université de BORDEAUX

L3 Mathématiques et Informatique appliquées aux Sciences Humaines et Sociales

Travaux d'Étude et de Recherche

La régression quantile comme alternative à la régression linéaire classique

Élèves :
Marie Meyer
Alexandra Millot

Enseignant :
Antoine Barbieri

Année académique 2022-2023

Remerciements

Avant tout développement, il apparaît judicieux de commencer ce rapport de TER par des remerciements ; à ceux qui nous ont beaucoup appris.

Tout d'abord, nous souhaitons exprimer notre sincère gratitude envers notre encadrant, Antoine Barbieri, pour avoir consacré de son temps à nous offrir une aide précieuse et à nous transmettre ses connaissances tout au long de la réalisation de ce projet d'étude.

Nous tenons également à remercier les responsables de la Licence d'avoir mis en place les TER lors de ce dernier semestre, ainsi que les responsables de TER notamment Bedr'Eddine Ainseba et à nouveau Antoine Barbieri.

Enfin, nous souhaitons faire part de notre gratitude à tous les intervenants de la Licence MIASHS qui nous ont accompagnés tout au long de ces trois années. Ils ont non seulement enrichi nos connaissances, mais ont également contribué à développer notre raisonnement scientifique et notre professionnalisme. Nous leur sommes reconnaissantes pour leur engagement envers notre réussite.

Table des matières

Introduction

Les modèles de régression ont pour objectif d'expliquer et de quantifier la variabilité d'une variable aléatoire d'intérêt, appelée variable réponse (Y), au travers des variables explicatives (X). On distingue différents modèles de régression dont celles qui nous intéressent : la régression linéaire et la régression quantile.

La régression linéaire introduit par Quetelet [1835], est aujourd'hui l'un des outils statistiques les plus couramment utilisés pour modéliser l'espérance conditionnelle de la variable réponse en fonction d'une ou plusieurs variables explicatives. Cependant, elle présente des limites, notamment lorsqu'il s'agit de modéliser des distributions de données asymétriques ou des données avec des valeurs extrêmes.

Dans ce TER nous allons nous poser la question : Dans quelle mesure la régression quantile peut-elle être considérée comme une alternative à la régression linéaire classique ?

Nous allons explorer une alternative à la régression linéaire classique : la régression quantile introduit par Koenker and Bassett [1978]. La régression quantile, tout comme la régression linéaire, ~~elle~~ modélise la relation entre une variable réponse et une ou plusieurs variables explicatives. Sa particularité première est qu'elle modélise non pas l'espérence conditionnelle conditonnelle mais les quantiles condtionnelles de la variable réponse en fonction des variables explicatives.

Ainsi, nous verrons comment elle permet de modéliser la relation entre les variables avec plus de flexibilité, surtout dans les situations où la régression linéaire ne serait pas adaptée. Pour cela, nous commencerons par aborder l'approche théorique de la régression quantile. En débutant, par un retour sur la régression linéaire classique avant d'aborder la régression quantile. Nous finirons cette approche par l'analyse des avantages et des inconvénients de la régression quantile par rapport à la régression linéaire classique. Enfin, dans une deuxième partie, nous étudierons une application de la régression quantile sur la pluviométrie à Bordeaux, afin de déterminer les conditions propices aux fortes précipitations pouvant causer des inondation.

1 Présentation de la régression quantile

1.1 Généralité sur les modèles de régression

Soit la relation entre la variable à expliquer (noté Y) et la variable explicative (noté X) est supposée linéaire. On parle alors, d'une régression linéaire pour les n observations de la forme :

$$Y = \beta_0 + x_i\beta_1 + \varepsilon_i, \qquad i = 1, 2, ..., n \tag{1}$$

où,

- Y est la variable aléatoire aléatoire réponse de Y associée à l'individu i ;
- x_i représente la valeur prise par la variable X associée à l'individu i, les x_i sont supposés non-aléatoires ;
- β_0 est le coefficient de régression appelé constante ou intercept ;
- β_1 est le coefficient de régression appelé pente ;
- ε_i sont les termes d'erreurs du modèle. Plus généralement, ε est une variable aléatoire indépendante non obsevée de X.

Si on utilise p variables explicatives pour expliquer la variable dépendante Y, on peut utiliser le modèle linéaire multiple sous la forme :

$$y = \beta_0 \sum_{j=1}^{p} x_j\beta_j + \varepsilon$$

On peut écrire ce modèle sous la forme matricielle suivante :

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

où,

- Y désigne le vecteur à expliquer de dimension n, tel que $y = (y_1, ..., y_n)^\mathsf{T}$;
- X est la matrice explicative de taille $n \times (p + 1)$;
- β est le vecteur de dimension $p + 1$ des paramètres à estimer du modèle ;
- ε est le vecteur aléatoire de dimension n, appelé erreur du modèle.

La notion de linéaire fait référence au fait que ce sont des modèles linéaires en leur paramètres.

Précisons les vecteurs et matrice du modèle linéaire :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1p} \\ 1 & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

1.2 Retour sur la régression linéaire

La régression linéaire est une méthode d'analyse statistique qui vise à modéliser l'espérance conditionnelle (le comportement moyen) de la variable réponse en fonction des variables explicatives.

La régression linéaire classique s'intéresse à expliquer le comportement moyen d'un phénomène (Y) en fonction des variables explicatives. Soit le modèle général donné en (1) est précisé tel que :

$$\mathbb{E}(Y|X) = X\beta$$

où $\mathbb{E}(Y|X)$ est "espérance conditionnelle de Y sachant X".

On distingue trois hypothèses fondamentales sur les erreurs :

i) On suppose que les erreurs ε_i, $i = 1, ..., n$ sont centrées : $\mathbb{E}(\varepsilon_i) = 0$;
ii) On suppose que les erreurs ε_i, $i = 1, ..., n$ sont décorrélées : $Corr(\varepsilon_i, \varepsilon_j) = 0$ avec $i \neq j$;
iii) On suppose que les erreurs ε_i, $i = 1, ..., n$ sont de variance constante : $Var(\varepsilon_i) = \sigma^2$;

Remarque. ε_i est supposée suivre une loi Normale de moyenne 0 et d'écart-type σ, tel que $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Nous pouvons représenter graphiquement les termes d'erreurs du modèle, comme le montre la Figure 1.

