1 Quantile $\mathbb{Q}_q$ of a random variable

Quantile of a sample: Given an unordered sample $y_1, y_2, ..., y_\ell$, we can construct a sorted sample $y^{(1)} \le y^{(2)} \le ... \le y^{(\ell)}$, where $y^{(i)}$ is the ith smallest value of the sample ($i = 1..\ell$), also known as the ith *order statistic*.

Informally, the q-quantile $y^{(q)}$ is the value that divides the ordered sample into two parts with proportions $q : (1-q)$. However, this definition is ambiguous. One practical approach is to use different formulas for $q \cdot \ell \notin \mathbb{N}$ and $q \cdot \ell \in \mathbb{N}$:

$$y^{(q)} := \begin{cases} y^{(\lceil q \cdot \ell \rceil)} & q \cdot \ell \text{ is not integer} \\ \frac{1}{2}\big(y^{(q\cdot\ell)} + y^{(q\cdot\ell+1)}\big) & q \cdot \ell \text{ is integer} \end{cases} \tag{1}$$

Quantile of a random variable: For a random variable Y, the quantile function, denoted either as $\mathbb{Q}_q[Y]$ or $y^{(q)}$, is defined as the inverse of its CDF:

$$\mathbb{Q}_q[Y] := \mathcal{F}_Y^{-1}(q) = \inf\{y \mid \mathcal{F}_Y(y) \ge q\}, \tag{2}$$

where inf denotes the infimum, which is the greatest lower bound, *i.e.*, $\mathbb{Q}_q[Y]$ is the smallest value y for which the probability $\mathbb{P}[Y \le y]$ is at least q.

Example: For a uniform distribution on interval $[a, b]$, the CDF is:

$$\mathcal{F}_Y(y^*) = \begin{cases} \frac{y^*-a}{b-a}, & y^* \in [a..b] \\ 0, & y^* < a \\ 1, & y^* > b \end{cases}$$

The corresponding quantile function is:

$$\mathbb{Q}_q[Y] = \mathcal{F}_Y^{-1}(q) = a + q \cdot (b-a).$$

Example: For a sample $\{y_1, ..., y_\ell\}$, the empirical CDF

$$\mathcal{F}_Y(y) := \frac{1}{\ell} \cdot \sum_{i=1}^{\ell} [\![y_i \le y]\!]$$

can be used in (2) to define quantiles $\mathbb{Q}_q$.

Some order statistics:

- $y^{(1)} = \min Y$ is the 1st order statistic
- $y^{(2)}$ is the 2nd order statistic (2nd smallest value)
- $y^{(\ell/2)}$ is the median, which divides the sample in half
- $y^{(\ell)} = \max Y$ is the last (ℓth) order statistic

NB: In $y^{(i)}$, values $i = 1..\ell$ are integers, while in $y^{(q)}$, values $q \in [0..1]$ are fractional, *e.g.*, $y^{(10)}$ refers to the 10th order statistic, while $y^{(0.1)}$ refers to the 0.1-quantile.

In practice, other definitions of quantiles $y^{(q)}$ are also used, *e.g.*, $y^{(q)} := y^{(\lceil q \cdot \ell \rceil)}$ for any ℓ

Quantile $\mathbb{Q}_q$ and probability $\mathbb{P}$: CDF maps real numbers $y \in \mathbb{R}$ to probabilities $p \in [0..1]$:

$$\mathcal{F}_Y(y) := \mathbb{P}[Y \le y] : y \to p. \tag{3}$$

Quantile $\mathbb{Q}_q$, being the inverse of CDF, maps probabilities $p \in [0..1]$ to real numbers $y \in \mathbb{R}$:

$$\mathbb{Q}_p[Y] = \mathcal{F}_Y^{-1}(p) : p \to y, \tag{4}$$

We specifically denote probability p as q to emphasize its connection to quantiles.

The meaning of $\mathbb{Q}_q[Y]$ is that it is the value of Y such that the probability of Y being less than or equal to $\mathbb{Q}_q[Y]$ is q:

$$\mathbb{P}\big[Y \le \mathbb{Q}_q[Y]\big] = q. \tag{5}$$

Conditional quantile $\mathbb{Q}_q[Y|X]$: The generalization of the quantile $\mathbb{Q}_q[Y]$ to the conditional case is straightforward; it's defined as the inverse of the conditional CDF:

$$\mathbb{Q}_q[Y|X] := \mathcal{F}_{Y|X}^{-1}(q) = \inf\big\{y \mid \mathcal{F}_{Y|X}(y) \ge q\big\}, \tag{6}$$

where $\mathcal{F}_{Y|X}(y) := \mathbb{P}[Y \le y|X]$ is the conditional CDF defined via the conditional PDF $f_{Y|X}(y) \equiv f_Y(y|X)$ (continuous case) or PMF $p_{Y|X}(y) \equiv p_{Y(y|X)}$ (discrete case).

