الجمهورية الجزائرية الديمقراطية الشعبية.
République Algérienne démocratique et populaire.

وزارة التعليم العالي والبحث العلمي.
Ministère de l'enseignement supérieur et de la recherche scientifique.

جامعة سعد دحلب البليدة .
Université SAAD DAHLAB de BLIDA.

كلية العلوم.
Faculté des Sciences.

قسم الرياضيات.
Département de Mathématiques.

MEMOIRE DE FIN D'ETUDE

EN VUE DE L'OBTENTION DU DIPLOME DE MASTER

EN MATHEMATIQUE.

OPTION : Modélisation Stochastique et Statistique.

Thème :

Méthodes d'estimation de régressions quantiles, application sur des données climatiques.

Soutenu par :

- **Chentouf Saida Khadidja .**
- **Bouazzouz Amina.**

Encadré par :

- *M^r* **A . RASSOUL.**

Composition de jury :

- *M^r* O.TAMI — Président
- *M^r* R. FRIHI — Examinateur

Année Universitaire 2017-2018.

DEDICACE

Avec les sentiments de la plus profonde humilité,
Je dédie ce modeste travail:
*A ma bien aimée **très chère mère**, symbole de l'amour et d'affection,*
celle qui m'a toujours encouragé.
*A mon **très cher père** qui est à l'origine de ce qui je suis.*
A ces deux êtres, qui tous ce qu'a de la valeur dans ce monde ne peuvent vouloir d'infiniment petit de leurs sacrifices.

*A ma **grand-mère** et mon **grand-père***

A mon cher frère, Rabah.

A ma chère sœur: Khaoula son mari Mohammed.

A mes oncles : Mohammed, Kamel, Moussa, Noureddine.

A mes tante : Fatima Zohra, Nabila, Safia , Salima et son fils Nabil.

A mes cousins

A mes amies

A toute la promotion Master2 MSS 2018.

BOULAZEOUL AMINA

REMERCIEMENTS

Avant tout, je remercie **Allah le tout puissant** de m'avoir donné la force et le courage d'arriver là ;

En témoignage de mon profond sentiment de respect et de reconnaissance, je tiens à présenter mes sincères remerciements à Mr. **Abdelazziz rassoul** d'avoir pris la responsabilité de me diriger dans ce projet et de me conseiller avec bienveillance et comme encadreur infatigable et à qui je dois en grande partie la réussite de ce rapport de stage.

Un remerciement aux membres du jury d'avoir accepté d'examiner ce modeste travail.

Je profite aussi de l'occasion pour remercier tous mes professeurs et enseignants et enseignantes commençant par mes études primaires terminant par mes études supérieures et je leur dis que ce travail n'aurait jamais pu voir le jour sans leurs efforts avec moi.

Enfin MERCI du fond du cœur à mes parents, mes sœurs et mes frères et tout ma famille, merci pour tout le temps que vous avez consacré à m'aider, merci pour votre soutien inconditionnel tout long de ces années d'étude, merci parce que vous m'encouragez et vous permettez ce que je suis aujourd'hui. Je vous dédie ce mémoire.

DEDICACE

Avec les sentiments de la plus profonde humilité,
Je dédie ce modeste travail:

A mon cher mari ***Abdelchakour*** *ainsi que toute ma* ***belle-famille****.*

A ma chère fille ***Meriem****.*

A ma bien aimée très chère ***mère****, symbole de l'amour et d'affection, celle qui m'a toujours encouragé.*

A mon très cher ***père*** *qui est à l'origine de ce qui je suis.*
A ces deux êtres, qui tous ce qu'a de la valeur dans ce monde ne peuvent vouloir d'infiniment petit de leurs sacrifices.

A mon cher frère, ***Abdou*** *et sa femme* ***Azhar****.*

A mes chères sœurs: ***Nafissa*** *(Mohammed, Mehdi ,Adam),*
Fatima*(Rihame,Lamis,Yanis,Djaber),*
Hadjira*(Fatima,Amine Assinette,Roudina) ,*
Ikram *son mari* ***Abdelkarim****.*

A mes cousins

A mes amies

À mon Cousin : ***Brahim***

A toute la promotion Master2 MSS 2018.

CHENTOUF SAIDA KHADIDJA

Table des matières

Table des figures

Liste des tableaux

يوفر الانحدار الكمي نموذجا إحصائيا أكثر اكتمالا من متوسط الانحدار و لديه ألان تطبيقات واسعة الانتشار ، لاسيما في المجال البيئي و المناخي ، و سيكون تطبيق الانحدار الكمي موضوع هذه الدراسة بدءا من الفصل الأول الذي يوضح مقدمة للمتغيرات العشوائية مع تذكير ببعض المفاهيم الأساسية حول قوانين الاحتمالية . الفصل الثاني يركز على الكميات الشرطية و الشرطية مع خصائصها و انحدارها الكمي. يقدم الفصل الثالث عرضا لبرمجيات R (عام ،التركيب و الحزم) بالإضافة إلى الانحدار الكمي تحت R و أخيرا نناقش مثالا يوضح تطبيق الانحدار الكمي لدراسة تأثير المتغيرات المستقلة (المؤشرات المناخية) و المتغير التابع (الترسيب السنوي في منظمة الحضنة ولاية المسيلة).

0.1 Résumé

La régression du quantile (RQ) offre un modèle statistique plus complet que la régression d'espérance et a maintenant des applications très répandues notamment dans le domaine environnemental et climatique, La théorie et l'application de la régression quantile feront l'objet des la présente étude en commençant par l'inférence statistiques sur la régression du quantile et la présentation des différents modèles. La deuxième partie concerne le package quantreg qui contient le code d'exécussion du régression quantile. Enfin, on termine par une application sur l'influence des quatres indices climatiques sur la quantité des précipitation sur le bassin d'ElHodhna situé dans la wilaya de Msila.

0.2 Abstract

The quantile regression (RQ) offers a more complete statistical model than the regression of hope and now has widespread applications especially in the environmental and climatic domain, the theory and application of quantile regression will be the subject of the This study begins with the statistical inference about quantile regression and the presentation of different models. The second part presents the quantum package that contains the execution code of the quantile regression. Finally, it ends with an application of the influence of the four climatic indices on the quantity of precipitation in the Elhodhna basin located in Msila province.

0.3 Notations et Abréviation

i.i.d : indépendante et identiquement distribuée.

$\mathbf{Y}$: La variable aléatoire d'intérêt.

$\mathbf{X}$: Le covariable aléatoire.

$\mathbf{N}$: Taille d'échantillon.

$\mathbf{n}$: Taille d'échantillon observé.

$\mathbf{p}$: La probabilité du quantile.

$\mathbf{Q}_F(\mathbf{p})$:La fonction quantile d'ordre $\mathbf{p}$.

$\mathbf{Q}_{F_n}$: Estimateur de $\mathbf{Q}_F(\mathbf{p})$.

$\mathbf{F}(\mathbf{y}|\mathbf{x})$: Fonction de répartition conditionnelle de $\mathbf{Y}$ sachant $\mathbf{X} = \mathbf{x}$.

$\mathbf{Qp}(\mathbf{x})$: Quantile conditionnel d'ordre $\mathbf{p}$ de $\mathbf{Y}$ sachant $\mathbf{X} = \mathbf{x}$.

$\hat{\mathbf{F}}_n$: Fonction de répartition empirique.

$\hat{\mathbf{F}}_n^{-1}$: Inverse généralisé de la fonction de répartition empirique.

$\hat{\mathbf{F}}_n(\mathbf{y}|\mathbf{x})$: Estimateur de la fonction de répartition conditionnelle.

$\hat{\mathbf{F}}_n^{-1}(\mathbf{y}|\mathbf{x})$: Inverse généralisé de l'estimateur de la fonction de répartition conditionnelle.

$\mathbf{Q}_{p,n}(\mathbf{x})$: Estimateur du quantile conditionnel d'ordre $\mathbf{p}$ de $\mathbf{Y}$ sachant $\mathbf{X} = \mathbf{x}$.

0.4 Introduction

Beaucoup d'études économiques empiriques se concentrent sur l'observation ou la modélisation de la moyenne et cette focalisation est souvent critiquée. La moyenne apporte une information essentielle mais néanmoins limitée.

Une de ces limites du recours à la moyenne est d'ordre plus technique : c'est le fait qu'elle s'avère parfois difficile à modéliser. Cela peut être le cas en présence de valeurs extrêmes ou aberrantes (dues par exemple à des erreurs de mesures), auxquelles la moyenne est bien plus sensible que les quantiles. Lorsque la distribution de la variable d'intérêt est très étalée, ce qui est par exemple le cas des revenus, la moyenne pourra beaucoup varier en fonction de l'échantillon utilisé. L'estimation de la moyenne est également compromise en présence de données censurées, c'est à dire lorsqu'on n'observe la variable d'intérêt qu'au delà ou en deçà d'un seuil fixe.

La régression quantile est l'un des outils dont dispose le statisticien pour répondre à ces limites inhérentes à la moyenne. Elle permet d'avoir une description plus précise de la distribution d'une variable d'intérêt conditionnelle à ses déterminants, comparativement à la régression linéaire qui se focalise sur la moyenne conditionnelle. Si son principe est ancien (La méthode des moindres déviations absolues, qui revient à s'intéresser à la médiane plutôt qu'à la moyenne comme dans les moindres carrés ordinaires, est aussi ancienne que la méthodes des moindres carrés ordinaires et remonte à Boscovitch, Laplace et Gauss.),elle a connu récemment un regain d'intérêt. Un ensemble de procédures préprogrammées en font aujourd'hui un outil simple d'utilisation.

Quelques applications typiques de la régression par quantiles : tableaux de référence médicale, analyse de survie, économie financière, détection de l'hétéroscédasticité et de la modélisation environnementale ou l'étude des données sur la pollution montrent que les modèles des niveaux moyens de concentration peuvent être moins pertinents du point de vue de la santé publique que les modèles comparables pour les quantiles supérieurs représentant des niveaux de concentration plus extrêmes; voir, par exemple, Pandey et Nguyen (1999) et Hendricks et Koenker.(1992).

L'hydrologie s'intéresse ainsi à la modélisation des précipitations et du débit des rives afin d'évaluer les risques de sécheresse pour aider à la conception des réservoirs et les risques de fortes précipitations pour concevoir des drains d'inondation et de ruissellement.

La modélisation des queues de distributions et la connaissance des quantiles extrêmes sont donc au cœur des statistiques hydrologiques.

Les régressions quantiles peuvent être aujourd'hui effectuées aisément

avec de nombreux logiciels statistiques en insistant sur les détails de leur implémentation pratique par les logiciels statistiques standards (R). Une application avec le logiciel R a été faite dans ce travail pour étudier l'impact des indices climatiques sur la précipitation annuelle au niveau de la région géographique EL_HODNA ou les données de précipitations s'étalent sur la période allant de 1979-2013 pour trois stations différentes.

Chapitre 1

Notion de variable aléatoire

1.1 Introduction

Dans de nombreuses expériences aléatoires, on n'est pas intéressé directement par le résultat de l'expérience, mais par une certaine fonction de ce résultat ;considérons par exemple l'expérience qui consiste à observer, pour chacune des n pièces produites par une machine, si la pièce est défectueuse ou non,nous attribuerons la valeur1 à une pièce défectueuse et la valeur 0 à une pièce en bon état. L'univers associé à cette expérience est $\Omega = \{0,1\}^n$.

Ce qui intéresse le fabricant est la proportion de pièces défectueuses produites par la machine. Introduisons donc une fonction de Ω dans $\mathbb{R}$ qui à tout $\omega = (\omega_1, \omega_2, ..., \omega_n)$ de Ω associe le nombre

$$X(\omega) = \sum_{i=1}^{n} \frac{\omega_i}{n}$$

qui correspond à la proportion de pièces défectueuses associée à l'observation de ω. Une telle fonction Xdéfinie sur Ω et à valeurs dans $\mathbb{R}$ s'appelle une variable aléatoire réelle.

1.2 Définitions

Définition 1.1 *Etant donné un univers Ω, une variable aléatoire réelle (v.a.r.) est une application de Ω dans $\mathbb{R}$:*

$$X : \omega \in \Omega \rightarrow X(\omega) \in \mathbb{R}.$$

Définition 1.2 *Soit Ω un univers muni d'une probabilité P, et soit X une v.a.r. On appelle loi de probabilité de X, notée P_X, l'application qui à toute*

partie A de $\mathbb{R}$ associe

$$P_X(A) = P(\{\omega \in \Omega : X(\omega) \in A\}).$$

Remarque 1.1 *Dans la suite du cours, on utilisera la notation abrégée :*

$$P(\{\omega \in \Omega : X(\omega) \in A\}) = P(X \in A).$$

De même, on notera $P(X = x)$ la probabilité $P(\{\omega \in \Omega : X(\omega) \in x\})$.

Définition 1.3 *La fonction de répartition de la v.a.r.X est définie par*

$$F_X(x) = P(X \leq x), x \in \mathbb{R}$$

Propriétés de la fonction de répartition :

1. $0 \leq F_X \leq 1$.
2. F_X *tend vers* 0 *en* $-\infty$ *et vers* 1 *en* $+\infty$.
3. F_X *est croissante.*
4. F_X *est continue à droite.*

Proposition 1.1 *On a l'identité*

$$P(a < X \leq b) = F_X(b) - F_X(a), \quad \forall a < b.$$

Remarque 1.2 *On montre facilement que F_X est continue si et seulement si $P(X = x) = 0$ pour tout $x \in \mathbb{R}$. On parle alors de loi diffuse ou de v.a.r.continue.*

Définition 1.4 *Soit X une v.a.r. de fonction de répartition F_X supposée strictement croissante de $I \subset \mathbb{R}$ dans $]0,1[$. Le quantile d'ordre $\alpha \in]0,1[$ de X est le nombre $x_\alpha \in I$ tel que $F_X(x_\alpha) = \alpha$, ce qui signifie que $P(X \leq x_\alpha) = \alpha$.*

Remarque 1.3

- $x_{\frac{1}{2}}$ *est appelé médiane de X. La médiane vérifie les deux égalités*

$$P(X \leq x_{\frac{1}{2}}) = \frac{1}{2} = P(X > x_{\frac{1}{2}})$$

- *Dans le cas où F_X n'est pas strictement croissante mais simplement croissante, on définit le quantile d'ordre α par x*

$$\alpha = \inf\{x \in \mathbb{R} : F_X(X) \geq \alpha\}.$$

1.3 Variables aléatoires discrètes

Définition 1.5 *Une v.a.r. X à valeurs dans un ensemble χ fini ou dénombrable est appelée v.a.r. discrète. Dans ce cas, la loi de X est déterminée par l'ensemble des probabilités :*

$$P_X(x) = P(X = x), x \in \chi$$

Ainsi, pour toute partie A de X, on a alors :

$$P_X(A) = P(X \in A) = \sum_{x \in A} P(X = x) \text{ et } P_X(\chi) = \sum_{x \in \chi} P(X = x) = 1.$$

Remarque 1.4 *Supposons que l'on observe la durée de vie T d'une ampoule électrique et que cette durée de vieT, exprimée en heures, satisfait pour tout $0 < a < b$,*

$$P(a < T \leq b) = \exp\left(\frac{-a}{100}\right) - \exp\left(\frac{-b}{100}\right).$$

On note X le nombre de périodes complètes de 100 heures que dure l'ampoule. Les valeurs possibles de X etant entières, la v.a.r. X est donc discrète. Calculons la fonction de répartition de X. Comme X est positive, on a

$$F_X(x) = P(X \leq x) = 0, \qquad \forall x < 0$$

De plus, pour tout $n \in \mathbb{N}$

$$P(X = n) = P(100n \leq T < 100(n+1)) = \exp(-n) - \exp(-(n+1))$$

Ainsi, on a donc pour tout $x \geq 0$:

$$P(X \leq x) = \sum_{n=0}^{[x]} P(X = n) = 1 - \exp(-([x]+1)).$$

On notera que la fonction F_X est une fonction en escalier.

1.3.1 Exemples de variables discrètes

Soit X une v.a.r. discrète prenant ses valeurs dans un ensemble $\{x_1, x_2, ..., x_n\}$, éventuellement infini. Alors la loi de X est caractérisée par l'ensemble des probabilités $P(X = x_i)$, c'est-à-dire les nombres réels positifs p_i tels que

$$P(X = x_i) = p_i \quad \text{avec} \quad 0 \leq p_i \leq 1 \text{ et } \quad \sum_{i=1}^{n} p_i = 1.$$

Loi de Bernoulli

On dit qu'une v.a.r. X à valeurs dans $\{0,1\}$ suit une loi de Bernoulli de paramètre $p \in]0,1[$, notée $B(p)$, si

$$P(X=1) = 1 - P(X=0) = p.$$

Par exemple, cette loi intervient lorsque l'on modélise l'état de fonctionnement d'un système. La probabilité que le système fonctionne vaut p et la probabilité que le système ne fonctionne pas vaut $1-p$. Cette loi s'applique aussi aux jeux de hasard de type binaire comme pile ou face

Loi Binomiale

On dit qu'une v.a.r. X à valeurs dans$\{0,1,...,n\}$ suit une loi binomiale de paramètres (n,p), notée $B(n,p)$, si

$$P(X=k) = C_n^k p^k (1-p)^{n-k} \quad , \quad 0 \leq k \leq n.$$

Cette loi intervient par exemple pour modéliser le nombre de pièces défectueuses dans un lot de n pièces, qui ont chacune une probabilité p d'être défectueuse, indépendamment les unes des autres.

Loi Géométrique

On dit qu'une v.a.r. X à valeurs dans $\mathbb{N}^*$ suit une loi géométrique de paramètre $p \in]0,1[$, notée $G(p)$, si

$$P(X=k) = p(1-p)^{k-1} \quad , \quad k \in \mathbb{N}^*$$

Cette loi permet de modéliser le nombre de réalisations indépendantes d'une expérience à 2 issues (succès-échec), jusqu'à l'obtention du premier succès, si à chaque réalisation la probabilité de succès est p.

Loi de Poisson

On dit qu'une v.a.r. X à valeurs dans $\mathbb{N}$ suit une loi de Poisson de paramètre $\lambda > 0$, notée $P(\lambda)$, si

$$P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad , \quad k \in \mathbb{N}.$$

Cette loi intervient comme comportement limite de la loi binomiale lorsque $n \to +\infty$ et $np \to \lambda$.

Elle intervient également pour modéliser des "événements rares". Soit N la variable aléatoire comptant le nombre d'occurrences d'un événement pendant une période donnée T. On suppose qu'un seul événement arrive à la fois, que le nombre d'événement se produisant pendant T ne dépend que de la durée de cette période et que les événements sont indépendants.

Si le nombre moyen d'événements (i.e. accidents) par unité de temps (i.e.semaine) est c, alors on démontre que la probabilité d'obtenir n événements pendant un temps T est :

$$P(N = n) = exp(-cT)\frac{(cT)^n}{n!}.$$

1.4 Variables aléatoires continues

Définition 1.6 *Soit X une v.a.r. qui prend un nombre infini non dénombrable de valeurs. Si F_X est une fonction continue, on dit que X est une v.a.r.continue. Dans ce cas, la loi de X est déterminée par l'ensemble des probabilités $P(a < X < b)$, pour tout $a < b$.*

Remarque 1.5 *Notons que l'on peut mettre $<$ ou $\leq$ dans ce qui précède car la variable étant continue, on a $P(X = x) = 0$ pour tout $x \in \mathbb{R}$. Exemple : Soit $\lambda > 0$. Une v.a.r. X de fonction de répartition*

$$F_X(x) = \begin{cases} 1 - \exp(-\lambda x) & si\ x \geq 0 \\ 0 & si\ \ x < 0 \end{cases}$$

est continue.

Définition 1.7 *Si l'on peut écrire la fonction de répartition d'une variable continue sous la forme*

$$F_X(t) = \int_{-\infty}^{t} f_X(x)dx,$$

où f_X est une fonction de $\mathbb{R}$ dans $\mathbb{R}$, alors on dit que f_X est la densité de probabilité de la v.a.r. X. Ceci implique que l'on a pour tout $a < b$:

$$P(a < X < b) = F_X(b) - F_X(a) = \int_b^a f_X(x)dx.$$

Cette intégrale étant positive pour tout $a < b$, il en résulte que $f_X \geq 0$. De plus, puisque $\lim_{t\to+\infty} F_X(t) = 1$, on a

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1.$$

Une densité de probabilité est donc une fonction positive ou nulle, d'intégrale 1, et qui caractérise la loi d'une v.a.r. continue. De plus, en tout point $x_0 \in \mathbb{R}$ ou F_X est dérivable, on a

$$f_X(x_0) = F'_X(x_0).$$

Exemple 1.1 *Dans l'exemple de la durée de vie T d'une ampoule électrique,T a pour densité de probabilité*

$$f(x) = \begin{cases} \exp(-x/100)/100 & \text{pour tout } x \geq 0 \\ 0 & \text{pourt tout } x < 0 \end{cases}.$$

Enfin, établir que deux v.a.r. (discrètes ou continues) X et Y ont même loi, c'est démontrer que l'on a l'égalité suivante :

$$P(a < X \leq b) = P(a < Y \leq b) \qquad a, b \in \mathbb{R}.$$

Ainsi, en faisant tendre a vers $-\infty$, on obtient le résultat suivant :

Théorème 1.1 *Deux v.a.r. à valeurs dans le même ensemble d'arrivée ont la même loi si et seulement si leurs fonctions de répartition sont égales.*

1.4.1 Exemples de variables continues

Soit X une v.a.r. continue. Alors la loi de$X$ est caractérisée par l'ensemble des probabilités

$$P(a < X < b) = \int_a^b f_X(x)dx,$$

où f_X est la densité de probabilité de X et a et b sont deux nombres réels, éventuellement infinis. Comme nous l'avons vu plus haut, il suffit de connaître cette densité pour connaître la loi de X.

