maximum de Vraisemblance de θ sont équivalents aux estimateurs des Moindres Carrés de θ. On pourra le montrer dans le cas de la régression linéaire. En revanche, certaines propriétés ne sont possibles que sous l'hypothèse de normalité des résidus.*

1.5.2 Estimation de θ dans le cas général

on estime θ par la méthode des moindres carrés. Elle consiste à poser :

$$\hat{\theta} = Arg \;\; min\|y - X\theta\|^2 \quad , \quad \theta \in \Re$$

(Cette écriture suppose que $\Re^n$ est muni de la norme euclidienne classique, autrement dit que l'on utilise le critére dit des moindres carrès ordinaires.) . On montre alors que ce probléme admet la solution unique :

$$\hat{\theta} = (X^t X)^{-1} X^t y \quad \text{(estimation)}$$

valeur observèe du vecteur aléatoire :

$$\hat{\theta} = (X^t X)^{-1} X^t Y \quad \text{(estimateur)}$$

Propriètès de $\hat{\theta}$:
$E(\hat{\theta}) = (X^t X)^{-1} X^t E(Y) = (X^t X)^{-1} X^t X\theta = \theta$; $\hat{\theta}$ est un estimateur sans biais de θ
$Var(\hat{\theta}) = \sigma^2 (X^t X)^{-1} X^t X (X^t X)^{-1} = \sigma^2 (X^t X)^{-1} = \dfrac{\sigma^2}{n} S_n^{-1}$ avec $S_n = \frac{1}{n}(X^t X)^{-1}$
(matrice des variances-covariances empiriques lorsque les variables X_i sont centrées). On obtient un estimateur convergent, sous réserve que :

$$\lim_{n \mapsto \infty} \det S_n = d \succ 0$$

1.5.3 l'intervalle de confiance de θ :

Selon les propriétés de $\hat{\theta}$,on écrit que :$\hat{\theta} \sim N(\theta, \sigma^2((X^tX)^{-1}))$ alors la v a $\frac{\hat{\theta}-\theta}{\sqrt{\sigma^2((X^tX)^{-1})}}$ est distribuée selon une loi $N(0,1)$ et la v a $\frac{(n-k)\hat{\sigma^2}}{\sigma^2}$ est distribuée selon une loi ψ^2_{n-k} . Ces deux v.a. étant indépendantes, que est distribuée selon une loi Student $(n-k)$ Si on note $t_{(1-\frac{\alpha}{2})}$ est le $(1-\frac{\alpha}{2})$ quantile de la distribution de Student$(n-k)$, l'intervalle de confiance de θ de sécurité $1-\alpha$ est défini par [2] :

$$I_c(\theta) = [\hat{\theta}(y) \mp t_{1-\alpha}\sqrt{\sigma^2((X^tX)^{-1})}]$$

Chapitre 2

Régression des modèles linéaire et des modèles non linéaire

La régression est un des méthodes les plus connues et les plus appliquées en statistique pour l'analyse de données quantitatives. Elle est utilisée pour établir une liaison entre une variable quantitative et une ou plusieurs autres variables quantitatives, sous la forme d'un modèle. Si on s'intéresse à la relation entre deux variables, on parlera de régression simple en exprimant une variable en fonction de l'autre. Si la relation porte entre une variable et plusieurs autres variables, on parlera de régression multiple. La mise en ùvre d'une régression impose l'existence d'une relation de cause à effet entre les variables prises en compte dans le modèle.[2]

La régression permet de détecter et de quantiter l'effet d'une variable indépendante X sur une variable dépendante Y . Dans le cadre d'une analyse de régression, on fait l'hypothése implicite que la variable indépendante, appelèe également variable prédictrice ou explicative, ou règresseur, est responsable d'une partie de la variation de la variable dépendante, mais qu'en revanche la variable dépendante n'affecte pas la variable indépen-

dante. Le modéle postulè dans le cadre d'une analyse de régression permet d'expliquer et de mesurer l'ifluence d'une variable quantitative sur une autre variable quantitative.[4]
La technique de régression linéaire est très utile et couramment employèe en sciences expérimentales dans la mesure ôu elle permet non seulement d'effectuer des tests d'hypothéses quant à l'effet d'une variable sur une autre, mais également de prédire les valeurs de la variable dèpendante, sous certaines conditions et avec une certaine tolérance, ce qui revient d'une certaine maniére à quantifier l'effet de la variable indépendante

2.1 Régression linéaire

Le terme régression est dû à Galton (1886) qui, de suite a observé l'influence de la taille des individus (personnes) sur leur poids.
La régression linéaire est une méthode statistique vraisemblablement la plus utilisée par les praticiens de toutes disciplines : la recherche d'une liaison entre deux ou plusieurs caractères est une démarche très courante en médecine. en psychologie. en physique. en économie ...etc
La régression linéaire (ou les modèles linéaires) est un outil statistique trés utilisé pour étudier la présence d'une relation entre une variable dépendante y (quantitative et continue) et une ou plusieurs variables indépendantes $x_1, x_2, ..., x_p$ (qualitatives et/ou quantitatives).[3]

2.1.1 Régression linéaire simple

La régression linéaire simple est une téchnique statistique qui permet d'expliquer et d'expliquer et exprimer une variable aléatoire y en fonction d'une auter variablealéatoire x et elle sert à prévoir les valeurs futures de y en fonction de x .
Pour décrire une relation linéaire entre deux variables quantitatives ou encore pour pouvoir prédire y pour une valeur donnée de x, nous utilisons une droite de régression :

$$y_i = ax_i + b + \varepsilon_i$$

Puisque tout modèle statistique n'est qu'une approximation (nous espérons la meilleure possible ! !), il y a toujours une erreur, notée ϵ dans le modèle, car le lien linéaire n'est jamais parfait.

S'il y avait une relation linéaire parfaite entre y et x, le terme d'erreur serait toujours égale à 0, et toute la variabilité de y serait expliquée par la variable indépendante x.

