Faculté des Sciences

Département de Mathématique

Régression quantile bayesienne

Année académique
2008-2009

Mémoire présenté par
Nadia Dardenne en vue
de l'obtention du grade de
Master en Statistiques, orientation
générale,à finalité spécialisée

Remerciements

Tout d'abord, je souhaiterais remercier mon promoteur M. Philippe Lambert pour sa disponibilité, ses remarques constructives ainsi que pour l'intérêt porté à mon travail tout au long de cette année.

Un grand merci ensuite à M^{me} Gentiane Haesbroeck pour son soutien, sa disponibilité tout au long de ces deux dernières années ainsi que pour sa motivation de faire en sorte que ce master continue d'être et soit amélioré.

Je tiens également à remercier M. Jean Schmets, et en son nom tout le département, pour son aide, d'un point de vue administratif mais aussi et surtout morale, lors des années académiques 2003-2004-2005.

Je souhaiterais, par la même occasion, remercier infiniment le Docteur Lampertz, le personnel du service d'oncologie et de l'hôpital du jour du Bois de l'Abbaye, ainsi que le professeur Carlier, pour leur franchise, leur gentillesse, leur soutien. Merci de m'avoir permis d'y croire...

Enfin, et surtout, merci à mes parents, ma famille, mes amis (ceux qu'on ne compte que sur les doigts d'une seule main comme on dit) d'avoir été et d'être toujours présent et ce quelque soit les circonstances. Je sais que cela n'a pas dû être facile pour vous de devoir supporter la maladie, les incertitudes, mes sautes d'humeur, mes angoisses. Merci pour tout...

Table des matières

Introduction

En régression, on s'intéresse à l'évolution de la moyenne d'une variable dépendante conditionnellement à des valeurs fixées des covariables où les paramètres de régression sont habituellement estimés par la méthode des moindres carrés.

Néanmoins, on pourrait désirer vouloir plus d'informations à propos de la distribution conditionnelle de la variable réponse. En effet, supposons vouloir étudier les dépenses alimentaires d'un ménage en fonction de son revenu. Via la méthode des moindres carrés, nous pourrions estimer l'effet qu'a, en moyenne, cette covariable (revenu) sur la variable dépendante (dépenses alimentaires), mais nous n'aurions aucune indication de l'effet de cette covariable chez les familles dépensant le moins (ou le plus) en alimentation. Or, il se pourrait que cet effet ne soit pas le même (peut être que, concernant les individus dépensant peu en alimentation, cette covariable a moins d'effet que chez ceux dépensant le plus...).
Une analyse plus complète de l'effet de la covariable peut ainsi être obtenue par estimation des quantiles conditionnels de la variable dépendante, et c'est ce en quoi consiste la régression quantile.

Ce mémoire se veut être une présentation de diverses approches de la régression quantile, approches aussi bien fréquentistes que bayesiennes et ce qu'elles soient paramétriques ou non paramétriques, en mettant l'accent néanmoins sur les méthodes bayesiennes trouvées dans la littérature.

Dans ce qui va suivre, nous commencerons par rappeler, dans le premier chapitre, en quoi consiste une régression, ainsi que la méthode des moindres carrés selon laquelle sont habituellement estimés les paramètres de régression. Nous rappellerons également la définition théorique et empirique d'un quantile. La dernière section de ce chapitre sera quant à elle consacrée à une courte introduction à la statistique bayesienne en y énonçant ses idées fondamentales, le lecteur n'étant peut être pas coutumier de cette approche de la probabilité, et en y mentionnant également ses différences de point de vue avec la statistique fréquentiste, ses avantages mais aussi ses inconvénients.

Le chapitre 2 constituera en un panorama des quantiles de régression en statistique fré-

quentiste. Nous y présenterons d'abord la théorie de Koenker [12], théorie la plus connue, la plus utilisée en pratique et basée sur le fait qu'un quantile d'ordre p peut être dérivé à partir d'un problème de minimisation d'une fonction appellée fonction de perte. Néanmoins, comme nous le verrons, cette théorie présente certaines limites, comme le fait de ne pas être adéquate lorsque la relation entre covariable et variable dépendante ne semble pas être linéaire, c'est pourquoi nous introduirons deux techniques non paramétriques de la régression quantile : l'une reposant sur l'utilisation des B-splines [3] et l'autre sur l'approche par polynômes locaux [30]. Enfin, ce chapitre se clôturera par la présentation d'une méthode se basant sur l'utilisation des modèles additifs généralisés [17], méthode présentant quelques avantages par rapport aux trois précédentes, notamment le fait de ne pas devoir résoudre autant de problèmes de minimisation que de quantiles conditionnels souhaités.

Les approches bayesiennes seront quant à elles présentées dans le chapitre 3 et reposent essentiellement sur l'idée de supposer une distribution asymétrique de Laplace de paramètre de localisation μ nul, de paramètre d'échelle σ^2 et de paramètre d'asymétrie p fixé selon le quantile conditionnel d'ordre p recherché comme distribution pour les erreurs du modèle de régression envisagé. Le choix d'une telle distribution se justifie dans le fait que, comme nous le démontrerons, le paramètre de localisation μ est le quantile d'ordre p d'une telle distribution. Dès lors, pour un p fixé, le quantile d'ordre p des erreurs est nul ainsi qu'il est supposé en régression quantile.
Nous expliciterons dans ce chapitre 3 essentiellement deux méthodes bayesiennes : la méthode de Yu et al (2001) [28] suivie de celle de E.G. Tsionas (2003) [21]. Ces deux approches reposent donc sur la même hypothèse concernant la distribution des erreurs, la seule différence étant la stratégie d'exploration de la distribution a posteriori. Enfin, pour conclure ce chapitre, nous présenterons les idées fondamentales d'une techniques bayesienne non paramétrique utilisée dans le cadre principalement de la régression médiane [13], cette approche permettant plus de flexibilité dans la modélisation des données car, ainsi qu'il sera mentionné, l'hypothèse d'une distribution asymétrique de Laplace comme distribution des erreurs présente certaines limites de modélisation.

Enfin, le dernier chapitre constituera en une sorte de conclusion avec des applications d'une partie des méthodes présentées dans ce mémoire sur divers échantillons de données reprenant trois cas de figure différents (relation linéaire entre variable dépendante et covariable, échantillon avec plus d'une covariable, relation non linéaire entre variable dépendante et covariable), et ce afin de montrer ce que chaque approche de la regression quantile peut apporter selon le cas de figure envisagé, mais aussi afin de signaler leurs éventuelles limites selon la structure des données.

Chapitre 1

Quelques rappels

1.1 Introduction

Dans ce mémoire, nous allons donc présenter différentes approches des quantiles de régression, en nous intéressant plus particulièrement aux méthodes bayesiennes.

Or, qu'est-ce-qu'un quantile de régression ? Quelle en est la définition mathématique et son interprétation ? De même, que peuvent apporter les méthodes bayesiennes par rapport aux méthodes fréquentistes ? Sont-elles préférables ?

Avant de répondre à ces questions, il convient de commencer ce premier chapitre par un rappel concernant les notions de quantile et de régression, et ce afin principalement de fixer certaines notations.

La dernière section sera quant à elle consacrée à un rappel sur la statistique bayesienne afin de familiariser le lecteur avec cette notion différente de la probabilité, et en y présentant ces principaux avantages par rapport au paradigme fréquentiste.

1.2 Régression

Lorsque la théorie, l'hypothèse scientifique laisse à supposer l'existence d'un lien entre une variable aléatoire Y dite variable dépendante et d'autres variables aléatoires $\mathbf{X}^T = (X_1, ..., X_q)$ appelées covariables ou variables explicatives, la technique de régression est dès lors fréquemment utilisée afin d'établir au mieux cette relation mais également dans le but de prédire les valeurs de Y pour des valeurs fixées des covariables.
Quand ce lien est supposé être linéaire, on parle de régression linéaire et le modèle suivant

est souvent considéré

$$\begin{aligned} Y|\mathbf{x} &= \mathbf{x}^T\boldsymbol{\beta} + \epsilon \\ \text{avec} \quad E[\epsilon] &= 0 \\ \text{et} \quad V[\epsilon] &= \sigma^2 \end{aligned} \tag{1.1}$$

pour une valeur $\mathbf{X} = \mathbf{x}$ donnée.
Le vecteur $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)^T$ est le vecteur des coefficients de régression, coefficients inconnus et à estimer.
Le terme ϵ est quant à lui une variable aléatoire de moyenne nulle, de variance constante, et de fonction de densité f_ϵ supposée inconnue bien que dans certains cas, cette densité soit supposée normale.
Notons que nous aurions pu inclure un intercept β_0 dans le modèle (1.1) qui s'écrirait, pour une valeur $\mathbf{X} = \mathbf{x}$ donnée,

$$Y|\mathbf{x} = \beta_0 + \mathbf{x}^T\boldsymbol{\beta} + \epsilon. \tag{1.2}$$

Du modèle (1.1), il vient

$$E[Y|\mathbf{X}{=}\mathbf{x}] = \mathbf{x}^T\boldsymbol{\beta},$$

autrement dit, l'espérance conditionnelle de Y pour une valeur de $\mathbf{X} = \mathbf{x}$ donnée est une fonction linéaire de ce vecteur et donc, typiquement, en régression, nous sommes intéressés par l'impact que pourraient avoir des variables explicatives sur la moyenne de la variable dépendante.

Les coefficients de régression sont estimés habituellement par la méthode des moindres carrés, ou régression L_2, qui, pour rappel, consiste, en considérant un échantillon aléatoire $\{y_1, ..., y_n\}$ de Y et $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ les vecteurs correspondant aux covariables où $\mathbf{x}_i = (x_{i1}, ..., x_{iq})^T$, à estimer les paramètres de régression en considérant le problème de minimisation suivant

$$\hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b}\in\mathbb{R}^q} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T\mathbf{b})^2,$$

qui consiste en fait à minimiser la somme des carrés des résidus.

Il est évident que la régression linéaire ne convient pas lorsque le relation entre variable dépendante et variable explicative n'est pas linéaire.

Enfin, notons que nous pouvons également effectuer des tests statistiques où sous l'hypothèse nulle, on suppose que le coefficient de régression relatif à une covariable est nulle. Un "non rejet" de cette hypothèse implique alors que nous ne pouvons rejeter l'idée que les

valeurs de la variable dépendante ne dépendent pas des valeurs prises par cette covariable.

Nous ne mentionnerons pas dans cette section les résultats théoriques relatifs aux tests statistiques réalisés dans le cadre de la régression linéaire. Si certains résultats doivent être mentionnés dans la suite, ils le seront dans les sections concernées.

1.3 Quantiles

Rappelons à présent la notion théorique de quantile.

Définition 1.3.1. Soit une variable aléatoire X de fonction de répartition

$F(x) = Prob(X \leq x)$.

Pour tout $0 < p < 1$, on définit le quantile d'ordre p de X par

$$Q_X(p) = \inf\{x : F(x) \geq p\}.$$

Si X est une variable aléatoire continue, le quantile d'ordre p est dès lors

$$Q_X(p) = F^{-1}(p).$$

Une autre définition des quantiles est la suivante.

Définition 1.3.2. Soit une variable aléatoire X. On appelle quantile d'ordre p, avec $p \in (0, 1)$, la valeur $Q(p)$ telle que

$$Prob(X \leq Q(p)) = p$$

Avec cette définition, la plupart des quantiles n'existent pas pour des distributions discrètes. Dès lors, la définition qui sera retenue tout au long de ce mémoire sera la définition (1.3.1).

Comme nous recherchons habituellement les quantiles d'une série d'observations, il convient d'en donner une définition empirique. Il existe plusieurs conventions possibles à ce sujet. Par soucis de cohérence avec le définition retenue dans le cadre théorique, nous retiendrons, dans le cadre empirique, la définition suivante.

Définition 1.3.3. Soit $\{x_1, ..., x_n\}$ une réalisation de la variable aléatoire X à laquelle peut être associé la réalisation $\{x_{(1)}, ..., x_{(n)}\}$ où les valeurs $x_{(i)}$ sont ordonnées.

Pour tout $0 < p < 1$, on définit le quantile empirique d'ordre p de X par

$$Q_n(p) = \inf\{x : F_n(x) \geq p\}$$

où $F_n(x)$ n'est rien d'autre que la distribution empirique de X qui est, pour rappel, définie par

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \Delta_{x_{(i)}}(x) \quad \text{où} \quad \Delta_{x_{(i)}}(x) = \begin{cases} 0 & \text{si} \quad x < x_{(i)}, \\ 1 & \text{si} \quad x \geq x_{(i)} \end{cases}$$

Par conséquent,

$$Q_n(p) = x_{(\lceil np \rceil)}\ ^{1}$$

Par exemple, considérons la médiane de la série ordonnée qui est le quantile d'ordre $\frac{1}{2}$. Par la définition (1.3.3), elle vaut

$$Q_n(0.5) = x_{(\lceil \frac{n}{2} \rceil)}.$$

1.4 Statistique bayesienne

Dans ce mémoire, bien que diverses approches dites fréquentistes, qu'elles soient paramétriques ou non paramétriques, des quantiles de régression seront présentées, notre intérêt se portera plus particulièrement sur les approches bayesiennes.
C'est pourquoi il est bon de rappeler en quelques lignes en quoi consiste, en toute généralité, la statistique bayesienne, en quoi elle se différencie de la statistique fréquentiste, et d'en mentionner ses principaux avantages et inconvénients.

Ce rappel se veut assez bref mais suffisant pour permettre au lecteur peu familier avec la statistique bayesienne d'en comprendre les grands principes afin de faciliter la lecture du chapitre consacré à l'approche bayesienne des quantiles de régression. Si des notions plus complexes de statistique bayesienne devaient être développées, elles le seront dans les sections concernées.

1.4.1 Principales différences entre l'approche bayesienne et fréquentiste

En statistique fréquentiste, la probabilité d'un évènement est vue comme la proportion de réalisation de cet évènement lorsque l'expérience qui lui est associée est répétée un grand nombre de fois. On peut constater que cette définition présente certaines limites. En effet, comment, par exemple, calculer la probabilité qu'"il pleuvra la semaine prochaine" en utilisant cette définition ?

1. Pour rappel, $\lceil x \rceil$ désigne le plus petit entier supérieur ou égal à x.

Ce point de vue est différent de celui envisagé en statistique bayesienne où la notion de probabilité peut être définie (voir [2]) comme la plausibilité qu'un évènement se réalise ou qu'une affirmation soit correcte compte tenu de quelconques informations préalables, informations étant souvent différentes d'une personne à une autre. Dès lors, la quantification de la plausibilité le sera également.
On constate donc que ce point de vue de la probabilité est plutôt subjectif, ce qui explique en partie pourquoi l'appoche fréquentiste est la plus utilisée. Néanmoins, cette notion de plausibilité se base sur un certain nombre de règles qui sont les suivantes.

1. le degré de plausibilité d'un évènement est un nombre réel et plus grand est ce nombre, plus grand est le degré de plausibilité,
2. consistence :
   - si la plausibilité d'un évènement peut être obtenue de plusieurs manières, elles doivent toutes fournir le même degré de plausibilité,
   - on doit toujours tenir compte de toute l'information qui nous est fournie,
   - des états équivalents de connaissances mènent à des plausibilités équivalentes,
3. le sens commun : pas de contradiction avec ce qui nous parait évident.

De ces règles peuvent être démontrées des propriétés bien connues en probabilités telles que la règle du produit, et en déduire ainsi le théorème de Bayes, mais aussi le fait que la plausibilité est toujours comprise entre 0 et 1 et bien d'autres (voir [2]).

Une autre différence entre la statistique bayesienne et fréquentiste est que, dans l'approche classique qu'est la statistique fréquentiste, le(s) paramètre(s) à estimer est (sont) inconnu(s) mais supposé(s) fixe(s), tandis qu'en bayesien, on cherche à évaluer la plausibilité de chaque valeur possible du (des) paramètre(s). La distribution de plausibilité qu'il lui est alors assignée traduit l'incertitude quant à la valeur de ce paramètre. Dès lors, la notion d'intervalle de confiance se voit être également différente selon l'approche envisagée.

En effet, supposons être dans le cas où un seul paramètre, θ, est à estimer. En fréquentiste, nous savons, avant même de collecter les données, que, si l'expérience est répétée un grand nombre de fois, la proportion d'intervalles de confiance construits de niveau $1-\alpha$ contenant θ est de $1-\alpha$. En d'autres termes, si α vaut 0.05, il y a 95% des intervalles de confiance de niveau $1-\alpha$ qui contiendront le paramètre θ. Une fois les données collectées, l'intervalle de confiance ainsi construit contient ou non θ sans que nous puissions le vérifier. Par contre, en bayesien, les intervalles de confiances sont interprétés comme des ensembles de valeurs plausibles pour le paramètre. Si nous reprenons notre exemple, il vient que la plausibilité que θ soit dans l'intervalle de confiance à 95% est de 0.95.

On constate donc que ces deux approches présentent une conception quelque peu différente de la probabilité.

1.4.2 Du théorème de Bayes à la statistique bayesienne

A présent, rappelons quelques notions de base de la statisque bayesienne en commençant par le théorème qui est le fondement du développement de cette branche de la statistique : le théorème de Bayes [2].

Définition 1.4.1. Soit deux variables aléatoires X et Y discrètes définies sur le même espace probabilisé, la probabilité conditionnelle que $X = x$ sachant que $Y = y$ est définie par

$$Prob(X = x|Y = y) = \frac{Prob(X = x, Y = y)}{Prob(Y = y)}$$

où $Prob(X = x, Y = y)$ n'est rien d'autre que la probabilité jointe que $X = x$ et $Y = y$. En d'autres termes, $Prob(X = x, Y = y) = Prob(X = x \cap Y = y)$.

On peut évidemment adapter cette définition dans le cas où les variables sont continues. Dans ce cas, les probabilités sont remplacées par les fonctions de densité et la définition s'écrit alors

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

où $p(x|y)$ est la densité conditionnelle de X, $p(y)$ la densité de Y et où $p(x, y)$ n'est rien d'autre que la densité jointe de X et Y.
La remarque que nous venons de faire vaut également pour le théorème suivant.

Théorème 1.4.1. *Soit deux variables aléatoires X et Y discrètes. Le théorème de Bayes s'énonce comme suit*

$$Prob(X = x|Y = y) = \frac{Prob(Y = y|X = x)Prob(X = x)}{Prob(Y = y)}$$

avec $Prob(Y = y) \neq 0$.

Ce théorème est donc à la base du méchanisme par lequel des connaissances a priori sont converties en connaissances a posteriori en se basant sur les données à disposition.

En effet, supposons vouloir estimer le paramètre θ à partir d'un échantillon de données D. En se basant sur le théorème de Bayes, il vient

$$p(\theta|D) = \frac{Prob(D|\theta)p(\theta)}{Prob(D)}$$

2. Thomas Bayes (1702-1761) : mathématicien britannique et pasteur de l'Église presbytérienne, connu pour avoir formulé le théorème portant son nom, le théorème de Bayes, et publié à titre posthume en 1763.

ou encore

$$p(\theta|D) \propto Prob(D|\theta)p(\theta)$$

étant donné que $Prob(D)$ est une quantité constante, indépendante du paramètre θ et où

- $p(\theta)$ est la distribution a priori du paramètre, distribution supposée continue ici,
- $Prob(D|\theta)$ est la vraisemblance de l'échantillon, la distribution de l'échantillon conditionnellement à θ
- et $p(\theta|D)$ la distribution a posteriori, distribution continue vu l'hypothèse de continuité de l'a priori.

La distribution a posteriori $p(\theta|D)$ contient toute l'information disponible à propos du paramètre θ, information que l'on peut résumer en recherchant mode, médiane, moyenne a posteriori, mais aussi variance a posteriori ou autres quantiles. On appelle intervalle de crédibilité à 95% un intervalle contenant 95% de la probabilité a posteriori.

Comme nous venons de le voir, une connaissance a priori du problème est nécessaire. Cette connaissance peut dépendre d'une situation à une autre, d'une personne à une autre et semble donc quelque peu subjective.

Néanmoins, il ne faut pas considérer une connaissance a priori du problème envisagé comme une nuisance mais plutôt comme une opportunité, une chance de pouvoir inclure de l'information préalablement connue du problème envisagé.

Bien sûr, nous n'avons pas toujours à disposition une connaissance a priori du problème.De même, dans certains cas, nous souhaitons inclure le moins de connaissance possible. Dès lors, dans ce cas, on peut avoir recours à des a priori dits impropres ou non informatifs. Dans ce cas, il faut s'assurer que la distribution a posteriori soit bien une densité, autrement dit, qu'elle vérifie bien les propriétés requises pour qu'elle puisse être considérée comme une densité.

Quelques questions s'imposent à nous concernant cette distribution a priori. En effet, quelle distribution a priori choisir de façon à ce que cet a priori soit le moins informatif possible ? Existe-t-il un critère permettant de répondre à cette question ?

La réponse est oui : c'est le principe du maximum d'entropie, principe qui suggère de choisir comme distribution a priori celle maximisant l'entropie sous certaines contraintes. Signalons pour terminer une propriété fort intéressante dérivée de ce principe du maximum d'entropie et concernant le cas où plusieurs paramètres sont à estimer. Dans ce cas, la distribution a priori jointe de ces paramètres est le produit des distributions a priori de chacun de ces paramètres, ce qui revient à dire, qu'a priori, il y a indépendance entre eux.

1.4.3 Tests d'hypothèses et intervalle de crédibilité

Nous avons déjà mentioné que la distribution a posteriori d'un paramètre θ contenait toute l'information disponible sur ce paramètre. Supposons à présent vouloir répondre à une question précise concernant ce paramètre, c'est-à-dire est-ce que, par exemple, θ est supérieur à une valeur fixée θ_0, cette valeur résultant généralement d'une étude préalable. Pour répondre à ce genre de questions, nous devons résoudre le test

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0.$$

En bayesien, nous résolvons ce genre de test en calculant la probabilité a posteriori de l'hypothèse nulle (voir [1])

$$Prob(\theta \leq \theta_0 | D) = \int_{-\infty}^{\theta_0} p(\theta|D) d\theta.$$

et il y a rejet de H_0 si cette probabilité est strictement inférieure à un niveau d'acceptation α (typiquement, α vaut 0.05%) préalablement choisi.

En fréquentiste, nous effectuons généralement un test d'hypothèse simple contre une alternative composée, c'est-à-dire

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

où nous avons recours au test de Student, si nous effectuons un test sur la moyenne, voire à des tests non paramétriques tel que celui de Wilcoxon ou le test du signe, et dans ce cas θ est la médiane, pour répondre à ce genre de question.

En bayesien, ce genre de question n'a pas vraiment de sens si nous envisageons une distribution a priori continue. En effet, dans la mesure où un a priori continu est envisagé, il en résulte une distribution a posteriori continue et donc nous ne pouvons effectuer le test en calculant la probabilité a posteriori de l'hypothèse nulle et la comparer avec α car cette probabilité est nulle, et ce quelque soit le θ_0 fixé.
Dès lors, on recherche l'intervalle de crédibilité à $(1 - \alpha)\%$ (typiquement l'intervalle de crédibilité à 95%) pour θ à partir de la distribution a posteriori de ce paramètre (voir [1] pour plus de détails). Si θ_0 appartient à cet intervalle, on conclut que θ_0 est une valeur plausible pour le paramètre avec une certaine crédibilité et on ne rejette pas l'hypothèse nulle. Dans le cas contraire, on conclut que θ_0 n'a pas de crédibilité comme valeur plausible pour le paramètre et on rejette l'hypothèse nulle en faveur de l'hypothèse alternative.

1.4.4 Avantages et inconvénients de la statistique bayesienne

Concluons ce chapitre en précisant en quoi l'approche bayesienne présente divers avantages, mais aussi quelques inconvénients, par rapport à l'approche fréquentiste.

Comme nous l'avons déjà signalé, cette approche permet d'incorporer une connaissance a priori du problème même si cela peut présenter néanmoins certains désavantages (dépendance par rapport à la distribution a priori choisie, si celle-ci est impropre, il faut être certain que ce ne soit pas le cas de la distribution a posteriori,...), elle permet aussi de voir les intervalles de confiance de manière plus intuitive comme nous l'avons déjà mentionné. Elle présente également d'autres qualités.

D'une part, l'inférence en statistique bayesienne ne repose pas sur des approximations asympotiques comme cela peut parfois être le cas en statistique fréquentiste. En effet, via la distribution a posteriori des paramètres, on peut obtenir un grand nombre d'informations telles que moyenne, variance, intervalle de crédibilité et calcul de probabilité (exemple :$Prob(\theta > 0.5)$,...) pour résoudre des tests d'hypothèses.

Malheureusement, il y a quelques inconvenients. Comme nous l'avons déjà mentionné, la distribution a posteriori contient toute l'information à propos du(des) paramètre(s). De plus, c'est à partir de cette distribution que nous recherchons les intervalles de crédibilité et/ou calculons des probabilités afin de résoudre des tests d'hypothèses. Si cette distribution a posteriori est connue (par exemple, une loi normale de paramètre de localisation μ et de paramètre d'échelle σ connus), nous n'avons pas besoin d'échantillonner l'a posteriori pour calculer moyenne, variance, quantile ou pour effectuer des tests etc. Par contre, si l'a posteriori n'est pas une distribution connue, nous devons pouvoir en générer un échantillon en espérant que les valeurs obtenues de la moyenne, variance etc ne soient pas trop éloignées de ce qu'elles seraient si cet a posteriori était connu. Afin de générer un tel échantillon, il faut dès lors recourir à des méthodes numériques. Nous reviendrons sur cette problématique et sur ces méthodes numériques dans le chapitre consacré à l'approche bayesienne des quantiles de régression ainsi que dans celui présentant la mise en pratique de certaines méthodes envisagées dans ce mémoire.

Chapitre 2

Panorama des quantiles de régression en statistique fréquentiste

2.1 Introduction

Avant d'envisager les quelques approches bayesiennes trouvées dans la littérature, nous allons passé en revue un certain nombre de méthodes fréquentistes sans toutefois les détailler amplement, le but ici étant d'en comprendre le(s) principe(s) et leur(s) éventuel(s) avantage(s) ou inconvénient(s). Ces méthodes sont de plus implémentées dans divers packages que nous mentionnerons dans les sections concernées et utilisables librement via le logiciel R [1].

La section 2.2 de ce chapitre sera consacrée à l'approche fréquentiste et paramétrique des quantiles de regression telle qu'elle est abordée par Koenker dans [12].
En effet, on ne peut envisager une approche bayesienne de la régression quantile sans connaître la théorie fréquentiste de Koenker car cette théorie est la plus connue et la plus utilisée en pratique.

Néanmoins, cette théorie présentant quelques limites, nous mentionnerons dans la section 2.3 deux approches non paramétriques, une faisant appel à une méthode dite localement polynomiale et une autre utilisant des B-splines.

Enfin, la dernière section sera consacrée à une méthode se basant sur les modèles additifs généralisés, cette méthode présentant quelques avantages par rapport aux précédentes.

1. logiciel libre disponible à l'adresse *http ://www.r-project.org/*.

2.2 Théorie de Koenker

Nous avons dans le chapitre précédent rappelé la définition la plus courante des quantiles, que ce soit d'un point de vue théorique ou empirique.
Néanmoins, les quantiles peuvent être également dérivés à partir d'un problème d'optimisation que nous allons exposer en détails, ce résultat étant fondamental car il est à la base de la théorie des quantiles de régression telle qu'envisagée par Koenker.

2.2.1 Autre approche des quantiles

Soit une variable aléatoire continue X de fonction de répartition $F(x) = Prob(X \leq x)$ et de fonction de densité $f(x)$.

Pour rappel, le quantile d'ordre p, $0 < p < 1$, est $F^{-1}(p)$=inf$\{x : F(x) \geq p\}$.