Loi uniforme

La loi uniforme sur un intervalle est la loi des "tirages au hasard" dans cet intervalle. Si $a < b$ sont deux réels, la loi uniforme sur l'intervalle $[a, b]$ est notée $\mathcal{U}([a, b])$. Elle a pour densité :

$$f_X(x) = \frac{1}{b-a} 1_{[a,b]}(x).$$

et la fonction de répartition est donnéer par

$$F(x) = \begin{cases} 0 & pour\ x < a \\ \frac{x-a}{b-a} & pour\ a \leq x < b \\ 1 & pour\ x \geq b \end{cases} .$$

Loi exponentielle

On dit que X suit une loi exponentielle de paramètre $\lambda > 0$, notée $\xi(\lambda)$, si la loi de X a pour densité

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases} ,$$

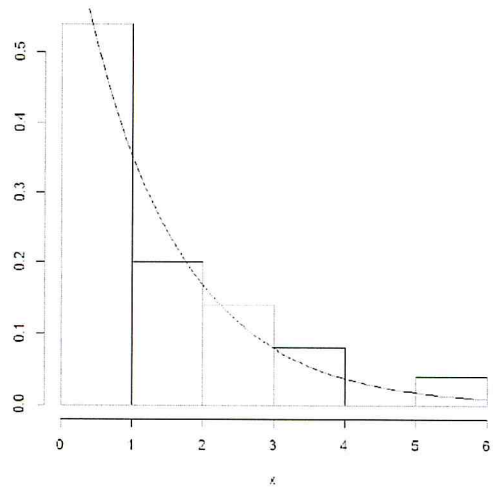


FIG. 1.1 – Histogramme et densité d'une loi exponentielle

La loi exponentielle est utilisée en fiabilité. Le paramètre λ représente le taux moyen de défaillance alors que son inverse $\theta = 1/\lambda$ est "le temps moyen de bon fonctionnement". La loi exponentielle s'applique bien aux matériels électroniques ou aux matériels subissant des défaillances brutales et la fonction de répartition est donnéer par :

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & si\ x \geq 0 \\ 0 & si\ x < 0 \end{cases} .$$

Loi Gamma

Soient $a > 0$ et $\lambda > 0$. On dit que X suit une loi Gamma de paramètres (a, λ), notée $\gamma(a, \lambda)$, si la loi de X a pour densité

$$f_X(x) = \begin{cases} \frac{\lambda^a}{\Gamma(a)} x^{a-1} \exp(-\lambda x) & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Où pour tout $a > 0$, la fonction gamma est donnée par

$$\Gamma(a) = \int_0^{+\infty} x^{a-1} \exp(-x) dx.$$

Le paramètre a est un paramètre de forme alors que le paramètre λ est un paramètre d'échelle. Pour n entier, $a = n/2$ et $\lambda = 1/2$,la loi $G(n/2, 1/2)$ est appelée loi du chi-deux à n degrés de liberté, et notée $\chi^2(n)$. Elle joue un rôle important en statistique, c'est la loi de la somme des carrés de n variables aléatoires indépendantes de loi $\mathcal{N}(0, 1)$. On l'utilise pour les variances empiriques d'échantillons gaussiens. La loi $G(1, \lambda)$ est la loi exponentielle $\xi(\lambda)$.

Loi Normale

Soient $\mu \in \mathbb{R}$ et $\sigma > 0$. On dit que X suit une loi normale de paramètres (μ, σ^2), notée $\mathcal{N}(\mu, \sigma^2)$, si la loi de X a pour densité

$$f_X(x) = \frac{1}{\sigma \sqrt[2]{2\pi}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \quad x \in \mathbb{R}$$

et la fonction de répartition est donnéer par

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt$$

Lois du χ^2, de Student et de Fisher

Par définition, la variable aléatoire, somme des carrés de ν variables indépendantes $\mathcal{N}(0, 1)$ suit une loi du χ^2 à ν degrés de liberté. Deux autres lois jouent des rôles important en statistique. La loi de Student à n degrés de liberté, $T(n)$ est la loi du rapport $X/(\sqrt{Y/n})$, où les variables aléatoires X et Y sont indépendantes , X de loi $\mathcal{N}(0, 1), Y$ de loi $\chi^2(n)$. Elle a pour densité :

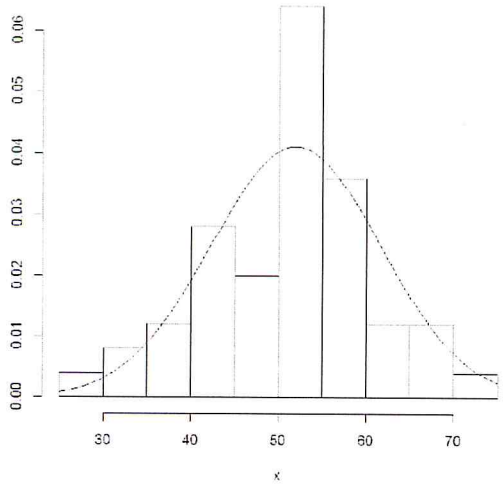


FIG. 1.2 – Histogramme et densité de la loi normale

Elle décrit la distribution de la moyenne empirique d'un échantillon gaussien.La loi de Fisher de paramètres n et m (entiers positifs), est la loi du rapport $(X/n)/(Y/m)$, où X et Y sont deux variables aléatoires indépendantes, de lois respectives $\chi^2(n)$ et $\chi^2(m)$. Elle caractérise la distribution de rapports de variances est et très présente en théorie des tests (analyse de variance et modèle linéaire). L'expression de sa densité est définie par un rapports de fonctions $\Gamma(x)$.

1.5 Caractéristiques des variables aléatoires

1.5.1 Espérance

Définition 1.8 *Soit X une v.a.r. et h une application de $\mathbb{R}$ dans $\mathbb{R}$. Donc $h(X)$ est elle aussi une v.a.r.*

- *Si X est discrète à valeurs dans un ensemble X, l'espérance de $h(X)$ est la quantité*

$$E(h(X)) = \sum_{x \in \chi} h(x) P(X = x),$$

pour vu que cette série converge (dans le cas où X est infini).

- *Si X est continue et admettant une densité f_X, l'espérance de $h(X)$ est la quantité*

$$E(h(X)) = \int_{-\infty}^{+\infty} h(x) f_X(x) dx,$$

à condition que cette intégrale soit convergente.

- *Notons que si $h(x) = x$, on obtient $E(X)$ appelée espérance mathématique (ou moyenne) de la v.a.r. X. Par ailleurs, si l'on définit la v.a.r. suivante :*

$$1_{\{X \in A\}} = \begin{cases} 1 & si \quad X \in A \quad (A \subset \mathbb{R}) \\ 0 & si\ non \end{cases},$$

qui est appelée fonction caractéristique de l'événement $\{X \in A\}$, alors l'espérance de cette v.a.r. est :

$$E(1_{\{X \in A\}}) = P(X \in A) = P_X(A),$$

d'où le lien étroit entre probabilité et espérance.

Propriétés

1. L'espérance est linéaire : pour tout $\alpha, \beta \in \mathbb{R}$ et pour toutes v.a.r. X et Y
$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y).$$
2. Si X est une v.a.r. constante égale à $a \in \mathbb{R}$, c'est-à-dire pour tout $\omega \in \Omega$, $X(\omega) = a$, alors
$$P(X = a) = 1 \quad \text{et} \quad E(X) = a.$$
3. L'espérance d'une v.a.r. positive est positive. En particulier, si $X \geq Y$ (ce qui signifie que pour tout $\omega \in \Omega$, $X(\omega) \geq Y(\omega)$), alors
$$E(X - Y) \geq 0$$
donc
$$E(X) \geq E(Y).$$
4. L'espérance d'une v.a.r. X est un indicateur de "localisation" de sa loi : $E(X)$ "valeur moyenne de X".

Néanmoins, la connaissance de l'espérance mathématique donne peu de renseignements sur cette v.a.r. Ainsi, il faut étudier "l'étalement" de sa loi, c'est-à-dire la dispersion de la v.a.r. X autour de sa moyenne $E(X)$

1.5.2 Variance et écart type

Définition 1.9 *Pour rendre positifs les écarts entre X et son espérance $E(X)$, un autre outil plus facile à manipuler que la valeur absolue, est à notre disposition : la mise au carré. On ne va donc pas calculer la moyenne des écarts mais la moyenne des écarts au carré. C'est ce qu'on appelle la variance.*

Définition 1.10 *La variance de la v.a.r. X est la quantité :*

$$Var(X) = E[(X - E(X))^2].$$

Vérifiée les propriétés :

- $Var(X) = E(X^2) - (E(X))^2$.
- $Var(aX + b) = a2Var(X)$ *pour tout* $a, b \in \mathbb{R}$.
- *En particulier,* $Var(X + b) = Var(X)$.

Afin d'être en mesure de comparer, en termes d'ordre de grandeur, variance et espérance, il faut prendre la racine carrée de la variance. C'est ce qu'on appelle l'écart-type.

Définition 1.11 *La racine carrée de $Var(X)$, notée σ_X, est appelée écart-type de X.*

Remarque 1.6

- *Si X est une v.a.r. telle que $E(X) = \mu$ et $Var(X) = \sigma^2$, alors la variable$Y = (X - \mu)/\sigma$ est d'espérance nulle et de variance 1. On dit que Y est centrée (d'espérance nulle) et réduite (de variance 1).*
- *Le moment d'ordre k est défini par*

$$m_k = \int_{-\infty}^{+\infty} (x - \mu)^k f(x) dx$$

- $\gamma_1 = \frac{m_3}{\sigma^3}$ *est le coefficient d'asymétrie (Skewness).*
- $\gamma_2 = \frac{m_4}{\sigma^4} - 3$ *est le coefficient d'aplatissement (Kurtosis).*

Résumés

- Loi uniforme $U[a, b]$:$E(X) = \frac{a+b}{2}$ et $Var(X) = \frac{(b-a)^2}{12}$.
- Loi de Bernoulli $B(p)$: $E(X) = p$ et $Var(X) = p(1 - p)$.
- Loi binomiale $B(n, p)$: $E(X) = np$ et $Var(X) = np(1 - p)$.
- Loi géométrique $G(p)$: $E(X) = \frac{1}{p}$ et $Var(X) = \frac{1-p}{p^2}$.
- Loi de Poisson $P(\lambda)$: $E(X) = Var(X) = \lambda$.
- Loi normale $\mathcal{N}(\mu, \sigma^2)$: $E(X) = \mu$ et $Var(X) = \sigma^2$.
- Loi exponentielle $E(\lambda)$: $E(X) = \frac{1}{\lambda}$ et $Var(X) = \frac{1}{\lambda^2}$

Chapitre 2

Estimation des quantiles

Dans cette section une revue de littérature sur les modèles statistiques pour l'estimation des quantiles sera présentée. Ce chapitre est divisé en deux parties.

- *La première partie :* consacré pour la modélisation des quantiles en utilisant l'inverse de la fonction de répartition
- *La deuxième partie :* est consacrée à la régression des quantiles

2.1 Quantiles basés sur l'inverse de la fonction de répartition

La fonction quantile d'une variable aléatoire est l'inverse de sa fonction de répartition. Soit Y une variable aléatoire à valeurs dans $\mathbb{R}$ et F_Y sa fonction de répartition. On appelle fonction quantile de Y, la fonction notée Q_Y , de $]0,1[$ dans $\mathbb{R}$ qui associe à $p \in]0,1[$ (p est la probabilité du quantile)

$$Q_Y(p) = F_Y^{-1}(p) = \inf\{y : F_Y(y) \geq p\} \tag{2.1}$$

avec F_Y^{-1} est l'inverse de la fonction de répartition F_Y. L'estimation de la fonction quantile Q_Y , repose sur l'estimation de la fonction de répartition F_Y. Cette dernière peut être estimée d'une façon non paramétrique, en utilisant la *fonction empirique* ou *les méthodes de lissage*, par exemple la méthode à noyaux ou celle basée sur les polynômes de Bernstein. Par contre, en hydroclimatologie, l'estimation de la fonction de répartition est souvent faite par le biais d'une fonction de répartition paramétrique. Plusieurs distributions paramétriques peuvent être employées dans ce cadre, *log-normale*, *gamma*, *GEV*, *lois de Halphen*,... etc.
Bien que le choix des distributions paramétriques soit vaste, les distributions

les plus utilisées pour la modélisation des extrêmes hydrologiques se résument aux distributions résultantes de la *théorie des valeurs extrêmes* (TVE). En effet, la TVE se base sur la description du comportement asymptotique des valeurs extrêmes. Plus formellement, considérons $Y = (Y_1, Y_2, ..., Y_n)$ un vecteur de n variables aléatoires i.i.d de fonction de répartition F_{Y_i} définie par :

$$F_{Y_i}(y_i) = P(Y_i \leq y_i), i = 1...n. \tag{2.2}$$

Pour approfondir le comportement des évènements extrêmes, on considère la variable aléatoire $M_n = \max(Y_1, Y_{2,.....} Y_n)$. Comme les variables aléatoires Y_i sont i.i.d., alors la fonction de répartition de M_n est définie par :

$$F_{M_n}(y) = P(M_n \leq y) = (F(y))^n. \tag{2.3}$$

Dans la pratique, il est difficile de calculer la fonction de répartition dans la formule (2.3). Le *théorème de Fisher et Tippett* [1928] donne une solution asymptotique pour le calcul de cettefonction de répartition. S'il existe deux suites de constantes $a_n > 0$ et $b_n \in \mathbb{R}$ et une distribution non dégénérée G telle que :

$$\lim_{n \rightsquigarrow +\infty} P\left\{\left(\frac{M_n - a_n}{b_n}\right) \leq y\right\} \rightarrow G(y) \tag{2.4}$$

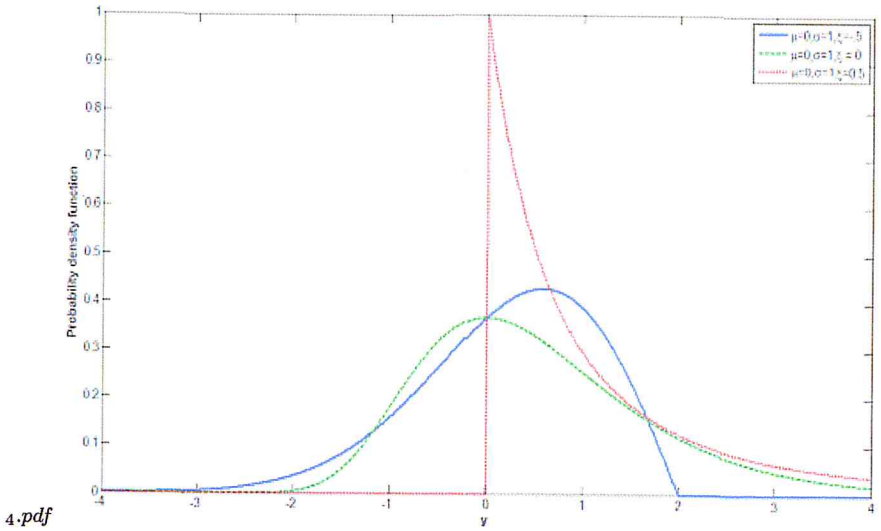
Alors G est de la forme :

$$G_{\mu,\sigma,\xi}(y) = \begin{cases} \exp\left[-\left(1 + \xi\left(\frac{y-\mu}{\sigma}\right)_+\right)^{-\frac{1}{\xi}}\right] & \text{si } \xi \neq 0 \\ \exp\left[-\exp\left(-\left(\frac{x-\mu}{\sigma}\right)_+\right)\right] & \text{si } \xi = 0 \end{cases} \tag{2.5}$$

où $y+ = \max(0, y)$ et μ, σ et ξ sont respectivement les paramètres de position (ou location), de dispersion (ou d'échelle) et de forme de la GEV. La distribution G s'appelle la loi généralisée des valeurs extrêmes (GEV : *Generalized Extreme Value*) et représente la première distribution issue de la TVE. La figure 1.1 représente la fonction de densité de la loi GEV dépendamment du paramètre de forme : le cas $\xi = 0$ correspond à la loi Gumbel, $\xi > 0$ à la loi de Fréchet et $\xi < 0$ à la loi Weibull.

L'approche basée sur la GEV a été critiquée dans la mesure où l'utilisation d'un seul maximum par année conduit à une perte d'information contenue dans les grandes valeurs observées dans un échantillon d'une variable aléatoire. Pour surmonter ce problème, Pickands [1975] a proposé la méthode des séries de durée partielle, ou excès au-delà d'un seuil (POT : Peak sover thresholds). La méthode POT consiste à utiliser, non seulement un maximum par variable aléatoire, mais toutes les observations qui dépassent

GEV



4.*pdf*

FIG. 2.1 – Illustration de la fonction de densité de la loi GEV selon différentes valeurs de ξ.

un certain seuil prédéfini et plus particulièrement les différences entre ces observations et le seuil fixé. Formellement, considérons

$$u \in \mathbb{R}, \mathbb{N}_u = card\{i; i = 1, ..., n | Y_i > u\}$$

et

$$Z_j = Y_j - u > 0, \text{ pour tout } j = 1, ..., N_u$$

où N_u représente le nombre des dépassements après le seuil u et Z_j sont les nouvelles variables. Le but, ici, est de définir à partir de la loi des $Y_i, i = 1, ..., n$, une loi conditionnelle par rapport au seuil u pour les variables Y_i, $j = 1, ..., N_u$ qui est définie par :

$$F_u(z) = F(Y - u \leq z \mid Y > u) = \frac{F(u+z) - F(u)}{1 - F(u)}; \quad z > 0 \qquad (2.6)$$

Pickands[1975] a proposé le résultat limite à la loi F_u. Ce résultat affirme que si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes (**Fréchet, Gumbel ou Weibull**) et lorsque le seuil u tend vers le point terminal z_F , alors il existe une fonction(u) strictement positive et un réel tels que

$$\lim_{u \to z_F} \sup_{0 \leq z \leq z_F - u} \left| F_u(z) - H_{\sigma(u),\xi}(z) \right| = 0 \tag{2.7}$$

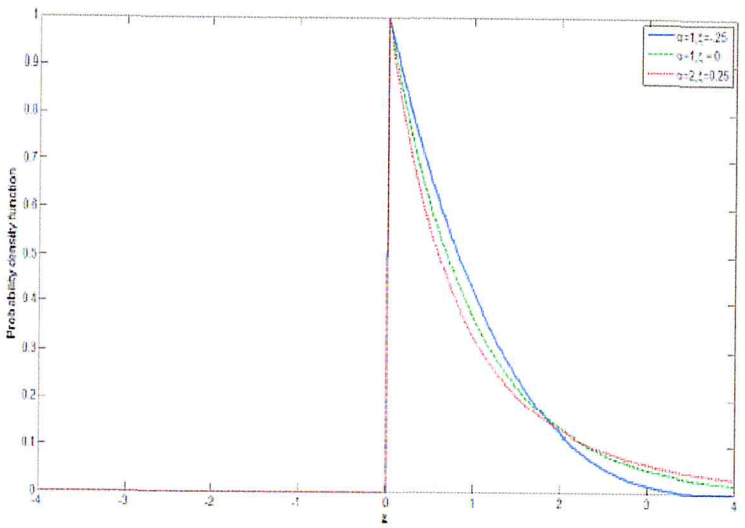
où $H_{\sigma(u),\xi}$, est la fonction de répartition de la loi de Pareto Généralisée (GPD : Generalized Pareto Distribution) et F_u est la fonction de répartition des excès au-delà du seuil u. Ainsi, pour u grand, la loi des excès est approchée par une loi GPD.

$$F_u \approx H_{\sigma(u),\xi}.$$

La distribution généralisée de Pareto s'écrit sous la forme

$$H_{\sigma(u),\xi}(z) \begin{cases} 1 - \left(1 - \xi \frac{z}{\sigma(u)}\right)^{\frac{-1}{\xi}} & \text{si } \xi \neq 0 \\ 1 - \exp\left(-\frac{z}{\sigma(u)}\right) & \text{si } \xi = 0 \end{cases} \tag{2.8}$$

où $z \geq 0$ si $\xi \geq 0$ et $0 \leq z \leq \frac{-\sigma(u)}{\xi}$ si $\xi < 0$, σ est le paramètre d'échelle et ξ est le paramètre de forme (La figure **??** montre une illustration de la fonction de densité de la loi de GPD pour différentes valeurs de σ et ξ)



Lois des excs et densit GPD

Plusieurs méthodes ont été fournies pour l'estimation des paramètres de la loi GEV et de la loi GPD. On trouve la méthode du maximum de vraisemblance [Smith, 1985], la méthode des moments [Christopeit, 1994] et la

méthode bayésienne [Christopeit, 1994]. Des estimateurs non paramétriques ont été aussi développés, comme l'estimateur de Pickands [1975] et l'estimateur de Hill [1975]. L'estimation des quantiles inconditionnels (stationnaires) pour les lois GEV et GPD est donnée par les formules suivantes :

$$\hat{Q}_Y^{GEV}(p) = G^{-1}_{\mu,\sigma,\xi}(p) = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - (-\log(p))^{-\xi}\right] & \text{si } \xi \neq 0 \\ \mu - \sigma \log(-\log(p)) & \text{si } \xi = 0 \end{cases}, \tag{2.9}$$

$$\hat{Q}_Y^{GPD}(p) = H^{-1}_{\mu,\sigma,\xi}(p) = \begin{cases} u - \frac{\sigma(u)}{\xi}\left(1 - p^{\xi}\right) & \text{si } \xi \neq 0 \\ u - \sigma(u)\log(p) & \text{si } \xi = 0 \end{cases}. \tag{2.10}$$

2.2 QUANTILE

Pour une variable Y, la fonction quantile se définit à partir de l'inverse de sa fonction de répartition. Quand cette fonction de répartition est strictement croissante, son inverse est défini sans ambiguïté. Mais une fonction de répartition reste constante sur tout intervalle dans lequel la variable aléatoire ne peut pas prendre de valeurs. De manière générale, soit $F(.)$ la fonction de répartition de la variable Y.

Définition 2.1 *On appelle fonction quantile d'ordre p de Y la fonction qui, à $p \in]0,1[$, associe*

$$Q_F(p) = F^{-1}(p) = \inf\{y : F(y) \geq p\} \tag{2.11}$$

où $F^{-1}(.)$ est souvent appelée l'inverse généralisé de $F(.)$.

