

# Régression quantile

Abouqateb Mouad, Cadon Mathilde, Elhadji Leye

October 2019

## Introduction

La régression ordinaire (OLS) ne s'intéresse qu'à la moyenne de la distribution conditionnelle, or il est possible de se trouver dans le cas de données où la moyenne ne suffit pas à expliquer les choses. C'est à ce moment que les régressions quantiles interviennent puisqu'elles prennent en compte l'ensemble de la distribution conditionnelle de la variable d'intérêt.

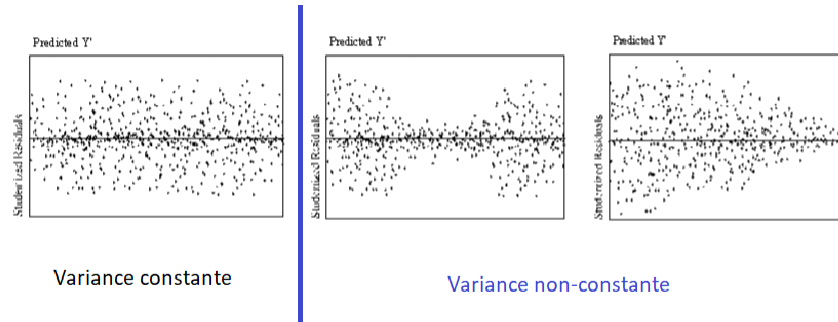


Figure 1: Exemple de jeux de données où la régression quantile est utile.

## 1 Rappels : Médiane / quantiles

### 1.1 Médiane

**Définition 1.** Soit  $y_1, \dots, y_n \in \mathbb{R}$ , on a :

$$Med_n(y_1, \dots, y_n) \in \operatorname{argmin}_{\mu \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |y_i - \mu|$$

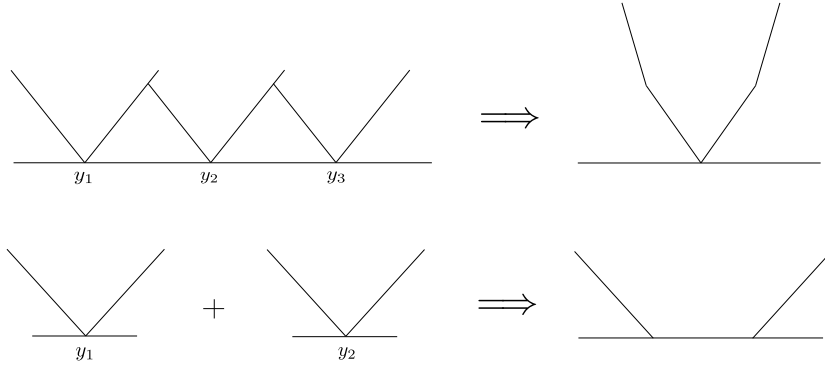


Figure 2: Fonction d'optimisation  $f : \mu \rightarrow \frac{1}{n} \sum_{i=1}^n |y_i - \mu|$  pour  $n = 2$  et  $n = 3$

Pour 3 observations  $y_1, y_2, y_3$  la fonction d'optimisation admet un unique minimum global.

Pour 2 observations  $y_1, y_2$  tout les points entre  $y_1$  et  $y_2$  sont des minimums globaux

Le problème d'optimisation dépend de la parité des points:

En pratique, pour  $y_{(1)} \leq \dots \leq y_{(n)}$  on prend:

$$Med_n(y_1, \dots, y_n) = \begin{cases} y_{\lfloor \frac{n}{2} \rfloor + 1} & \text{si } n \text{ est impair} \\ \frac{y_{\lfloor \frac{n}{2} \rfloor} + y_{\lfloor \frac{n}{2} \rfloor + 1}}{2} & \text{si } n \text{ est pair} \end{cases}$$

**Remarque :**

$$f : \begin{cases} \mathbb{R} & \longrightarrow \mathbb{R} \\ \mu & \longmapsto \frac{1}{n} \sum_{i=1}^n |y_i - \mu| \end{cases}$$

La fonction  $f$  est une fonction convexe. Il y aura donc une solution optimale.

