RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
Ministère de L'enseignement Supérieur et de la Recherche Scientifique
UNIVERSITÉ IBN KHALDOUN TIARET
FACULTÉ DE MATHÉMATIQUES ET DE L'INFORMATIQUES
Département de Mathématiques

MÉMOIRE DE MASTER

Spécialité :
« Mathématiques »

Option :
«Analyse fonctionnelle et »
équations différentielles

Présenté Par :

MEKKI Maroua & BOUBAKOUR Nacira

Intitulé :

La fonction de régression et les méthodes les plus courantes de l'estimation

Soutenu publiquement le 15 / 06 / 2023
à Tiaret devant le jury composé de :

Mr. BENIA Kheïreddine	MCB	U. Ibn Khaldoun Tiaret	Président
Mr. BENALLOU Mohamed	MCB	U. Ibn Khaldoun Tiaret	Encadreur
Mr. REZZOUG Nadir	MCB	U. Ibn Khaldoun Tiaret	Examinateur

Année universitaire :2022/2023

Dédicace

je dédie ce modeste travail

A mes très chers parents pour leur amour inestimable, leur

confiance, leur soutien, leurs sacrifices et toutes les valeurs qu'ils

ont su m'accorder

A mes frères et sœurs

qui m'ont apporté leurs encouragement et leurs

soutiens.

A toutes mes amies

Maroua

Dédicace

je dédie ce modeste travail

A mes très chers parents pour leur amour inestimable, leur

confiance, leur soutien, leurs sacrifices et toutes les valeurs qu'ils

ont su m'accorder

A mes frères et sœurs

qui m'ont apporté leurs encouragement et leurs

soutiens.

A toutes mes amies

Nacira

Remerciements

Au nom de ALLAH Clément et Miséricordieux.

En premier lieu, nous remercions ALLAH, le tout puissant et miséricordieux, qui nous a donné la force, le courage et la patience d'accomplir ce modeste travail

Nos remerciements et nos profondes gratitudes vont à notre Encadreur Monsieur BENALLOU MOHAMED pour ses précieux conseils et pour tout le soutien et l'orientation. D'avoir bien voulu diriger notre travail, d'avoir donner le meilleur de son savoir, de son aide, et surtout d'avoir fait preuve de beaucoup de patience, son aide durant toute la période du travail.

Nous tenons aussi à remercier les membres du jury pour leur précieux temps accordé à l'étude de notre mémoire.

Nous remercions nos enseignants pour leurs efforts , nos parents et nos proches pour l'amour et le soutien contant qu'ils nous ont témoigné tout au long de notre parcours. Merci à toutes et touts nos amis pour leurs encouragements.

Résumé

L'objectif de ce travail est la présentation de la fonction de régression et les méthodes d'estimations avec une étude détaillée de chaque cas, à savoir, la régression linéaire (simple & multiple) avec l'estimation paramétrique basée sur la méthode de moindre carrée et l'autre la régression non linéaire et l'estimation non paramétrique et le plus connues, à noyau.

Table des matières

Table des figures

.

Introduction

La statistique est la science dont l'objet est de recueillir, de traiter et d'analyser des données issues de l'observation de phénomènes aléatoires, c'est-à-dire dans lesquels le hasard intervient. L'analyse des données est utilisée pour décrire les phénomènes étudiés, faire des prévisions et prendre des décisions à leur sujet. En cela, la statistique est un outil essentiel pour la compréhension et la gestion des phénomènes complexes. Les méthodes statistiques se répartissent en deux classes :

- La statistique descriptive, statistique exploratoire ou analyse des données, a pour but de résumer l'information contenue dans les données de façon synthétique et efficace. Elle utilise pour cela des représentations de données sous forme de graphiques, de tableaux et d'indicateurs numériques (par exemple des moyennes). Elle permet d'étudier et de suggérer des hypothèses pour une étude ultérieure plus sophistiquée. Les probabilités n'ont ici qu'un rôle mineur.
- La statistique inférentielle va au delà de la simple description des données. Elle a pour but de faire des prévisions et de prendre des décisions au vu des observations. En général, il faut pour cela proposer des modèles probabilistes du phénomène aléatoire étudié et savoir gérer les risques d'erreurs. Les probabilités jouent ici un rôle fondamental. L'informatique et la statistique sont deux éléments du traitement de l'information, l'informatique acquiert et traite l'information tandis que la statistique l'analyse. Les deux disciplines sont donc étroitement liées.

L'objectif de ce mémoire est de présenter une méthode la plus utilisée de la statistique inférentielle : la régression. Nous souhaitons aborder de manière simultanée les fondements théoriques et les questions inévitables que l'on se pose lorsque l'on modélise des phénomènes réels. En effet, comme pour toute méthode statistique, il est nécessaire de comprendre précisément la méthode et de savoir la mettre en œuvre. Si ces deux objectifs sont atteints, il sera alors aisé de transposer ces acquis à d'autres méthodes, moyennant un investissement modéré, tant théorique que pratique. Les grandes étapes - modélisation, estimation, choix de variables, examen de la validité du modèle choisi - restent les mêmes d'une méthode à l'autre. Nous étudierons chaque cas séparément, sont d'abord, la régression linéaire avec les deux branches simple et multiple ici on estime paramétriquement à cause du nombre fini de paramètres dans le modèle linéaire, on se limite parmis les méthodes paramétrique d'estimation à la méthode de moindre carrée et du maximum de vraisemblance, puis on passe au second cas, qui est la régression non linéaire, ici on présente les méthodes d'estimations non paramétrique et les plus connus parmis elles, c'est la méthode à noyau.

Notre travail est structuré comme suit :

Le premier chapitre présente les notions de base qui est l'estimation et ses différentes variantes, cette notion a une importance capitale en statistiques, en premier lieu on rappelé les définitions et les lois classiques et le principe de l'estimation, puis on décrit en détails deux types fameux d'estimation : l'estimation paramétrique et l'estimation non paramétrique.

Le chapitre suivant introduit la régression linéaire (simple & multiple) avec les méthodes d'estimation paramétrique concentrons sur la méthode de moindre carrée.

Le dernier chapitre est consacré à l'estimation non paramétrique de la régression non linéaire.

Chapitre 1

Le modèle de régression : présentation

Ce chapitre permet de reprendre certaines notions de base des variables aléatoires, et leurs propriétés, ensuite la définition de la regression et l'estimation.

1.1 généralités

1.1.1 Échantillon

En statistique, un échantillon est un ensemble d'individus extraits d'une population étudiée de manière à ce qu'il soit représentatif de cette population, au moins pour l'objet de l'étude. Pour ce faire, on peut le tirer de façon aléatoire, par un ensemble de méthodes mathématiquement très contraignantes, ou quand ces méthodes se révèlent impossibles à appliquer, par des méthodes pratiques comme la méthode des quotas.

- En traitement des signaux, on parle alors d'échantillonnage de signal.
- Chez les boulangers, l'échantillon accompagné de sa souche, était la baguette de bois qui permettait de comptabiliser le nombre de pains livrés au client mais non encore payés.

Nous allons voir que si une variable aléatoire suit une certaine loi, alors ses réalisations (sous forme d'échantillons) sont encadrées avec des probabilités de réalisation. Par exemple, lorsque l'on a une énorme urne avec une proportion p de boules blanches alors le nombre de boules blanches tirées sur un échan-tillon de taille n est parfaitement défini. En pratique, la fréquence observée varie autour de p avec des probabilités fortes autour de p et plus faibles lorsqu'on s'éloigne de p.

Nous allons chercher à faire l'inverse : **l'inférence statistique** consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon issu de cette population. Les caractéristiques de l'échantillon, une fois connues, reflètent avec une certaine marge d'erreur possible celles de la population.

1.1.2 Espace probabilisé

Une expérience est appellé "aléatoire" s'il est impossible de prévoir à l'avance son résultat et si elle est répétée dans des conditions identiques. On appelle ensemble associé à une expérience aléatoire l'énsemble fondamentale $\Omega = \{$tous résultat possibles de cette expérience$\}$. Un espace de probabilité ou espace probabilisé est la donnée d'une probabilité à tout événement, il permet la modélisation quantitative de l'expérience aléatoire étudiée. Formellement, c'est un triplet $(\Omega, F, \mathbb{P})$, F est un ensemble des événements ou tribu sur Ω, et $\mathbb{P}$ est une probabilité sur F.

1.1.3 Variable aléatoire

Une variable aléatoire est une fonction définie sur l'ensemble des éventualités, c'est-à-dire l'ensemble des résultats possibles d'une expérience aléatoire.

Une variable aléatoire est souvent à valeurs réelles (gain d'un joueur dans un jeu de hasard, durée de vie) et on parle alors de variable aléatoire réelle : $X : \Omega \to X(\Omega) \in \mathbb{R}$. La variable aléatoire peut aussi associer à chaque éventualité un vecteur de $\mathbb{R}^n$ ou $\mathbb{C}^n$, et on parle alors de vecteur aléatoire : $X : \Omega \to X(\Omega) \in \mathbb{R}^n$ ou $X : \Omega \to X(\Omega) \in \mathbb{C}^n$. La variable aléatoire peut encore associer à chaque éventualité une valeur qualitative (couleurs, Pile ou Face), ou même une fonction (p.e. une fonction de $C(\mathbb{R}_+, \mathbb{R}^d)$, et on parlera alors de processus stochastique.

Ce furent les jeux de hasard qui amenèrent à concevoir les variables aléatoires, en associant à une éventualité (résultat du lancer d'un dé, d'un tirage à pile ou face, d'une roulette, ...) un gain. Cette association éventualité-gain a donné lieu par la suite à la conception d'une fonction de portée plus générale. Le développement des variables aléatoires est associé à la théorie de la mesure.

Définition 1.1 *Soient $(\Omega, F, \mathbb{P})$ un espace probabilisé et $(E, \mathcal{E})$ un espace mesurable. On appelle variable aléatoire de Ω vers E, toute fonction mesurable X de Ω vers E.*

Cette condition de mesurabilité de X assure que l'image réciproque par X de tout élément B de la tribu $\mathcal{E}$ possède une probabilité et permet ainsi de définir, sur $(E, \mathcal{E})$, une mesure de probabilité, notée $\mathbb{P}_X$, par $\mathbb{P}_X(B) = \mathbb{P}(\mathbb{X}^{-1}(B)) = \mathbb{P}(X \in B)$.

La mesure $\mathbb{P}_X$ est l'image, par l'application X, de la probabilité $\mathbb{P}$ définie sur (Ω, F).

Définition 1.2 *La probabilité $\mathbb{P}_X$ est appelée loi de probabilité de la variable aléatoire X .*

Quelques variables aléatoires réelles

En guise d'introduction aux définitions concernant les variables aléatoires réelles, il semble intéressant de présenter brièvement une famille de variables très utilisées.

Outre la variable certaine qui prend une valeur donnée avec une probabilité égale à 1, la variable aléatoire réelle la plus simple est appelée **variable de Bernoulli**. Celle-ci peut

prendre deux états, qu'il est toujours possible de coder 1 et 0, avec les probabilités p et $1-p$. Une interprétation simple concerne un jeu de dé dans lequel on gagnerait 100 dinars en tirant le six ($p = 1/6$). Sur une séquence de parties, la moyenne des gains tend vers p lorsque le nombre de parties tend vers l'infini.

Si on considère qu'une partie est constituée par n tirages au lieu d'un seul, le total des gains est une réalisation d'une **variable binomiale** qui peut prendre toutes les valeurs entières de 0 à n. Cette variable a pour moyenne le produit np.

Variable aléatoire réelle discrète

Définition 1.3 *On dit qu'une v.a.r* ***discrète*** *si elle ne prend qu'un nombre fini ou infini dénombrable de valeurs, formellement*

$$X \in \{x_i, i \in K \subset \mathbb{N}\}$$

Ainsi le résultat d'un lancer de dé cubique est une variable aléatoire réelle discrète car elle ne peut prendre que 6 valeurs : $1, 2, 3, 4, 5, 6$. Le résultat de deux lancers de dés cubiques est une variable aléatoire discrète car elle ne peut prendre que 36 valeurs possibles : les couples $(1,1), (1,2), \ldots, (2,1), (2,2), \ldots, (6,5), (6,6)$.

De même, la variable aléatoire donnant le nombre minimal de lancers nécessaires pour obtenir un premier 6 avec un dé cubique est une variable aléatoire discrète car on peut obtenir le premier 6 au premier lancer ($X = 1$), au second ($X = 2$), au 20^e ($X = 20), \ldots$, au n^e $(X = n), \ldots$ L'ensemble des valeurs possibles pour X est donc infini et dénombrable.

Dans ce cas,la loi de la variable aléatoire X est la loi de probabilité sur l'ensemble des valeurs possibles de X qui affecte la probabilité $\mathbb{P}(X = x_i)$ au singleton $\{x_i\}$. En pratique, l'ensemble des valeurs que peut prendre X est $\mathbb{N}$ ou une partie de $\mathbb{N}$.

1. L'espérance mathématique (moment d'ordre 1) de la v.a.r discrète X, notée $\mathbb{E}(X)$ est définie par (si la série $\sum_{i \in K} x_i \mathbb{P}(x_i)$ est absolument converge ou lorsque K est fini) :

$$\mathbb{E}(X) = \sum_{i \in K} x_i \mathbb{P}(X = x_i).$$

2. Le nombre :

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

lorsqu'il existe, est appelé variance de X, et l'écart type de X est :

$$\sigma(X) = \sqrt{Var(X)}.$$

Variables aléatoires réelle continue

Définition 1.4 *Une v.a.r prend des valeurs sur un ensemble infini non dénombrable des points, est dit continues si elle existe une fonction f non négative, définie pour toute valeur x appartenant à $\mathbb{R}$ et vérifiant, pour toute partie A de $\mathbb{R}$, la propriété :*

$$\mathbb{P}(X \in A) = \int_A f(x)dx,$$

telle que : $\int_{\mathbb{R}} f(x)dx = 1$.

La fonction f est appelée la densité de probabilité de la variable aléatoire X.

1. L'espérance mathématique de la v.a.r continue X, définie sur l'espace probabilisé $(\Omega, F, \mathbb{P})$,est donnée par l'integrale, si elle converge :

$$\mathbb{E}(X) = \int_\Omega X d\mathbb{P} = \int_\Omega x d\mathbb{P} dx,$$

que l'on peut écrire, si f est la densité de probabilité de X

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f(x)dx.$$

2. La variance, ou carré de l'écart -type σ, est donnée par l'intégrale si cette intégrale et la densité f existent

$$\mathbb{E}[(X - \mathbb{E}(X))^2] = Var(X) = \sigma^2 = \int_\Omega [x - \mathbb{E}(X)]^2 f(x)dx.$$

1.1.4 Lois usuelles

Loi normale ou loi de Gauss

Une variable aléatoire réelle X suit une loi normale (ou loi gaussienne, loi de Laplace-Gauss) d'espérance μ et d'écart type σ (nombre strictement positif, car il s'agit de la racine carrée de la variance σ^2) si cette variable aléatoire réelle X admet pour densité de probabilité la fonction f définie, pour tout nombre réel x, par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

Une telle variable aléatoire est alors dite variable gaussienne.

Une loi normale sera notée de la manière suivante $\mathcal{N}(\mu, \sigma)$ car elle dépend de deux paramètres μ (la moyenne) et σ (l'écart-type). Ainsi si une variable aléatoire X suit $\mathcal{N}(\mu, \sigma)$ alors

$$\mathbb{E}(X) = \mu \quad \text{et } V(X) = \sigma^2.$$

Lorsque la moyenne μ vaut 0, et l'écart-type σ vaut 1, la loi sera notée $\mathcal{N}(0,1)$ et sera appelée loi normale standard ou centrée réduite. Sa fonction caractéristique vaut $e^{-t^2/2}$. Seule la loi $\mathcal{N}(0,1)$ est tabulée car les autres lois (c'est-à-dire avec d'autres paramètres) se déduise de celle-ci à l'aide du théorème suivant : Si Y suit $\mathcal{N}(\mu,\sigma)$ alors $Z = \frac{Y-\mu}{\sigma}$ suit $\mathcal{N}(0,1)$.

On note Φ la fonction de répartition de la loi normale centrée réduite :

$$\Phi(x) = \mathbb{P}(Z < x)$$

avec Z une variable aléatoire suivant $\mathcal{N}(0,1)$.

Propriétés et Exemples :

$\Phi(-x) = 1 - \Phi(x)$

$\Phi(0) = 0,5$; $\Phi(1,645) \approx 0,95$; $\Phi(1,960) \approx 0,9750$.