Remarque 2.1 *Pour certaines valeurs de p, on donne un nom particulier aux quantiles, par exemple , pour $p = 0.5$ le quantile appelé médiane, pour $p = 0.25, 0.75$ le quantile appelé quartile, pour $p = 0.1; 0.9$ le quantile appelé décile et pour $p = 0.01; 0.09$ le quantile appelé centile,..., etc.*

Remarque 2.2 *$Q(Y)$ est une fonction non décroissante de p c'est- à-dire $Q_1(Y) \leq Q_2(Y)$ pour $p_1 < p_2$.*

Théorème 2.1 *Soit $(Y_1, ..., Y_n)$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées de fonction de répartition commune F avec une densité continue f. Alors si $f(Q_F(p)) > 0$, on a*

$$\sqrt{n}\left(Q_{F_n}(p) - Q_F(p)\right) = \sqrt{n}\left(\hat{F}_n^{-1}(p) - F^{-1}(p)\right) \xrightarrow{\mathcal{L}} N(0, \sigma^2)$$

où

$$\sigma^2 = p(1-p)/f\left(Q_F(p)\right)^2$$

Preuve: Le théorème central limite pour les variables aléatoires i.i.d implique que pour tout y dans le support de F on a :

$$\sqrt{n}(\hat{F}_n(y) - F(y)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

où :

$$\sigma^2 = F(y)(1 - F(y)).$$

Posant

$$y = Q_F(p) = F^{-1}(p),$$

donc on obtient :

$$\sqrt{n}(\hat{F}_n(Q_F(p)) - F(Q_F(p))) = \sqrt{n}(\hat{F}_n(F^{-1}(p)) - F(F^{-1}(p))) \xrightarrow{\mathcal{L}} \mathcal{N}(0, p(1-p))$$

ceci implique que

$$\sqrt{n}(\hat{F}_n(\hat{F}_n^{-1}(p)) - F(\hat{F}_n^{-1}(p))) \xrightarrow{\mathcal{L}} \mathcal{N}(0, p(1-p))$$

(voir D.Andrews,(1996) Handbook of Econometrice (vol.4)). alors

$$\sqrt{n}(p - F(\hat{F}_n^{-1}(p))) \xrightarrow{\mathcal{L}} \mathcal{N}(0, p(1-p)).$$

Un développement limité de Taylor donne

$$F(\hat{F}_n^{-1}(p)) = F(F^{-1}(p)) + f(\hat{Q}_F(p))(\hat{F}_n^{-1}(p) - F^{-1}(p)),$$

où $\hat{Q}_F(p)$ est une point sur le segment entre $Q_F(p)$ et $Q_{F_n}(p)$. En reécrivant la dernière formule en supposant que $f(\hat{Q}_F(p)) > 0$, on obtient

$$\sqrt{n}(\hat{F}_n^{-1}(p) - F^{-1}(p)) = -\frac{\sqrt{n}}{f(\hat{Q}_F(p))}(p - F(\hat{F}_n^{-1}(p))).$$

Lorsque $Q_{F_n}(p) \to Q_F(p)$ en probabilité $\hat{Q}_F(p) \to Q_F(p)$ en probabilité. Alors d'aprés Slutsky

$$\sqrt{n}(\hat{F}_n^{-1}(p) - F^{-1}(p)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

où :

$$\sigma^2 = p(1-p)/(f(Q_F(p))^2.$$

D'où le résultat. ■

2.3 Quantile conditionel

Définition 2.2 *Considérons deux variables quantitatives continues : une variable Y, appelée variable d'intérêt, et une variable X, appelée covariable. Soit $p \in]0,1[$. Le quantile conditionnel d'ordre p de la variable Y sachant que $X = x$ est défini de la manière suivante :*

$$Q_P(x) = F^{-1}(p|x) = \inf\{y : F(y|x) \geq p\}; \tag{2.12}$$

où $F(.|x)$ désigne la fonction de répartition conditionnelle de Y sachant que $X = x$, avec

$$F(y|x) = E\left[1_{\{Y \leq y\}}|X = x\right]$$

2.3.1 Quantiles basés sur la régression des quantiles

La régression des quantiles est une méthode statistique qui permet d'étudier l'impact de différentes covariables sur l'ensemble de la distribution de la variable d'intérêt. Contrairement à la régression ordinaire qui se rapproche des moyennes conditionnelles de la variable d'intérêt par rapport aux valeurs des covariables, la régression des quantiles donne une estimation des quantiles. Dans la régression ordinaire, le coefficient de régression représente le changement opéré dans la variable d'intérêt par unité de changement dans la covariable associée à ce coefficient. Dans la régression des quantiles, le coefficient de régression fournit une estimation du changement d'un quantile spécifique de la variable d'intérêt par unité de changement de la covariable. Considérons une variable aléatoire Y (liée à une covariable X) de fonction de répartition $F_Y(y) = P(Y \leq y)$. Le quantile d'ordre p est défini par :

$$Q_p(Y) = \inf\{y : F_Y(y) \geq p\}.$$

On considère la distribution conditionelle $F(Y|X = x)$, la régression des quantiles tente d'évaluer comment les quantiles conditionnels

$$Q_p(Y|X) = \inf\{y : F_{Y|X}(y) \geq p\}$$

changent lorsque la covariable X varie.

Dans le cas de la régression des quantiles linéaires, les quantiles conditionnels prennent la forme suivante :

$$Q_p(Y|X) = X'\beta_p\ , \tag{2.13}$$

où à chaque valeur de p correspond un coefficient β_p. L'expression (2.13) peut s'écrire d'une manière

équivalente :

$$Y = X'\beta_p + \varepsilon \qquad \text{avec} \qquad Q_p(\varepsilon|X) = 0. \tag{2.14}$$

Pour bien comprendre le principe de la régression des quantiles, il est bien utile de détailler comment on peut estimer les quantiles à partir du modèle de l'équation (2.14). En effet, l'estimation des régressions quantiles part de l'observation cruciale que le quantile d'ordre p est le résultat de la minimisation suivante (voir [Koenker & Bassett, 1987] pour la preuve)

$$\beta_p(Y) = \arg\min_{\beta} E\left[\rho_p(Y - X'\beta)\right], \tag{2.15}$$

où ρ_p est une fonction de perte définie par

$$\rho_p(u) = u\left(p - 1_{\{u<0\}}\right)$$

Cette estimation peut sembler moins intuitive que l'approche directe, qui utilise la statistique d'ordre $Y_{(1)} < ... < Y_{(n)}$ en estimant $Q_p(Y)$ par $\hat{Q}_p(Y)$ $= Y_{[np]}$ où $[np]$ est le plus petit entier supérieur ou égal à np. L'intérêt de cette méthode est qu'elle peut s'étendre facilement à un cadre conditionnel où on modélise le quanti conditionnel de la variable d'intérêt Y comme une fonction explicative des covariables X :

$$Q_p(Y \mid X = x) = \arg\min_{\beta} E\left[\rho_p(Y - X'\beta) \mid X = x\right]. \tag{2.16}$$

Dans la régression des quantiles, les coefficients régresseurs peuvent se définir comme suit :

$$\beta_p = \arg\min_{\beta} E\left[\rho_p(Y - X'\beta)\right] \tag{2.17}$$

On peut noter l'analogie avec le modèle de régression ordinaire, qui modélise l'espérance conditionnelle de Y par une forme linéaire en X :

$$E(Y \mid X) = X'\beta$$

Un estimateur de l'espérance d'une variable aléatoire Y conditionnel à X pouvant être obtenu par la fonction de perte quadratique

$$\arg\min_{\beta} E\left[(Y - X'\beta)^2 \mid X = x\right]$$

La fonction de perte quadratique est donc remplacée, dans la régression quantile, par la fonction de perte ρ_p.

2.3.2 Interprétation des régressions quantiles

La manière dont les distributions conditionnelles se modifient en fonction des variables explicatives renvoie à plusieurs questions. La première est simplement de décrire comment les quantiles conditionnels se modifient en fonction de ces déterminants, sans tenter de savoir si ce sont des personnes « comparables » aux différents quantiles conditionnels.

Les régressions quantiles sont des outils développés pour répondre à cette question. A condition de faire l'hypothèse simple mais restrictive d'invariance des rangs, elles peuvent permettre de répondre à une deuxième question, un peu plus précise, qui est de déterminer quelle est la variation de la variable d'intérêt correspondant à une variation marginale d'un de ces déterminants, pour les personnes qui se trouvent à un certain niveau de la distribution conditionnelle de la variable d'intérêt Y .

En général, les régressions quantiles ne donnent pas d'éléments pour répondre à une question encore différente, qui est d'estimer la distribution des effets de ce déterminant X sur la variable d'intérêt Y . Pour bien comprendre que les réponses à ces trois questions peuvent être différentes, il peut être utile de recourir à un exemple très simple .

Exemple 2.1 *Supposons qu'on s'intéresse à l'effet d'une variable binaire X sur une variable Y . La « population » à laquelle on s'intéresse est composée de cinq types en proportion identique , qu'on identifiera par les lettres de A à E . Suivant la valeur de X , ces personnes peuvent avoir des valeurs différentes de Y .*

- *Lorsque $X = 0$, les personnes ont , suivant leur type , $Y^A = 1, Y^B = 2$, $Y^C = 4$, $Y^D = 5$, ou $Y^E = 9$.*
- *Lorsque $X = 1$, la variable d'intérêt prend les valeurs $Y^A = 4, Y^B = 6, Y^C = 5, Y^D = 11$ ou $Y^E = 10$, suivant les types.*
- *Les effets individuels correspondant à un passage de $X = 0$ à $X = 1$ sont donc $\Delta Y^A = 3$, $\Delta Y^B = 4$, $\Delta Y^C = 1$, $\Delta Y^D = 6$ et $\Delta Y^E = 1$.*

On peut résumer ces données dans le tableau suivant :

Une régression quantile d'ordre 0.5 (regression médiane) de Y sur X, mesure l'écart entre la médiane de la distribution de Y conditionnelle à $X = 0$ et la médiane de la distribution conditionnelle à $X = 1$. Il vaut donc ici $2(6 - 4)$. Cette valeur est différente de l'effet pour les individus médians quand $X = 0$ (i.e., tels que $Y = q_{0.5}(Y \mid X = 0)$) de passer de $X = 0$ à $X = 1$: ces individus sont ceux de types C, pour lesquels on a $\Delta Y^C = 1$. Cette différence vient du fait qu'ici, les individus ne sont pas ordonnés (en termes de Y) de la même manière lorsque $X = 0$ et $X = 1$. Les individus

TAB. 2.1 – Exemple sur la régression des quantiles

Type	Valeur de Y si $X = 0$	Valeur de Y si $X = 1$	ΔY
A	1	4	3
B	2	6	4
C	4	5	1
D	5	11	6
E	9	10	1
Médiane	4	6	3
Moyenne	4.2	7.2	3

de type C sont « devant » ceux de type B lorsque $X = 0$ mais derrière eux lorsque $X = 1$.

L'hypothèse d'invariance des rangs n'est donc pas vérifiée, et on ne peut pas interpréter le coefficient de la régression quantile d'ordre 0.5 comme l'effet d'un passage de $X = 0$ à $X = 1$ pour les individus médians quand $X = 0$. Enfin, ces deux valeurs sont encore différentes de la médiane de la distribution des effets individuels ΔY , qui vaut 3. Ceci est lié à la non linéarité 9 des quantiles.

La médiane de la distribution de ΔY ne correspond pas à la différence des médianes des distributions de la variable d'intérêt, ou en termes mathématiques,

$$q_{0.5}(\Delta Y) \neq q_{0.5}(Y \mid X = 1) - q_{0.5}(Y|X = 0).$$

Ces remarques permettent de bien cadrer les usages qui peuvent être faits, ou non, des résultats d'une régression quantile. Il est utile de faire le lien avec ceux obtenus par une régression linéaire classique. L'objet principal de celle-ci est de modéliser la manière dont la moyenne conditionnelle varie en fonction de déterminants sur notre exemple, cela correspondrait à $3(= 7.2 - 4.2)$. Du fait de la linéarité de la moyenne, la différence des moyennes conditionnelles correspond également à l'effet moyen de l'augmentation de X, c'est-à-dire à la moyenne de ΔY . En revanche, cette différence ne correspond pas à ce que gagnerait à passer de $X = 0$ à 1 une personne dont la valeur de Y est proche de la moyenne conditionnelle lorsque $X = 0$ (dans notre exemple, il s'agit encore de C). Contrairement aux régressions quantiles, la condition d'invariance des rangs ne suffit pas ici pour obtenir une telle interprétation.

2.4 Autres propriétés de Quantiles et Quantile Régression

2.4.1 Propriétés d'équivariance de base

Soit A une matrice de dimension $k \times k$ matrice non singulière, $\gamma \in \mathbb{R}^k$ et $\alpha > 0$ est une constante, soit $\hat{\beta}(p; y, X)$ être l'estimateur dans la régession du quantile d'ordre p sur les observations (y, X), puis pour tout $p \in [0, 1]$, alors :

i) $\hat{\beta}(p; \alpha y, X) = \alpha\hat{\beta}(p; y, X)$;

ii) $\hat{\beta}(p; -\alpha y, X) = -\alpha\hat{\beta}(1-p; y, X)$;

iii) $\hat{\beta}(p; y + \gamma, X) = \hat{\beta}(p; y, X) + \gamma$;

iv) $\hat{\beta}(p; y, XA) = A^{-1}\hat{\beta}(p; y, X)$.

2.4.2 Propriété d'équivariance

Les quantiles sont équivariants à monotone transformations. Supposons que g est une fonction croissante sur $\mathbb{R}$. Alors pour toute variable Y

$$Q_p(g(Y)) = g(Q_p(Y)).$$

2.5 Des estimateurs plus adaptés à certaines situations

Même dans le cadre du modèle de translation simple (2.14), dans lequel les coefficients correspondant à une régression linéaire et à une régression quantile sont les mêmes, il peut être préférable d'utiliser cette dernière. La première raison est qu'elle est robuste aux valeurs aberrantes ou à des erreurs très dispersées. Intuitivement, cette propriété est due au fait que les quantiles sont moins sensibles que la moyenne à la présence de valeurs très grandes. Considérons tout d'abord le cas des valeurs aberrantes. Supposons que la variable Y^* vérifie le modèle de translation simple (2.14), mais que dans de très rares cas, les données observées ne correspondent pas à la variable d'intérêt Y mais à une variable erronée, éventuellement corrélée avec les variables explicatives X. Formellement, on observe

$$Y = AX'\delta + (1 - A)Y^*$$

où A est une variable inobservée valant 1 lorsque Y est abberrant, 0 sinon, avec

$$P(A = 1|X, \varepsilon) = p$$

petit.

Une régression linéaire de la variable observée (avec erreur) Y sur X donnera une estimation biaisée de notre paramètre d'intérêt, puisqu'elle sera égale (pour un échantillon de taille tendant vers l'infini) à

$$\gamma MCO = \gamma + p(\delta - \gamma).$$

Si δ est très différent de γ, le terme de biais peut être important même lorsque la probabilité d'observer des observations erronées p est faible. En revanche, on peut montrer que si X' est très grand, l'estimateur de l'effet de $X_k(k > 1)$ obtenu par une régression quantile vaut bien γ_k . En d'autres termes, la présence de valeurs aberrantes n'affecte pas les résultats de la régression quantile, sauf les coefficients de la constante.

Dans un même ordre d'idée, toujours dans le cas du modèle de translation simple, les résultats obtenus par régression quantile seront plus précis en général lorsque les résidus sont très dispersés. Un exemple extrême est celui où " n'a pas d'espérance, ce qui se produit lorsque " peut prendre des valeurs très grandes avec une probabilité importante. Cette situation n'est pas si rare en pratique. Les lois de Cauchy et certaines lois de Pareto (utilisées pour modéliser les hauts salaires ou les patrimoines), par exemple, n'ont pas d'espérance. Dans ce cas, l'estimateur des moindres carrés ordinaires n'est pas convergent : même pour des échantillons énormes, il pourra prendre des valeurs très différentes du vrai paramètre β. A l'inverse, l'estimateur obtenu par régression quantile sera convergent.

2.6 Principes statistiques et mise en oeuvre pratique

2.6.1 Définition de l'estimateur et propriétés statistiques

Pour bien comprendre le principe des régressions quantiles, il est utile de détailler comment on peut estimer les quantiles d'une variable d'intérêt Y à partir d'un échantillon $(Y_i)_i = 1...n$ de variables supposées $i.i.d.$ La manière la plus intuitive de calculer l'estimateur standard $\hat{Q}_p(Y)$ consiste à ordonner ces n variables, le quantile d'ordre p étant fourni par la $[np^{ieme}]$ observation

où $[np]$ est le plus petit entier supérieur ou égal à np. Mais il est plus utile, pour le passage aux régressions quantiles.

$$\hat{Q}_p(Y) = \arg\min_b \frac{1}{n} \sum_{i=1}^{n} \rho_p(Y_i - b) \tag{2.18}$$

où $\rho_p(.)$ est une « fonction test » définie par

$$\rho_p(u) = (p - 1_{\{u<0\}})u. \tag{2.19}$$

Par exemple, pour $p = 1/2$, c'est-à-dire si on s'intéresse à la médiane, la fonction test correspond simplement à la (demi-) valeur absolue. La solution du programme de minimisation ci-dessus correspond alors bien à la médiane.

L'intérêt de cette définition est de s'étendre simplement au cadre conditionnel qui nous intéresse, où l'on modélise le quantile conditionnel de la variable d'intérêt Y comme une fonction des variables explicatives X. En effet, il suffit de remplacer $\hat{Q}_p(Y)$ et b dans (2.18) par respectivement $Q_p(Y|X)$ et une fonction $b(X)$. Dans le cas des régressions quantiles classiques, on peut se limiter aux fonctions linéaires puisqu'on suppose que

$$Q_p(Y \mid X) = X'\beta_p$$

On a alors :

$$\hat{\beta}_p = \arg\min_\beta E\left[\rho_p(Y - X'\beta)\right]. \tag{2.20}$$

On peut noter l'analogie avec le modèle de régression linéaire classique, qui modélise l'espérance conditionnelle de Y par une forme linéaire en X :

$$E(Y \mid X) = X'\beta_0 \tag{2.21}$$

L'espérance d'une variable aléatoire pouvant être obtenue par

$$E(Y) = \arg\min_a E[(Y - a)^2] \tag{2.22}$$

le coefficient β_0 est défini par

$$\beta_0 = \arg\min_\beta E[(Y - X'\beta)^2] \tag{2.23}$$

La fonction de perte quadratique qui est utilisée dans une régression linéaire par les moindres carrés ordinaires est donc remplacée, dans la régression quantile, par la fonction test $\rho_p(.)$. Celle-ci augmentant de manière linéaire et non quadratique avec le résidu, les très grands écarts sont beaucoup moins

pénalisés, ce qui explique la robustesse de la régression quantile aux valeurs extrêmes ou aberrantes.

Cette estimation peut se faire pour tout quantile d'ordre p, où $p \in [0, 1]$. Il existe donc en principe une infinité de régressions quantiles possibles. En pratique, le nombre de quantiles qu'on estime dépendra de la taille de l'échantillon. Il est bien entendu illusoire de tenter d'approcher très finement une distribution avec un nombre fini d'observations : le nombre de quantiles empiriques distincts sera probablement restreint. Le choix de modéliser l'ensemble des percentiles ou simplement les quartiles et la médiane dépendra non seulement du degré de précision souhaitée pour décrire la distribution mais aussi des données disponibles.

2.6.2 Algorithmes utilisés

Il n'existe pas de solution explicite à (2.18), si bien qu'il faut résoudre ce programme numériquement. Un problème est que la fonction objectif n'est ni différentiable (la fonction n'est pas dérivable en 0) ni strictement convexe. Les algorithmes standards tels que celui de Newton Raphson ne peuvent donc pas être utilisés directement. Cependant, on peut reformuler (2.18) comme un programme linéaire :

$$\min_{(\beta,u,v)\in\mathbb{R}^p\times\mathbb{R}_+^{2n}} p1'u + (1-p)1'v \qquad et \qquad X\beta + u - v - Y = 0,$$

où $X = (X_{1,\ldots\ldots,}X_n)', Y = (Y_{1,\ldots\ldots,}Y_n)'$ et 1 est un vecteur de 1 de taille n. La méthode du simplexe a été jusqu'à récemment la méthode la plus classique pour résoudre ce type de problèmes linéaires. Cependant, elle devient coûteuse en temps de calcul lorsque le nombre d'observations augmente et elle n'est donc indiquée que pour de petits échantillons. Pour des échantillons plus conséquents, les méthodes de points intérieurs sont plus performantes pour résoudre ces programmes linéaires .

2.6.3 Propriétés asymptotiques de l'estimateur et estimation de la précision

Les propriétés asymptotiques de $\hat{\beta}_p$ sont délicates à établir car, contrairement à l'estimateur des moindres carrés, il n'existe pas de forme explicite pour $\hat{\beta}_p$. Pour plus de détails, on se référera par exemple à l'ouvrage de **Koenker(2005)**. Nous nous contentons ici du résultat principal sur la loi asymptotique de $\hat{\beta}_p$.