- ***Les hypothéses relatives à ce modèle***

$H_1 : E(\xi_i) = 0, \forall_i$

$H_2 : Var(\xi_i) = \sigma_i^2, Cov(\xi_i, \xi_j) = 0, \forall_i$

- ***Estimation des parametres***

l'estimation des paramméteres de régression s' écrit :

$$\widehat{y_i} = \widehat{b} + \widehat{a} \;\; x_i$$

on cherche de $\widehat{b}$ et $\widehat{a}$ dans cette equiation qui minimuse l' erreur quadratique moyenne $\sum_{i=1}^{n} \varepsilon_i^2$ cette méthode s'appelle Moinder carré ordinaire on écrit :

$$(\widehat{a}, \widehat{b}) = \min \sum_{i=1}^{n} \varepsilon_i^2$$

$\widehat{a}, \widehat{b}$ sont estimés par :

$$\widehat{a} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x} \;\; \overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \qquad , \qquad \widehat{b} = \overline{y} - \widehat{a}\overline{x}$$

- ***L'intervalle de confiance :***

saus l'hypothése que les résidus sont Gaussien (normale) c'est à dire $\varepsilon_i \sim N(0, \sigma^2)$ on a :

✓ $\dfrac{(n-2)S_\varepsilon^2}{\sigma_\varepsilon^2} \quad \sim \quad \psi_{n-2}^2$

✓ $\dfrac{\hat{a} - a}{\sqrt{var(\hat{a})}} \;\; \sim \;\; T_{n-1}$

ce ci permet de tester les hypothéses de millité d'un dexes paraméteres ainsi que de construire les intervalles de confiance de a et b au niveau de confionce $(1 - \alpha)$

donc l'intervalle de confionce de a est : $]\hat{a} - t_{n-2}^{\alpha\backslash 2}\sqrt{var(\hat{a})}; \hat{a} + t_{n-2}^{\alpha\backslash 2}\sqrt{var(\hat{a})}$ [avece $t_{n-2}^{\alpha\backslash 2}$: est fractile d' order $(1 - \alpha \setminus 2)$ pour la loi de student .

la méme choser avec l' intervalle de $\hat{b}$ on a :

$$(n-2)\frac{S_\varepsilon^2}{\sigma_\varepsilon^2} \quad \sim \quad \psi_{n-2}^2$$

- ***Test de signification du modéle :***

Analyse de la qualité du modèle pour évaluer la qualité d'ajustement du modèle on utilise le coefficient de détermination R^2 on définit l'équation d' analyse de variance suivant :

$$SCT = SCE + SCR$$

Óu :

$SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2$ (somme des carrées totale)

$SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ (somme des carrées expliquée)

$SCR = \sum_{i=1}^{n}(\hat{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\varepsilon_i^2$ (somme des carrés resoluelle)

donc la qualité du moéle est mesurée par une quantité appelé confficient de détermination D Óu :

$$D = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

cette qantité indique les proncentages des y expliquée par x.

Théorème 2.1.1 *La variance de régression peut également s'écrire*

$$S_{y^*}^2 = r^2 S_y^2$$

où r^2 est le coefficient de détermination.

Preuve.

$$\begin{aligned} S_{y^*}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{y} - \frac{S_{xy}}{S_x^2}(x_i - \bar{x}) - \bar{y})^2 \\ = \frac{S_{xy}^2}{S_x^2} \\ = \mathrm{r}^2 S_y^2 \end{aligned}$$

■

- ***Teste de validation du modéle :***

ce teste mesure l'apportglobal de x sur la détermination de y dons ce cos on veut testes l'hypothése :

$$\begin{cases} \mathrm{H}_0 : & a = 0 \\ \mathrm{H}_1 : & a \neq 0 \end{cases}$$

Les indicateurs de variabilité sont résumés dans le tableau d'analyse de la variance ci-dessous :

TABLE 2.1 – Tableau d'analyse de la variance de test de validation

source	Degrés de liberté	Somme des carrés	Somme des carrés moyens	Stat de Fisher
modèle	1	SSM	SSM	$F_{cal} = \dfrac{SSM}{S^2}$
erreur	$n-2$	SSR	$S^2 = \dfrac{SSR}{n-2}$	/
total	$n-1$	SST	$S^2 = \dfrac{SST}{n-1}$	/

On accept H_0 si :

$$F_{cal} \prec f^{1-\alpha}_{(1,n-2)}$$

óu : $f^{1-\alpha}_{(1,n-2)}$ est la fractile d'ordre $(1-\alpha)$ de la loi de ficher de $(1, n-2)$ d'egrée de liberté .

- **La prévision** :

Un des buts de la régression est de faire de la prévision, c'est-à-dire de prévoir la variable à expliquer y en présence d'une nouvelle valeur de la variable explicative x. Soit donc x_{n+1} une nouvelle valeur, pour laquelle nous voulons prédire y_{n+1}. Le modèle est toujours le même :

$$y_{n+1} = \theta_1 + \theta_2 x_{n+1} + \varepsilon_{n+1}$$

avec : $E(\varepsilon_{n+1}) = 0$ et $Var(\varepsilon_{n+1}) = 0$, $Cov(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, ..., n$ .Il est naturel de prédire la valeur correspondante via le modèle ajusté :

$$\hat{y}_{n+1} = \hat{\theta}_1 + \hat{\theta}_2 \; x_{n+1} + \varepsilon_{n+1}$$

Deux types d'erreurs vont entacher notre prévision : la première est due à la non-connaissance de ε_{n+1}, la seconde à l'incertitude sur les estimateurs $\hat{\theta}_1$ et $\hat{\theta}_2$.

Proposition 2.1.2 *L'erreur de prévision $\varepsilon_{n+1} = y_{n+1} - \hat{y}_{n+1}$ satisfait les propriétés suivantes :*

$$\begin{cases} E(\ _{n+1}\ ) = 0 \\ Var(\ _{n+1}) = \sigma^2(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}) \end{cases}$$

Preuve. Pour l'espérance, il suit d'utiliser le fait que $n+1$ est centrée et que les estimateurs $\hat{\theta}_1$et $\hat{\theta}_2$ sont sans biais :

$$E(\hat{\varepsilon}_{n+1}) = E(\theta_1 - \hat{\theta}_1) + E(\theta_2 - \hat{\theta}_2) \; x_{n+1} + e(\varepsilon_{n+1}) = 0$$

Nous obtenons la variance de l'erreur de prévision en nous servant du fait que y_{n+1} est fonction de ε_{n+1} seulement tandis que $\hat{y}_{n+1}$ est fonction des autres erreurs (ε_i) pour $1 \prec i \prec n$