Remarque 1.1 *La somme de deux variables gaussiennes indépendantes est elle-même une variable gaussienne (stabilité) : Soient X et Y deux variables aléatoires indépendantes suivant respectivement les lois $\mathcal{N}(\mu_1,\sigma_1)$ et $\mathcal{N}(\mu_2,\sigma_2)$. Alors, la variable aléatoire $X+Y$ suit la loi normale $\mathcal{N}\left(\mu_1+\mu_2, \sqrt{\sigma_1^2+\sigma_2^2}\right)$.*

Loi du $\mathcal{X}^2$(khi-deux)

Définition 1.5 *Soit $Z_1, Z_2, \cdots, Z_v$ une suite de variables aléatoires indépendantes de même loi $\mathcal{N}(0,1)$. Alors la variable aléatoire $\sum_{i=1}^{v} Z_i^2$ suit une loi appelée loi du Khi-deux à v degrés de liberté, notée $\mathcal{X}^2(v)$.*

Proposition 1.1

1. *Sa fonction caractristique est $(1-2it)^{-v/2}$.*
2. *La densité de la loi du $\mathcal{X}^2(v)$ est $f_v(x) = \begin{cases} \frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1}e^{-x/2} & pour\ x>0 \\ 0 & sinon \end{cases}$,*
où Γ est la fonction Gamma d'Euler définie par $\Gamma(r) = \int_0^\infty x^{r-1}e^{-x}dx$.
3. *L'espérance de la loi du $\mathcal{X}^2(v)$ est égale au nombre v de degrés de liberté et sa variance est $2v$.*
4. *La somme de deux variables alatoires indpendantes suivant respectivement $\mathcal{X}^2(v_1)$ et $\mathcal{X}^2(v_2)$ suit aussi une loi du $\mathcal{X}^2$ avec v_1+v_2 degrés de liberté.*

Loi de Student

Définition 1.6 *Soient Z et Q deux variables aléatoires indépendantes telles que Z suit $\mathcal{N}(0,1)$ et Q suit $\mathcal{X}^2(v)$. Alors la variable aléatoire : $T = \frac{Z}{\sqrt{Q/v}}$ suit une loi appelée **loi de Student** à degrés de liberté, notée $St(v)$.*

Proposition 1.2
1. *La densité de la loi de Student à degrés de liberté v est : $f(x) = \frac{1}{\sqrt{\pi v}} \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \frac{1}{(1+x^2/v)^{\frac{1+v}{2}}}$.*
2. *L'espérance n'est pas définie pour $v = 1$ et vaut 0 si $v = 2$. Sa variance n'existe pas pour $v = 2$ et vaut $\frac{v}{v} - 2$ pour $v = 3$.*
3. *La loi de Student converge en loi vers la loi normale centrée réduite.*

Remarque 1.2 *pour $v = 1$, la loi de Student s'appelle loi de* ***Cauchy****, ou loi de* ***Lorentz.***

Loi de Fisher-Snedecor

Définition 1.7 *Soient Q_1 et Q_2 deux variables aléatoires indépendantes telles que Q_1 suit $\mathcal{X}^2(v_1)$ et Q_2 suit $\mathcal{X}^2(v_2)$ alors la variable aléatoire : $F = \frac{Q_1/v_1}{Q_2/v_2}$ suit une loi de* ***Fisher-Snedecor*** *à (v_1, v_2) degrés de liberté, notée $F(v_1, v_2)$.*

Proposition 1.3 *La densité de la loi $F(v_1, v_2)$ est*

$$f(x) = \frac{\Gamma\left(\frac{v_1+v_2}{2}\right)}{\Gamma(v_1/2)\Gamma(v_2/2)} \left(\frac{v_1}{v_2}\right)^{v_1/2} \frac{x^{v_1/2-1}}{\left(1+\frac{v_1}{v_1}x\right)^{\frac{v_1+v_2}{2}}} \text{ si } x > 0 \text{ (0 sinon)}$$

Son espérance n'existe que si $v_2 \geq 3$ et vaut $\frac{v_2}{v_2-2}$. Sa variance n'existe que si $v_2 \geq 5$ est vaut $\frac{2v_2^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}$.

Proposition 1.4
1. *Si X suit une loi de Fisher $F(v_1, v_2)$ alors $\frac{1}{F}$ suit une loi de Fisher $F(v_2, v_1)$*
2. *Si T suit une loi de Student à degrés de liberté v alors T^2 suit une loi de Fisher $F(1, v)$*

1.2 L'estimation et la covergence

L'estimation consiste à donner des valeurs approximatives aux paramètres d'une population à l'aide d'un échantillon de n observations issues de cette population. On peut se tromper sur la valeur exacte, mais on donne la "meilleure valeur" possible que l'on peut supposer. Alors on cherche à ce qu'un estimateur soit sans biais, convergent, efficace et robuste.

Exemple d'estimateurs

Si l'on cherche à évaluer la taille moyenne des enfants de 10 ans, on peut effectuer un sondage sur un échantillon de la population des enfants de 10 ans (par exemple en s'adressant à des écoles réparties dans plusieurs milieux différents). La taille moyenne calculée sur cet échantillon, appelée moyenne empirique, sera un estimateur de la taille moyenne des enfants de 10 ans.

Si l'on cherche à déterminer le pourcentage d'électeurs décidés à voter pour le candidat A, on peut effectuer un sondage sur un échantillon représentatif. Le pourcentage de votes favorables

à A dans l'échantillon est un estimateur du pourcentage d'électeurs décidés à voter pour A dans la population totale.

Si l'on cherche à évaluer la population totale de poissons dans un lac, on peut commencer par ramasser n poissons, les baguer pour pouvoir les identifier ultérieurement, les relâcher, les laisser se mélanger aux autres poissons. On tire alors un échantillon de poissons du lac, on calcule la proportion p de poissons bagués. La valeur n/p est un estimateur de la population totale de poissons dans le lac. S'il n'y a aucun poisson bagué dans l'échantillon, on procède à un autre tirage.

Un estimateur est très souvent une moyenne, une population totale, une proportion ou une variance.

Remarque 1.3 *Un estimateur ne doit évidemment jamais dépendre de θ , il ne dépend que des observations empiriques.*

L'estimation est habituellement divisée en deux composantes principales : l'estimations paramétrique et l'estimation non-paramétrique.

1.2.1 Estimation paramètrique

Un estimateur de T est une fonction $T_n : x \to T_n(x, X_1, \cdots, X_n)$ mesurable par rapport à l'observation $(X_1, \cdots, X_n)$. Si l'on sait à priori que T appartient à une famille paramètrique $\{T(x, \theta), \theta \in \Theta\}$ où $\Theta \in \mathbb{R}^d$ et $T(\cdot, \cdot)$ est une fonction continue, on parle alors d'estimation paramètrique, car estimer T revient à estimer le paramètre fini dimensionnel θ.

Au phénomène étudié, nous associons maintenant un modèle statistique P_θ qui dépend d'un paramètre θ. Pour se faire une idée de la valeur inconnue du paramètre θ, a partir des observation $(X_1, \cdots, X_n)$ qui sont i.i.d, on calcule ensuite une certaine valeur numérique, que l'on considérera comme valeur approchée de θ qu'on appellera une estimateur de θ.

Dans ce cas ou'il n'y a pas d'estimateur évident, on cherche un estimateur par la méthode de vraisemblance, ou par la méthode des moments,...et

1.2.2 Estimation non-paramètrique

Par opposition, en statistique non paramétrique, le modèle n'est pas décrit par un nombre fini de paramètres. Divers cas de figures peuvent se présenter, comme par exemple :

- On s'autorise toutes les distributions possibles, i.e. on ne fait aucune hypothèse sur la forme/nature/type de la distribution des variables aléatoires
- On travaille sur des espaces fonctionnels, de dimension infinie. Exemple : les densités continues sur $[0, 1]$, ou les densités monotones sur $\mathbb{R}$.
- Le nombre de paramètres du moèle n'est pas fixé et varie (augmente) avec le nombre d'observations.
- Le support de la distribution est discret et varie (augmente) avec le nombre d'observations.

L'avantage principale de l'estimation non-paramètrique à un ensemble fini d'observation est de ne pas nécessiter d'hypothèses à priori sur l'appartenence à une famille de lois connues. L'estimation ne concerne pas les paramètres permettant de sélectionner une loi, mais directement la fonction elle_même (d'où le terme non-paramètrique)

1.2.3 Qualité d'un estimateur

Un estimateur est une valeur $\widehat{\theta}$ calculée sur un échantillon tiré au hasard, la valeur $\widehat{\theta}$ est donc une variable aléatoire possédant une espérance $\mathbb{E}\left(\widehat{\theta}\right)$ et une variance $Var\left(\widehat{\theta}\right)$. On comprend alors que la valeur $\widehat{\theta}$ puisse fluctuer selon l'échantillon. Elle a de très faibles chances de coïncider exactement avec la valeur exacte θ qu'elle est censée représenter. L'objectif est donc de maîtriser l'erreur commise en prenant la valeur x pour la valeur X.

Biais

Une variable aléatoire fluctue autour de son espérance. On souhaite donc que l'espérance de $\widehat{\theta}$ soit égale à θ, soit qu'en "moyenne" l'estimateur ne se trompe pas.

Définition 1.8 $Biais\left(\widehat{\theta}\right) = \mathbb{E}\left(\widehat{\theta}\right) - \theta.$

Lorsque l'espérance de l'estimateur coincide avec la vraie valeur, l'estimateur est dit sans biais.

Exemple 1.1 *L'estimateur choisi précédemment sur la taille moyenne des enfants de 10 ans est un estimateur sans biais mais celui des poissons comporte un biais : le nombre de poissons estimé est en moyenne supérieur au nombre de poissons réels.*

Erreur quadratique moyenne

L'erreur quadratique moyenne est l'espérance du carré de l'erreur entre la vraie valeur et sa valeur estimée.

Définition 1.9 $MSE\left(\widehat{\theta}\right) = \mathbb{E}\left(\left(\widehat{\theta} - \theta\right)^2\right).$

Remarque 1.4 *L'erreur quadratique moyenne peut être écrite comme une somme de la variance et du carré du biais de l'estimateur*

$$MSE\left(\widehat{\theta}\right) = Var(\widehat{\theta}) + \left(\mathbb{E}\left(\widehat{\theta}\right) - \theta\right)^2$$

Définition 1.10 *L'erreur moyenne quadratique intégrée MISE :*

$$\begin{aligned} MISE(\hat{r}, r) &= \int MSE(r(x), \hat{r}(x))dx \\ &= \int \mathbb{B}iais(\hat{r}(x))^2 dx + \int \mathbb{V}ar(\hat{r}(x))dx \end{aligned} \tag{1.1}$$

Définition 1.11 *On dit qu'un estimateur $\hat{r}$ de r est ponctuellement consistant en moyenne quadratique si :*

$\lim_{n\to\infty} MSE(r(x), \hat{r}(x)) = 0$

Définition 1.12 *On dit qu'un estimateur $\hat{r}$ de r est uniformément consistant en moyenne quadratique si :*

$\lim_{n\to\infty} MISE(r(x), \hat{r}(x)) = 0$

Définition 1.13 *On dit qu'un estimateur $\hat{r}$ de r est asymptotiquement normal si :*

$\hat{r} \to \mathcal{N}(\mathbb{E}(\hat{r}), \mathbb{V}ar(\hat{r}))$ *en loi.*

1.2.4 Comparaison d'estimateurs

Définition 1.14 *On dit que l'estimateur $\widehat{\theta}_1$ domine l'estimateur $\widehat{\theta}_2$ si pour tout $\theta \in \Theta$, $MSE\left(\widehat{\theta}_1\right) \leq MSE\left(\widehat{\theta}_2\right)$.*

Définition 1.15 *On dit qu'un estimateur est admissible s'il n'existe aucune estimateur le dominant.*

Définition 1.16 *Soit $\widehat{\theta}_1$, $\widehat{\theta}_2$ deux estimateurs sans biais de θ, $\widehat{\theta}_1$ est dit plus efficace que $\widehat{\theta}_2$ si :*

$Var\ (\widehat{\theta}_1) < Var\ (\widehat{\theta}_2)$, $\forall \theta \in \Theta$.

1.2.5 Convergence

On souhaite aussi pouvoir, en augmentant la taille de l'échantillon, diminuer l'erreur commise en prenant $\widehat{\theta}_n$ à la place de $\widehat{\theta}$. Si c'est le cas, on dit que l'estimateur est convergent (on voit aussi consistant), c'est-à-dire qu'il converge vers sa vraie valeur. La définition précise en mathématique est la suivante :

Définition 1.17 *L'estimateur $\widehat{\theta}_n$ est convergent s'il* ***converge en probabilité*** *vers θ, soit :* $\lim\limits_{n\to\infty} \mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| > \varepsilon\right) = 0$, $\forall \varepsilon > 0$.

On l'interprète comme le fait que la probabilité de s'éloigner de la valeur à estimer de plus de ε tend vers 0 quand la taille de l'échantillon augmente. Cette définition est parfois écrite de manière inverse :

Définition 1.18 *L'estimateur $\widehat{\theta}_n$ est convergent s'il* ***converge en probabilité*** *vers θ, soit :* $\lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| \leq \varepsilon\right) = 1, \forall \varepsilon > 0.$

Il existe enfin un type de convergence plus forte, la convergence presque sûre, définie ainsi pour un estimateur :

Définition 1.19 *L'estimateur $\widehat{\theta}_n$ est fortement convergent s'il* ***converge presque sûrement*** *vers θ, soit :* $\mathbb{P}\left(\lim_{n\to\infty} \widehat{\theta}_n = \theta\right) = 1.$

Exemple 1.2 *La moyenne empirique est un estimateur convergent de l'espérance d'une variable aléatoire. La loi faible des grands nombres assure que la moyenne converge en probabilité vers l'espérance et la loi forte des grands nombres qu'elle converge presque sûrement.*

Lois des Grands Nombres

Ces lois décrivent le comportement asymptotique de la moyenne de l'échantillon. Elles sont de deux types : **lois faibles** mettant en jeu la convergence en probabilité et **lois fortes** relatives à la convergence presque sûre.

Théorème 1.1 *Si $(X_1, \cdots, X_n)$ est un échantillon d'une v.a.r X tel que $\mathbb{E} \mid X \mid < \infty$, alors*

$$\begin{array}{lll} \text{loi faible} & \overline{X}_n \xrightarrow{P} \mu & \text{quand } n \longrightarrow \infty \\ \text{loi forte} & \overline{X_n} \xrightarrow{p.s} \mu & \text{quand } n \longrightarrow \infty \end{array}$$

ou $\mu := \mathbb{E}(X), \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1.2.6 Taux de convergence

Efficience

La variable aléatoire fluctue autour de son espérance. Plus la variance $Var\left(\widehat{\theta}_n\right)$ est faible, moins les variations sont importantes. On cherche donc à ce que la variance soit la plus faible possible. C'est ce qu'on appelle l'efficience d'un estimateur.

Robustesse

Il arrive que lors d'un sondage, une valeur extrême et rare apparaisse (par exemple un enfant de 10 ans mesurant $1,80m$). On cherche à ce que ce genre de valeur ne change que de manière très faible la valeur de l'estimateur. On dit alors que l'estimateur est robuste.

Exemple 1.3 *En reprenant l'exemple de l'enfant, la moyenne n'est pas un estimateur robuste car ajouter l'enfant très grand modifiera beaucoup la valeur de l'estimateur. La médiane par contre n'est pas modifiée dans un tel cas.*

1.2.7 Quelque estimateur classique

Soit X une variable aléatoire de moyenne μ et d'écart-type σ

1. On prend en général comme estimateur de la moyenne μ la moyenne empirique $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ qui est un estimateur sans biais. Son estimation $\overline{x}$ est la moyenne observée dans une réalisation de l'échantillon.

2. $\widetilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ est un estimateur consistant de σ^2(mais biaisé).

3. $S^2 = \frac{n}{n-1}\widetilde{S}^2$est un estimateur sans biais et consistant de σ^2. Son estimateur est $s^2 = \frac{n}{n-1}\sigma_e^2$ où $\sigma_e$est l'ecart-type observé dans une réalisation de l'échantillon.

Remarque 1.5 *: Si la moyenne μ de X est connue, $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ est un meilleur estimateur de σ^2 que S^2.*

1.3 Présentation générale de la fonction de regression

1.3.1 Historique

L'origine du mot régression vient de Sir Francis Galton. En 1885, travaillant sur l'hérédité, il chercha à expliquer la taille des fils en fonction de celle des pères. Il constata que lorsque le père était plus grand que la moyenne,*taller than mediocrity* , son fils avait tendance à être plus petit que lui et,a contrario,que lorsque le père était plus petit que la moyenne,*shorter than mediocrity*,son fils avait tendance à être plus grand que lui. Ces résultats l'ont conduit à considérer sa théorie de *regression toward mediocrity*. Cependant l'analyse de causalité entre plusieurs variables est plus ancienne et remonte au milieu du $XVIII^e$ siècle. En 1757, **R. Boscovich**, né à Ragussa, l'actuelle Dubrovnik, proposa une méthode minimisant la somme des valeurs absolues entre un modèle de causalité et les observations. Ensuite **Legendre** dans son célèbre article de 1805, ≪ Nouvelles méthodes pour la détermination des orbites des comètes ≫,introduit la méthode d'estimation par moindres carrés des coefficients d'un modèle de causalité et donna le nom à la méthode. Parallèlement, **Gauss** publia en 1809 un

travail sur le mouvement des corps célestes qui contenait un développement de la méthode des moindres carrés, qu'il affirmait utiliser depuis1795 (**Birkes et Dodge**, 1993).

1.3.2 Le modèle

La régression est un ensemble de méthodes statistique très utilisés pour analyser la relation d'une variable par rapport à une ou plusieurs autres. On recourt à une estimation des paramètres inconnus du modèle de régression par un ajustement mathématique du modèle spécifié en fonction des données récoltées. En économétrie, elle est notamment utilisée à des fins de prévision économique. À partir d'un ensemble de valeurs expérimentales, qui peuvent être représentées par des points sur un graphique, on cherche à calculer la courbe qui reproduit le mieux les variations de la grandeur à étudier, c'est-à-dire celle qui s'ajuste "au mieux" au nuage de points. La régression est donc l'opération qui consiste à ajuster une droite (ou une autre courbe mathématique) "le plus près possible" d'un certain nombre de points observés.

Une des méthodes les plus employées pour obtenir un modèle estimé est celle des « moindres carrés ». Sous certaines hypothèses, les estimateurs obtenus par la méthode des moindres carrés sont les meilleurs estimateurs pouvant être obtenus parmi ceux linéaires et non biaisés. Disposant de l'estimation de cette courbe, on peut alors effectuer des interpolations, pour calculer l'ordonnée de points intermédiaires. D'autres méthodes ont été plus récemment développées pour éviter des problèmes causés par des données atypiques ou sortant du cadre des hypothèses classiques.

Les étapes généralement suivies pour obtenir une régression sont :

1. Cadrage du problème,
2. Sélection des variables pertinentes,
3. Récolte des données,
4. Spécification du modèle,
5. Sélection de la méthode d'ajustement,
6. Ajustement du modèle aux données,
7. Validation du modèle estimé,
8. Utilisation du modèle.

Le processus de construction du modèle de régression est souvent itératif, car lors de la validation ou de l'utilisation du modèle il n'est pas rare de devoir revenir sur les étapes précédentes selon les résultats obtenus.

Les tableurs permettent, pour affiner les corrélations, d'utiliser un modèle linéaire comme la régression linéaire — une droite d'équation $Y = a * X + b$ — mais aussi des modèles logarithmiques ou exponentiels. Le modèle exponentiel est utilisé pour illustrer des phénomènes dont les variations sont très rapides : l'accroissement de la variable est ici proportionnel à sa

valeur.