Théorème 2.2 *Supposons que*

$$\varepsilon_p = Y - X'\beta_p$$

admette, conditionnellement à X, une densité en

$$f_{\varepsilon_p|X}(0 \mid X) \quad \textit{et que} \quad J_p = E\left[f_{\varepsilon p|X}(0 \mid X)XX'\right]$$

soit inversible. Alors

$$\sqrt{n}(\hat{\beta}_p - \hat{\beta}) \xrightarrow{d} \mathcal{N}(0, p(1-p)J_p^{-1}E\left[XX'\right]J_p^{-1}) \tag{2.24}$$

Dans le cas du modèle de translation simple (2.14)*, la variance asymptotique prend une forme particulièrement simple. On a en effet,*

$$\varepsilon_p = \varepsilon - Q_p(\varepsilon)$$

et la variance asymptotique V_{as} s'écrit plus simplement :

$$V_{as} = \frac{p(1-p)}{f_\varepsilon(Q_p(\varepsilon))^2}E\left[XX'\right]^{-1}$$

Cette variance est très proche de celle des MCO avec résidus homoscédastiques, si ce n'est que

$$\sigma^2 = V(\varepsilon)$$

est remplacé par

$$p(1-p)/f_\varepsilon(Q_p(\varepsilon))^2$$

Le terme de densité est logique : seuls les résidus autour de $Q_p(\varepsilon)$ vont apporter de l'information sur la valeur du quantile conditionnel de Y . Ce résultat explique que même dans le cas de translation simple, il peut être parfois préférable d'utiliser une régression quantile pour certaines distributions des termes inobservés ε. L'estimation par régression quantile sera plus précise qu'une estimation par MCO lorsque

$$p(1-p)/f_\varepsilon(Q_p(\varepsilon))^2 < \sigma^2$$

En dehors du modèle restrictif de translation simple, la variance asymptotique est plus complexe à estimer que dans le cadre d'un modèle de régression linéaire simple. Plusieurs méthodes d'inférence ont été proposées pour construire des tests ou des intervalles de confiance sur β_p, et il n'existe pas à l'heure actuelle de consensus sur la méthode à utiliser. On trouvera dans ***Kocherginsky, et al.(2005)*** *une présentation générale de ces méthodes et une discussion pratique des cas où certaines sont plus ou moins indiquées. Le choix dépend des hypothèses plus ou moins restrictives qu'on accepte de faire sur le modèle sous-jacent (modèle de translation...), de la taille de l'échantillon ou du nombre de variables du modèle.*

Certaines méthode s'appuient sur une estimation directe de la variance asymptotique en partant de la formule (2.24). La difficulté principale de cette approche est la présence de la densité conditionnelle $f_{\varepsilon p|X}(0|X)$, qui est délicate à estimer . Dans le cadre restrictif d'un modèle de *translation-échelle*, une méthode basée sur les tests de rang est parfois utilisée (**cf.Koenker,2005**). Il est surtout courant de s'appuyer sur des méthodes de bootstrap. Elles consistent à générer des échantillons «$factices$» par des tirages avec remise à partir de l'échantillon initial, et à effectuer une régression quantile sur ces échantillons. L'inconvénient de ces méthodes est qu'elles sont aussi coûteuses en temps de calcul. Ce dernier augmente à la fois avec la taille de l'échantillon et le nombre de variables explicatives. Une solution récente(«**Markov Chain Marginal Bootstrap**»,*ou* **MCMB**) a été proposée par **He & Hu(2002)** pour résoudre en partie ce problème quand le nombre de variables explicatives est important. Ces méthodes ne sont pas toujours performantes sur de petits échantillons.

Enfin, l'un des intérêts de la régression quantile étant de ne pas supposer a priori que les variables explicatives ont un effet homogène sur l'ensemble de la distribution de la variable d'intérêt, il est tout à fait possible de tester cette hypothèse à partir des estimations obtenues. Par exemple, l'homogénéité de l'effet de l'une des variables X_k correspond à l'égalité des coefficients $\beta_{K,p_1,.....},\beta_{K,p_m}$ (où $(p_1,.....p_m)$ peuvent être par exemple l'ensemble des déciles), ce qui peut se tester simplement. Un tel test s'appuie sur la distribution jointe asymptotique de $(\hat{\beta}_{p_1},.....,\hat{\beta}_{p_m})$, donnée par le résultat suivant :

$$\sqrt{n}(\hat{\beta}_{p_K}-\beta_{p_K})_{K=1}^{m} \xrightarrow{D} \mathcal{N}(0,V), \tag{2.25}$$

où V est une matrice par bloc dont le bloc $V_{k,l}$ vérifie

$$V_{k,l}=[p_k \cap p_l - p_k p_l]\, J_{kl}^{-1} E\left[XX'\right] J_{kl}^{-1} \tag{2.26}$$

Ce résultat est une généralisation du théorème1 à plusieurs quantiles.

2.7 Avantages de la régression quantile

2.7.1 Pourquoi la régression quantile ?

Raison1 : La régression quantile nous permet d'étudier l'impact de prédicteurs sur différents quantiles de la distribution de la réponse, et donc fournit une image complète de la relation entre Y et X.

- Y_i : vitesses maximales des vents des cyclones tropicaux dans l'Atlantique Nord

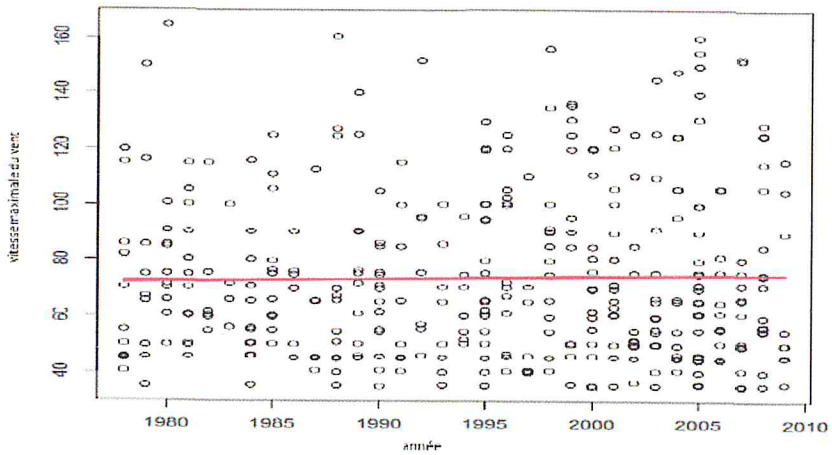


FIG. 2.2 – vitesses maximales des vents des cyclones tropicaux dans l'Atlantique Nord

– X_i : année1978 – 2009
LS estimation de la pente : 0.095
$Q - valeur$: 0.569
Aucune tendance significative dans la moyenne!
Q : Les quantiles de la vitesse maximale du vent changent-ils avec le temps ?

p^{ime} quantile :$Q_p(Y) = \{y : P(Y < y) = p\}$.
$p = 0.95 : 0.009$
$p = 0.75 : 0.100$
$p = 0.50 : 0.718$
$p = 0.25 : 0659$
Raison2 : robuste aux valeurs aberrantes dans y observations ?
Raison3 : l'estimation et l'inférence sont sans distribution ?

2.8 Extensions

2.8.1 Les régressions quantiles dans les modèles linéaires

$$Q(Y|X) = X'\beta(p), \quad 0 < p < 1$$

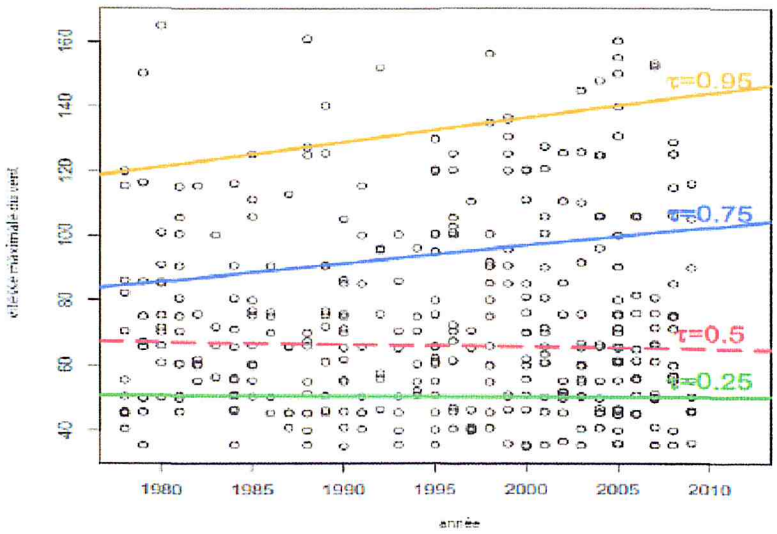


FIG. 2.3 – tendance dans la moyenne des vents

où $\beta(p) = (\beta_1(p), ..., \beta_k(p))'$ est le coeffcient quantile qui peut dépend de p, le premier élément de X est celui correspondant à l'interception pour que

$$Q(Y|X) = \beta_1(p) + x_2\beta_2(p) + ...x_k\beta_k(p);$$

$\beta(p)$ est le changement marginal dans le p-quantile en raison de la changement marginal en x.

Remarque 2.3 *$Q(Y|X)$ est une fonction non décroissante de p pour tout x donné.*

2.8.2 Modèle de décalage à l'échelle de l'emplacement

$$Y_i = \alpha + Z_i'\beta + (1 + Z_i'\gamma)\varepsilon_i \ \varepsilon_i \overset{iid}{\sim} F(.)$$

La fonction quantile conditionnelle

$$Q(Y|X_i) = \alpha(p) + Z_i'\beta(p).$$

où

$\alpha(p) = \alpha + F^{-1}(p)$ est non décroissante de p

$\beta(p) = \beta + \gamma F^{-1}(p)$ peut dépendre de p. Autrement dit, la covariable est autorisé à avoir un impact différent sur les Distributions quantiles différents du Y .

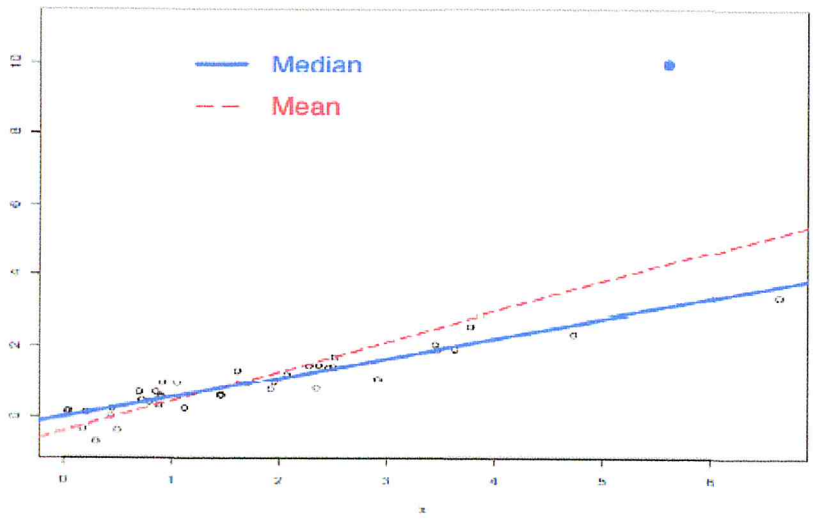


FIG. 2.4 – Régression de la médiane et la moyenne

2.8.3 Modèle de changement de lieu

$\gamma = 0$, pour que $\beta(p) = \beta$ est constant à travers niveaux quantiles.

2.8.4 Les régressions quantiles instrumentales

Comme en régression linéaire, il arrive fréquemment que certaines composantes des variables X soient a priori endogènes. Par exemple, dans une étude sur l'impact d'un dispositif de formation sur le salaire, le fait de participer à ce dispositif peut être lié à des caractéristiques inobservées qui influent également le salaire. Dans ce cas, l'estimateur $\hat{\beta}_p$ défini par (2.18) ne mesure pas l'effet causal du dispositif de formation.

En revanche, on peut disposer d'instruments affectant ces variables mais pas directement les composantes inobservées de la variable d'intérêt (représentées par le résidu ε_P). Plus précisément, si l'on se place dans le cadre de la régression quantile précédente,

$$Y = X'\beta_p + \varepsilon_P$$

on suppose qu'il existe des variables Z corrélées à X et telles que

$$Q_P(\varepsilon_P|Z) = 0 \tag{2.27}$$

Cette hypothèse est l'équivalent de l'hypothèse $E(\varepsilon|Z) = 0$ en régression linéaire instrumentale.

Il est utile de distinguer, parmi les variables explicatives X, les variables a priori endogènes, notées $X_1 \in \mathbb{R}^q$ (c'est à dire telles que $Q_p(\varepsilon_P \mid X_1) \neq 0$), et les variables exogènes X_2. Supposons qu'on dispose de $Z_1 \in \mathbb{R}^r$ (avec $r \geq q$) variables supplémentaires aux explicatives, telles que $Q_p(\varepsilon_P|Z) = 0$, avec $Z = (Z_1, X_2)$. Cette condition implique que :

$$Q_p(Y - X_1'\beta_{1p}|Z) = X_2'\beta_{2p} \tag{2.28}$$

Cette propriété est à la base d'une méthode proposée récemment par **Chernozhukov & Hansen** (2008). L'équation (2.28) signifie que dans une régression quantile de $Y - X_1'\beta_{1p}$sur Z_1 et X_2, le coefficient de Z_1 est égal à 0. L'idée de **Chernozhukov et Hansen** est alorsd' « inverser » la régression quantile, en estimant β_{1p} par le paramètre $\hat{\beta}_{1p}$ permettantd'obtenir, dans la régression quantile de $Y - X_1'\hat{\beta}_{1p}$ sur Z un coefficient égal à 0 pour Z_1. En pratique, les auteurs proposent l'algorithme suivant :

a. Définir une grille sur β_{1p}, $\{b_1, ..., b_2\}$.

b. Pour $j = 1$ à J :

- Calculer les estimateurs de régression quantile de $Y - X_1'b_j$ sur (Z_1, X_2). soit $(\hat{\gamma}(b_j), \hat{\beta}_{2p}(b_j))$ les estimateurs correspondants.

c. Calculer la statistique de Wald correspondant au test de $\gamma(b_j) = 0$:

$$W_n(b_j) = n\hat{\gamma}(b_j)'\hat{V}_{as}^{-1}(\hat{\gamma}(b_j))\hat{\gamma}(b_j)$$

d. Définir l'estimateur de β_p par :

$$\hat{\beta}_{1p} = \underset{j=1....J}{\arg\min}\, W_n(b_j), \qquad \hat{\beta}_{2p} = \hat{\beta}_{2p}(\hat{\beta}_{1p})$$

L'intérêt de cet algorithme est qu'il ne s'appuie que sur des régressions quantiles classiques. Il peut donc être mis en oeuvre simplement avec des logiciels standards. La commande Stata ivqreg a d'ailleurs été introduite récemment. En pratique, la grille doit être suffisamment fine pour ne pas altérer les propriétés asymptotiques de l'estimateur (**cf.Chernozhukov & Hansen** , 2008 pour plus de détails). Pour que le temps de calcul reste raisonnable, le nombre de variables endogènes doit donc être petit ($Q = 1$ ou 2).

Notons que d'autres solutions existent pour estimer des régressions quantiles instrumentales. Abadie et al. (2002) proposent de recourir à une approche par régression quantile pondérée dans le cas où la variable endogène

X_1 et l'instrument Z_1 sont binaires. On peut estimer directement les coefficients en s'appuyant sur l'équation (2.28) et la méthode des moments généralisés. Comme toujours, la difficulté est évidemment de trouver un instrument valide, c'est-à-dire vérifiant (2.27).

2.8.5 Les régressions quantiles dans les modèles non linéaires

Nous considérons ici des extensions de la régression linéaire quantile aux modèles nonlinéaires de la forme β

$$Y = g(X'\beta_0 + \varepsilon) \tag{2.29}$$

où g est une fonction non-linéaire connue. Deux exemples importants sont le modèle binaire, pour lequel $g(x) = 1\{x > 0\}$, et le modèle à censure fixe, pour lequel

$$g(x) = max(s, x), \text{ou } g(x) = min(s, x))$$

avec s une constante connue. Ce dernier modèle est souvent utilisé pour modéliser la consommation d'un bien, qui prend la valeur nulle quand il n'est pas consommé. Ceci peut être rationalisé par l'existence d'une valuation implicite du bien c^* par les consommateurs éventuels, qui ne consomment ce bien que lorsque cette valuation est strictement positive. On observe donc la consommation $c = max(0, c^*)$. Dans ces modèles, il est difficile d'utiliser des restrictions de la forme $E(\varepsilon|X) = 0$ car en général, $E(Y|X) \neq g(X'\beta_0)$. L'approche standard consiste alors à imposer des hypothèses paramétriques sur la distribution des résidus. Par exemple, il est fréquent de supposer l'indépendance entre X et ε et la normalité de ces derniers (on parle alors de modèle probit lorsque $g(x) = 1\{x > 0\}$ et de modèle tobit lorsque $g(x) = max(0, x)$). Ces hypothèses sont cependant restrictives et souvent difficiles à justifier.

Une approche alternative à ces hypothèses paramétriques est de recourir à des restrictions sur les quantiles. En effet, on peut facilement étendre les restrictions sur les quantiles des termes de perturbations à une transformation non linéaire(**??**), valable pour toute variable aléatoire U et toute fonction g croissante et continue à gauche. Ainsi, si l'on impose dans le modèle non linéaire (2.29) la restriction $Q_p(\varepsilon|X) = 0$ et que g est croissante continue à gauche, on obtient

$$Q_p(Y \mid X) = g(Q_p(X'\beta_0 + \varepsilon \mid X)) = g(X'\beta_0)$$

Par le même argument que celui développé dans la section 3, il s'ensuit que

$$\beta_0 \in \arg\min_{\beta} E\left[\rho_p(Y - g(X'\beta))\right]$$

Comme précédemment, on estime alors β_0 par

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho_p(Y_i - g(X_i'\beta)) \tag{2.30}$$

L'estimateur défini par (2.30) est très proche de celui de la régression quantile linéaire, la différence étant simplement l'ajout dans le programme de la fonction g. Pour le modèle de censure fixe pour lequel $g(x) = max(s, x)$, $\hat{\beta}$ est $\sqrt{n}-$convergent, et peut être estimé par une application itérative de régressions quantiles linéaires (**cf.par exemple Buchinsky,1994**). Pour la médiane ($p = 1/2$), l'estimateur proposé par **Powell** (1984) (« censored LAD estimator », i.e. l'estimateur des moindres valeurs absolues censuré) est implémenté sous Stata via la commande clad. L'estimation de la médiane n'est pas suffisante si la censure se trouve très haut par rapport à la distribution. **Hong & Chernozhukov(2002)** proposent donc un estimateur en trois étapes qui peut être utilisé pour l'ensemble des quantiles. Cette méthode est décrite dans**Fack & Landais (2009)** qui l'appliquent pour modéliser l'impact des incitations fiscales sur les dons aux oeuvres. Ces dons sont en effet minoritaires, puisque environ 15% des ménages déclarent un tel don, et une modélisation de l'effet moyen ne permettrait pas de mettre en évidence un effet.

2.9 Tests en régression quantile

Les tests présentés dans cette section ont été développés par **Furno** (2011). Ils portent globalement sur l'hypothèse d'invariance des coefficients de régressions comme décrit par **Johnston et Di Nardo** (1997), en mettant en confrontation un modèle de régression contrainte et un ou plusieurs autres modèles de régression non-contrainte. **Chow** (1960) utilise la même démarche lorsqu'il analyse des données temporelles. Nous présentons ci-dessous ces tests.

2.9.1 Test de qualité de prédiction

Soit n la taille de l'échantillon d'étude. L'évaluation de la qualité de prédiction d'un modèle construit est déterminante lorsqu'on veut prédire une réponse correspondant aux variables indépendantes en présence dans le modèle. Le test de qualité de prédiction consiste d'abord à estimer les paramètres du modèle sous les n_1 (n_1, n_2) premières observations de l'échantillon d'étude, puis d'utiliser le modèle construit pour prédire les valeurs de la variable

dépendante des données restantes. Cette démarche permet de comparer la constance des coefficients estimés entre les modèles de régression contrainte (le modèle est identique pour toutes les n observations) et non-contrainte (le modèle reste identique pour seulement n_1 observations) définis dans le tableau 2.2)

TAB. 2.2 – Modéles comparés pour le test de qualité de prédiction

	Modèles de régression
Contrainte	$Q_{y_i}(p\|x_i) = \beta_0(p) + \beta_1(p)x_{1i} + ... + \beta_k(p)x_{pi},\ i = 1, ..., n$
Non-contrainte	$Q_{y_i}(p\|x_i) = \beta'_0(p) + \beta'_1(p)x_{1i} + ... + \beta'_k(p)x_{pi},\ i = 1, ..., n_1 (n_1 \approx \frac{n}{2})$

Autrement dit, cela permet d'apprécier la stabilité des paramètres à estimer sans tenir compte de la taille d'échantillon utilisée. Ainsi, on peut appréhender l'homogénéité des paramètres sous forme d'un test d'hypothèse :

$$H_0 \ : \begin{pmatrix} \beta_0(p) \\ . \\ . \\ . \\ \beta_k(p) \end{pmatrix} = \begin{pmatrix} \beta'_0(p) \\ . \\ . \\ . \\ \beta'_k(p) \end{pmatrix}$$

H_1 : Au moins un des $\beta_j(p)$ est différent de $\beta'_j(p),\ j = 0, ..., p.$

La statistique de test C^1 est donnée par :

$$C^1 = \frac{(\tilde{v}(p) - \hat{v}(p))/\Delta_{(ddl_{H_0}, ddl_{H_1})}}{\hat{v}(p)/ddl_{H_1}} \sim F(n_2, n_1 - (p+1)), \tag{2.31}$$

où

$$\Delta(ddl_{H_0}, ddl_{H_1}) = n - (p+1) - (n_1 - (p+1) = n_2$$
$$ddl_{H_1} = n_1 - (p+1)$$

$F(n_2, n_1 - (p+1))$ est la **loi de Fisher** à n_2 et $(n_1 - (p+1))$ degrés de liberté, $\tilde{v}(p)$ et $\hat{v}(p)$ sont les fonctions objectives des modèles contraint et non-contraint (donnés dans le Tableau 2.2). La règle de décision est de rejeter H_0 quand

$$C^1 > F_{1-\alpha}(n_2, n_1 - (p+1))$$

avec

$$F_{1-\alpha}(n_2, n_1 - (p+1))$$

le quantile d'ordre $(1-\alpha)$ de la **loi de Fisher** avec n_2 et $n_1 - (p+1)$ degrés de liberté.

Principe d'application du test

1. Estimer les paramètres du modèle en utilisant la méthode RQ pour :
   (a) l'échantillon complet (n observations) ;
   (b) le sous-échantillon de n_1 premières observations ;
2. Évaluer la fonction objective pour chacun des deux modèles contraint et non-contraint ;
3. Calculer la statistique de test C^1 ;
4. Rejeter l'hypothèse nulle au seuil si la valeur calculée de

$$C^1 > F_{1-\alpha}(n_2, n_1 - (p+1)).$$

2.9.2 Test de stabilité des paramètres

Le test de stabilité sur les paramètres ou encore le test de changement structurel vise à constater statistiquement des modifications de comportements dans l'échantillon soumis à l'étude. Pour ce faire, on doit caractériser au mieux ce qui permet de définir les sous-échantillons à confronter (en utilisant des informations externes ou une variable indépendante disponible dans la donnée, comme variable de segmentation) et déceler la nature du changement survenu (modification ou non des coefficients relatifs aux variables indépendantes).