The meaning of the conditional quantile $\mathbb{Q}_q[Y|X]$ is that it is the value of Y such that the probability of Y being less than or equal to $\mathbb{Q}_q[Y|X]$ given X is q:

$$\mathbb{P}\big[\, Y \le \mathbb{Q}_q[Y|X] \mid X \,\big] = q. \tag{7}$$

For a continuous random variable Y, the probability of Y being less than or equal to y^* is given by the cumulative distribution function (CDF):

$$\mathcal{F}_Y(y^*) := \mathbb{P}[Y \le y^*] = \int_{y=-\infty}^{y^*} f_Y(y)\,\mathrm{d}y,$$

where $f_Y(y)$ is the probability density function (PDF).

For a discrete random variable Y, the CDF is defined as:

$$\mathcal{F}_Y(y^*) := \mathbb{P}[Y \le y^*] = \sum_{y \le y^*} p_Y(y)$$

where $p_Y(y)$ is the probability mass function (PMF).

Some important quantiles:

- $\mathbb{Q}_0[Y] = \min Y$ is the minimum value
- $\mathbb{Q}_{1/4}[Y]$ is the 1st quartile (Q_1)
- $\mathbb{Q}_{1/2}[Y]$ is the median or 2nd quartile (Q_2)
- $\mathbb{Q}_{3/4}[Y]$ is the 3rd quartile (Q_3)
- $\mathbb{Q}_1[Y] = \max Y$ is the maximum value

Percentiles are also quantiles, *e.g.* $\mathbb{Q}_{0.95}[Y]$ is the 95th percentile.

Technically, $\mathbb{Q}_q[Y]$ is a function of q, and it is usually denoted as $Q_Y(p)$, similar to PDF $f_Y(y)$ and CDF $\mathcal{F}_Y(y)$.

However, the notation $\mathbb{Q}_q[Y]$ is used here to emphasize the analogy between quantiles $\mathbb{Q}_q[Y]$ and expectation $\mathbb{E}[Y]$.

2 Quantile loss $\mathcal{L}_q$

Check-loss: Consider an asymmetric loss function parameterized by $q \in (0, 1)$:

$$\begin{aligned}\mathcal{L}_q(\varepsilon) &:= \begin{cases} q \cdot \varepsilon & \varepsilon \geq 0 \\ -(1-q) \cdot \varepsilon & \varepsilon < 0 \end{cases} \\ &= \varepsilon \cdot q \cdot [\![\varepsilon \geq 0]\!] - \varepsilon \cdot (1-q) \cdot [\![\varepsilon \leq 0]\!],\end{aligned} \tag{8}$$

This loss function is also called the *pinball loss* and *quantile loss* (more on this below)

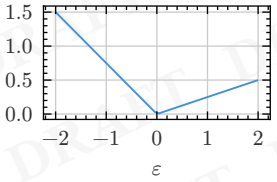
where $\varepsilon := y - \hat{y}$ is the error term (residual) and $\hat{y}$ is the prediction of a regression model.

Strictly speaking, this is an estimation of the error: $\hat{\varepsilon} := y - \hat{y}$; for different estimations of $\hat{y}$, there are different $\hat{\varepsilon}$

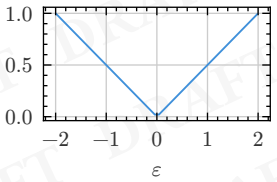
Constant model: Let's first consider the simplest case, where we look for a^* in the family of all constant models $a^* \in \{a \mid a = \text{const}\}$.

The empirical risk (expected check-loss) can be expressed as:

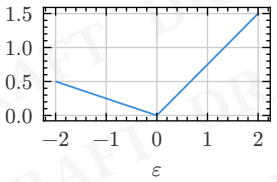
$$\begin{aligned}\mathcal{R}(a) &= \int \mathcal{L}_q(\varepsilon = y - a)\, \mathrm{d}\mathcal{F}(\boldsymbol{x}, y) && a(\boldsymbol{x}) \to a = \text{const} \\ &= \int \mathcal{L}_q(\varepsilon = y - a)\, \mathrm{d}\mathcal{F}(y) && \text{as } a \text{ is not a function of } \boldsymbol{x} \\ &= \int\limits_{y-a\geq 0} \mathcal{L}_q(y-a)\, \mathrm{d}\mathcal{F}(y) + \int\limits_{y-a<0} \mathcal{L}_q(y-a)\, \mathrm{d}\mathcal{F}(y) && \text{nonoverlapping regions: } \varepsilon \geq 0 \text{ and } \varepsilon < 0 \\ &= \int\limits_{y\geq a} (y-a)\cdot q\, \mathrm{d}\mathcal{F}(y) - \int\limits_{y<a} (y-a)\cdot(1-q)\, \mathrm{d}\mathcal{F}(y) \to \min_a && \text{expand } \mathcal{L}_q(\varepsilon) \text{ according to (8)}\end{aligned} \tag{9}$$