Koenker démontre dans [12] que le quantile d'ordre p peut être dérivé d'un problème de minimisation qui est un problème typique de la théorie de la décision. Pour ce faire, définissons d'abord la fonction de perte $\rho_p(.)$ (représentée sur le graphique (2.1)), par

$$\rho_p(u) = u(p - I(u < 0)) = \begin{cases} u\,p & \text{si} \quad u \geq 0 \\ u(p-1) & \text{sinon} \end{cases} \tag{2.1}$$

avec $0 < p < 1$.

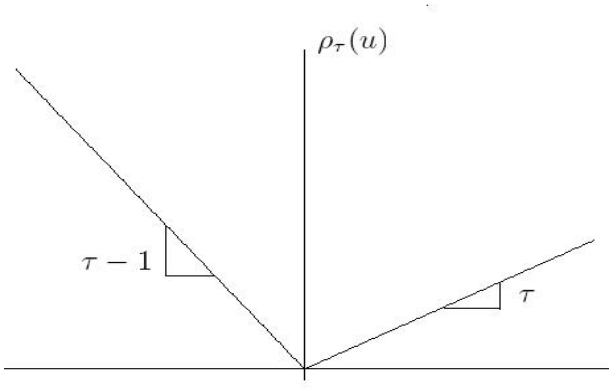

Figure 2.1 – Fonction de perte ρ_p.

Le problème de minimisation envisagé consiste à chercher la valeur $\hat{x}$ de l'argument minimisant la perte attendue. Cela revient à minimiser

$$E_{\rho_p}(X-\hat{x}) = (p-1)\int_{-\infty}^{\hat{x}} (x-\hat{x})\, dF(x) + p\int_{\hat{x}}^{+\infty} (x-\hat{x})\, dF(x).$$

Pour alléger l'écriture, notons $g(\hat{x}) = E_{\rho_p}(X-\hat{x})$.

Dès lors, par le théorème de dérivation des intégrales paramètriques [2], il vient

$$\begin{aligned} \frac{dg(\hat{x})}{d\hat{x}} &= (1-p)\int_{-\infty}^{\hat{x}} dF(x) - p\int_{\hat{x}}^{+\infty} dF(x) \\ &= (1-p)F(\hat{x}) - p(1-F(\hat{x})) \\ &= F(\hat{x}) - p \end{aligned}$$

Comme

$$\frac{d^2 g(\hat{x})}{d\hat{x}^2} = f(\hat{x}) \geq 0 \quad \forall \hat{x},$$

car f est une fonction de densité, le minimum de la fonction $g(\hat{x})$ est atteint en $\hat{x} = F^{-1}(p)$. Autrement dit, la perte moyenne atteint son minimum en le quantile théorique d'ordre p.

Dans le cas où la distribution F est remplacée par la distribution empirique $F_n(x)$, on cherche à déterminer $\hat{x}$ qui minimise

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

pour un échantillon indépendant$\{x_1, ..., x_n\}$ de la variable X. La solution de ce problème d'optimisation est dès lors le quantile empirique d'ordre p.

En conclusion, rechercher le quantile d'ordre p revient à rechercher $\xi(p)$ tel que

$$\xi(p) = \min_{\xi\in\mathbb{R}} \sum_{i=1}^{n} \rho_p(x_i - \xi). \tag{2.2}$$

Remarque 2.2.1. *Pour rappel, la médiane d'une série d'observation $\{x_1, ..., x_n\}$ peut être vue comme étant la solution du problème de minimisation suivant*

$$\tilde{x} = arg\min_{\xi\in\mathbb{R}} \sum_{i=1}^{n} |x_i - \xi|,$$

qui correspond au problème d'optimisation (2.2) dans le cas où p vaut 0.5. On constate donc que la méthode fournie pour calculer les quantiles d'ordre p dans ce chapitre est une sorte de généralisation de celle donnée pour la médiane.

2. Notons que les conditions d'application de ce théorème tel qu'énoncé dans [18] sont vérifiées dans le cas où X est une variable continue, ses fonctions de répartition et de densité étant dès lors continues.

D'un point de vue interprétation, on cherche donc à minimiser une somme pondérée constituée de termes positifs. En effet, la fonction de perte définie en 2.1 peut être réécrite sous la forme suivante

$$\rho_p(u) = u(p - I(u < 0)) = \begin{cases} |u|\, p & \text{si} \quad u \geq 0 \\ |u|(1-p) & \text{sinon} \end{cases} \tag{2.3}$$

On désire donc minimiser

$$\sum_{x_i < \xi} |x_i - \xi|(1-p) + \sum_{x_i \geq \xi} |x_i - \xi| p.$$

Dès lors, si, par exemple, une surestimation du paramètre ξ est trois fois plus coûteuse qu'une sous-estimation de ce paramètre, en d'autres termes, si $1-p$ est trois fois plus grand que p, il faut, afin de minimiser cette somme, choisir ξ tel que $Prob(X \geq \xi)$ soit trois fois plus grande que $Prob(X < \xi)$ pour compenser. Autrement dit, il faut choisir comme valeur de ξ le premier quartile, c'est-à-dire le quantile d'ordre 0.25.

2.2.2 Du quantile au quantile de régression

Le fait que la moyenne d'un échantillon [3] $\{y_1, ..., y_n\}$ d'une variable aléatoire Y soit solution du problème de minimisation

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2$$

suggère que, si nous exprimons la moyenne conditionnelle de Y pour un vecteur de régresseur $\mathbf{x}$ donné comme étant égale à $\mathbf{x}^T \boldsymbol{\beta}$, les coefficients de régression $\boldsymbol{\beta}$ soient estimés en résolvant

$$\min_{\mathbf{b} \in \mathbb{R}^q} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \mathbf{b})^2.$$

Dès lors, étant donné que le quantile d'ordre p est la solution $\hat{\xi}(p)$ de

$$\min_{\xi \in \mathbb{R}} \sum_{i=1}^{n} \rho_p(y_i - \xi),$$

on est amené à spécifier le quantile conditionnel d'ordre p comme étant $Q_Y(p|\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}(p)$, où $\hat{\boldsymbol{\beta}}(p)$ est solution de

$$\min_{\mathbf{b} \in \mathbb{R}^q} \sum_{i=1}^{n} \rho_p(y_i - \mathbf{x}_i^T \mathbf{b}).$$

Nous aboutissons donc à la définition d'un quantile de régression.

3. Nous reprenons les mêmes notations que celles utilisées dans le chapitre 1.

Définition 2.2.1. Dans le contexte tel que défini ci-dessus, le quantile de régression d'ordre p, $\hat{\boldsymbol{\beta}}(p)$, est la solution du problème de minimisation

$$\min_{\mathbf{b}\in\mathbb{R}^q} \sum_{i=1}^{n} \rho_p(y_i - \mathbf{x}_i^T\mathbf{b})$$

pour tout $0 < p < 1$ et un vecteur de régresseur $\mathbf{x}$ donné.

Cette définition s'étend dans un cadre plus théorique.

Définition 2.2.2. Le quantile de régression d'ordre p, $\hat{\boldsymbol{\beta}}(p)$, est la solution de

$$\min_{\mathbf{b}\in\mathbb{R}^q} E\left[\rho_p(Y - \mathbf{x}^T\mathbf{b})\right]$$

où Y est une variable aléatoire et $\mathbf{x}$ un vecteur de régresseur donné appartenant à $\mathbb{R}^q$.

Remarque 2.2.2. *Lorsque nous résolvons le problème de minimisation évoqué ci-dessus, la valeur de p est fixée. Dès lors, nous optenons les solutions pour un p fixé. Si nous voulons obtenir différents quantiles de régession, nous devons donc résoudre autant de problèmes de minimisation que de nombre de quantiles de régression voulus. Il semble donc complexe de trouver toutes les solutions pour chaque valeur distincte de p. Toutefois, nous avons à notre disposition un package nommé "quantreg" dans le logiciel R, package mis au point par Koenker, et que nous utiliserons dans la suite lorsque nous appliquerons les diverses méthodes présentées sur différents échantillons de données.*

Remarque 2.2.3. *Nous avions déjà remarqué précédemment que la méthode fournie dans cette section pour calculer les quantiles d'ordre p était une généralisation de celle donnée pour la médiane.*
On peut également voir la régression quantile comme une généralisation de la régression L_1 qui, pour rappel, consiste à estimer les paramètres de régression $\hat{\boldsymbol{\beta}}$ à partir du problème de minimisation

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^q} \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i^T\boldsymbol{b}|$$

en considérant les mêmes notations que celles utilisées dans le premier chapitre de ce mémoire.

Nous ne développerons pas les nombreuses propriétés des quantiles de régression, le lecteur intéressé par une étude approfondie des quantiles de régression dans le cadre fréquentiste pouvant se référer à [12] ou à [9]. Néanmoins, nous allons citer un résultat important concernant les tests d'hypothèses.

2.2.3 Inférence

Comme en régression linéaire, on pourrait se demander si toutes les covariables sont nécessaires au modèle. En d'autres termes, on se demande si on ne pourrait pas supposer certains coefficients de régression nuls. On désire donc effectuer le test statistique

$$H_0 : \beta_j(p) = 0 \quad \text{versus} \quad H_1 : \beta_j(p) \neq 0$$

où sous l'hypothèse nulle, le coefficient de régression relatif à la j$^{\text{ème}}$ covariable ($j = 1, ..., q$ selon nos notations) est supposé nul.

En général, ce genre d'hypothèse est testée en exploitant les résultats asymptotiques concernant les intervalles de confiance. Il en est de même pour la théorie des quantiles de régression où nous avons le résultat suivant.

Théorème 2.2.1. *Considérons à nouveau le modèle classique envisagé en régression linéaire*

$$Y = \boldsymbol{X}^T\boldsymbol{\beta} + \epsilon$$

en supposant que les résidus, ϵ, soient identiquement et indépendamment distribués selon une loi de densité f_ϵ et de fonction de répartition F_ϵ tel que $f_\epsilon(F_\epsilon^{-}1(p)) > 0$, p étant fixé, et en supposant également que $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T \equiv Q_n$ converge vers une matrice définie positive notée Q_0, c'est-à-dire $\lim_{n\to\infty} Q_n = Q_0$. Dans ces conditions, il vient

$$\sqrt{n}(\hat{\boldsymbol{\beta}}(p) - \boldsymbol{\beta}(p)) \xrightarrow{L} N\left(\ 0,\ \ \frac{(1-p)p}{f_\epsilon^2(F_\epsilon^{-1}(p))}Q_0\ \right).$$

On remarque donc que la variance asymptotique dépend de la fonction de densité des résidus du modèle, densité qui peut s'avérer difficile à estimer (on doit alors avoir recours à des méthodes non paramétriques).

2.2.4 Interprétation

Avant de présenter d'autres approches, il est bon de comprendre l'interprétation donnée aux quantiles de régression via quelques exemples qui se veulent assez simples mais qui

permettront néanmoins de montrer les limites de la théorie de Koenker.
Nous détaillerons plus amplement ces exemples dans le dernier chapitre de ce mémoire.

Nous venons de voir que le quantile conditionnel d'ordre p de Y était considéré comme une fonction linéaire de $\mathbf{x}$, en d'autre termes, que $Q_Y(p|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}(p)$. Dès lors,

$$Prob(Y < Q_y(p|\mathbf{x})|\mathbf{X} = \mathbf{x}) = p$$

et donc le quantile d'ordre p de la distribution des résidus doit être nul.
On constate donc que la droite de régression cherchée, $\mathbf{x}^T\hat{\boldsymbol{\beta}}(p)$ est telle qu'il y a un pourcentage p de points en dessous de cette droite et un pourcentage $1-p$ au dessus.

Passons à quelques exemples afin d'illuster cette notion de quantile de régression.

Exemple 2.2.1. *Dans un premier temps, envisageons l'échantillon suivant "Engel Data" (Koenker et Basset (1982)), échantillon de 235 observations avec comme variable dépendante la dépense alimentaire annuelle du ménage en euros et comme covariable le salaire annuel du ménage en euros.*
Sur le graphique (2.2) *sont représentées les droites de régressions relatives aux valeurs de* $p = \{0.05, 0.25, 0.75, 0.95\}$ *ainsi que celle correspondant à la médiane et la droite de régression linéaire obtenue par la méthode des moindres carrés.*
Une étude plus détaillée de ce graphique suivra mais on peut déjà constater que, par exemple, pour une valeur de la covariable de 20, 25% des dépenses alimentaires seraient inférieures à 11.849 et 75% inférieures à 14.427. Par contre, pour un revenu annuel de 80, 25% des dépenses alimentaires seraient inférieures à 40.295 et 75% inférieures à 53.078. En d'autres termes, les personnes ayant un revenu de 20 ont 25% de chance de dépenser 28.446 en moins et 75% de chances de dépenser 35.651 de moins que ceux ayant un revenu de 80. Ceci tend à confirmer que plus le salaire est élévée, plus les dépenses le sont également et ce d'autant plus qu'on s'intéresse aux quantiles conditionnels plus élevés.

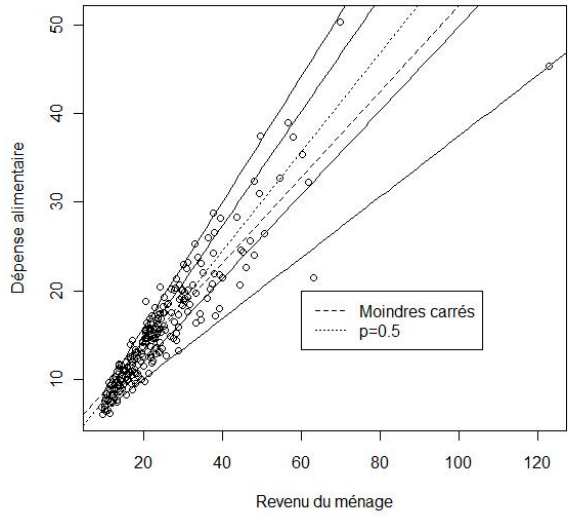

Figure 2.2 – Représentation des doites de régression pour $p = (0.05, 0.25, 0.5, 0.75, 0.95)$ et droite obtenue par la méthode des moindres carrés.

Exemple 2.2.2. *Examinons à présent un autre exemple basé sur l'échantillon "mcycle" disponible dans le package "MASS" du logiciel R. Les données viennent de mesures expérimentales reprenant l'accélération de la tête d'un mannequin utilisé pour des "crash test" de moto afin de tester l'efficacité des casques. La variable dépendante est donc l'accélération de la tête exprimée en G et la covariable est le temps (en millisecondes) après l'impact.*
La représentation des données (graphique (2.3)*) montre que supposer une relation linéaire liant variable dépendante et covariable n'est pas des plus adéquates.*
Dès lors, nous devons recourir à des méthodes plutôt non paramétriques telles que, par exemple, l'utilisation de B-splines pour déterminer les quantiles de régression et obtenir dans ce dernier cas, le graphique (2.4) *pour des valeurs de* $p = \{0.25, 0.5, 0.75\}$.

Il est donc important, vu qu'en pratique nous pouvons tomber sur des échantillons du même type que celui de l'exemple (2.2.2), de présenter quelques méthodes non paramétriques dans le cadre de la régression quantile.

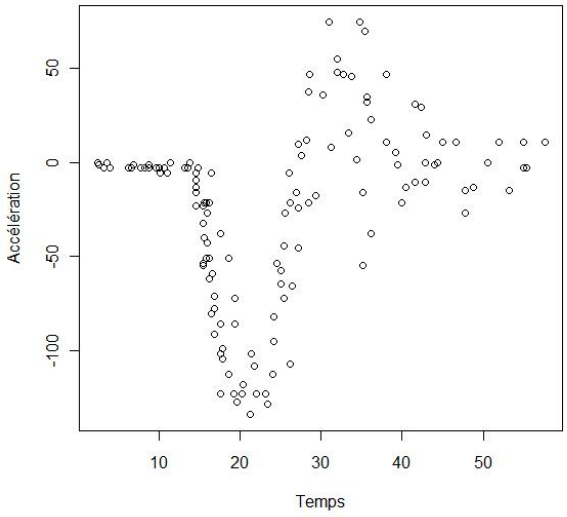

Figure 2.3 – Echantillon "mcycle".

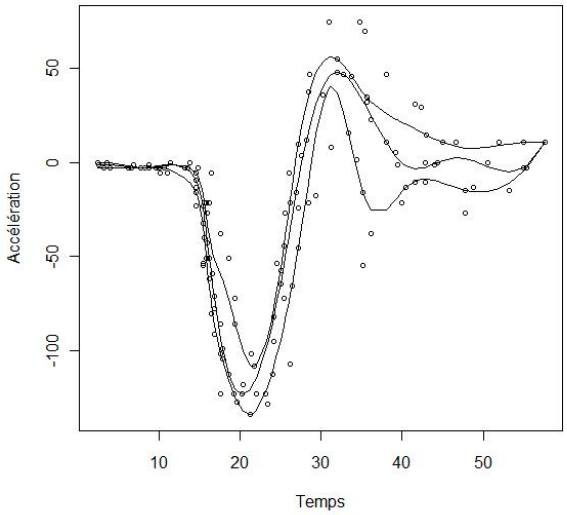

Figure 2.4 – Représentation des courbes de régression pour $p = (0.25, 0.5, 0.75)$ et obtenues par la méthode des B-splines.

2.3 Approche non paramétrique

Dans cette section, deux techniques dites non paramétriques vont être présentées dans le cadre de la régression quantile :

- l'approche localement polynomiale, et
- la méthode des B-splines.

Notons que ces méthodes sont déjà exploitables en régression "classique" [4], nous ne faisons donc ici que les généraliser dans le cadre de la régression quantile.

2.3.1 Approche localement polynomiale

Soit Y la variable réponse et X une covariable. Nous avons donc à disposition l'échantillon d'observations $\{(x_i, y_i), i = 1, ..., n\}$.
Pour rappel, nous désirons estimer le quantile d'ordre p de la variable réponse conditionnellement à une valeur donnée de la covariable, c'est à dire estimer $Q_Y(p|x)$ pour $X = x$.

Comme nous venons de le mentionner, la méthode présentée ici n'est rien d'autre qu'une généralisation de celle considérée en régression "classique". Dès lors, avant d'expliciter en quoi consiste l'approche localement polynomiale en régression quantile, nous allons rappeler quelques rudiments de cette technique dans le cadre de la régression "classique".

De la régression...

Supposons avoir le modèle

$$Y_i = f(x_i) + \epsilon_i$$

où $f(u)$ est une fonction inconnue et ϵ_i le terme d'erreur tel que les ϵ_i soient identiquement et indépendamment distribués selon une loi de densité inconnue, de moyenne nulle et de variance σ^2.

L'idée de la méthode localement polynomiale est d'approximer localement $f(x)$ en utilisant le théorème de Taylor.

Nous considérerons ensuite une fonction de poids $w_i(x) = K(x_i - x/h)/h$ où K est une fonction noyau et h un paramètre d'échelle (*bandwith*) qui contrôle le degré de lissage (*smoothing*) d'une fonction que l'on désire estimer.

4. Nous appellons régression "classique" la régression telle qu'elle est rappellée au chapitre 1.

Pour rappel, un noyau K est une fonction continue et bornée telle que

$$K(-u) = K(u),$$

$$K(u) \geq 0 \, \forall u,$$

$$\int_{\mathbb{R}} K(u)du = 1,$$

autrement dit, K est une fonction de densité symétrique.

Quelques exemples de noyaux sont

- le noyau uniforme : $K(u) = \frac{1}{2} I(|u| \leq 1)$,
- le noyau Epanechnikov : $K(u) = \frac{3}{4}(1 - u^2) I(|u| \leq 1)$,
- le noyau gaussien : $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2})$.

En utilisant le théorème de Taylor, on peut approximer linéairement $f(u)$ par

$$\begin{array}{rcl} f(u) & \approx & f(x) + (u - x) f'(x) \\ & \approx & \beta_0 + \beta_1 (u - x) \end{array}$$

pour u au voisinage de x.
On peut évidement étendre l'approximation de $f(u)$ par une fonction quadratique, cubique, etc.
On constate donc qu'estimer localement $f(x)$ revient à calculer $\hat{\beta}_0$ tandis que l'estimation de $f'(x)$ est donné par $\hat{\beta}_1$, et donc obtenir $\hat{f}(x)$ revient à chercher $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$ tel que

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^2} \sum_{i=1}^{n} w_i(x)(y_i - (\beta_0 + \beta_1(x_i - x)))^2.$$

...à la régression quantile

Dans le cas de la régression quantile, $f(x)$ est le quantile d'orde p de Y conditionnellement à $X = x$. On a donc

$$f(x) = Q_Y(p|x).$$

Suivons la même démarche qu'en régression, c'est-à-dire approximons localement $Q_Y(p|x)$ par une fonction linéaire du type $\beta_0 + \beta_1(u - x)$. Dans ce cas, une estimation $\hat{Q}_Y(p|x)$ est donnée par $\hat{\beta}_0$, où $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$ sont les solutions de

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^2} \sum_{i=1}^{n} w_i(x) \rho_p(y_i - \beta_0 - \beta_1(x_i - x)).$$

Comme nous l'avons mentionné dans le cadre de la régression, on peut étendre l'approximation de $f(u)$ à condition que la fonction f soit suffisamment dérivable. Dès lors, on a le

problème de minimisation suivant

$$\arg \min_{\beta \in \mathbb{R}^{q+1}} \sum_{i=1}^{n} w_i(x) \rho_p(y_i - \beta_0 - \beta_1(x_i - x) - ... - \beta_q(x_i - x)^q).$$

Un des principaux soucis de cette méthode non paramétrique est le choix de h. En effet, choisir un h trop "grand" impliquerait un sur-lissage (*oversmoothing*) tandis que choisir un h trop "petit" reviendrait à "ajuster du bruit" et impliquerait une grande variance pour l'estimateur.

Dans le livre de Koenker, aucune solution n'est mentionnée à ce sujet. Par contre, dans l'artice de Yu et Jones [30], un choix optimal de h est basé sur la minimisation de l'erreur quadratique moyenne. Les auteurs présentent également une autre technique faisant intervenir deux noyaux et donc deux *bandwith* à déterminer. Cette méthode ne sera pas explicitée ici mais les lecteurs intéressés par cette seconde technique peuvent se référer à [30].

Exemple 2.3.1. *Reprenons l'exemple 2.2.2 de la section précédente. Afin de déterminer les quantiles de régression, nous pouvons utiliser la méthode présentée dans cette section. Elle est d'ailleurs implémentée dans le package "quantreg" sur base d'un noyau gaussien et uniquement dans le cas d'une seule variable explicative.*

Le graphique (2.5) *représente la régression médiane pour trois valeurs de h différentes, ce qui permet de vérifier ce que nous disions à propos du choix de ce paramètre. Il apparait qu'une valeur de h égale à 1 semble être un bon compromis entre les deux effets mentionnés ci-dessus en regard des courbes de régression représentées sur ce graphique.*

Le graphique (2.6) *représente quant à lui la régression quantile pour des valeurs de p égale à 0.25, 0.5 et 0.75 et pour une valeur de h fixée à 1. On constate que, même si une valeur de h de 1 ne semblait pas être un mauvais choix pour la médiane, il n'en est pas de même pour les quantiles 0.25 et 0.75. Dès lors, il faut choisir un h différent pour chaque quantile, ce qui montre la complexité de cette méthode non paramètrique.*

Un moyen moins fastidieux est l'utilisation de B-splines, méthode d'ailleurs préférée par Koenker en toute généralité, et non uniquement dans cet exemple précis.

2.3.2 Méthode des B-splines

Avant d'expliciter en quoi consiste la régression quantile via l'utilisation des B-splines, commençons par introduire la notion de B-splines en régression "classique".

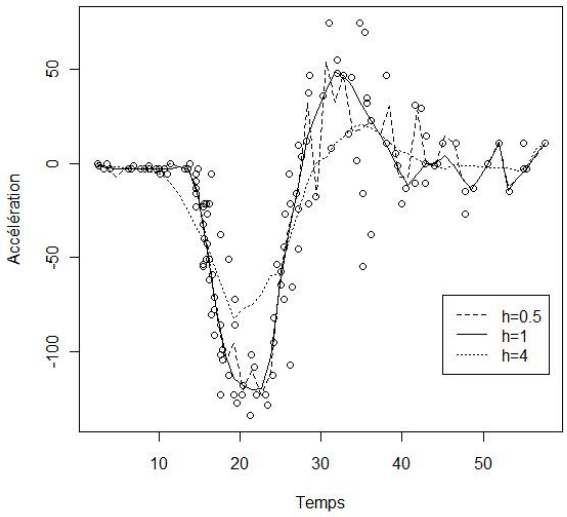

Figure 2.5 – Régression médiane par polynômes locaux pour trois valeurs différentes de h.

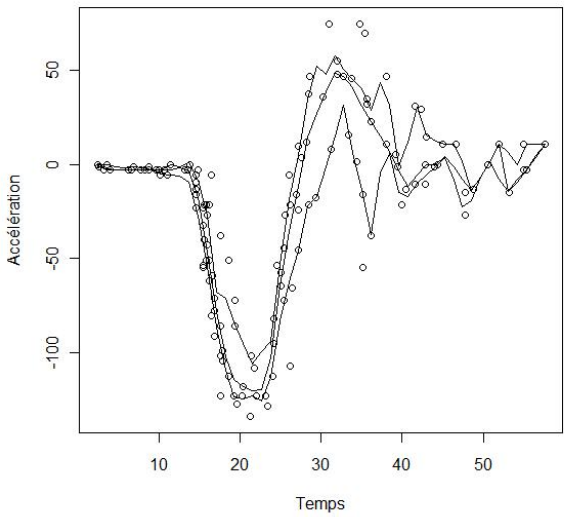

Figure 2.6 – Régression quantile par polynômes locaux pour des valeurs de p fixées à 0.25, 0.5 et 0.75 et $h = 1$.

Un B-spline consiste en plusieurs fonctions polynomiales connectées d'une certaine façon. Avant d'en donner une définition plus précise, regardons quelques exemples provenant de [3].

A l'extrême gauche de l'exemple 1 du graphique (2.7) est représenté un B-spline de degré 1. Il consiste en deux morceaux de fonction linéaire, l'une allant de x_1 à x_2, l'autre allant de x_2 à x_3. Ces trois points sont appellés noeuds et on peut consater qu'à gauche du premier noeud ainsi qu'à droite du dernier, le B-spline est nul.
Sur la partie droite du graphique, trois autres B-splines de degré 1 sont représentés, chacun basé sur trois noeuds.

Passons à l'exemple 2 de ce même graphique. Sur la gauche est représenté à présent un B-spline de degré 2. Ce dernier consiste désormais en trois fonctions quadratiques jointes aux deux noeuds x_2 et x_3, et est basé sur les quatres noeuds x_1, x_2, x_3 et x_4. Sur la partie droite de cet exemple sont à nouveau représentés trois B-splines mais à présent de degré 2.

On remarque, quelque soit l'exemple considéré, que les B-splines se chevauchent, et que, pour un point x fixé, il y a deux (resp.trois) B-splines de degré 1 (resp.2) non nuls.
Ceci nous amène à la définition générale des B-splines, mais avant nous devons spécifier un domaine sur lequel seront construits les B-splines.

Soit une suite de points $x_1, ..., x_n$. On se donne un domaine compact délimité par x_{min} et x_{max} et divisié en un certain nombre n' d'intervalles de longueurs égales et délimités par $n' + 1$ noeuds, et sur lesquels seront construits les B-splines (nous envisageons donc des noeuds équidistants, tel que dans [3] mais nous pourrions également considérer des noeuds non équidistants).

Définition 2.3.1. Un B-spline de degré q consiste en $q + 1$ fonctions polynomiales de degré q tel que

- les fonctions polynomiales sont jointes en q noeuds dits intérieurs,
- aux noeuds de connection, les dérivées jusqu'à l'ordre $q - 1$ de ces fonctions sont continues,
- le B-spline est positif sur le domaine recouvert par $q+2$ noeuds et vaut zéro ailleurs,
- excepté aux bords du domaine, le B-spline se chevauche avec $2q$ B-splines voisins,
- pour un point x fixé, il y a $q + 1$ B-splines non nuls.