Le modèle de regression est donc

$$Y = m(X) + \varepsilon,$$

où la variable d'erreur ε vérifie $\mathbb{E}[\varepsilon|X = x] = 0$ pour toute valeur x de X , de sorte que la fonction de régression $m(x)$ s'interprète encore de façon directe en termes d'espérance conditionnelle : $m(x) = \mathbb{E}[Y|X = x]$. Dans la suite, nous décrivons des méthodes qui permettent d'estimer la fonction de régression m de façon paramétrique ou non paramétrique sur la base de copies indépendantes $(X_1, Y_1), \cdots, (X_n, Y_n)$ du vecteur aléatoire $(X; Y)$. Un exemple de telle fonction de régression m et d'échantillon aléatoire associé est donné a la **(Figure 1.1)**.

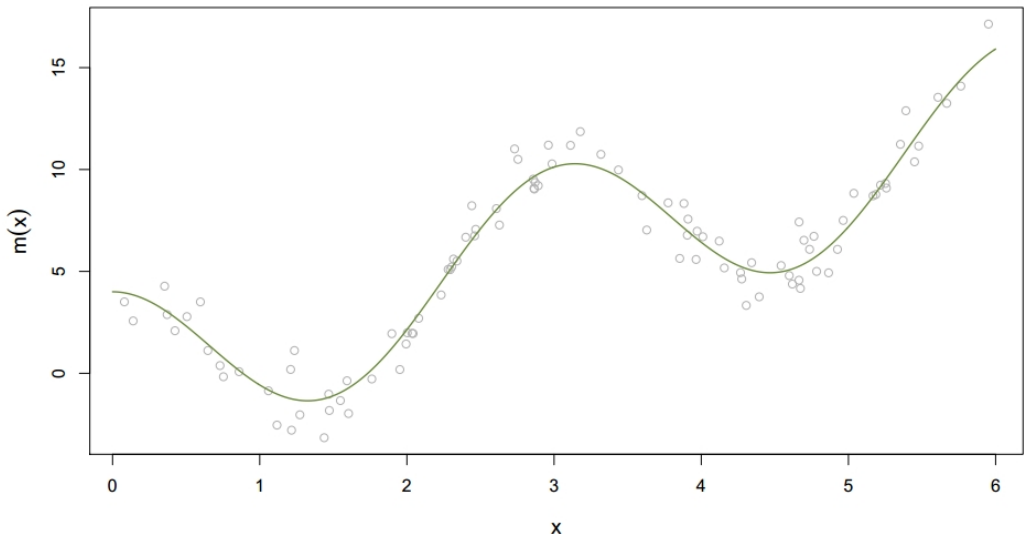


Fig 1.1 { Graphe de la fonction de régression $m(x) = 2x\sin(2x) + 4\cos(2x)$ et un échantillon aléatoire (X_i, Y_i), $i = 1, \cdots, 100$, engendré depuis le modèle $Y = m(X) + \varepsilon$, avec $\varepsilon \sim \mathcal{N}(0; 1)$ indépendant de $X \sim$Unif$(0; 6)$.

Chapitre 2

La régression linéaire et l'estimation paramétrique

Dans ce chapitre, nous allons analyser la régression linéaire (simple et multiple) : nous pouvons la voir comme une technique statistique permettant de modéliser la relation linéaire entre une variable explicative (notée X) et une variable à expliquer (notée Y). Cette présentation va nous permettre d'exposer la régression linéaire dans un cas simple afin de bien comprendre les enjeux de cette méthode, les problèmes posés et les réponses apportées.

2.1 La régression linéaire simple

L'approche la plus classique consiste a postuler que le lien entre Y et X est linéaire, au sens où il existe des nombres réels β_0 et β_1 tels que :

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

où la variable d'erreur ε vévifie $\mathbb{E}[\varepsilon|X = x] = 0$, pour toute valeur x de X . Cette approche est paramétrique puisque la fonction de régression $m(x) = E[Y|X = x] = \beta_0 + \beta_1 X$ est connue dès qu'un nombre fini de paramètres β_0, β_1 le sont. Dans une optique de prédiction de Y sur la base de X , l'estimateur usuel de $\beta = (\beta_0, \beta_1)$, à savoir l'estimateur des moindres carrés, permettra de construire le prédicteur

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X.$$

Exemple 2.1 .

La pollution de l'air constitue actuellement une des préoccupations majeures de santé publique. De nombreuses études épidémiologiques ont permis de mettre en évidence l'influence

sur la santé de certains composés chimiques comme le dioxyde de souffre SO_2, le dioxyde d'azote NO_2, l'ozone O_3, ou des particules sous forme de poussières contenues dans l'air. L'influence de cette pollution est notable sur les personnes sensibles (nouveau-nés, asthmatiques,personnes âgées). La prévision des pics de concentration de ces composés est donc importante.

Nous allons nous intéresser plus particulièrement à la concentration en ozone. Nous possédons quelques connaissances *a priori* sur la manière dont se forme l'ozone, grâce aux lois régissant les équilibres chimiques. La concentration de l'ozone sera fonction de la température ; plus la température sera élevée, plus la concentration en ozone va augmenter. Cette relation très vague doit être améliorée afin de pouvoir prédire les pics d'ozone.

Afin de mieux comprendre ce phénomène, l'association **Air Breizh** (surveillance de la qualité de l'air en Bretagne) mesure depuis 1994 la concentration en O_3 (en $\mu g/ml$) toute les 10 minutes et obtient donc le maximum journalier de la concentration en O_3 , noté dorénavant $\boldsymbol{O_3}$. Air Breizh collecte également à certaines heures de la journée des données météorologiques comme la température, la nébulosité, le vent....

Le tableau suivant donne les 10 premières mesures effectuées.

Individu	$\boldsymbol{O_3}$	$T12$
1	63, 3	13, 4
2	89, 6	15
3	79	7, 9
4	81, 2	13, 1
5	88	14, 1
6	68, 4	16, 7
7	139	26, 8
8	78, 2	18, 4
9	113, 8	27, 2
10	41, 8	20, 6

Tableau 2.1. 10 données de température à $12h$ et teneur en ozone.

Nous allons donc chercher à expliquer le maximum de $\boldsymbol{O_3}$ de la journée par la température à $12h$. D'un point de vue pratique, le but de cette régression est double :

– ajuster un modèle pour expliquer la concentration en $\boldsymbol{O_3}$ en fonction de $T12$;
– prédire les valeurs de concentration en $\boldsymbol{O_3}$ pour de nouvelles valeurs de $T12$.

Avant toute analyse, il est intéressant de représenter les données. Voici donc une représentation graphique des données. Chaque point du graphique ($Fig.2.1$) représente, pour un jour donné, une mesure de la température à $12h$ et le pic d'ozone de la journée.

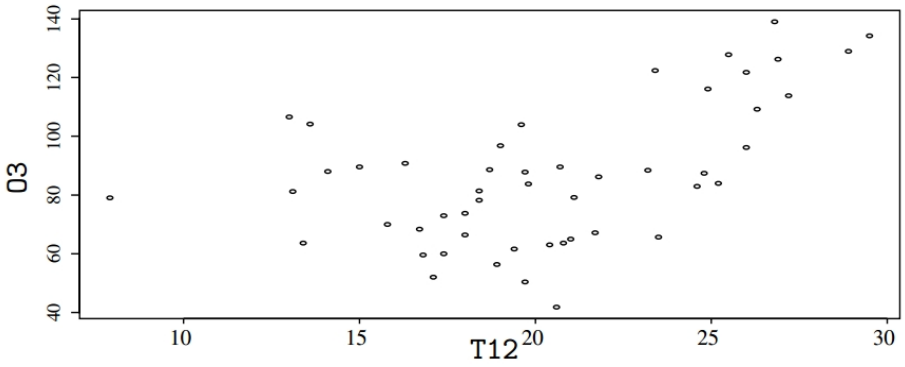


Fig. 2.1. 50 données journalières de température et O_3.

Pour analyser la relation entre les x_i (température) et les y_i (ozone), nous allons chercher une fonction f telle que

$$y_i \approx f(x_i)$$

Pour définir $\approx$, il faut donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données et une classe de fonctions $\mathcal{G}$ dans laquelle est supposée se trouver la vraie fonction inconnue. Le problème mathématique peut s'écrire de la façon suivante :

$$\arg\min_{f\in\mathcal{G}} \sum_{i=1}^{n} l\left(y_i - f(x_i)\right) \tag{2.1}$$

où n représente le nombre de données à analyser et $l(.)$ est appelée *fonction de coût* ou encore *fonction de perte*.

2.1.1 Modélisation mathématique

Nous venons de voir que le problème mathématique peut s'écrire de la façon suivante (cf. équation 2.1) :

$$\arg\min_{f\in\mathcal{G}} \sum_{i=1}^{n} l\left(y_i - f(x_i)\right)$$

Nous allons discuter du choix de la fonction de coût et de l'ensemble $\mathcal{G}$. Nous présenterons des graphiques illustratifs bâtis à partir de 10 données fictives de température et de concentration en ozone.

Choix du critère de qualité et distance à la droite

De nombreuses fonctions de coût $l(.)$ existent, mais les deux principales utilisées sont les suivantes :

- $l(u) = u^2$: coût quadratique .
- $l(u) = |u|$: coût absolu.

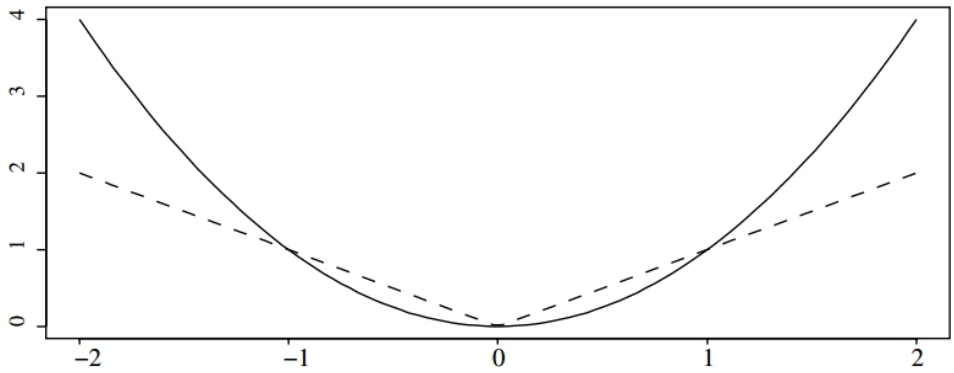


Fig. 2.2.Coût absolu (pointillés) et coût quadratique (trait plein).

Ces fonctions sont positives, symétriques, elles donnent donc la même valeur lorsque l'erreur est positive ou négative et s'annulent lorsque u vaut zéro.

La fonctionl peut aussi être vue comme la distance entre une observation (x_i, y_i) et son point correspondant sur la droite $(x_i, f(x_i))$ **(voir fig. 2.3).**

:/Users/chkinfo/AppData/Local/Temp/graphics/RVN0A602$_{4.pdf}$

Fig. 2.3. Distances à la droite : coût absolu (pointillés) et distance d'un point à une droite.

Par point correspondant, nous entendons « évalué » à la même valeur x_i. Nous aurions pu prendre comme critère à minimiser la somme des distances des points (x_i, y_i) à la droite (cf. fig. 1.3), mais ce type de distance n'entre pas dans le cadre des fonctions de coût puisqu'au point (x_i, y_i) correspond sur la droite un point $(x_i', f(x_i'))$ d'abscisse et d'ordonnée différentes.

Il est évident, que par rapport au coût absolu, le coût quadratique accorde une importance plus grande aux points qui restent éloignés de la droite ajustée, la distance étant élevée au carré (cf. fig. 1.2). Sur l'exemple fictif, dans la classe $\mathcal{G}$ des fonctions linéaires, nous allons minimiser $\sum_{i=1}^{n} (y_i - f(x_i))^2$ (coût quadratique) et $\sum_{i=1}^{n} |y_i - f(x_i)|$ (coût absolu). Les droites ajustées sont représentées sur le graphique ci-dessous :

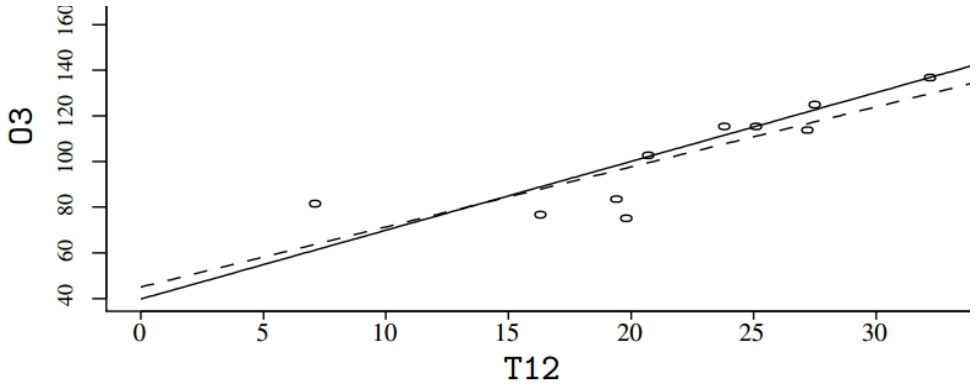


Fig. 2.4. 10 données fictives de température et O_3, régressions avec un coût absolu (trait plein) et quadratique (pointillé).

La droite ajustée avec un coût quadratique propose un compromis où aucun point n'est très éloigné de la droite : le coût quadratique est sensible aux points aberrants qui sont éloignés de la droite. Ainsi (fig. 1.4) le premier point d'abscisse approximative $7°C$ est assez éloigné des autres. La droite ajustée avec un coût quadratique lui accorde une plus grosse importance que l'autre droite et passe relativement donc plus près de lui. En enlevant ce point (de manière imaginaire), la droite ajustée risque d'être très différente : le point est dit in-fluent et le coût quadratique peu robuste. Le coût absolu est plus robuste et la modification d'une observation modifie moins la droite ajustée. Malgré cette non-robustesse, le coût quadratique est le coût le plus sou-vent utilisé, ceci pour plusieurs raisons : historique, calculabilité, propriétés mathématiques. En 1800, il n'existait pas d'ordinateur et l'utilisation du coût quadratique permettait de calculer explicitement les estimateurs à partir des données. A propos de l'utilisation d'autres fonctions de coût, voici ce que disait **Gauss (1809)** : « Mais de tous ces principes, celui des moindres carrés est le plus simple : avec les autres, nous serions conduits aux calculs les plus complexes ». Les lecteurs intéressés par le coût absolu peuvent consulter le livre de **Dodge & Rousson (2004)**.

Choix des fonctions à utiliser

Si la classe $\mathcal{G}$ est trop large, par exemple la classe des fonctions continues (C_0), un grand nombre de fonctions de cette classe minimisent le critère (2.1). Ainsi toutes les fonctions de la classe qui passent par tous les points (interpolation), quand c'est possible, annulent la quantité $\sum_{i=1}^{n} l\left(y_i - f(x_i)\right)$.

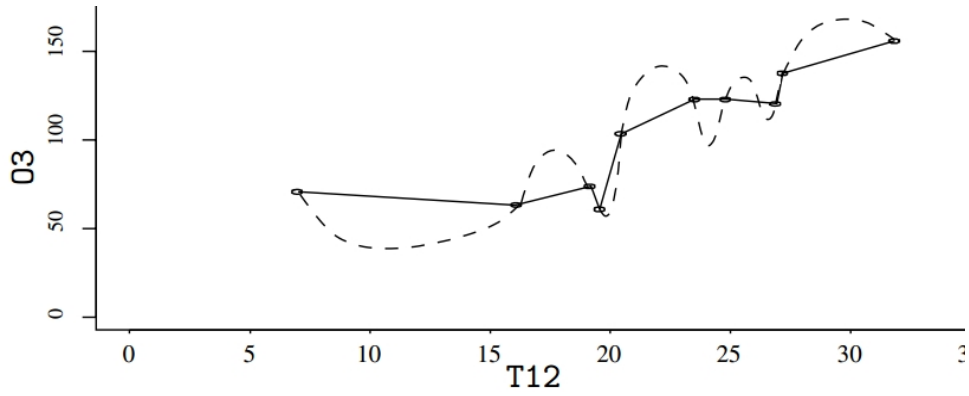


Fig. 2.5. Deux fonctions continues annulant le critère (2.1).

La fonction continue tracée en pointillés sur la figure (fig. 1.5) semble inappropriée bien qu'elle annule le critère (2.1). La fonction continue tracée en traits pleins annule aussi le critère (2.1). D'autres fonctions continues annulent ce critère, la classe des fonctions continues est trop vaste. Ces fonctions passent par tous les points et c'est là leur principal défaut. Nous souhaitons plutôt une courbe, ne passant pas par tous les points, mais possédant un trajet harmonieux, sans trop de détours. Bien sûr le trajet sans aucun détour est la ligne droite et la classe $\mathcal{G}$ la plus simple sera l'ensemble des fonctions affines. Par abus de langage, on emploie le terme de fonctions linéaires. D'autres classes de fonctions peuvent être choisies et ce choix est en général dicté par une connaissance a priori du phénomène et (ou) par l'observation des données.

Ainsi une étude de régression linéaire simple débute toujours par un tracé des observations (x, y). Cette première représentation permet de savoir si le modèle linéaire est pertinent. Le graphique suivant représente trois nuages de points différents.