1a: Check loss $\mathcal{L}_{0.25}(\varepsilon)$



1b: Check loss $\mathcal{L}_{0.5}(\varepsilon)$



1c: Check loss $\mathcal{L}_{0.75}(\varepsilon)$

Risk minimization: The integral is split at $a = a^*$ into two independent regions: $(-\infty..a^*)$ and $[a^*..+\infty)$. By differentiating both integrals with respect to a, we can find a^*:

$$\begin{aligned}\frac{\partial}{\partial a}\mathcal{R}(a) &= q \cdot \int_{y\geq a} \frac{\partial}{\partial a}(y-a)\, \mathrm{d}\mathcal{F}(y) - (1-q)\cdot \int_{y<a} \frac{\partial}{\partial a}(y-a)\, \mathrm{d}\mathcal{F}(y) && \text{constants} \\ &= -q \cdot \int_{y=a^*}^{+\infty} \mathrm{d}\mathcal{F}(y) + (1-q)\cdot \int_{y=-\infty}^{a^*} \mathrm{d}\mathcal{F}(y) && \mathrm{d}\mathcal{F}(y) = f(y)\, \mathrm{d}y \\ &= -q\cdot(1-\mathcal{F}_Y(a)) + (1-q)\cdot\mathcal{F}_Y(a) = -q + \mathcal{F}_Y(a) && \mathcal{F}_Y(a) \equiv \mathcal{F}(Y=a)\end{aligned} \tag{10}$$

At the extreme point $a = a^*$, the derivative of the risk is zero:

$$-q + \mathcal{F}_Y(a^*) = 0. \tag{11}$$

Thus, the optimal constant model a^* is the q-quantile of the random variable Y:

$$a^* = \mathcal{F}_Y^{-1}(q) = \mathbb{Q}_q[Y]. \tag{12}$$

For a pair $(\boldsymbol{x}, y)$ taken from the joint distribution $f(\boldsymbol{x}, y)$, a function $\hat{y}(\boldsymbol{x}) = a^*(\boldsymbol{x})$ that minimizes $\mathcal{R}(a = a^*)$ can be found by minimizing $\mathcal{R}(a)$:

$$\begin{aligned}\mathcal{R}(a) &:= \mathbb{E}_{(\boldsymbol{x},y)\sim f(\boldsymbol{x},y)}\left[\mathcal{L}_q(a(\boldsymbol{x}), y)\right] \\ &= \mathbb{E}_{(\boldsymbol{x},y)\sim f(\boldsymbol{x},y)}\left[\mathcal{L}_q(y - a(\boldsymbol{x}))\right] \\ &= \int \mathcal{L}_q(y-a(\boldsymbol{x}))\cdot f(\boldsymbol{x},y)\cdot \mathrm{d}\boldsymbol{x}\,\mathrm{d}y \\ &= \int \mathcal{L}_q(y-a(\boldsymbol{x}))\cdot \mathrm{d}\mathcal{F}(\boldsymbol{x},y) \to \min_a\end{aligned}$$

Implications: We assumed that a is a constant function of $\boldsymbol{x}$ and derived the optimal constant model $\hat{y}(\boldsymbol{x}) = a^*$ that minimizes the empirical risk (expected check-loss) $\mathcal{R}(a)$. Notably, if we differentiate $\mathcal{R}(a)$ with respect to any general function $a(\boldsymbol{x})$, the result remains the same.

Minimizing the check loss $\mathcal{L}_q(\varepsilon)$ for a regression model $\hat{y}(\boldsymbol{x})$ is equivalent to finding the q-quantile of the random variable Y. Therefore, the algorithm a^* derived from solving the minimization problem $\mathcal{R} = \mathbb{E}[\mathcal{L}_q] \to \min$ effectively predicts the q-quantile of Y.

Some implications of minimizing (8):

- $\mathbb{E}[Y] = \arg\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{x}\in X^\ell} \{y(\boldsymbol{x}) - \hat{y}(\boldsymbol{x}|\boldsymbol{\theta})\}^2$
- $\operatorname{med} Y = \arg\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{x}\in X^\ell} |y(\boldsymbol{x}) - \hat{y}(\boldsymbol{x}|\boldsymbol{\theta})|$
- $\mathbb{Q}_q[Y] = \arg\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{x}\in X^\ell} \mathcal{L}_q(y(\boldsymbol{x}) - \hat{y}(\boldsymbol{x}|\boldsymbol{\theta}))$

Quantile parameter q: By using the check loss $\mathcal{L}_q(\varepsilon)$, we can train a regression model $\hat{y}_q(\boldsymbol{x})$ that predicts the q-quantile of the random variable Y given the input $\boldsymbol{x}$:

$$\hat{y}_q(\boldsymbol{x}) = \mathbb{Q}_q[Y|X = \boldsymbol{x}], \tag{13}$$

where $\hat{y}_q$ **depends both on hyperparameter q** and on the input $\boldsymbol{x}$. This means that **predictions $\hat{y}_q(\boldsymbol{x})$ are different for different values of q.**

Likewise, the error term (residual) depends on q:

$$\varepsilon_q(\boldsymbol{x}) = \mathbb{Q}_q[Y|X = \boldsymbol{x}] - y(\boldsymbol{x}) = \hat{y}_q - y, \tag{14}$$

and the check loss in (8) is actually $\mathcal{L}_q(\varepsilon) \equiv \mathcal{L}_q(\varepsilon_q)$.