On note $B_j(x; q)$ la valeur au point x du j$^{\text{ème}}$ B-spline pour une grille de noeuds équidistants donnée.

L'index d'un B-spline est relié à un noeud, plus précisément, il donne l'index du noeud qui caractérise la position du B-spline. Les auteurs de [3] prennent comme convention de choisir comme noeud caractérisant la position du B-spline, le noeud à partir duquel le B-spline devient non nul.

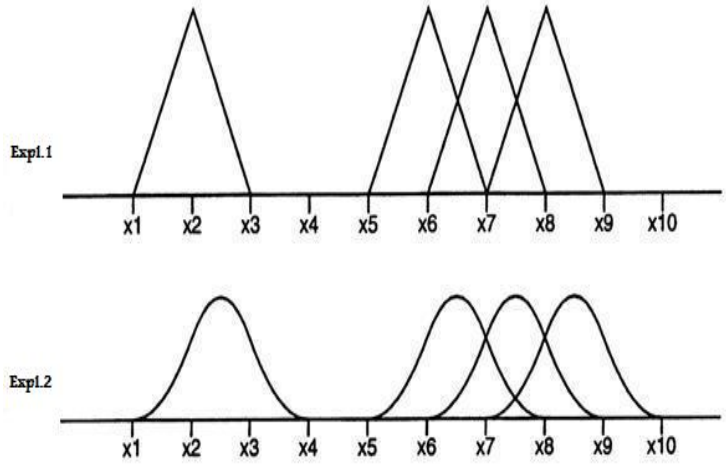

Figure 2.7 – Exemples de B-splines de degré 1 et 2.(Graphiques provenant de [3])

Passons à l'utilisation des B-splines en régression.

2.3.3 Utilisation des B-splines en régression

Alors qu'en régression linéaire, la variable réponse Y est vue comme une combinaison linéaire des covariables, ici, c'est une combinaison de B-splines de degré q pour un q fixé. La fonction objectif à minimiser est dès lors, en considérant une seule variable explicative,

$$\sum_{i=1}^{n} \left\{ \; y_i - \textstyle\sum_{j=1}^{m} \alpha_j B_j(x_i; q) \; \right\}^2 .$$

De là, il vient en régression quantile

$$\sum_{i=1}^{n} \rho_q(y_i - \sum_{j=1}^{m} \alpha_j B_j(x_i; q)).$$

C'est ce modèle qui est illustré dans l'exemple 2.2.2 avec un nombre de noeuds égal à 15. Cependant, le choix du nombre de noeuds est problématique. En effet, trop de noeuds mène à un *overfitting* de l'ensemble des données, c'est-à-dire à un biais de l'estimateur quasiment nul mais présentant une variance élévée, tandis que dans le cas contraire, trop

peu de noeuds mène à une faible variance mais un biais élévé. Il n'existe pas vraiment de solution quant au nombre idéal de noeuds à choisir afin d'obtenir un compromis entre ces deux effets. Une alternative envisagée dans [3] est d'utiliser un grand nombre de noeuds équidistants tout en posant une pénalité sur la dérivée seconde de la combinaison linéaire des B-splines ($\sum_{j=1}^{m} \alpha_j B_j(x_i;q)$) afin de prévenir l'*overfitting*. C'est cette voie qui est choisie dans le cadre de la régression.
Dès lors, on inclut dans le modèle une pénalité sur la dérivée seconde, plus précisément, on pénalise la courbure sur un intervalle donné et la fonction objectif devient, dans le cas de la régression,

$$\sum_{i=1}^{n} \left\{ \; y_i - \textstyle\sum_{j=1}^{m} \alpha_j B_j(x_i;q) \; \right\}^2 + \lambda \int_{x_{min}}^{x_{max}} \left\{ \; \textstyle\sum_{j=1}^{m} \alpha_j B_j''(x;q) \; \right\}^2 dx,$$

avec $\lambda > 0$.

En régression quantile, cela donne

$$\sum_{i=1}^{n} \rho_q(y_i - \sum_{j=1}^{m} \alpha_j B_j(x_i;q)) + \lambda \int_{x_{min}}^{x_{max}} \left\{ \; \textstyle\sum_{j=1}^{m} \alpha_j B_j''(x;q) \; \right\}^2 dx.$$

Intuitivement, vu que nous considérons un grand nombre de noeuds, la "courbe" de régression va interpoler un grand nombre de points, ce qui aboutira à un biais quasiment nul et une grande variance de l'estimateur. En introduisant une pénalité sur la dérivée seconde, c'est-à-dire sur la courbure, on réduit ce phénomène. Néanmoins, si l'on considère des valeurs de λ très grandes, la courbe de régression va se rapprocher d'une droite linéaire, impliquant un grand biais et une petite variance de l'estimateur. Tandis que pour des valeurs de λ petites, la pénalité a peu d'impact et on risque donc l'*overfitting*. Il faut donc choisir une valeur de λ permettant un compromis entre ces deux effets.

Remarque 2.3.1. *Il est possible, dans le package "quantreg" de Koenker, de traiter la régression quantile en imposant une pénalité mais cette méthode se veut plus générale que celle expliquée ci-dessus. En effet, Koenker envisage le problème de minimisation*

$$\min_{g \in G} \sum_{i=1}^{n} \rho_p(y_i - g(x_i)) + \lambda \int (g''(x))^2 dx$$

où G est un espace de Sobolev de fonctions C^2 dont le carré de la dérivée seconde est intégrable.

Le lecteur intéressé par une extension dans le cas bivarié (deux variables explicatives) ainsi que par la mise en pratique de ces méthodes se réfèrera à [12].

Avant de clôturer cette section, illustrons à l'aide d'un petit exemple ce que nous venons de dire à propos du choix des valeurs de λ.

Exemple 2.3.2. *On considère à nouveau l'échantillon de données "mcycle" que nous avons déjà traité par la méthode des polynômes locaux et des B-splines. Les graphiques représentent la courbe de régression pour le quantile d'ordre p où p est fixé à 0.25 et ce pour différentes valeurs de λ. On remarque que plus la valeur de λ est grande, c'est-à-dire plus la pénalité est élevée, moins la courbe de régression interpole de données, ce qui confirme ce que nous avions énoncé à propos du choix de la pénalité. On remarque en outre qu'une valeur de λ égale à 1 semble être, au vu des graphiques, un bon compromis entre les deux effets possibles dû au choix d'une valeur de λ trop petite ou trop grande.*

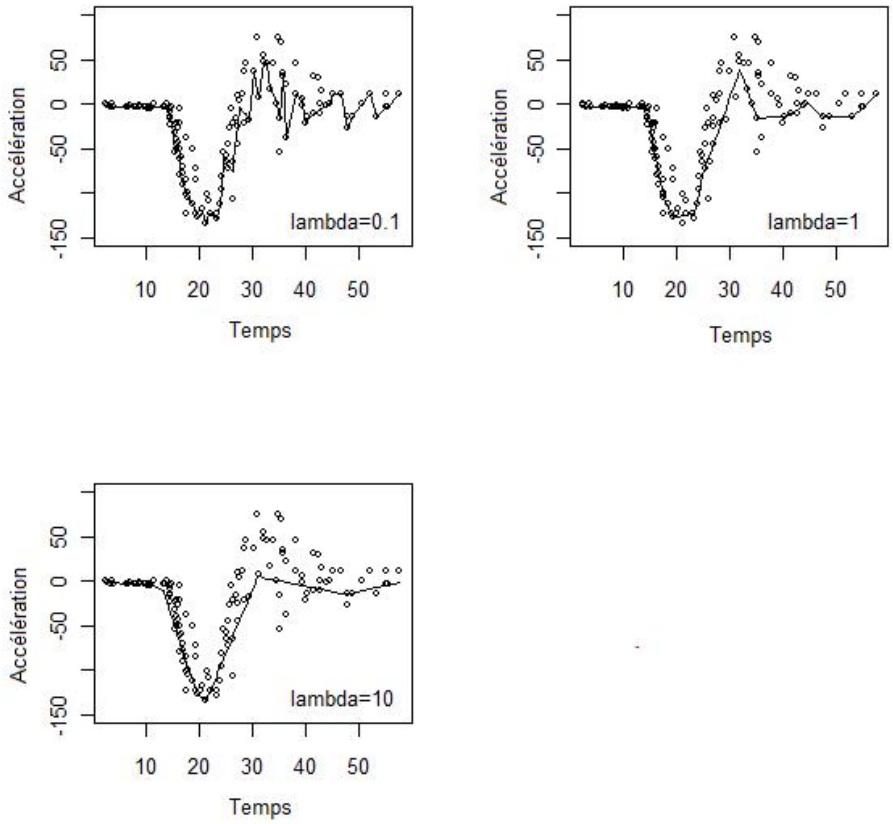

Figure 2.8 – Régression quantile d'ordre 0.25 pour différentes pénalités.

2.4 Utilisation des modèles additifs généralisés

Terminons ce chapitre consacré à un panorama des techniques fréquentistes en régression quantile par une méthode basée sur les modèles généralisés.

Dans les modèles généralisés, la variance d'une variable aléatoire Y de fonction de densité appartenant à la famille exponentielle dépend d'un paramètre de localisation μ, celui-ci étant égal à $E[Y]$, et d'un paramètre de dispersion ϕ. En général, pour une distribution appartenant à la famille exponentielle, le coefficient de dissymétrie et de kurtosis sont également dépendants de μ. Dès lors, dans les modèles généralisés, ces paramètres ne sont pas modélisés explicitement mais implicitement à travers leur dépendance en μ.

Dans [17], les auteurs présentent un modèle général de régression dans lequel la composante systématique et aléatoire sont plus flexibles. En effet, ils développent un modèle appelé *modèle généralisé additif pour la localisation, la dispersion et la forme* (*generalized additive model for location, scale and shape*) dans lequel l'hypothèse concernant la famille exponentielle est étendue en une famille de distributions très générale. Ainsi, la composante systématique du modèle est également étendue et tous les paramètres de la distribution conditionnelle de Y peuvent être modélisés comme une fonction paramétrique et/ou non paramétrique des covariables et/ou des effets aléatoires.

2.4.1 Modèles généralisés additifs pour la localisation, la dispersion et la forme

Soient $i = 1, ..., n$ observations indépendantes Y_i de fonction de densité $f_Y(y_i|\boldsymbol{\theta}^i)$ conditionnelle à $\boldsymbol{\theta}^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i})^T = (\mu_i, \sigma_i, \nu_i, \tau_i)^T$ un vecteur de quatre paramètres.
Un modèle généralisé additif pour la localisation, la dispersion et le forme suppose que chacun de ces quatre paramètres puisse être une fonction des covariables.
Les quatre paramètres $(\mu_i, \sigma_i, \nu_i, \tau_i)$ sont donc les paramètres de la distribution conditionnelle de Y. Les deux premiers paramètres μ_i et σ_i caractérisent en général les paramètres de localisation et d'échelle, tandis que les deux autres paramètres caractérisent la forme (asymétrie et kurtosis).

Soit $\mathbf{y}^T = (y_1, ..., y_n)$ un vecteur de n observations de la variable dépendante. Pour $k = 1, ..., 4$, considérons la fonction de lien $g_k(.)$ reliant les paramètres aux covariables et effets aléatoires par le modèle suivant

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.4}$$

c'est-à-dire, avec quatre paramètres,

$$\begin{aligned} g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \textstyle\sum_{j=1}^{J_1} \mathbf{Z}_{j1}\boldsymbol{\gamma}_{j1}, \\ g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \textstyle\sum_{j=1}^{J_2} \mathbf{Z}_{j2}\boldsymbol{\gamma}_{j2}, \\ g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \textstyle\sum_{j=1}^{J_3} \mathbf{Z}_{j3}\boldsymbol{\gamma}_{j3}, \\ g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \textstyle\sum_{j=1}^{J_4} \mathbf{Z}_{j4}\boldsymbol{\gamma}_{j4}. \end{aligned}$$

où $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$, $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_n)^T$, $\boldsymbol{\nu} = (\nu_1, ..., \nu_n)^T$, $\boldsymbol{\tau} = (\tau_1, ..., \tau_n)^T$ et $\boldsymbol{\eta}_k = (\eta_{1k}, ..., \eta_{nk})^T$ sont des vecteurs de longueur n, $\boldsymbol{\beta}_k^T = (\beta_{1k}, ..., \beta_{J'_k k})$ est un vecteur de J'_k paramètres, $\mathbf{X}_k$ est une matrice d'éléments connus et de dimension $n \times J'_k$, $\mathbf{Z}_{jk}$ est une matrice d'éléments fixés de dimension $n \times q_{jk}$ et $\boldsymbol{\gamma}_{jk}$ un vecteur de variables aléatoires de dimension q_{jk}.
Les auteurs de [17] supposent en outre que $\boldsymbol{\gamma}_{jk} \overset{ind}{\sim} N_{q_{jk}}(\mathbf{0}, \mathbf{G}_{jk}^-)$ où $\mathbf{G}_{jk}^-$ est la matrice inverse (généralisée) d'une matrice symétrique $\mathbf{G}_{jk} = \mathbf{G}_{jk}(\boldsymbol{\lambda}_{jk})$ de dimension $q_{jk} \times q_{jk}$ et qui dépend d'un vecteur d'hyperparamètres $\boldsymbol{\lambda}_{jk}$.
Le modèle (2.4) est appellé *modèle généralisé additif pour la localisation, la dispersion et la forme (GAMLSS).*

Le vecteur de paramètres $\boldsymbol{\beta}_k$ ainsi que les paramètres des effets aléatoires $\boldsymbol{\gamma}_{jk}$ ($j = 1, ..., J_k$ et $k = 1, 2, 3, 4$) sont estimés en maximisant une fonction de vraisemblance pénalisée $l_q(\boldsymbol{\beta}, \boldsymbol{\gamma})$ qui est donnée par

$$l_q(\boldsymbol{\beta}, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}, \boldsymbol{\gamma}) - \frac{1}{2} \sum_{k=1}^{q} \sum_{j=1}^{J_k} \lambda_{jk} \boldsymbol{\gamma}_{jk}^T G_{jk} \boldsymbol{\gamma}_{jk}$$

où q est le nombre de paramètres (quatre dans notre cas), et où

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \textstyle\sum_{i=1}^{n} \log f_Y(y_i|\boldsymbol{\theta}^i) \\ &= \textstyle\sum_{i=1}^{n} \log f_Y(y_i|\mu_i, \sigma_i, \nu_i, \tau_i). \end{aligned}$$

Nous n'expliciterons pas l'algorithme mettant en oeuvre cette procédure, celle-ci étant implémentée par les auteurs dans le package "gamlss" disponible à nouveau dans le logiciel *R*.
Le lecteur néanmoins intéressé par l'algorithme utilisé peut consulter [17].

Revenons à la distribution conditionnelle de Y. Comme nous l'avons déjà signalé, dans les modèles généralisés, cette distribution devait appartenir à la famille exponentielle. Ici, nous ne sommes pas aussi restrictif, autorisant la distribution conditionnelle à appartenir à une famille plus générale de distributions dont celles implémentées dans le package des auteurs sont reprises dans le tableau (2.9) tiré de [17] .

Number of parameters	*Distribution*
Discrete, one parameter	Binomial
	Geometric
	Logarithmic
	Poisson
	Positive Poisson
Discrete, two parameters	Beta–binomial
	Generalized Poisson
	Negative binomial type I
	Negative binomial type II
	Poisson–inverse Gaussian
Discrete, three parameters	Sichel
Continuous, one parameter	Exponential
	Double exponential
	Pareto
	Rayleigh
Continuous, two parameters	Gamma
	Gumbel
	Inverse Gaussian
	Logistic
	Log-logistic
	Normal
	Reverse Gumbel
	Weibull
	Weibull (proportional hazards)
Continuous, three parameters	Box–Cox normal (Cole and Green, 1992)
	Generalized extreme family
	Generalized gamma family (Box–Cox gamma)
	Power exponential family
	t-family
Continuous, four parameters	Box–Cox *t*
	Box–Cox power exponential
	Johnson–Su original
	Reparameterized Johnson–Su

Figure 2.9 – Distributions implémentées dans *gamlss*.

Etablissons à présent le lien avec la régression quantile par un exemple.

Exemple 2.4.1. *Considérons l'échantillon de données "Dutch Girls" disponible sur le site*[5] *des auteurs. L'étude consiste à établir le lien entre l'indice de masse corporelle (BMI) et l'âge. Une représentation est fournie par le graphique* (2.10)*. On remarque que supposer une relation linéaire entre l'âge et la paramère de localisation de la variable dépendante n'est pas approprié. Il en est peut être de même pour les autres paramètres. Se basant sur une étude antérieure dans laquelle le paramètre de kurtosis n'était pas modélisé de la manière la plus adéquate, les auteurs ont fait le choix de considérer comme distribution conditionnelle pour la variable dépendante* Y*, une distribution t de Box-Cox*[6]*, c'est-à-dire* $Y \sim BCT(\mu, \sigma, \nu, \tau)$ *où*

$$\begin{aligned} g_1(\mu) &= h_1(x) \\ g_2(\sigma) &= h_2(x) \\ g_3(\nu) &= h_3(x) \\ g_4(\tau) &= h_4(x), \end{aligned}$$

$g_k(.)$ *étant la fonction de lien et* $h_k(x)$ *une fonction non paramétrique (spline cubique) de la covariable* x *(l'âge dans notre exemple). Les courbes de régression pour les quantiles d'ordre* p *égal à 0.05, 0.25, 0.5, 0.75 et 0.95 (cette liste est non exhaustive) sont obtenues, premièrement, en recherchant* $(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau})$ *pour chaque valeur de la covariable, ce qui se fait après avoir estimé, par maximisation de la fonction de vraisemblance que nous venons d'introduire, les paramètres de régression, et ensuite en substituant ces estimations* $(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau})$ *dans*

$$y_p = \begin{cases} \mu(1 + \sigma\nu t_{\tau,p})^{1/\nu} & si \quad \nu \neq 0, \\ \mu \exp(\sigma t_{\tau,p}) & sinon, \end{cases}$$

où $t_{\tau,p}$ *est le quantile d'ordre* p *d'une distribution de Student avec* τ *degré de liberté.*
Une représentation est donnée par (2.11)*.*
On remarque donc que, contrairement aux méthodes présentées antérieurement, nous n'avons pas autant de modèles que de quantiles souhaités. En effet, après avoir estimé μ, σ, ν *et* τ*, il*

5. http ://www.gamlss.com/
6. Cette distribution est définie dans l'annexe A.

suffit de modifier le quantile de la distribution de Student pour obtenir le quantile conditionnel souhaité. Néanmoins l'étape consistant en la recherches des paramètres optimaux des fonctions non paramètriques $h_k(x)$ (nombre de noeuds...) peut s'avérer longue et difficile et ce d'autant plus que l'échantillon de données présente une forme particulière.

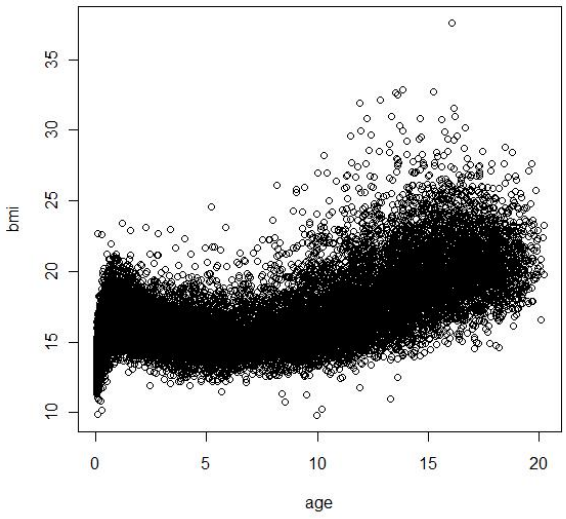

Figure 2.10 – BMI en fonction de l'âge.

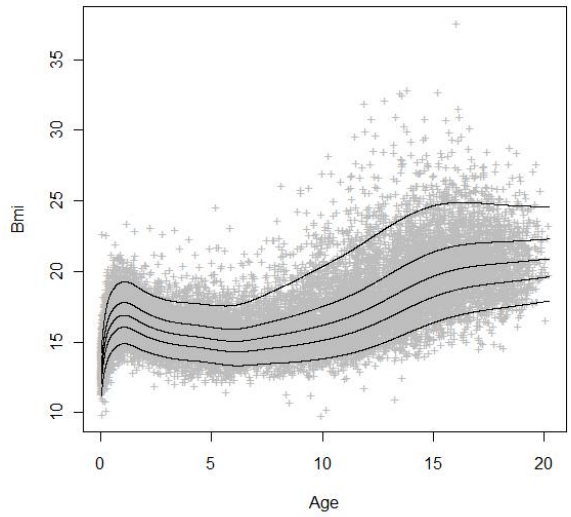

Figure 2.11 – Régression quantile pour p égal à 0.05, 0.25, 0.50, 0.75 et 0.95.

Chapitre 3

Approches bayesiennes

3.1 Introduction

Pour rappel, en considérant le modèle linéaire suivant

$$Y|\mathbf{x} = \mathbf{x}^T\boldsymbol{\beta} + \epsilon$$

pour une valeur des covariables $\mathbf{X} = \mathbf{x}$, le quantile conditionnel d'ordre p de Y est défini comme étant

$$Q_Y(p|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}(p)$$

pour une valeur de p fixée.

Alors qu'en régression "classique", les résidus sont supposés être distribués indépendamment selon une loi de densité inconnue mais de moyenne nulle et de variance constante, en régression quantile, ils sont également supposés être distribués selon une loi de densité inconnue mais dont le quantile d'ordre p est nul, c'est-à-dire distribués selon une loi de densité $f_p(.)$ telle que

$$\int_{-\infty}^{0} f_p(\epsilon)d\epsilon = p. \tag{3.1}$$

Partant de cette constatation, les méthodes bayesiennes présentées dans ce chapitres suggèrent de considérer une loi asymétrique de Laplace de paramètre de localisation μ nul, de paramètre d'échelle σ positif, et de paramètre d'asymétrie p comme loi de densité pour les résidus du modèle, et de là en dériver la distribution a posteriori des paramètres de la régression quantile. Ce choix peut cependant sembler très arbitraire et ne réside que dans le fait que cette distribution présente la caractéristique voulue. En effet, le paramètre de localisation μ de cette distribution étant le quantile d'ordre p, en contraignant ce paramètre

à 0, nous vérifions dès lors (3.1).

Néanmoins, comme nous le verrons, cette loi de densité manque de flexibilité, p déterminant à la fois l'asymétrie de cette distribution de Laplace et l'ordre du quantile conditionnel d'intérêt. Il semble en outre peu crédible que les observations suivent cette distribution particulière, c'est pourquoi nous présenterons également les idées générales des méthodes bayesiennes non paramétriques permettant donc une plus grande flexibilité de cette distribution.

Cette loi asymétrique de Laplace va en définitive jouer un rôle essentiel en régression quantile bayesienne, cette distribution apparaissant d'une quelconque manière dans toutes les approches des quantiles de régression qui seront présentées ici. C'est pourquoi la première section de ce chapitre lui sera longuement consacrée. On y présentera ses principales caractéristiques (fonction de répartition, moyenne, variance...) ce qui permettra entre autre de justifier son utilisation en régression quantile, c'est-à-dire de vérifier que le paramètre de localisation μ est bien le quantile d'ordre p de cette distribution de Laplace.

Les deux sections suivantes seront consacrées aux méthodes bayesiennes de Yu et al [28] [25] et de Tsionas [21]. Bien que ces deux méthodes reposent sur la même hypothèse concernant la loi de densité des résidus, loi asymétrique de Laplace, la stratégie d'exploration de l'a posteriori sera différente, la méthode de Yu et al reposant sur l'utilisation de l'algorithme de Metropolis (1953), celle de Tsionas sur l'algorithme de Gibbs (1984).

Enfin, la dernière section de ce chapitre sera consacrée à la présentation des idées fondamentales des méthodes bayesiennes non paramétriques utilisées dans le cadre de la régression quantile et suggérées par Kottas et al. Ces idées ont pour but, comme nous venons de le mentionner, d'obtenir une plus grande flexibilité de la distribution des erreurs mais vu leur côté très technique, ne seront pas implémentées ni appliquées dans ce mémoire. Nous présenterons principalement ces idées dans le cadre de la régression médiane [13], bien qu'elles furent développées, et continuent d'ailleurs de l'être, dans un cadre plus général de la régression quantile mais les articles de référence à ce sujet [14] [19] viennent ou ne sont pas encore à l'heure actuelle publiés.

3.2 Loi asymétrique de Laplace

3.2.1 Fonction de densité

Une variable aléatoire U est dite suivre une loi asymétrique de Laplace [29] si sa fonction de densité est donnée par

$$f(u;\mu,\sigma,p) = \frac{p(1-p)}{\sigma}\exp\{-\rho_p(\frac{u-\mu}{\sigma})\}$$

où $\rho_p(x)$ n'est rien d'autre que la fonction de perte déjà définie dans le chapitre 2, définition (2.1), par

$$\rho_p(x) = x(p - I(x < 0)),$$

avec

- p : paramètre d'asymétrie tel que $0 < p < 1$,
- σ : paramètre d'échelle tel que $\sigma > 0$,
- μ : paramètre de localisation tel que $-\infty < \mu < +\infty$.

Notation : $U \sim ALD(\mu,\sigma,p)$

Remarquons que la fonction définie ci-dessus est bien une fonction de densité.
En effet, on a

1. $f(u;\mu,\sigma,p) \geq 0$ car cette fonction est, à une constante positive près, une fonction exponentielle,
2. $f(u,\mu,\sigma,p)$ est continue et ce pour la même raison évoquée au point 1,
3. $\int_{-\infty}^{+\infty} f(u;\mu,\sigma,p)du = 1$, ce qui est vérifié ci-dessous.

   En effet,

   $$\int_{-\infty}^{+\infty} f(u;\mu,\sigma,p)du = \frac{p(1-p)}{\sigma}\left\{ \int_{-\infty}^{\mu} e^{(-\frac{(x-\mu)}{\sigma}(p-1))}du + \int_{\mu}^{+\infty} e^{(-\frac{(x-\mu)}{\sigma}(p))}du \right\}.$$

   Dès lors,

   $$\int_{-\infty}^{+\infty} f(u;\mu,\sigma,p)du = \frac{p(1-p)}{\sigma}\left\{ \left[\frac{-\sigma}{p-1}e^{(-\frac{(x-\mu)}{\sigma}(p-1))}\right]_{-\infty}^{\mu} + \left[\frac{-\sigma}{p}e^{(-\frac{(x-\mu)}{\sigma}(p))}\right]_{\mu}^{+\infty} \right\}$$

   ce qui donne

   $$\begin{aligned} \int_{-\infty}^{+\infty} f(u;\mu,\sigma,p)du &= \frac{p(1-p)}{\sigma}\{\frac{-\sigma}{p-1} + \frac{\sigma}{p}\} \\ &= 1 \end{aligned}$$

Remarquons que pour une valeur de p de $1/2$, on retrouve la fonction de densité de la loi standard de Laplace

$$f(u) = \frac{1}{4\sigma}\exp(\frac{-|u-\mu|}{2\sigma})$$

qui est symétrique par rapport au paramètre de localisation μ.

Par contre, pour d'autres valeurs de p, cette loi est asymétrique, avec une asymétrie vers la gauche, c'est-à-dire étalement des observations vers la droite de la distribution pour des valeurs de p inférieures à $\frac{1}{2}$ et une asymétrie vers la droite, et donc étalement des observations vers la gauche, lorsque p est supérieur à $\frac{1}{2}$, ce qui est représenté respectivement sur les graphiques (3.1) et (3.2) en considérant un paramètre de localisation nul et un paramètre d'échelle égal à 1.

Le graphique (3.3) représente quant à lui la fonction de densité pour des valeurs du paramètre d'échelle variant de 0.1 à 1, et ce pour un paramètre d'asymétrie et de localisation respectivement de 0.75 et 0.