En considérant G groupes (Un groupe est un sous-échantillon de taille n_i obtenu du tri des données selon la variable de segmentation, telle que $\sum_{i=1}^{G} n_i = n$) de n_i observations($n_i > p+1$), le test de stabilité repose sur la comparaison des modèles de régression contrainte (les coefficients sont les mêmes quel que soit le groupe étudié) et non-contrainte (les coefficients ne sont pas nécessairement identiques sur les différents groupes) pour déterminer si, sur ces groupes étudiés, les coefficients sont identiques. Le tableau 2.3 présente ces différents modéles

Cette comparaison informe sur la nécessité ou non de distinguer les modèles de régression à $(p+1)$ paramètres dans les G groupes, en appréhendant la question sous les tests d'hypothèses suivants :

$$H_0 : \begin{pmatrix} \beta_0(p) \\ . \\ . \\ \beta_k(p) \end{pmatrix} = \begin{pmatrix} \beta'_0(p) \\ . \\ . \\ \beta'_k(p) \end{pmatrix} = \ldots\ldots\ldots = \begin{pmatrix} \beta_0^G(p) \\ . \\ . \\ \beta_k^G(p) \end{pmatrix},$$

TAB. 2.3 – Modéles comparés pour le test de stabilité des paramétres.les différent sous-groupes ont été constitués à partir d'unvariabledesegmentation

	Modèles de régression
Contrainte	$Q_{y_i}(p\|x_i) = \beta_0(p) + \beta_1(p)x_{1i} + ... + \beta_k(p)x_{ki},\ i = 1, ..., n$
Non-contrainte	$Q_{y_i}(p\|x_i) = \beta_0^1(p) + \beta_1^1(p)x_{1i} + ... + \beta_k^1(p)x_{ki},\ i = 1, ..., n_1$
Non-contrainte	$Q_{y_i}(p\|x_i) = \beta_0^G(p) + \beta_1^G(p)x_{1i} + ... + \beta_k^G(p)x_{ki},\ i = 1, ..., n_G$

H_1 : Au moins un coefficients différe des autres.

La statistique de test s'appuie sur les sommes pondérées des valeurs absolues de résidus des modèles de régression contrainte et non-contrainte. Elle s'écrit :

$$\tilde{C}^1 = \frac{\left[\tilde{v}(p) - \sum_{i=1}^{G} \hat{v}_i(p)\right] / \Delta_{(ddl_{H_0}, ddl_{H_1})}}{(\sum_{i=1}^{G} \hat{v}_i(p))/ddl_{H_1}} \sim F((G-1)(p+1), n - G(p+1)),$$

où

$$\Delta_{(ddl_{H_0}, ddl_{H_1})} = n - (p+1) - \sum_{i=1}^{G}(n_i - (p+1)) = \sum_{i=1}^{G-1}(p+1) = (G-1)(p+1)$$

et

$$ddl_{H_1} = \sum_{i=1}^{G}(n_i - (p+1)) = n - G(p+1)$$

et $\tilde{v}(p), \hat{v}_i(p)$ sont les fonctions objectives des modèles de régression contrainte et non-contrainte (donnés dans le tableau 2.3). Sous H_0, la statistique $\tilde{C}^1$ suit la **loi de Fisher** à $(G-1)(p+1)$ et $(n - G(p+1))$ degrés de liberté. La région critique $(R.C)$ du test s'écrit : $R.C.$:

$$\tilde{C}^1 > F_{1-\alpha}((G-1)(p+1), n - G(p+1))$$

où $F_{1-\alpha}((G-1)(p+1), n - G(p+1))$ est le quantile d'ordre $(1-\alpha)$ de la loi de Fisher à $(G-1)(p+1)$ et $n - G(p+1)$ degrés de liberté .

Principe d'application du test

1. Regrouper les données selon la variable de segmentation ;
2. Effectuer la régression avec la méthode RQ sur :
   (a) l'échantillon complet (n observations) ;

(b) chacun des G groupes ;

3. Évaluer la fonction objective pour toutes les $(G+1)$ régressions ;
4. Calculer la statistique de test $\tilde{C}^1$;
5. Conclure avec $R.C$: en rejetant l'hypothèse nulle (H_0) au seuil , si la valeur calculée de $\tilde{C}^1 > F_{1-\alpha}((G-1)(p+1), n-G(p+1))$ (c'est-à-dire que l'on ne peut pas affirmer l'homogénéité des paramètres inter-groupes).

2.9.3 Test de Wald pour les hypothèses générales linéaires

Définir le vecteur coeffcient

$$\theta = (\beta(p_1)^t .....\beta(p_2)^t)^t.$$

- Hypothèse nulle

$$H_0 \ : R\theta = r.$$

- Statistique de test

$$T_n = n(R\hat{\theta} - r)^t (RV^{-1}R^t)^{-1}(R\hat{\theta} - r)$$

où V est la matrice de bloc (2.26).
- Sous

$$H_0, T_n \xrightarrow{d} \chi^2_q$$

- où q est le rang de R

2.10 Classement Score Test

Considéronz le modéle

$$Q_p(Y|X_i, Z_i) = X_i^t\beta(p) + Z_i^t\gamma(p).$$

et hypothéses

$$H_0 : \gamma(p) = 0 \ \ v.s \ H_1 : \gamma(p) \neq 0$$

ici $\beta(p) \in \mathbb{R}^k$ et $\gamma(p) \in \mathbb{R}^q$.

2.10.1 Fonction Score

$$S_n = n^{-\frac{1}{2}} \sum_{i=1}^{n} z_i^* \psi_p \left\{ y_i - X_i^t \hat{\beta}(p) \right\}$$

- $\psi_p(u) = p - I(u < 0)$;
- $Z^* = (z_i^*) = Z - X(X^t \Psi X)^{-1} X^t \Psi Z$;
- $\Psi = diag(f_i \{Q_p(Y|X_i, Z_i)\})$;
- $\hat{\beta}(p)$ sont les estimations du coefficient quantile obtenues sous H_0 .

2.10.2 Propriété asymptotique

Sous H_0 as $n \longrightarrow \infty$,

$$S_n = \mathcal{N}(0, M_n^{\frac{1}{2}})$$

où

$$M_n = n^{-1} \sum_{i=1}^{n} z_i^* z_i^{*t} p(1-p).$$

2.10.3 Statistiques de test de Score

$$T_n = S_n^t M_n^{-1} S_n \xrightarrow{d} \chi_q^2$$

sous H_0 .

2.10.4 Simplification pour i.i.d paramètres

- $$Z^* = (z_i^*) = \left\{ I - X(X^t X)^{-1} X^t \right\} Z$$

  les résidus par projeter Z sur X ;
- $$M_n = p(1-p) n^{-1} \sum_{i=1}^{n} z_i^* z_i^{*t}$$
- donc pas besoin d'estimer les paramétres de nuisance $f_i \{Q_p(Y|X_i, Z_i)\}$.

2.10.5 Construction de l'intervalle de confiance de $\gamma(p)$

- $\gamma(p)$ est un paramétre d'intérét correspondant à l'un des covariables .
- l'intervalle de confiance de $\gamma(p)$ peut étre construit par de test de score de rang .
- considérez les hypotheses :

$$H_0 : \gamma(p) = \gamma_0 \text{ v.s } H_1 : \gamma(p) \neq \gamma_0$$

où $\gamma(p)$ est un scalaire prédéfini .
- Rejeter H_0 si $T_n \geq \chi^2_\alpha(1)$, le $(1-\alpha)^{éme}$ quantile de $\chi^2(1)$,et vice versa.
- la collection de tous les γ_0 pour le quel H_0 n'est pas rejété est pris pour étre le $(1-\alpha)^{éme}$ l'intervalle de confiance de $\gamma(p)$.

2.11 Implémentation dans le paquet R quantreg

```
fit=rq(y~x,tau)
#assuming iid errors
summary.rq(fit, se="rank", tau, alpha= 0.05, iid=TRUE)
#assuming non iid errors
summary.rq(fit, se="rank", tau, alpha= 0.05, iid=FALSE)
#Outputs : estimation and (1-alpha) CI for each coefficient .
```

2.12 Présentation de R

2.12.1 Généralités

R est un systéme qui est communément appelé langage et logiciel, il permet de réaliser des analyses statistiques.

Plus particuliérement, il comporte des moyens qui rendent possibles la manipulation des données, les calculs et les représentations graphiques. *R* a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes.

En effet *R* posséde :

- un systéme éfficace de manipulation et de stockage des données
- différents opérateurs pour le calcul sur tableaux, en particulier les matrices
- un grand nombre d'outils pour l'analyse des données et les méthodes statistiques

- des moyens graphiques pour visualiser les analyses
- un langage de programmation simple et performant comportant : conditions,boucles, moyens d'entrées sorties, possibilité de définir des fonctions récursives.
- La conception de R a été fortement influencée par deux langages :
  - *S* qui est un langage développé par les *AT&T Bell Laboratories* et plus particuliérement par *Rick Becker*, *John Chambers* et *Allan Wilks*. *S* est un langage de haut niveau et est un environnement pour l'analyse des données et les représentations graphiques. *S* est utilisable à travers le logiciel $S - Plus$ qui est commercialisé par la société *Insightful* ($http : //www.splus.com/$). $S - Plus$ est un des logiciels de statistiques les plus populaires et il s'est imposé comme une référence dans le milieu statistique.
  - *Scheme* de *Sussman* ($http : //www.schemers.org/$) est un langage fonctionnel, le principe fondamental de ce langage est la récursivité. *R*, le langage obtenu est trés semblable à *S*, l'exécution et la sémantique sont dérivées de *Scheme*.

Le noyau de *R* est écrit en langage machine interprété qui a une syntaxe similaire au langage *C*, mais qui est réellement un langage de programmation avec des capacités identique au langage *Scheme*.

La plupart des fonctions accessibles, par l'utilisateur dans *R*, sont écrites en *R*. (Le systéme est lui-meme écrit en *R*). Pour les taches intensives les langages *C*, C^{++} et Fortran ont été utilisés et liés pour une meilleure efficacité.

R permet aux utilisateurs d'accroitre les possibilités du logiciel en créant de nouvelles fonctions. Les utilisateurs expérimentés peuvent écrire du code en *C* pour manipuler directement des objets *R*.

R comporte un grand nombre de procédures statistiques.

Parmi elles, nous avons : les modéles linéaires, les modéles linéaires généralisés, la régression non-linéaire, les séries chronologiques, les tests paramétriques et non paramétriques classiques, ...

Il y a également un grand nombre de fonctions fournissant un environnement graphique flexible afin de visualiser et créer divers genres de présentations de données.

Les utilisateurs pensent souvent que *R* est un syst'eme de statistique. Les concepteurs et développeurs préférent dire que c'est un environnement dans lequel des techniques statistiques sont exécutées. *R* peut étendre ses fonctions par l'intermédiaire de modules. Il existe à l'heure actuelle douze modules fournis lorsque *R* est distribué ($sousUnix$) et d'autres sont disponibles par l'intermédiaire du $CRAN$. Ils sont disponibles pour des buts spécifiques et présentent une large gamme de statistiques modernes. (analyse descriptive

des donn´ees multidimensionnelles, arbres de régression et de classification, graphiques en trois dimensions, etc...)

R est développé pour pouvoir être utilisé avec les systèmes d'exploitation *Unix*, *GNU/Linux*, *Windows* et *MacOS*. R posséde un site officiel à l'adresse $http://www.R-project.org/$, c'est un logiciel libre qui est distribué sous les termes de la <**GNU Public Licence**> (régle du copyleft) et il fait partie intégrante du **projet GNU.**

2.12.2 Créateurs de R

R a été initialement créé par **Robert Gentleman** et **Ross Ihaka** du département de statistique de l'Université d'Auckland en Nouvelle Zélande.

2.12.3 Le CRAN

Le <**Comprehensive R Archive Network**> (**CRAN**) est un ensemble de sites qui fournit ce qui est nécessaire à la distribution de R, ses extensions, sa documentation, ses fichiers sources et ses fichiers binaires.

Le **CRAN** est similaire à **CPAN** pour le langage **Perl** ou **CTAN** pour **TEX/LATEX**.

Le site maître du **CRAN** est situé en *Autriche* à *Vienne*, nous pouvons y accéder par l'URL : $http://cran.r-project.org/$

2.12.4 Le point fort de R

Ce logiciel étant de domaine public son point fort est représenté par le développement d'applications, de modules qui sont mis à la disposition de tous les utilisateurs et développeurs. Son réseau international de développement est en perpétuelle évolution.

L'intérêt majeur de R est qu'il est ouvert à tous. De ce fait, tout le monde peut apporter sa pierre 'a l'édifice .

Cela laisse envisager de phénoménales extensions au systéme. Le potentiel de R semble donc énorme.

2.12.5 Pourquoi utiliser R ?

Tout d'abord R est un logiciel **gratuit** et à code source ouvert (open-source). Il fonctionne sous $UNIX$ (et *Linux*), Windows et Macintosh. C'est donc un logiciel **multi − plates − formes**. Il est développé dans la mouvance des logiciels libres par une communauté sans cesse plus vaste de bénévoles motivés.

Tout le monde peut d'ailleurs contribuer a son amélioration en y intégrant de nouvelles fonctionnalites ou méthodes d'analyse non encore implémentées. Cela en fait donc un logiciel en rapide et constante évolution.

C'est aussi un outil trés puissant et trés complet, particulierernent bien adapts pour la mise en oeuvre informatique de **méthodes statistiques**. Il est plus difficile d'accés que certains autres logiciels du marche (comme *SPSS* ou *Minitab* par exemple), car il n'est pas concu pour être utilisé à l'aide de «clics» de souris dans des menus. L'avantage en est toutefois double :

- l'approche est **pédagogique** puisqu'il faut maitriser les méthodes statistiques pour parvenir à les mettre en oeuvre ;
- l'outil est trés éffieace lorsque l'on domine le langage *R* puisque l'on devient alors capable de créer ses propres outils, ce qui permet ainsi d'opérer

des analyses trés sophistiquees sur les données.

2.12.6 R et les statistiques

R est un logiciel dans lequel de nombreuses techniques statistiques modernes et classiques ont ete implementees. Les methodes les plus courantes permettant

de realiser une analyse statistique telles que :

- statistique descriptive ;
- tests d'hypotheses ;
- analyse de la variance ;
- methodes de regression lineaire (simple et multiple) ;
- - etc.

sont enchassees directement dans le coeur du systems. Notez également que laplupart des methodes avancees de statistique sont aussi disponibles au travers de modules externes appeles *packages*. Ceux-ci sont faciles a installer directement a partir d'un menu du logiciel. Ils sont tous regroupes sur le site internet du $ComprehensiveRArchiveNetwork$ $(CRAN)(http$: $//cran.r - project.org)$sur lequel vous pouvez les consulter. Ce site fournit aussi, pour certainsngrands domaines d'étude, une liste comrnentee des packages associes aces themes (appelee $TaskView$), ce qui facilite ainsi la recherche d'une methode statistique particuliere. Par ailleurs, une documentation detaillee en anglais de chaque package est disponible sur le $CRAN$.

Il est par ailleurs utile de noter que les methodes statistiques les plus recentes y sont regulierement ajoutees par la cornmunaute statistique ellemerne.

2.12.7 R et les graphiques

Une des grandes forces de R reside dans ses capacites, bien supeneures a celles des autres logiciels courants du marche, a combiner un langage de programmation avec la possibilite de realiser des graphiques de qualite.

Les graphiques usuels s'obtiennent aisement au moyen de fonctions predefinies. Ces dernieres possedent de tres nombreux parametres permettant par exemple d'ajouter des titres, des legendes, des couleurs, etc.

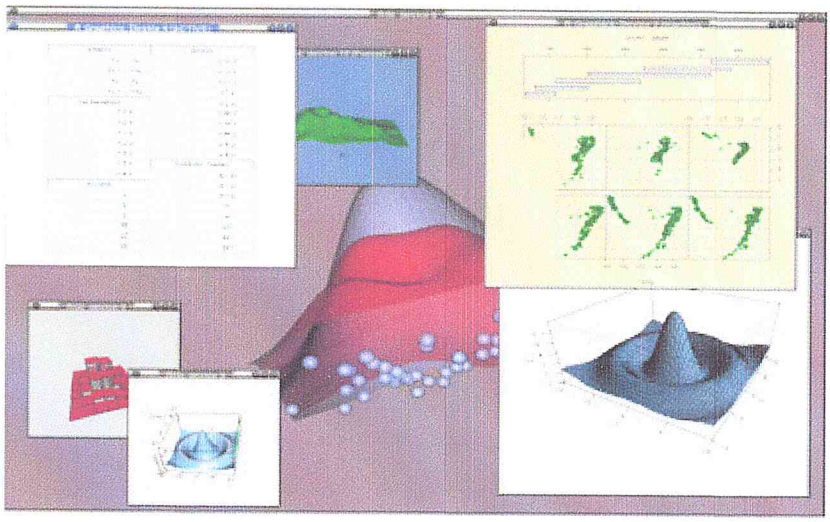
Mais il est egalement possible d'effectuer des graphiques plus sophistiques permettant de representer des donnees complexes telles que des courbes de surface ou de niveau, des volumes affiches avec un effet **3D**, des courbes de densite, et bien d'autres choses encore.

Il vous est egalement possible d'y ajouter des formules mathernatiques. Vous pouvez aussi agencer ou superposer plusieurs graphiques sur une rneme fenetre, et utiliser de nombreuses palettes de couleur.

Vous pouvez obtenir une demonstration des possibilites graphiques de R en tapant successivement les commande suivantes :

```
demo (image)
example (contour)
demo (graphics)
demo (persp)
demo (plotmath)
demo (Hershey)
require (lattice) # Charge Ie package que vous devez avoir
# prealablement installe en passant par Ie
# menu Packages/Installer le(s) package(s).
demo (lattice)
example (wireframe)
require (rgl) # Meme remarque que ci-dessus.
demo (rgl)   # Possibilite d'interaction avec la souris.
example (persp3d)
```

La figure ci-dessous presents quelques-uns de ces graphiques



Quelques possibilites graphiques offertes par R

2.12.8 Installation du logiciel *R*

Le logiciel est télécharger sur le site web officiel de *R* (*http* : *//www.r − project.org/*), il faut ensuite se diriger dans download **CRAN**. Choisissez un site miroir proche de chez vous

(par exemple : en France), les téléchargements seront probablement plus rapides, vous trouvez ensuite un encadrement légendé *download and install R.*

Sous Windows XP :

Commencez par télécharger le logiciel *R* (fichier **R − x − win.exe** où **x** est le numéro de la derniére version disponible) à l'aide de votre navigateur web usuel à l'adresse suivante : **http : //cran.r − project.org/bin/windows/base/**

Enregistrez ensuite ce fichier exécutable sur le bureau de Windows puis doublecliquez sur le fichier **R − x − win.exe** dont voici l'icône :

Le logiciel s'installe alors et vous n'avez plus qu'à suivre les instructions qui s'affichent et à conserver les options proposées par d´efaut.

Lorsque l'icône est ajoutée sur le bureau, l'installation peut être considérée comme terminée.

2.12.9 Instalation des packages

Définition 2.3 *Un package R se présente sous la forme d'un fichier compressé qui regroupe des fonctions, les documente et peut contenir du package facilite l'utilisation et le partage de fonction de R. Autrement dit, un package est un compilation certains d'outils sont déjà présents dans l'installation de base de R. En effet, lors de l'installation de R, un dossier Library s'est crée par défaut, sous Windows fichiers binaires pré-compilés.*

Il comprend les packages de base de *R*. Mais d'autres packages qui vous seront utiles pour réaliser vos analyses statistiques seront à télécharger puis à installer.

Sur le **CRAN**, le réseau officiel de distribution de code et de documentation *R*. Plus de **3000** bibliothéques *R* sont en ce moment (il y'a des mois) disponibles sur le site **CRAN**.

D'autres bibliothéques sont disponibles sur la page Web de chercheurs ou en annexe de publications scientifiques.

L'installation du module (**package**) **packageade4** doit se faire une seule fois, lors de la premiére utilisation. Pour les utilisations subséquentes, il suffira de charger le package aprés l'ouverture du logiciel *R*. On suppose qu'on dispose d'une connexion internet pour effectuer ces opérations. On peut ensuite travailler hors-connexion.

- Ouvrir le logiciel R en cliquant sur l'icône créé sur votre bureau lors du téléchargement. Dans la fenêtre qui apparaît, aller dans menu "packages / Choisir le site miroir "**CRAN**"(pas trop loin si possible), par exemple : *france lyon* et cliquer sur **OK** (voir figure 1.3).
- Menu " **packages / Installer des packages**" , Dans la liste des packages, choisir **ade4** et cliquer sur **OK**.
- Menu < **packages/Mettre à jour les packages** > , pour obtenir la derniere version des packages déjà installés.

2.12.10 Charger un package déjà installé

Aller dans le menu **"packages -Charger le package"** puis sélectionner le package ou taper la commande :

`library( nom package)`

`library(nom-de-la-bibliotheque) # par exemple : library(MASS)`

`library() # permet aussi de lister les packages installes sur notre machine.`

2.12.11 Notions élémentaires

Appel de R

On rentre dans R en tapant dans la fenêtre *Unix* : R et on le quitte en tapant :

`quit(), q() ou Control D`

Une question est pos´ee lorsqu'on veut quitter.

`Save workspace image? [y/n/c] :`

- `y` permet de quitter et de sauvegarder le travail effectué
- `n` permet de quitter sans sauvegarder
- `c` permet d'annuler la fermeture de R

Le prompt de R est par défaut le caractére `'>'`, signifie qu'il est en attente d'une commande.

Les objets courants sont sauvés dans le fichier `'.RData'`

L'aide de R

Il est possible d'obtenir de l'aide sous différents formats (**html et texte**).R nous propose une aide en ligne en tapant la commande :

`> help.start()`

L' aide pour une commande particuli'ere est possible en tapant :

`>?nom-commande`

Les commandes :

`> help(nom-commande)`

`>?"nom-commande"`

`> help("nom-commande")`

donnent le même résultat avec la possibilité pour les deux derniéres d'obtenir l'aide pour les caractéres spéciaux.

2.12.12 Créer et manipuler des données

Voici quelques définitions importantes dans la manipulation des données.