3 Expectation $\mathbb{E}$ and median $\mathbb{Q}_{1/2}$

Minimization of MSE: The expectation $\mathbb{E}[Y]$ is the average value of a random variable Y. It can be found by minimizing quadratic loss (MSE):

$$\mathbb{E}[Y|X=\boldsymbol{x}^*] = \arg\min_a \mathbb{E}[(Y-a(X))^2 \mid X=\boldsymbol{x}^*], \tag{15}$$

which holds for both conditional $\mathbb{E}[Y|X]$ and unconditional $\mathbb{E}[Y]$ expectations.

The algorithm a^* that minimizes the average quadratic loss has the lowest MSE among all possible estimators and sometimes is called the *minimum mean squared error* (MMSE) estimator, which is more commonly known as the *least squares* (LS) estimator.

In other words, minimization of quadratic loss is **one of (many) possible ways** to find a good model a^*. During training, this model learns how to predict conditional expectation $\mathbb{E}[Y|X=\boldsymbol{x}^*]$ for a given $\boldsymbol{x}^*$, then we use it to predict the expectation $\mathbb{E}[Y|X=\boldsymbol{x}']$ for previously unseen data points $\boldsymbol{x}'$.

This estimator has good theoretical guarantees (*e.g.*, unbiasedness, minimum variance, *etc.* under certain conditions) and because of that is the first choice for most regression problems.

Differentiating the quadratic loss with respect to a gives:

$$\begin{aligned}
&\frac{\partial}{\partial a}\mathbb{E}[(Y-a(X))^2 \mid X=\boldsymbol{x}^*] \\
&= \frac{\partial}{\partial a}\mathbb{E}[Y^2 - 2Ya(X) + a(X)^2 \mid X=\boldsymbol{x}^*] \\
&= \mathbb{E}[-2Y + 2a(X) \mid X=\boldsymbol{x}^*] \\
&= -2\mathbb{E}[Y \mid X=\boldsymbol{x}^*] + 2a(\boldsymbol{x}^*) = 0
\end{aligned}$$

Rearranging gives:

$$a(\boldsymbol{x}^*) = \mathbb{E}[Y|X=\boldsymbol{x}^*]$$

Minimization of MAE: An alternative estimator a is obtained when instead of minimizing the quadratic term ε^2, we replace it with absolute difference $|\varepsilon|$, which is equivalent to minimizing mean absolute error (MAE). Interestingly, this gives us the median of the random variable Y:

$$\mathbb{Q}_{1/2}[Y|X=\boldsymbol{x}^*] = \arg\min_a \mathbb{E}[|Y-a(X)| \mid X=\boldsymbol{x}^*] \tag{16}$$

Indeed, MAE is directly connected to the quantile loss $\mathcal{L}_q(\varepsilon)$. For $q=1/2$, the quantile loss is simply the absolute value of the error ε (we ignore $1/2$ factor):

$$\mathcal{L}_{1/2}(\varepsilon) = \begin{cases} 1/2\cdot\varepsilon & \varepsilon \geq 0 \\ -1/2\cdot\varepsilon & \varepsilon < 0 \end{cases} = \frac{|\varepsilon|}{2} \tag{17}$$

For $q=1/2$, the quantile $\mathbb{Q}_{1/2}[Y]$ corresponds to the value y^* such that $\mathbb{P}[Y \leq y^*] = 1/2$; *i.e.*, the value y^* cuts the distribution of Y in half. This is what the median ($Q_{1/2}$) of a random variable Y is.

Given training data $(\boldsymbol{x}^*, y^*) \in (X,Y)^\ell$, the empirical estimation according to (15) and (16) can be expressed as:

$$\mathcal{R}(a) = \frac{1}{\ell}\cdot\sum_{(\boldsymbol{x}^*,y^*)\in(X,Y)^\ell} \underbrace{\mathcal{L}(y^*-a(\boldsymbol{x}^*))}_{\substack{(Y-a(X))^2 \mid X=\boldsymbol{x}^* \\ \text{or} \\ |Y-a(X)| \mid X=\boldsymbol{x}^*}} \to \min_a.$$

This model is also referred to as the Least Absolute Deviations (LAD) estimator.

Laplace distribution: Quadratic loss is derived from assuming a Gaussian distribution of Y. Formally, absolute loss comes from assuming a Laplace distribution of Y:

$$f_Y(y) = \frac{1}{2b}\cdot e^{-\frac{|y-\mu|}{b}}, \tag{18}$$

where b is the scale parameter and μ is the mean. As the Laplace distribution is symmetric, the mean $\mathbb{E}[Y]$ is equal to the median $\mathbb{Q}_{1/2}[Y]$.

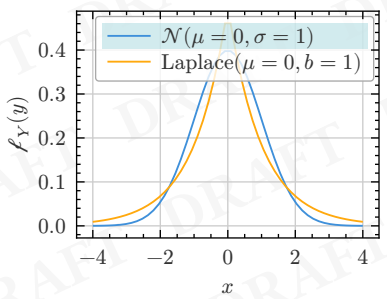


Figure 2: PDF of standard Normal and Laplace distributions. Laplace has heavier tails.