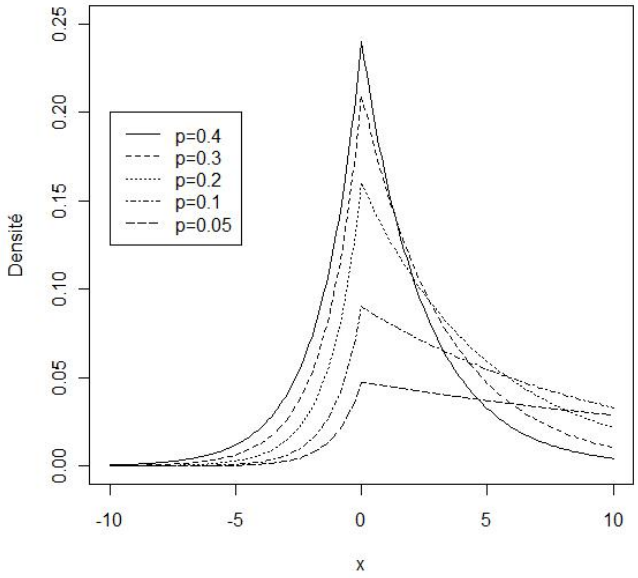

Figure 3.1 – Représentation de la fonction de densité pour des valeurs de p inférieures à 0.5.

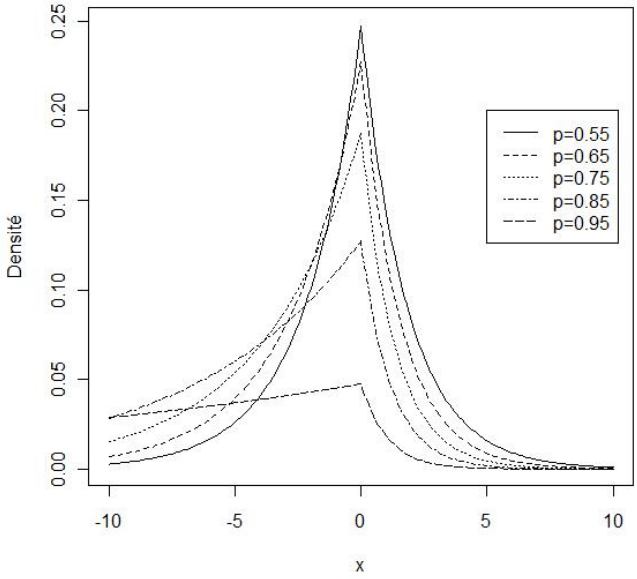

Figure 3.2 – Représentation de la fonction de densité pour des valeurs de p supérieures à 0.5.

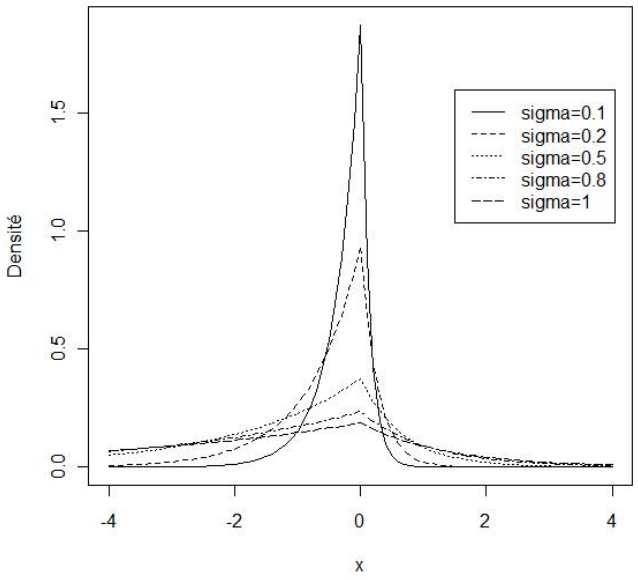

Figure 3.3 – Représentation de la fonction de densité pour différentes valeurs du paramètre d'échelle.

3.2.2 Fonction de répartition

Soit une variable aléatoire $X \sim ALD(\mu, \sigma, p)$.

Sa fonction de répartition est dès lors donnée par

$$F(x;\mu,\sigma,p) = \begin{cases} p \quad \exp(\frac{1-p}{\sigma}(x-\mu)) \quad \text{si} \quad x \leq \mu, \\ 1-(1-p)\exp(\frac{-p}{\sigma}(x-\mu)) \quad \text{si} \quad x > \mu. \end{cases} \tag{3.2}$$

En effet, par définition

$$F(x;\mu,\sigma,p) = \int_{-\infty}^{x} f(t;\mu,\sigma,p)dt$$

D'où, si $x \leq \mu$, il vient

$$\begin{aligned} F(x;\mu,\sigma,p) &= \int_{-\infty}^{x} \frac{p(1-p)}{\sigma}\exp(\frac{-(t-\mu)(p-1)}{\sigma}) \quad dt \\ &= \left[\frac{\frac{p(1-p)}{\sigma}\exp(-(t-\mu)(p-1))}{\frac{-(p-1)}{\sigma}}\right]_{-\infty}^{x} \\ &= \left[p \quad \exp(\frac{-(t-\mu)(p-1)}{\sigma})\right]_{-\infty}^{x} \\ &= p \quad \exp(\frac{-(x-\mu)(p-1)}{\sigma}) - 0 \\ &= p \quad \exp(\frac{1-p}{\sigma}(x-\mu)) \end{aligned}$$

Le second résultat de (3.2) s'obtient de manière similaire.

Une représentation graphique de cette fonction est disponible page suivante (graphique (3.4)).

De ces résultats, on en déduit que

$$\begin{aligned} Pr(X \leq \mu) &= F(\mu;\mu,\sigma,p) \\ &= p \end{aligned}$$

et

$$\begin{aligned} Pr(X > \mu) &= 1 - F(\mu;\mu,\sigma,p) \\ &= 1-p \end{aligned}$$

ce qui montre que le paramètre de localisation μ est, comme nous l'annoncions en introduction de ce chapitre, le quantile d'ordre p de la distribution, et ce quelque soit la valeur de σ.
Cette caractéristique importante de la distribution asymétrique de Laplace telle que définie ici est à la base de son utilisation en régression quantile.

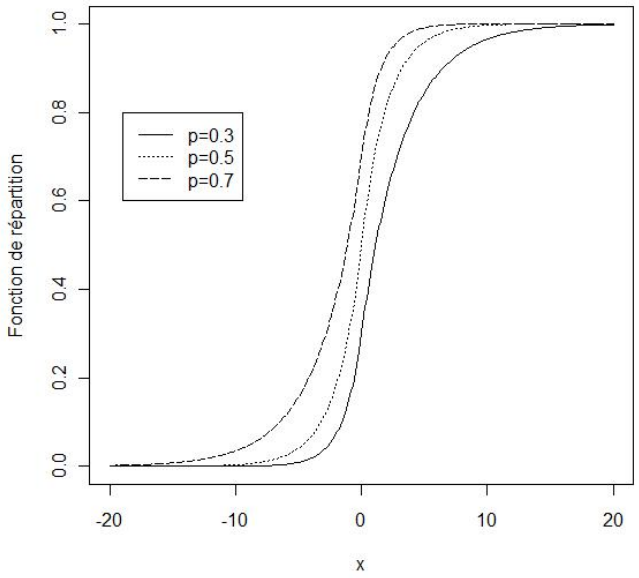

Figure 3.4 – Représentation de la fonction de répartition pour différentes valeurs de p.

Il peut s'avérer utile de pouvoir générer des données distribuées selon cette loi. Pour les distributions standards que sont la loi normale, la loi de Poisson, la loi de Student etc, ceci se fait aisément via des commandes préalablement implémentées dans le logiciel R. Concernant la loi asymétrique de Laplace telle que nous l'avons définie dans ce chapitre, cette dernière n'est pas implémentée dans le logiciel R et dès lors, pour générer des données selon cette distribution, deux possibilités s'offrent à nous.

Premièrement, nous pouvons utiliser la propriété suivante.

Propriété 3.2.1. *Si $Y \sim U(0,1)$, et si F est une fonction de répartition connue, alors $X = F^{-1}(Y) \sim F$ (avec F strictement croissante).*

L'inverse de la fonction de répartition étant donnée par

$$F^{-1}(x;\mu,\sigma,p) \quad = \quad \begin{cases} \mu + \frac{\sigma}{1-p}\log(\frac{x}{p}) & \text{si} \quad 0 \leq x \leq p, \\ \mu - \frac{\sigma}{p}\log(\frac{1-x}{1-p}) & \text{si} \quad p < x \leq 1, \end{cases} \tag{3.3}$$

il suffit donc de générer une variable aléatoire uniforme et de calculer l'inverse de la fonction de la répartition en cette variable pour obtenir une variable aléatoire distribuée selon une loi asymétrique de Laplace.

Une autre possibilité est de se baser sur les propriétés suivantes.

Propriété 3.2.2. *Considérons deux variables aléatoires ξ et η indépendantes et distribuées selon une loi exponentielle de moyenne égale 1, c'est-à-dire $\xi \sim Exp(1)$ et $\eta \sim Exp(1)$. Dans ce cas, il vient que $\frac{\xi}{p} - \frac{\eta}{1-p} \sim ALD(0, 1, p)$.*

Démonstration. Soit $\xi \sim Exp(1)$ et $\eta \sim Exp(1)$.
Les fonctions de densité de ξ et η sont respectivement

$$f_\xi(\xi) = e^{-\xi}\chi_{[0,+\infty[}(\xi)$$

$$f_\eta(\eta) = e^{-\eta}\chi_{[0,+\infty[}(\eta)$$

En posant

$$X = \frac{\xi}{p}$$

et

$$Y = \frac{\eta}{1-p},$$

on obtient

$$\begin{aligned} f_X(x) &= p\, f_\xi(px) \\ &= pe^{-px}\chi_{[0,+\infty[}(x) \\ f_Y(y) &= (1-p)\, f_\eta((1-p)y) \\ &= (1-p)e^{-(1-p)y}\chi_{[0,+\infty[}(y) \end{aligned}$$

Vu l'indépendance entre les deux variables aléatoires, il vient

$$\begin{aligned} f_{X,Y}(x,y) &= f_X(x)\, f_Y(y) \\ &= p(1-p)e^{-px}e^{-(1-p)y} \quad \text{avec} \quad x \geq 0 \quad \text{et} \quad y \geq 0. \end{aligned}$$

En effectuant le changement de variables suivant

$$\begin{cases} Z &= X - Y \\ W &= X \end{cases} \Leftrightarrow \begin{cases} X &= W \\ Y &= W - Z \end{cases}$$

la distribution jointe de Z et W est dès lors

$$f_{Z,W}(z,w) = p(1-p)e^{-pw}e^{-(1-p)}(w-z) \quad \text{avec} \quad w \geq z \geq 0.$$

En intégrant sur le domaine de définition de W, il vient

$$\begin{aligned} f_Z(z) &= \begin{cases} \int_z^{+\infty} p(1-p)e^{-w+(1-p)z} & \text{si} \quad z \geq 0 \\ \int_0^{+\infty} p(1-p)e^{-w+(1-p)z} & \text{sinon} \end{cases} \\ &= p(1-p)e^{-z(p-I(z\leq 0))} \end{aligned}$$

En d'autres termes, $Z \sim ALD(0,1,p)$. □

Propriété 3.2.3. *Si $X \sim ALD(0,1,p)$, alors $Y = \mu + \sigma X \sim ALD(\mu,\sigma,p)$.*

De plus, si $X \sim ALD(\mu,\sigma,p)$, alors $Y = \alpha + \beta X \sim ALD(\alpha+\beta\mu, \beta\sigma, p)$.

Démonstration. Ces résultats s'obtiennent par un simple changement de variables. □

Il suffit donc de simuler deux variables aléatoires indépendantes selon une loi exponentielle standard pour obtenir, par combinaison linéaire, la loi asymétrique de Laplace voulue.

3.2.3 Moyenne, variance et coefficient de dissymétrie

Afin d'obtenir les valeurs de la moyenne, variance et par extension d'autres moments non centrés d'un certain ordre, une possibilité est de dériver successivement la fonction génératrice des moments $E\left[e^{tx}\right]$ et de calculer cette dérivée en $t = 0$.

Ici, la fonction génératrice des moments étant égale à

$$E\left[e^{tx}\right] = p(1-p)\frac{e^{\mu t}}{(p-\sigma t)(\sigma t + 1 - p)},$$

on en déduit donc, en dérivant successivement cette fonction et en l'évaluant ensuite en $t = 0$, les valeurs de la moyenne et la variance qui sont

$$\begin{aligned} E\left[X\right] &= \mu + \tfrac{\sigma(1-2p)}{p(1-p)}, \\ V\left[X\right] &= \tfrac{\sigma^2(1-2p+2p^2)}{(1-p)^2p^2}. \end{aligned}$$

Remarquons que ces résultats peuvent également se déduire du résultat suivant fournit par les auteurs de [29]

$$E\left[(X-\mu)^k\right] = k!\sigma^k p(1-p)\left(\frac{1}{p^{k+1}} + \frac{(-1)^k}{(1-p)^{k+1}}\right), \tag{3.4}$$

qui peut se démontrer par récurrence sur k.

Enfin, ce dernier résultat permet également de calculer plus facilement le coefficient de dissymétrie de Fisher, $\gamma_1 = \frac{\mu_3}{\sigma^3}$. En effet, le moment centré d'ordre 3, μ_3, n'étant rien d'autre que

$$\begin{aligned} E\left[(X - E\left[X\right])^3\right] &= E\left[(X - \mu + \mu - E\left[X\right])^3\right] \\ &= E\left[(X-\mu)^3\right] + 3E\left[(X-\mu)^2\right](\mu - E\left[X\right]) \\ &\quad +3E\left[X-\mu\right](\mu - E\left[X\right])^2 + (\mu - E\left[X\right])^3, \end{aligned}$$

nous obtenons, après calculs,

$$\gamma_1 = \frac{-2(p^3 - (1-p)^3)}{((1-p)^2 + p^2)^{3/2}},$$

ce qui implique que pour des valeurs de p supérieures à 0.5, nous avons un coefficient négatif tandis que ce coefficient est positif pour des valeurs de p inférieures à 0.5. Notons qu'il est bien égal à zéro lorsque p vaut 0.5. Le graphique (3.5) témoigne de cette tendance et ceci coïncide avec nos observations du début de section.

Terminons cette section consacrée à la distribution asymétrique de Laplace en recherchant l'estimateur par maximum de vraisemblance de μ.

Considérons l'échantillon aléatoire $\underset{\sim}{X}$ de taille n tel que $X_i \overset{iid}{\sim} ALD(\mu, \sigma, p)$. L'estimateur par maximum de vraisemblance de μ est

$$\begin{aligned} \hat{\mu} &= \arg\max_{\mu\in\mathbb{R}} \tfrac{p^n(1-p)^n}{\sigma^n} e^{-\sum_{i=1}^n \rho_p(x_i-\mu)} \\ &= \arg\min_{\mu\in\mathbb{R}} \textstyle\sum_{i=1}^n \rho_p(x_i - \mu). \end{aligned}$$

On retrouve la définition du quantile d'ordre p telle que défini dans le chapitre 2.
Ce résultat n'a rien d'étonnant étant donné que μ est, comme nous l'avions remarqué, le quantile d'ordre p de la distribution.

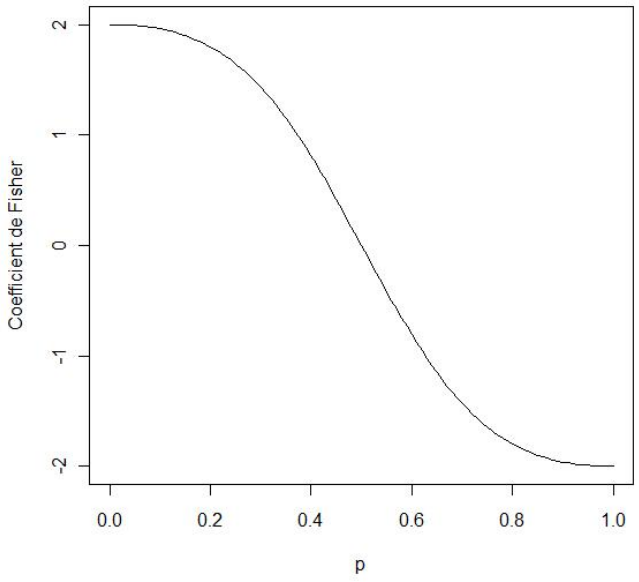

Figure 3.5 – Evolution du coefficient de Fisher en fonction de p

Comme nous l'avons signalé en introduisant ce chapitre, les méthodes bayesiennes qui seront présentées dans les deux sections qui vont suivre reposent sur l'hypothèse d'une distribution asymétrique de Laplace pour les résidus, de paramètre de localisation μ nul, de paramètre d'échelle σ positif et de paramètre d'asymétrie p. Ainsi, on a bien que le quantile d'ordre p des résidus est nul.

3.3 Approche de Yu et al (2001)

3.3.1 Présentation de cette approche

La méthode présentée dans cette section repose donc sur l'idée de supposer comme distribution pour les résidus du modèle une loi asymétrique de Laplace de paramètre de localisation μ égal à 0, de paramètre d'échelle σ égal à 1, et de paramètre d'asymétrie p fixé selon le quantile conditionnel souhaité. On remarque que Yu et al [28] suppose donc un paramètre d'échelle σ fixé à 1, ce qui peut paraître plutôt restrictif comme hypothèse. Néanmoins, ici, la seule condition souhaitée est que la distribution des erreurs ait un quantile d'ordre p nul, ce qui est le cas en considérant un paramètre de localisation μ nul et ce quelque soit la valeur de σ.

Considérer une telle loi asymétrique de Laplace pour les erreurs revient en fait à considérer comme distribution conditionnelle pour la variable dépendante [1] Y la distribution asymétrique de Laplace suivante

$$f(y;\boldsymbol{\beta}) = p(1-p)\exp\{-\rho_p(y-\mathbf{x}^T\boldsymbol{\beta})\}.$$

On remarque dès lors qu'estimer le quantile conditionnel d'ordre p de Y, $Q_Y(p|\mathbf{x})$, revient à estimer le paramètre de localisation μ égal à $\mathbf{x}^T\boldsymbol{\beta}$ d'une telle distribution.

Bien sûr, cette distribution est un choix possible, on pourrait ne pas contraindre σ à 1, ce qui sera le cas dans la seconde méthode bayesienne présentée dans ce mémoire, voire envisager une autre distribution pour les erreurs pour autant que le quantile d'ordre p de cette distribution soit nul.

En statistique bayesienne, on s'intéresse à la densité a posteriori des paramètres, ici les quantiles de régression pour un p fixé. Vu les rappels mentionnés dans le premier chapitre de ce mémoire, cette densité est donnée par

$$p(\boldsymbol{\beta}|y) \propto L(y|\boldsymbol{\beta})p(\boldsymbol{\beta})$$

où $p(\boldsymbol{\beta})$ est la densité a priori de $\boldsymbol{\beta}$ et $L(y|\boldsymbol{\beta})$ la fonction de vraisemblance qui vaut dans ce cas

$$\begin{aligned} L(y|\boldsymbol{\beta}) &= \prod_{i=1}^{n} f(y_i;\mu_i) \\ &= \prod_{i=1}^{n} p(1-p)\exp\{-\rho_p(y_i-\mathbf{x}_i^T\boldsymbol{\beta})\} \\ &= p^n(1-p)^n\exp\{-\sum_{i=1}^{n}\rho_p(y_i-\mathbf{x}_i^T\boldsymbol{\beta})\} \end{aligned} \tag{3.5}$$

1. A nouveau, nous reprenons les mêmes notations que celles des chapitres précédents.

Bien qu'en théorie, rien ne nous empêche d'utiliser n'importe quelle distribution a priori, il semble logique, en l'absence d'informations sur les paramètres à estimer, de considérer une densité a priori impropre uniforme pour chaque composante de $\boldsymbol{\beta}$. Dans ce cas, vu le principe du maximum d'entropie, la densité jointe a priori est donnée par $p(\boldsymbol{\beta}) \propto 1$ et la densité a posteriori sera proportionnelle à la vraisemblance.
Néanmoins, en utilisant une densité a priori impropre, il faut s'assurer que la densité a posteriori soit une fonction de densité, c'est-à-dire vérifie les conditions rappellées dans la section 2 de ce chapitre, ce qui est prouvé dans le théorème suivant.

Théorème 3.3.1. *Si la fonction de vraisemblance est donnée par* (3.5) *et* $p(\boldsymbol{\beta}) \propto 1$*, la densité a posteriori de* $\boldsymbol{\beta}$*,* $p(\boldsymbol{\beta}|y)$*, sera une densité non impropre. En d'autres termes,*

$$0 < \int \pi(\boldsymbol{\beta}|y) d\boldsymbol{\beta} < \infty$$

ou de manière équivalente

$$0 < \int L(y|\boldsymbol{\beta}) p(\boldsymbol{\beta}) d\boldsymbol{\beta} < \infty$$

Le lecteur intéressé par la démonstration du théorème peut se référer à [28].

Une fois les distributions a posteriori des paramètres obtenues, il devient dès lors possible de rechercher les moyennes a posteriori de même que les écart-types ainsi que les intervalles de crédibilité. On peut en outre également obtenir les distributions a posteriori de combinaisons linéaires des paramètres.

Il apparait cependant que les densités a posteriori des composantes de $\boldsymbol{\beta}$ ne sont pas des lois de densités connues. Dès lors, pour en évaluer les propriétés, nous allons recourir aux méthodes dites MCMC (*Markov Chain Monte Carlo*). La section suivante présente la méthode MCMC qui sera utilisée ici, son implémentation étant disponible en annexe (appendice B).

3.3.2 Implémentation

Rappel théorique sur les méthodes MCMC

Commençons par rappeler en quoi consiste les méthodes MCMC en général.

Plaçons-nous dans le cas univarié où seul un paramètre est à estimer.

Soit un paramètre inconnu à estimer θ et une valeur initiale θ^0 choisie du paramètre. Les méthodes MCMC proposent différentes façons de construire des chaînes θ^t pour

$t = 1, 2, ...$ où θ^t est généré à partir d'une distribution $T_t(\theta^t|\theta^{t-1})$ dite de transition et ne dépendant que de θ^{t-1}.
Après une période de "burn-in" correspondant aux premières iérationss, θ^{s+m} (m=1,2,...) peut être vu comme un échantillon aléatoire identiquement mais pas indépendamment distribué selon la distribution a posteriori $p(\theta|data)$.

Dans ce mémoire, nous utiliserons l'algorithme de Métropolis (1953).
Pour rappel, cet algorithme fonctionne comme suit

1. Choisir une valeur initiale θ^0 du paramètre
2. Pour $t = 1, 2, ...$
   - Générer θ_{prop} à partir d'une distribution symétrique $q_t(\theta_{prop}|\theta^{t-1})$ appelée loi de proposition ou loi proposante.
   - Calculer $prob = \min\left\{ 1, \ \frac{p(\theta_{prob}|data)}{p(\theta^{t-1}|data)} \right\}$.
   - Poser $\theta^t = \theta_{prob}$ avec une probabilité $prob$ sinon $\theta^t = \theta^{t-1}$(en d'autres termes, on accepte le candidat θ_{prob} avec une certaine probabilité $prob$, sinon on le rejette au profit de la valeur du paramètre généré à l'étape précédente).

Il existe une généralisation de cet algorithme, l'algorithme de Metropolis-Hastings (1970), qui permet l'utilisation de distributions asymétriques comme loi de proposition.
La seule différence dans l'algorithme se trouve dans le calcul de la probabilité d'acceptation qui devient

$$prob = \left\{ 1, \ \frac{p(\theta_{prob}|data)q_t(\theta^{t-1}|data)}{p(\theta^{t-1}|data)q_t(\theta_{prob}|data)}. \right\}$$

On remarque de suite que lorsque la loi de proposition est symétrique, nous retrouvons le calcul précédent.

Remarque 3.3.1. *Il est recommandé dans [5] de viser un taux d'acceptation de 0.20 lorsque tous les paramètres sont mis à jour simultanément et de 0.40 sinon. L'écart-type de la loi proposante nous permettra d'atteindre de tels taux d'acceptation. En effet, on peut noter que si l'écart-type considéré est trop petit, beaucoup de candidats seront acceptés et l'espace du paramètre Θ sera parcouru lentement. La chaîne générée est dite alors pauvrement mixée (poorly mixing), restant dans de petites régions de l'espace du paramètre pendant un certain temps, ce qui se traduira par une large auto-corrélation dans la chaîne générée. Par contre, si l'écart-type de la loi de proposition est trop élevé, beaucoup de candidats seront rejettés et à nouveau, l'espace du paramètre sera parcouru lentement, ce qui se traduira dans ce cas par des périodes constantes dans la chaîne générée.*

Remarque 3.3.2. *En général, il y a plus d'un paramètre à estimer. Dès lors, la loi de proposition sera une loi multivariée (typiquement, une loi multinormale ou de Student multivariée).*

Dans ce mémoire, deux implémentations de l'algorithme Metropolis-Hasting sont envisagées. L'une consiste à mettre à jour tous les paramètres à estimer simultanément, et ce via une loi de proposition multinormale tandis que la seconde consiste en une mise à jour paramètre par paramètre via une loi de proposition gaussienne. Ces implémentations ont été élaborées pour une utilisation dans le logiciel R et sont largement détaillées en annexe (appendice B).
De plus, elles ont été implémentées de façon à permettre une utilisation assez large, ceci dans le but de permettre aux lecteurs intéressés de modifier à leur guise certains paramètres telle que la distribution a posteriori (en choisissant une autre distribution conditionnelle ou d'autres a priori), certains choix initiaux etc. Une mise en application de ces algorithmes sera présentée au chapitre suivant.
Remarquons que, concernant la mise à jour simultanée des paramètres, nous avons choisi comme loi de proposition une loi multinormale, mais nous aurions pu tout aussi bien opter pour une loi de Student multivariée. De même, concernant la mise à jour paramètre par paramètre, nous aurions également pu opter pour une loi de Student plutôt qu'une gaussienne voire même envisager des lois de propositions non symétriques et recourir dès lors à l'utilisation de l'algorithme de Metropolis-Hastings (1970). Notons que les auteurs ont, dans les exemples mentionnés dans [28] et [25], opté pour une mise à jour paramètre par paramètre via une loi de proposition gaussienne.

Terminons cette section consacrée à l'approche de Yu par mentionner d'autres articles du même auteur sur la régression quantile. Il faut savoir que Yu a, avec d'autres auteurs, développé largement la régression quantile en statistique fréquentiste non paramétrique en se basant sur la méthode des polynômes locaux (voir [11] [22] [23] [26] [30]) mais il a aussi developpé la méthode présentée ici dans le cas de la régression Tobit [2](voir [24]). Enfin l'article [27] est un résumé des diverses méthodes fréquentistes paramétriques et non paramétriques ainsi que de l'approche bayesienne reposant sur la distribution asymétrique de Laplace de la régression quantile.

2. Le modèle Tobit, développé par James Tobin (1918-2002) est utilisé principalement en économétrie dans le cas où, par exemple, nous souhaitons trouver le montant financier qu'un individu ou une famille dépense dans l'achat d'un bien immobilier en fonction de variables socio-économiques (revenu,...). Il s'avère dès lors que nous n'avons aucun renseignement sur la dépense en biens immobiliers pour les consommateurs qui n'achètent pas, nous n'en possédons que pour ceux qui achètent réellement un bien immobilier. On a donc un échantillon pour lequel l'information sur la variable dépendante n'est disponible que pour certaines observations, c'est-à-dire un échantillon censuré. C'est pourquoi un modèle de régression Tobit est aussi appellé un modèle de régression censuré.