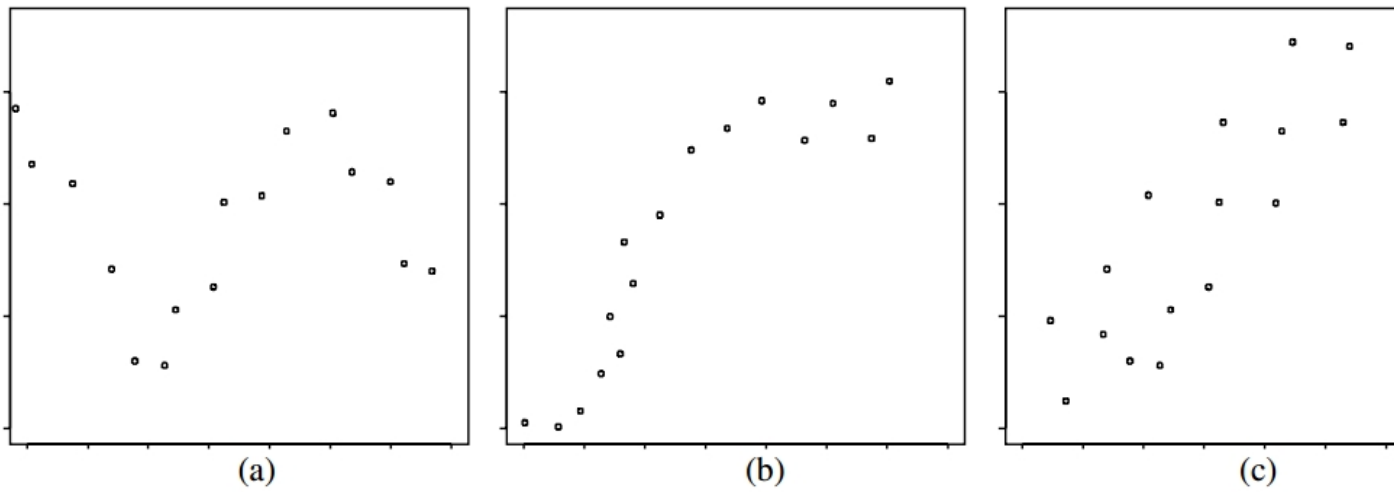


Fig. 2.6. Exemples fictifs de tracés : (a) fonction sinusoïdale, (b) fonction croissante sigmoïdale et (c) droite.

Au vu du graphique, il semble inadéquat de proposer une régression linéaire pour les deux premiers graphiques, le tracé présentant une forme sinusoïdale ou sigmoïdale. Par contre, la modélisation par une droite de la relation entre X et Y pour le dernier graphique semble correspondre à la réalité de la liaison. Dans la suite de cette section, nous prendrons $\mathcal{G} = \{f : \ f(x) = ax + b, \quad (a, b) \in \mathbb{R}^2\}$.

2.1.2 Modélisation statistique

Lorsque nous ajustons par une droite les données, nous supposons implicitement qu'elles étaient de la forme

$$Y = \beta_0 + \beta_1 X$$

Dans l'exemple, nous supposons donc un modèle où la concentration d'ozone dépend linéairement de la température. Nous savons pertinemment que toutes les observations mesurées ne sont pas sur la droite. D'une part, il est irréaliste de croire que la concentration de l'ozone dépend linéairement de la température et de la température seulement. D'autre part, les mesures effectuées dépendent de la précision de l'appareil de mesure, de l'opérateur et il arrive souvent que, pour des valeurs identiques de la variable X, nous observions des valeurs différentes pour Y. Nous supposons alors que la concentration d'ozone dépend linéairement de la température mais cette liaison est perturbée par un «bruit ». Nous supposons en fait que les données suivent le modèle suivant :

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2.2}$$

L'équation (2.2) est appelée **modèle de régression linéaire** et dans ce cas précis **modèle de régression linéaire simple**. Les β_j, $j = 0, 1$, appelés les paramètres du modèle (constante de régression et coefficient de régression), sont fixes mais inconnus, et nous voulons les estimer. La quantité notée ε est appelée bruit, ou erreur, et est aléatoire et inconnue.

Afin d'estimer les paramètres inconnus du modèle, nous mesurons dans le cadre de la régression simple une seule variable explicative ou variable exogène X et une variable à expliquer ou variable endogène Y. La variable X est souvent considérée comme non aléatoire au contraire de Y. Nous mesurons alors n observations de la variable X, notées x_i, où i varie de 1 à n et n valeurs de la variable à expliquer Y notées y_i.

Nous supposons que nous avons collecténcouples de données (x_i, y_i) où y_i est la réalisation de la variable aléatoire Y_i. Par abus de notation, nous confondrons la variable aléatoire Y_i et sa réalisation, l'observation y_i. Avec la notation ε_i, nous confondrons la variable aléatoire avec sa réalisation. Suivant le modèle (2.2), nous pouvons écrire

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

où

– *les x_i sont des valeurs connues non aléatoires ;*
– *les paramètres β_j, $j = 0, 1$ du modèle sont inconnus ;*
– *les ε_i sont les réalisations inconnues d'une variable aléatoire ;*
– *les y_i sont les observations d'une variable aléatoire.*

.

2.1.3 Estimateurs des moindres carrés

(estimateurs des MC)
On appelle estimateurs des moindres carrés **(MC)** de β_j, les estimateurs $\widehat{\beta}_j$ obtenus par minimisation de la quantité

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

où $\mathbf{1}$ est le vecteur de $\mathbb{R}^n$ dont tous les coefficients valent 1. Les estimateurs peuvent également s'écrire sous la forme suivante :

$$\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \underset{(\beta_0,\beta_1)\in\mathbb{R}\times\mathbb{R}}{\arg\min}\ S(\beta_0, \beta_1).$$

Calcul des estimateurs de β_j, quelques propriétés

La fonction $S(\beta_0, \beta_1)$ est strictement convexe. Si elle admet un point singulier, celui-ci correspond à l'unique minimum. Annulons les dérivées partielles, nous obtenons un système d'équations appelées « équations normales » :

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

La première équation donne

$$\widehat{\beta}_0 n + \widehat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$

et nous avons un estimateur de l'ordonnée à l'origine

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}, \tag{2.3}$$

La seconde équation donne

$$\widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

En remplaçant $\widehat{\beta}_0$ par son expression (2.3) nous avons une première écriture de

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \overline{y}}{\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \overline{x}},$$

et en utilisant la nullité de la somme $(x_i - \overline{x})$, nous avons d'autres écritures pour l'estimateur de la pente de la droite

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \overline{y})}{\sum_{i=1}^{n} x_i (x_i - \overline{x})} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) y_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{2.4}$$

Pour obtenir ce résultat, nous supposons qu'il existe au moins deux points d'abscisses différentes. Cette hypothèse notée $\mathcal{H}_1$ s'écrit $x_i \neq x_j$ pour au moins deux individus. Elle permet d'obtenir l'unicité des coefficients estimés $\widehat{\beta}_0$et $\widehat{\beta}_1$.Nous pouvons maintenant estimer la droite de régression par la formule

$$\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X.$$

Si nous évaluons la droite aux points x_i ayant servi à estimer les paramètres, nous obtenons des y_i et ces valeurs sont appelées les valeurs ajustées. Si nous évaluons la droite en d'autres points, les valeurs obtenues seront appelées les valeurs prévues ou prévisions. Représentons les points initiaux et la droite de régression estimée. La droite de régression passe par le centre de gravité du nuage de points $(\overline{x}, \overline{y})$ comme l'indique l'équation (2.3).

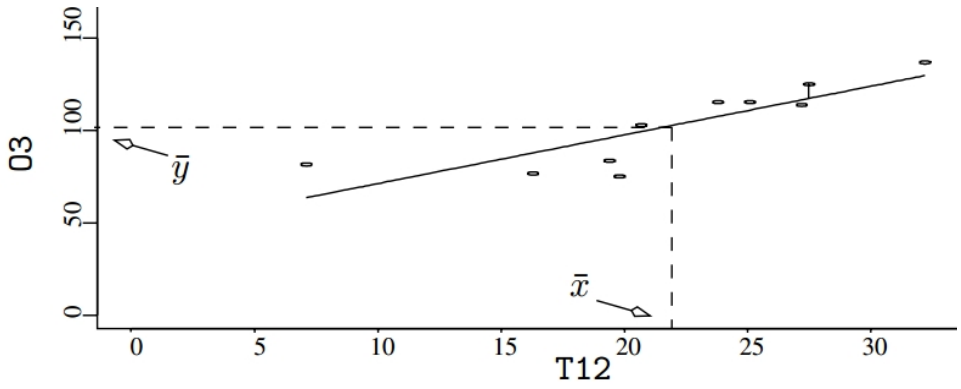


Fig. 2.7. Nuage de points, droite de régression et centre de gravité.

Nous avons réalisé une expérience et nous avons mesuré n valeurs (x_i, y_i). A partir de ces n valeurs, nous avons obtenu un estimateur de β_0 et de β_1. Si nous refaisions une expérience, nous mesurerions n nouveaux couples de données (x_i, y_i). A partir de ces données, nous aurions un nouvel estimateur de β_0 et de β_1. Les estimateurs sont fonction des données

mesurées et changent donc avec les observations collectées **(fig. 2.8)**. Les vraies valeurs de β_0 et β_1 sont inconnues et ne changent pas.

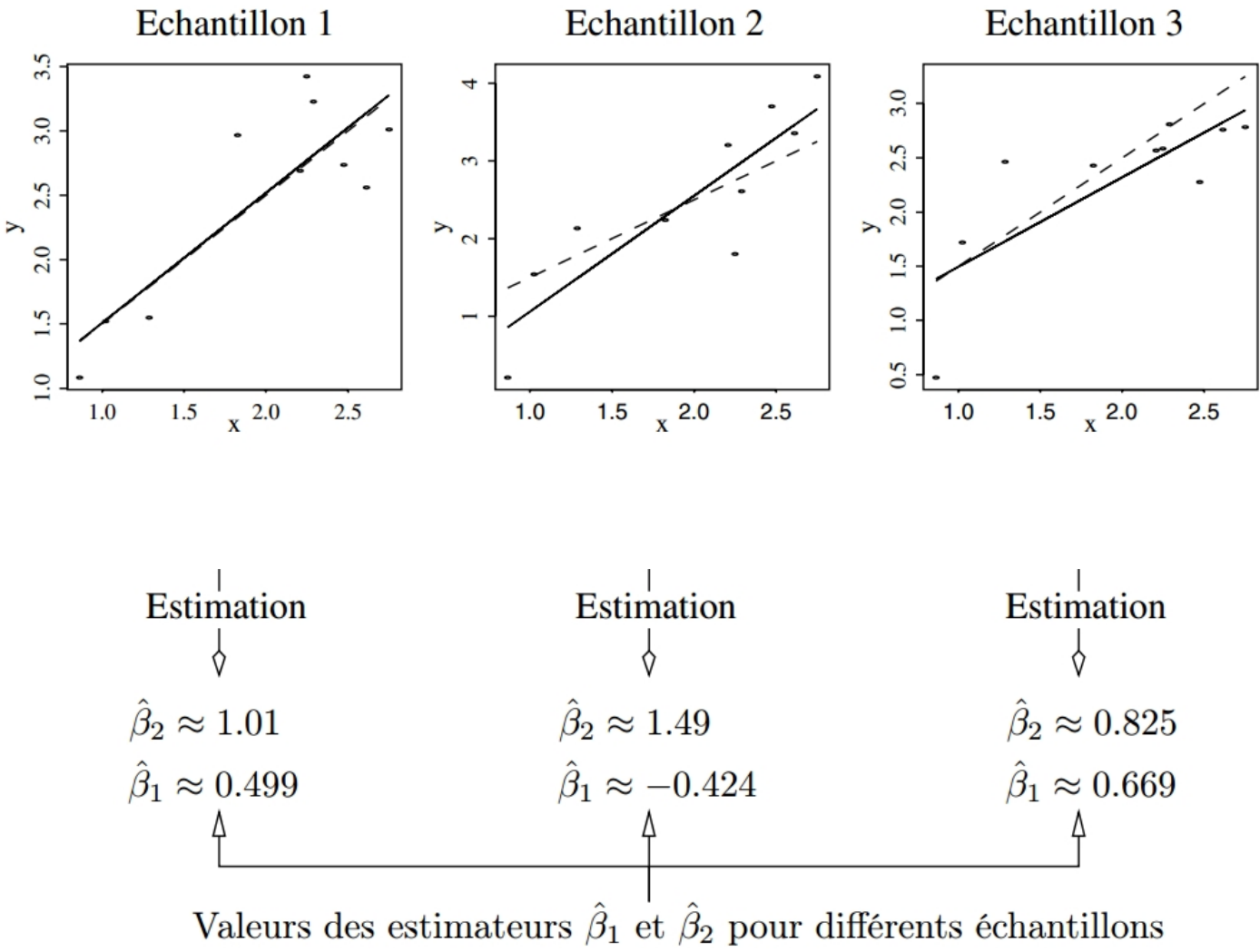


Fig. 2.8. Exemple de la variabilité des estimations. Le vrai modèle est $Y = X + 0.5 + \varepsilon$, où ε est choisi comme suivant une loi $\mathcal{N}(0, 0.25)$. Nous avons ici 3 répétitions de la mesure de 10 points (x_i, y_i), Le trait en pointillé représente la vraie droite de régression et le trait plein son estimation.

Le statisticien cherche en général à vérifier que les estimateurs utilisés admettent certaines propriétés comme :

– un estimateur $\widehat{\beta}$ est-il sans biais ? Par définition $\widehat{\beta}$ est sans biais si $\mathbb{E}\left[\widehat{\beta}\right] = \beta$. En moyenne sur toutes les expériences possibles de taille n, l'estimateur $\widehat{\beta}$ moyen sera égal à la valeur inconnue du paramètre.

– un estimateur $\widehat{\beta}$ est-il de variance minimale parmi les estimateurs d'une classe définie ? En d'autres termes, parmi tous les estimateurs de la classe, l'estimateur utilisé admet-il parmi toutes les expériences la plus petite variabilité ?

Pour cela, nous supposons une seconde hypothèse notée $\mathcal{H}_2$ qui s'énonce aussi comme suit : les erreurs sont centrées, de même variance et non corrélées entre elles. Elle permet de calculer les propriétés statistiques des estimateurs. $\mathcal{H}_2$: $\mathbb{E}[\varepsilon_i] = 0$, pour $i = 1, \cdots, n$ et $\text{cov}(\varepsilon_i, \varepsilon_j) = \delta_{i,j}\sigma^2$, où $\delta_{i,j} = 1$ si $i = j$ et $\delta_{i,j} = 0$ lorsque $i \neq j$.

Nous avons les propriétés de ces estimateurs

Proposition 2.1 ***(Biais des estimateurs)***
$\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont des estimateurs sans biais de β_0 et $\beta 1$, c'est-à-dire que $\mathbb{E}\left[\widehat{\beta}_0\right] = \beta_0$ et $\mathbb{E}\left[\widehat{\beta}_1\right] = \beta_1$

Proposition 2.2 ***(Variances de $\widehat{\beta}_0$ et $\widehat{\beta}_1$)***

Les variances et covariance des estimateurs des paramètres valent :

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

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

$$Cov\left(\widehat{\beta}_0\ ,\ \widehat{\beta}_1\right) = -\frac{\sigma^2 \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

Cette proposition nous permet d'envisager la précision de l'estimation en utilisant la variance. Plus la variance est faible, plus l'estimateur sera précis. Pour avoir des variances petites, il faut avoir un numérateur petit et (ou) un dénominateur grand. Les estimateurs seront donc de faibles variances lorsque : la variance σ^2 est faible. Cela signifie que la variance de Y est faible et donc les mesures sont proches de la droite à estimer ;

– la quantité $\sum_{i=1}^n (x_i - \overline{x})^2$ est grande, les mesures x_i doivent être dispersées autour de leur moyenne ;

– la quantité $\sum_{i=1}^n x_i^2$ ne doit pas être trop grande, les points doivent avoir une faible moyenne en valeur absolue. En effet, nous avons

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

L'équation (2.3) indique que la droite des MC passe par le centre de gravité du nuage $(\overline{x}, \overline{y})$. Supposons $\overline{x}$ positif, alors si nous augmentons la pente, l'ordonnée à l'origine va diminuer et vice versa. Nous retrouvons donc le signe négatif pour la covariance entre $\widehat{\beta}_0$ et $\widehat{\beta}_1$. Nous terminons cette partie concernant les propriétés par le théorème de Gauss-Markov qui indique que, parmi tous les estimateurs linéaires sans biais, les estimateurs des MC possèdent la plus petite variance.

Théorème 2.1 ***(Gauss-Markov)***

Parmi les estimateurs sans biais linéaires en Y, les estimateurs $\widehat{\beta}_j$ sont de variance minimale.

2.1.4 Estimateurs du maximum de vraisemblance

La vraisemblance vaut

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

Ce qui donne pour la log-vraisemblance :

$$\log\mathcal{L}\left(\beta_0,\beta_1,\sigma^2\right) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}S(\beta_0,\beta_1)$$

Nous voulons maximiser cette quantité par rapport aux trois variables $(\beta_0,\beta_1,\sigma^2)$. Les deux premières variables n'apparaissent que dans le terme en $-S(\beta_0,\beta_1)$, qu'il faut donc minimiser. Or on a déjà vu que cette quantité est minimale lorsqu'on considère les estimateurs des moindres carrés, c'est-à-dire pour $\beta_1=\hat{\beta}_1$ et $\beta_2=\hat{\beta}_2$.

Bilan : les estimateurs du maximum de vraisemblance de β_1 et β_2 sont égaux aux estimateurs des moindres carrés.

Ceci étant vu, il reste simplement à maximiser $\log\mathcal{L}(\hat{\beta}_1,\hat{\beta}_2,\sigma^2)$ par rapport à σ^2 . Calculons donc la dérivée par rapport à σ^2 :

$$\frac{\partial\log\mathcal{L}(\hat{\beta}_0,\hat{\beta}_1,\sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}S(\hat{\beta}_1,\hat{\beta}_2) = -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_2 x_i\right)^2$$

D'où l'on déduit que l'estimateur du maximum de vraisemblance de σ^2 est :

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

L'estimateur du maximum de vraisemblance de σ^2 est donc biaisé. On a en effet $\mathbb{E}[\hat{\sigma}^2_{mv}] = \dfrac{n-2}{n}\sigma^2$, mais ce biais est d'autant plus négligeable que le nombre d'observations est grand.