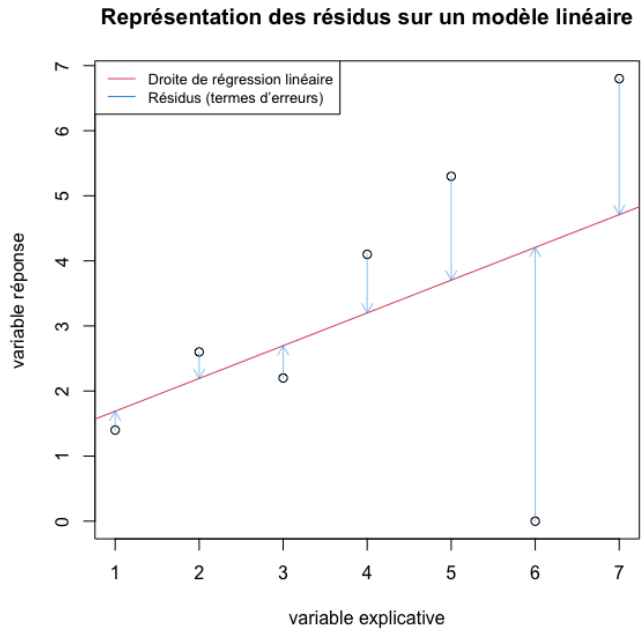

FIGURE 1 – Représentation des résidus

1.2.1 Estimation du modèle

La méthode la plus connue pour estimer le modèle (1) est la méthode des moindres carrés (voir annexe). Elle est basée sur la fonction de coût quadratique suivante :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \underbrace{(y_i - x_i^{\mathsf{T}}\beta)}_{\varepsilon_i^2}{}^2 \tag{2}$$

En régression linéaire classique, la fonction de coût quadratique est généralement utilisée pour évaluer la performance du modèle et ajuster les coefficients de régression. Elle mesure l'erreur entre les valeurs réelles observées et les prévisions du modèle, en prenant la somme des carrés des différences entre la valeur réelle et la valeur prédite de la variable réponse Y.

Remarque. L'objectif de la régression linéaire classique est de minimiser cette fonction de coût en trouvant les coefficients de régression β qui réduisent au maximum l'erreur entre les prévisions et les valeurs réelles.

1.2.2 Inférence statistique

L'inférence fait référence à la manière dont les paramètres d'un modèle de régression sont estimés à partir des données, ainsi qu'à la manière dont les prévisions sont formulées à partir de ce modèle.

Afin de vérifier la validité du modèle et de s'assurer que les résultats obtenus sont fiables et interprétables, on doit pouvoir démontrer que trois hypothèses fondamentales sur les erreurs (vu en page 5) sont vérifiées.

Des tests peuvent être effectués à cet effet. Cependant, dans le cadre de ce TER, nous nous limiterons à les citer sans rentrer dans les détails :

- Test de normalité : le test de Shapiro-Wilk ou le test de Kolmogorov-Smirnov peut être utilisé pour vérifier si les erreurs suivent une distribution normale ;
- Test de linéarité : le test de Durbin-Watson peut être utilisé pour vérifier si les erreurs sont corrélées, ce qui peut indiquer une non-linéarité de la relation entre les variables ;
- Test d'homoscédasticité : le test de Breusch-Pagan ou le test de White peut être utilisé pour vérifier si la variance des erreurs est constante ;
- Test d'indépendance : le test de Durbin-Watson peut également être utilisé pour vérifier l'indépendance des erreurs.

On peut alors supposer que dans la suite de variables aléatoires, les ε_i en plus d'être indépendants, de même loi, centrées et de même variance, sont distribuées suivant une loi $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Cela implique que la variable Y suit également une distribution normale $Y \sim \mathcal{N}(X\beta, \sigma^2)$, car la somme de variables aléatoires normales est elle-même normale.

Estimateurs du maximum de vraisemblance

Sous l'hypothèse de normalité, on est en mesure de définir la vraisemblance du modèle et de l' pour estimer les paramètres et faire de l'inférence statistique L'estimation par maximum de vraisemblance à $\theta = (\beta, \sigma)^\mathsf{T}$ est basée sur la vraisemblance du modèle linéaire gaussien :

$$\begin{aligned} \mathcal{L}(\theta; Y) &= \prod_{i=1}^{n} f_Y(y_i; \theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\mathsf{T}\beta)^2\right\} \qquad (3) \\ \Leftrightarrow \ln \mathcal{L}(\theta; Y) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\mathsf{T}\beta)^2 \end{aligned}$$

où, $f_Y(y_i; \theta)$ est la densité de la loi Normale où $(Y_i | X_i = x_i) \sim \mathcal{N}(x_i^\top \beta; \sigma^2)$.

On cherche les estimateurs $\hat{\beta}$ et $\hat{\sigma}$ qui maximisent la log vraissemblance.

Remarque. Maximiser en β la vraisemblance définie en (3) revient au problème d'optimisation présenté en (2). Les estimateurs du maximum de vraisemblance sont donc égaux aux estimateurs des moindres carrées ordinaires.

1.3 La régression quantile

La régression quantile est une méthode d'analyse statistique qui vise à modéliser les quantiles conditionnelles de la variable réponse en fonction des variables explicatives.

1.3.1 Les quantiles et quantile conditionnelle

En statistique, les quantiles sont les valeurs qui divisent une distribution en intervalles de probabilité égaux.