Définition 2.4 ***(objet)*** *un objet est un espace dans lequel vous pouvez stocker tout ce qui vous intéresse.*

Définition 2.5 ***(vecteur)*** *un vecteur est un objet d'un même mode pour toutes les valeur qui le constituent.*

Définition 2.6 ***(matrice)*** *une matrice est un objet d'un même mode pour toutes les valeur qui la constituent.*

Définition 2.7 ***(liste)*** *une liste est un objet permettant de stocker des objet qui peuvent être hétérogénes, c'est-à-dire qui n'ont pas tous le même mode ou la même longueur*

Définition 2.8 ***(tableau)*** *un tableau de données, ou* ***data.frame*** *en anglais, est une liste particuliére dont les composantes sont même longueur et dont les modes peuvent être différents(est une collection de vecteurs de même longueur). Un data.frame est un tableau à double entrée : les lignes sont les individus sur les quels les mesures sont faites et les colonnes sont les variables.*

2.12.13 Les objets : création et types Création

Simplement par affectation avec les opérateurs `<-`, `->` en lui donnant un nom :

```
b <- sqrt(2) (cree l'objet b)   # (racine carree)
x <-b (x recoit la valeur b)
x=b (x recoit la valeur b)
b-> x (x recoit la valeur b)
```

Si un objet níexiste pas l'affectation le crée. Sinon l'affectation écrase la valeur précédente.

Affichage de la valeur díun objet

```
a (affiche le contenu de a)
q (affiche le contenu de la fonction q qui permet de quitter)
print(a) (affiche le contenu de a)
```

Vecteur

On déclare nos données sous forme matricielle (vecteur ligne ou matrice) dans plusieurs

façons : la fonction `c( )` pour un vecteur (de type ligne) :

```
v=c(1,2,3,4)
[1] 1 2 3 4
```

Pour un vecteur colonne, il faut utiliser la fonction `as.matrix()` :

```
> as.matrix(c(1,2,3,4))
[,1]
[1,] 1
[2,] 2
[3,] 3
[4,] 4
```

l'opérateur : (prioritaire sur les opérations au sein díune expression)

```
x=1 :10
x
[1] 1 2 3 4 5 6 7 8 9 10
```

La fonction `seq` de différents arguments `seq(form,to )` : `form` le début de la séquence, `to` pour la fin :

```
seq(from=1, to=10)
[1] 1 2 3 4 5 6 7 8 9 10
seq(from=1.5, to=4.7)
[1] 1.5 2.5 3.5 4.5
```

`seq(from, to, by= )` : la même chose et by pour le pas ;

```
seq(from=1, to=10, by=2)
[1] 1 3 5 7 9
```

`seq(from, to, length.out=)` : il créé une séquence de from jusqu'au to de la longueur length :

```
> seq(from=1, to=10, length=3)
[1] 1.0 5.5 10.0
```

Matrices

La fonction `matrix()` : remplissage par colonne par défaut :

`matrix(data = NA, nrow = 1, ncol =1 , byrow = FALSE, dimnames = NULL)`

`data` : les données pour remplir la matrice ;

`nrow` : le nombre de lignes et ncol le nombre de colonnes ;

`dimnames` : une liste de dimension **2** avec les noms des lignes et/ou colonnes

```
matrix(data=1 :6,nr=2,nc=3,
dimnames=list(c("row1", "row2"), c("C.1", "C.2", "C.3")))
      C.1  C.2  C.3
row1    1    3    5
row2    2    4    6
```

Les listes

la liste est un objet hétérogéne. C'est donc un ensemble ordonné d'objets qui ne sont pas forcément tous du même mode ni de même longueur. Les objets sont appelés composants de la liste. Remarquons que

- les composants peuvent être de mode $\neq$,
- les composants peuvent avoir un nom,
- la liste n'est pas une classe.

Création `tablev<- c("Paris", "Lyon")`
`list <- list (1 :3,ville="tablev") les composants peuvent avoir 1 nom`

Forme une liste à **5** éléments, qui sont des vecteurs numériques de longueur **1**, égaux respectivement à 1,2,3,4,5 :
`as.list(1 :5), conversion explicite`

Remarque 2.4 *On peut poser la question à R si l'objet est une liste* `is.list(x)` *et la réponce est* `TRUE` *ou* `FALSE`.

Créer un tableau de données sous R

Pour créer un tableau de données sous R, il faut utiliser la fonction **data.frame**. Cette fonction permet de concaténer des vecteurs de mêmes longueur et de modes différents.

Exemple : Saisissez deux vecteurs notés mat et phy :
`mat<-c(19.6,17.6,18.2,16.0)`
`phy<-c(19.1,17.8,18.7,16.1)`
puis construisez le tableau de données, notées, à l'aide de la fonction **data.frame** en tapant la ligne de commande suivant :
`res<-data.frame(mat,phy)`
Il est possible de donner des noms aux lignes du tableau de données avec l'option **row.names** qui doit fournir un vecteur de mode caractére et de longueur égale au nombre de lignes du tableau de données.

Calcul arithmétique et fonctions simples

Les opérations suivantes sont effectuées composante par composante :
- la somme des deux vecteurs avec $+$;
- le produit avec $*$;
- le rapport avec $/$;
- puissance avec $\hat{}$;

Voici quelques fonction R simple :
- `sum(x)` : sommes des composantes de `x` ;
- `prod(x)` : produit des composantes de `x` ;
- `max(x)` : maximum des composantes de `x` ;
- `which.max(x)` : retourne l'indice du maximum des composantes de `x` ;
- `range(x)` : idem que `c(min(x),max(x))` ;
- `length(x)` : nombre d'éléments dans `x` ;
- `mean(x)` : la moyenne des éléments dans `x` ;
- `median(x)` : la médiane des éléments dans `x` ;
- `var(x)` : la variance (corrigée) des éléments dans `x` ;
- `cor(x)` : matrice de corrélation si `x` est un `data-frame` et `1` sinon ;
- `cov(x,y)` : la covariance entre `x` et `y` ;

Le résultat est une valeur, sauf pour `range(), var(), cor() et cov()`.
- `round(x,n)` : arrondi les éléments de `x` à `n` chiffres aprés la virgule ;
- `rev(x)` : inverse l'ordre de `x` ;
- `sort(x)` : trie les éléments de `x` dans ordre croissante ;
- `rank(x)` : rangs des éléments de `x` ;
- `cumsum(x)` : un vecteur avec les sommes cumulées de composantes de `x` ;
- `cumprod(x)` : idem pour le produit ;
- `table(x)` : retourne un tableau avec les effectifs de différentes valeurs de `x` ;
- `which(x==a)` : retourne un vecteur des indices de `x` pour les quels l'opération de comparaison est vraie ;

```
x=c(2,6,10,60,34)
which(x<35)
[1] 1 2 3 5  #les indices pour les quels on a vrai la omparaison
x[which(x<35)]
[1] 2 6 10 34
```

2.12.14 Graphiques

Possibilité de voir des exemples de graphiques avec **demo (graphics)** ou **demo (persp).** Lorsqu'une fonction graphique est tapée sur la console, une fenêtre graphique va s'ouvrir avec le graphe demandé.

Partitionner une fenêtre graphique

`par(mfcol=c(nr,nc))` :
on partitionne la fenêtre en une matrice de `nr` lignes et `nc` colonnes et le remplissage est réalisé par colonne.

`mfrow` : idem mais les graphiques sont dessinés en ligne ;
`layout()` : pour des partitions plus complexes
`layout(matrix(c(1,2,3,4),2,2)) # pour inserer dans un graphique`
`layout(matrix(c(1,1,2,1),2,2),c(3,1),c(1,3))`
`layout.show(2)  # et visualiser la partition cree`
`plot (x)` : graphe des valeurs de `x` (sur l'ordonnée) en fonction des valeurs de `x` ;
`plot (x,y)` : graphe des valeurs de `y` (l'ordonnée) en fonction des valeurs de `x` ;
`pie(x)` : "camembert" des valeurs de `x` ;
`boxplot(x)` : boxplot de `x` ;
`hist(x)` : histogramme de `x` (pour `x` quantitative) ;
`barplot(x)` : diagramme en colonnes (pour `x` qualitative) Pour chaque fonction, on a plusieurs options mais certaines sont communes :
`type` : "p" : points, "l" : lignes, "b" les deux, "h" : lignes verticales, "s" : escaliers ; for stair steps ;
`xlab, ylab` : noms des axes, variables caractéres entre " " ;
`main` : variable de type caractére ; sub : sous-titre ;
`points(x,y)` : ajoute des points ;
`lines(x,y)` : idem mais avec des lignes ;
`segments(x0,y0,x1,y1)` : trace une ligne entre les points `(x0,y0)` et `(x1,y1)`
`abline(a,b)` : trace une ligne de pente `b` et ordonnée à l'origine `a` ;
`legend(x,y,legend)` : ajoute une légende au point de coordonnées `(x, y)` avec les symboles donnés par legend.

2.12.15 Les fonctions

Structure générale pour créer des fonctions :

La syntaxe générale de définition d'une fonction est la suivante :
`nom-fonction<-function(arg1[=expr1],arg2[=expr2]...)`
{ blocs díinstructions }
Les accolades permettent de séparer les instructions par rapport à la signature de la fonction, les crochets, eux permettent de spécifies valeurs par défaut des arguments de façon facultative.

Exemple 2.2 *voici un exemple de simulation de la fonction* $\texttt{f(x)} = \texttt{x}^2$:

```
carre=function(x) {x*x}
carre(2)
[1] 4
```

Lois de probabilité usuelles

R fournit un jeu trés complet de tables statistiques. Les fonctions fournies nous donnent la densité de probabilité cumulée (ou fonction de répartition), $P(X \leq x)$, la densité de

probabilité, le quantile (ou fonction de répartition inverse) et des série de nombres pseudo aléatoires générées suivant la loi de probabilité considérée.

TAB. 2.4 – Abbreviations des lois de probabilités dans R

Loi de probabilité	Nomes R	Arguments supplémontaires
Beta	beta	Shape1,Shape2,ncp
Binomiale	binom	size,prob
Cauchy	cauchy	location,scale
Khi2	chisq	df,ncp
Exponentielle	exp	rate
Fisher	f	df1,df2,ncp
Gamma	gamma	shape,scale
Géométrique	geom	prob
Hypergéométrique	hyper	m,n,k
Log-normale	lonrm	meanlog,sdlog
Logistique	logis	location,scale
Négative binomiale	nbinom	size,prob
Normale	norm	mean,sd
Poisson	pois	lambda
Student	t	df,ncp
Uniforme	unif	min,max
Weibull	weibull	shape,scale

Pour les lois *Beta*, *Khi2*, *Fisher* et *Student* ncp désigne le paramètre de décentrage.

Quatre préfixes sont possibles pour obtenir les sorties désirées :

- **d** : correspond à la densité de probabilité.
- **p** : correspond à la fonction de répartition (valeur inverse du quantile).
- **q** : correspond au quantile.
- **r** : correspond à la génération aléatoire de nombre.

Le premier argument est `'x'`(une valeur) pour dnom-loi, `'q'`(un quantile) pour pnom-loi, et `'n'`(un nombre) pour rnom-loi.

Exemples avec la loi normale (arguments par défauts : mean=0, sd=1) :

```
dnorm(0.37)
[1] 0.3725483
# donne l'ordonnée de la courbe de densité à la valeur 0.37.
```

```
pnorm(0)
[1] 0.5
# donne la probabilité de la fonction de répartition au quantile
0.
round(qnorm(0.975),2)
[1] 1.96
# donne le quantile pour la probabilité de 0.975 de la fonction
de répartition.
rnorm(3)    # engendre 3 valeurs de la loi normale
[1] -1.5651925 1.2417975 0.9014982
```

2.13 La regréssion quantile sous R

2.13.1 Qu'est-ce que la régression quantile ?

La régression quantile est une technique statistique destinée à estimer et à faire des inférences sur les fonctions quantiles conditionnelles. Tout comme les méthodes de régression linéaire classiques basées sur la minimisation des sommes des carrés des résidus permettent d'estimer les modèles de fonctions moyennes conditionnelles, les méthodes de régression quantile offrent un mécanisme d'estimation des modèles pour la fonction médiane conditionnelle et toute la gamme des autres fonctions conditionnelles. En complétant l'estimation des fonctions moyennes conditionnelles par des techniques d'estimation d'une famille entière de fonctions quantiles conditionnelles, la régression quantile est capable de fournir une analyse statistique plus complète des relations stochastiques entre variables aléatoires.

2.13.2 Qu'est-ce qu'une vignette ?

Cette annexe a été rédigée dans le format **Sweave de Leisch** (2003). Sweave est une mise en œuvre conçue pour *R* du style de programmation alphabétisé préconisé par **Knuth** (1992). Le format permet une interaction naturelle entre le code écrit en format *R*, la sortie de ce code et un commentaire sur le code. Les documents **Sweave** prétraités par *R* pour produire un document **LATEX** qui peut ensuite être traité par des méthodes conventionnelles. Beaucoup de packages *R* ont maintenant des vignettes **Sweave** décrivant leurs fonctionnalités de base. Des exemples de vignettes peuvent être trouvés pour beaucoup d'entre elles des packages *R* y compris celui-ci pour les packages `quantreg` dans le source.

répertoire de distribution inst/doc.

2.13.3 Package Quantreg

Une fois que R est en cours d'exécution, l'installation de packages supplémentaires est assez simple. Pour installer le package de régression quantile à partir de R, il suffit de taper simplement

```
install.packages("quantreg")
```

A condition que votre machine dispose d'une connexion Internet appropriée et que vous ayez les droits d'écriture dans les répertoires système appropriés, l'installation du package devrait se faire automatiquement. Une fois que le package quantreg est installé, il doit être rendu accessible à la session R courante par la commande

bibliothèque(quantreg)

Ces procédures donnent accès à une grande variété de progiciels spécialisés pour l'analyse statistique. Au fur et à mesure que nous avançons, une variété d'autres forfaits seront sollicités. Les services d'aide en ligne sont disponibles selon deux modalités. Si vous savez exactement ce que vous recherchez et que vous souhaitez simplement vérifier les détails d'une commande particulière, vous pouvez, par exemple, essayer

```
help(package = "quantreg")
help(rq)
```

La première commande donne un bref résumé des commandes disponibles dans le package, et la seconde demande des informations plus détaillées sur une commande spécifique. Un raccourci pratique pour cette dernière commande est de taper simplement `?rq`. Plus généralement, on peut initier une session d'aide du navigateur web avec la commande

```
help.start()
```

et naviguer comme vous le souhaitez. L'approche par navigateur est mieux adaptée aux enquêtes exploratoires, alors que l'approche en ligne de commande est mieux adaptée aux enquêtes de confirmation.

Une caractéristique précieuse des fichiers d'aide de R est que les exemples utilisés pour illustrer sont exécutables et peuvent donc être collées dans une session R ou exécutées en tant que groupe avec une commande telle que

```
exemple(rq)
```

Les exemples pour la commande `rq` de base incluent une analyse des données de Brownlee stackloss data : d'abord la régression médiane, puis la première régression quantile est calculée, puis le processus complet de régression quantilee. Une caractéristique curieuse de cet ensemble de données souvent analysées, mais qui est très difficile à trouver sans ajustement de régression quantile, est le fait que les 8 des 21 points tombent exactement sur un hyperplan dans l'espace.

Le deuxième exemple dans le fichier d'aide `rq` calcule une valeur univariée

pondérée. en utilisant des données générées de façon aléatoire. Les données originales d'**Engel** (1857) sur la relation entre les dépenses alimentaires et le revenu des ménages sont prises en compte dans le troisième exemple. Les données sont tracées, puis six lignes de régression quantile ajustées sont superposées sur le nuage de points.Le dernier exemple illustre l'imposition de contraintes d'inégalité sur les coefficients de régression quantile à l'aide d'un ensemble de données simulées.

Examinons plus en détail les résultats de la régression médiane pour l'exemple d'Engel. Exécution

```
data(engel)
fit1<-rq(y~x, tau = 0.5, data = engel)
```

assigne le résultat du calcul de régression médiane à l'ajustement de l'objet `fit1`. Dans la commande `rq()` il y a aussi beaucoup d'options. Le premier argument est un "formula" qui spécifie le modèle désiré. Dans ce cas, nous voulions s'adapter à un modèle linéaire bivarié simple et donc la formule est juste `y~x`; si nous avions eu deux covariables, nous pourrions dire, par exemple, `y~x+z`. Variables factorielles c'est-à-dire, variables prenant seulement quelques valeurs discrètes sont traitées spécialement par la formule et donnent lieu à un groupe de variables indicatrices (factices).

Si nous aimerions voir un résumé concis du résultat, nous pouvons simplement taper

```
fit1
Call :
rq(formula = y ~x, tau = 0.5, data = engel)
Coefficients :
(Intercept) x
81.4822474 0.5601806
Degrees of freedom : 235 total ; 233 residual
```

Par convention pour toutes les routines d'ajustement du modèle linéaire R, nous ne voyons que les coefficients estimés et quelques informations sur le modèle en cours d'estimation. Pour obtenir une évaluation plus détaillée du modèle ajusté, il est possible d'utiliser

```
summary(fit1)
Call : rq(formula=y~x, tau=0.5, data=engel)
tau : [1] 0.5
Coefficients :
```

Le tableau qui en résulte donne l'interception et la pente estimées dans la première colonne et les intervalles de confiance pour ces paramètres dans les deuxième et troisième colonnes. Par défaut, ces intervalles de confiance sont calculés par la méthode d'inversion de rang.

TAB. 2.5 – Coefficients de la régression des quantiles pour l'exemple des données engel

	coefficients	lower bd	upper bd
(Intercept)	81.48225	53.25915	114.01156
x	0.56018	0.48702	0.60199

Pour extraire les résidus ou les coefficients de la relation ajustée, on peut écrire

```
r1<-resid(fit1)
c1<-coef(fit1)
```

Ils peuvent ensuite être facilement utilisés dans les calculs ultérieurs.

2.13.4 Orientation d'objet

Une brève digression sur le rôle de l'orientation objet dans R vaut peut-être la peine à ce stade. Les expressions dans R manipulent les objets.

Les objets peuvent être des données sous forme de vecteurs, de matrices ou de tableaux d'ordre supérieur, mais les objets peuvent aussi être des fonctions ou des collections d'objets plus complexes.

Les objets ont une classe et cet identificateur de classe aide à reconnaître leurs caractéristiques spéciales et permet aux fonctions d'agir sur eux de manière appropriée.

Ainsi, par exemple, la fonction `summary` lorsqu'on opère sur un objet de la classe `rq` comme produit par la fonction `rq` peut agir très différemment sur l'objet que si l'objet était d'une autre classe, par exemple, `lm`, indiquant qu'il était le produit de l'ajustement des moindres carrés. Le résumé d'une structure de données comme une matrice ou un cadre de données aurait une autre intention et un autre résultat. Dans les dialectes antérieurs de S et R, les méthodes pour diverses classes étaient distinguées en ajoutant le nom de la classe à la méthode séparée par un point. Ainsi, la fonction `summary.rq` résumerait un objet `rq`, et `summary.lm` résumerait un objet `lm`. Dans un cas comme dans l'autre,

l'objectif principal de la fonction `summary` résumé est de produire des preuves inférentielles pour accompagner les estimations ponctuelles des paramètres.

De même, le tracé de diverses classes d'objets R peut être effectué par l'expression `plot(x)` avec l'espoir que la commande `plot` reconnaîtra la classe de l'objet `x` et procédera en conséquence. Plus récemment, **Chambers** (1998) a introduit une élégante élaboration du cadre de classe, de méthode et de répartition pour S et R.

L'assignation des objets est généralement effectuée par l'opérateur <- et, une fois assignés, ces nouveaux objets sont disponibles pour la durée de la session R ou jusqu'à ce qu'ils soient explicitement retirés de la session.

R est un langage open source et donc tous les fichiers source décrivant la fonctionnalité du langage sont en fin de compte accessibles à l'utilisateur individuel, et les utilisateurs sont libres de modifier et d'étendre la fonctionnalité du langage de la manière qu'ils jugent appropriée.

Pour ce faire, il faut être capable de trouver des fonctions et de les modifier. Cela nous amène quelque peu au-delà de la portée tutoriel de cette vignette ; cependant, il suffit de dire que la plupart des fonctions du package `quantreg` que vous trouverez dans ce qui suit peuvent être vues en tapant simplement le nom de la fonction, peut-être concaténée avec un nom de classe.

2.13.5 Inférence formelle

Il existe plusieurs méthodes alternatives d'inférence sur les coefficients de régression quantile. Comme alternative aux intervalles de confiance de rank-inversion, on peut obtenir un tableau plus conventionnel de coefficients, d'erreurs types, de statistiques `t` et de valeurs `p` à l'aide de la fonction `summary` :

```
summary(fit1, se = "nid")
Call : rq(formula = y~x, tau = 0.5, data = engel)
tau : [1] 0.5
Coefficients :
```

TAB. 2.6 – Coefficients de la régression des quantiles pour $\alpha = 0.5$

	Value	Std.Error	t value	Pr(>|t|)
(Intercept)	81.48225	19.25066	4.23270	0.00003
x	0.56018	0.02828	19.81032	0.00000

Les erreurs-types rapportées dans ce tableau sont calculées pour la formule sandwich de régression quantile et en utilisant la règle de bande passante **Hall- Sheather**.

Pour obtenir la version noyau **Powell** de l'estimation de la matrice de covariance, on spécifie l'option `se="ker"` dans la commande summary. Il est également possible de contrôler les largeurs de bande utilisées avec l'option de largeur de bande. Une autre option disponible dans `summary.rq` est de calculer les erreurs standard bootstrapped. Ceci est accompli en spécifiant l'option `se="boot"` . Il y a actuellement trois variantes du bootstrap disponibles : le bootstrap standard `(x, y)`, la version **Parzen**, **Wei et Ying** (1994), et le bootstrap marginal de la chaîne **Markov** de **He** et **Hu** (2002) et **Kocherginsky**, **He** et **Hu** (2004).

Il est également possible de spécifier `m` des `n` versions du bootstrap dans lesquelles la taille de l'échantillon des échantillons de bootstrap est différente de (généralement plus petite que) la taille de l'échantillon d'origine.