2.2 La régression linéaire multiple

La modélisation de la concentration d'ozone dans l'atmosphère évoquée au section 1 est relativement simpliste.En effet,des variables météorologiques autres que la température peuvent expliquer cette concentration,comme par exemple le rayonnement, la précipitation ou encore le vent qui déplace les masses d'air. L'association Air Breizh mesure ainsi en même temps que la concentration d'ozone les variables météorologiques susceptibles d'avoir une influence sur celle-ci Voici quelques-unes de ces données :

Individu	$\boldsymbol{O_3}$	$T12$	Vx	$Ne12$
1	63.6	13.4	9.35	7
2	89.6	15	5.4	4
3	79	7.9	19.3	8
4	81.2	13.1	12.6	7
5	88	14.1	-20.3	6
6	68.4	16.7	-3.69	7
7	139	26.8	8.27	1
8	78.2	18.4	4.93	7
9	113.8	27.2	-4.93	6
10	41.8	20.6	-3.38	8

Tableau 2.2. 10 données journalières.

La variable Vx est une variable synthétique représentant le vent. Le vent est normalement mesuré en degré (direction) et mètre par seconde (vitesse). La variable créée est la projection du vent sur l'axe est-ouest, elle tient compte de la direction et de la vitesse. La variable $Ne12$ représente la nébulosité mesurée à 12 heures. Pour analyser la relation entre la température ($T12$), le vent (Vx), la nébulosité à midi ($Ne12$) et l'ozone ($\boldsymbol{O_3}$),nous allons chercher une fonction f telleque

$$\boldsymbol{O_{3i}} \approx f(T12_i, Vx_i, Ne12_i).$$

Afin de préciser le sens de $\approx$, il faut définir un critère positif quantifiant la qualité de l'ajustement de la fonction f aux données.Cette notion de coût permet d'appréhender de manière aisée les problèmes d'ajutement économique dans certains modèles.Minimiser un coût nécessite la connaissance de l'espace sur lequel on minimise,donc la classe de fonctions $\mathcal{G}$ dans laquelle nous supposerons que se trouve la vraie fonction inconnue. Le problème mathématique peut s'écrire de la façon suivante :

$$\arg\min_{f\in\mathcal{G}} \sum_{i=1}^{n} l(y_i - f(x_{i_1}, ..., x_{ip})),$$

où n représente le nombre de données à analyser et $l(.)$ est appelée fonction de coût. La fonction de coût sera la même que celle utilisée précédemment, c'est-à dire le coût quadratique. En ce qui concerne le choix de la classe $\mathcal{G}$,

$$\mathcal{F} = \Big\{ f : f(x_1, ..., x_p) = \sum_{j=1}^{p} \beta_j x_j \quad avec \quad \beta_j \in \mathbb{R},\ j \in \{1, ..., p\} \Big\}.$$

2.2.1 Modélisation

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 variables explicatives sont en nombre fini.Nous supposons donc que les données collectées suivent le modèle suivant :

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, ..., n \tag{2.5}$$

où

– les x_{ij} sont des nombres connus, non aléatoires. La variables x_{i1} peut valoir 1 pour tout i à n. Dans ce cas, β_1 représente la constante (intercept dans les logiciels anglo-saxons). En statistiques,cette colonne de 1

est presque toujours présente.

– les paramètres à estimer β_j du modèle sont inconnus.
– les ε_i sont des variables aléatoires inconnues.

En utilisant l'écriture matricielle de (2.5), nous obtenons la définition suivante

$$Y_{n\times 1} = X_{n\times p}\ \beta_{p\times 1} + \varepsilon_{n\times 1} \tag{2.6}$$

où :

– *Y est un vecteur aléatoire de dimension n,*
– *X est une matrice de taille $n \times p$ connue, appelée matrice du plan d'expérience, X est la concaténation des p variables X_j : $X = (X_1 \,|X_2|\, ...\, |X_p|)$. Nous noterons la i^e ligne du tableau X par le vecteur ligne $x_i' = (x_{i1}, ...x_{ip})$;*
– *β est le vecteur de dimension p des paramètres inconnus du modèle ;*
– *ε est le vecteur centré,de dimension n des erreurs.*

Nous supposons que la matrice X est de plein rang .Cette hypothèse sera notée $\mathcal{H}_1$. Comme, en général , le nombre d'individus n est plus grand que le nombre de variables explicatives p, le rang de la matrice X vaut p.

La présentation précédente revient à supposer que la fonction liant Y aux variables explicatives X est un hyperplan représenté (**Fig 2.9**).

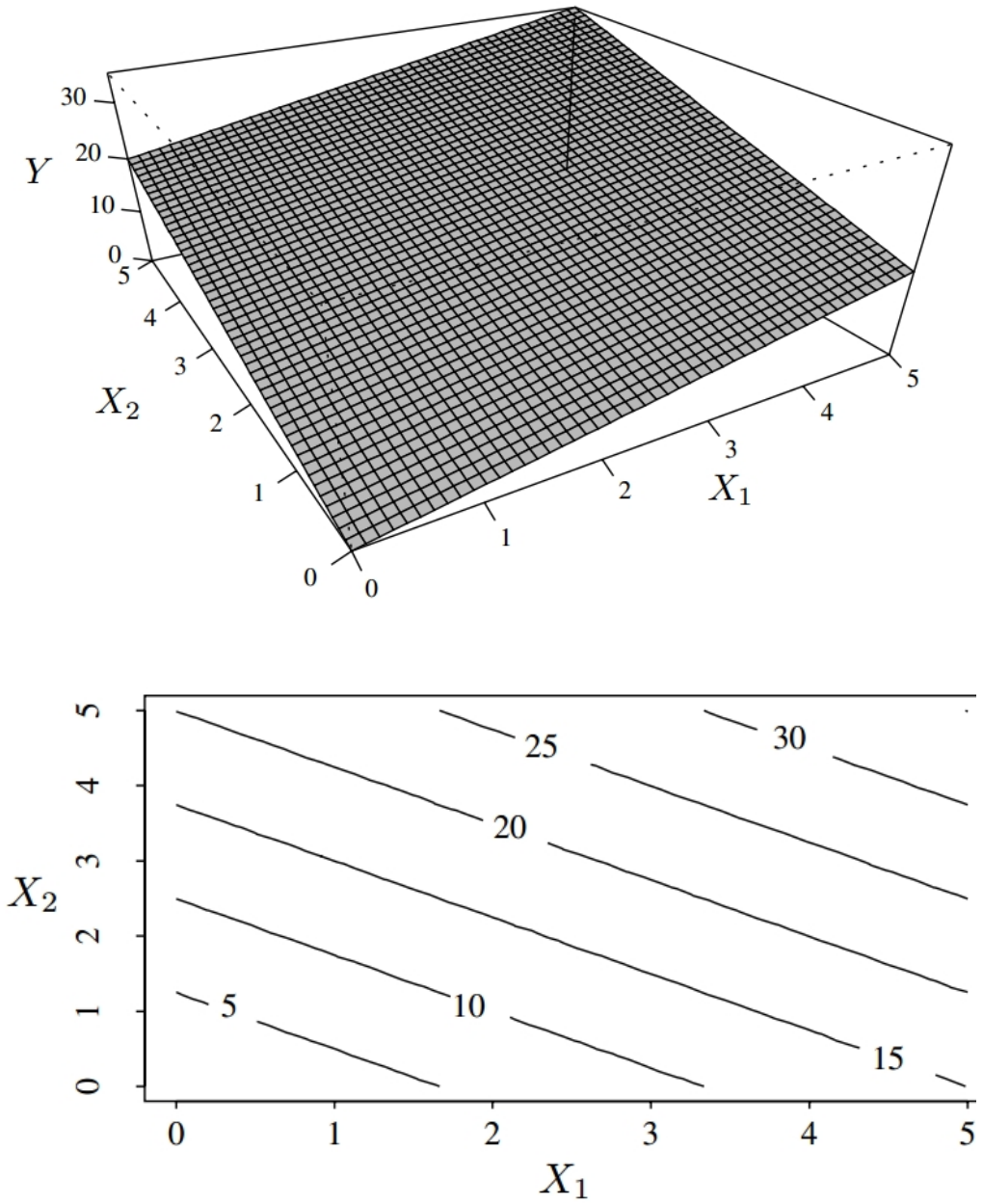


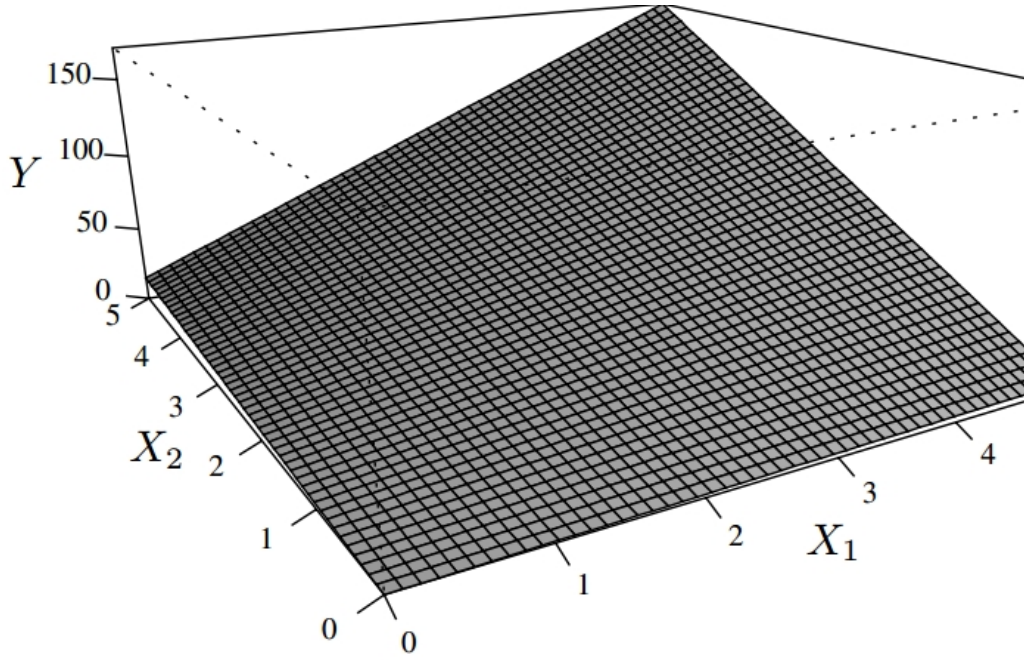
Fig 2.9. Représentation géométrique de la relation $Y = 3X_1 + 4X_2$.

Il est naturel dans nombre de problémes de penser que des interactions existent entre les variables explicatives.Dans l'exemple de l'ozone,nous pouvons penser que la température et

le vent interagissent .Pour modéliser cetteinteration,nous écrivons en général un modèle avec un produit entre les variables explicatives qui interagissent, Ainsi,pour deux variables,nous avons la modélisation suivante :

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i, \qquad i = 1, ..., n.$$

Les produits peuvent s'effectuer entre deux variables définissant des interactions d'ordre 2,entre trois variables définissant des interactions d'ordre 3, etc..D'un point de vue géométrique,cela donne (**Fig 2.10**)



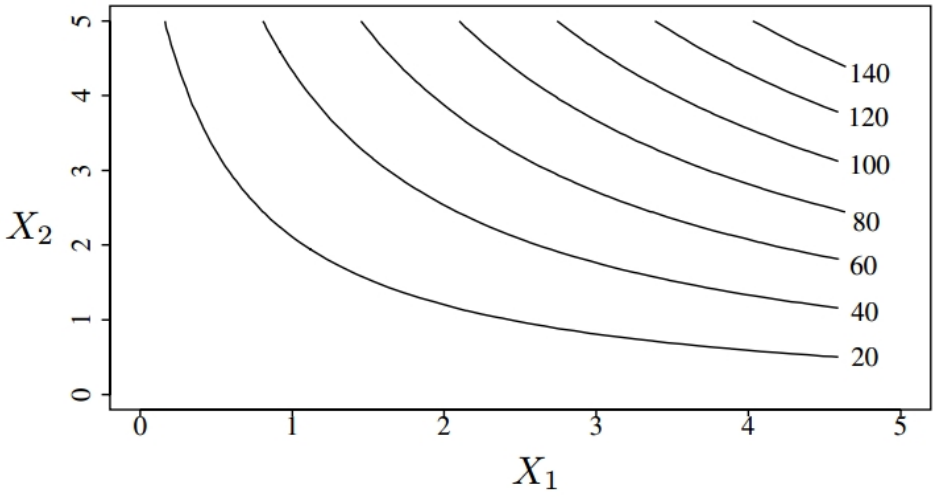
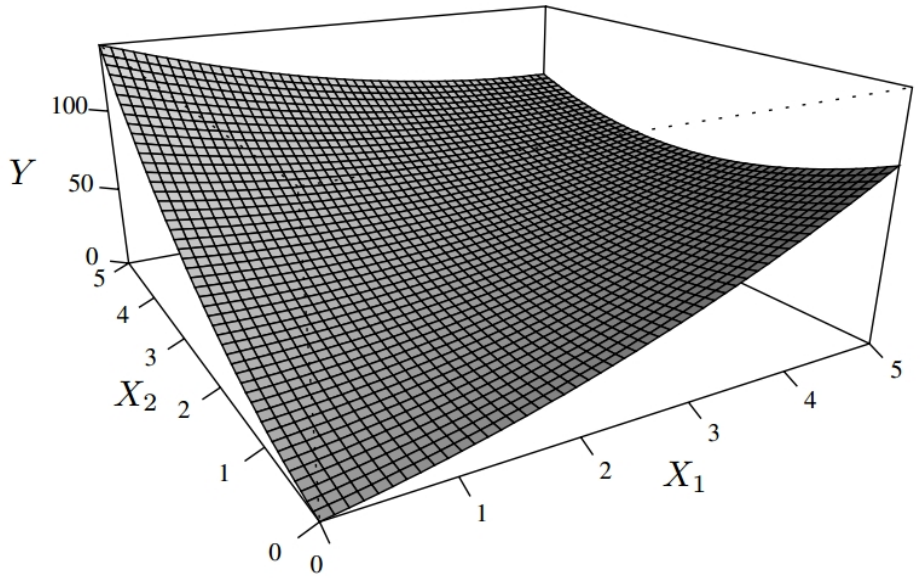


Fig 2.10. Représentation géométrique de la relation $y = X_1 + 3X_2 + 6X_1X_2$.

Cependant ce type de modélisation rentre parfaitement dans le cadre de la régression multiple. Les variables d'interaction sont des produits de variables connues et sont donc connues. Dans l'exemple précédent, la troisième variable explicative X_3 sera tout simplement le produit. De même, d'autres extensions peuvent êtrre utilisées comme le modèle de régression polynômial. En reprenant notre exemple à deux variables explicatives X_1 et X_2, nous pouvons proposer le modèle polynômial de degré 2 suivant :

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2 + \varepsilon_i, \qquad i = 1, ..., n.$$

Ce modèle peut être remis dans la formulation de la section précédente en posant $X_3 = X_1X_2$, $X_4 = X_i^2$ et $X_5 = X_2^2$. L'hyper surface ressemble alors à (**Fig 2.11**)



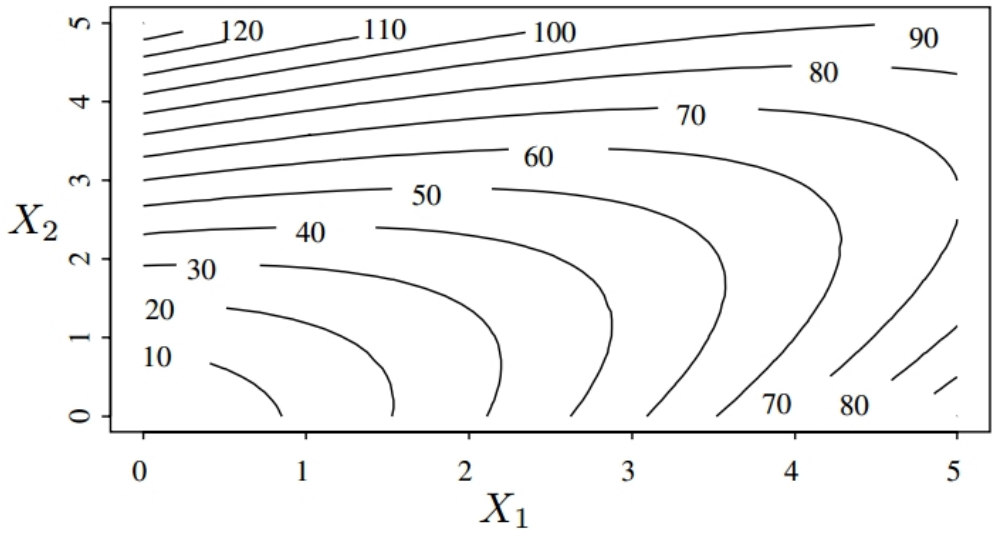


Fig 2.11. Représentation géométrique de la relation
$Y = 10X_1 + 8X_2 - 6X_1X_2 + 2X_1^2 + 4X_2^2$.

En conclusion nous pouvons considérer que n'importe quelle transformation connue et fixée des variables explicatives (logarithme ,exponentielle,produit etc.) rentre dans le modèle de régression multiple.La transformée d'une variable explicative X_1 par une fonction connue et fixe (log par exemple) devient $X_i = \log(X_1)$et le modèle reste donc un modèle de régression multiple.Par contre une transformation comme $\exp\{-r(X_1 - k)\}$ qui est une fonction non linéaire de deux paramètres inconnus r et k il est impossible de calculer $\exp\{-r(X_1 - k)\}$ et donc de la noter X_i. Ce type de relation est traité dans Antoniadis et al.(1992).

