# HMMA307: Advanced Linear Modeling

## Chapter 7 : Quantile Regression

**Yani Bouaffad  Ryma Lakehal  Loïc Sauton**

https://github.com/ybouaffad/HMMA307_CM_Quantile_Regression

Université de Montpellier

# Table of Contents

# Introduction

Classical linear regression estimates the mean response of the dependent variable dependent on the independent variables. There are many cases, such as skewed data, multimodal data, or data with outliers, when the behavior at the conditional mean fails to fully capture the patterns in the data.

# Table of Contents

# Reminder : Median/Quantiles

## Median Definition

Let $y_1, \ldots, y_n \in \mathbb{R}$, we have :

$$Med_n(y_1, \ldots, y_n) \in \arg\min_{\mu \in \mathbb{R}} \quad \frac{1}{n} \sum_{i=1}^{n} |y_i - \mu|$$

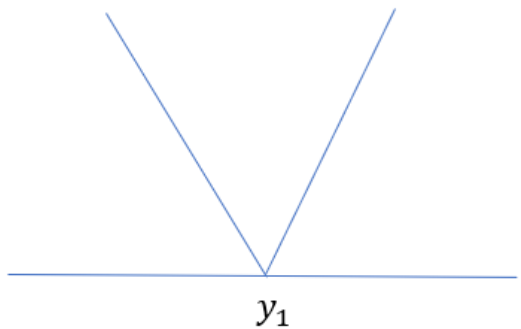


Figure: Optimization function for n=1

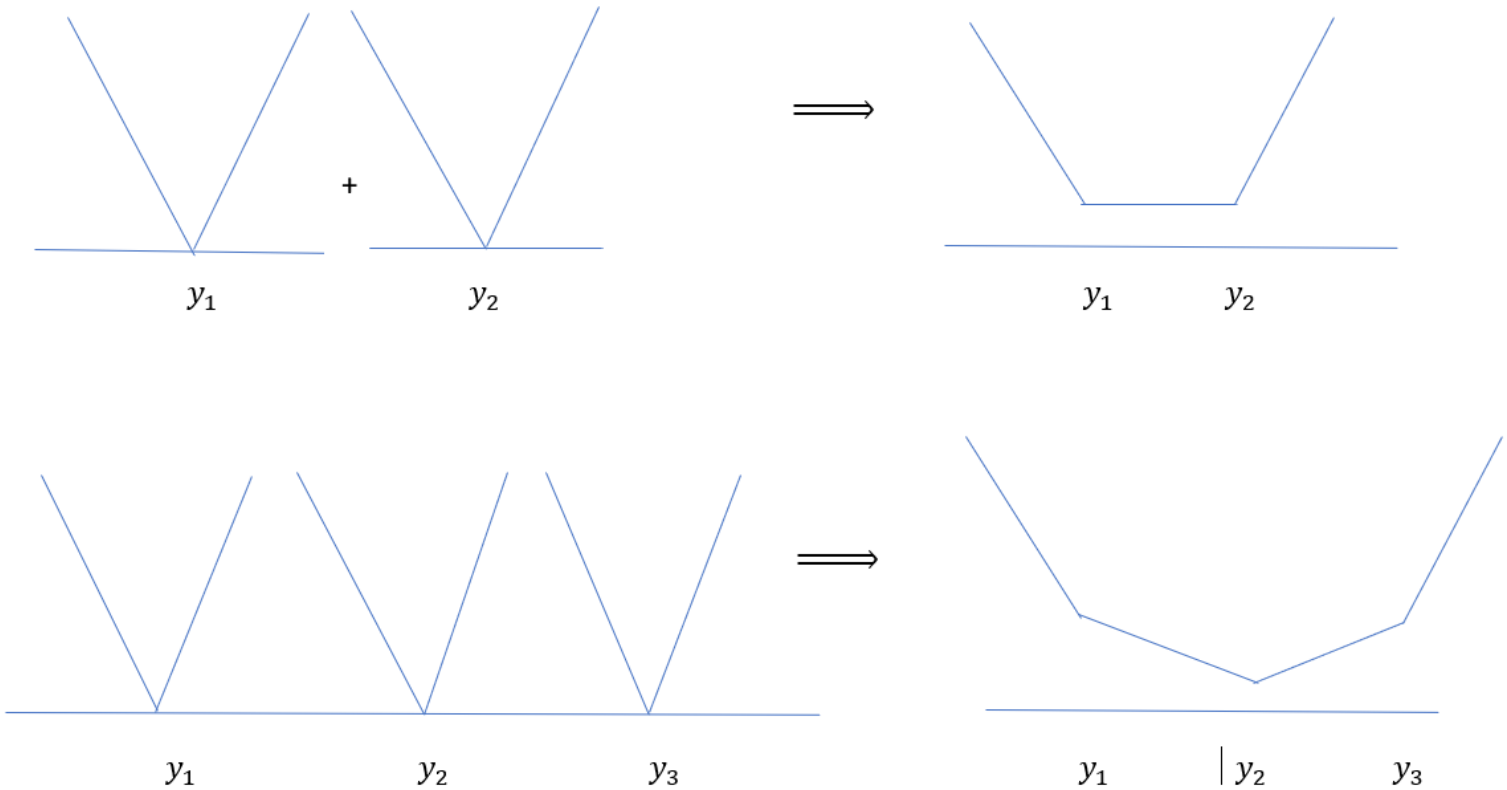


Figure: Optimization function for n=2 and for n=3

for 2 observations $y_1; y_2$ all points between $y_1$ and $y_2$ are global minimums
for 3 observations $y_1; y_2; y_3$ the optimization function admits a unique global minimizer

The optimization problem depends on the parity of the points :
In practice, for $y_{(1)} \leqslant \cdots \leqslant y_{(n)}$ we have :

$$\mathrm{Med}_n(y_1, \ldots, y_n) = \begin{cases} y_{\lfloor \frac{n}{2} \rfloor + 1} & \text{if } n \text{ is odd} \\ \frac{y_{\lfloor \frac{n}{2} \rfloor} + y_{\lfloor \frac{n}{2} \rfloor + 1}}{2} & \text{if } n \text{ is even} \end{cases}$$

**Remark:**

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

The function f is a convex function. So there will be an optimal solution.
Our optimization function is not always smooth, this implies the non-existence of the gradient in the critical points.

# Quantile

## Quantiles Definition

Let $Y$ be a real valued random variable with cumulative distribution function $F_Y(y) = P(Y \leq y)$. The $\alpha$-th quantile of $Y$ is given for $\alpha \in ]0,1[$ by :

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

**Remark :** Let us define the loss function $l_\alpha$

$$l_\alpha : \left| \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \left\{ \begin{array}{ll} -(1-\alpha)x & \text{si } x \leqslant 0 \\ \alpha x & \text{si } x \geqslant 0 \end{array} \right. \end{array} \right.$$