Certains quantiles portent des noms spécifiques. En considérant $\tau \in]0;1[$, l'ordre du quantile, nous retrouvons, le premier quartile d'ordre $\tau = 0.25$. Le second quartile est le plus couramment utilisé, plus connu sous le nom de médiane, son ordre est de $\tau = 0.5$. Ainsi, celui de niveau $\tau = 0.75$ est appelé le troisième quartile. Les quantiles dont les niveaux sont des multiples de un dixième (resp. un centième) sont appelés déciles (resp. centiles). La Figure 2 illustre ces trois exemples.

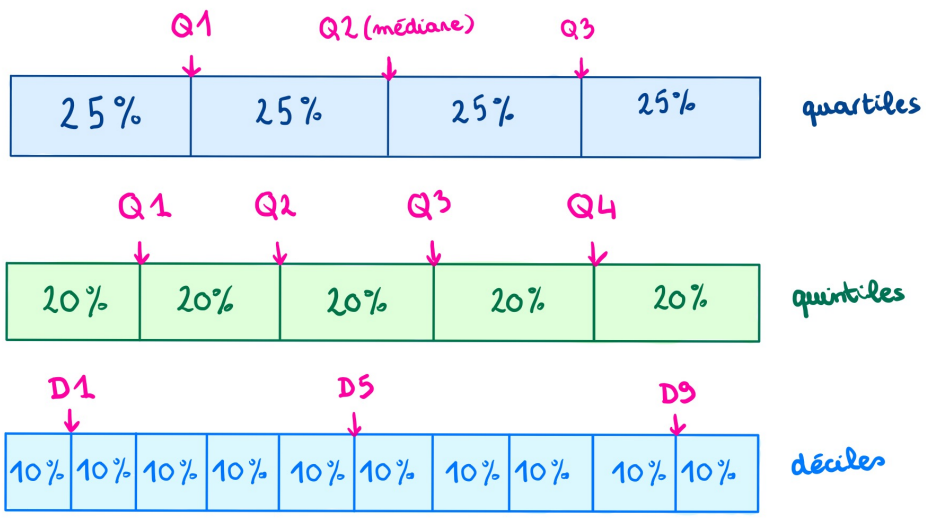

FIGURE 2 – Représentation des quantiles : la première ligne représente les quartiles où la population est divisée en 4 parties égales ; la deuxième ligne représente les quintiles où la population est divisée en 5 parties égales ; la troisième ligne représente les déciles où la population est divisée en 10 parties égales

Soit Y une variable aléatoire (discrète ou continue) qui a pour fonction de répartition F_Y déterminée par $F_Y(y) = \mathcal{P}(Y \leq y)$. Pour tout $\tau \in]0;1[$ et $F_Y(y)$ continue et strictement croissante, le quantile d'ordre τ de Y est défini par :

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

Les quantiles expriment les seuils en-dessous desquels se trouvent $(\tau \times 100)$ % de la donnée.

Si, d'autre part, Y a une fonction de distribution de probabilité conditionnelle à X, notée par $F_{Y|X}(y)$. On définit le quantile conditionnel d'ordre τ de la variable Y sachant X = x :

$$Q_{Y|X}(\tau) = F_{Y|X}^{-1}(y) = \inf\{y : F_{Y|X}(y) \geq \tau\}$$

Remarque. Les régressions quantiles tentent d'évaluer comment les quantiles conditionnels de Y varie lorsque X varie.

1.3.2 Le modèle de régression quantile

Le modèle de régression quantile standard, liant une ou plusieurs variables explicatives avec la variable réponse a été présenté par Koenker and Bassett [1978]. Il suppose que ces quantiles de la distribution condition- nelle ont une forme linéaire :

$$Q_{y|x}(\tau) = x^{\mathsf{T}}\beta_\tau + \varepsilon$$

où,

- $Q_\tau(y|x)$ la τ-ème fonction de régression quantile de y sachant x.
- y désigne le vecteur à expliquer de dimension n, tel que $y = (y_1, ..., y_n)^{\mathsf{T}}$;
- x est la matrice explicative de taille $n \times (p + 1)$;
- β est le vecteur de dimension $p+1$ des paramètres à estimer du modèle au quantile d'ordre τ ;
- ε sont les termes d'erreurs du modèle.

Cette expression peut s'écrire de manière équivalente :

$$Y_i = x_i^{\mathsf{T}}\beta_\tau + \varepsilon_i, \qquad i = 1, ..., n \tag{4}$$

où ε_i est l'erreur sur l'individus i dont la distribution est restreinte pour avoir le τ-ème quantile égal à zéro, soit $\int_{-\infty}^{0} f_\tau(\varepsilon_i)d\varepsilon_i = \tau$. $f_\tau(\epsilon_i)$ correspond à la densité de probabilité de l'erreur ε_i pour le τ-ème quantile.

Remarque. Une différence importante réside dans le fait que les coefficients sont spécifiques à τ et susceptibles de varier d'un quantile à l'autre. Cela fournit une information supplémentaire qui ne peut être obtenue à partir d'une simple régression linéaire. En effet, l'intérêt peut porter sur l'ensemble de la distribution $Y|X$ contrairement à la régression classique qui se concentre essentiellement sur $\mathbb{E}(Y|X)$.

1.3.3 L'estimation des paramètres

Avant de procéder à l'estimation de β_τ, nous tenons à souligner qu'il existe deux approches pour estimer ce paramètre : la méthode paramétrique et la méthode non paramétrique. Nous nous intéressons ici à l'approche non paramétrique considérée comme une alternative aux problèmes posés par les hypothèses restrictives de la modélisation paramétrique. En effet, la méthode paramétrique part du principe que les données suivent une distribution connue, ce qui n'est souvent pas le cas en pratique.

L'estimation de β_τ s'écrit :

$$\hat{\beta}_\tau = \underset{\beta_\tau \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta_\tau) \tag{5}$$

où :

$$\rho_\tau(u) = (\tau - \mathbb{1}\{u < 0\})u \tag{6}$$

est la fonction de perte à minimiser.