Ainsi un modèle linéaire ne veut pas foreément dire que le lien entre variables explicatives et la variable à expliquer est linéaire mais que le *modèle est linéaire en les paramètres.*

2.2.2 Estimateurs des moindres carrés MC

Définition 2.1 ***(Estimateur des MC)***
on appelle estimateur des moindres carrés (noté MC) $\hat{\beta}$ de β la valeur suivante :

$$\hat{\beta} = \underset{\beta_1,\ldots,\beta_p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \underset{\beta \in \mathbb{R}^p}{\arg\min} (Y - X\beta)'(Y - X\beta).$$

Théorème 2.2 ***(Expression de l'estimateur des MC)***
Si l'hypothèse $\mathcal{H}_1$ est vérifiée, l'estimateur des MC $\hat{\beta}$ de β vaut

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

Calcul de $\hat{\beta}$

Il est intéressant de considérer les variables dans l'espace des variables ($\mathbb{R}^n$). Ainsi, Y vecteur colonne, définit dans $\mathbb{R}^n$ un vecteur $\overrightarrow{OY}$ d'orgine O et d'extrémité Y. Ce vecteur a pour coordonnées $(y_1, ..., y_n)$. La matrice X du plan d'expérience est formée p vecteurs colonnes . Chaque vecteur X_j définit dans $\mathbb{R}^n$ un vecteur $\overrightarrow{OX_j}$ d'origine O et d'extrémité X_j. Ce vecteur a pour coordonnées $(x_{1i}, ..., x_{nj})$. Ces p vecteurs linéairement indépendants (hypothèse $\mathcal{H}_1$) engendrent un sous espace vectoriel de $\mathbb{R}^n$,noté dorénavant $\Im(X)$,de dimension p

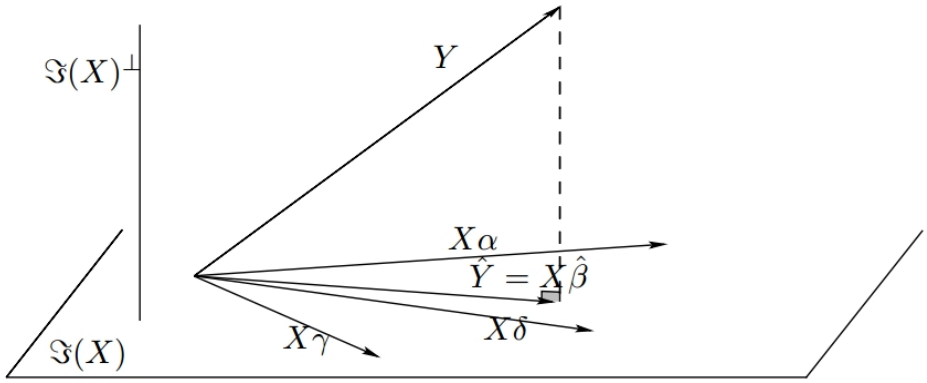


Fig. 2.12. Représentation dans l'espace des variables.

Cet espace $\Im(X)$, appelé image de X, est engendré par les colonnes de X. Il est parfois appelé espace des solutions. Ainsi, tout vecteur $\overrightarrow{v}$ de $\Im(X)$ s'écrit de façon unique sous la forme suivante :

$$\overrightarrow{v} = \alpha_1 \overrightarrow{X_1} + ... + \alpha_p \overrightarrow{X_p} = X\alpha.$$

Selon le modèle (2.6), le vecteur Y est la somme d'un élément de $\Im(X)$ et d'un bruit, élément de $\mathbb{R}^n$, qui n'a aucune rasion d'appartenir à $\Im(X)$. Minimiser $S(\beta)$ revient à chercher un élément de $\Im(X)$ qui soit le plus proche de Y, au sens de la norme eulidienne classique. Par définition, cet unique élément est appelé projection orthogonale de Y sur $\Im(X)$. Il sera noté $\hat{Y} = P_X Y$ où P_X est la matrice de projection orthogonale sur $\Im(X)$. Dans la littérature anglo saxonne, cette matrice est souvent notée H et est appelée ≪hat matrix≫ car elle met des ≪ hat ≫ sur Y. Par souci de cohérence de l'écriture, nous noterons l'élément courant (i, j) de P_X, h_{ij}. L'élément $\hat{Y}$ de $\Im(X)$ est aussi noté $\hat{Y} = X\hat{\beta}$, où $\hat{\beta}$ est l'estimateur des MC de β. L'espace orthogonal à $\Im(X)$ noté $\Im(X)^{\perp}$ est souvent appelé espace des résidus. Le vecteur $\hat{Y} = P_X Y$ contient les valeures ajustées par le modèle de Y .

– Calcul de $\hat{\beta}$ par projection :

Trois possibilités de calcul de $\hat{\beta}$ sont proposées.

- La première consiste à connaître la forme analytique de P_X.La matrice de projection orthogonale sur $\Im(X)$ est donnée par :

  $P_X = X(X'X)^{-1}X'$
  et,comme $P_XY = X\hat{\beta}$,nous obtenons $\hat{\beta} = (X'X)^{-1}X'Y$.

- La deuxième méthode utilise le fait que le vecteur Y de $\mathbb{R}^n$ se décompose de façon unique en une partie sur $\Im(X)$ et une partie sur $\Im(X)^{\perp}$,cela s'écrit :

  $Y = P_XY + (I - P_X)Y.$

La quantité $(I - P_X)Y$ étant un élément de $\Im(X)^{\perp}$ est orthogonale à tout élément v quelconque de $\Im(X)$. Rappelons que $\Im(X)$ est l'espace engendré par les colonnes de X, c'est-à-dire que toutes les combinaisons linéaires de variables $X_1, ...X_P$ sont éléments de $\Im(X)$ ou encore que, pour tout $\alpha \in \mathbb{R}^p$, nous avons $X\alpha \in \Im(X)$. Les deux vecteurs v et $(I - P_X)Y$ étant orthogonaux, le produit scalaire entre ces deux quantités est nul, soit :

$$\begin{aligned}
\langle v, (I - P_X)Y\rangle &= 0 \quad \forall v \in \Im(X) \\
\langle X\alpha, (I - P_X)Y\rangle &= 0 \quad \forall \alpha \in \mathbb{R}^{\mathbb{P}} \\
\alpha'X'(I - P_X)Y &= 0 \\
X'Y &= X'P_XY \quad avec \;\; P_XY = X\hat{\beta} \\
X'Y &= X'X\hat{\beta} \quad X \;\text{ de rang plein} \\
\hat{\beta} &= (X'X)^{-1}X'Y.
\end{aligned}$$

Nous retrouvons $P_X = X(X'X)^{-1}X'$,matrice de projection orthogonale sur l'espace engendré par les colonnes de X. Les propriétés caractéristiques d'un projecteur orthogonal ($P'_X = P_X$ et $P_X^2 = P_X$) sont vérifiées.

- La denière façon de procéder consiste à écrire que le vecteur $(I - P_X)Y$ est orthogonal à chacune des colonnes de X qui engendre $\Im(X)$:

$$\begin{cases} \langle X_1, Y - X\hat{\beta}\rangle = 0 \\ \vdots \\ \langle X_p, Y - X\hat{\beta}\rangle = 0 \end{cases} \Leftrightarrow X'Y = X'X\hat{\beta}.$$

Soit $P_X = X(X'X)^{-1}X'$ la matrice de projection orthogonale sur $\Im(X)$,la matrice de projection orthogonale sur $\Im(X)^{\perp}$ est $P_{X_{\perp}} = (I - P_X)$.

– Calcul matriciel

Nous pouvons aussi retrouver le résultat précédent de manière analytique en écrivant la fonction à minimiser $S(\beta)$:

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

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

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

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

$$-2X'Y + 2X'\hat{\beta} = 0$$

c'est -à-dire $\hat{\beta} = (X'X)^{-1}X'Y$.
Pour s'assurer que ce point $\hat{\beta}$ est bien un minimum strict, il faut que la dérivée seconde soit une matrice définie positive.Or la dérivée seconde s'écrit

$$\frac{\partial^2 S(\beta)}{\partial \beta^2} = 2X'X,$$

et X est de plein rang donc $X'X$ est inversible et n'a pas de valeur proprenulle.La matrice $X'X$ est donc définie.De plus $\forall z \in \mathbb{R}^{\mathbb{P}}$, nous avons

$$z'2X'Xz = 2\langle Xz, Xz\rangle = 2\|Xz\|^2 \geq 0$$

$(X'X)$ est donc bien définie positive et $\hat{\beta}$ est bien un minimum strict.

Interprétation

Nous venons de voir que $\hat{Y}$ est la projection de Y sur le sous-espace engendré par les colonnes

de X .Cette projection existe et est unique même si l'hypothèse $\mathcal{H}_1$ n'est pas vérifiée . L'hypothèse $\mathcal{H}_1$ nous permet d'obtenir un $\hat{\beta}$ unique ; dans ce cas , s'intéresser aux coordonnées de $\hat{\beta}$ a un sens , et ces coordonnées sont les coordonnées de $\widehat{Y}$ dans le repère $X_1, ..., X_p$.Ce repère n'a aucune raison d'être orthogonal et donc $\hat{\beta}_j$ n'est pas la coordonnée de la projection de Y sur X_j . Nous avons

$$P_X Y = \hat{\beta}_1 X_1 + ... + \hat{\beta}_P x_P.$$

calculons la projection de Y sur X_j.

$$\begin{aligned} P_{X_j} Y &= P_{X_j} P_X Y \\ &= \hat{\beta}_1 P_{X_j} X_1 + ... + \hat{\beta}_p P_{X_j} X_p \\ &= \hat{\beta}_j X_j + \sum_{i \neq j} \hat{\beta}_i P_{X_j} X_i \end{aligned}$$

Cette dernière quantité est différente de $\hat{\beta} X_j$ sauf si X_j est ortogonal à toutes les autres variables

Lorsque toutes les variables sont orthogonales deux à deux, il est clair que $(X'X)$ est une matrice diagonale

$$(X'X) = \deg(\| X_1 \|^2, ..., \| X_p \|^2) \tag{2.1}$$

Quelques propriétés statistiques

Le statisticien cherche à vérifier que les estimateurs des MC que nous avons construits admettent de bonnes propriétés au sens statistique. Dans notre cadre de travail, cela peut se résumer en deux parties :l'estimateur des MC est-il sans biais et est-il de variance minimale dans sa classe d'estimateurs ?

Pour cela , nous supposons une seconde hypothèse notée $\mathcal{H}_2$ indiuant que les erreurs sont centrées, de même variance (homoscédasticité) et non corrélées entre elles.L'écriture de cette hypothèse est $\mathcal{H}_2 : \mathbb{E}(\varepsilon) = 0, \sum_\varepsilon = \sigma^2 I_n$ avec I_n la matrice identité d'ordre n . Cette hypothèse nous permet de calculer

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

L'estimateur des MC est donc sans biais.Calculons sa variance

$$\mathbb{V}(\hat{\beta}) = \mathbb{V}((X'X)^{-1}X'Y) = (X'X)^{-1}X'\mathbb{V}[Y]\,X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

Proposition 2.3 ***($\hat{\beta}$ sans biais)*** *:*
L'estimateur $\hat{\beta}$ des MC est un estimateur sans biais de β et sa variance vaut $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$.

Remarque 2.1 *Lorsque les variables sont orthogonales deux à deux, les composautes de $\hat{\beta}$ ne sont pas corrélées entre elles puisque la matrice $(X'X)$ est diagonale. Le théorème de Gauss-Maekov, nous indique que parmi tous les estimateurs linéaires sans biais de β, l'estimateur obtenu par MC admet la plus petite variance.*

Théorème 2.3 ***(Gauss-Markov)***
L'estimateur $\hat{\beta}$ des MC est optimal parmi les estimateurs linéaires sans biais de β.

2.2.3 Estimateurs du Maximum de Vraisemblance

Nous allons commencer par faire le lien entre l'estimateur du maximum de vraisemblance et l'estimateur des moindres carrés. Commençons par remarquer que les y_i sont eux-mêmes gaussiens :

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \Rightarrow y_i = x_i'\beta + \varepsilon_i \sim \mathcal{N}(x_i'\beta, \sigma^2)$$

et mutuellement indépendants puisque les erreurs ε_i le sont. La vraisemblance s'en déduit :

$$\begin{aligned}\mathcal{L}(\beta_1, \beta_2, \sigma^2) = \prod_{i=1}^{n} f_Y(y_i) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \parallel Y - X\beta \parallel^2\right]\end{aligned}$$

D'où l'on déduit la log-vraisemblance :

$$\log \mathcal{L}(Y, \beta, \sigma^2) = -\tfrac{n}{2}\log 2\pi\sigma^2 - \tfrac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2} \parallel Y - X\beta \parallel^2$$

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

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

Une fois ceci fait, on veut maximiser sur $\mathbb{R}^*_+$ une fonction de la forme $\varphi(x) = a + b\log x + \frac{c}{x}$, ce qui ne pose aucun souci en passant par la dérivée :

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

d'où il vient :

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

On voit donc que l'estimateur $\hat{\sigma}^2_{mv}$ du maximum de vraisemblance est biaisé, mais d'autant moins que le nombre de variables explicatives est petit devant le nombre n d'observations. Nous continuerons à considérer l'estimateur $\hat{\sigma}^2$ des moindres carrés vu au section précédente et nous conserverons aussi la notation adoptée pour les résidus $\hat{\varepsilon}_i$, de sorte que :

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

Chapitre 3

La régression non linéaire et l'estimation non paramétrique

Ce chapitre est consacré à la présentation de la dexième version de la regression qui est la regression non linéaire et les méthodes d'estimation non paramtrique, principalement la méthode à noyau. Cette méthode est très utile lorsqu'on veut d'écrire la relation entre une variable à expliquer Y et une variable explicative X, sans supposer une forme particulière. On discute la qualité de l'estimateur en donnant la consistance et la convergence.

3.1 Modèle non paramétrique

Soit un échantillon aléatoire (par opposition au cadre déterministe) composé des couples $(x_i, y_i); i = 1, ..., n$ où les (x_i) représentent les valeurs observées de la variable explicative X qui est une variable aléatoire réelle de loi donnée par la fonction de répartition F supposé dérivable, de dérivée f (qui est donc la densité de X par rapport à la mesure de Lebesgue sur $\mathbb{R}$), et les y_i représentent celles de la variable dépendante Y (dite aussi la variable d'intérêt). Alors, le modèle de régression non paramétrique univariée est donné par :

$$y_i = r(x_i) + \varepsilon; \quad i = 1, ..., n. \tag{3.1}$$

Où $r(.)$ est une fonction inconnue appelé *fonction de régression* et les ε_i sont les erreurs aléatoires réelles, appelées "bruit ", sous les hypothèses suivantes :

1. $\mathbb{E}(\varepsilon_i) = 0, \forall i$,
2. X et ε, sont indépendants
3. L'homogénieté des variance, i.e, $\text{Var}(\varepsilon_i, \varepsilon_i) = \sigma^2, \forall i$
4. La non auto-corrélation des erreurs, i.e, Cov $(\varepsilon_i, \varepsilon_j) = 0$, pour $i \neq j$

Définition 3.1 *(Modèle non paramétrique)*

Nous pouvons retenir une approche non paramétrique dans laquelle on va estimer la relation entre le niveau moyen de Y et toutes les valeurs réalisées de X. Nous ne supposons aucune

forme spécifique sur la fonction de régression, tel que, $r(x)$ est la moyenne conditionnelle de la courbe de régression, c'est-à -dire :

$$\mathbb{E}\left(Y \mid X = x\right) = r(x) \tag{3.2}$$

L'hypothèse de base est que $(X;Y)$ est un vecteur alééatoire de $\mathbb{R}^2$. Le paramètre fonctionnel du modèle, que nous cherchons à estimer, est la fonction r de régression :

$$r : \mathbb{R} \to \mathbb{R}$$

Remarque 3.1 *Le principal avantage de cette approche est qui elle ne nécessite aucune hypothèse a priori sur la forme du lien entre X et YAvec une approche non paramétrique, on aboutit à :*

1. *Une représentation graphique de la relation entre X. et Y .*
2. *Il n'existe pas de forme analytique de la fonction de lien $r(x)$.*

Tout le problème consiste alors à estimer cette fonction de régression. C'est donc un problème non paramétrique.

3.2 Estimation par la méthode du noyau

La classe des estimateurs non linéaires regroupe la majorité des estimateurs de la régression, comme l'estimation par fonctions splines ou B-splines, l'estimation par des polynômes par morceaux, et par la méthode du noyau. Dans cette section, nous présenterons le célèbre estimateur à noyau de la régression introduit par **Nadaraya et Watson** et quelques unes de ses propriétés essentielles. Sans oublier l'étude de la consistance et la convergence de cet estimateur.

3.2.1 Principe de la méthode

Parmi les estimateurs non-paramétriques, l'estimateur à noyau est, de loin, le plus populaire. Les méthodes à noyaux permettent de trouver des fonctions de décision non linéaires, tout en s'appuyant fondamentalement sur des méthodes linéaires. Grâce à l'utilisation de fonctions noyau, il devient ainsi possible d'avoir le meilleur traitement des problèmes non linéaires. Le problème consiste à estimer la fonction de régression en tous points $x_1, x_2, ..., x_n$. Le principe de la méthode du noyau repose en fait sur des méthodes de lissage, elle donne, pour estimateur de $\mathbb{E}(Y/X = x) = r(x)$, une moyenne pondérée des valeurs y_i pour les i dont le point x_i est proche du point x pour laquelle on veut estimer $r(x)$.

Définition 3.2 *Un estimateur à noyau **Parzen et Rosenblatt** de la densité f est une fonction $\widehat{f}$ définie par :*

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

Où $(h)_{n\geq 1}$ est une suite de réels positifs appelés paramètres de lissage ou largeur de la fenêtre, qui tend vers 0 quand n tend vers l'infini, et K est une densité de probabilité appelée noyau.