3.4 Approche de Tsionas (2003)

Pour rappel, et selon nos notations, la régression quantile telle qu'introduite par Koenker est basée sur l'idée, en considérant le modèle

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i,$$

que le quantile de régression d'ordre p est solution du problème de minimisation

$$\min_{\boldsymbol{\beta}} \left[\sum_{\{i|y_i \geq \mathbf{x}_i^T \boldsymbol{\beta}\}} p|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| + \sum_{\{i|y_i < \mathbf{x}_i^T \boldsymbol{\beta}\}} (1-p)|y_i - \mathbf{x}_i^T \boldsymbol{\beta}| \right].$$

Yu et al (2001) suggéraient de considérer une loi asymétrique de Laplace de paramètre de localisation μ égal à 0, de paramètre d'échelle σ égal à 1 et de paramètre d'asymétrie p fixé selon le quantile conditionnel recherché.

E. G. Tsionas (2003) a entamé pratiquement la même démarche en considérant donc comme distribution pour les erreurs du modèle

$$f(\epsilon) \propto \tau^{-1} \exp\left[-\tau^{-1}|\epsilon|\{pI_{[0,+\infty)}(\epsilon) + (1-p)I_{(-\infty,0)}(\epsilon)\}\right]$$

la distribution asymétrique de Laplace telle que définie dans la section 1 de ce chapitre, si ce n'est que nous avons noté ici le paramètre d'échelle τ au lieu de σ et que l'auteur n'a pas considéré le terme $p(1-p)$, ce terme étant constant vu que p est fixé.
On constate donc que contrairement à Yu et al (2001), Tsionas (2003) ne contraint pas le paramètre d'échelle à 1.

La vraisemblance devient dès lors

$$L(y|\boldsymbol{\beta}, \tau) \propto \tau^{-n} \exp\left\{ -\tau^{-1} \sum_{i=1}^{n} |y_i - \mathbf{x}_i^T \boldsymbol{\beta}| \{pI_{[0,+\infty)}(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + (1-p)I_{(-\infty,0)}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})\} \right\}.$$

L'auteur a également opté pour des densités a priori impropres pour $\boldsymbol{\beta}$ et τ, c'est-à-dire

$$p(\boldsymbol{\beta}) \propto 1$$

$$p(\tau) \propto \tfrac{1}{\tau}$$

et dès lors, la densité a priori jointe pour ces paramètres devient, par le principe du maximum d'entropie,

$$p(\boldsymbol{\beta}, \tau) \propto \frac{1}{\tau}.$$

Ceci nous amène à la densité jointe a posteriori suivante

$$p(\boldsymbol{\beta},\tau|y) \propto \tau^{-(n+1)} \exp\left\{ -\tau^{-1} \sum_{i=1}^{n} |y_i - \mathbf{x}_i^T\boldsymbol{\beta}| \{ pI_{[0,+\infty)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + (1-p)I_{(-\infty,0)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) \} \right\}.$$

On remarque donc que, jusqu'à présent, la démarche suivie par Tsionas (2003) est pratiquement la même que celle de Yu et al (2001), si ce n'est qu'ici le paramètre d'échelle n'est pas contrait à 1 comme nous l'avons déjà signalé. Dès lors, la distribution a posteriori de $\boldsymbol{\beta}$ est obtenue en intégrant sur le domaine de τ et est de la forme

$$p(\boldsymbol{\beta}|y) \propto \left\{ \sum_{i=1}^{n} |y_i - \mathbf{x}_i^T\boldsymbol{\beta}| \{ pI_{[0,+\infty)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + (1-p)I_{(-\infty,0)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) \} \right\}^{-n}.$$

A nouveau, cette distribution n'est pas connue et nous devons recourir à des méthodes numériques telles que le MCMC pour en évaluer les propriétés.
Néanmoins, l'auteur n'explorera pas l'a posteriori via une méthode MCMC mais va utiliser un artifice qui consiste à voir la loi de Laplace comme un mélange de lois normales, et ce dans le but de faciliter l'exploration de l'a posteriori.

En particulier, il considère la densité suivante, en posant $\sigma = \sqrt{2\tau}$

$$f(\epsilon_i|w_i) \propto (\sigma^2 w_i)^{-\frac{1}{2}} \exp\left\{ -\frac{\epsilon_i^2}{2\sigma^2 w_i} [pI_{[0,+\infty)}(\epsilon_i) + (1-p)I_{(-\infty,0)}(\epsilon_i)] \right\} \tag{3.6}$$

où les variables w_i sont non observées et distribuées selon une loi exponentielle de moyenne égale à 1

$$p(w_i) = \exp(-w_i).$$

Ainsi, en intégrant sur ces variables latentes w_i, on retrouve la loi asymétrique de Laplace pour les erreurs. Autrement dit,

$$f(\epsilon_i) = \int_0^{+\infty} f(\epsilon_i|w_i) f(w_i) dw_i$$

où $f(\epsilon_i)$ est la loi asymétrique de Laplace telle qu'envisagée dans cette section.

On peut remarquer que la densité introduite en (3.6) est à nouveau une loi symétrique pour p égal à 0.5 et asymétrique vers la gauche (resp. droite) pour $p < 0.5$ (resp. $p > 0.5$) sinon. Néanmoins, nous avons à présent une forme quadratique des ϵ_i.

En considérant à nouveau des distributions a priori non informatives pour les paramètres $\boldsymbol{\beta}$ et σ, il vient la distribution jointe a posteriori suivante

$$p(\boldsymbol{\beta},\sigma,w|y) \propto \sigma^{-(n+1)} \prod_{i=1}^{n} w_i^{-1/2} \exp \quad \{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{w_i} [pI_{[0,+\infty)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + (1-p)I_{(-\infty,0)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - w_i] \}.$$

et les distributions conditionnelles a posteriori suivantes

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{(y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}{w_i}|\boldsymbol{\beta},w,y\sim\chi_n^2$$

$$p(w_i|\boldsymbol{\beta},\sigma,y)\propto w_i^{-1/2}\exp\{-q_i(\boldsymbol{\beta},\sigma)w_i^{-1}-w_i\}$$

avec $q_i(\boldsymbol{\beta},\sigma)=(\frac{1}{2\sigma^2})(y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2$.

La distribution conditionnelle a posteriori de $\boldsymbol{\beta}$ est ensuite donnée par

$$p(\boldsymbol{\beta}|\sigma,w,y)\propto\exp\ \{-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}\frac{(y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}{w_i}[pI_{[0,+\infty)}(y_i-\mathbf{x}_i^T\boldsymbol{\beta})+(1-p)I_{(-\infty,0)}(y_i-\mathbf{x}_i^T\boldsymbol{\beta})-w_i]\}.$$

Cette distribution est à nouveau inconnue et nécessite l'usage de méthodes numériques. E.G. Tsionas présente dans son article [21] une procédure [3] permettant de pourvoir générer les composantes de $\boldsymbol{\beta}$, non pas toutes simultanément mais composante par composante via l'algorithme de Gibbs (1984), cet algorithme présentant l'avantage, par rapport à l'algorithme de Métropolis (1953) d'éviter les problèmes de monitoring, c'est-à-dire le problème concernant l'atteinte d'un taux d'acceptation de 0.20 ou 0.40 selon que les paramètres soient mis à jour simultanément ou non en modifiant la variance (ou la matrice de variance-covariance dans le cas multivarié) de la loi de proposition.

Nous allons rappeller brièvement comment fonctionne cet algorithme de Gibbs (1984).

Gibbs sampler

En toute généralité, on désire générer θ à partir de $p(\theta)$.

Soit $\theta=(\theta_1,...,\theta_J)$ une subdivision de θ en J sous-vecteurs.

A l'itération t, l'algorithme est

– Pour chaque composante, c'est-à-dire pour $j=1,...,J$,

3. Nous avons implémenté cette procédure dans le logiciel *R*, malheureusement, par manque de capacité de stockage, nous n'avons pu effectué un grand nombre d'itérations, et dès lors, les estimations obtenues dépendent fortement des conditions initiales choisies. C'est pourquoi nous n'avons pas appliqué cette méthode dans le chapitre 4. Notons que les auteurs n'ont pas utilisé ce logiciel mais ont implémenté la procédure via un langage de programmation dit de *GAUSS*.

– Générer θ_j^t à partir de $p(\theta_j^t|\theta_1^t, ..., \theta_{j-1}^t, \theta_{j+1}^{t-1}, ..., \theta_J^{t-1})$.

On constate donc que chaque composante est mise à jour conditionnellement aux mises à jours de l'étape précédente des autres composantes, ce qui explique pourquoi cette procédure est requise dans la méthode bayesienne de la régression quantile présentée dans cette section.
Après un nombre, (disons T), suffisamment grand d'itérations, l'ensemble $\{\theta^{T+1}, \theta^{T+2}, ..., \theta^{T+M}\}$ peut être vu comme un échantillon aléatoire distribué selon une distribution $p(\theta)$.

Terminons cette section par quelques remarques.
Premièrement, notons que la distribution a posteriori marginale des paramètres de régression sera la même quelque soit la méthode (utilisation ou non de variables latentes w_i) envisagée. En effet, la distribution marginale des erreurs étant identique dans les deux cas (distribution asymétrique de Laplace), il en sera de même quant à la distribution marginale a posteriori des $\boldsymbol{\beta}$. Néanmoins, la seconde méthode qui consiste à voir cette distribution asymétrique comme un mélange de lois normales via l'introduction de variables latentes w_i est une méthode plus efficace, d'un point de vue algorithmique, qui permet d'explorer plus facilement l'a posteriori et qui évite tout problème de monitoring en utilisant l'algorithme de Gibbs (1984) plutôt que celui de Metropolis (1953).
Enfin, rappellons que les deux approches bayesiennes présentées dans ce chapitre (approche de Yu et al (2001) et Tsionas (2003)) se basent sur la même hypothèse concernant la distribution des erreurs si ce n'est que, premièrement, Tsionas ne contraint pas le paramètre d'échelle à 1 et, deuxième, envisage une technique d'exploration de l' a posteriori différente.

3.5 Approche non paramétrique

Dans les deux premières sections de ce chapitre, nous avons présenté deux méthodes qui considéraient une loi asymétrique de Laplace pour les erreurs, loi de paramètre de localisation μ nul, de paramètre d'échelle σ quelconque dans l'approche de Tsionas (2003) et contraint à 1 dans celle de Yu et al (2001) et ce pour un paramètre d'asymétrie p fixé.
Pour rappel, le choix d'une telle distribution était motivé par le fait que, en régression quantile, les erreurs sont supposées être distribuées identiquement et indépendamment selon une loi de densité dont le quantile d'ordre p est nul, ce qui est vérifié avec la loi asymétrique de Laplace citée ci-dessus.

Le problème du choix d'une telle distribution est que le paramètre p détermine à la fois l'ordre du quantile d'intérêt et l'asymétrie de la distribution conditionnelle supposée limitant donc sa flexibilité à modéliser l'asymétrie de la distribution conditionnelle. En effet, pour $p = 0.5$, c'est-à-dire dans le cas de la régression médiane, cette distribution est symétrique.
De plus, pour un p fixé, la distribution asymétrique de Laplace a une forme spécifique à

gauche et à droite du paramètre de localisation μ qui, selon toute vraisemblance, n'est pas celle des données.

Pour plus de flexibilité, Kottas et al (2005) envisagent une estimation non paramétrique de la fonction de densité des erreurs basée sur un mélange de processus de Dirichlet. Plus précisément, ils envisagent une distribution mélangeante de lois uniformes.
Nous ne détaillerons cette approche que dans le cadre de la régression médiane [13], bien que les auteurs l'aient développée dans un cadre plus général, et même dans le cas de données censurées, car un des articles de référence à ce sujet [14] n'est pas encore publié [4], l'autre [19] venant de l'être très récemment.

Avant de continuer l'explication de cette approche dans le cadre de la régression médiane et faisant appel à ce qu'on nomme la statistique bayesienne non paramétrique, il est bon d'en rappeler voire d'en énoncer les bases. Nous commencerons par définir ce que nous appellons une distribution de Dirichlet ainsi que le contexte où est habituellement utilisée cette distribution. Ensuite, nous présenterons la notion de processus de Dirichlet, de même que celle concernant un mélange de processus de Dirichlet telles qu'elles sont présentées dans [10] et [13].

3.5.1 Notion de base

Distribution binomiale et distribution Beta

Soit n expériences indépendantes ayant deux résultats possibles : le succès ou l'échec. Soit π la probabilité de succès, et donc $1-\pi$ la probabilité d'échec, et soit Y le nombre de succès au cours de ces n expériences. Y ne peut donc prendre comme valeurs que les entiers de 0 à n.

Dans ces conditions, on dit que la variable aléatoire Y est distribuée selon une loi binomiale de paramètre n, avec $n \in \mathbb{N}_0$, et de paramètre π, $\pi \in [0,1]$, ce qui se note $Y \sim B(n,\pi)$. La distribution de Y conditionnellement à π est dès lors

$$f(y|\pi) = \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{n-y}$$

où

$$\left(\begin{array}{c} n \\ y \end{array} \right) = \frac{n!}{y!(n-y)!}.$$

4. Cet article est néanmoins disponible à l'adresse *http ://users.soe.ucsc.edu/ thanos/research.html.*

En statistique fréquentiste, π est fixé et la fonction de densité définie ci-dessus est la distribution de Y définie pour toutes les valeurs possibles de Y en $y = 0, ..., n$.

Dans un cadre bayesien, on s'intéresse à la distribution de la plausibilité a posteriori pour π sachant y. Par le théorème de Bayes, cette densité a posteriori vaut

$$g(\pi|y) \propto g(\pi)f(y|\pi)$$

où $g(\pi)$ est la distribution a priori pour π.

Si nous n'avons aucune connaissance préalable au sujet de π, nous pouvons opter pour un a priori uniforme.
Dans le cas contraire, une distribution Beta de paramètre a et b est un a priori (conjugué [5]) possible. La densité correspondante est donnée par

$$g(\pi; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1} \quad \text{pour} \quad 0 \leq \pi \leq 1,$$

avec $a; b > 0$.

Dès lors,

$$g(\pi|y) \propto \pi^{a+y-1}(1-\pi)^{b+n-y-1} \quad \text{pour} \qquad 0 \leq \pi \leq 1.$$

On a donc $\pi|y \sim Beta(a+y, b+n-y)$, ce qui implique donc que le choix d'une distribution Beta comme a priori revient à construire l'a priori sur base d'une étude antérieure dans laquelle a succès et b échecs furent observés.

Distributions multinomiale et de Dirichlet

La distribution multinomiale est définie lorsque les n expériences indépendantes évoquées précédemment ont K $(K > 2)$ résultats possibles, avec des probabilités respectives π_k $(k = 1, ..., K)$.

Dans ces conditions, il vient

$$(Y_1, ..., Y_K) \sim Mult(n; \pi_1, ..., \pi_K) \quad \text{avec} \quad \sum_{k=1}^{K} \pi_k = 1$$

où y_k est le nombre de fois que le $k^{\text{ème}}$ résultat a été observé durant les K expériences.

5. Avec un a priori conjugué, l'a posteriori appartient à la même famille de distributions que l'a priori. Notons que de tels a priori conjugués n'existent que pour les paramètres des distributions appartenant à la famille exponentielle.

La distribution a priori conjuguée pour $\pi = (\pi_1, .., \pi_K)$ est la distribution de Dirichlet, qui est une généralisation de la distribution Beta,

$$g(\pi; \alpha_1, ...\alpha_K) = \frac{\Gamma(\alpha_+)}{\Gamma(\alpha_1)...\Gamma(\alpha_K)} \pi_1^{\alpha_1 - 1} ... \pi_K^{\alpha_K - 1} \quad \text{avec} \quad \alpha_+ = \alpha_1 + ... + \alpha_K, \alpha_k > 0 \, \forall k.$$

ce qui se note $\pi \sim Dir(\alpha_1, ..., \alpha_K)$. La distribution a posteriori vaut dès lors

$$g(\pi|y) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k + y_k - 1},$$

c'est-à-dire $\pi|y \sim Dir(\alpha_1 + y_1, ..., \alpha_K + y_K)$.

Processus de Dirichlet

Soit $F(.)$ une fonction de répartition inconnue définie sur Ω un espace probabilisé. On dit que F a comme a priori un processus de Dirichlet de paramètre $c_0 F_0$, ce qui se note $F \sim DP(c_0 F_0)$, si, pour n'importe quelle partition $B_1, B_2, ..., B_k$ de Ω telle que $\Omega = B_1 \cup B_2 \cup ... \cup B_k$ avec des B_j disjoints, la distribution jointe de $(F(B_1), F(B_2), ..., F(B_k))$ est une loi de Dirichlet de paramètres $(c_0 F_0(B_1), c_0 F_0(B_2), ..., c_0 F_0(B_k))$ où, si nous considérons les B_j comme étant des intervalles disjoints de la forme $B_j = (b_{1j}, b_{2j}]$, $F(B_j)$ et $F_0(B_j)$ vallent respectivement

$$\begin{aligned} F(B_j) &= F(b_{2j}) - F(b_{1j}) \\ F_0(B_j) &= F_0(b_{2j}) - F_0(b_{1j}). \end{aligned}$$

Ici, $F_0(.)$ est une fonction de répartition spécifiée sur Ω qui joue le rôle d'hyperparamètres et qui est appelé la *mesure de base* d'un processus de Dirichlet, tandis que c_0 est un paramètre de précision sur lequelle nous reviendrons dans la suite.

D'un point de vue interprétation, on a donc une fonction de répartition $F(.)$ inconnue évaluée en chaque B_j et dont nous avons une connaissance a priori, une distribution de Dirichlet.

Le théorème suivant caractérise l'a posteriori en utilisant donc comme a priori un processus de Dirichlet.

Théorème 3.5.1. *Soit l'échantillon $\{y_1, ..., y_n\}$ provenant d'une variable aléatoire de fonction de répartition $F(.)$ et supposons que $F \sim DP(c_0 F_0)$. Dès lors*

$$F|y \sim DP\left(\; c_0 F_0(.) + \textstyle\sum_{i=1}^{n} \delta_{y_i}(.) \; \right)$$

où $y = (y_1, ..., y_n)^T$ *et* $\delta_{yi}(.)$ *est la fonction caractéristique telle que*

$$\delta_{y_i}(y) = \begin{cases} 1 & si \quad y = y_i, \\ 0 & sinon. \end{cases}$$

Donc, la distribution a posteriori de $F(.)$ *est un processus de Dirichlet de paramètres*

$$c_0 F_0(.) + \sum_{i=1}^{n} \delta_{y_i}(.) = c_0 F_0(.) + n F_n(.),$$

où $F_n(.)$ *est la fonction de répartition empirique définie comme étant*

$$F_n(y) = \frac{\#\, de\ y_j\ de\ l'échantillon \leq y}{n}.$$

Dans la pratique, les intervalles $B_1, B_2, ..., B_k$ sont construits après avoir eu connaissance de l'échantillon $\{y_1, ..., y_n\}$ de tel façon que B_j contienne au moins une donnée y_i.

Avant de décrire en quoi consiste un mélange de processus de Dirichlet, considérons un simple exemple de processus de Dirichlet permettant de fixer les idées.

Exemple 3.5.1. *Soient* $y_1, y_2, ..., y_5 \overset{iid}{\sim} F(.)$ *où l'on suppose que* $F \sim DP(c_0, F_0)$ *avec* $c_0 = 0.1$ *et* $F_0(y) = 1 - exp(-y)$ *(autrement dit,* $F_0(.)$ *est la fonction de répartition d'une loi exponentielle de moyenne 1). Supposons en outre que* $y_1 = 1$, $y_2 = 0.7$, $y_3 = 0.8$, $y_4 = 1.2$ *et* $y_5 = 1.3$ *et considérons* $B_1 = \{y : 0 < y \leq 1\}$, $B_2 = \{y : 1 < y \leq 1.25\}$, *et* $B_3 = \{y : 1.25 < y < \infty\}$. *On remarque que ces intervalles forment bien une partition disjointe de* $\Omega = [0, +\infty)$.
On a donc $F(B_1) = F(1) - F(0) = p_1, F(B_2) = F(1.25) - F(1) = p_2, F(B_3) = F(\infty) - F(1.25) = p_3$, *avec* $p_1 + p_2 + p_3 = 1$.

De plus,

$$(p_1, p_2, p_3) \sim Dir(c_0 F_0(B_1), c_0 F_0(B_2), c_0 F_0(B_3)),$$

avec

$$\begin{aligned} F_0(B_1) &= F_0(1) - F_0(0) = 1 - \exp(-1) = 0.632 \\ F_0(B_2) &= F_0(1.25) - F_0(1) = \exp(-1) - \exp(-1.25) = 0.081 \\ F_0(B_3) &= F_0(\infty) - F_0(1.25) = \exp(-1.25) = 0.287 \end{aligned}$$

Par le théorème (3.5.1)*, il vient*

$$(p_1, p_2, p_3)|y \sim Dir(c_0F_0(B_1) + nF_n(B_1), c_0F_0(B_2) + nF_n(B_2), c_0F_0(B_3) + nF_n(B_3)).$$

avec, par définition de F_n*,*

$$\begin{aligned} F_n(B_1) &= F_n(1) - F_n(0) = \tfrac{3}{5} \\ F_n(B_2) &= F_n(1.25) - F_n(1) = \tfrac{1}{5} \\ F_n(B_3) &= F_n(\infty) - F_n(1.25) = \tfrac{1}{5} \end{aligned}$$

et donc $F_n(B_1) + F_n(B_2) + F_n(B_3) = 1$*.*

Finalement, on obtient

$$(p_1, p_2, p_3)|y \sim Dir(3.063, 1.008, 1.029).$$

Mélange de processus de Dirichlet

Soit une fonction de répartition $F(.;\boldsymbol{\theta})$ telle que $\boldsymbol{\theta} \in \Omega$ et dont la fonction de densité associée est notée $f(.;\boldsymbol{\theta})$.
Si G est propre, c'est-à-dire si G vérifie les conditions pour être considérée comme étant une fonction de répartition, on définit la distribution de mélange suivante

$$F(.;G) = \int F(.;\boldsymbol{\theta})G(d\boldsymbol{\theta}) \tag{3.7}$$

où $G(d\boldsymbol{\theta})$ est vue comme étant la distribution conditionnelle de $\boldsymbol{\theta}$ selon G.
En dérivant les deux membres de l'égalité (3.7) par rapport à l'argument (.), il vient

$$f(.;G) = \int f(.;\boldsymbol{\theta})G(d\boldsymbol{\theta}).$$

Dans ce contexte, soit $D = \{Y_i, i = 1, ..., n\}$ un échantillon de $F(.;G)$ et supposons que, pour chaque Y_i (i=1,...,n), nous introduisions des variables latentes $\boldsymbol{\theta}_i$ telles que les Y_i soient indépendants conditionnellement aux $\boldsymbol{\theta}_i$. Supposons en outre que les $\boldsymbol{\theta}_i$ soient indépendants et identiquement distribués conditionnellement à G.
En marginalisant sur $\boldsymbol{\theta}_i$, il vient que les Y_i sont toujours indépendants mais conditionnellement à G avec comme densité jointe

$$\prod_{i=1}^{n} f(y_i;G) = \prod_{i=1}^{n} \int f(y_i;\boldsymbol{\theta}_i)G(d\boldsymbol{\theta}_i).$$

En spécifiant de plus que $G \sim DP(c_0 G_0)$, nous avons entièrement spécifié le modèle bayesian appellé *mélange de processus de Dirichlet*. Au niveau interprétation, c_0 reflète notre connaissance a priori au sujet de la similure entre G et la mesure de base G_0. En d'autres termes, plus c_0 est elevé, plus on peut avoir confiance en notre approximation. Une répresentation de l'effet du choix de c_0 est donnée sur (3.6).
Notons que nous pourrions également spécifier un a priori pour c_0.

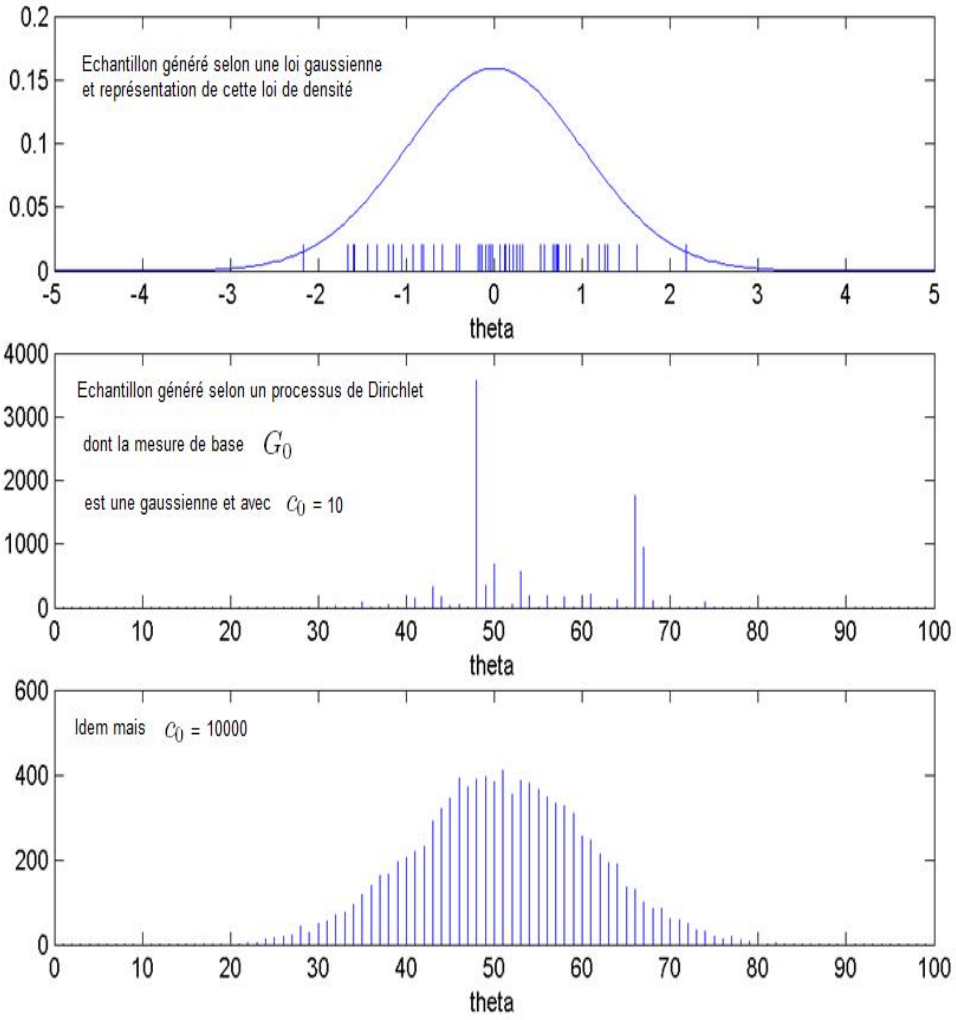

Figure 3.6 – Processus de Dirichlet.(Graphique tiré de [20])

3.5.2 Régression médiane basée sur un mélange de processus de Dirichlet

Comme nous l'annoncions précédemment, nous allons, dans cette section, présenter les idées générales de la régression médiane basées sur l'utilisation d'un mélange de processus de Dirichlet [13].

Dans un premier temps, nous allons introduire une famille de distributions semiparamétrique de médiane nulle plus flexible.