Notre fonction d'optimisation n'est pas toujours lisse, cela implique la non existence du gradient dans les points critiques.

Nous pouvons également reformuler le problème :

$$\begin{aligned} & \min_{\mu, t_1, \dots, t_n} \left( \frac{1}{n} \sum_{i=1}^n t_i \right) \\ \text{s.c: } & \forall i \in \{1, \dots, n\} \quad \begin{cases} -t_i \leq y_i - \mu \leq t_i \\ 0 \leq t_i \end{cases} \end{aligned}$$

Les deux problèmes ainsi posés sont équivalents.

Dans les années 1940, George Dantzig s'intéresse à la programmation linéaire (PL). L'intérêt principal de ces méthodes est leurs robustesses (STAT 593).

la robustesse d'un estimateur est sa capacité à ne pas être perturbé par une petite modification dans les données ou dans les paramètres du modèle choisi pour l'estimation.

Le package "CVXPY" de Python contient diverses méthodes numériques de résolution pour ces problèmes.

### Contrepartie "théorique":

Soit  $Y$  une variable aléatoire réelle, alors

$$Med(Y) \in \operatorname{argmin}_{\mu \in \mathbb{R}} \mathbb{E}_Y(|Y - \mu|)$$

Rappel :

$$\mathbb{E}(Y) = \operatorname{argmin}_{\mu \in \mathbb{R}} \mathbb{E}_Y(|Y - \mu|^2)$$

## 1.2 Quantile

**Définition 2.** (Usuelle) Le quantile de niveau  $\alpha$  est défini,  $\forall \alpha \in ]0, 1[$ ,

$$q_\alpha(y_1, \dots, y_n) = \inf\{t \in \mathbb{R} : \mathbb{F}_n(t) \geq \alpha\}$$

avec

$$\mathbb{F}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{y_i \leq t}$$

Interprétation pour  $\alpha = 0.75$  :

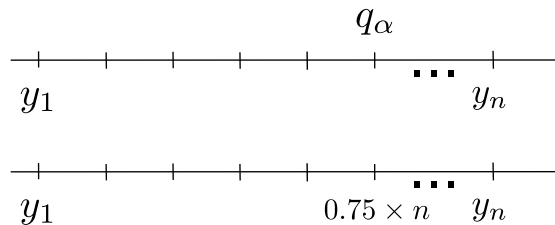


Figure 3: Schéma d'explication du quantile  $\alpha = 0.75$

Pour une variable aléatoire réelle  $Y$ ,

$$q_\alpha(Y) = \inf\{t \in \mathbb{R} : \mathbb{F}_Y(t) \geq \alpha\}$$

où  $\mathbb{F}_y$  est la fonction de répartition de  $Y$

**Propriété 1.**

$$q_\alpha(Y) \in \operatorname{argmin}_{\mu \in \mathbb{R}} \underbrace{\alpha \int_{y > \mu} |y - \mu| d\mathbb{F}_Y(y) + (1 - \alpha) \int_{y \leq \mu} |y - \mu| d\mathbb{F}_Y(y)}_{\psi(\mu)}$$

**Preuve 1.**

$$\begin{aligned} \psi(\mu) &= \alpha \int_{\mu}^{+\infty} (y - \mu) d\mathbb{F}_Y(y) + (1 - \alpha) \int_{-\infty}^{\mu} (\mu - y) d\mathbb{F}_Y(y) \\ \psi'(\mu) &= -\alpha \int_{\mu}^{+\infty} d\mathbb{F}_Y(y) + (1 - \alpha) \int_{-\infty}^{\mu} d\mathbb{F}_Y(y) \\ \psi'(\mu) &= -\alpha \mathbb{P}(Y \geq \mu) + (1 - \alpha) \mathbb{P}(Y \leq \mu) \\ \psi'(\mu) &= -\alpha(1 - \mathbb{F}_Y(\mu)) + (1 - \alpha)\mathbb{F}_Y(\mu) \end{aligned}$$