Reprenons l'exemple de Givord and D'Haultfœuille [2013], pour $\tau = 0.5$, c'est-à-dire si l'on s'intéresse à la médiane, la fonction de perte revient à considérer le coût absolu.

correspond simplement à la (demi-) valeur absolue.

Dans la Figure 3, on peut voir la représentation graphique de quelques fonctions ρ_τ pour différentes valeurs de τ.

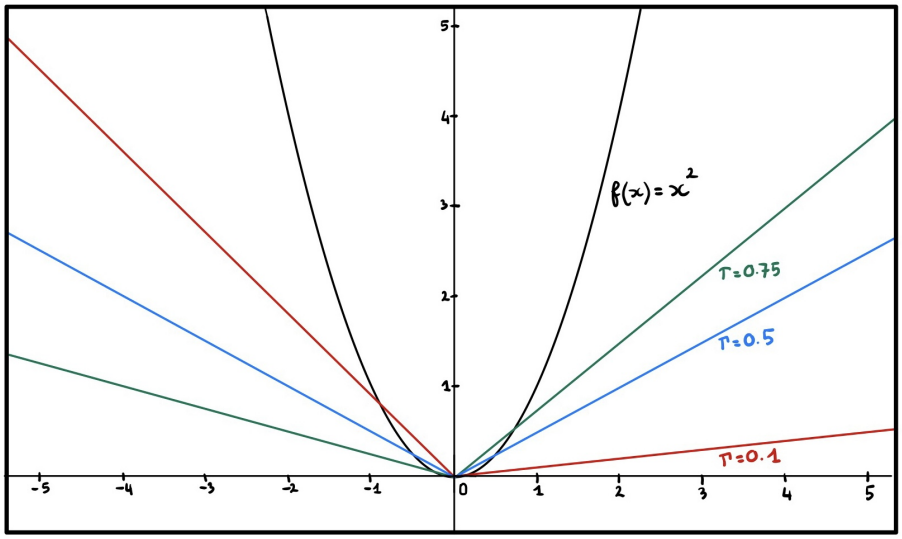

FIGURE 3 – Fonctions ρ_τ avec différentes valeurs de τ et fonction quadratique

1.3.4 Inférence statistique

Dans le cadre de l'inférence statistique, nous faisons une hypothèse de distribution sur la variable aléatoire étudiée. Contrairement à la régression linéaire classique, qui suppose des erreurs de régression indépendantes et identiquement distribuées, la régression quantile se base sur la distribution asymétrique de Laplace ($\mathcal{AL}$). Cette distribution vue dans les articles Koenker and

Machado [1999] et Yu and Moyeed [2001], permet de modéliser la fonction de vraisemblance utilisée pour estimer les paramètres du modèle de régression et pour produire des intervalles de confiance.

Nous disons qu'une variable aléatoire Y est distribuée sous la forme d'un distributeur $\mathcal{AL}$ avec des paramètres μ, σ et τ, et nous l'écrivons comme $Y \sim \mathcal{AL}(\mu, \sigma, \tau)$, si la densité de probabilité correspondante est donnée par :

$$f(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\}$$

où :

- τ est le paramètre d'asymétrie tel que $\tau \in]0; 1[$;
- σ est le paramètre d'échelle tel que $\sigma \in \mathbb{R}_*^+$;
- μ est le paramètre de locatisation tel que $\mu_i = x_i^\mathsf{T}\beta_\tau \in \mathbb{R}$.

Concrètement dans le modèle de régression quantile (4), ε_i correspondent à des variables aléatoires indépendantes et identiquement distribuées provenant de la distribution $\varepsilon_i \sim \mathcal{AL}(0, \sigma, \tau)$.

On a comme équivalence :

$$\begin{aligned} & Y_i \sim \mathcal{AL}(\mu, \sigma, \tau) \\ \Leftrightarrow \quad & \varepsilon_i \sim \mathcal{AL}(0, \sigma, \tau) \end{aligned}$$

Dans le contexte de la régression, l'estimation des paramètres est obtenue en maximisant la vraisemblance qui en découle de la densité de probabilité donnée dans l'équation (7). Pour n observations, nous avons alors :

$$\mathcal{L}(\theta; y) = \prod_{i=1}^{n} f(y_i|\mu_i, \sigma, \tau) = \left(\frac{\tau(1-\tau)}{\sigma}\right)^n \exp\left\{-\frac{1}{\sigma}\sum_{i=1}^{n}\rho_\tau(y_i - \mu_i)\right\} \tag{7}$$

Remarque. L'utilisation de la distribution $\mathcal{AL}$ en régression quantile vient du lien direct entre maximiser la vraisemblance de l'équation (7) et la minimisation de l'équation (5).

1.4 Avantages et inconvénients de la régression quantile par rapport à la régression linéaire classique

Cette dernière sous-partie a pour objectif de distinguer les avantages et les inconvénients de la régression quantile par rapport à la régression linéaire classique.

1.4.1 Résoudre le problème des valeurs extrêmes

Les valeurs extrêmes peuvent avoir un impact significatif sur les résultats d'une analyse de régression linéaire. Elles peuvent fausser les estimations des paramètres, augmenter la variabilité des estimations et affecter la validité des conclusions tirées de l'analyse. Il est donc important de savoir détecter ces valeurs afin de prendre en compte leurs impactes sur lors de l'interprétation des résultats.

En régression quantile, cet impact est moindre. En effet, la fonction de perte quadratique (2) utilisée dans la régression linéaire est remplacée par la fonction de perte quantile (6), qui s'accroît linéairement et non quadratiquement avec le résidu. En raison de cette caractéristique, les écarts très importants sont beaucoup moins pénalisés, ce qui explique la robustesse de la régression quantile face aux valeurs extrêmes ou aberrantes.

Nous rappelons que pour $\tau = 0.5$, la fonction ρ revient au coût absolu, ce qui revient à minimiser la somme des valeurs absolues des erreurs.

La Figure 4 montre l'impact considérable des valeurs extrêmes sur la moyenne par rapport à la médiane. La moyenne tend visuellement vers les valeurs extrêmes ce qui n'est pas le cas de la médiane.