Famille semiparamétrique

Considérons une fonction de densité $f(.;\theta)$ continue, unimodale et symétrique autour de 0 (cette densité pourrait être la densité d'une Laplace de médiane nulle telle que nous l'avons définie dans ce chapitre). Bien que θ est un paramètre arbitraire, dans la suite, nous le considérons comme étant un paramètre d'échelle tel que $\theta > 0$.
Nous définissons ensuite la famille de distributions

$$p(.;\theta,\gamma) = \frac{1}{\gamma} f(\frac{.}{\gamma};\theta) I_{(-\infty,0)}(.) + \gamma f(.\gamma;\theta) I_{[0,+\infty)}(.), \tag{3.8}$$

où $\gamma > 0$. Pour des valeurs de $\gamma \neq 1$, nous avons des distributions asymétriques dont l'asymétrie dépend de la valeur de γ avec une dissymétrie à gauche pour $\gamma < 1$, et une dissymétrie à droite sinon (graphique (3.7)). Lorsque γ vaut 1, les distributions sont évidemment symétriques étant donné que $p(.,\theta,\gamma) = f(.;\gamma)$. La médiane des distributions de cette nouvelle famille est toujours 0, mais leur fonction de densité est discontinue en cette valeur.
Enfin, on peut remarquer que ces distributions sont toujours unimodales, de mode égal à 0.
La fonction de répartion de (3.8) est quant à elle donnée par

$$P(.;\theta,\gamma) = F(\frac{.}{\gamma};\theta) I_{(-\infty,0)}(.) + F(.\gamma;\theta) I_{[0,+\infty)}(.),$$

où $F(.;\theta)$ est la fonction de répartition de $f(.;\theta)$.

A présent, présentons quelques propriétés de cette famille de distributions.
Supposons une variable aléatoire ϵ telle que $\epsilon \sim p(.;\theta,\gamma)$. Le moment non centré d'ordre k ($k > 0$) de ϵ est donné par

$$E\left[\epsilon^k|\theta,\gamma\right] = (\gamma^{-k} + (-1)^k\gamma^k) m_k(\theta),$$

avec $m_k(\theta) = \int_0^{+\infty} u^k f(u;\theta) du$. On remarque donc que $E\left[\epsilon^k|\theta,\gamma\right]$ aura une valeur finie si le moment correspondant de $f(;\theta)$ est fini.

Quant à la moyenne $E[\epsilon|\theta,\gamma]$ et variance $V[\epsilon|\theta,\gamma]$ de ϵ, elles sont données respectivement par

$$\begin{aligned} E[\epsilon|\theta,\gamma] &= (\tfrac{1}{\gamma}-\gamma)m_1(\theta) \\ V[\epsilon|\theta,\gamma] &= (\tfrac{1}{\gamma^2}+\gamma^2)(m_2(\theta)-m_1^2(\theta))+2m_1^2(\theta) \end{aligned}$$

si $m_1(\theta)$ et $m_2(\theta)$ prennent des valeurs finies.
Enfin, le quantile d'ordre p de ϵ, $Q_p(\theta,\gamma)$, est également donné à partir de celui de $f(.;\theta)$, $r_p(\theta)$. En effet, on a

$$Q_p(\theta,\gamma) = \begin{cases} \gamma r_p(\theta) & \text{si} \quad 0<p<0.5 \\ \frac{r_p(\theta)}{\gamma} & \text{si} \quad 0.5<p<1. \end{cases}$$

Revenons au processus de mélange de Dirichlet. Pour une distribution G propre (dans le sens définit précédemment), considérons

$$f(.;\theta,\gamma) = \int p(.;\theta,\gamma)dG(d\theta). \tag{3.9}$$

En "mélangeant" sur θ, on préserve le fait que la médiane est toujours nulle tout en obtenant plus de flexibilité dans le modèle. Comme précédemment, on suppose que $G \sim DP(c_0G_0)$.

Venons à l'utilisation d'une telle famille et l'application d'un mélange de processus de Dirichlet en régression médiane.

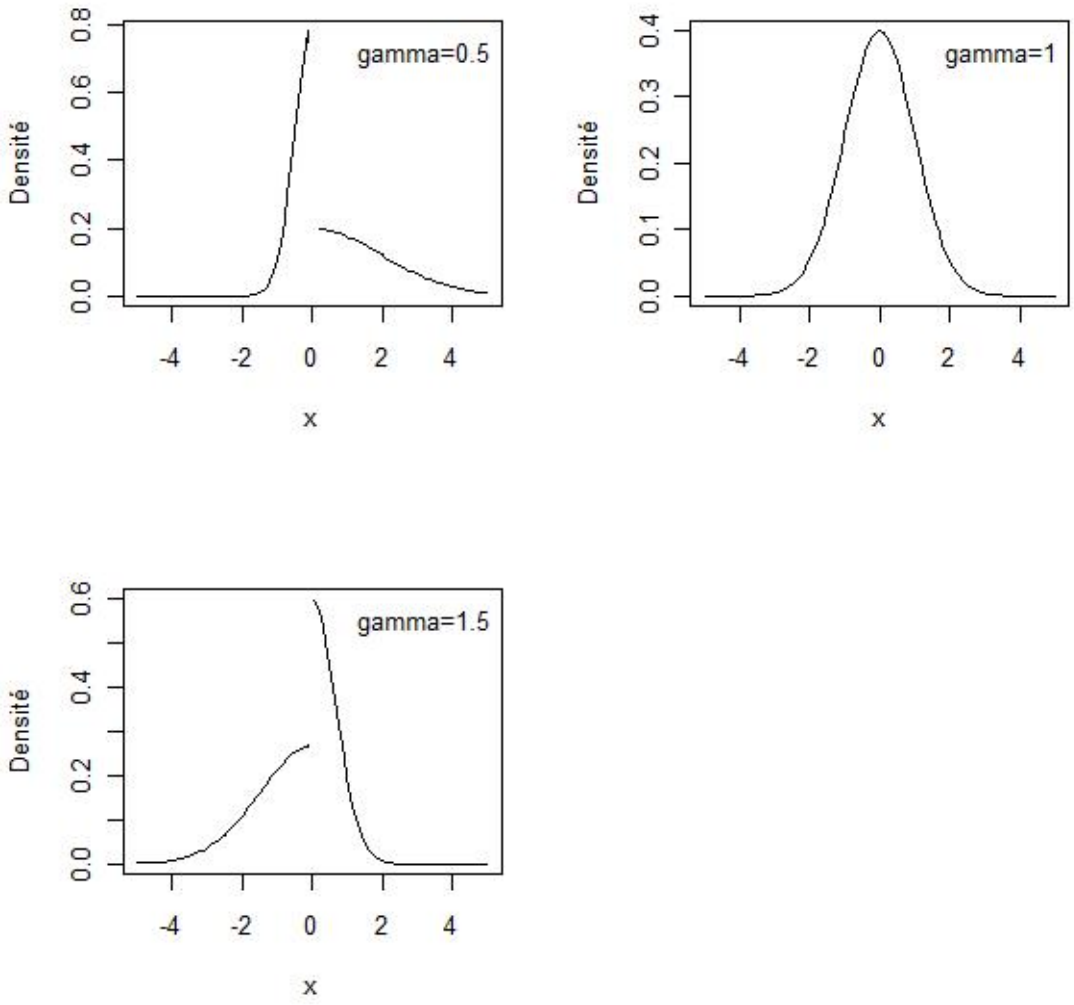

Figure 3.7 – Représentation de la famille semiparamétrique de distributions pour $\theta = 1$ et où $f(.;\theta)$ est une loi gaussienne.

Régression médiane

Selon nos notations, le modèle de régression envisagé est

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i,$$

où les erreurs ϵ_i sont supposées être distribuées identiquement et indépendamment selon une distribution de médiane nulle.
On peut également inclure un intercept dans le modèle qui devient dès lors

$$y_i = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i,$$

avec évidemment la même hypothèse concernant la distribution des erreurs que nous modélisons selon (3.9) en considérant

$$p(.;\theta,\gamma) = f_N(.|0,\theta\phi) I_{(-\infty,0)}(.) + f_N(.|0,\frac{\theta}{\phi}) I_{[0,+\infty)}(.)$$

où $f_N(.|\mu,\sigma^2)$ est la densité d'une loi normale de paramètre de localisation μ et de paramètre d'échelle σ^2 et où nous avons reparamétrisé γ en $\phi = \gamma^2$.
Pour compléter la modélisation, supposons que $G \sim DP(c_0 G_0)$ et spécifions des a priori paramétriques pour $\alpha, \boldsymbol{\beta}$ et ϕ. De plus, les auteurs considèrent comme a priori pour $(\alpha,\boldsymbol{\beta})^T$ et pour ϕ respectivement une loi multinormale et une loi Gamma de moyenne a/b. Dès lors, il vient

$$\begin{array}{rcl} Y_i|\alpha,\boldsymbol{\beta},\phi,\theta_i & \overset{ind}{\sim} & p(y_i - \alpha - \mathbf{x}_i^T\boldsymbol{\beta};\theta_i,\phi), \quad i = 1,...,n, \\ \theta_i & \overset{iid}{\sim} & G, \\ G & \sim & DP(c_0 G_0), \\ (\alpha,\boldsymbol{\beta})^T & \sim & N_{p+1}(\boldsymbol{\mu},\boldsymbol{\Sigma}), \\ \phi & \sim & Gamma(a,b), \end{array} \tag{3.10}$$

où la mesure de base G_0 est une distribution Gamma inverse de moyenne $t/(s-1)$, c'est-à-dire $G_0 \sim IGamma(t,s)$ avec $s > 1$ et où les composantes du vecteur $(\alpha,\boldsymbol{\beta})^T$ sont supposées être a priori indépendantes, et donc

$$\boldsymbol{\Sigma} = diag(\sigma_0^2, \sigma_1^2, ..., \sigma_q^2)$$

avec $\boldsymbol{\mu} = (\mu_0, \mu_1, ..., \mu_q)^T$ où tous ces hyperparamètres sont supposés fixes et où q covariables sont envisagées.

Comme nous l'avons mentionné, cette méthode fut étendue dans le cas de données censurées [13] mais aussi dans un cadre plus général de la régression quantile [14] et [19]. Notons en outre que dans [19], cette approche fut développée dans le cas de la régression Tobit mais aussi lorsque la relation unissant variable réponse et covariable ne semble pas être linéaire.
Ces approches devenant de plus en plus techniques, nous n'irons pas plus loin dans leurs

explications, les détails se trouvant dans les articles réferrés avec notamment quelques applications.

Terminons ce chapitre par une application de cette procédure, application tirée de [13].

Exemple 3.5.2. *Cet exemple consiste à étudier la performance de la méthode que nous venons d'introduire à estimer la distribution suivante qui est un mélange de lois normales,* $p_1 f_N(.|\mu_1, \sigma_1^2) + p_2 f_N(.|\mu_2, \sigma_2^2) + (1 - p_1 - p_2) f_N(.|\mu_3, \sigma_3^2)$, *avec* $p_1 = 0.435$, $p_2 = 0.43$, $\mu_1 = -0.4$, $\sigma_1 = 0$, $\mu_2 = 0$, $\sigma_2 = 1.5$, $\mu_3 = 5$, $\sigma_3 = 3$.
Sur le graphique (3.8), *on remarque (droite en trait plein) que cette distribution est unimodale et asymétrique. Sa médiane étant égale à 0, nous pouvons la modéliser en utilisant* (3.9).
Les auteurs ont ensuite simulé un échantillon de taille 250 provenant de la distribution de mélange et ont ensuite appliqué la procédure (3.10) *en posant* $c_0 = 1$, $s = 2$, $t = 8.027$, $a = b = 2.5$, *et ce en excluant les coefficients de régression. Ils obtiennent le graphique* (3.8) *où sont représentées la densité prédite a priori (en pointillé) ainsi que celle a posteriori (représentée par des tirets).*

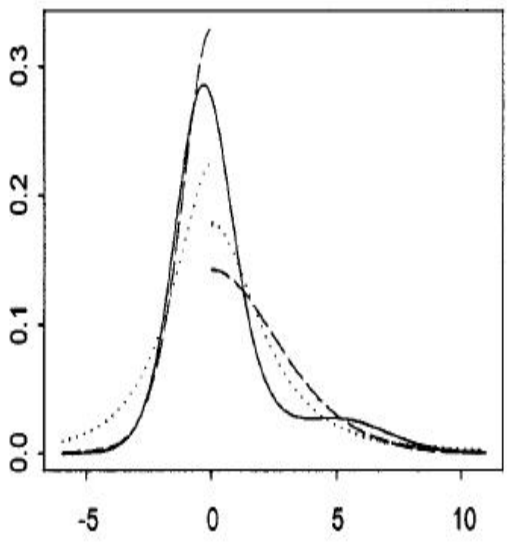

Figure 3.8 – Estimation de densité.(Graphique tiré de [13])

Chapitre 4

Applications

4.1 Introduction

Dans ce dernier chapitre, nous allons reprendre deux des exemples portant sur les échantillons "Engel data" et "mcycle", déjà succintement présentés au cours des chapitres précédents en les détaillant plus longuement (prédiction et interprétation des résultats pour l'échantillon "Engel data" ; discussion du choix des "bandwith" dans l'approche par polynômes locaux et utilisation du modèle GAMLSS pour l'échantillon "mcycle") mais surtout en appliquant une méthode bayesienne si la structure des données le permet.

Nous introduirons aussi un troisième échantillon de données nommé "stackloss" et disponible dans le logiciel *R* comportant plus d'une variable explicative. En effet, la méthode fréquentiste paramétrique de Koenker ainsi que les méthodes bayesiennes pouvant être appliquées dans le cas où plus d'une covariable est envisagée, il est bon d'en montrer une application.

Remarquons que nous n'avons pas choisi ces trois exemples au hasard. En effet, ils reprennent quelques principaux cas de figure ayant lieu en pratique :

- une seule covariable dont la relation avec la variable dépendante semble être linéaire ("Engel data"). Dès lors, on peut envisager d'appliquer une méthode fréquentiste paramétrique et les approches bayesiennes ne devraient pas poser de problèmes,

- une seule covariable et structure des données particulière où l'on doit avoir recours à des méthodes fréquentistes non paramétriques ("mcycle") aussi les approches bayesiennes présentées dans le chapitre précédent seront-elles d'application ?

- enfin, un exemple avec plus d'une covariable ("stackloss") montrant que la méthode

de Koenker et les approches bayesiennes peuvent être utilisées dans ce contexte.

Remarquons que, dans nos exemples, les covariables sont continues, néanmoins, les approches de la régression quantile présentées dans ce mémoire restent d'application lorsque les covariables sont catégorielles.

4.2 Exemple 1 : Engel data

4.2.1 Approche de Koenker

Dans le chapitre 2, nous avions déjà présenté cet ensemble de données où pour rappel nous étudions la relation entre la dépense alimentaire annuelle en euro et le salaire annuel en euro également.
Vu la représentation des données, supposer une relation linéaire entre ces deux variables ne semble pas inapproprié. Le graphique (4.1) représente les droites de régression pour des valeurs de $p = \{0.05, 0.25, 0.75, 0.95\}$, ainsi que celle correspondant à la médiane et la droite de régression obtenue par la méthode des moindres carrés. On peut remarquer que ces deux dernières ne se confondent pas, suggérant que la distribution des dépenses alimentaires conditionnellement au revenu n'est pas symétrique.
Le tableau (4.1) reprend les estimations obtenues par la méthode de Koenker des coefficients de régression (nous avons envisagé un modèle avec intercept) pour les différentes valeurs de p envisagées ainsi que les bornes des intervalles de confiance à 95%. On constate que les valeurs plausibles pour $\hat{\beta}_1(p)$ sont toutes strictement positives et ce quelque soit la valeur de p, ce qui implique que chacun des quantiles conditionnels considérés pour la variable dépendante augmente avec le salaire annuel, et ce d'autant plus rapidement que l'on s'intéresse à des quantiles d'ordre plus élevé. Nous pouvions déjà remarquer cette tendance sur le graphique (4.1).
Nous avons également mentionné les estimations des paramètres de régression par la méthode des moindres carrés dans le tableau (4.2), estimations différentes de celles de la médiane ce qui est logique au vu du graphique (4.1).

4.2.2 Approche de Yu et al (2001)

L'intérêt de ce mémoire étant l'application de méthodes bayesiennes dans le cadre des quantiles de régression, voyons les résultats donnés par la méthode de Yu et al (2001) en utilisant la première implémentation de l'algorithme présenté en annexe, c'est-à-dire celle consistant à mettre à jour tous les paramètres simultanément.
Comme nous le mentionnons dans cette annexe, nous devons choisir des valeurs initiales pour les quantiles de régression ainsi que d'une matrice de variance-covariance pour ces paramètres. Dans cet exemple, nous avons opté pour les estimations données par la mé-

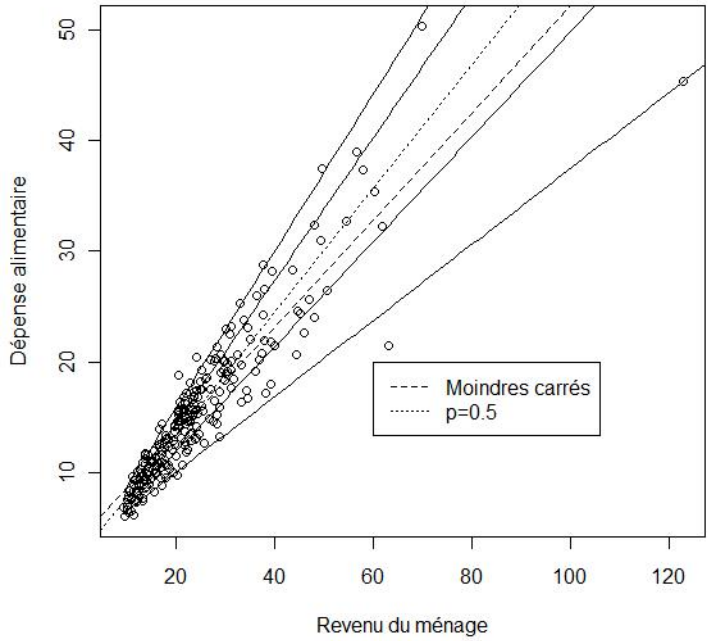

Figure 4.1 – Régression quantile pour $p = (0.05, 0.25, 0.5, 0.75, 0.95)$ et par la méthode des moindres carrés.

Quantiles de régression	Estimation	Borne inf.	Borne sup.
$\hat{\beta}_0(0.05)$	3.096	2.437	3.235
$\hat{\beta}_1(0.05)$	0.343	0.343	0.390
$\hat{\beta}_0(0.25)$	2.367	1.829	2.977
$\hat{\beta}_1(0.25)$	0.474	0.420	0.494
$\hat{\beta}_0(0.5)$	2.020	1.320	2.826
$\hat{\beta}_1(0.5)$	0.560	0.487	0.602
$\hat{\beta}_0(0.75)$	1.547	0.812	2.660
$\hat{\beta}_1(0.75)$	0.644	0.580	0.690
$\hat{\beta}_0(0.95)$	1.589	1.147	2.072
$\hat{\beta}_1(0.95)$	0.710	0.674	0.734

TABLE 4.1 – Estimation des quantiles de régression.

Coefficient de régression	Estimation	Borne inf.	Borne sup.
$\hat{\beta}_0$	3.656	2.881	4.431
$\hat{\beta}_1$	0.485	0.457	0.513

TABLE 4.2 – Estimation par la méthode des moindres carrés.

thode de Koenker, ainsi que pour la matrice de variance-covariance obtenue par bootstrap et fournie via le package "quantreg". Notons que la convergence des chaînes MCMC est garantie quelles que soient les conditions initiales ou la matrice de variance-covariance retenue. L'exemple suivant sera traité en considérant différentes valeurs initiales afin d'en déterminer l'impact sur la convergence des chaînes.
Revenons à notre échantillon "Engel data". Nous avons effectué 10000 itérations et avons visé à chaque fois un taux d'acceptation de 0.20 tel que recommandé dans [5].
Sur le graphique (4.2), les droites de régression sont représentées pour les deux méthodes, c'est-à-dire la méthode de Koenker (droites discontinues) et celle de Yu et al (2001)(droites en trait plein) selon la procédure explicitée ci-dessus. On remarque que ces deux approches donnent des résultats allant dans le même sens avec des droites très proches voire se confondant.
Expliquons comment ces droites de régression ont été obtenues selon la méthode bayesienne envisagée ici.
Pour une valeur fixée de p, nous avons donc généré 10000 paramètres de la régression quantile $\boldsymbol{\beta}(p) = (\beta_0(p), \beta_1(p))^T$ via la méthode MCMC mettant à jour les paramètres simultanément comme nous venons de le signaler.
Ensuite, pour chaque estimation de ces paramètres $\hat{\boldsymbol{\beta}}(p)$, nous avons estimé le quantile conditionnel d'ordre p pour une valeur fixée de la covariable.
En effet, rappellons que

$$Q_Y(p|\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}(p)$$

où Y est ici la dépense alimentaire annuelle et où $\mathbf{x}$ est un vecteur dont la première composante est 1 et la seconde une valeur fixée de la covariable, le salaire annuel.
Dès lors, il vient

$$\hat{Q}_Y(p|\mathbf{x}) = \mathbf{x}^T \hat{\boldsymbol{\beta}}(p).$$

Nous obtenons donc un échantillon de taille 10000 du quantile conditionnel d'ordre p de la variable réponse pour une valeur fixée de p et de la covariable. Or, pour représenter les droites de régression, nous désirons une estimation ponctuelle de ce quantile conditionnel. Pour ce faire, nous choisissons la moyenne de cet échantillon. En d'autres termes, l'estimation du quantile conditionnel d'ordre p est donné par la moyenne de l'échantillon reprenant les estimations de ce quantile conditionnel pour chaque valeur des paramètres de régression générés, et ce pour une valeur fixée de p et de la covariable.
Remarquons que le choix de la moyenne est peut être discutable. En effet, ne vaut-il pas mieux considérer la médiane qui est plus robuste ? Néanmoins, comme nous le verrons, le

choix entre moyenne ou médiane importe peu, la distribution des paramètres étant quasiment symétriques.
Nous avons ensuite réitéré la procédure pour d'autres valeurs fixées de la covariable puis de p pour obtenir les différentes droites de régression.

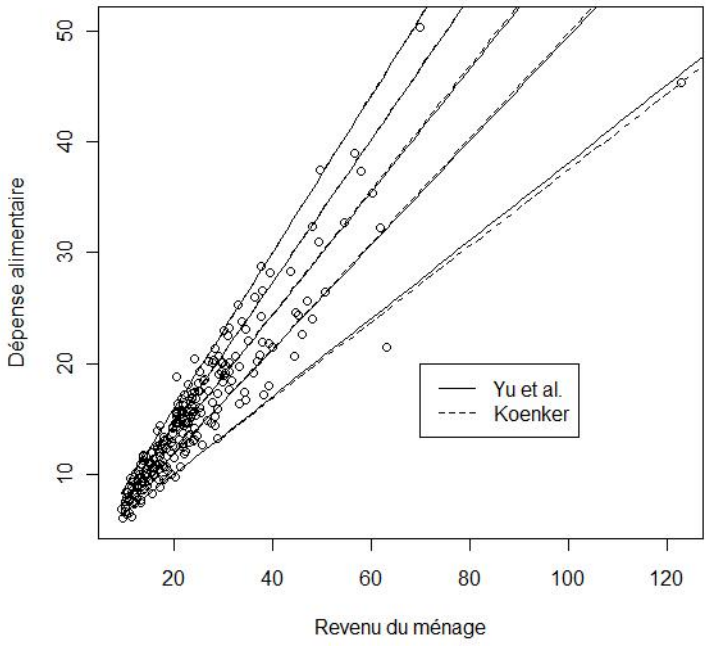

Figure 4.2 – Régression quantile pour des valeurs de p égales à 0.05, 0.25, 0.5, 0.75 et 0.95.

Prédiction

Nous venons de représenter les droites de régression selon les méthodes de Koenker et de Yu et al (2001).
Nous pouvons, via ces droites et les estimations des paramètres de régression, prédire, comme nous l'avions déjà fait dans le chapitre 2 de ce mémoire, certaines dépenses alimentaires pour des valeurs fixées du revenu. Par exemple, en se basant uniquement sur les estimations des paramètres de régression données par la méthode fréquentiste, il s'avère que pour une valeur de la covariable de 20, il y a 25% de chance que la dépense alimentaire soit inférieure à 11.849 tandis que pour un revenu annuel de 80, il y a 75% de chance que la dépense alimentaire soit inférieure à 40.295.

En bayesien, nous pouvons avoir une image plus complète du quantile conditionnel d'ordre 0.25 étant donné que nous pouvons obtenir la distribution de ce quantile pour une valeur fixée de la covariable.
Dans un premier temps, vérifions la convergence des chaînes des paramètres de la régression quantile d'ordre 0.25 générés (toujours selon la démarche expliquée dans la sous-section précédente).
Au vu des graphiques (4.3), il s'avère que nous n'avons pas de présence de "burn in", ce qui n'est guère surprenant vu les conditions initiales choisies (estimations des paramètres de régression et matrice de variance-covariance données par la méthode de Koenker). On remarque en outre que l'auto-corrélation diminue avec le nombre d'itérations (graphique (4.4)), ce qui est souhaité. Dans le tableau (4.3) sont repris moyenne, médiane et intervalle de crédibilité à 95% de ces paramètres de régression $\hat{\boldsymbol{\beta}}(0.25) = (\hat{\beta}_0(0.25), \hat{\beta}_1(0.25))^T$. Premièrement, on peut remarquer que les valeurs médiane et moyenne sont proches, ce qui indique une certaines symétrie dans la distribution marginale de chacun de ces paramètres. On peut ensuite noter que les valeurs moyennes ne sont pas très éloignées des estimations données par la méthode Koenker. Enfin, notons que les valeurs crédibles pour $\hat{\beta}_0$ (resp. $\hat{\beta}_1$) sont à 95% comprises entre 1.699 et 3.119 (resp. 0.435 et 0.502).

A présent, recherchons la distribution du quantile conditionnel d'ordre 0.25 et ce pour une valeur de la covariable fixée à, dans un premier temps, 20 et ensuite à 80. Pour ce faire, nous suivons la même démarche que celle expliquée précédemment pour obtenir les droites de régression.
Les distributions de ces quantiles conditionnels sont visibles sur les graphiques (4.5). On remarque que, pour un revenu annuel de 20, la distribution du quantile conditionnel est plutôt symétrique avec une valeur moyenne et médiane très proches respectivement de 11.815 et 11.817, et dont 95% des valeurs plausibles sont comprises entre 11.492 et 12.113. Notons que la valeur moyenne de 11.815 est assez proche de l'estimation de 11.849 obtenue en fréquentiste, ce qui pouvait déjà se déduire en regardant les droites de régression obtenues selon ces deux méthodes, droites quasiment confondues. D'un point de vue interprétation, on déduit de ces nouveaux résultats que pour un revenu annuel de 20, il y a 25% de chance que la dépense soit en moyenne inférieure à 11.815. Pour un revenu de 80, il y a 25% de chance que la dépense soit en moyenne inférieure à 40.090, ce qui est à nouveau proche de l'estimation donnée par Koenker, ce quantile conditionnel étant à 95% compris entre 37.771 et 41.943, ce qui est bien plus élévée que lorsque le salaire annuel est de 20. La distribution de ce quantile conditionnel est plutôt dissymérique à droite avec une valeur médiane de 41.182 légèrement supérieure à la valeur moyenne.
En conclusion, les personnes possédant un revenu annuel de 20 ont donc 25% de chance de dépenser 28.241 en moins que ceux ayant un revenu de 80, ce qui n'est guère surprenant. On pourrait conclure à cette même conclusion pour d'autres quantiles conditionnels au vu du graphique (4.2) en remarquant que, chez les personnes les plus dépensières, l'effet de la covariable est plus marqué.

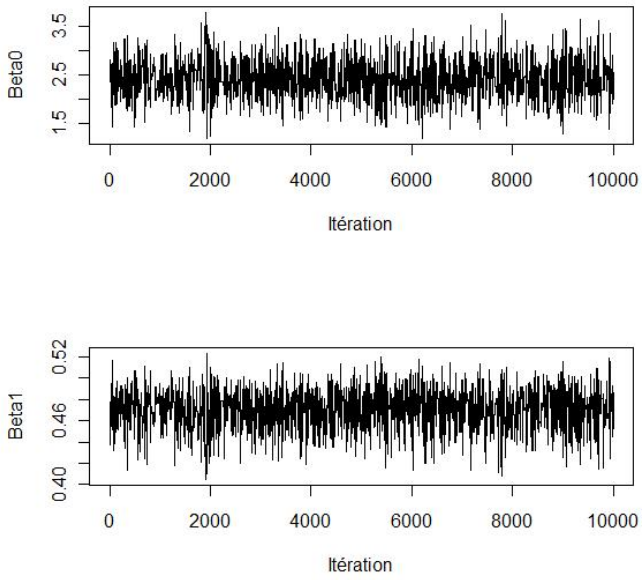

Figure 4.3 – Chaînes.