On veut trouver l'endroit du point critique :  $\psi'(\mu) = 0$

$$\begin{aligned} \psi'(\mu) = 0 &\iff 0 = \mathbb{F}_Y(\mu)(\alpha + 1 - \alpha) - \alpha \\ &\iff \alpha = \mathbb{F}_Y(\mu) \\ (\text{inversion}) &\iff \mu = \mathbb{F}_Y^{-1}(\alpha) \end{aligned}$$

**Remarque :**

On définit la fonction perte  $l_\alpha$  :

$$\begin{aligned} l_\alpha : \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto \begin{cases} -(1 - \alpha)x & \text{si } x \leq 0 \\ \alpha x & \text{si } x \geq 0 \end{cases} \\ l_\alpha : x &\rightarrow \alpha|x| \mathbb{1}_{x > 0} + (1 - \alpha)|x| \mathbb{1}_{x \leq 0} \end{aligned}$$

On a: A encadrer

$$\psi(\mu) = \mathbb{E}_Y(l_\alpha(Y - \mu))$$

En effet:

$$\begin{aligned} \psi(\mu) &= \alpha \int_{y > \mu} |y - \mu| d\mathbb{F}_Y(y) + (1 - \alpha) \int_{y \leq \mu} |y - \mu| d\mathbb{F}_Y(y) \\ &= \int \alpha|y - \mu| \mathbb{1}_{y > \mu} + (1 - \alpha)|y - \mu| \mathbb{1}_{y \leq \mu} d\mathbb{F}_Y(y) \\ &= \int (l_\alpha(y - \mu)) d\mathbb{F}_Y(y) \end{aligned}$$

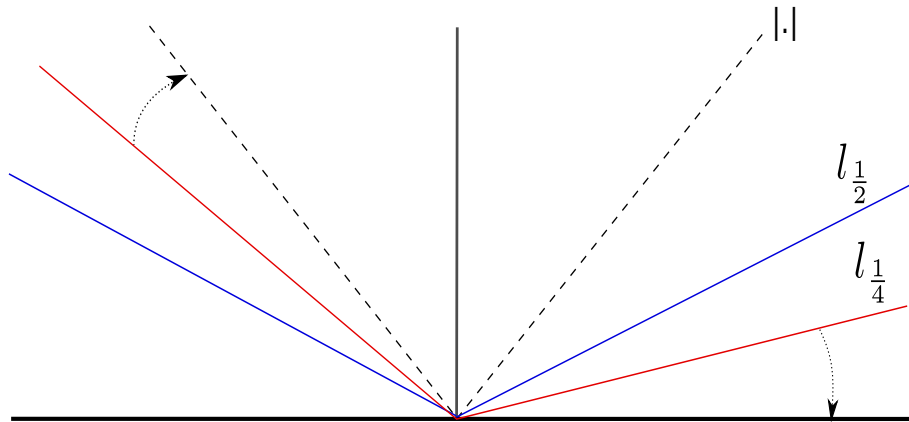


Figure 4: La perte flipper "pinball loss"

La fonction coût qu'on a défini  $l_\alpha$ , s'appelle **pinball loss**, sa forme et sous forme d'une trajectoire de balle de flipper, de pente qui dépend de  $\alpha$ .