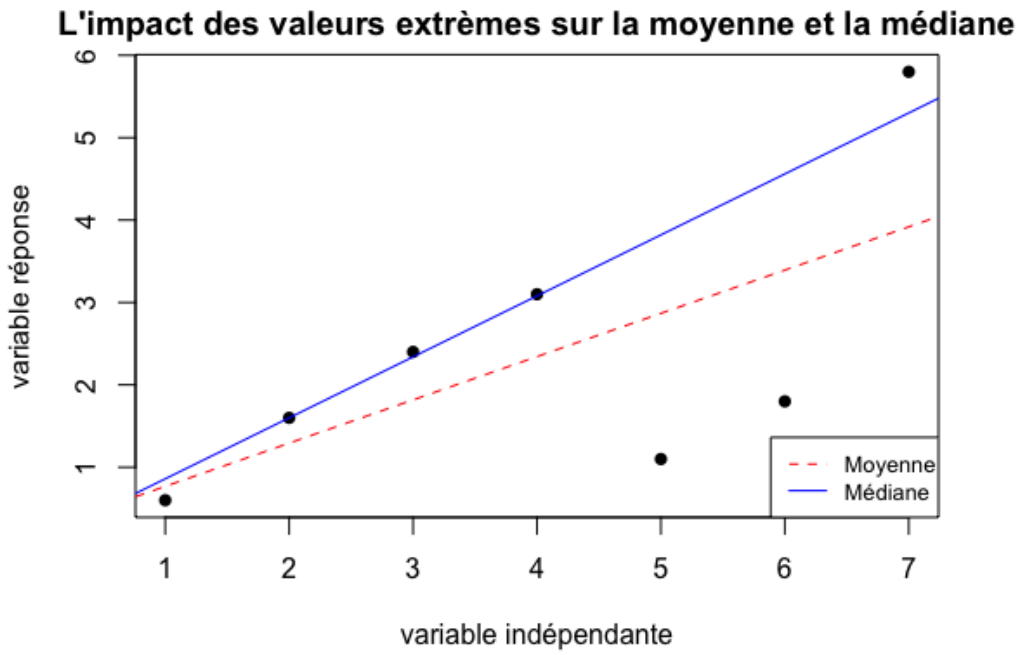

FIGURE 4 – Représentation graphique de l'impact des valeurs extrèmes sur la droite de régression linéaire (correspondant à la moyenne), comparé à l'impact sur la régression quantile d'ordre 0.5 (la médiane)

1.4.2 Résoudre l'hétéroscédasticité

En régression linéaire, la présence d'hétéroscédasticité (voir Figure 5) est une préoccupation majeure dans l'analyse de régression et l'analyse de la variance. L'héréoscédasticité est lorsque les variances des résidus des variables examinées sont différentes. D'après l'étude de Goldberger [1964], cette variabilité de la variance peut conduire à des estimations biaisées des paramètres du modèle et à des intervalles de confiance incorrects.

Pour déterminer la présence d'hétéroscédasticité, il est possible de réaliser le test statistique de Breush-Pagan qui supposent que les erreurs de modélisation ont toutes la même variance, correspondant à l'une des trois hypothèses sur les erreurs. Ce test est disponible sur R avec le package `lmtest` de Zeileis and Hothorn [2002].

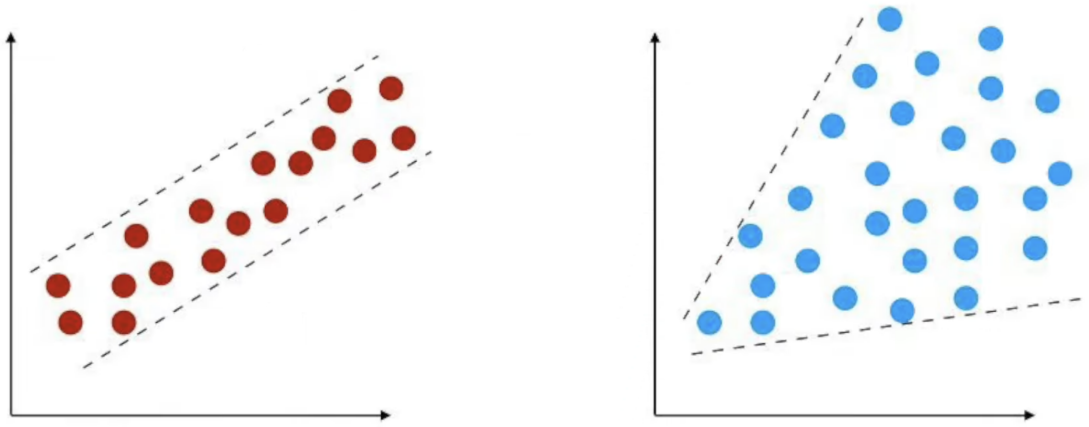

FIGURE 5 – Tracé avec des données aléatoires montrant l'homoscédasticité à gauche et l'hétéroscédasticité à droite.

Dans le but de résoudre le problème d'hétéroscédasticité du modèle, d'après Chen and Qin [2010], on aura tendance à utiliser celui de régression quantile d'ordre 0.5 (correpsondant à la médiane). En effet, contrairement à la régression linéaire classique qui minimise les carrés des déviations des résidus, la régression quantile minimise les déviations absolues des résidus. Cette approche permet à la régression quantile de mieux gérer l'hétéroscédasticité en limitant l'impact des observations ayant des résidus plus importants sur l'estimation des paramètres du modèle.

1.4.3 Résoudre la distribution non normale

L'une des hypothèses fondamentales sur les erreurs est celle de normalité des résidus. Souvent, dans le cas où cette hypothèse est rejetée, la régression linéaire conduit à des estimations biaisées et des intervalles de confiance erronés. Cela peut également entrainter le rejet de l'hypothèse sur l'homoscédasticité vue précédement.