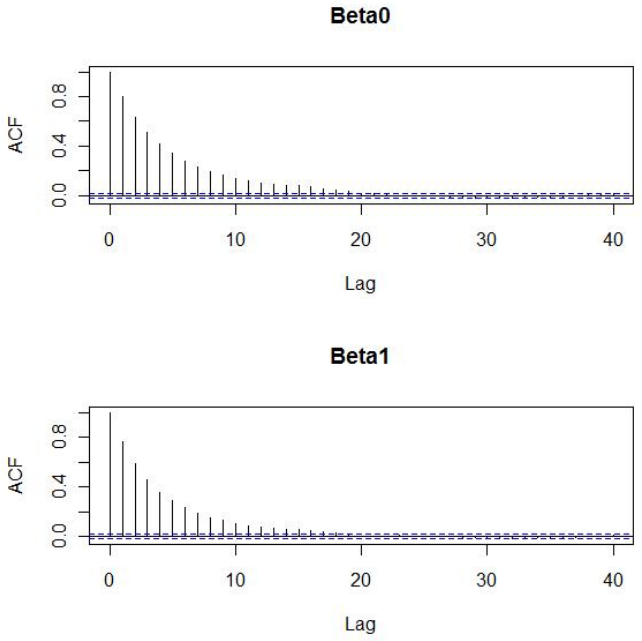

Figure 4.4 – Auto-corrélation.

Quantiles de régression	Moyenne	Quantile 0.025	Médiane	Quantile 0.975
$\hat{\beta}_0(0.25)$	2.390	1.699	2.379	3.119
$\hat{\beta}_1(0.25)$	0.471	0.435	0.473	0.502

TABLE 4.3 – Estimation des quantiles de régression pour $p = 0.25$.

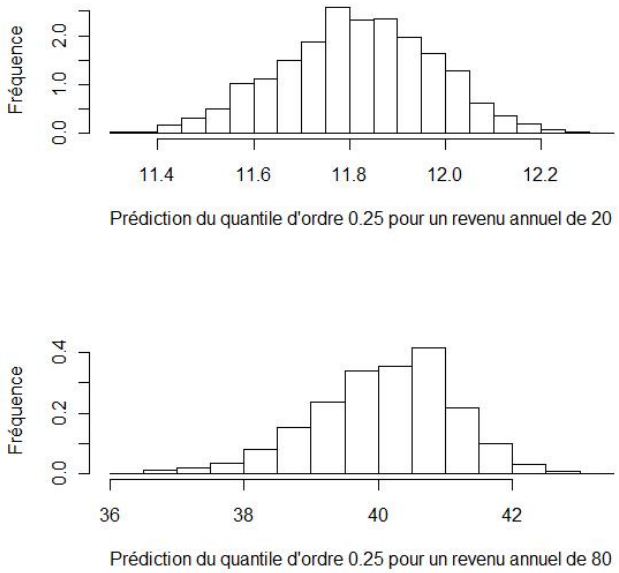

Figure 4.5 – Prédictions.

Cet exemple ne compte qu'une seule variable explicative, ce qui ne veut pas dire que nous ne pouvons pas envisager un exemple avec plus d'une covariable. C'est d'ailleurs ce que nous allons faire dans la section suivante où la méthode bayesienne utilisée sera plus longuement détaillée notamment en déterminant l'influence du choix des conditions initiales sur la convergence des chaînes.

4.3 Exemple 2 : "Stackloss"

4.3.1 Présentation des données

L'échantillon "stackloss" est composé de 21 données qui sont en fait 21 jours consécutifs durant lesquels fut observé la transformation par une plante de l'ammoniaque en acide nitrique, l'acide ainsi produit étant ensuite absorbé dans un tube. Les variables étudiées sont les suivantes :

- la variable dépendante *stack loss* qui est le pourcentage d'ammoniaque ($\times$ 10) non transformé en acide nitrique,
- les trois covariables
  - le flux d'air vers la plante,
  - la température de l'eau,
  - la concentration en acide circulant dans le tube.

4.3.2 Etude du quantile conditionnel d'ordre 0.75

Nous n'allons pas effectuer l'étude pour tous les quantiles possibles, mais nous allons plutôt nous intéresser à un quantile particulier, le quantile d'ordre 0.75, les autres cas se traitant de manière similaire.

Signalons tout d'abord les estimations des paramètres de régression données par la méthode de Koenker. Elles sont reprises dans le tableau (4.4). On constate qu'hormis pour la première covariable (flux d'air), 0 est une valeur plausible pour les trois paramètres de régression (notons en outre que le paramètre de régression correspondant à la covariable représentant la concentration en acide est d'ailleurs estimé à 0). Voyons ce qu'il en est avec la méthode proposée par Yu et al.

Impact des conditions initiales

Dans un premier temps, considérons comme valeurs initiales des paramètres les estimations données par la méthode Koenker et choisissons également une matrice de variance-covariance fournie par cette méthode et obtenue par boostrap. Nous avons effectué 10000

Quantiles de régression	Estimation	Borne inf.	Borne sup.
$\hat{\beta}_0(0.75)$	-54.190	-61.163	8.484
$\hat{\beta}_1(0.75)$	0.871	0.533	1.206
$\hat{\beta}_2(0.75)$	0.983	-0.538	1.782
$\hat{\beta}_3(0.75)$	0.000	-0.517	0.053

TABLE 4.4 – Estimation des quantiles de régression pour $p = 0.75$.

itérations. Comme nous pouvons le constater sur le graphique (4.6), en considérant ces valeurs initiales et cette matrice de variance-covariance, nous ne visualisons pas de phase de "burn in", et on peut constater que l'auto-corrélation diminue avec le nombre d'itérations (4.7). Enfin, le graphique (4.8) représente quant à lui les distributions jointes des paramètres de la régression quantile.

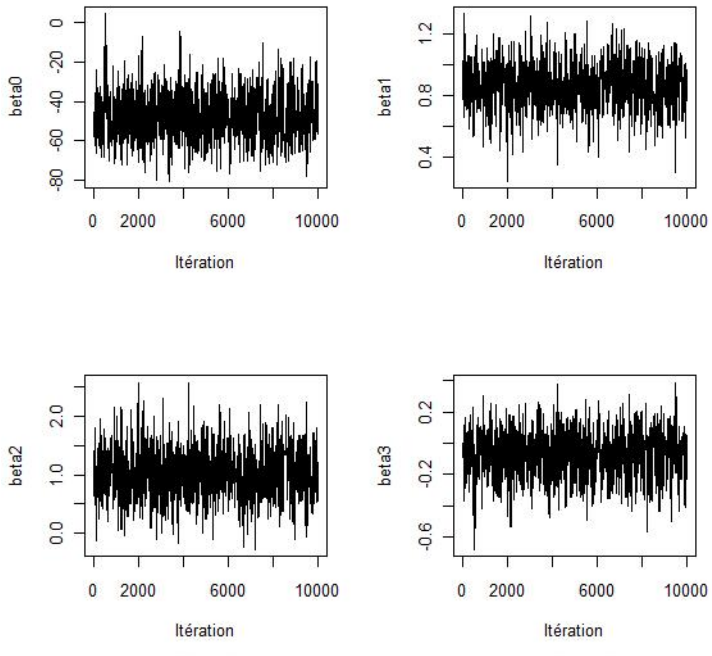

Figure 4.6 – Chaînes.

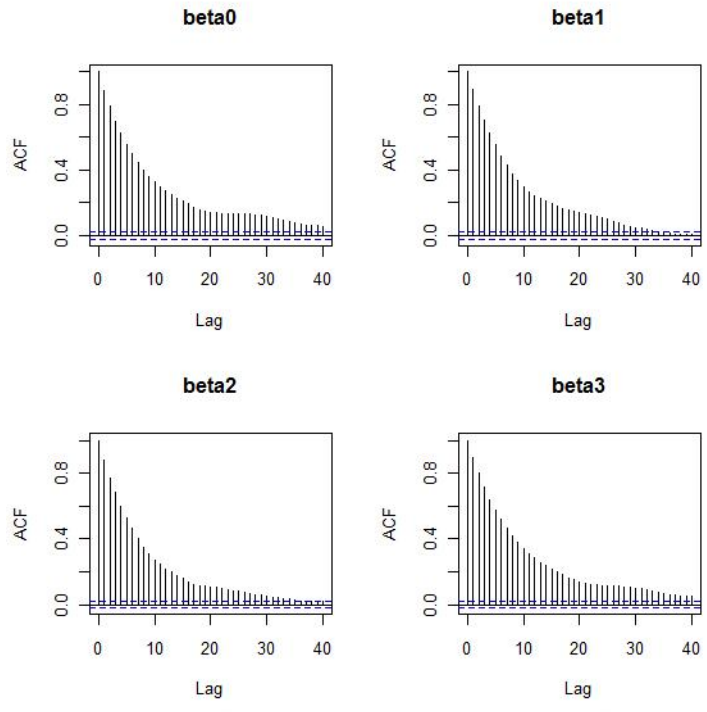

Figure 4.7 – Auto-corrélation.

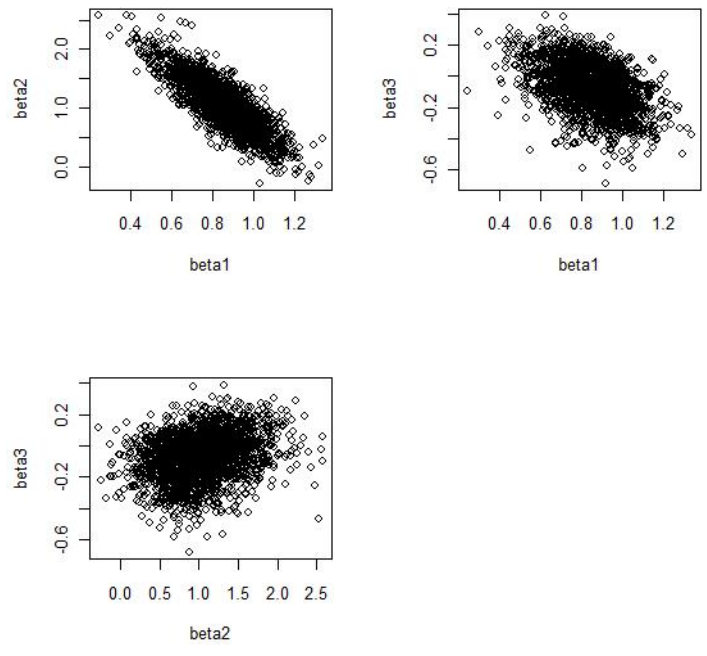

Figure 4.8 – Distributions jointes.

Quantiles de régression	Moyenne	Quantile 0.025	Médiane	Quantile 0.975.
$\hat{\beta}_0(0.75)$	-48.325	-66.920	-49.666	-20.938
$\hat{\beta}_1(0.95)$	0.856	0.587	0.851	1.150
$\hat{\beta}_2(0.95)$	1.042	0.250	1.045	1.855
$\hat{\beta}_3(0.95)$	-0.065	-0.410	-0.043	0.178

TABLE 4.5 – Estimation par la méthode Yu et al des quantiles de régression pour $p = 0.75$

Comme nous l'avons déjà mentionné, rien ne nous empêche de choisir d'autres valeurs initiales. En considérant comme valeurs initiales 1 pour chaque paramètre, nous voyons apparaître une phase de "burn in" (graphique (4.9)). Nous pouvons observé le même genre de phénomène en considérant d'autres valeurs initiales de celles choisies en premier lieu.
Dès lors, afin d'estimer moyenne, médiane, intervalle de crédibilité, et afin d'obtenir la distribution de chaque paramètre, nous devons ignorer les premières valeurs itérées à cause de cette phase de "burn in".
En mettant de côté les 1000 premières valeurs générées, nous obtenons les renseignements sur la distribution de chaque paramètre de la régression quantile repris dans le tableau (4.5). On constate à présent que seule la covariable relative à la concentration en acide ne semble pas être indispensable au modèle vu que 0 est à nouveau une valeur plausible pour le paramètre de régression relatif à cette variable explicative.
Le graphique (4.10) nous permet de visualiser l'auto-correlation après avoir retiré ces 1000 premières valeurs. On constate dès lors que l'auto-correlation diminue avec le nombre d'itérations.
Enfin, les graphiques (4.11) représentent la distribution de chaque paramètre de la régression quantile, toujours en ayant mis de côté les 1000 premières valeurs générées. On constate que les distributions pour $\hat{\beta}_1$ et $\hat{\beta}_2$ sont plutôt symétriques, ce qui n'est guère étonnant vu les valeurs très proches de la médiane et de la moyenne.
On remarque par contre une légère dissymétrie pour $\hat{\beta}_0$ et $\hat{\beta}_3$, dissymétrie à gauche pour $\hat{\beta}_0$ avec donc une valeur moyenne supérieure à celle correspondant la médiane, et une dissymétrie à droite où la valeur moyenne est dans ce cas inférieure à celle de la moyenne, comme nous pouvions déjà le constater dans le tableau (4.5).

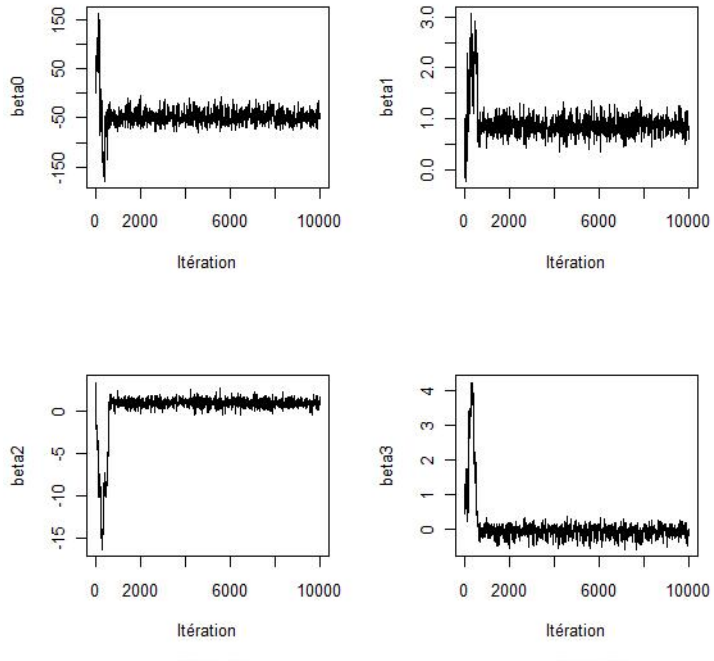

Figure 4.9 – Chaînes avec 1 comme valeur initiale pour chaque paramètre.

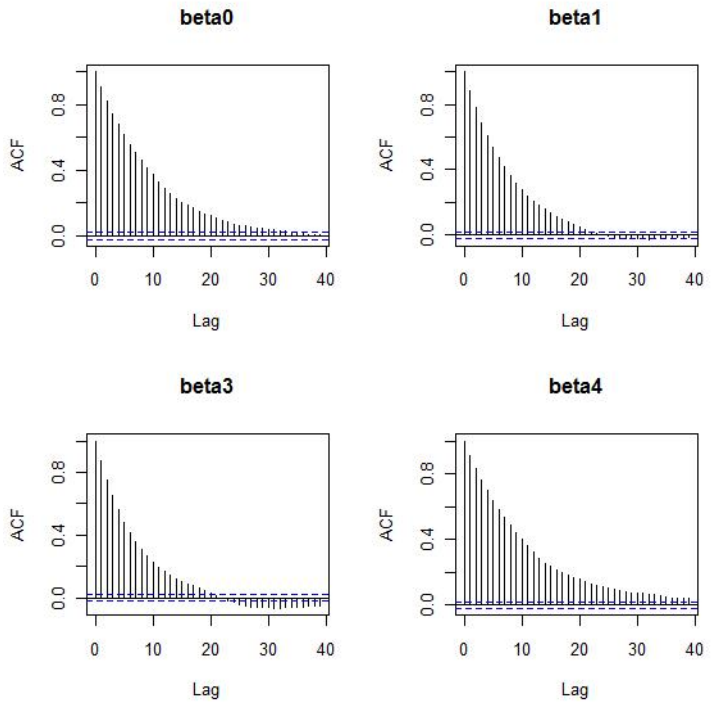

Figure 4.10 – Auto-corrélation avec 1 comme valeur initiale pour chaque paramètre.

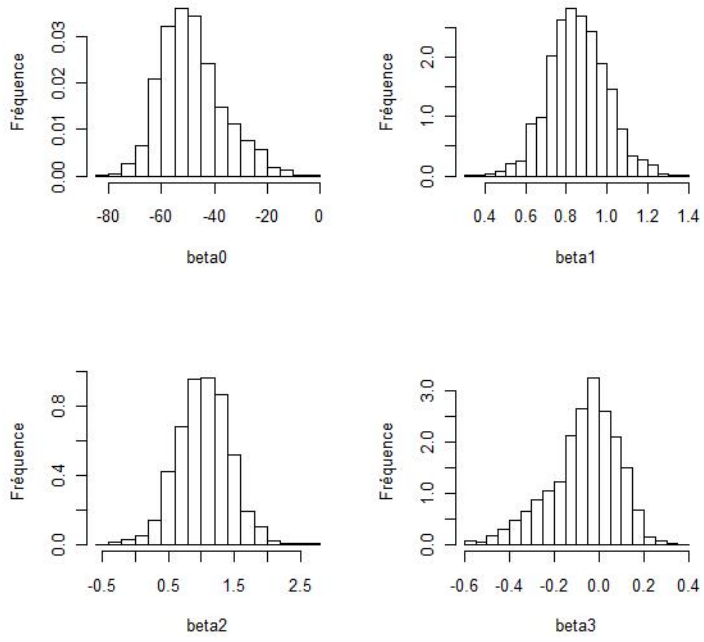

Figure 4.11 – Distribution de chaque quantile de régression avec 1 comme valeur initiale.

En régression, nous désirons connaître l'impact de chaque covariable sur l'espérance de la variable dépendante. Nous pouvons agir de même vis à vis des quantiles. Si nous calculons la probabilité que chaque coefficient soit strictement positif, on obtient une probabilité de 1 pour le paramètre de la régression quantile relatif à la première covariable, 0.990 pour la seconde et 0.356 pour la dernière. Ceci signifie donc que le quantile conditionnel d'ordre 0.75 de la variable dépendante tend à augmenter avec les deux premières covariables tandis que la troisième a plutôt un effet négatif voire nul, ce qui n'est guère étonnant en regard des intervalles de crédibilité.

Prédiction

Enfin, pour clôturer cet exemple, recherchons la distribution du quantile conditionnel d'ordre 0.75 pour des valeurs fixées des covariables. Pour rechercher cettte distribution, la démarche est quasiment la même que celle présentée dans l'exemple précédent, à savoir estimer le quantile conditionnel d'ordre 0.75 pour chaque paramètre généré, en omettant les 1000 premières valeurs, et ce pour des valeurs fixées à 50, 20 et 80 pour les covariables et en tenant compter d'un intercept.
Nous obtenons dès lors la distribution représentée sur la figure (4.12). On constate que cette distribution est dissymétrique à gauche avec une valeur moyenne 10.097 et médiane de 9.955. On notera également que 95 % des valeurs plausibles pour ce quantile conditionnel sont comprises entre 8.163 et 12.919.
On en conclut que, pour ces valeurs fixées des covariables, il y a 75% de chance que le pourcentage d'ammoniaque soit en moyenne inférieurs à 0.101%.

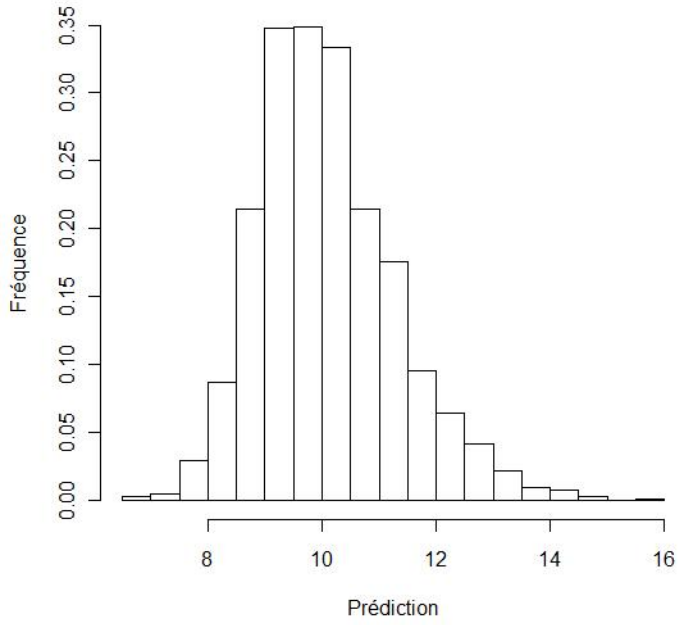

Figure 4.12 – Prédiction pour des valeurs des covariables de 50, 20 et 80, et une valeur de p de 0.75.

Remarque 4.3.1. *Nous aurions pu utiliser la seconde version de l'algorithme donnée en annexe, cela n'aurait rien changé quant aux résultats obtenus. Néanmoins, si nous avions opté pour cette version, nous aurions dû centrer les variables afin de ne pas obtenir une trop forte corrélation entre les chaînes générées.*

4.4 Exemple 3 : "Mcycle"

4.4.1 Méthodes non paramétriques

Cet échantillon de données fut traité par la méthode des B-splines, avec et sans pénalité, ainsi que par polynômes locaux dans le chapitre 2. Il s'est avéré, lorsque nous avions appliqué la méthode des splines avec pénalité, qu'une pénalité de 1 semblait être un bon compromis entre biais et variance de l'estimateur, et ce pour le quantile conditionnel d'ordre 0.25. De même, lors de la méthode par polynômes locaux, nous avions conclu qu'une valeur du *"bandwith"* h de 1 était à nouveau, au vu du graphique, un bon compromis entre biais et variance de l'estimateur pour la régression médiane mais ce n'était pas le cas pour les quantiles conditionnels d'ordre 0.25 et 0.75.
Au vu de (4.13) et (4.14), il semble que des valeurs de h égales à 2 pour les quantiles conditionnels d'ordre 0.25 et 0.75 ne soient pas trop mauvaises dans le sens qu'elles semblent à nouveau satisfaire le compromis souhaité entre biais et variance de l'estimateur. Sur le graphique (4.15) sont représentées les trois courbes de régressions pour p égal à 0.25, 0.5 et 0.75 et un *"bandwith"* respectivement de 2, 1 et 2.

Remarquons que sur le graphique (4.15) ainsi que le graphique (2.4) présenté dans la section 2 du chapitre 2, les courbes de régression se croisent, ce qui peut ne pas sembler logique. En effet, les courbes quantiles ne devraient pas se croiser mais il arrive que leurs estimations se croisent.

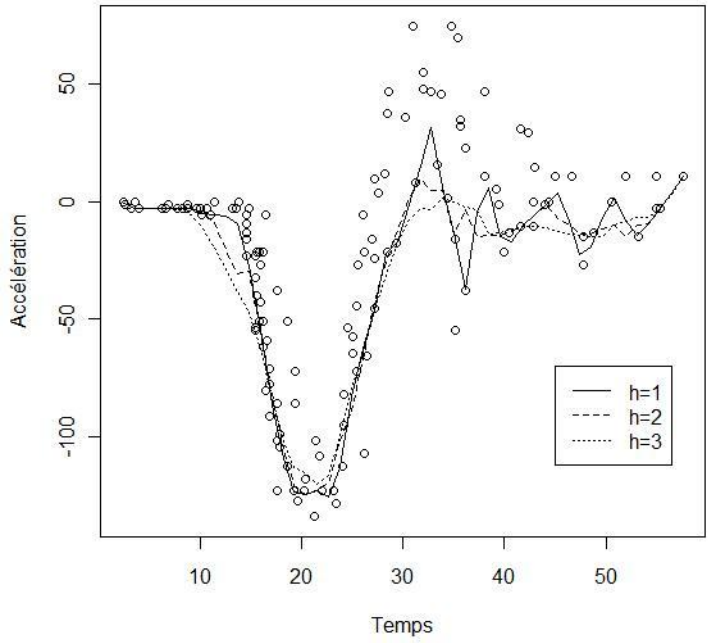

Figure 4.13 – Régression quantile pour p égal à 0.25 et différentes valeurs de h.

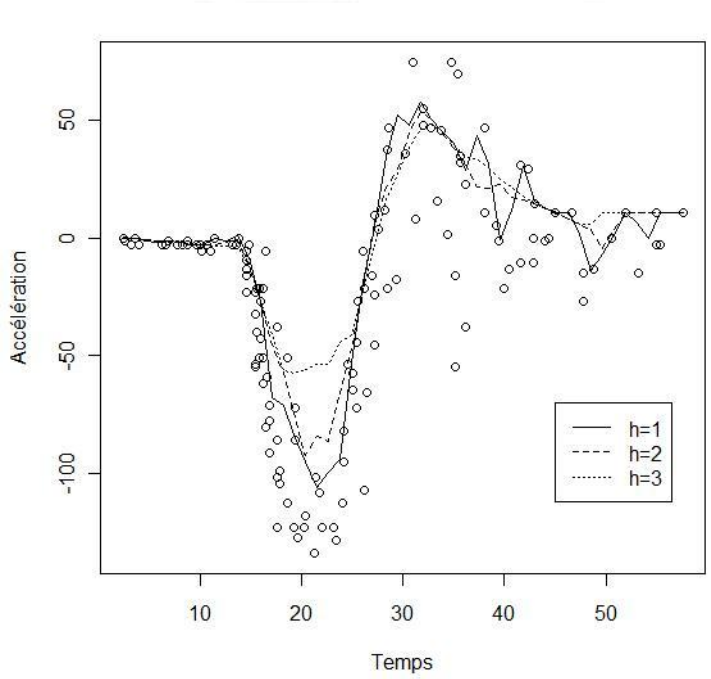

Figure 4.14 – Régression quantile pour p égal à 0.75 et différentes valeurs de h.

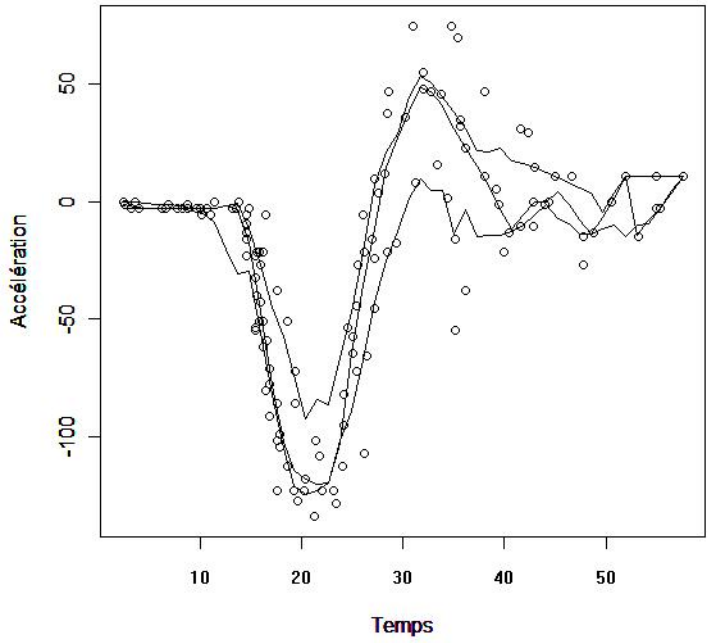

Figure 4.15 – Régression quantile pour p égal à 0.25, 0.5 et 0.75 et h valant respectivement 2, 1 et 2.