Conditions sur le noyau K :

Il y a un certain nombre de conditions qui sont considérées comme usuelles pour les noyaux et qui permettent d'analyser le risque de l'estimateur à noyau K (pour plus de détails voir [1]) :

1. K est bornée ; i.e, $\sup_{u\in\mathbb{R}} \mid K(u) \mid \leq M < \infty$;
2. $\lim_{|u|\mapsto\infty} |u| K(u) = 0$;
3. $\int |K(u)| du < \infty$;
4. $\int K(u) du = 1$;
5. $\forall u \in \mathbb{R}, K(u) = K(-u)$;
6. $\int_{-\infty}^{+\infty} u^2 K(u) du < \infty$;
7. $\int_{\mathbb{R}} uK(u) du = 0$.

quelques exemples des noyaux :

1. *Noyau rectangulaire (uniforme) :*

$$K_1(u) = \begin{cases} 1/2, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

2. *Noyau triangulaire :*

$$K_2(u) = \begin{cases} (1 - |u|), & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

3. *Noyau parabolique ou d'Epanechnikov :*

$$K_3(u) = \begin{cases} \frac{3}{4}(1 - u^2) & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

4. *Noyau "biweight" quadratique :*

$$K_4(u) = \begin{cases} \frac{15}{16}(1 - u^2), & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

5. *Noyau gaussien*

$$K_5(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}); \qquad \forall u \in \mathbb{R}$$

6. *Noyau cubique :*

$$K_6(u) = \begin{cases} \frac{35}{32}(1-u^2)^3, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

3.2.2 Construction de l'estimateur :

Dans cette section, on donne deux méthodes différentes de la construction de l'estimateur à noyau

I- Méthode classique :

Proposition 3.1 *En supposant que la densité f soit continue, l'estimateur à noyau de la densité f, en un point x, est construit en comptant le nombre d'observation dans l'intervalle autour de x qui est donné par $[x-h, x+h]$. Alors, on écrit :*

$$\begin{aligned} \widehat{f}(x) &= \tfrac{1}{2nh} \textstyle\sum_{i=1}^{n} \mathbf{1}_{[x-h;x+h]}(x_i) \\ &= \tfrac{1}{nh} \textstyle\sum_{i=1}^{n} K\left(\tfrac{x-x_i}{h}\right) \end{aligned}$$

Preuve. On sait que :$x_i \in [x-h, x+h]$ donc :

$$\begin{aligned} x - h \leq x_i \leq x + h \\ 1 \geq \frac{x - x_i}{h} \geq -1 \end{aligned} \tag{3.3}$$

On fait un changement de variable :

$$\begin{aligned} \hat{f}(x) &= \frac{1}{2nh} \sum_{i=1}^{n} 1_{[x-h,x+h]}(x_i) \\ &= \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} 1_{[x-h,x+h]}(x_i) \\ &= \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} 1_{[-1,+1]}\left(\frac{x-x_i}{h}\right) \\ &= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \end{aligned}$$

■

Où $K(u) = \frac{1}{2}1_{[-1,+1]}(u)$

Nous reprenons le modèle

$$r(x) = \mathbb{E}(Y/X = x) = \frac{g(x)}{f(x)} = \frac{\int y f_{X,Y}(x,y)dy}{\int f_{X,Y}(x,y)dy} \tag{3.4}$$

où $f_{X,Y}(.,.)$ est la densité jointe sur $\mathbb{R}^2$ et nous désignons par $f(x)$ la densité marginale de X (par rapport à la mesure de Lebesgue sur $\mathbb{R}$), aussi :

$$\begin{aligned} g(x) &= \textstyle\int_{\mathbb{R}} y f_{X,Y}(x,y)dy \\ &= \lim_{h\to 0} \frac{1}{2h} \textstyle\int_{x-h}^{x+h} \int_{\mathbb{R}} y F_{X,Y}(x,y)(dx,dy) \\ &= \lim_{h\to\infty} \frac{1}{2h} \mathbb{E}[Y 1_{(|x_i - x|\leq h)}] \end{aligned}$$

où $F_{X,Y}(.,.)$ est la fonction de répartition du (X,Y)

$$\hat{g}(x) = \frac{1}{n} \textstyle\sum_{i=1}^{n} y_i \frac{1_{(|x_i - x|\leq h)}}{2h}$$

donc $r(x)$ est estimé par

$$\hat{r}(x) = \frac{\hat{g}(x)}{\hat{f}(x)} = \frac{\sum_{i=1}^{n} y_i \frac{\mathbf{1}_{(|x_i - x|\leq h)}}{2h}}{\sum_{i=1}^{n} \frac{1_{(|x_i - x|\leq h)}}{2h}}$$

Cet estimateur se présente sous la forme d'une moyenne locale pondérée des valeurs y_i , mais il présente le désavantage d'être discontinu. Sa généralisation naturelle est l'estimateur à noyau, défini comme suit :

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)} \mathbf{1}_{\sum_{i=1}^{n} K(\frac{x_i - x}{h})\neq 0}, \forall x \tag{3.5}$$

tel que,

$$\hat{g}(x) = \frac{1}{nh} \textstyle\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right)$$

l'équation (3.5) est l'estimateur de Nadarya Watson [N-W].
où **1**. désigne la fonction indicatrice. On rappelle que, pour tout évènement $\mathcal{A}$ Borel-mesurable on a,

$$1(\mathcal{A}) := \begin{cases} 1, & si \ \ \mathcal{A} \ \ est \ \ vérifié, \\ 0, & sinon \end{cases}$$

De manière similaire, nous pouvons définir l'estimateur Nadarya Watson par :

$$\hat{r}_n^{NW}(x) := \begin{cases} \dfrac{\sum_{i=1}^n y_i K(\frac{x_i - x}{h})}{\sum_{i=1}^n K(\frac{x_i - x}{h})}, & \text{lorsque } \sum_{i=1}^n K(\frac{x_i - x}{h}) \neq 0 \quad \text{est vérifié,} \\ \dfrac{1}{n}\sum_{i=1}^n y_i, & \text{sinon} \end{cases}$$

Le noyau K détermine la forme du voisinage autour du point x et la fenètre h contrôle la taille de ce voisinage, c'est à dire le nombre d'observations prises pour effectuer la moyenne locale. Intuitivement, il est naturel que la fenètre h soit prépondérante pour la consistance de l'estimateur [N-W]. Si on pose

$$W_i(x) = \frac{K\left(\dfrac{x_i - x}{h}\right)}{\sum_{i=1}^n K\left(\dfrac{x_i - x}{h}\right)}$$

où $K(.)$ désigne une fonction noyau, $h > 0$ un paramètre de lissage (bandwidth parameter), alors, on peut réécrire l'équation (3.5) comme suit :

$$\hat{r}(x) = \sum_{i=1}^n y_i W_i(x),$$

dans ce cas l'estimateur $\hat{r}(x)$ de $r(x)$ est dit estimateur linéaire de la fonction régression non-paramétrique, et $W_i(x)$ appelée fonction de poids

II- Optimisation de l'espérance conditionnelle :

On veut rappeler la relation suivante :

$$\hat{r}(x) = \arg\min_a \mathbb{E}\left[(Y-a)^2 | X = x\right] \tag{3.6}$$

le minima de cette égalité est trouvée en différenciant l'espérance : $\mathbb{E}[(Y-a)^2|X=x]$ par rapport à a, en égalant le résultat à 0 et, finalement, en isolant a :

En effet

$$\begin{aligned}\frac{\partial}{\partial a}\mathbb{E}[(Y-a)^2|X=x] &= -2\mathbb{E}[(Y-a)|X=x] \\ &= -2\mathbb{E}[Y|X=x]+2a \\ &= 0 \\ &\Rightarrow a=\mathbb{E}(Y|X=x)\end{aligned}$$

Le fait que la dérivée seconde, soit positive mène à la conclusion que cette valeur a est bien un minimum, et non un maximum. Ce lien entre la fonction de régression $r(x)$ et l'optimisation d'une espérance conditionnelle sera exploité à maintes reprises au cours de ce mémoire.

L'estimateur est alors, donné par :

$$\begin{aligned}\hat{r}(x_0) &= \arg\min_a \mathbb{E}[(Y-a)^2|X=x_0] \\ &= \arg\min_a \sum_{i=1}^{n} \frac{(Y_i-a)^2 K(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n} K(\frac{x_i-x_0}{h})} \\ &= \arg\min_a W_i(x_0)(Y-a)^2\end{aligned}$$

De manière analogue à la démonstration de la proposition (12) , il est possible de trouver l'estimateur du noyau. En effet, cet estimateur est obtenu en dérivant la version empirique $\mathbb{E}[(Y-a)^2|X=x]$ par rapport à a.

On pose

$$m(x_0) = \sum_{i=1}^{n} \frac{(Y-a)^2 K(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n} K(\frac{x_i-x_0}{h})}$$

Alors :

$$\begin{aligned}
\frac{\partial m}{\partial a}(x_0) &= \frac{\partial}{\partial a}\sum_{i=1}^{n}\frac{(Y-a)^2K(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})} \\
&= -2\sum_{i=1}^{n}\frac{(Y-a)K(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})} \\
&= -2\sum_{i=1}^{n}\frac{Y_iK(\frac{x_i-x_0}{h})-aK(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})} \\
&= -2\sum_{i=1}^{n}\frac{Y_iK(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})} + 2a\sum_{i=1}^{n}\frac{K(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})} \\
&= 0 \\
\Rightarrow a &= \frac{\sum_{i=1}^{n}Y_iK(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})}
\end{aligned}$$

La dérivée seconde $\frac{\partial^2 m}{\partial a^2}(x_0) = 2\sum_{i=1}^{n}\frac{K(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})}$ est positive. Certitude que cette valeur a est bien une valeur minimale.

Remarque 3.2 *1. $W_i(x)$ représente des poids dépendant de x et de $(x_1, ..., x_n)$. La valeur de $W_i(x)$ dépend du type de l'estimateur considéré et ne dépend pas des observations y_i.*

2. Généralement, plus les points x_i sont proches de x_0 ; plus le poids sera important : $W(x_0)$ est donc décroissante dans la distance $|x_0 - x_i|$

3. Pour tout x, les poids $W(x_i)$ satisfont àă la relation suivante :

$$\sum_{i=1}^{n} W(x_i) = 1 \tag{3.7}$$

3.3 Biais et variance de l'estimateur

3.3.1 Biais de l'estimateur

Le traitement du biais est purement analytique et repose essentiellement sur le développement de Taylor. Il nous faut supposer certaines conditions de régularités sur les fonctions

$g(.)$ et $f(.)$ qui détermineront l'ordre du biais asymptotique en fonction du paramètre de lissage h [1] .

Proposition 3.2 *Supposons que $g(.)$ et $f(.)$ sont de classe $C^2(\mathbb{R})$ et que le noyau K est d'ordre 2, i.e, tel que : les conditions $2, 4, 6$ de noyau K sont vérifiées. Nous avons alors,lorsque $h \to 0$ et $nh \to \infty$*

$$\begin{aligned}\mathbb{B}iais(\hat{r}(x)) &= \mathbb{E}[\hat{r}(x) - r(x)] \\ &= \frac{h^2}{2}(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}) \int u^2 K(u)du + o(h^2)\end{aligned}$$

Preuve.

$$\begin{aligned}\mathbb{B}iais(\hat{r}(x)) &= \mathbb{E}(\hat{r}(x)) - r(x) \\ &= \mathbb{E}\left[\left(K\left(\frac{x-X}{h}\right)\right)\right]^{-1}\left[\int \frac{1}{h}K\left(\frac{x-t}{h}\right)g(t)dt - g(x) + g(x) - r(x)\int \frac{1}{h}K\left(\frac{x-t}{h}\right)f(t)\right. \\ &\backsimeq \frac{1}{h^2}[f(x)]^{-1}\left[g''(x) - r(x)f''(x)\right]\int_{\mathbb{R}} u^2 k(u)du \\ &= \frac{1}{h^2}\left[r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right]\int_{\mathbb{R}} u^2 K(u)du\end{aligned}$$

■

3.3.2 Variance de l'estimateur

Les hypothèses (1),(3) de noyau sont supposés vérifier le fait que $K(.)$ soit de carré intégrable. Nous posons, par convenance que :

$$\sigma^2(x) = \mathbb{V}ar[Y|X = x],$$

lorsque cette expression est bien définie, on a :

Proposition 3.3 *On suppose que $E[Y^2] \leq \infty$. A chaque point de continuité des fonctions $g(x), f(x)$ et $\sigma^2(x)$, tel que $f(x) > 0$,*

$$\begin{aligned}\mathbb{V}ar(\hat{r}(x)) &= \mathbb{E}[(\hat{r}(x)) - \mathbb{E}(\hat{r}(x))^2] \\ &= \frac{1}{nh}\frac{\sigma^2(x)}{f(x)}\int K^2(u)du + o(\frac{1}{h})\end{aligned}$$

Preuve. En utilisant le lemme de Bochner on obtenons aisément :

$$\begin{aligned}\mathbb{V}ar[\hat{f}_n(x)] &= \frac{1}{nh^2}\{\mathbb{E}\left[Y^2K^2(\left(\frac{x-X}{h}\right)\right] - \left(\mathbb{E}\left[YK\left(\frac{x-X}{h}\right)\right]\right)^2\} \\ &= \frac{1}{nh}\int_{\mathbb{R}} K^2(u)s(x-hu)du - h\left[\int_{\mathbb{R}} K(u)r(x-hu)du\right]^2 \\ &= \frac{1}{nh}s(x)\int_{\mathbb{R}} K^2 du(1+o(1))\end{aligned}$$

Amroun

De même,

$$\mathbb{E}[\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\hat{g}(x) - \mathbb{E}[\hat{g}(x)]] = \frac{1}{nh} g(x) \int_{\mathbb{R}} K(u)^2(u)du(1 + o(1))$$

Soit le vecteur

$$A(x) = \begin{pmatrix} \hat{f}(x) \\ \hat{g}(x) \end{pmatrix}$$

et $\sum[A(x)]$ sa matrice de variance covariance. Il s'ensuite :

$$\sum[A_n(x)] = \frac{1}{nh} \begin{pmatrix} f(x) & g(x) \\ g(x) & s(x) \end{pmatrix} \int_{\mathbb{R}} K^2(u)du(1 + o(1))$$

En remarquant que,

$$\begin{pmatrix} -\dfrac{g(x)}{f(x)^2} & \dfrac{1}{f(x)} \end{pmatrix} \begin{pmatrix} f(x) & g(x) \\ g(x) & s(x) \end{pmatrix} \begin{pmatrix} -\dfrac{g(x)}{\{f(x)\}^2} \\ \dfrac{1}{f(x)} \end{pmatrix} \frac{s(x)}{\{f(x)\}^2} - \frac{\{g(x)\}^2}{\{f(x)\}^3}$$

On obtient alors,

$$\begin{aligned} \mathbb{V}r(\hat{r}(x)) &= \frac{1}{nh}[\frac{s(x)}{\{f(x)\}^2 - \dfrac{\{r(x)\}^2}{\{f(x)\}^3}}] \int_R K^2(u)du(1 + o(1)) \\ &= \frac{1}{nh}[\frac{\sigma^2(x)}{\{f(x)\}} K^2(u)du](1 + o(1)) \end{aligned}$$

■

Remarque 3.3 $o(1)$ *dans la démonstration précédente vérifiée l'égalité suivante :*

$$o(1) = o(h) + o((nh)^{-1})$$

3.4 Consistance

La consistance est la propriété la plus importante d'un estimateur. On distingue une consistance forte et aussi une consistance faible. L'estimateur à noyau de la fonction de régression dépend de deux paramètres : la fenêtre h et le noyau k. Le noyau K établit l'aspect du voisinage de x et h contrôle la taille de ce voisinage, donc h est le paramètre prédominant pour avoir de bonnes propriétés asymptotiques, néanmoins le noyau K ne doit pas être négligé en se basant alors sur l'étude du biais, et de la variance

3.4.1 Consistance faible

Théorème 3.1 *Supposons que le noyau K vérifier les hypothèses (1,6), que $E(Y^2) < \infty$ et que : $f(x)$ est strictement positive. Si $h \longrightarrow 0$ et $nh \longrightarrow +\infty$ (quand $n \longrightarrow \infty$), alors $\widehat{r}(x)$ est un estimateur consistant de $r(x)$.*

Preuve. Nous déduisons du théorème de Bochner ((Voir Annexe) que ,lorsque $h \longrightarrow 0$

$$\begin{aligned}
\mathbb{E}[\hat{f}(x)] &= \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\right] \\
&= \frac{1}{h}\mathbb{E}\left[K\left(\frac{x-X}{h}\right)\right] \\
&= \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-t}{h}\right) f(t)dt \to f(x)\int_{R} K(t)dt \\
&= f(x)
\end{aligned}$$

Nous constatons que le biais de l'estimateur converge vers zéro quand la fenêtre tend vers zéro, de plus, on constate qu'il ne dépend pas du nombre de variables, il dépend surtout du noyau K . Donc $\hat{f}$ est un estimateur asymptotiquement sans biais. D'autre part, comme les x_i sont indépendantes et identiquement distribuées, il vient que :

$$\begin{aligned}
\mathbb{V}ar(\hat{f}(x)) &= \frac{1}{n}\mathbb{V}ar\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right] \\
&\leq \frac{1}{n}\mathbb{E}\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right]^2 \\
&= \frac{1}{h}\int_{-\infty}^{+\infty}\frac{1}{h}K^2\left(\frac{x-t}{h}\right) f_X(t)dt
\end{aligned}$$

D'après le théorème de Bochner :
$\int_{-\infty}^{+\infty}\frac{1}{h}K^2(\frac{x-t}{h})f(t)dt \longrightarrow f(t)\int_R k^2(t)dt < \infty$,quand $n \to \infty$. ■

Donc

$$\mathbb{V}ar(\hat{f}(x)) \to 0 \text{ quand } nh \to \infty.$$

On obtient alors que $\hat{f}(x)$ est un estimateur consistant de $f(x)$, il suffit maintenant de montrer que $\hat{g}(x)$ est un estimateur consistant aussi de

$$g(x) = \int_{-\infty}^{+\infty} y f_{X,Y}\,(x,y)\,dy.$$

Preuve. Nous avons

$$\begin{aligned}
\mathbb{E}[\hat{g}(x)] &= \mathbb{E}[\frac{1}{nh}\sum_{i=1}^{n} Y_i K(\frac{x-x_i}{h})] \\
&= \frac{1}{h}\mathbb{E}[YK(\frac{x-X}{h})] \\
&= \frac{1}{h}\int_R \mathbb{E}(Y/X=x)K(\frac{x-t}{h})f(t)dt \\
&= \frac{1}{h}\int_{\mathbb{R}} r(t)K(\frac{x-t}{h})f(t)dt \\
&= \frac{1}{h}\int_{\mathbb{R}} K(\frac{x-t}{h})r(t)f(t)dt \\
&= \frac{1}{h}\int_{\mathbb{R}} K(\frac{x-t}{h})g(t)dt \longrightarrow g(x)
\end{aligned}$$