■

2.1.2 Régression linaire multiple

la régression linaire multiple est une généralisation de régresion linaire semple son but est d'étudier et modéliser la relation entre une variable expliquée (y) et plusieurs variables explicatives $(x_1, x_2, ..., x_n)$

le modéle de la RLM en s'ecrit sous la forme : $y_i = b_0 + b_1 x_{i1} + ... + b_p x_{ip} + \varepsilon_i$ donc la forme matricielle comme suite :

$$y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12}...x_{1n} \\ 1 & x_{21} & x_{22}...x_{2n} \\ & . \ . & ..... \\ 1 & x_{n1} & x_{n2}...x_{nn} \end{pmatrix} . \begin{pmatrix} b_0 \\ b_1 \\ ... \\ b_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{pmatrix}$$

avec : $Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$ et $X = \begin{pmatrix} 1 & x_{11} & x_{12}...x_{1n} \\ 1 & x_{21} & x_{22}...x_{2n} \\ & . & . \ ..... \\ 1 & x_{n1} & x_{n2}...x_{nn} \end{pmatrix}$ et $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$ La matrice X est connue : c'est la matrice des variables explicatives ; Y est observé : c'est le vecteur des données correspondant à la variable à expliquer ; mais θ_i est inconnu (il est même à estimer et à tester) : c'est le vecteur des coeficients de la relation linéaire. De son côté, le vecteur ε des résidus n'est pas observé.[3]

- ***Estimation des parametres***

en utilisant la méthode de MCO on calcule la statistique $\hat{\theta} = (\hat{b_0}, \hat{b_1}, ..., \hat{b_n})^t$ pour minimise la quantite $\varepsilon^t\varepsilon$ danc :

$$\hat{\theta} = (X^tX)^{-1}X^tY.$$

propriétés de les estimateurs :

⋆ $E(\hat{\theta}) = \theta$ (estimateur sans biais)

⋆ $var(\hat{\theta}) = \sigma^2(X^tX)^{-1}$

⋆ $cov(\hat{\theta}, \varepsilon) = 0$

preuve :

•− $\mathrm{E}(\theta) = \theta$ on a :

$$\hat{\theta} = (X^tX)^{-1}X^tY.$$

et : $Y = X\ \theta + \varepsilon$

donc :

$$\begin{aligned} \hat{\theta} &= (X^tX)^{-1}X^t(X\ \theta + \varepsilon). \\ &= (X^tX)^{-1}X^tX\ \theta + (X^tX)^{-1}X^t\ \ \varepsilon \\ &= \theta + (X^tX)^{-1}X^t\ \ \varepsilon \end{aligned}$$

alors :

$$\begin{aligned} E(\hat{\theta}) &= E(\theta) + E((X^tX)^{-1}X^t \ \ \varepsilon) \\ &= \theta + E((X^tX)^{-1}X^t \ \ \varepsilon) \\ &= \theta + E((X^tX)^{-1}X^t)E(\varepsilon) \end{aligned}$$

donc :

$$E(\hat{\theta}) = \theta$$

•− $\quad \mathrm{Var}(\theta) = \sigma_\varepsilon^2(X^tX)^{-1}$

on à : $\hat{\theta} = \theta + (X^tX)^{-1}X^t \ \ \varepsilon$

alors :

$$\begin{aligned} var(\hat{\theta}) &= var(\theta) + var((X^tX)^{-1}X^t \ \ \varepsilon) \\ &= (X^tX)^{-1}X^t((X^tX)^{-1}X^t)^t var(\ \ \varepsilon) \\ &= (X^tX)^{-1}\sigma^2 \end{aligned}$$

ou : $\quad \sigma^2 = \dfrac{1}{n-(p+1)}\varepsilon^t\varepsilon$

- ***L'intervalle de confiance***

Selon les propriétés de $\hat{\theta}$ on à écrit que : $\hat{\theta} \ \ \sim N(\theta, \sigma^2(X^tX)^{-1})$ le variable aléatoire $\dfrac{\hat{\theta}-\theta}{\sqrt{\sigma^2(X^tX)^{-1}}}$ est distribuée selon une loi $N(0,1)$ est v.a $\dfrac{(n-k)\hat{\sigma}^2}{\sigma^2}$ distribuée selon une loi ψ^2_{n-k}

Si on note $t_{(1-\frac{\alpha}{2})}$ est le $(1-\frac{\alpha}{2})$ -quantile de la distribution de Student $(n-k)$, l'intervalle de confiance de j de sécurité $1-\alpha$ est défini par :

$$I_c(\theta) = [\ \ \hat{\theta} \pm \sqrt{\sigma^2(X^tX)^{-1}} \ \ ]$$

- ***Tests du modéles***

Test de signification du modéle soit l'hypothése :

$$\begin{cases} \mathrm{H}_0 : \quad b_j = b_{j_0} \\ H_1 : \quad b_j = b_{j_0} \end{cases}$$

au niveau de confiance $(1-\alpha)$ on accepte H_0 si :

$$T_{cal} = \frac{|\hat{b} - b|}{\sqrt{var}} \prec t^{-\frac{\alpha}{2}}_{n-(p+1)}$$

Test de la validation du modéle on a l' equiation d'analyse de varionce est donneé par :

$$SCT = SCE + SCR$$

avec :

1° $SCT = Y^tY - n\bar{Y}^2$

2° $SCE = \hat{Y}^tY - n\bar{Y}^2$

3° $SCR = Y^tY - \hat{\theta}^tX^tY$

Pour une qualité d'ajustement du modéle on utilise le cofficient de détermination $D \in [0,1]$ alors : $D = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$

Les indicateurs de variabilité sont résumés dans le tableau d'analyse de la variance ci-dessous :

TABLE 2.2 – Tableau d'analyse de la variance de les indicateurs de variabilité

source	Degrés de liberté	Somme des carrés	Somme des carrés moyens	Stat de Fisher
modèle	p	SCM	$CME = \frac{SCE}{p}$	$F_{cal} = \frac{CME}{CMR}$
erreur	$n-(p+1)$	SCR	$S^2 = CMR = \frac{SSR}{n-(p+1)}$	/
total	$n-1$	SCT	$S^2 = \frac{SCT}{n-1}$	/

On accept H_0 si :

$$F_{cal} \prec f^{1-\alpha}_{(p,n-(p+1))}$$

la prévision

le modéle est ecrit :