$$l_\alpha : x \to \alpha|x|\ \mathbb{1}_{x>0} + (1-\alpha)|x|\ \mathbb{1}_{x\leqslant 0}$$

# Quantiles

A specific quantile can be found by minimizing the expected loss of $Y - \mu$ with respect to $\mu$

$$\min_{\mu \in \mathbb{R}} \mathbb{E}(l_\alpha(Y - \mu)) =$$

$$\min_{\mu \in \mathbb{R}} \left\{ (\alpha - 1) \int_{-\infty}^{\mu} (y - \mu) dF_Y(y) \qquad + \alpha \int_{\mu}^{\infty} (y - \mu) dF_Y(y) \right\}$$

This can be shown by setting the derivative of the expected loss function to $0$ and letting $q_\alpha$ be the solution of

$$0 = (1 - \alpha) \int_{-\infty}^{q_\alpha} dF_Y(y) - \alpha \int_{q_\alpha}^{\infty} dF_Y(y).$$

This equation reduces to

$$0 = F_Y(q_\alpha) - \alpha$$

and then to

$$F_Y(q_\alpha) = \alpha.$$

Hence $q_\alpha$is $\alpha$ th quantile of the random variable Y.

# Table of Contents

# Quantile regression

$y_1, \ldots, y_n \in \mathbb{R}$ observations, $x_1, \ldots, x_n \in \mathbb{R}^p$ explanatory variables

- The quantile regression is described by the following equation:

$$y_i = x_i^\top \beta^\alpha + \varepsilon_i$$

  where $\beta^\alpha$ is the vector of unknown parameters associated with the $q^{th}$ quantile
- The OLS minimizes $\sum_i \varepsilon_i^2$, the sum of squares of the model prediction error $\varepsilon_i$
- The median regression, also called least absolute-deviation regression minimizes $\sum_i |\varepsilon_i|$
- The quantile regression minimizes

$$\mathbb{E}(l_\alpha(Y - X\beta)) = \sum_i \alpha\,|\varepsilon_i| + \sum_i (1-\alpha)\,|\varepsilon_i|$$

  a sum that gives the asymmetric penalties $\alpha\,|\varepsilon_i|$ underprediction and $(1-\alpha)\,|\varepsilon_i|$ overprediction

# Quantile regression

**Remark :**

$$\varepsilon_i = y_i - x_i^\top \beta^\alpha$$

**Definition**

Let $\alpha \in ]0,1[$. The $\alpha$th quantile regression estimator $\hat{\beta}^\alpha$ of $\beta$

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

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

# Table of Contents

# Advantages

- Flexibility for modeling data with heterogeneous conditional distributions.
- Median regression is more robust to outliers than the OLS regression.
- Richer characterization and description of the data: can show different effects of the independent variables on the dependent variable depending across the spectrum of the dependent variable.