Interprétation :

Cas  $\alpha$  petit  $\simeq 0$  :

Les erreurs (x) positives ne sont pas ou peu pénalisées.  
 $\Rightarrow$  on peut prendre un point vers la gauche de l'échantillon

Cas  $\alpha \simeq \frac{1}{2}$  :

On récupère la médiane

Cas  $\alpha$  grand  $\simeq 1$  :

Les erreurs négatives ne sont pas ou peu pénalisées.  
 $\Rightarrow$  on peut prendre un point à droite de l'échantillon

Explications

On cherche à minimiser :

$$\widehat{\mu}_\alpha : \mu \rightarrow \frac{1}{n} \sum_{i=1}^n l_\alpha(y_i - \mu)$$

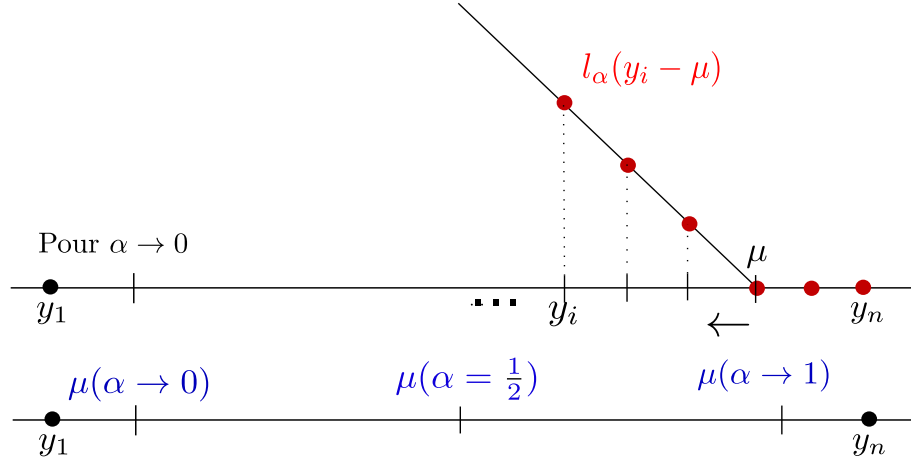


Figure 5: Explication pour le choix de  $\mu$ , cas où  $\alpha = 0$  pour le premier schéma

La perte "flipper",  $l_\alpha$ , permet de donner plus de poids aux observations les plus lointaines de  $\mu$ .

La régression quantile étant une régression biaisée, le paramètre  $\alpha$  règle le niveau de la pente de pondération ( $\alpha$  d'un côté,  $(1 - \alpha)$  de l'autre) de la fonction de perte associée à la régression et reflète ainsi le déséquilibre désiré dans la prévision quantile.

Si  $\alpha \simeq 0$ , il y a une pénalisation avec la distance du point à  $\mu$ . Il faut minimiser la fonction donc on va avoir tendance à faire bouger le  $\mu$  vers la gauche.

Si  $\alpha \simeq 1$ , il va falloir faire bouger le  $\mu$  vers la droite pour minimiser la fonction.

**Théorème 1.**

$$\mu_\alpha \in \underset{\mu}{\operatorname{argmin}} \mathbb{E}_Y(l_\alpha(Y - \mu))$$

est le quantile  $q_\alpha(Y)$

**Remarque :**

$\mu \longrightarrow l_\alpha(\mu)$  est convexe pour tout  $\alpha \in ]0, 1[$ .

Preuve par moyenne de fonction convexe.

Contre partie empirique :

$$q_{\alpha,n}(y_1, \dots, y_n) \in \underset{\mu \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n l_\alpha(y_i - \mu)$$

**Remarque :**

Ce problème est aussi formulable comme un LP.

$$\begin{aligned} \min_{u,v,\mu} & \alpha \sum_{i=1}^n u_i + (1-\alpha) \sum_{i=1}^n v_i \\ \text{s.c.} & \begin{cases} u_i - v_i = y_i - \mu \\ u_i \geq 0 \\ v_i \geq 0 \end{cases} \end{aligned}$$

Interprétation:

On sépare les parties  $\geq 0$  et  $\leq 0$  dans les valeurs absolues.