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i1} + ... + \hat{b}_p x_{ip}$$

donc la prévision est :

$$\hat{y}_p = \hat{b}_0 + \hat{b}_1 x_{p1} + ... + \hat{b}_p x_{pp}$$

où : $(x_{p1} + x_{p2} + ... + x_{pp})^t = X_p^t$ est une vecteur donné on a :
$Var(\hat{y}_p) = S_\varepsilon^2\{X_p^t(X^tX)^{-1}X_p + 1\}$ donc l'intervalle de conffionce de y_p est donné par :

$$y_p \in]\hat{y}_p - t^{1-\frac{\alpha}{2}}_{(n-(p+1))}\sqrt{Var(\hat{y}_p)}; \hat{y}_p + t^{1-\frac{\alpha}{2}}_{(n-(p+1))}\sqrt{Var(\hat{y}_p)}[$$

2.1.3 Analyse de la variance (ANOVA)

Analyse de la variance d'un seul facteur

on dispose dans ce cas une variable quantitative à expliquer et un seul facteur explictif .On note :
A :le facteur et $(A_1, A_2, ..., A_p)$ ses modalités (types , niveau),chaque niveau A_j est contient n_j mesure de y alors $y = (y_{1j}, y_{2j}, ..., y_{n_j j})$.
C'est le cas le plus simple rencontré lorsqu'il n'y a qu'un facteur agissant sur les résultats. Le schéma est alors le suivant :

TABLE 2.3 – analyse de la variance d'un seul facteur

facteur de A	facteur de A_1	.....	facteur de A_j	.....	facteur de A_p
les observation	y_{11}	.....	y_{1j}	.....	y_{1p}
:	:	:	:	.....	:
	$y_{n_i 1}$	.....	$y_{n_i j}$	.....	$y_{n_i p}$
rolation	$\bar{y_{.1}}$	.....	$\bar{y_{.j}}$	.....	$\bar{y_{.p}}$

avec :

y_{ij} : est la i^{me} observation du j^{me} niveau (type).

n_j : est la taille de l'echantillon dans la j^{me} niveau .

$\bar{y_{.j}}$: est la moyenne de j^{me} niveau.

où :

$$\bar{y_{.j}} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \quad ; \quad j = 1, ..., p$$

n :est la taille totale où :

$$n = \sum_{j=1}^{p} n_j$$

$\bar{y_{..}}$: est la moyenne totale où :

$$\bar{y_{..}} = \frac{1}{n} \sum_{j=1}^{p} n_j \bar{y_{.j}}$$

Modéle d'ANOVA(1) le modéle s'ecrit théoriquement :

$$y_{ij} - y_{..} = (y_{ij} - y_{.j}) + (y_{.j} - y_{..})$$

où :

$y_{..} = \frac{1}{n} \sum_{j=1}^{p} n_j y_{.j}$

en tempe l'observation on a :

$$y_{ij} - \bar{y_{..}} = (y_{ij} - \bar{y_{.j}}) + (y_{.j} + \bar{y_{..}})$$

avec :

$y_{ij} - \bar{y_{..}}$: est une ecart Total (SCT).

$y_{ij} - \bar{y_{.j}}$: est une ecart Résidale (SCR).

$y_{.j} + \bar{y_{..}}$: est une ecart factoriel (SCA) .

Tableau dANOVA (1)

TABLE 2.4 – Tableau dANOVA (1)

source	Degrés de liberté	SDM	Somme des c m	Stat de Fisher
fact A	$p-1$	SCA	$CMA = \frac{SCA}{p-1}$	$F_{cal} = \dfrac{CMA}{CMR}$
Résidale	$n-p$	SCR	$S^2 = CMR = \dfrac{SCR}{n-p}$	/
total	$n-1$	SCT	$S^2 = CMT = \dfrac{SCT}{n-1}$	/

Test d'egalité des moyennes on veut tester d'effet de la facteur A sur la variable y tell que :

$$\begin{cases} \text{H}_0 : & \mu_1 = \mu_2 = ... = \mu_p \\ H_1 : & \exists \ \ \mu_i \neq \mu_j \ \ ; \ \ i \neq j \end{cases}$$

on accepte H_0 si :

$$F_{cal} = \frac{CMA}{CMR} \prec f^{1-\alpha}_{(p-1,n-p)}$$

le rejette H_0 implique dans ce cas on utilise les teste de comparaison de deux couples $(\mu_k; \mu_l)$.alors :

$$\begin{cases} \text{H}_0 : & \mu_k = \mu_l \\ H_1 : & \exists \ \ \mu_l \neq \mu_k \end{cases}$$

on a donc le test est une test de comparaisons de stident à faire :

$$T_{cal} = \frac{|\bar{y_k} - \bar{y_{.l}}|}{\sqrt{(\frac{1}{n_k} + \frac{1}{n_l})CMR}}$$

on accepte H_0 si :

$$T_{cal} \prec t^{1-\frac{\alpha}{2c_p^2}}_{(n-p)}$$

Analyse de la variance de deux facteur

dans ce cas les données sont regroupées selon deux facteurs A, B ,notons $A_1, A_2, ..., A_p$ les types du facteur (A).et $B_1, B_2, ..., B_q$ les types du facteur B .

avec :

TABLE 2.5 – tableau d'analyse de la variance de deux facteur

	B_1	.....	B_j	....	B_q	moy
A_1	$y_{111}, y_{112}, ..., y_{11r}$	....	$y_{1j1}, y_{1j2}, ..., y_{1jr}$	...	$y_{1q1}, y_{1qr}, ..., y_{1qr}$	$\bar{y_{1..}}$
...	...	...	...	...	...	...
A_i	$y_{i11}, y_{i12}, ..., y_{i1r}$	...	$y_{ij1}, y_{ij2}, ..., y_{ijr}$	...	$y_{iq1}, y_{iq2}, ..., y_{iqr}$	$\bar{y_{i..}}$
...	...	...	....	...	...	..
A_p	$y_{p11}, y_{p12}, ..., y_{p1r}$	...	$y_{pj1}, y_{pj2}, ..., y_{pjr}$	...	$y_{pq1}, y_{pq2}, ..., y_{pqr}$	$\bar{y_{p..}}$
moy	$\bar{y_{1..}}$	...	$\bar{y_{.j.}}$	...	$\bar{y_{.q.}}$	$\bar{y_{...}}$

$$\bar{y_{ij.}} = \frac{1}{r}\sum_{k=1}^{r} y_{ijk}$$
$$\bar{y_{i..}} = \frac{1}{qr}\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk}$$
$$\bar{y_{.j.}} = \frac{1}{pr}\sum_{i=1}^{p}\sum_{k=1}^{r} y_{ijk}$$
$$\bar{y_{...}} = \frac{1}{pqr}\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} y_{ijk}$$

2.2 MODÈLES NON LINÉAIRES, MAIS LINÉARISABLES

Il existe de très nombreux exemples de modèles définis par une combinaison non linéaire des paramètres, combinaison qui peut néanmoins être linéarisée. Quelques fonctions classiques sont rassemblées dans le tableau ci-dessous, qui ne donne qu'un très petit aperçu de la variété des situations rencontrées en pratique .