Likelihood: Assuming observations $(\boldsymbol{x}^*, y^*)$ are i.i.d. and algorithm a predicts the conditional mean $\mu = \mathbb{E}[Y|X=\boldsymbol{x}^*] = \mathbb{Q}_{1/2}[Y|X=\boldsymbol{x}^*]$, the likelihood function is given by:

$$\mathbb{L} = \prod_{(\boldsymbol{x}^*,y^*)\in(X,Y)^\ell} f_Y(y^*|\boldsymbol{x}^*) = \prod_{(\boldsymbol{x}^*,y^*)\in(X,Y)^\ell} \frac{1}{2b}\cdot e^{-|y^*-a(\boldsymbol{x}^*)|/b}, \tag{19}$$

Maximizing the likelihood function is equivalent to minimizing MAE:

$$\begin{aligned}
\log\mathbb{L} &= \sum_{(\boldsymbol{x}^*,y^*)\in(X,Y)^\ell}\left(-\log 2b - \frac{|y^*-a(\boldsymbol{x}^*)|}{b}\right) \\
&= -\frac{1}{b}\underbrace{\sum_{(\boldsymbol{x}^*,y^*)\in(X,Y)^\ell}|y^*-a(\boldsymbol{x}^*)|}_{\ell\cdot\text{MAE}} - \underbrace{\ell\cdot\log 2b}_{\text{const}} \to \max_a
\end{aligned} \tag{20}$$

Median is more robust to outliers than mean:

- In an ordered sample $\{y_1,\ldots,y_\ell\}$, adding an outlier y' shifts the mean **proportionally** to its magnitude:
$$\Delta\mathbb{E}[Y] = \frac{y'}{\ell+1}.$$
- In the worst case an **extreme** outlier $y' \ll y_1$ or $y' \gg y_\ell$ can only shift the median to the adjacent element:
$$y^{(\ell/2-1)} - y^{(\ell/2)} \leq \Delta\mathbb{Q}_{1/2}[Y] \leq y^{(\ell/2+1)} - y^{(\ell/2)}.$$

Measures of central tendency: Expectation $\mathbb{E}[Y]$ and median $\mathbb{Q}_{1/2}[Y]$ are two distinct measures of central tendency for a random variable Y. In some cases, they are equal, but in general they are not. This distinction leads to two different regression models:

Quantile regression is not limited to $q=1/2$; we can construct a regression model for any conditional quantile $\mathbb{Q}_q[Y|X]$

- In ordinary least squares (LS), we predict the expected value of a random variable:

$$\hat{y}(\boldsymbol{x}) = \mathbb{E}[Y|X = \boldsymbol{x}]. \tag{21}$$

- In median regression, we build a model that predicts the conditional median:

$$\hat{y}(\boldsymbol{x}) = \mathbb{Q}_{1/2}[Y|X = \boldsymbol{x}]. \tag{22}$$

4 Quantile regression

Quantile regression was introduced by Roger Koenker and Gilbert Bassett in (Koenker & Bassett, 1978).

For a short overview and examples see (Koenker & Hallock, 2001) and (Koenker, 2005) for details.

Probabilistic model: Suppose the distribution of the data $(\boldsymbol{x}, y)$ is modeled as a joint distribution $\mathscr{f}(\boldsymbol{x}, y)$. Our goal is to predict the quantile $\mathbb{Q}_q[Y] = (\cdot)(\boldsymbol{x})$ for a given $\boldsymbol{x}$, *i.e.*, to predict the conditional quantile $\mathbb{Q}_q[Y|X = \boldsymbol{x}]$.

Optimization problem: The empirical risk is defined as the average quantile loss (8) over the distribution $\mathscr{f}(\boldsymbol{x}, y)$. By minimizing the empirical risk, we can find the optimal model $a^*(\boldsymbol{x})$ that predicts the quantile $\mathbb{Q}_q[Y|X = \boldsymbol{x}]$:

$$a^*(\boldsymbol{x}) = \arg\min_a \underbrace{\mathbb{E}_{(\boldsymbol{x},y)\sim\mathscr{f}(\boldsymbol{x},y)}\big[\mathcal{L}_q(y - a(\boldsymbol{x}))\big]}_{\mathcal{R}(a)}. \tag{23}$$

Practical reformulation: From the theoretical expression of the empirical risk, we can derive a practical reformulation of the quantile regression problem.

For a specific pair $(\boldsymbol{x}^*, y^*)$ drawn from the joint distribution $\mathscr{f}(\boldsymbol{x}, y)$ represented by a training set $(X, Y)^\ell$, the empirical risk can be expressed via the check loss (8):

$$\mathcal{R}(a) = \frac{1}{\ell} \cdot \sum_{(\boldsymbol{x}^*,y^*)\in(X,Y)^\ell} \mathcal{L}_q(y^* - a(\boldsymbol{x}^*)) \to \min_a. \tag{24}$$

The model $a(\boldsymbol{x}) \equiv a(\boldsymbol{x}|\boldsymbol{\theta}; q)$ can be any general regression model supporting custom loss functions or the quantile loss $\mathcal{L}_q$ specifically.