L'utilisation de la régression quantile est alors une bonne alternative dans ce cas là. La raison est qu'elle suppose la distribution asymétrique de Laplace pour les résidus. En effet, elle minimise la somme des valeurs absolues des résidus (5) plutôt que la somme des carrés des résidus (2), ce qui la rend moins sensible à la distribution des résidus et plus robuste aux données aberrantes.

1.4.4 L'intérêt et les inconvénients de la modélisation des quantiles

La régression quantile se distingue par sa capacité à considérer non seulement la médiane, mais également tout autre quantile souhaité. En effet, selon l'objectif visé, modéliser le quantile conditionnel peut être plus pertinent que d'examiner la moyenne conditionnelle, notamment lorsqu'on souhaite étudier les extrémités de la distribution. L'analyse des quantiles élevés et faibles revêt alors un intérêt particulier, tel que pour appréhender les revenus les plus élevés et les plus faibles, bien qu'il soit un exemple parmis tant d'autres.

L'un des inconvénients majeurs de la régression quantile réside dans la complexité des calculs nécessaires pour obtenir l'estimation des paramètres, en raison de l'algorithme d'optimisation

numérique à mettre en œuvre pour minimiser la fonction de perte. De plus, la régression quantile n'étant pas couramment utilisée, elle manque de visibilité.

1.4.5 Résumé des avantages et des inconvénients des régressions quantile et linéaire

Pour résumer, nous avons mis en place le tableau des avantages et des inconvénients des régressions quantile et linéaire :

Régression quantile		Régression linéaire	
Avantages	Inconvénients	Avantages	Inconvénients
- Moins impacté par les valeurs aberrantes ;	- Calcul complexe des estimateurs ;	- Calcul des estimateurs plus simple ;	- Impacté par les valeurs aberrantes ; (Conséquence : estimations biaisés)
- Utilisation possible avec hétéroscédasticité ;	- Manque de visibilité ;	- Régression la plus connue	- Ne peut être possible qu'en présence d'homoscédasticité ; (Conséquence : estimations biaisés)
- Peut modéliser plusieurs quantiles.		- Meilleure précision si les 3 hypothèses sur les erreurs sont respectées.	- Les résidus doivent suivre une loi Normale ; (Conséquence : estimations biaisés)
			- Ne peut modéliser que la moyenne.

TABLE 1 – Tableau résumant les avantages et les inconvénients des régressions quantile et linéaire l'une par rapport à l'autre.

1.5 Les différents packages

Nous utilisons l'environnement de développement RStudio pour étudier ces deux régressions, car il est bien adapté au traitement de données et à l'analyse statistique. Pour étendre les fonctionnalités de base de R, nous avons dû télécharger des packages R spécifiques pour la modélisation des données et d'autres besoins associés. Les packages sont des bibliothèques de fonctions et de données que l'on peut télécharger directement sur R pour étendre les possibilités du logiciel.

Pour l'utilisation de la régression linaire, il est nécessaire de télécharger le package `stats` de R Core Team [2023].

Pour utiliser la régression linéaire, il est indispensable de télécharger soit le package `lqmm` de Geraci [2014], soit le package `quantreg` de Koenker [2023]. Dans le cadre de notre TER, nous allons opter pour le package `lqmm`, parce qu'il repose sur la distribution asymétrique de Laplace.

La représentation graphique des résultats se fait grâce au package `ggplot2` de Wickham [2016].

La vérification de la présence d'hétéroscédasticité, peut être faite avec le package `performance` de Lüdecke et al. [2021] en utilisant la fonction check_heteroscedasticity().

2 Application sur la pluviométrie

2.1 Présentation des données

Dans cette seconde partie, nous allons appuyer nos propos théoriques en appliquant nos méthodes sur des données réelles provenant de deux bases distinctes : les données de la bouée houle du Cap Ferret (campagne 03302 Cap Ferret) et les données météorologiques Infoclimat mesurées sur le site de Bordeaux. Les données ainsi que leurs unitées sont les suivantes :

- precipitation_cum : précipitation cumulée en millimètre (mm) ;
- temperature : la température en degrés Celsius (°C) sur Bordeaux ;
- pression : pression atmosphérique pascal (Pa) ;
- point2rose : point de rosée [1] en degrés Celsius (°C) ;
- vent_vit_rafale : vitesse des rafales de vent en kilomètres par heure (km/h).

À partir de celles-ci, nous allons chercher à déterminer les conditions météorologiques favorables ou défavorables lors de fortes pluies en nous intéressant à la precipitation cumulée, c'est-à-dire la force des précipitations, en l'occurrence la pluie. Étudier les fortes pluies présente un intérêt particulier car cela permet de faire le lien avec le phénomène d'inondation.

Avant d'entamer l'analyse, il convient de préciser que notre étude se focalisera principalement sur la régression quantile d'ordre $\tau = 0.9$, c'est-à-dire le neuvième décile. Nous nous intéresserons en effet aux valeurs élevées de précipitation, correspondant aux fortes pluies, d'où l'importance de bien choisir le quantile à étudier. De plus, une étude de la moyenne ne serait pas pertinente dans notre contexte, car elle ne reflèterait pas le comportement des précipitations lors de fortes pluies.

Nous utiliserons le logiciel R et deux packages spécifiques pour effectuer nos analyses : le package `lqmm` pour la régression quantile et le package `stats` pour la régression linéaire. Le code utilisé est présenté en annexe.

Cette application sera divisée en plusieurs parties. Tout d'abord, une analyse univariée sera réalisée, suivie d'une analyse multivariée. Enfin, nous conclurons en interprétant les résultats pour déduire les conditions météorologiques favorables à des inondations.

1. Le point de rosée est la température à laquelle l'air doit être refroidi pour que l'humidité qu'il contient commence à se condenser.

2.2 Analyse univarié

Nous commençons notre application par une analyse univarié. Cette première partie consiste a modéliser la variable réponse précipitation cumulée en fonction des variables explicatives température, rafale de vent, point de rosée et pression, de manière indépendante les unes des autres.