TABLE 2.6 – tableau de transformation des fonction à un modèle linéaire

Fonction	Transformation	Modèle linéarisé
$y = \alpha_0 \exp \alpha_1 x$	$Y = \ln y$	$Y = y + \alpha_1 x$,$y = \ln \alpha_0$
$y = \alpha_0 + \alpha_1 x$	$X = \ln x$	$Y = \alpha_0 + \alpha_1 X$
$y = \frac{x}{(-\alpha_1 + \alpha_0 x)}$	$X = \frac{1}{x}, Y = \frac{1}{y}$	$Y = \alpha_0 - \alpha_1 X$
$y = \frac{\exp(\alpha_0 + \alpha_1 x)}{1 + (\alpha_0 + \alpha_1 x)}$	$Y = \ln(\frac{y}{1-y})$	$Y = \alpha_0 + \alpha_1 X$

2.3 RÉGRESSION NON LINÉAIRE

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.

Définition 2.3.1 *Nous allons partir de p paramètres θ_k estimés et utiliser une méthode d'itération . Sur chaque itération les poids seront considérés constants et la fonction sera linéarisée pour chacun des n points sur l'ensemble des paramètres. La fonction dépend des x_i et des paramètres θ_k.*
Le modèle de régression linéaire s'écrit :

$$y_i = f(x_i, \theta_i) + \varepsilon_i \qquad ; i = 1...k$$

-La loi de probabilité sur ε_i est une loi normale, centrée réduite et de variance σ^2 finie.
-Les ε_i sont indépendants entre eux.
-y_i représente l'observation i de la variable dépendante
-θ représente un vecteur à p composantes de paramètres généralement inconnus.
-La fonction f est la fonction de régression, la plupart du temps non linéaire. Elle dépend d'une variable réelle x et de paramètres θ.[8]

2.3.1 Estimation des paramétres

Comme en régression linéaire, les paramètres d'un modèle de régression non linéaire sont estimés en minimisant la somme des carrés des résidus du modèle. C'est-à-dire qu'on cherche à minimiser l'expression suivante :

$$SSR(\theta) = \sum_{i=1} (y_i - f(x_i, \theta))^2$$

Pour cela, il faut dériver cette somme par rapport à chacun de ses paramètres et chercher les solutions qui annulent les dérivés.
On peut réécrire ceci sous forme vectorielle :

$$SSR(\theta) = (y - f(x, \theta))(y - f(x, \theta))'$$
$$SSR(\theta) = yy' - 2yf(x, \theta)' + f(x, \theta)'f(x, \theta)$$

On dérive cette expression par rapport à toutes les composantes du vecteur θ à p paramètres, et on annule toutes les dérivées partielles.
On obtient les conditions du premier ordre qui doivent être vérifiées pour toute estimation du vecteur θ qui correspond à un minimum intérieur de $SSR(\theta)$. Ces conditions du premier ordre, ou équations normales, sont :

$$-2yF(x,\hat{\theta})' + 2F(x,\hat{\theta})'f(x,\theta) = 0 \qquad (I)$$

Où la matrice $F(x,\theta)$ de dimension $n \times p$ est composée d'éléments du type :

$$F_i(x,\theta) = \frac{\partial f_i(x,\theta)}{\partial \theta_i}$$

Le fait que chaque vecteur de l'équation (I) possède p éléments implique l'existence de p équations normales déterminant les p composants de θ. Finalement, on obtient des équations qu'on ne peut la plupart du temps pas résoudre de manière analytique. Cependant, il existe des algorithmes qui permettent d'estimer les paramètres. Nous intéresserons ici à l'algorithme de Gauss-Newton, qui est utilisé par défaut par la fonction de R . [8]

2.3.2 Méthode de Gauss-Newton

Dans les problèmes de moindres carrés non linéaires, la fonction à minimiser prend en général la forme :

$$f(x) = \tfrac{1}{2}\sum_{i=1}^{n} f(x)_i^2$$

Pour appliquer la méthode de Newton à la minimisation de $f(x)$, on doit calculer le Hessien de f, qui dans ce cas précis prend une forme particulière. D'une part, la gradient de f est :

$$\nabla f(x)_i = \sum_{i=1}^{n} \nabla f(x)_i f(x)_i$$

et le Hessien de g est donné par :

$$\nabla^2 f(x)_i = \sum_{i=1}^{n} \nabla f(x)_i \nabla f(x)_i^t + \sum_{i=1}^{n} f(x)_i \nabla^2 f(x)_i$$

donc La matrice obtenue ;

$$H(x) = \sum_{i=1}^{n} \nabla f(x)_i \nabla f(x)_i^t$$

est semi-définie positive et la plupart du temps, avec $m \succ n$, elle est définie positive. La méthode obtenue de la méthode de Newton en ramplacant $\nabla f(x)_i$ par $H(x)$ est la méthode de Gauss-Newton :

$$\begin{cases} \mathrm{x}_0 \text{ et donné} \\ \mathrm{H}_k = \sum_{i=1}^{n} \nabla f(x)_i \nabla f(x)_i^t \\ \mathrm{x}_{k+1} = x_k + H_k^t \nabla f(x)_k \end{cases}$$

Chapitre 3

Application sur R :

3.1 Application 01 : Régression linéaire

on veut étudier la relation entre la fréquence cardiaque maximun Y et l'age X chez des coureurs en la tableau suivant contient les valeurs de deux variables chez de 30 hommes (je pris les résultats dans centre sportif du mandat de l'adrar) :

Table 3.1 – Tableau des donné de la fréquence cardiaque maximun et l'age de 30 hommes

X_i	40	32	85	38	45	55	55	44	49	50	32	25	58	45	30
y_i	187	180	190	181	194	190	184	180	195	189	182	202	191	195	188
X_i	60	55	47	44	33	30	47	50	44	28	60	60	48	36	41
y_i	188	195	180	182	185	181	194	195	192	187	188	185	189	189	190