In LS regression, only one prediction $\hat{y}(\boldsymbol{x}) = \mathbb{E}[Y|X = \boldsymbol{x}]$ exists, with a single residual $\varepsilon(\boldsymbol{x}) \coloneqq y(\boldsymbol{x}) - \hat{y}(\boldsymbol{x})$.

In quantile regression, $\hat{y}_q(\boldsymbol{x})$ is parameterized by q, producing multiple possible predictions $\mathbb{Q}_q[Y|X = \boldsymbol{x}]$ for the same random variable, with corresponding residuals $\hat{\varepsilon}_q(\boldsymbol{x}^*) \coloneqq \hat{y}_q(\boldsymbol{x}^*) - y(\boldsymbol{x}^*)$

Linear quantile regression: The conditional quantile $\mathbb{Q}_q[Y|X]$ can be modeled as a linear function of predictors $\boldsymbol{x}$:

$$\mathbb{Q}_q[Y|X = \boldsymbol{x}] = \langle \boldsymbol{x}, \boldsymbol{\beta}\rangle, \quad \beta_j \equiv \beta_j(q) \tag{25}$$

where $\boldsymbol{\beta}(q)$ is a vector of regression coefficients, and $\beta_j(q) = \beta_{j|q} \in \mathbb{R}$ are regression coefficients for the feature $\boldsymbol{x}^j$ and a *predefined hyperparameter* q. Coefficients $\beta_j(q)$ are estimated by minimizing the empirical risk:

$$\mathcal{R}(\boldsymbol{\beta}) = \frac{1}{\ell} \cdot \sum_{\boldsymbol{x}\in X^\ell} \mathcal{L}_q(y(\boldsymbol{x}) - \langle \boldsymbol{x}, \boldsymbol{\beta}\rangle) \to \min_{\boldsymbol{\beta}}. \tag{26}$$

Neural quantile regression: Neural networks inherently support custom loss functions and can model conditional quantiles $\mathbb{Q}_q[Y|X]$ as well (Figure 4). A model predicting conditional quantiles $\mathbb{Q}_q[Y|X]$ must be trained with a quantile loss, which can be easily implemented:

```
class QuantileLoss(L.LightningModule):
    def __init__(self, q: float):
        super().__init__()
        self.q = q

    def forward(self, y_pred, y_true):
        epsilon = y_true - y_pred
        return T.where(
            epsilon >= 0,
            self.q * epsilon,
            (self.q - 1) * epsilon,
        ).mean()
```

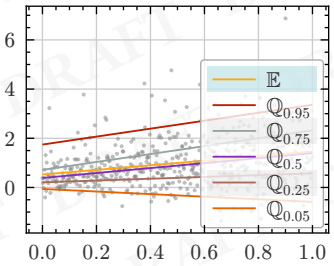


Figure 3: Linear quantile regression for non-normaly distributed noise

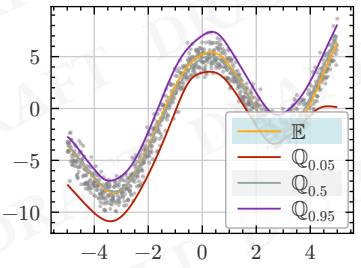


Figure 4: Quantile regression performed by a neural network

Gradient boosting quantile regression: Quantile loss (8) is differentiable if $\varepsilon \neq 0$:

$$\frac{\partial}{\partial\varepsilon}\mathcal{L}_q(\varepsilon) = \begin{cases} q & \varepsilon > 0 \\ -(1-q) & \varepsilon < 0 \end{cases}, \tag{27}$$

thus, gradient boosting can approximate the quantile function $\mathbb{Q}_q[Y|X]$ to handle non-linear dependencies between features and quantiles (Figure 5).

Ensemble models: Multiple base algorithms $a_t(\boldsymbol{x})$ can be combined to create an ensemble model

$$A(\boldsymbol{x}) = \frac{1}{T} \cdot \sum_{t=1}^{T} a_t(\boldsymbol{x}). \tag{28}$$

If each base algorithm $a_t(\boldsymbol{x})$ is trained to predict quantiles $\mathbb{Q}_q[Y|X]$, the ensemble $A(\boldsymbol{x})$ will estimate the expectation of the quantile $\mathbb{E}\big[\mathbb{Q}_q[Y|X]\big]$.

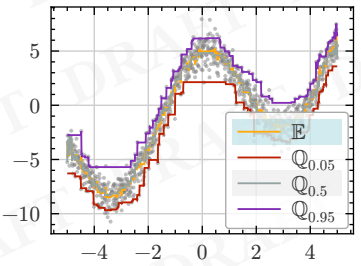


Figure 5: Quantile regression performed by a gradient boosting model

5 Convergence and reliability of quantile regression parameters

Linear quantile regression: For linear quantile regression (25), the conditional quantile $\mathbb{Q}_q[Y|X]$ is modeled as a linear function of predictors $\boldsymbol{x}$. The theoretical properties of linear quantile parameters $\hat{\boldsymbol{\beta}}(q)$ such as convergence and variance can be derived, though the analysis is more complex than for traditional Gaussian regression.