2.2.1 Représentation graphique et interprétation

Pour débuter notre analyse, nous avons tracé à l'aide du package `ggplot2` du logiciel R, les différentes droites de régression quantile pour chaque décile, ainsi que la droite de régression linéaire. Ceci nous permettra d'analyser le comportement des courbes les unes par rapport aux autres.

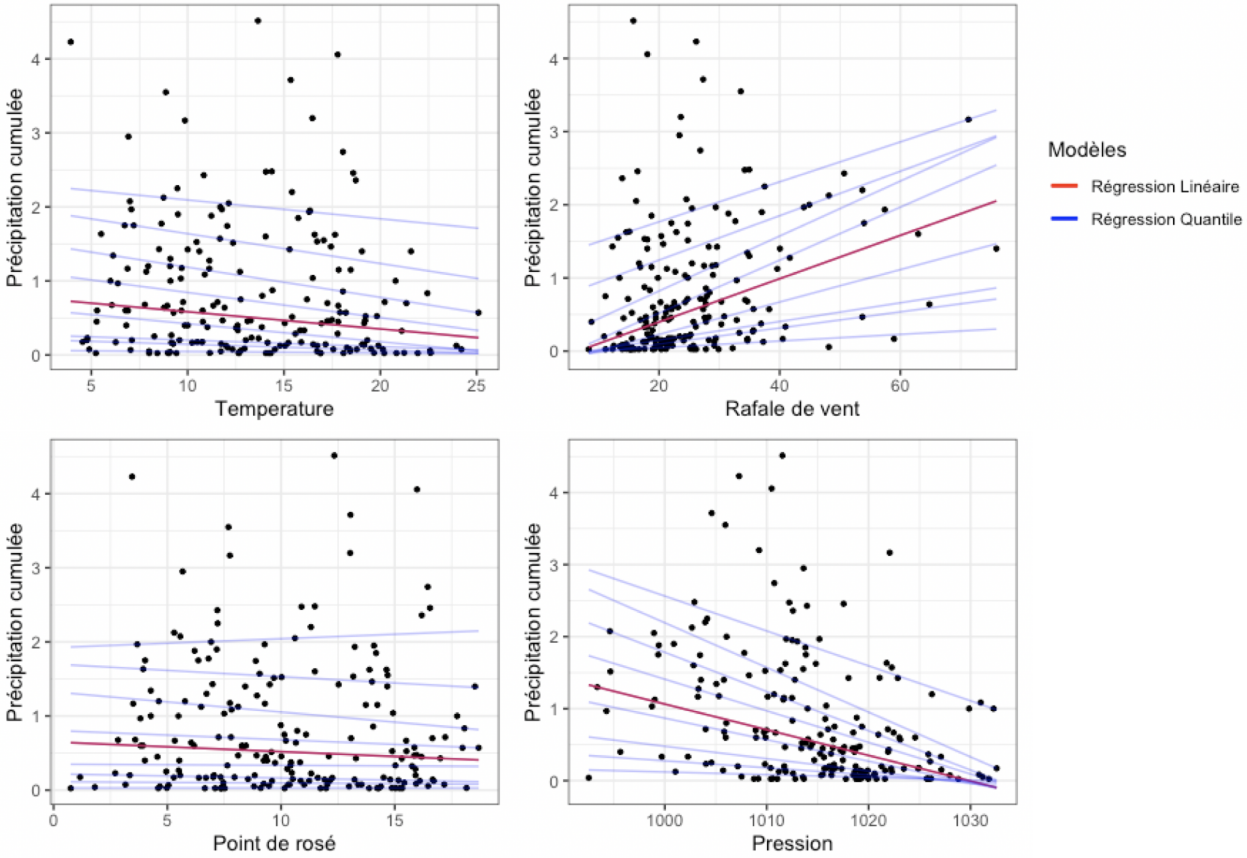

FIGURE 6 – Graphiques représentant la droite de régression linéaire et les droites de régression quantile pour chaque décile, la variable réponse précipitation cumulée et les variables dépendantes température, rafale de vent, point de rosé et pression.

Pour que la régression linéaire soit correctement utilisée, il est nécessaire de satisfaire la condition d'une distribution normale. Dans ce contexte, les droites de régression pour la moyenne et la médiane devraient être confondues, ce qui n'est pas observé ici. On observe que la droite de régression linéaire est influencée par les valeurs extrêmes.

Visuellement, on peut observer sur la Figure 6 un effet d'hétéroscédasticité qui n'est pas compatible avec la régression linéaire classique.

Les deux problèmes évoqués précédemment démontrent à quel point l'utilisation de la régression linéaire peut conduire à des résultats erronés.

En ce qui concerne le neuvième décile, nous constatons que l'impact sur notre variable réponse n'est pas uniforme pour chaque décile représenté. En effet, si nous comparons la pente du neuvième décile avec celles des autres déciles pour la variable explicative point de rosée, nous constatons que la première est positive tandis que les autres sont négatives. Autrement dit, les fortes pluies sont bien plus fréquentes lorsque les rafales de vent sont élevés.

2.2.2 Formule de régression quantile

À l'aide du package `lqmm`, nous pouvons dresser le tableau des estimations et p valeurs associées à chaque variable indépendante :

Covariables	Intercept	Estimation	p valeur
vent_vit_rafale	0.290	0.063	2.595e-10
pression	41.501	-0.039	1.747e-07
point2rose	0.937	0.110	0.0014509
temperature	1.175	0.069	0.0004453

TABLE 2 – Tableau répertoriant les estimations des modèles de régression univarié pour un quantile d'ordre 0.9

Les résultats obtenus nous permettent d'interpréter que, pour un intervalle de confiance de 95%, les variables point de rosée, vitesse des rafales de vent, pression atmosphérique et température ont des effets significatifs sur les précipitations cumulées. Par exemple, une augmentation d'un pascal de la pression atmosphérique entraînerait une diminution de 0.039 millimètre du neuvième décile des précipitations cumulées, toutes choses étant égales par ailleurs.