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

H_0 : il y à une rolation dans la fréquence et l'age de le homme

H_1 : il n y pas une rolation dans la fréquence et l'age de le homme

- ***calcule les estimateur*** $(\hat{a}, \hat{b})$ ***:***

on calculer $X_i^2, Y_i^2, XY, \bar{X}, \bar{Y}$:

$\sum_{i=1}^{30} X_i = 1366$

$\sum_{i=1}^{30} Y_i = 5648$

$\sum_{i=1}^{30} X_i Y_i = 257466$

$\sum_{i=1}^{30} X_i^2 = 66812$

$\sum_{i=1}^{30} Y_i^2 = 1064230$

Donc la moyenn des hommes est :

$\bar{X} = \frac{1}{n}\sum_{i=1}^{30} X_i \quad = \frac{1366}{30} = 45.53333$

$\bar{Y} = \frac{1}{n}\sum_{i=1}^{30} Y_i \quad = \frac{5648}{30} = 188.2667$

$S_x^2 = \frac{1}{n}(\sum_{i=1}^{30} X_i^2 - n\bar{X}^2) \quad = \frac{1}{30}(66812 - 30 \times (2073.284)) = 153.7822$

alors l'equation de régresion est :

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

et :

avec :

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

$= \frac{85.34912}{4613.466}$

alors :

$$\hat{a} = 0.06367$$

et :

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

$$= 188.2667 - (0.06367) \times 45.53333$$

$$\hat{b} = 185.36762$$

donc l'equiation de régresion est :

$$\hat{Y_i} = 185.36762 + 0.06367X_i$$

- ***calcule de résidus*** (ε_i) ***:***

en calcul résidus (ε_i) an a la résidus calcule par la relation suivant :

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

donc pour calculer le taux d'erreur comprend le tableau suivant :

TABLE 3.2 – Tableau des donné de la résius

$\hat{y}_i$	187.9	187.4	190.7	187.7	188.2	188.8	188.8	188.1	188.4	188.5
ε_i	-0.9	-7.4	-0.7	-6.7	5.7	1.1	-4.8	-8.1	6.5	0.4
ε_i^2	0.8	54.8	0.6	46.1	33.2	1.2	23.7	66.7	42.4	0.2
$\hat{y}_i$	187.4	186.9	189.1	188.2	187.2	189.1	188.8	188.3	188.1	187.4
ε_i	-5.4	15.1	1.9	6.7	0.7	-1.1	6.1	-8.3	-6.1	-2.4
ε_i^2	29.2	3.7	45.7	0.52	1.41	37.5	69.8	38.2	6.09	39.4
$\hat{y}_i$	187.2	188.3	188.5	188.1	187.1	189.1	189.1	188.4	187.6	187.9
ε_i	-6.2	5.6	6.4	3.8	-0.1	-1.1	-4.18	0.57	1.34	2.02
ε_i^2	31.8	41.5	14.6	0.02	1.4	17.5	17.53	0.33	1.79	4.08

alors la somme de résidu ou carré est :

$\sum_{i=1}^{n} \varepsilon_i^2 = 881.1651$

on a :

$$\varsigma_\varepsilon^2 = \frac{1}{n-2} \sum_{i=1}^{n} \varepsilon_i^2$$

alors : $\varsigma_\varepsilon^2 = \frac{1}{30-2} 881.1651 = 31.47018$

- ***l'intervalle de confiance de*** $\hat{a}$ ***:***

l'intervalle de confiance de $\hat{a}$ sécurité par :

$$I_c(a) =]\ \hat{a} - t_{n-2}^{1-\frac{\alpha}{2}} \sqrt{var(\hat{a})}\ \ ;\ \ \hat{a} + t_{n-2}^{1-\frac{\alpha}{2}} \sqrt{var(\hat{a})}\ \ [$$

alors pour $\alpha = 0.05$ on à pour le table de student à degrés de liberté :
$t_{n-2}^{1-\frac{\alpha}{2}} = t_{28}^{0.05} = 2.0423$
alors $I_c(a) =]\ 0.06367 - 2.0423\sqrt{0.0004}\ ;\ \ 0.06367 + 2.0423\sqrt{0.0004}[=]0.06888551; 0.1084545[$
.
donc $\alpha \in I_c(a)$ alors on rejette H_0

- ***Teste de signification du modéle :***

$T_{cal} = 2.903533$ car :
$T_{cal} = \frac{\hat{a}}{\sqrt{var(\hat{a})}} = \frac{0.06367}{\sqrt{0.0004}} = 2.903533$
et $t_{n-2}^{1-\frac{\alpha}{2}} = 2.0423$ alors
$T_{cal} > t_{n-2}^{1-\frac{\alpha}{2}} = 2.0423$
alors on rejette H_0 donc il n y pas une rolation dans la fréquence et l'age de le homme

- ***Test de signification globale du modéle :***

on à la table d'ANOVA est :

TABLE 3.3 – Tableau d'analyse de la variance de test de validation de l'application président

source	D. de liberté	Somme des carrés	Somme des carrés moyens	Stat de Fisher
modèle	1	$SCM = 1207468$	$SSM = 1207468$	$F_{cal} = 144.989$
erreur	28	$SCR = 233172$	$S^2 = 8328$	/
total	29	$SCT = 1440640$	$S^2 = 49677.24$	/

et on à : $f_{(1,28)}^{1-\alpha} = 3.7$ donc $F_{cal} = 144.989 \succ f_{(1,28)}^{1-\alpha} = 3.73$ alors on rejette H_0.

3.2 APPLICATION 02 : RÉGRESION NON LINÉAIRE

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 ecrit par :

$$Y = bX^a$$

donc en liniérement de cet modéle 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 :

$$Y^{'} = a + b^{'} X^{'}$$

danc l'equation de régresion est :

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

$\bar{Y^{'}} = \hat{a} + \hat{b} \bar{X^{'}}$ 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 résilta donne le tableau suivant :

TABLE 3.4 – Tableau des donné du poide x et taille y de l'application président

	poids (X)	taille (Y)	$X' = \ln X$	$Y' = \ln Y$	$X'.Y'$
01	46	152	3.8286	5.0238	19.2346
02	78	158	4.3567	5.0626	22.05462
03	85	160	4.4426	5.0752	22.5472
04	85	162	4.4426	50876	22.6024
05	85	158	4.4426	5.0626	22.4913
06	85	159	4.4426	5.0689	22.5194
07	85	161	4.4426	5.0814	22.5749
08	95	165	4.5539	5.1059	23.2519
09	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 hypothéses relatives à ce modèle :***