Cette approche de "sous-échantillonnage" présente un certain nombre d'avantages, dont le moindre n'est pas qu'elle peut être considérablement plus rapide que l'échantillonnage complet. `n` sur la version `n`. Par défaut, le résumé produit également une estimation des composants en estimant la matrice de covariance complète des paramètres estimés et de ses éléments constitutifs.

Pour plus de détails, voir la documentation pour `summary.rq`. Dans le cas de l'affaire bootstrap, la matrice complète des réplications bootstrap est également disponible.

Il existe plusieurs options pour la routine d'ajustement de base `rq`. Une option importante qui contrôle le choix de l'algorithme utilisé dans l'ajustement est `method`. par défaut est `method = "br"`, qui invoque une variante de **Barrodale** et **Roberts** (1974) algorithme simplex décrit par **Koenker** et **d'Orey** (1987).

Pour les problèmes comportant plus de quelques milliers d'observations, il vaut la peine de prendre en considération `method = "fn"`, qui invoque l'algorithme de **Frisch-Newton** décrit par **Portnoy** et **Koenker** (1997).

Plutôt que de se déplacer autour de l'extérieur de l'édifice de la contrainte définie comme la méthode simplex, l'approche par points intérieurs incarnée par la méthode l'algorithme de **Frisch-Newton** s'enfouit à l'intérieur de la contrainte fixée vers l'extérieur.

Au lieu de prendre les étapes de descente les plus raides à chaque intersection des bords extérieurs, il prend des étapes de Newton basées sur une forme Lagrangienne de la fonction objective. Des formes spéciales de **Frisch-Newton** sont disponibles pour les problèmes qui incluent des contraintes d'inégalité linéaire et pour les problèmes avec des matrices de conception clairsemées. Pour des problèmes extrêmement importants avec des observations plausiblement échangeables, `method` = "pfn" met en œuvre une version de l'algorithme de **Frisch-Newton.**

avec une étape de prétraitement qui peut accélérer considérablement les choses. Dans les problèmes de taille modérée où l'option simplex par défaut est assez simplex pratique, l'approche de programmation paramétrique pour trouver l'inversion de rang. les intervalles de confiance peuvent être assez lents. Dans de tels cas, il peut être avantageux d'essayer l'une des autres méthodes d'inférence basées sur l'estimation de l'asymptotique. covariance ou de considérer le bootstrap. Les deux approches sont décrites plus en détail dans les pages qui suivent.Fournir une visualisation un peu plus élaborée de l'exemple d'**Engel**, Considérer un exemple qui superpose plusieurs quantile

conditionnel estimé. sur le nuage de points d'**Engel**. Dans la figure (2.5) qui en résulte, la médiane. La ligne de régression apparaît sous la forme d'une ligne continue et la ligne des moindres carrés sous la forme d'une ligne pointillée. Les autres lignes de régression quantile apparaissent en gris.

Notez que le tracé de l'image de la lignes ajustées est facilement accomplie par la convention selon laquelle la commande `abline` recherche une paire de coefficients qui, s'ils sont trouvés, sont traités comme la pente, et l'interception de la ligne tracée. Il existe de nombreuses options qui peuvent être utilisées pour faire avancer les choses affiner l'intrigue. Le bouclage sur les quantiles est également géré de manière pratique par R `syntaxe`. Il est souvent utile de calculer des régressions quantile sur un ensemble discret de τs ; ceci peut être accompli en spécifiant `tau` comme vecteur dans `rq` :

```
data(engel)
xx<-income -mean(income)
fit1<-rq(foodexp~xx,tau=2 :98/100,data=engel)
plot(summary(fit1))
```

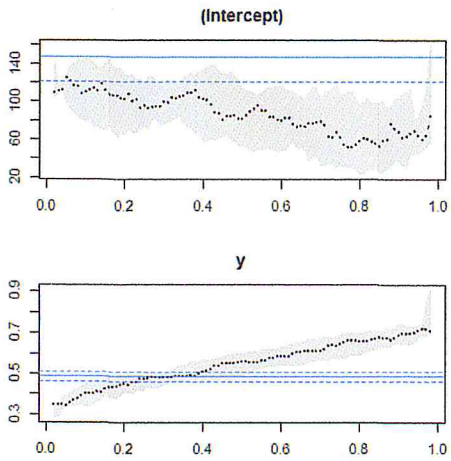


FIG. 2.5 – Diagramme de dispersion et ajustement par regression quantile des données Engel sur les depenses alimentaires.

Le graphique montre un nuage de points des données d'Engel sur les depenses alimentaires par rapport au revenu des menages pour un echantillon de 235 menages belges de la classe ouvriere du 19eme siecle. Les lignes de

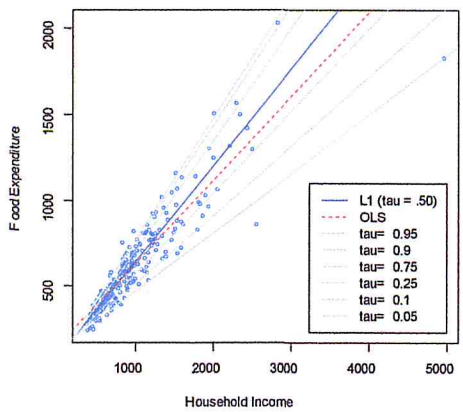
regression{0.05, 0.1, 0.25, 0.75, 0.90, 0.95} engris, l'ajustement median en noir solide et l'estimation des moindres carres de la fonction moyenne conditionnelle en tant que ligne pointillée.

Démonstration d'un tracé de courbe d'Engel dans l'espace d'échantillonnage.

```
data(engel)
plot(foodexp~income, data = engel, cex= .5, col = "blue",
+ xlab = "Household Income", ylab = "Food Expenditure")
z<-rq(foodexp~income, tau= .50, data = engel)
abline(z, col = "dark blue")
abline(lm(foodexp ~income, data = engel), lty=2, col="red")
taus<-c(.05,.1,.25,.75,.90,.95)
nt<-length(taus)
for( i in 1 :length(taus)) {
 + abline(rq(foodexp~income, tau=taus[i], data = engel), col="gray")
+}
legend("bottomright",c("L1 (tau=.50)", "OLS", paste("tau= ", formatC(rev(taus
= c("dark blue", "red", rep("gray", nt)), lty = c(1,2, rep(1, nt)),inset=0.03)
```

Les résultats peuvent être résumés sous la forme d'un graphique :

```
postscript("engelcoef.ps", horizontal = FALSE,width=6.5, height
= 3.5)
plot(fit1, nrow = 1, ncol = 2)
dev.off()
```



Droite de rgression et interval de confiance

Il est à noter que la variable du revenu du ménage a été centrée à sa valeur moyenne pour cette parcelle, de sorte que l'interception est en réalité un concept central et estime la fonction quantile des dépenses alimentaires conditionnelle au revenu moyen.

ou en produisant une table au format LaTex.

```
latex(fit2, caption = "Engel's Law", transpose = TRUE)
```

La commande `postscript` qui précède le tracé indique à R que les instructions pour le tracé doivent être écrites dans un format `postscript` encapsulé et placées dans le fichier `engelcoef.ps`. Ces fichiers sont ensuite commodément inclus dans les documents $LATEX$, par exemple. La commande `dev.off()` ferme le périphérique `postscript` courant et termine la figure.

Les lignes horizontales dans les diagrammes de coefficients de la figure (2.13.5) représentent l'ajustement des moindres carrés et l'intervalle de confiance qui lui est associé. Dans le cas d'un échantillon, nous savons que l'intégration de la fonction quantile sur l'ensemble du domaine $[0, 1]$ donne la moyenne de la distribution (de l'échantillon) :

$$\mu = \int_{-\infty}^{+\infty} x dF(x) = \int_0^1 F^{-1}(t) dt \ .$$

De même, dans les diagrammes de coefficients, on peut s'attendre à ce que l'intégration des coefficients individuels produise un effet moyen à peu près égal à l'effet moyen estimé par les coefficients associés. coefficient des moindres carrés. Il convient toutefois d'être prudent quant à cette interprétation dans des situations très hétérogènes. En ce qui concerne les données d'**Engel**, il convient de noter que l'option l'interception des moindres carrés est nettement supérieure à n'importe laquelle des courbes de régression de quantile ajustées dans notre diagramme de dispersion initial. L'ajustement des moindres carrés est fortement affecté. par les deux observations périphériques avec des dépenses alimentaires relativement basses ; leurs L'attraction incline la ligne ajustée de façon à ce que son interception soit tirée vers le haut. En fait, le L'interception pour le modèle **Engel** est difficile à interpréter parce qu'il nous demande de considérer les dépenses alimentaires pour les ménages à revenu nul. Centrage de la covariable

de sorte que la moyenne des observations soit égale à zéro, comme nous l'avons fait avant de calculer `fit1` pour le tracé des coefficients, rétablit une interprétation raisonnable du paramètre d'interception. Après centrage, l'estimation des moindres carrés de l'interception est une prédiction de la dépense alimentaire moyenne pour un ménage ayant un revenu moyen, et la régression quantile intercept $\hat{\alpha}(\tau)$ est une prédiction du quantile de l'alimentation τth

TAB. 2.7 – Coefficients de la régression des quantiles pour différentes valeurs de α

Quantiles	(Intercept)	x
	124.880	0.343
0.05	(98.302, 130.517)	(0.343, 0.390)
	95.484	0.474
0.25	(73.786, 120.098)	(0.420, 0.494)
	81.482	0.560
0.50	(53.259, 114.012)	(0.487, 0.602)
	62.397	0.644
0.75	(32.745, 107.314)	(0.580, 0.690)
	64.104	0.709
0.90	(46.265, 83.579)	(0.674, 0.734)

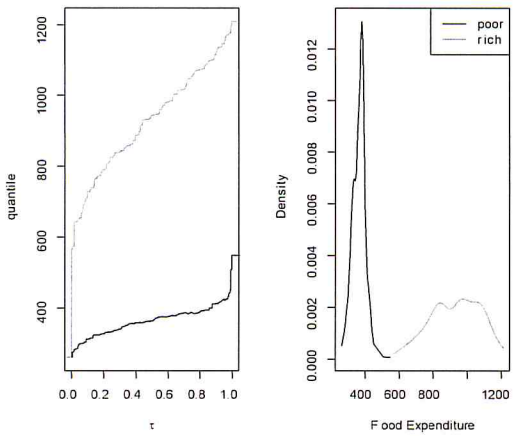
les dépenses des ménages ayant un revenu moyen. Dans la terminologie de **Tukey,** "intercept" est devenu un "centercept". (voir tableau 2.7).

La commande `latex` produit un tableau au format $LATEX$ qui peut être facilement inclus dans les documents. Dans de nombreux cas, la forme tracée des résultats fournira un affichage plus économique et informatif. Il convient de souligner une fois de plus que parce que les fonctions de régression quantile et, en fait, toutes les fonctions R sont open source, les utilisateurs peuvent toujours modifier les fonctions disponibles pour obtenir les effets spéciaux requis pour une application particulière. Lorsque de telles modifications semblent être de l'ordre de d'application générale, il est souhaitable de les communiquer à l'auteur du paquet afin qu'elles puissent être partagées avec l'ensemble de la communauté. Si nous voulons voir toutes les solutions de régression par quantiles distincts pour un on peut spécifier un τ en dehors de la plage $[0, 1]$; par exemple,

```
z<-rq(y~x,tau = -1)
```

Cette forme de la fonction exécute les étapes de programmation paramétrique nécessaires pour trouver le chemin d'échantillonnage complet du processus de régression quantile. L'objet retourné est de classe `rq.process` et a plusieurs composants : la solution primaire dans `z$sol`, et la solution double dans `z$dsol`. En mode interactif, en tapant le nom d'un objet R, le programme imprime l'objet d'une manière raisonnablement intelligible, déterminée par la méthode d'impression désignée pour la classe de l'objet. Encore une fois, le tracé est souvent un moyen d'affichage plus informatif. est une méthode de tracé spéciale pour les objets de la classe `rq.process`.

EstimEstimation des fonctions quantile conditionnelles de y à des valeurs spécifiques de x est également assez facile. Dans le code suivant, nous traçons le tracé de l'estimation empirique



Estimation des fonctions
conditionnelles de quantile et de densit
pour les dpenses alimentaires sur la
base des donnes d'Engel.

Deux estimations sont présentées, l'une pour les ménages relativement pauvres, avec un revenu de 504,5 francs belges, et l'autre pour les ménages relativement aisés, avec 1538,99 francs belges.

data(engel)

```
## do *NOT* attach()
## Poor is defined as at the .1 quantile of the sample distn
## Rich is defined as at the .9 quantile of the sample distn
x.poor<-quantile(engel[,"income"], .10)
x.rich<-quantile(engel[,"income"], .90)
z<-rq(foodexp~income,tau=-1, data = engel)
ps<-z$sol["tau",]
coefs<-z$sol[4 :5,]
qs.poor<-c(c(1,x.poor) %*% coefs)
qs.rich<-c(c(1,x.rich) %*% coefs)
## now plot the two quantile functions to compare
par(mfrow=c(1,2))
plot(c(ps,ps),c(qs.poor,qs.rich),type="n",xlab=expression(tau),ylab="quanti
plot(stepfun(ps,c(qs.poor[1],qs.poor)),do.points=FALSE,add=TRUE)
plot(stepfun(ps,c(qs.poor[1],qs.rich)),do.points=FALSE,add=TRUE,
+ col.hor = "gray", col.vert = "gray")
## now plot associated conditional density estimates
```

```
## weights from ps (process)
ps.wts <- (c(0,diff(ps)) + c(diff(ps),0)) / 2
ap <- akj(qs.poor, z=qs.poor, p = ps.wts)
ar <- akj(qs.rich, z=qs.rich, p = ps.wts)
plot(c(qs.poor,qs.rich), c(ap$dens,ar$dens), type="n",
+ xlab= "Food Expenditure", ylab= "Density")
lines(qs.rich, ar$dens, col="gray")
lines(qs.poor, ap$dens, col="black")
legend("topright",c("poor","rich"),lty=c(1,1),col=c("black","gray"))
```

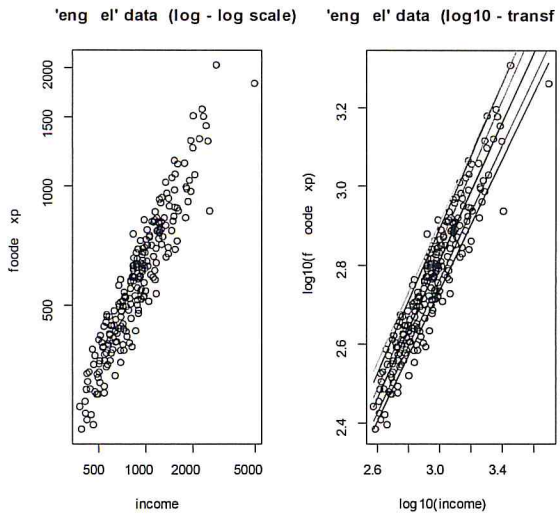


FIG. 2.6 – Représentation des données engel

Notez que l'indicateur $log = "xy"$ produit un tracé à la figure (2.6) avec $log - log$ *axes* et pour la commodité de l'étiquetage de l'axe, ces logarithmes sont à la base 10, et donc l'axe le montage subséquent est également spécifié comme base 10 bûches pour le traçage, même bien que les logarithmes de base 10 ne soient pas naturels et ne seraient jamais utilisés dans les rapports. résultats numériques. Cela ressemble beaucoup plus à une régression d'erreur iid classique. bien qu'un certain écart par rapport à la symétrie soit visible. Un intéressant serait d'effectuer des tests formels pour les écarts par rapport à l'iid. de l'hypothèse du type considéré précédemment.

2.13.6 EN SAVOIR PLUS SUR LES TESTS

Examinons maintenant d'autres formes de tests formels. Une première question naturelle est la suivante : Les relations de régression de quantile

estimées sont-elles conformes à l'hypothèse de décalage d'emplacement qui suppose que toutes les fonctions de quantile conditionnel ont les mêmes paramètres de pente ? Pour commencer, supposons que nous estimons simplement les ajustements de quartile pour les données d'**Engel** et que nous examinons la sortie par défaut

```
fit1<-rq(y~x, tau = 0.25)
fit2<-rq(y~x, tau = 0.5)
fit3<-rq(y~x, tau = 0.75)
```

Rappelons que `rq` ne produit que des estimations de coefficients et qu'un résumé est nécessaire pour évaluer la précision des estimations. C'est très bien pour juger si les covariables sont significatives à des quantiles particuliers, mais supposons que nous voulions tester que les pentes étaient les mêmes aux trois quartiles ? Cela se fait avec l'option anova comme suit :

```
anova(fit1, fit2, fit3)
Quantile Regression Analysis of Variance Table
Model : y ~x
Test of Equality of Slopes : tau in { 0.25 0.5 0.75 }
Df Resid Df F value Pr(>F)
1 2 701 15.557 2.452e-07 ***
Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Voici un exemple d'une classe générale de tests proposés par **Koenker et Bassett.** (1982*a*). Il peut être instructif de regarder le code de la commande `anova.rq` à voir comment ce test est effectué.

L'approche Wald est utilisée et l'approche asymptotique. la matrice de covariance est estimée en utilisant l'approche de **Hendricks et Koenker.** (1991). Il illustre également une syntaxe générale de test en *R* adaptée au présent situation.

Si vous avez estimé deux modèles avec des spécifications de covariables différentes, mais que vous avez le même τ, alors `anova(f0,f1)` devrait tester si le modèle plus restreint est correcte. Notez que cela suppose qu'ils sont imbriqués, avec des ajustements, disons `f0` et `f1`.

Il faut cependant faire attention de vérifier que l'hypothèse qui est visée est la suivante vraiment celui que la commande `anova` comprend (voir `?anova.rq` pour plus de détails). détails sur la version de régression quantile de ce document). Une variété d'autres options sont les suivantes décrite dans la documentation de la fonction `anova.rq`.

2.13.7 RÉGRESSION QUANTILE NON LINÉAIRE

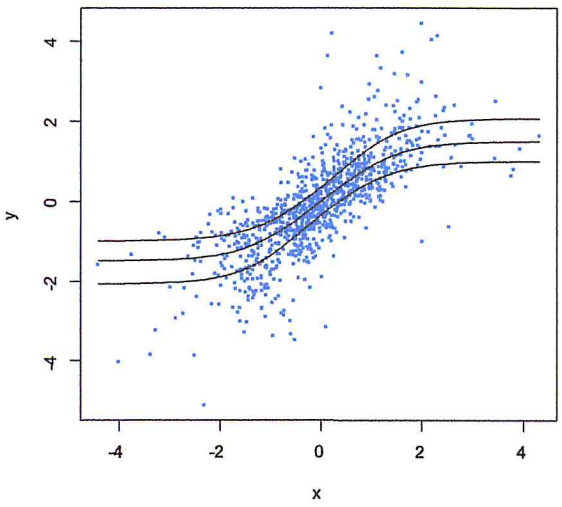
Les modèles de régression quantitative avec des fonctions de réponse non linéaires dans les paramètres peuvent être estimés avec la fonction `nlrq`.

Pour de tels modèles, la spécification de la formule du modèle est un peu plus ésotérique que pour les modèles linéaires ordinaires. mais suit les conventions de la commande *R* nls pour les moindres carrés non linéaires. estimation. Pour illustrer l'utilisation de `nlrq`, considérez le problème de l'estimation du quantile fonctions du modèle de la copule de Frank. Nous commençons par la définition de certains paramètres et la génération de données à partir du modèle de Frank :

```
n <- 1000
df <- 8
delta <- 8
x <- sort(rt(n, df))
u <- runif(n)
v <--log(1-(1-exp(-delta))/(1 + exp(-delta *
+ pt(x, df))*((1/u)-1)))/delta
y <- qt(v, df)
```

Nous traçons les observations de la figure 2.13.7, superposons trois fonctions quantile conditionnelles, puis estimons les mêmes trois fonctions quantile et traçons leurs courbes estimées comme les courbes en pointillés :

```
plot(x, y, pch = ".", col = "blue", cex = 3)
us <- c(0.25, 0.5, 0.75)
```



Estimation quantitative conditionnelle non linaire du modle de la copule de Frank

Les courbes solides sont les fonctions de quantile conditionnel réel et les courbes estimées correspondantes sont indiquées par des courbes enpointillés.