2.3 Analyse multivarié

Nous continuons notre application par une analyse multivarié. C'est-à-dire, qui cherche à expliquer la relation entre la variable réponse en fonction un ensemble de variables explicatives.

Pour ce modèle nous utiliserons les variables explicatives température, rafale de vent, point de rosée et pression. Nous débutons cette analyse en déterminant les estimations et les p-valeurs associées, qui indiquent la significativité statistique de chaque coefficient de régression. Pour cela, nous utilisons la régression quantile d'ordre 0.1, 0.5 (la médiane) et 0.9, ainsi que la régression linéaire, afin de les comparer (Table 3).

Covariables du modèle	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$		moyenne	
	$\hat{\beta}$	p	$\hat{\beta}$	p	$\hat{\beta}$	p	$\hat{\beta}$	p
Intercept	41.344	2.2e-7	41.344	2.4e-6	41.344	6.3e-9	41.344	4.8e-8
vent_vit_rafale	0.014	0.03	0.014	0.01	0.014	0.016	0.014	0.01
point2rose	0.226	4.6e-6	0.226	5.5e-7	0.226	8.6e-8	0.226	3.9e-8
pression	- 0.041	1.8e-7	- 0.040	2.9e-6	- 0.039	1.0e-8	0.040	7.8e-8
temperature	- 0.208	4.3e-6	- 0.208	1.1e-7	- 0.208	2.4e-9	0.208	1.2e-8

TABLE 3 – Tableau répertoriant tous les estimations du modèle de régression quantile d'ordre 0.1, 0.5 (la médiane) et 0.9, ainsi que du modèle de régression linéaire.

À la vue de la table 3 nous constatons que la vitesse de vent en rafale, le point de rosée et la température ainsi que l'intercept on la même estimation quelque soit l'ordre du quantile modélisé. Hormis, la pression où nous constatons une augmentation de 0,001 de l'estimation à chaque ordre de quantile que nous avons modélisé. Cela nous indique que pour la même valeurs de la pression il y a 1 à 2 unités de précipitation cumulée supplémentaire. Ce qui nous amène à penser que plus la pression atmosphérique à Bordeaux est élevé plus il est probable d'observer là-bas de fortes pluies.

Dans notre cas, passer par la régression quantile d'ordre 0.5 pour contrer l'effet des valeurs extrêmes sur la régression linéaire n'a pas grand intérêt. Cela nous est montré dans la Table 3, les estimations des variables explicatives donnée par la régression quantile d'ordre 0.5 et celle de la régression linéaire sont égaux.

2.4 Interprétation des résultats

Les résultats des modèles de régression quantile effectués nous montrent que les conditions météorologiques ont un impact identique quelque soit la fréquence de la pluie. Il n'y a donc pas de conditions particulières aux phénomènes d'inondation provoqués par de fortes pluies. Néanmoins, une forte la pression atmosphérique est une caractéristique météorologique des fortes pluies.

Conclusion

La régression quantile est considérée comme une alternative à la régression linéaire classique dans certains cas, en raison des limites liées aux hypothèses fondamentales sur les erreurs de la régression classique. Contrairement à la régression classique, les hypothèses rejetées ne sont pas des limites pour la régression quantile, ce qui la rend plus robuste et permet une meilleure estimation des paramètres.

Par ailleurs, la modélisation de l'espérance conditionnelle de la variable réponse en fonction des variables explicatives peut ne pas être adaptée à toutes les situations. Dans certains cas, il peut être plus pertinent de modéliser un quantile spécifique pour étudier le comportement de la variable réponse en fonction des variables explicatives.

Pour toutes ces raisons, il est essentiel de connaître la régression quantile et de savoir l'utiliser efficacement. Ainsi, la réalisation de ce TER nous a permis d'introduire les méthodes et l'utilisation de cette méthode de régression, qui nous semble importante de mettre en lumière.

Annexes

Références

Song Xi Chen and Yingli Qin. A two-step estimator for high-dimensional regression with heteroscedasticity. *The Annals of Statistics*, 38(5) :2998–3028, 2010.

Marco Geraci. Linear quantile mixed models : The lqmm package for laplace quantile regression. *Journal of Statistical Software*, 57(13) :1–29, 2014. doi : 10.18637/jss.v057.i13.

Pauline Givord and Xavier D'Haultfœuille. La régression quantile en pratique. *Méthodologie statistique*, page 11, 2013. URL `https://www.insee.fr/fr/statistiques/fichier/1381107/doc_regression_quantile.pdf`.

Goldberger. Théorie économétrique. *New York : John Wiley and Sons*, pages 238–243, 1964.

Roger Koenker. *quantreg : Quantile Regression*, 2023. URL `https://CRAN.R-project.org/package=quantreg`. R package version 5.95.

Roger Koenker and José Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448) :1296–1310, 1999. URL `http://dx.doi.org/10.1080/01621459.1999.10473882`.

Roger W Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1) :33–50, 1978. URL `https://EconPapers.repec.org/RePEc:ecm:emetrp:v:46:y:1978:i:1:p:33-50`.

Daniel Lüdecke, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. performance : An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60) :3139, 2021. doi : 10.21105/joss.03139.

A. Quetelet. *Sur l'homme et le développement de ses facultés, ou, Essai de physique sociale*. Number vol. 1 in Sur l'homme et le développement de ses facultés, ou, Essai de physique sociale. Bachelier, 1835. URL `https://books.google.fr/books?id=VXkZAAAAYAAJ`.

R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

Hadley Wickham. *ggplot2 : Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Keming Yu and Rana A. Moyeed. Bayesian quantile regression. *Statistics and Probability Letters*, 54(4) :437–447, 2001. URL `https://www.sciencedirect.com/journal/statistics-and-probability-letters/vol/54/issue/4`.

Achim Zeileis and Torsten Hothorn. Diagnostic checking in regression relationships. *R News*, 2(3) :7–10, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.