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

H_1 : il n y pas une rolation dans la poids (X) et taille (Y)de le homme

- ***Teste de signification du modéle :***

$T_{cal} = 3.813533$ car :

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

et $t_{n-2}^{1-\frac{\alpha}{2}} = 2.0423$ alors

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

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

3.2.1 • l'application sous R :

On cherche à expliquer les variations de y par celles d'une fonction linéaire de x à partir de 30 observations de chacune des variables, i.e. à ajuster le modèle , on peut faire cet exemple laide du logiciel R, partir les étapes suivantes :

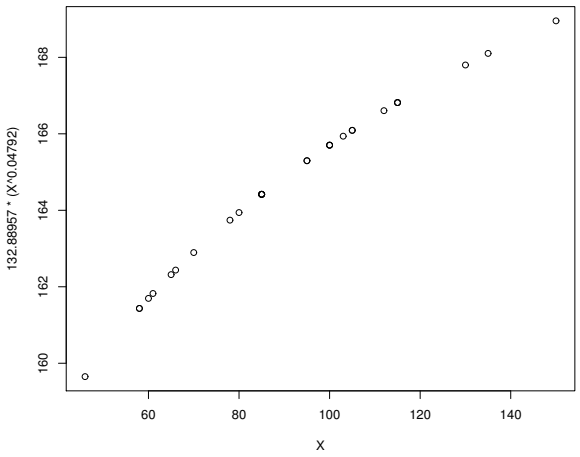

FIGURE 3.1 – nuage de pionts de modéle non linéaire de cet application

```
>x<-c(40,32,85,38,45,55,55,44,49,50,32,25,58,45,30,60,55,47,
44,33,30,47,50,44,28,60,60,48,36,41)
> x
 [1] 40 32 85 38 45 55 55 44 49 50 32 25 58 45 30 60 55 47 44 33 30 47
50 44 28
[26] 60 60 48 36 41
>y<-c(187,180,190,181,194,190,184,180,195,189,182,202, 191,195,
188,188,195,180,182,185,181,194,195,192,187,188,185,189,189,190)
> y
 [1] 187 180 190 181 194 190 184 180 195 189 182 202 191 195 188 188
195 180 182
[20] 185 181 194 195 192 187 188 185 189 189 190
> sum(x)
[1] 1366
> sum(y)
[1] 5648
> sum(x^2)
[1] 66812
> sum(y^2)
[1] 1064230
> sum(x*y)
[1] 257466
> mean(x)
[1] 45.53333
> mean(y)
[1] 188.2667

> regression=lm(y~x); regression

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
  185.36762      0.06367

>  plot(x,y)
> text(70,195, substitute(y=a*x+b, list(a=0.06367 , b=185.36762)))
```

FIGURE 3.2 – l'application sur R d'un modèle linéaire

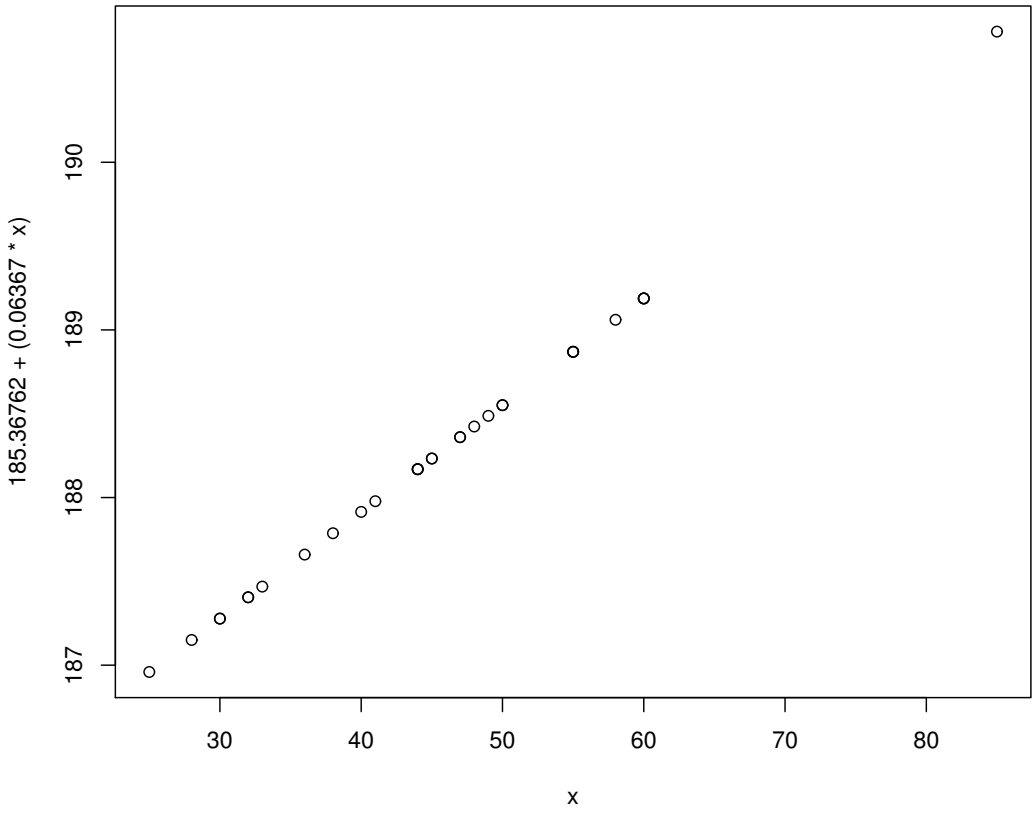

FIGURE 3.3 – nuage de point de la application d'un modèle linèaire

```
> regression.res<-residuals(regression)
> shapiro.test(regression.res)

        Shapiro-Wilk normality test

data:  regression.res
W = 0.9526, p-value = 0.1988

> par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
> plot(regression)
```