4.4.2 Modèle *GAMLSS*

Comme nous venons de le mentionner, il arrive que les estimations des courbes de régression se croisent, ce qui semble contraire à toute logique. Afin de remédier à ce soucis, nous pourrions traiter ces données via le modèle *GAMLSS* présenté dans la dernière section du chapitre 2. Via cette méthode, les courbes ne se croiseront pas. En effet, comme nous l'avons déjà expliqué, dans cette méthode, nous cherchons à modéliser les paramètres de localisation, d'échelle et de forme en fonction des covariables. Une fois les paramètres de régression estimés par maximisation d'une fonction de vraisemblance pénalisée, nous pouvons estimer ces paramètres de localisation, d'échelle et de forme pour chaque valeur de la covariable. Ensuite, nous recherchons les quantiles conditionnels d'ordre p en substituant ces estimations dans, si nous considérons comme distribution conditionnelle pour la variable dépendante une loi de Box-Cox [1],

$$y_p = \begin{cases} \mu(1+\sigma\nu t_{\tau,p})^{1/\nu} & \text{si} \quad \nu \neq 0, \\ \mu \exp(\sigma t_{\tau,p}) & \text{sinon,} \end{cases}$$

où $t_{\tau,p}$ est le quantile d'ordre p d'une distribution de Student avec τ degré de liberté.

Dès lors, pour une valeur fixée de la covariable, seuls les quantiles de la loi de Student varient avec p. Or, nous avons toujours $t_{\tau,p'} > t_{\tau,p''}$ pour $p' > p''$, ce qui explique pourquoi, ici, les courbes de régression ne se croisent pas.

Revenons à l'échantillon de données "mcycle". Comme nous pouvons le voir sur le graphique (4.16), la distribution des données a une structure particulière (peut-être une mixture de lois normales). En effet, bien qu'elle semble dissymétrique à droite, on remarque la présence d'un certain nombre de valeurs inférieures à 100. Cette distribution ne ressemblant a priori à aucune distribution connue, nous avons finalement [2] opté pour :

- une loi normale comme distribution conditionnelle pour la variable dépendante

$$f(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\}$$

  où $-\infty < y < +\infty$, $-\infty < \mu < +\infty$, $\sigma > 0$ avec $E[Y] = \mu$ et $V[Y] = \sigma^2$.

- et pour les fonctions de lien

$$g_1(\mu) = h_1(x)$$
$$g_2(\sigma) = h_2(x)$$

  $g_k(.)$ étant la fonction de lien et $h_k(x)$ une fonction non paramétrique (spline cubique) de la covariable x, le temps. Plus précisément, $h_1(.)$ et $h_2(.)$ sont des fonctions splines

1. Nous pourrions évidemment considérer d'autres distributions conditionnelles.
2. Bien qu'un grand nombre de distributions soit disponible dans le package "gamlss", soit elles ne sont pas utilisables dans ce cas (les données pouvant prendre des valeurs aussi bien négatives que positives), soit elles fournissaient des résultats "moins bons" (AIC plus élevé, résidus plus élevés en valeur absolue...).

de degrés de liberté 17.875 et 4.039. Ces modélisations semblaient les plus optimales en se basant sur le critère de l'AIC (AIC égal dans ce cas à 1122.345).

Nous pouvons observer les courbes de régression sur le graphique (4.17) pour des valeurs de p de 0.25, 0.5 et 0.75 où, omme attendu, ces courbes se croisent pas, ainsi que les estimations du paramètre de localisation et d'échelle pour des valeurs fixées de la covariable sur le graphique (4.18).

Enfin, dans le tableau (4.6) sont repris quelques renseignements concernant les résidus du modèle. On remarque que les résidus sont en moyenne pratiquement nuls avec une variance de 1. On notera également la légère dissymétrie à droite et un coefficient de kurtosis quasiment égale à celui d'une loi normale.

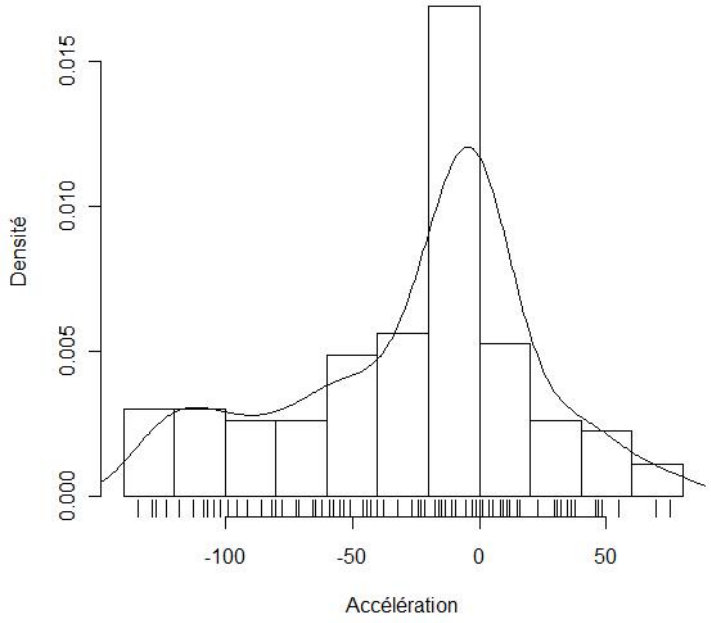

Figure 4.16 – Distribution de la variable "Accélération".

Moyenne	-0.005
Variance	1.008
Coef. d'asymétrie	0.134
Coef. de kurtosis	2.938

TABLE 4.6 – Résumé sur la distribution des résidus.

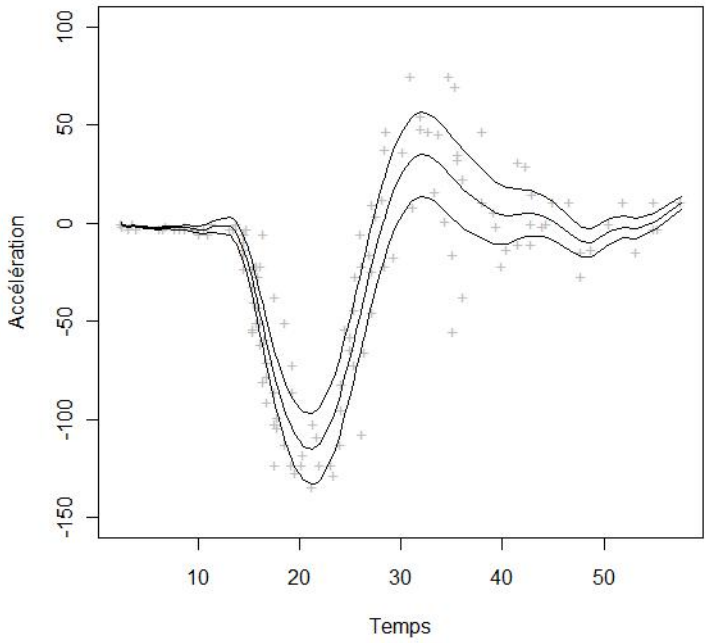

Figure 4.17 – Régression quantile pour des valeurs de p égales à 0.25, 0.5 et 0.95.

4.4.3 Méthodes bayesiennes

Qu'en est-il à présent des méthodes bayesiennes présentées dans ce mémoire ? Sont-elles applicables vu la structure particulière des données ?
La réponse est malheureusement non. En effet, de même que nous ne pouvions appliquer une méthode fréquentiste paramètrique vu cette structure des données, nous ne pourrons pas appliquer la méthode proposée par Yu et al ni même celle de E.G. Tsionas telles qu'elles sont présentées dans le chapitre 3. Notons cependant que ces méthodes pourraient être exploitées en utilisant les B-splines comme covariables mais pour ce faire, il faudrait

les généraliser afin notamment d'inclure un a priori pour modéliser la pénalité. On pourrait également avoir recours aux méthodes bayesiennes non paramétriques.

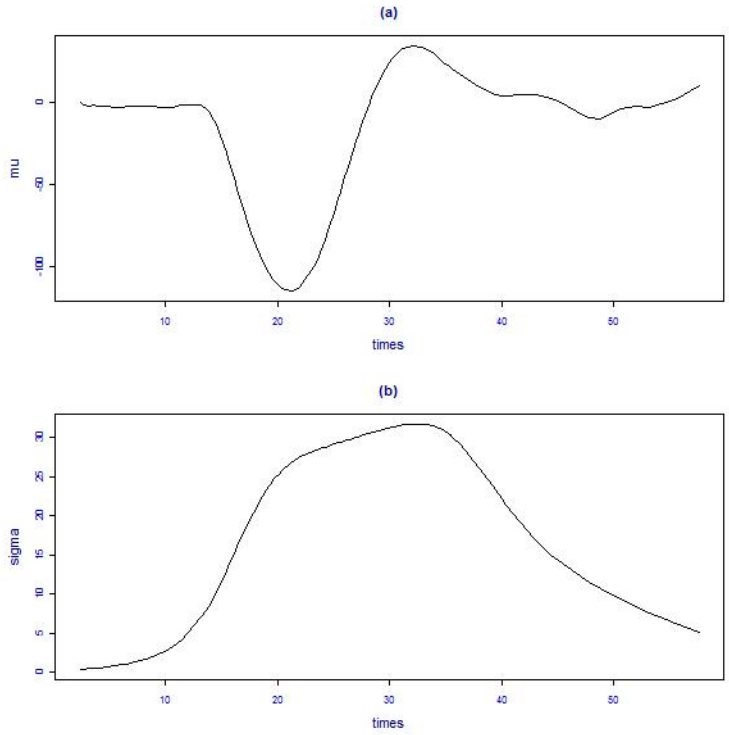

Figure 4.18 – Modélisation des différents paramètres.

Conclusion

Nous avons présenté dans ce mémoire diverses approches de la régression quantile, qu'elles soient fréquentistes ou bayesiennes que nous avons, pour un certain nombre d'entres elles, appliquées sur divers échantillons de données reprenant différents cas de figure (une ou plusieurs covariables, relation linéaire ou non entre variable dépendante et covariable), et ce afin de conclure aux avantages et inconvénients de ces méthodes selon le cas de figure envisagé.

On en conclut qu'en fréquentiste, là où la relation unissant la variable dépendante avec la covariable semblait être linéaire, la méthode paramétrique de Koenker pouvait être appliquée sans difficultés. Par contre, dans le cas contraire, nous devions avoir recours à des techniques non paramétriques (utilisation de B-splines, approche localement polynômiale) avec les problèmes que posent de telles techniques (choix adéquat du nombre de noeuds lors de l'utilisation de B-splines, de la pénalité, ou encore du *bandwith* si l'on opte pour l'approche par polynômes locaux).

Concernant les méthodes bayesiennes dites paramétriques, nous avons mentionné que les méthode de Yu et al (2001) et Tsionas (2003) telles qu'elles sont présentées dans ce mémoire ne sont d'application que lorsque la relation entre variable dépendante et covariable semble être linéaire, et dans ce cas (échantillon "Engel data"), nous avons obtenu des résultats proches de ceux donnés par la méthode paramétrique de Koenker avec les avantages de l'approche bayesienne (distribution des paramètres de régression, distribution de la prédiction et ainsi en déduire moyenne, intervalle de crédibilité,...).

Néanmoins, toutes les techniques de la régression quantile que nous venons de citer présentent le même désavantage qui est de devoir résoudre autant de problèmes de minimisation (en fréquentiste) ou de réappliquer l'algorithme de Metropolis (1953) ou de Gibbs (1984)(en bayesien) autant de fois que de quantiles conditionnelles d'ordre p souhaités, pour une valeur à chaque fois différente et fixée de p, ce qui peut s'avérer fastidieux. De plus, bien que théoriquement les droites de régressions ne devraient pas se croiser, il se pourrait que leurs estimations se croisent ce qui semble contraire à toute logique.

Pour remédier à ce soucis, une autre méthode fut présentée, méthode basée sur les

modèles additifs généralisés, qui permet une modélisation de la moyenne mais aussi du paramètre d'échelle, d'asymétrie et de kurtosis de la distribution conditionnelle de la variable dépendante. Les modélisations ne doivent pas être obligatoirement linéaires, cette technique permettant des modélisations non paramétriques. Malheureusement, la recherche des paramètres optimaux de ces fonctions non paramétriques peut s'avérer être une étape longue et fastidieuses. Néanmoins, comme nous l'avons déjà expliqué dans ce mémoire, par cette méthodes, les "droites" de régression ne se croiseront pas, ce qui respecte une certaine logique. De plus, cette approche permet de modéliser un grand nombre d'échantillons de données bien que nous n'ayons pu exploiter complétement cette méthode dans l'exemple "mcycle", la distribution des données ayant une forme très particulière.

Comme nous venons de le signaler ci-dessus, les méthodes de Yu et al (2001) et Tsionas (2003) telles qu'elles sont présentées ici ne sont d'application que lorsque la variable dépendante semble dépendre linéairement de la covariable. On pourrait toutefois généraliser ces méthodes pour qu'elles soient d'application lorsque cette relation n'est pas linéaire en utilisant des B-splines comme covariables et en incluant un a priori pour modéliser la pénalité. On pourrait également appliquer une méthode bayesienne non paramétrique [19] Notons enfin que ces deux approches sont également applicables lorsque plus d'une covariable est présente.

Il est important de mentionner que le fil conducteur des méthodes bayesiennes présentées ici était de supposer une distribution asymétrique de Laplace de paramètre de localisation μ nul, de paramètre d'échelle σ et de paramètre d'asymétrie p fixé selon l'ordre du quantile conditionnel souhaité, cette distribution vérifiant l'hypothèse faite sur la distribution des résidus du modèle, à savoir distribution dont le quantile d'ordre p est nul. Nous pourrions peut être envisager d'autres distributions pour les erreurs pour autant que cette hypothèse soit vérifiée. Un incovénient de cette distribution asymétrique de Laplace est qu'elle présente une forme particulière avec le rôle double joué par p (ordre du quantile conditionnel souhaité et paramètre d'asymétrie). Pour plus de flexibilité dans la distribution des données, nous avons présenté succintement une méthode bayesienne non paramétrique basée sur un mélange de processus de Dirichlet dans le cadre de la régression médiane, méthode généralisée pour d'autres quantiles et qui est toujours à l'heure actuelle en voie de développement. Nous n'avons pas appliqué cette méthode sur un ensemble de donnée, celle-ci étant plus technique mais un exemple est disponible dans le chapitre concerné.
Remarquons que, bien que cette méthode permet plus de flexibilité dans la modélisation des données, elle n'est pas utilisable dans le cas où la distribution des données est bimodale.

Enfin, on a pu constater, au vu des nombreuses références citées dans ce mémoire, que la théorie des quantiles de régression est vaste, et est toujours en développement en particulier en statistique fréquentiste non paramétrique (nombreux articles de Yu à ce sujet) et en statistique bayesienne (notamment les travaux de Kottas dans le domaine du non paramètrique).

Appendices

Annexe A

Distribution t de Box-Cox

Soit Y une variable aléatoire positive et distribuée selon une loi t de Box-Cox, ce qui se note $Y \sim BCT(\mu, \sigma, \nu, \tau)$, et soit une variable aléatoire Z distribuée selon une distribution de Student à τ ($\tau > 0$) degré de liberté.

La fonction de densité de Y est dès lors

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{y^{\nu-1} f_T(z)}{\mu^\nu \sigma F_T(\frac{1}{\sigma|\nu|})}$$

pour $y > 0$, $\mu > 0$, $\sigma > 0$ et $-\infty < \nu < +\infty$; et où $f_T(t)$ et $F_T(t)$ sont respectivement la fonction de densité et de répartition d'une variable aléatoire T distribuée selon une loi de Student à τ degrés de liberté.

Annexe B

Implémentation de la méthode proposée par Yu et al

Dans les deux sections qui vont suivre vont être présentées deux implémentations de l'algorithme Metropolis (1953).

La première implémentation consiste en une mise à jour de tous les paramètres simultanément tandis que dans la seconde, la mise à jour se fait paramètre par paramètre.

B.1 Première implémentation

Pour rappel, la distribution a posteriori de β est

$$p(\boldsymbol{\beta}|\mathbf{y}) \propto p^n(1-p)^n \exp\{-\sum_{i=1}^{n} \rho_p(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\}$$

où

$$\begin{aligned} \rho_p(x) &= x(p - I(x<0)) \\ &= \tfrac{|x|+(2p-1)x}{2} \end{aligned}$$

La fonction $\rho_p(x)$ est calculable via la fonction suivante, fonction ne dépendant que de p et de x.

```
rho<-function(p,x){
(abs(x)+(2*p-1)*x)/2}
```

Ensuite, le logarithme de la vraisemblance est obtenu via la fonction suivante, fonction ne dépendant à nouveau que de p, du vecteur des paramètres de régression (B), de la matrice des variables explicatives (X), et du vecteur d'observations (y).

```
loglikelihood<-function(p,B,X,y){
n=length(y)
w=y-X%*%B
n*log(p)+n*log(1-p)+(-sum(rho(p,w)))
}
```

Comme nous avons opté pour une distribution a priori impropre uniforme, nous obtenons directement le logarithme de la distribution a posteriori.

```
logposterior<-function (p,B,X,y){
loglikelihood(p,B,X,y)}
```

Une fois que nous avons établi ces trois premières fonctions, nous pouvons commencer la procédure décrite au chapitre 3, section 3.3.2.

Décrivons à présent les étapes principales de cette procédure en les référrant aux commandes de l'algorithme implémenté, disponible dans son entièreté page 100, par un dièse numéroté.

Nous créons donc une fonction MH (#1) qui dépend
- du choix initiaux des paramètres de régression (*beta.init*),
- de p qui est fixé,
- d'une matrice de variance-covariance entre les paramètres de régression (*sigma*)
- d'un paramètre *sd* qui nous permettra d'atteindre un taux d'acceptation de 0.20,
- du nombre d'itérations effectuées (M),
- de la matrice des variables explicatives (X),
- du vecteur d'observations (y),
- de la distribution a posteriori des paramètres de régression (*logposterior*).

Avant de commencer l'algorithme Metropolis (1953) proprement dit, nous devons initialiser une matrice (avec un nombre de lignes égal à M et un nombre de colonnes égal au nombre de paramètres de régression à estimer) qui reprendra tous les paramètres de regressions générés (#2) et dont la première ligne sera constituée des choix initiaux de ces paramètres.
Nous initialisons également le taux d'acceptation à zéro (#3) ainsi qu'un vecteur reprenant les densités a posteriori calculées en les paramètres de régression générés (#4), et ce afin de ne pas recalculer plusieurs fois ces valeurs. A nouveau, la première composante de ce

vecteur est la densité calculée en les paramètres initialement choisis (#5).

Ensuite, on génère les paramètres de régression simultanément selon une loi multinormale de matrice de variance-covariance égale à $sd^2 * sigma$ (#6) et on cacule la densité a posteriori pour ces paramètres (#7). On reprend ensuite la densité pour les valeurs des paramètres acceptés à l'étape précédente (#8) afin de calculer la probabilité d'acceptation (#9) et ce dans le but de déterminer si il y a rejet ou acceptation des paramètres de régression ainsi générés (#10).
Si il y a acceptation, on stoke la densité calculée en les paramètres acceptés ainsi que ces paramètres (#11). Sinon, les paramètres stockés sont ceux de l'étape précédente et donc on garde alors la densité a posteriori calculées en ces paramètres. (#12)

Enfin, on demande de rendre, dans une liste, les paramètres de régression acceptés et le taux d'acceptation (#13).

Comme il fut déjà mentionné lorsque nous avons rappelé le principe de l'algorithme de Metropolis (1953), il est conseillé de viser un taux d'acceptation de 0.20. Le paramètre sd nous permettra d'atteindre ce taux comme nous l'avons déjà expliqué dans la section concernée au rappel de cette méthode MCMC.

La fonction MH :

```
MH<-function(beta.init,p,sigma,sd,M,X,y,logposterior){#1
beta=matrix(nrow=M,ncol=length(beta.init))#2
beta[1,]=beta.init
n.accept<-0#3
lpost=NULL#4
lpost[1]=logposterior(p,beta[1,],X,y)#5
for(i in 2:M){
beta.prop<-beta[i-1,]+rmnorm(1,rep(0,length(beta.init)),sd^2*sigma)#6
lprop<-logposterior(p,beta.prop[1,],X,y)#7
lcurr<-lpost[i-1]#8
prob=min(1,exp(lprop-lcurr))#9
accept=(runif(1)<=prob)#10
if (accept)#11
{
n.accept=n.accept+1
beta[i,]=beta.prop
lpost<-c(lpost,lprop)
}
else {beta[i,]=beta[i-1,]#12
lpost<-c(lpost,lcurr)
}
}
return(list(beta=beta,taux=n.accept/(M-1)))#13
}
```

B.2 Seconde implémentation

La seconde implémentation diffère de la première dans le sens, qu'à présent, les paramètres sont mis à jour les uns après les autres.

Dès lors, nous ne devrons plus fournir de matrice de variance-covariance mais une variance pour chaque paramètre.

On crée donc une fonction $MHbis$ (#1) qui dépend à présent
- du choix initiaux des paramètres, comme précédemment (*beta.init*),
- d'un p fixé,
- d'un vecteur *sd* qui permettra de viser un taux d'acceptation de 0.40 par paramètre (ici, *sd* est un vecteur dont le nombre de composantes est égal au nombre de paramètres à estimer),
- d'un vecteur *sigma* reprenant les variances pour chaque paramètre,
- du nombre d'itérations voulues (M),
- de la matrice des variables explicatives (X),
- du vecteur d'observations (y),
- de la distribution a posteriori des paramètres de régression (*logposterior*).

Remarquons que la distribution a posteriori étant la même que lors de la première implémentation, nous ne la redéfinissons pas dans cette section.

Comme lors de la première implémentation, nous devons passer par une phase d'initialisation avant d'entamer la procédure en tant que telle (#2). On peut remarquer que la seule différence avec la première implémentation est l'initialisation du taux d'acceptation. En effet, avant, nous recherchions un taux global de 0.20 tandis qu'ici, nous visons un taux de 0.40 par paramètre. Dès lors, *n.accept* est un vecteur donc chaque composante est initialisée à zéro.

Après cette phase d'initialisation, on peut commencer la procédure proprement dite.
A l'étape i, le vecteur beta.prop est constitué des paramètres stockés à l'étape précédente (#3).
Ensuite, pour le paramètre j, on génère une valeur et on en calcule la probabilité a posteriori afin de déterminer si la valeur générée sera rejetée ou non (#4). Si on l'accepte, le taux d'acceptation pour le paramètre j augmente de un et on stocke ce paramètre dans la matrice (#5). De même, on stocke sa probabilité a posteriori.
Si on rejette, on stocke le paramètre j de l'étape précédente de même que la probabilité a posteriori qui reste dès lors inchangée (#6). On garde ensuite dans beta.prop[j], la valeur acceptée pour le paramètre j (#7).
Enfin, on demande de rendre tous les quantiles de régression stockés et les taux d'acceptations pour chaque paramètre (#8).

La fonction $MHbis$:

```
MHbis<-function(beta.init,p,sd,sigma,M,X,y,logposterior){#1
nb=length(beta.init)
beta=matrix(nrow=M,ncol=length(beta.init))#2
beta[1,]=beta.init
n.accept<-rep(0,nb)
lcurr<-logposterior(p,beta[1,],X,y)
for(i in 2:M){#3
beta.prop<-beta[i-1,]
for(j in 1:nb){#4
beta.prop[j]<-beta[i-1,j]+rnorm(1,0,sd[j]*sigma[j])
lprop<-logposterior(p,beta.prop,X,y)
prob=min(1,exp(lprop-lcurr))
accept=(runif(1)<=prob)
if (accept)#5
{
n.accept[j]=n.accept[j]+1
beta[i,j]=beta.prop[j]
lcurr<-lprop
}
else {#6
beta[i,j]=beta[i-1,j]
lcurr<-lcurr
}
beta.prop[j]<-beta[i,j]#7
}
}
return(list(beta=beta,taux=n.accept/(M-1)))#8
}
```

Bibliographie

[1] William M. Bolstad. *Introduction To Bayesian Statistics*. Wiley, second edition, 2007.

[2] R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1) :1–13, 1946.

[3] Paul H. C Eilers and Brian D. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2) :89–121, 1996.

[4] Ali Gannoun, Jérôme Saracco, and Keming Yu. Comparison of kernel estimators of conditional distribution function and quantile regression under censoring. *Statistical Modelling*, 7(4) :329–344, 2007.

[5] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, second edition, 2004.

[6] Paul Gérard. *Modèles linéaires*, Année académique 2007-2008. Notes de cours.

[7] Gentiane Haesbroeck. *Probabilité et Statistique I*, Année académique 2005-2006. Notes de cours.

[8] Gentiane Haesbroeck. *Probabilité et Statistique II*, Année académique 2006-2007. Notes de cours.

[9] Gaëlle Hoffait. Quantiles de régression. Master's thesis, Université de Liège, Faculté des sciences, Département de Mathématique, Année académique 2003-2004.

[10] Joseph G. Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian Survival Analysis*. Springer, 2001.

[11] M. C. Jones and Keming Yu. Improved double kernel local linear quantile regression. *Statistical Modeling*, 7(4) :377–389, 2007.

[12] Roger Koenker. *Quantile Regression.* Cambridge University Press, 2005.

[13] Athanasios Kottas and Alan E. Gelfand. Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, 96(456) :1458–1468, December 2001.

[14] Athanasios Kottas and Milovan Krnjajic. Bayesian semiparametric modeling in quantile regression. *Scandinavian Journal of Statistics*, 36(2) :297–319, June 2009.

[15] Philippe Lambert. *Introduction to Bayesian analysis*, Année académique 2007-2008. Notes de cours.

[16] Tony Lancaster and Sung Jae Jun. Bayesian quantile regression. Technical report, 2008.

[17] R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and schape. *Appl. Statist.*, 54(3) :507–554, 2005.

[18] Jean Schmets. *Analyse mathématique : Introduction au Calcul Intégral. Notes de cours de la première candidature en sciences mathématiques et physiques.* Editions Derouaux, Année académique 1993-1994.

[19] Matthew Taddy and Athanasios Kottas. A bayesian nonparametric approach to inference for quantile regression. Technical report, 2007.

[20] Volker Tresp. *Dirichlet Process and Nonparametric Bayesian Modeling*, 2008. Notes de cours.

[21] Efthymios G. Tsionas. Bayesian quantile inference. *Journal of Statistical Computation and Simulation*, 73(9) :659–674, 2003.

[22] Keming Yu. Smoothing regression quantile by combining k-nn estimation with local linear kernel fitting. *Statistica Sinica*, 9 :759–774, 1999.

[23] Keming Yu. Quantile regression using rjmcmc algorithm. *Computational Statistics & Data Analysis*, 40(2) :303–315, 2002.

[24] Keming Yu and Stander Julian. Bayesian analysis of a tobit quantile regression model. *Journal of Econometrics*, 137(1) :260–276, 2007.

[25] Keming Yu, P. Van Kerm, and J. Zhang. Bayesian quantile regression : An application to the wage distribution in 1990s britain. *The Indian Journal of Statistics*, 67(2) :359–377, 2005.

[26] Keming Yu and Zudi Lu. Local linear additive quantile regression. *Scandinavian Journal of Statistics*, 31(3) :333–346, 2004.

[27] Keming Yu, Zudi Lu, and Julian Stander. Quantile regression : applications and current research areas. *Statistician*, 52(3) :331–350, 2003.

[28] Keming Yu and Rana A. Moyeed. Bayesian quantile regression. *Statistics and Probability Letters*, 54 :437–447, 2001.

[29] Keming Yu and Jin Zhang. A three-parameter asymmetric laplace distribution and its extension. *Communication in Statistics-Theory and Methods*, 34 :1867–1879, 2005.

[30] Kerming Yu and M. C. Jones. Local linear quantile regression. *Journal of the American Statistical Association*, 93(441) :228–237, 1998.