$$\begin{cases} x = x_+ - x_- \\ |x| = x_+ + x_- \end{cases}$$

## 2 Régression quantile

$y_1, \dots, y_n \in \mathbb{R}$  observations

$x_1, \dots, x_n \in \mathbb{R}^p$  variables explicatives

**Définition 3.** Soit  $\alpha \in ]0, 1[$ . On appelle régression quantile les coefficients

$$\hat{\beta}^\alpha \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n l_\alpha(y_i - x_i^T \beta)$$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$$

- $\hat{\beta}^\alpha = \hat{\beta}^\alpha(X, y)$
- $\hat{\beta}^\alpha(X, c.y) = c\hat{\beta}^\alpha(X, y)$  avec  $c > 0$ , équivariance d'échelle (scale equivariant)
- $\hat{\beta}^\alpha(X, y + Xv) = \hat{\beta}^\alpha(X, y) + v$  pour  $v \in \mathbb{R}^p$  équivariance en regression

Rappel  $h : \mathbb{R} \rightarrow \mathbb{R}$  fonction croissante.

$$q_{h(Y)}(\alpha) = h(q_Y(\alpha))$$

**Preuve 2.** Utiliser la définition des quantiles.

Cette propriété permettra de créer des estimateurs pertinents quand on observe indirectement des données.

**Exemple:**

$$h(y) = \min(\max(y, \tau); \bar{\tau})$$

C'est le cas ou l'on n'observe pas  $y_1, \dots, y_n$  mais

$$\begin{cases} \underline{\tau} & \text{si } y_i \leq \underline{\tau} \\ y_i & \text{si } \underline{\tau} \leq y_i \leq \bar{\tau} \\ \bar{\tau} & \text{si } y_i \geq \bar{\tau} \end{cases}$$

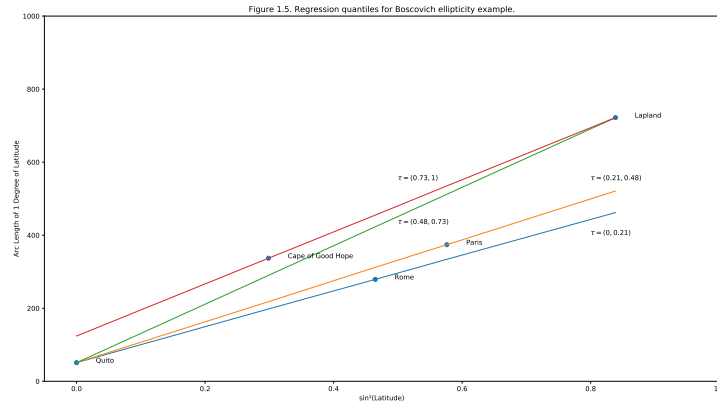
(Kroenker 2005)

Autre cas d'intérêt de la régression quantile : le cas hétéroscédastique c'est à dire  $(y_i - x_i^T \beta^*)$  ont des variances différentes.

**N.B:** L'estimateur  $\beta$  dépend du niveau de quantile  $\alpha$  que l'on spécifie.

**Exemple:**

Prenons l'exemple du jeu de données de Boscovich sur l'aplatissement de la Terre. On observe les régressions quantile (le quantile est de niveau  $\tau$ )



A l'aide du package 'CVXPY', nous avons estimé le paramètre "Arc length of 1 degree of Latitude" en fonction du  $\sin^2$  de la Latitude des villes. Nous avons ainsi obtenu 4 droites de régressions, correspondant à toutes les valeurs possibles de  $\tau$ . Il n'y a, en effet, que 4 régressions différentes pour toutes les



valeurs possibles des quantiles. Pour plus de précisions, se référer au notebook associé.

## **Conclusion**

Comparé à la régression linéaire classique, la régression quantile nous offre deux avantages principaux: Robustesse face aux valeurs aberrantes, et la flexibilité par rapport à la distribution des données.