```
for (i in 1 :length(us)) {
u<-us[i]
v<--log(1-(1-exp(-delta))/(1+exp(delta*pt(x,df))*((1/u)-1)))/delta
+ lines(x, qt(v, df))}
Dat <- NULL
Dat$x <- x
Dat$y <- y
deltas <- matrix(0, 3, length(us))
FrankModel <- function(x, delta, mu, sigma, df,
+ tau) {
+ z <- qt(-log(1 - (1 - exp(-delta))/(1 + exp(-delta *
+ pt(x, df)) * ((1/tau) - 1)))/delta, df)
+ mu + sigma * z
+ }
for (i in 1 :length(us)) {
+ tau = us[i]
+ fit <- nlrq(y ~FrankModel(x, delta, mu, sigma,
+ df = 8, tau = tau), data = Dat, tau = tau,
+ start = list(delta = 5, mu = 0, sigma = 1),
+ trace = TRUE)
+ lines(x, predict(fit, newdata = x), lty = 2,
+ col = "green")
+ deltas[i, ] <- coef(fit)
+ }
```

Chapitre 3

Application

3.1 Introduction

Les précipitations principale source d'eau présentent de fortes variabilités spatio-temporelle et une tendance à la baisse sur une grande partie de la méditerranée , en Algérie (**Taibiet al 2013 , MeddietTalia 2007**), au Maroc (**Benassi 2001 , Singlaet al 2010**) en Tunisie (**Kingumbiet al 2005**), en Espagne (**DeLuis et al 2000 , Sinogaet al 2011**), en Italy (**Longobardi et Villani 2009, Caloieroet al 2011**) et en Grèce (**Xoplakiet al 2000, Feidaset al 2007**). Identifier l'origine et les causes de cette modification du régime pluviométrique nécessite une analyse de la variabilité climatique. Il s'agit notamment de comprendre et de représenter les liens entre les facteurs climatiques et l'évolution spatio-temporelle des précipitations.

Dans un contexte de changement climatique l'analyse de la variabilité des précipitations a fait l'objet de plusieurs études et à différentes échelles de temps.

L'analyse des tendances des précipitations dans la région méditerranéenne montre une baisse significative des précipitations à partir des années **1970** (**Xoplakiet al., 2000 Knippertzet al, 2003a , New et al, 2001, Rodrigo and Trigo 2007, Singlaet al, 2010, Meddiet al, 2007**). Cette tendance à la baisse est plus importante en hiver (**Jacobeit, 2000 et Giorgi, 2002**). Selon **Demmaket al. (2001)**, l'Algérie a connu au cours des 25 dernières années (**1975-1998**), une sécheresse intense et persistante qui a touché l'ensemble du territoire , et a été particulièrement rude dans l'Ouest du pays.

L'analyse des séries temporelles de précipitations indique une rupture à partir des années **1970**, et la décennie **1980** a été la plus déficitaire (**Medjrab 2005 , Meddiet al 2007 , Bekkoussaet al 2008 , Meddiet al 2010**

, Taibi 2011).

Par ailleurs, de nombreuses études ont montré que les variations du régime pluviométrique dans le bassin méditerranéen sont liées à la circulation atmosphérique générale tels que : l'Oscillation Nord Atlantique *NAO* **(Salameh 2008 , Xoplaki et al 2004 , Lopez et al 2010 , Brandimarte et al 2011)** *ENSO* **(Meddi 2010 , Kiladis and Diaz 1989 , Rodo et al 1997 , Van Oldenborgh et al 2000 , Lloyd-Hughes and Saunders 2002 , Knippertz et al 2003)**, *MO* **(Mediterranean oscillation) (Conte 1989)** et *WeMO* **(West MediterraneanOscilation) (Martin-Vide and Lopez-Bustins 2006)**.

En Algérie, l'étude menée par **Meddi et al (2010)** montre que la variabilité temporelle des précipitations de l'Ouest du pays est influencée par *ENSO*. Il n'existe pas d'autres études qui mettent en évidence l'influence de la circulation atmosphérique générale sur les précipitations dans le Nord de l'Algérie.

Les facteurs (**indices**) climatiques à l'origine de la variabilité temporelle pluviométrique du Nord de l'Algérie, soumis aux conditions climatiques méditerranéennes et aux influences atlantiques sont donc mal maitrisés.

Ce travail vise à mettre en évidence l'influence de quatre modes de circulation atmosphérique générale sur la variabilité des précipitations du **Nord-Ouest** de l'Algérie à l'échelle annuelle et mensuelle.

3.2 Les indices climatiques

Pour expliquer la variabilité pluviométrique de la zone d'étude nous avons analysé les indices climatiques de 4 modes de la circulation atmosphérique : *ENSO*, *NAO*, *MO* et *WeMO*. Ces indices climatiques représentent une différence dépression calculée entre deux points (un de haute pression et l'autre de basse pression).

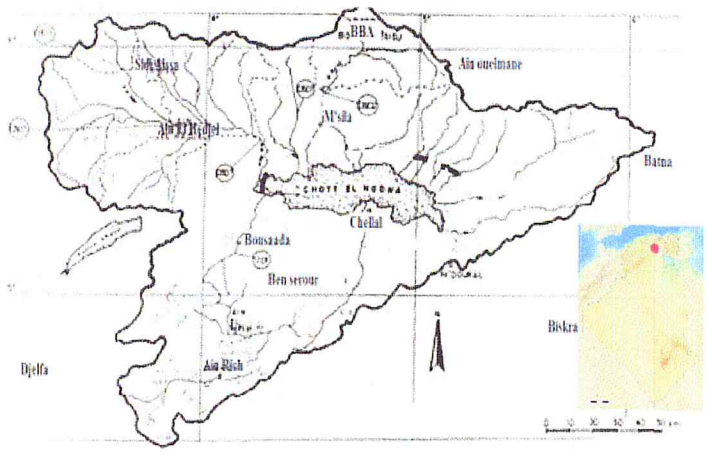
- L'indice **NAO** caractérise la circulation atmosphérique de l'hémisphère Nord et représente la différence de pression entre l'anti-cyclone des Açores et la dépression d'Islande.
- L'indice **MO** représente une circulation atmosphérique régionale qui caractérise le bassin Méditerranéen. Il correspond à la différence de pression entre Alger et le Caire.
- L'indice **WeMO** a été défini comme la différence dépression entre les régions du nord de la péninsule italienne et le sud-ouest de la péninsule ibérique afin de caractériser la variabilité climatique de la région Ouest de la méditerranée.
- Le phénomène **ENSO** caractérise quant à lui la circulation atmosphé-

rique de l'hémisphère Sud mais il a été démontré que des épisodes intenses d'ENSO ont une influence sur la circulation atmosphérique générale.

3.3 Zone d'étude

La plaine du **Hodna** est enserrée dans un cadre montagneux, le mot désigné pour ses habitants une région aux caractères précis : ce sont les plaines situées à l'Est et au Nord d'un vaste Sebkha plaines. L'un des plus vastes ensembles des zones arides et steppiques qu'ils existent dans le mondes, Le bassin versant du **Hodna** c'est un bassin endoréique, situé au cœur de la steppe Algérienne, il est parmi les 5 premiers bassins versant de l'Algérie située au sud des hauts plateaux ayant pour capitale M'sila, située au Sud-Est à 248 km d'Alger il est occupée la région centrale du territoire algérien, il couvre près de 25920 km^2, avec 1,094,000 habitants jusqu'au 2008, on trouve le bassin entre deux séries de montagnes, autour d'une cuvette fermée à 400 m d'altitude, il est limité par :

- Au Nord par Atlas tellienne
- Au Sud par Atlas saharien
- Au Nord-Est par les montagnes de Bibans et les plateaux Sétifien
- Au Nord-Ouest par les hautes plaines Algéroises
- Au Sud-Ouest les monts de Boussaada, terminaison des monts des Oueleds Nail



Localisation gographique du bassin l'Hodna
(F.A.O,1973)

3.3.1 Réseaux hydrographiques

Le réseau hydrographique est l'ensemble des cours d'eau, affluents et sous affluents, permanents ou temporaire, par lesquels toutes les eaux de ruissellement et convergent vers un seul point de vidange du bassin versant (exutoire) (Blagoune).

Le réseau hydrographique est tout diversifié, mais il se distingue par sa dégradation et le régime des oueds Hodnéens se caractérise aussi par l'existence d'un grand nombre d'année séche entrecoupé par un petit nombre d'année humide avec crues violentes.

Les oueds (cours d'eau temporaires à écoulement principal sous forme de crue) se déversent dans la dépression du Chottel-Hodna.Deux grands réseaux convergent vers cette dépression : au nord, l'oued Ksob draine les eaux des versants des Monts du Hodna, au sud l'oued Boussaada, l'oued Chair et l'oued Melh drainent celles des versants d'Atlas saharien .

3.3.2 Géologie et Géomorphologie

Les principaux caractères géologiques et géomorphologiques du territoire envisage peuvent ainsi résumes :

- **Les reliefs :** sont composes d'une alternance de marnes argileuses et de niveaux calcaires relevant du Cénomanien.
- **Les glacis (chebket) :** surfaces plus ou moins planes constituées par des dépôts alluviaux du quaternaire.
- **Les dépressions :** zones de concentration des eaux de ruissellement et de décantation des particules solides, elles correspondent à deux types selon leur caractère sale (sebkha, chott) ou non sale (daya).
- **Les dunes :** amas de sable quartzeux, souvent riche en matériel argileux .

3.3.3 Climat

Hodna est une steppe plus chaude et plus sèche que les hautes steppes de l'Algérois et de l'Oranie.Sa position au fond d'une cuvette lui valent une sècheresse et une pluviométrie capricieuse qui préfigurent le Sahara.

La situation géographique nous donnons une année climatique partagée entre deux grandes saisons : une saison fraiche et relativement humide et une saison sèche.

Les données de base de cette étude sont constituées de pluviométriques mensuels, température moyenne mensuelle et humidité mensuelle provenant

de trois stations météorologiques **(Tableau 3.1)** de l'ANRH de M'sila elles sont utilisées pour le calcul de l'indice de sécheresse (RDI) et sa tendance.

Les données des trois stations Beniou et Ben serrour et Chellal couvrent une période de $1979 - 2013$, le choix de ces stations est également effectué tenant compte la disponibilité des données.

3.4 Présentation des trois stations

TAB. 3.1 – Cordonnées des trois stations

Nom de la station	*Code*	*Longitude(m)*	*Latitude(m)*	*Altitude*	*Période d'observation*
Baniou	051801	4.380	35.440	404	$1979 - 2013$
Bensrour	052101	4.687	35.438	369	$1979 - 2013$
Chellal	050701	5	35.438	371	$1979 - 2013$

3.4.1 La pluviométrie

C'est l'un des paramètres principaux du régime hydrologique puisque c'est le facteur générateur de l'écoulement. Les données de précipitations s'étalent sur la période allant de $1979 - 2013$ pour les trois stations.

Les précipitations se caractérisent par une répartition saisonnière avec deux maxima en automne et au printemps. Cependant, les moyennes pluviométriques ne font apparaitre ni le caractère aléatoire des pluies, ni l'importance des averses de forte intensité qui provoquent le ruissellement et accentuent le déséquilibre du bilan hydrique déjà déficitaire.

3.4.2 Variation annuelle des précipitations

La variation annuelle des précipitations pour la période de 1979-2013 est représentée sur le graphe IV.3, et dans **le tableau 3.**.

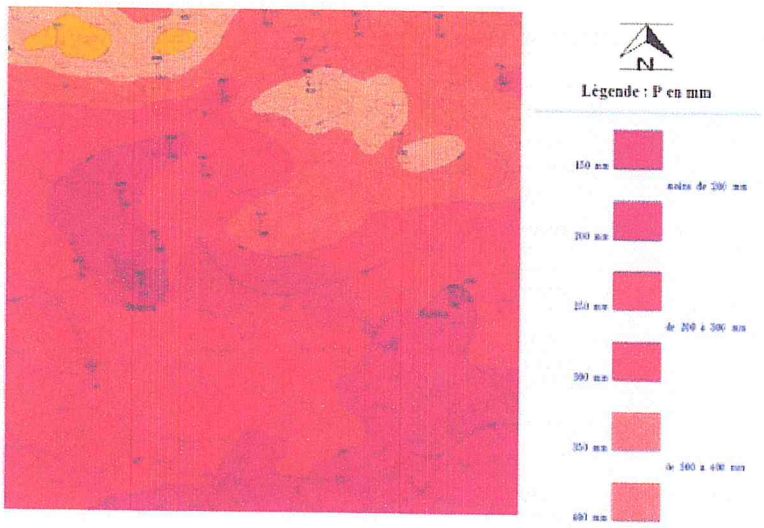
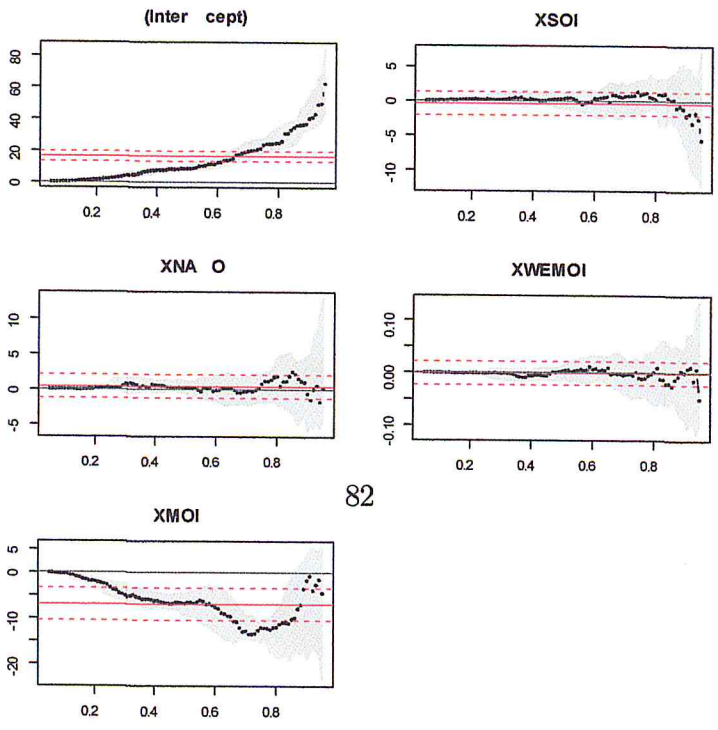


FIG. 3.1 – Carte pluviométrique (ANRH1/500000 :1922/60-1969/89).

TAB. 3.2 – Résumé sur la série statistique des précipitations annuelles

Précipitation	
min	0.000
1st Qu	1.978
median	9.664
mean	16.325
3rd Qu	22.037
max	139.420



TAB. 3.3 – Résumé sur les indices climatiques

	SOI	NAO	WEMOI	MOI
min	-3.60	-3.18000	1.0 Min	-1.52879
1st Qu	-0.50	-0.68500	53.0	-0.32388
Median	0.05	0.07804	109.5	0.01219
Mean	0.06	0.04037	119.2	-0.01403
3rd Qu	0.70	0.79250	184.2	0.34031
Max	2.90	2.63000	258.0	1.26489

TAB. 3.4 – Quelques résidues

Residuals	
min	-26.000
1Q	-12.707
median	-7.276
3Q	6.309
max	120.994

TAB. 3.5 – Coefficients et paramètres dela régression pour $\alpha = 0.25$

Coefficients				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.405796	1.912941	8.576	< 2e-16
XSOI	-0.389544	1.034327	-0.377	0.70665
XNAO	0.392425	1.001916	0.392	0.69550
XWEMOI	-0.001423	0.013634	-0.104	0.91692
XMOI	-6.893406	2.127378	-3.240	0.00129

TAB. 3.6 – Intervalles deconfiances des paramètres estimés pour $\alpha = 0.25$

Coefficients			
	coefficients	lower bd	upper bd
(Intercept)	2.64947	1.64412	4.54144
XSOI	0.29172	-0.28799	0.83992
XNAO	0.22739	-0.29067	0.97283
XWEMOI	-0.00198	-0.00749	0.00243
XMOI	-3.30789	-5.31525	-1.88689

TAB. 3.7 – Coefficients de la régression des quantiles pour $\alpha = 0.5$

Coefficients			
	coefficients	lower bd	upper bd
(Intercept)	8.65687	7.79178	11.65063
XSOI	11.65063	-1.22968	1.21103
XNAO	0.27101	-0.74966	1.17205
XWEMOI	0.00280	-0.01402	0.01372
XMOI	-6.72582	-9.31927	-4.63549

TAB. 3.8 – Intervalle de la confiance des paramètres estimés pour $\alpha = 0.5$

Coefficients			
	coefficients	lower bd	upper bd
(Intercept)	23.72181	18.81299	27.68654
XSOI	0.94094	-1.98627	2.94865
XNAO	0.79844	-1.80775	2.18885
XWEMOI	-0.00976	-0.04884	0.02386
XMOI	-12.04009	-15.42727	-11.16646

TAB. 3.9 – Coefficients de la régression des quantiles pour $\alpha = 0.75$

Coefficients			
	coefficients	lower bd	upper bd
(Intercept)	2.64947	1.64412	4.54144
XSOI	0.29172	-0.28799	0.83992
XNAO	0.22739	-0.29067	0.97283
XWEMOI	-0.00198	-0.00749	0.00243
XMOI	-3.30789	-5.31525	-1.88689

TAB. 3.10 – Intervalle dela confiances des paramètres estimés pour $\alpha = 0.75$

Coefficients			
	coefficients	lower bd	upper bd
(Intercept)	23.72181	18.81299	27.68654
XSOI	0.94094	-1.98627	2.94865
XNAO	0.79844	-1.80775	2.18885
XWEMOI	-0.00976	-0.04884	0.02386
XMOI	-12.04009	-15.42727	-11.16646

TAB. 3.11 – Intervalle de la confiance pour $\alpha = 0.25$ avec iid=True

Coefficients			
	coefficients	lower bd	upper bd
(Intercept)	2.64947	1.53214	4.72060
XSOI	0.29172	-0.35913	1.02287
XNAO	0.22739	-0.34232	1.13000
XWEMOI	-0.00198	-0.00967	0.00544
XMOI	-3.30789	-5.52209	-1.69060

TAB. 3.12 – Intervalle de confiance pour $\alpha = 0.25$ avec iid=False

Coefficients			
	coefficients	lower bd	upper bd
(Intercept)	2.64947	1.48025	4.77829
XSOI	0.29172	-0.36405	1.02287
XNAO	0.22739	-0.34232	1.20901
XWEMOI	-0.00198	-0.00968	0.00548
XMOI	-3.30789	-5.90780	-1.03061

Chapitre 4

CONCLUSION

Nous avons abordé, dans ce rapport de mémoire, deux parties : Une partie consacrée à l'estimation des quantiles conditionnels et une autre, à l'application sur des données réelles.

Dans la première partie, nous avons étudié les méthodes d'estimation du quantile

conditionnel : La régression quantile constante(1990),

Nous recommandons la régression quantile linéaire locale comme une approche directe, et nous avons remarqué que ces estimateurs peuvent se mettre sous une forme unifiée.

La performance de notre estimateur est montrée par des expériences numériques sur des

données simulées et des données réelles.

Dans la deuxième partie, et après avoir présenté les méthodes, nous avons montré la performance de ces méthodes au moyen des expériences numériques sur des données météorologiques.

L'Algérie depuis plusieurs années et comme tous les pays méditerranéens surtout a connu un changement de climat très remarquable ,La présente étude s'est focalisé sur l'analyse des données des pluies aux échelles mensuelles enregistrées dans trois (03) stations météorologiques toutes situées dans la région EL HODNA de l'Algérie pendant la période (1979-2013) d'une part, et leur relation avec quatre indices climatiques, d'autre part. Pour cela nous avons appliqué la méthode des régressions quantiles par le logiciel R.

les résultats fournis en se basant aussi sur une analyse des variances ANOVA ont révélé que la variable MOI est le facteur déterminant parmis les autres, ce dernier est négativement corrélé avec la variable indépendante la précipitation mensuelle, quant autres facteurs WeMOI , SOI et NAO n'ont pas une influence sur la réponse (précipitation mensuelle) .

La représentation graphique d'une sortie sous R pour la précipitation

montre que les courbes relatives aux deux facteurs SOI et NAO sont dans l'intervalle de confiance ou leurs impacts n'apparaissaient que pour le troisième quantile 75% d'où l'utilité de la méthode des quantiles de régression, tandis que one ne constate aucune influence de la variable WeMOI sur la réponse.

Bibliographie

[1] MATHERON, Georges. Les variables régionalisées et leur estimation : une application de la théorie des fonctions aléatoires aux sciences de la nature. Masson et CIE, 1965.

[2] Deheuvels, P. (1974). Valeurs extrémales d'échantillons croissants d'une variable aléatoire réelle. Ann. Inst. H. Poincaré, 10, 89-114.

[3] Fréchet, M. (1928). Sur la loi de probabilité de l'écart maximum. In Annales de la societe Polonaise de Mathematique. [sn].

[4] FERNIQUE, Xavier M. Fonctions aléatoires gaussiennes, vecteurs aléatoires gaussiens. Département de mathématiques et d'informatique, Université de Sherbrooke, 1994.

[5] BLOCH, Ernst. Le principe espérance. Gallimard, 1976.

[6] Spiegel, M. R., Ergas, A., & Marcotorchino, J. F. (1972). Théorie et applications de la statistique. New York : McGraw-Hill.

[7] Ouarda, T., St-Hilaire, A., & Bobée, B. (2008). Synthèse des développements récents en analyse régionale des extrêmes hydrologiques. Revue des sciences de l'eau/Journal of Water Science, 21(2), 219-232.

[8] Koenker, R., & Hallock, K. F. (2001). Quantile regression. Journal of economic perspectives, 15(4), 143-156.

[9] GANNOUN, Ali, SARACCO, Jérôme, et YU, Keming. Nonparametric prediction by conditional median and quantiles. Journal of statistical Planning and inference, 2003, vol. 117, no 2, p. 207-223.

[10] Koenker, Roger W., and Vasco d'Orey. "Algorithm AS 229 : Computing regression quantiles." Journal of the Royal Statistical Society. Series C (Applied Statistics) 36.3 (1987) : 383-393.

[11] KOENKER, Roger et HALLOCK, Kevin F. Quantile regression. Journal of economic perspectives, 2001, vol. 15, no 4, p. 143-156.

[12] CHERNOZHUKOV, Victor et HANSEN, Christian. Instrumental variable quantile regression : A robust inference approach. Journal of Econometrics, 2008, vol. 142, no 1, p. 379-398.

[13] Fack, G., & Landais, C. (2009). Les incitations fiscales aux dons sont-elles efficaces ?. Economie et statistique, 427(1), 101-121.

[14] D'haultfœuille, X., & Givord, P. (2014). La régression quantile en pratique. Economie et statistique, 471(1), 85-111.

[15] Schennach, S. M. (2008). Quantile regression with mismeasured covariates. Econometric Theory, 24(4), 1010-1043.

[16] Koenker, R., & Bassett Jr, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. Econometrica : Journal of the Econometric Society, 43-61.

[17] Révillion, S., & Tuffreau, A. (1994). Les Industries laminaires au paléolithique moyen : actes de la table ronde internationale organisée par l'ERA 37 du CRA-CNRS à Villeneuve-d'Ascq, 13 et 14 novembre 1991 (Vol. 18). CNRS.

[18] GIBRAT, Robert. Les inégalités économiques : applications : aux inégalitês des richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc : d'une loi nouvelle : la loi de l'effet proportionnel. Librairie du Recueil Sirey, 1931.

[19] DUVAL, R. Graphiques et équations. In : Annales de Didactique et de Sciences cognitives. 1988. p. 235-253.

[20] De Micheaux, P. L., Drouilhet, R., & Liquet, B. (2011). Le logiciel R : Maitriser le langage-Effectuer des analyses statistiques. Springer Science & Business Media.

[21] Ricordeau, G., Boccard, R., Damiani, C., & Van Willigen, A. (1961). Relations entre la quantité de lait consommé par les agneaux et leur croissance. In Annales de zootechnie (Vol. 10, No. 2, pp. 113-125).

[22] Huet, S., Jolivet, E., & Messéan, A. (1992). La régression non-linéaire : méthodes et applications en biologie. Quae.