Par le théorème de Bochner. De plus , si $s(x) = \int y^2 f_{X,Y}\,(x,y)dy$
$\mathbb{V}ar(\hat{g}(x)) = \mathbb{E}[\hat{g}(x)]^2 - \mathbb{E}^2[\hat{g}(x)] \simeq \frac{1}{h}s(x)\int_{\mathbb{R}} K^2(t)dt \longrightarrow 0$ quand $n \longrightarrow \infty$
donc

$$\lim_{n\to\infty} \mathbb{V}ar(\hat{g}(x)) = 0$$

d'où $\hat{r}(x) = \frac{\hat{g}(x)}{\hat{f}(x)} \to \frac{g(x)}{f(x)} = r(x)$ en probabilité

■

3.4.2 Absence de biais asymptotique

Théorème 3.2 *Sous les conditions (1,6) et si $f(x)$ est strictement positive , il vient que :*

a *Lorsque Y est bornée p.s $\;,h \longrightarrow 0\;$ et $nh \longrightarrow \infty (quand\ n \longrightarrow \infty)$ alors*

$$\mathbb{E}(\hat{r}) = \mathbb{E}[\hat{g}(x)]/\mathbb{E}[\hat{f}(x)] + o((nh)^{-1}) \tag{3.8}$$

b *Lorsque $\mathbb{E}[Y^2] < \infty, h \longrightarrow 0 etnh^2 \longrightarrow \infty (quand\ n \longrightarrow \infty)$ alors :*

$$\mathbb{E}(\hat{r}(x)) = \mathbb{E}[\hat{g}(x)]/\mathbb{E}[\hat{f}(x)] + o((n^{\frac{1}{2}}h)^{-1} \tag{3.9}$$

(a),(b) et le théorème de Bochner impliquent que $\hat{r}(x)$ est un estimateur asymptotiquement sans biais de $r(x)$

Preuve. En utilisant l'identité suivante

$$\frac{1}{\hat{f}(x)} = \frac{1}{\mathbb{E}[\hat{f}(x)]} - \frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\{\mathbb{E}[\hat{f}(x)]\}^2} + \frac{\{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\}^2}{\hat{f}(x)\{\mathbb{E}[\hat{f}(x)]\}^2}$$

On multiplie par $\hat{g}(x)$ des deux côtés, puis on passe à l'espérance

$$\begin{aligned}\mathbb{E}[\hat{r}(x)] &= \frac{\mathbb{E}[\hat{g}(x)]}{\mathbb{E}[\hat{f}(x)]} - \mathbb{E}\left[\hat{g}(x)\left(\frac{\hat{f}(x)-\mathbb{E}[\hat{f}(x)]}{\{\mathbb{E}[\hat{f}(x)]\}^2}\right)\right] \\ &\quad + \mathbb{E}\left[\hat{g}(x)\frac{\{\hat{f}(x)-\mathbb{E}[\hat{f}(x)]\}^2}{\hat{f}(x)\{\mathbb{E}[\hat{f}(x)]\}^2}\right] \\ &= \frac{\mathbb{E}[\hat{g}(x)]}{\mathbb{E}[\hat{f}(x)]} - \frac{a_n(x)+b_n(x)}{\{E[\hat{f}(x)]\}^2}\end{aligned}$$

tel que

$$a_n(x) = [\mathbb{E}(\hat{g}(x)) - \mathbb{E}(\hat{g}(x))] - [\hat{f}(x) - \mathbb{E}(\hat{f}(x))];$$

et

$$b_n(x) = \mathbb{E}[(\hat{f}(x))^{-1}\hat{g}_n(x)(\hat{f}(x)) - \mathbb{E}(\hat{f}(x))^2]$$

. Soit $s(x) = \int_{\mathbb{R}} y^2 f_{X,Y}(x,y)dy$. Nous calculons la variance asymptotique de $\hat{g}(x)$ puis $\hat{f}(x)$ via le théorème de Bochner

$$\begin{aligned}\mathbb{V}ar[\hat{g}(x)] &= \frac{1}{nh}\int_R K^2(u)f(x-uh)du - \frac{1}{n}\{\int_{\mathbb{R}} K(u)f(x-uh)du\}^2 \\ &\approx \frac{1}{nh}s(x)\int_R K^2 du \\ \mathbb{V}ar[\hat{f}(x)] &= \frac{1}{nh}\int_{\mathbb{R}} K^2(u)f(x-uh)du - \frac{1}{n}\{\int_R K(u)f(x-uh)du\}^2 \\ &\approx \frac{1}{nh}f(x)\int_{\mathbb{R}} K^2(u)du\end{aligned}$$

En utilisant l'inégalité de Cauchy-Schwartz (Voir Annexe) combinée aux formules ci-dessus, on obtient

$$a_n(x) = o((nh)^{-1}) \tag{3.10}$$

Lorsque la variable Y est bornée, i.e $|y| \leqslant M$ pour une certaine constante M fixée, nous remarquons que l'estimateur de [N-W] est lui aussi naturellement borné,

$$\frac{\hat{g}(x)}{\hat{f}(x)} = \frac{\sum_1^n y_i K(\frac{x-x_i}{h})}{\sum_1^n K(\frac{x-x_i}{h})} \leqslant \frac{\sum_1^n MK(\frac{x-x_i}{h})}{\sum_1^n K(\frac{x-x_i}{h})} = M \tag{3.11}$$

Cette dernière inégalité permet de borner $b_n(x)$

$$b_n(x) \leqslant M \times \mathbb{E}[\{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\}^2] \approx \frac{M}{nh}f(x)\int_{\mathbb{R}} K^2(u)du = o((nh)^{-1}) \tag{3.12}$$

Les relation (3.10) et (3.12) entraînent (3.8).Pour démontrer le cas b ; il suffit de remarquer que la relation (3.10) est toujours valable mais la relation (3.12) devient

$$\begin{aligned}|b_n(x)| &\leqslant \mathbb{E}[\max_{1\leq i\leq n}|Y_i|\,\{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\}^2] \\ &\leqslant \{\sum_1^n Y_i^2\}^{\frac{1}{2}} \times \{\mathbb{E}[\{\hat{f}(x) - E[\hat{f}(x)]\}^4]\}^{\frac{1}{2}} \\ &= \sqrt{n}\{\mathbb{E}[Y_i^2]\}^{\frac{1}{2}} \times o((nh)^{-1})\end{aligned}$$

$$= o((n^{\frac{1}{2}}h)^{-1}) \tag{3.13}$$

Les relations (3.10) et (3.13) impliquent (3.9), d'où le résultat énoncé

■

Remarque 3.4

1- *Dans l'estimation de r(x) par le noyau rectangulaire, le même poids est accordé à toutes les observations comprise entre $x-h$ et $x+h$. Dans les autres noyaux , le poids d'une observation est d'autant plus fort qu'elle est proche de x.*

2- *$\widehat{r}$ a les mêmes propriétés de continuité et de différentiabilité que K. Par exemple, si K est le noyau gaussien $\widehat{f}$ admet des dérivées de tous ordres*

3- *Pour choisir quel noyau prendre et surtout choisir le paramètre de lissage h, il faut étudier la qualité de l'estimation de r par $\widehat{r}$*

Exemple 3.1 *Dans le cas d'une fonction noya Epanechnikov,on a :*
$\begin{cases} \frac{3}{4}(1-u^2), & \mid u \mid \leq 1 \\ 0, & \mid u \mid > 1 \end{cases}$
avec $u = (x_i - x_0)/h$.Donc si $|x_i - x_0| > h, alors\ u \in\ ]-\infty, 1[\cup]1, +\infty[\,, K(u) = 0$ et par conséquent $W_i(x_0) = 0$

Les vitesses de convergence dépendent de deux paramètres : la fonction de noyau K dont l'efficacité est peu influente et le paramètre de lissage h, dont le choix est crucial aussi bien pour l'approche ponctuelle que pour la globale que nous exposons ci après.

1. $f(.)$ *et* $r(.)$ *de classe* $C^2(\mathbb{R})$
2. $\lim\limits_{n\to\infty} h = 0$,
3. $|Y|$ *est bornée*,
4. K satisfait les propriétés (1,4,6,7) :

3.4.3 Consistance forte

Parzen a établi la normalité asymptotique, ainsi que la convergence uniforme en probabilité. Son travail est un outil important et été largement développé par plusieurs chercheurs (Devroye et Györfi (1985)[4] , Silverman (1986) [5]). En (1976) Nadaraya [6] a énoncé le théorème de la consistance forte de l'estimateur.

3.5 Choix du noyau et paramètre de lissage

Comme nous vous le disons l'estimateur r_n dépend de deux paramètres : le noyau K et la largeur de la fenêtre h .

3.5.1 Etude de critère d'erreur quadratique moyenne de $r_n(x)$

L'erreur quadratique moyenne EQM (en anglais :mean squared error MSE) est une mesure permettant d'évaluer la similarité de r_n par rapport à la fonction de régression inconnue r, au point x. Notre but est de minimiser

$$MSE(r_n(x)) = \mathbb{E}[r_n(x) - r(x)]^2$$

Le développement de cette expression faite précédèmment, nous donne

$$MSE(r_n(x)) = var[r_n(x)] + [biais(r_n(x))]^2$$

Nous constatons d'une part que les expressions du biais de $r_n(x)$ et de la variance de $r_n(x)$

$$MSE(r_n(x)) = \frac{h_n^4}{4}\left[(r''(x) + 2r'(x)\frac{f'_X(x)}{f_X(x)}(u^2K(u)) + 0(1)\right]^2 + \frac{1}{nh_n}(\frac{\sigma^2(x)}{f_X(x)}[K^2(u)] + (1 + 0(1)))$$

où

$[u^2K^q(u)] = \int t^pK^q(t)dt$

Pour trouver donc un compromis entre le biais et la variance nous minimisions par rapport à h_n l'expression de l'erreur quadratique moyenne asymptotique AMSE (asymptotic mean square error) donnée par :

$$AMSE[r_n(x)] = \frac{h_n^4}{4}\left\{r''(x) + 2r'(x)\frac{f_X(x)}{f_X(x)}\right\}^2 [u^2K(u)]^2 + \frac{1}{nh_n}(\frac{\sigma^2(x)}{f_X(x)}[K^2(u)])$$

Comme AMSE est fonction convexe la fenêtre $h^{MSE}_{opt(r_n(x))} = \arg min_h(AMSEr_n(x))$ est solution de l'équation suivante :

$$\frac{\partial}{\partial h_n}\left[\frac{h_n^4}{4}\left\{r''(x) + 2r'(x)\frac{f'_X(x)}{f_X(x)}\right\}^2 [u^2K(u)]^2 + \frac{1}{nh_n}(\frac{\sigma^2(x)}{f_X}(x)[K^2(u)]\right] = 0$$

lorsque $[r''(x) + 2r'(x)\frac{f'_X(x)}{f_X(x)}]^2[u^2K(u)] \neq 0$

d'où

$$h^{MSE}_{opt(r_n(x))} = n^{-1/5}\left\{\frac{\frac{\sigma^2(x)[K^2(x)]}{f_X(x)}}{\left\{\{r''(x)+2r'(x)\frac{f'_X(x)}{f_X(x)}\}^2[tK]^2\right\}}\right\}^{1/5}$$

MISE (mean integrated squared error)

$$MISE[r_n(x)] = \mathbb{E}\left[\int_{\mathbb{R}}(r_n(x) - r(x))^2 dx\right]$$

En appliquant le théorème de Fubini, on a

$$MISE[r_n(x)] = \left[\int_{\mathbb{R}} E(r_n(x) - r(x))^2 dx\right]$$

Sous les mêmes hypothèses que les propositions (2.2.1) et (2.2.3),on a

$$AMISE[r_n(x)] = \frac{h_n^4}{4}\int\left\{r''(x) + 2r'(x)\frac{f'_X(x)}{f_X(x)}\right\}^2 dx[u^2K(u)] + \frac{1}{nh_n}\int\frac{\sigma^2(x)}{f_X(x)}dx[K^2(u)]$$

Théorème 3.3 ***(AMISE sous condition de continuité)***
Supposons que :

$$\exists\beta > 0 \quad telle \quad que \quad \inf_{x\in C} f(x) > \beta$$

et $h \to 0, \quad nh \to \infty$ *et K est borné , intégrable , positif ,symétrique et a support compact on a :*

$$AMISE[r_n(x)] \longrightarrow 0$$

la fenêtre $h^{MISE}_{opt(r_n(x))}$ *minimisant l' AMISE du critère global est :*

$$h^{MISE}_{opt(r_n(x))} = n^{-1/5}\left\{\frac{\int\frac{\sigma^2(x)}{f_X(x)}[K^2(x)]dx}{\int\left\{r''(x) + 2r'(x)\frac{f'_X(x)}{f_X(x)}\right\}^2 dx[tK]^2}\right\}^{1/5}$$

Un travail similaire se fait pour le choix optimum du paramètre de lissage dans le cas de l'estimateur de Parzen-Rosenblatt,nous obtenons

$$h^{MSE}_{opt(f_n(x))} = n^{-1/5}\left\{\frac{f_X(x)[K^2]}{\{\{f''(x)\}^2[t^2K]^2\}}\right\}^{1/5}$$

$$h^{MISE}_{opt(f_n(x))} = n^{-1/5}\left\{\frac{[K^2]}{\int_{\mathbb{R}}(f''_X(x))^2 dx[t^2K]^2}\right\}^{1/5}$$

Nous notons que l'expression de h_n *optimal , minimisant asymptotiquement les quatre critères d'erreurs la forme*

$$h_{opt} = Cn^{-1/5}$$

où la constante C est en fonction de la distribution et de termes aléatoires inconnues

CONCLUSION

Ce mémoire porte sur l'étude des estimateurs de la fonction de régression paramétrique et non paramétrique. On a présenté deux méthodes permettant d'effectuer de la régression paramétrique : par moindre carrée et le maximum de vraisemblance, et l'estimateur à noyau pour l'approche non paramétrique qui dépend de deux paramètres le noyau K et le paramètre de lissage h. À travers les résultats obtenus, nous concluons que : le noyau K est peu influent sur l'estimateur, par contre le paramètre h joue un rôle très important et son choix est crucial. On trouve plusieurs méthodes pour le choix de h optimale et nous avons choisi celles qui sont plus utilisées : celle minimisant les erreurs MSE et MISE.

On conclue que ces méthodes numérique sont développé de tel sortes, qu'on peut toujours trouver le meilleur estimateur selon les données à main.

ANNEXE

LES THÉORÈME UTILISÈS :

A.1 Théorème de Bochner

Soit $K : (\mathbb{R}^m, \mathcal{B}^m) \to (\mathbb{R}, \mathcal{B})$ une fonction mesurable, où $\mathcal{B}^p$ est la tribu borélienne de $\mathbb{R}^p$, vérifiant : $\exists\ M$ (constante) telle-que,

$$\forall z \in \mathbb{R}^m,$$
$$|K(z)| \leq M \Longrightarrow \int_{\mathbb{R}^m} |K(z)|\, dz < \infty.$$

Et

$$\|z\|\, |K(z)| \to 0 \text{ quand } \|z\| \to \infty.$$

Par ailleurs, soit $g :(\mathbb{R}^m, \mathcal{B}^m) \to (\mathbb{R}, \mathcal{B})$ une fonction tq

$$\int_{\mathbb{R}^m} |g(z)|\, dz < \infty.$$

Si g est continue, et si $0 < h \to 0$, quand $n \to \infty$ alors :

$$\lim_{n\to\infty} \tfrac{1}{h^m} \int_{\mathbb{R}^m} K(\tfrac{z}{h}) g(x-z) dz = g(x) \int_{\mathbb{R}^m} K(z) dz.$$

Si g est uniformément continue alors la convergence ci dessus est uniforme.

A.2 Théorème d'Inégalité Cauchy Shwartz

Soit $(H, <.,.>)$ une espace prèhilbertien. Alors

$$|<\mu, \theta>|^2 \leq |<\mu, \theta>|, \ \ \forall \mu, \theta \in H.$$

ou bien

$$|<\mu, \theta>| \leq \sqrt{<\mu, \theta>}.\sqrt{<\mu, \theta>}.$$

A.3 Théorème de Fubini

Si $f(x, y)$ est une fonction Bochner-intégrable sur $X \times Y$, Alors la fonction $\int f dy$ (de x) est déterminé presque partout sur X est Bochner-intégrable sur X . De même, la fonction $\int f dx$ (de y) est déterminé presque partout sur Y . En outre,

$$\int \int f dx dy = \int dy \int f dx = \int dx \int f dy.$$

Bibliographie

[1] Amroun, S, 2011.Sur L'Estimation De La Courbe De Régression De La Moyenne, Mémoire De Magister en Mathématiques Appliquées université de Bêjaïa.

[2] Arnak, S. Statistique avancée : méthodes non paramétriques. École centrale paris de Paris.

[3] Beghriche, H,. L'estimation de la Fonction de Régression, Mémoire de magistère en mathématique université de Mentouri-Constantine.

[4] Devroye, L. and Györfi, L, (1985). Nonparametric Density Estimation. New York.

[5] Silverman, B. W. 1986. Density Estimation for Statistics and Data Analysis. Chap-man Hall, London.

[6] Nadaraya, E. A. 1976. On the nonparametric estimator of Bayesian risk in the classification problem. Proc. AN. Georg SSR, 82(2), 277-280 (in Russian).

[7] Pierre-André Cornillon & Éric Matzner-Løber, Régression Théorie et applications, Springer-Verlag France, Paris, 2007, ISBN-10 : 2-287-39692-6Springer Paris Berlin Heidelberg New York.

[8] Kadi Abir & Zelliche Warda, Analyse de la régression non paramétrique, Mémoire de fin d'études (Université Mohammed Seddik Ben Yahia - Jijel) Département de Mathématique, 2019.

[9] Cours apprentissage non paramétrique en régression. https ://www.math.univ-toulouse.fr/ besse/Wikistat/pdf/st-m-app-non-param.pdf

[10] https ://fr.wikipedia.org/wiki/Hémoglobine_glyquée.