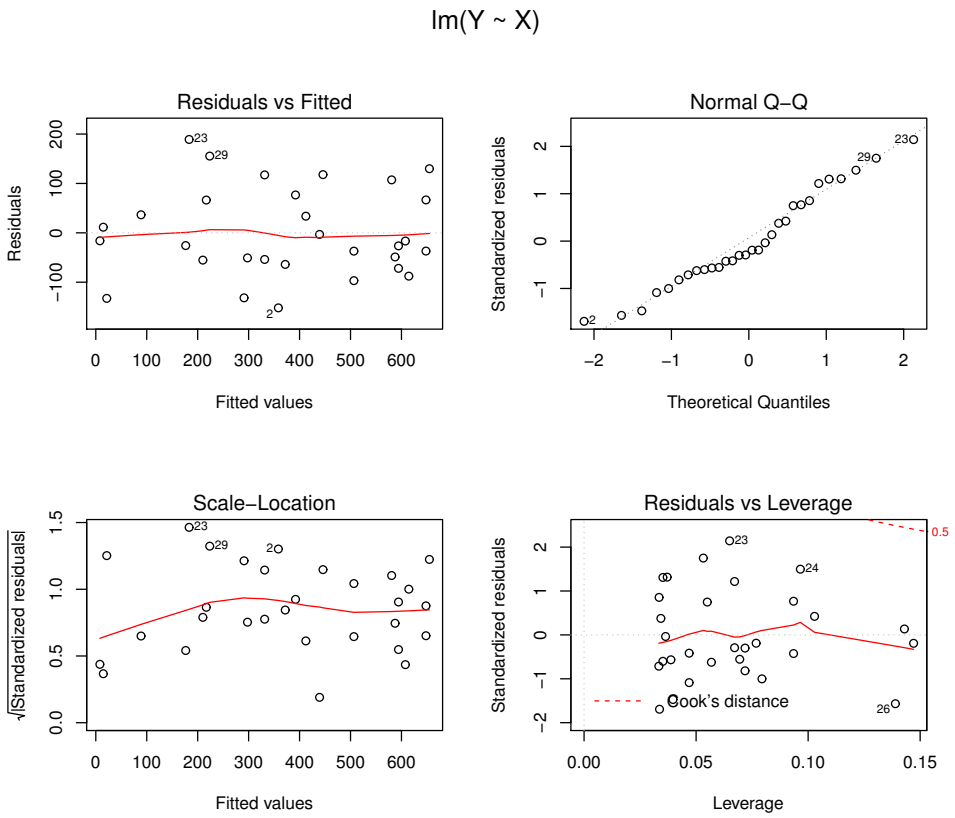

FIGURE 3.4 – Résultat de la fonction plot

```
> anova(regression)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X          1 1207468 1207468     145 1.373e-12 ***
Residuals 28  233172    8328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion

La régression linéaire et non linéaire est des modèles statistiques les plus employés : son champ d'application s'étend de la description et de l'analyse des données expérimentales jusqu'à la prévision, et il est aussi utilisé pour l'interpolation, ou pour l'aide à la mise en évidence de relations causales, par exemple. Il est par conséquent indispensable que le praticien possède une solide connaissance des prérequis, de la portée et des limites du modèle linéaire et non linéaire.

Pour cela dans ce mémoire on a traité les caractéristique de La régression linéaire et non linéaire (estimation , méthode de M.V ,intervale de confiance,analyse de la variance , la prévision...) les principales et ses propriétés que leurs les application de régression linéaire et non linéaire.Puis, on a présenté les méthodes d'estimation de La régression linéaire et non linéaire .

Bibliographie

[1] Benjamin Jourdain :Probabilitès et statistique ,septembre 2013

[2] C.Chouquet : Laboratoire de statistique et probabilités -2009-2010

[3] Christophe Lalanne ,S.Georges : Statistique appliquées à l'expérimentation

[4] Droesbeke(jj).Lejeune(M) :Société française de statistique . paris 2005

[5] D. Chessel J. Thioulouse :Fiche d'utilisation du logiciel R(Modèle linéaire généralisé)

[6] J.R. Lobry : Règression non linéaire avec le logiciel R.

[7] Jean .Pierre : Cours d'econométrée et régression

[8] Jean-Jacques Rousseau ,TP de Statistiques : Utilisation du logiciel R, Année 2006-2007

[9] Le Modéle Linéaire Gaussien Général ,Application aux plans factoriels, aux modèles mixtes et aux modèles pour données répètées (mars 2010)

[10] Mark.Asch : Régression linéaire et non linéaire 2010

[11] Mark Asch : TADE : régression non linéaire avec R ; Janvier 2011. Module TADE, EDSS, Université de Picardie Jules Verne.

[12] Ricco Rakotomala :Pratique de la Régression Logistique

[13] katell.Mellac : Méthodes d'annalyse de données en régression non linéaire

[14] G. COLLETAZ et C. HURLIN : Modèles Non Linéaires et Prévisions ,Novembre 2006

[15] Samuel Kotz, N. Balakrishnan :Encyclopedia of Statistical Sciences , Campell B.Read, Brani Vidakovics, Norman L.Johnson John Wiley Sons United Stated 2006

[16] Yves .Tillé : Résumé du Cours de Modéles de Régression , 2011

ملخص

هذا العمل مكرس لدراسة النماذج الإحصائية الخطية و الغير خطية، و بالتحديد نموذج الانحدار الخطي و الغير خطي، لأن دراسة هذه النماذج تمكننا من التقدير و التنبؤ بقيم مستقبلية انطلاقا من بيانات كاملة لإيجاد الحلول المناسبة في حالة وجود مشكلة ما .

لهذا سوف نستخدم نموذجين للتقدير، الانحدار الخطي و الغير خطي، ويتعلق النموذج الأول بالانحدار البسيط و المتعدد وكيفية التقدير باستعمالهما انطلاقا من بيانات كاملة، الثاني يتعلق بالتقدير بواسطة الانحدار الغير الخطي ،مع تطبيقات تجريبية لكل من النموذجين.

الكلمات المفتاحية : الانحدار، نسبة الخطأ، التقدير ، خطي و غير خطي ، مقدر كازي-نيوتن.

Résumé

Ce travail est consacré à l'étude des modèles statistiques linéaires et non-linéaires, la régression spécifiquement linéaire et modèle 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 et non linéaire, Le premier modèle est régression simple et le deuxième est multi-régression , 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: mise à niveau inférieur, erreur, estimation du rapport, linéaires et non linéaires, destinés Kazi-Newton.

Abstract

This work is devoted to the study of linear and nonlinear statistical models, specifically linear and nonlinear regression model, because the study of these models enables 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 models of estimation, linear and nonlinear regression. The first model relates to the simple and multiple regression and how to estimate using them from complete data. The second relates to estimation by nonlinear regression, with experimental applications for both model

Keywords : regression, error ratio, estimation, linear and nonlinear, Kazi-Newton, estimator.