Parameter expectation: All regression coefficients $\beta_j = \beta_j(q)$ are functions of q. Under appropriate conditions (independent observations with finite second moments), the asymptotic distribution of the quantile regression estimator $\hat{\boldsymbol{\beta}}(q)$ is unbiased:

$$\hat{\boldsymbol{\beta}}(q) \to \boldsymbol{\beta}(q), \tag{29}$$

i.e., theoretically the estimator $\hat{\boldsymbol{\beta}}(q)$ converges to the expected value of the parameter $\boldsymbol{\beta}(q)$ as the sample size ℓ approaches infinity:

$$\hat{\boldsymbol{\beta}}(q) \to \mathbb{E}[\boldsymbol{\beta}(q)]). \tag{30}$$

Parameter variance: The estimator $\hat{\boldsymbol{\beta}}(q)$ is asymptotically normally distributed with variance and mean $\boldsymbol{\beta}(q) = \mathbb{E}[\boldsymbol{\beta}(q)]$ according to (30)

$$\mathbb{D}[\boldsymbol{\beta}(q)] \to \underbrace{\frac{1}{\ell}}_{\text{I}} \cdot \underbrace{q \cdot (1-q)}_{\text{II}} \cdot \underbrace{D^{-1}\Omega D^{-1}}_{\text{III}}. \tag{31}$$

The variance in (31) depends on three terms:

1. The 1st multiplier determines the convergence rate of the estimator $\hat{\boldsymbol{\beta}}(q)$ as a function of the sample size ℓ; the larger the sample size, the smaller the variance.
2. The 2nd multiplier depends on the quantile q. As q approaches the tails (0 or 1), this term decreases, which would seemingly lower the variance. It reduces variance if isolated, however, this is not the primary contributor to overall variance.
    1. The 3rd multiplier is the sandwich variance estimator, which depends on both the estimated parameters $\hat{\boldsymbol{\beta}}(q)$ and the robust variance matrix Ω. Typical formulations include:

$$D(\hat{y}_q) = \frac{1}{\ell} \cdot \sum_{\boldsymbol{x} \in X^\ell} \hat{f}_{Y|X}(\hat{y}_q(\boldsymbol{x})) \cdot \boldsymbol{x}\boldsymbol{x}^\mathsf{T} \tag{32}$$

$$\hat{\Omega} = \frac{1}{\ell} \cdot \sum_{\boldsymbol{x} \in X^\ell} (q - [\![y(\boldsymbol{x}) \le \hat{y}_q(\boldsymbol{x})]\!]) \cdot \boldsymbol{x}\boldsymbol{x}^\mathsf{T} \tag{33}$$

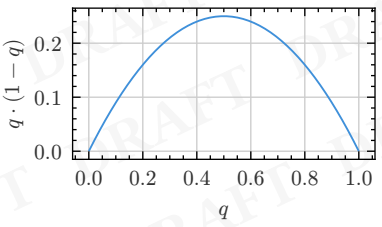


Figure 6: $q \cdot (1-q)$ term in (31) reaches its maximum at $q = 0.5$

Consequently, the **variance of estimated parameters $\hat{\boldsymbol{\beta}}(q)$ increases as q approaches 0 or 1**. In practice, predictions near the median are typically more precise, while predictions for extreme quantiles (*e.g.*, 0.01 or 0.99) are less reliable.

While $q \cdot (1-q)$ decreases near the tails, the sandwich term $D^{-1}\Omega D^{-1}$ becomes poorly estimated and tends to dominate.

Bad statistical guarantee: While ordinary least squares (OLS) estimates benefit from the Gauss-Markov theorem, which establishes OLS as the best linear unbiased estimator (BLUE) under classical assumptions, quantile regression follows different asymptotic properties.

Quantile regression estimators remain unbiased and consistent, but their variance behavior is more complex. As shown in equation (31), the variance depends on both the quantile level q and the underlying data distribution through the sandwich estimator term $D^{-1}\Omega D^{-1}$.

In practice, quantile regression estimates exhibit higher statistical variability than OLS estimates, particularly for extreme quantiles (*e.g.*, $q < 0.1$ or $q > 0.9$). This occurs because:

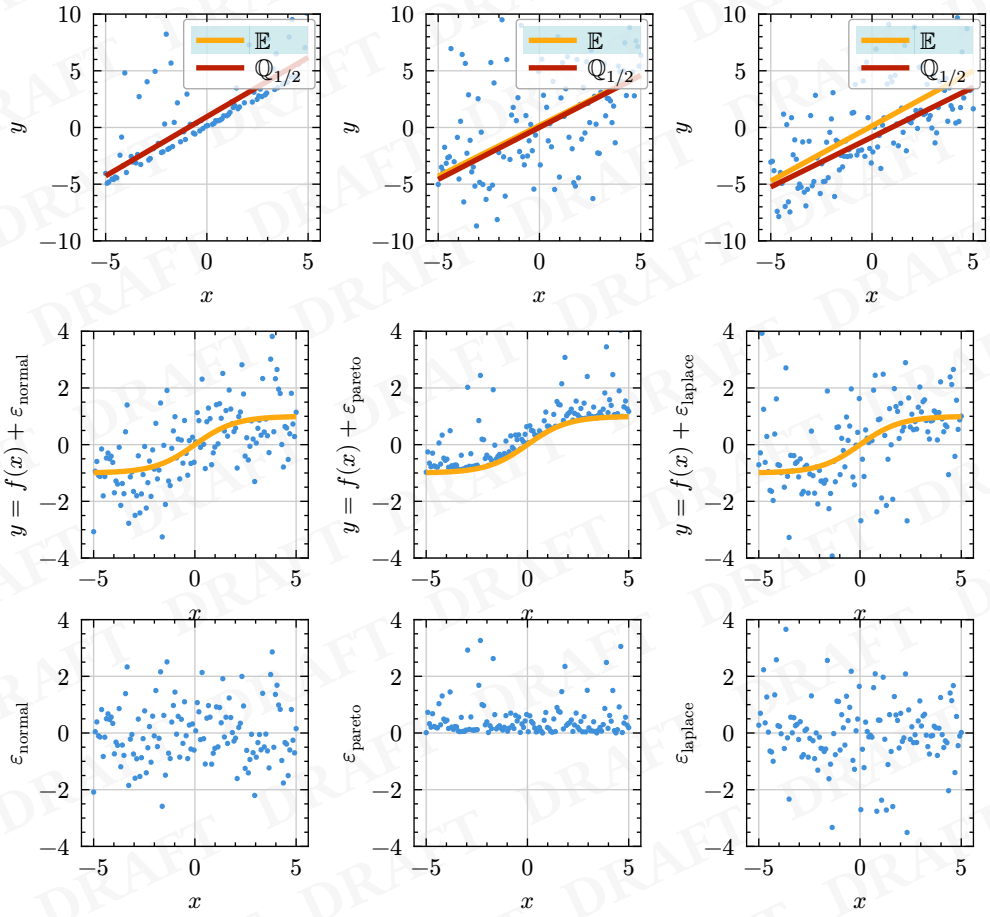
1. The sparsity of data in the tails leads to less reliable sandwich term estimation
2. The conditional density at extreme quantiles becomes more difficult to estimate accurately
3. The effective sample size for determining extreme quantiles is effectively reduced

This statistical efficiency trade-off is a necessary cost of gaining robustness to outliers and insights into the complete conditional distribution.

The variance of the quantile regression estimator is larger than that of OLS, especially for extreme quantiles.

6 Robustness of quantile regression

Non-normality (skew, heavy tails, multimodality): Quantile regression models **conditional quantiles**, capturing skewed or heavy-tailed distributions **without relying on normality assumptions**. OLS assumes normality and may produce misleading results when this assumption is violated.



Heteroscedasticity: Quantile regression does not assume homoscedasticity (constant variance). Instead, it models different parts of the conditional distribution independently, allowing for varying spread (*e.g.*, wider or narrower intervals) across predictors. OLS assumes homoscedasticity (or equal weight of all observations).

Robustness to outliers and noise: By focusing on quantiles rather than the mean, quantile regression reduces sensitivity to random noise and outliers, emphasizing specific distributional trends. Quantile regression also **does not assume any specific noise distribution**. In OLS, a few outliers can have a pronounced effect on parameter estimates.

Censoring: Censoring arises when the response variable y is not fully observed. For instance, in clinical trials, the exact value of y may be unavailable for some patients. If a patient exits a longitudinal study, we only know they survived up to time y, but their true survival time might be much longer (Figure 7).

In standard Gaussian regression, censoring results in a bias in estimates since observations are truncated. Quantile regression allows to model different quantiles q of the distribution: some areas of the distribution may not be affected by censoring, while others are. By choosing appropriate quantiles, we can obtain reliable estimates even in the presence of censoring.

However, in general it's better to experiment with different losses and other models, that work with censoring implicitly.

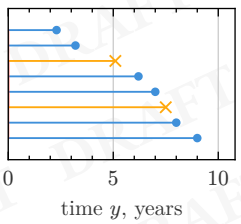


Figure 7: Time-to-death plot from the start of a clinical trial. Circles represent patients whose exact time-to-death is known, while crosses represent patients who withdrew from the study.

Invariance: Quantile regression is **invariant to monotonic transformations** of y like logarithm or square root. In OLS this is not the case, although transformations are sometimes used to normalize data.

7 Interpretation of linear quantile coefficients $\hat{\beta}_j(q)$

Impact on target variable: Quantile regression coefficients $\beta_j(q)$ represent the impact of a unit change in predictor $\boldsymbol{x}^j$ on the response variable at specific quantiles. Unlike OLS coefficients, they capture how features influence different parts of the target distribution.

By examining how $\hat{\beta}_j(q)$ varies across quantile levels, we can guess how predictors affect various segments of the conditional distribution, revealing effects that are not directly observable in standard regression.

Data: The ACTG 320 clinical trial, initiated in 1997 by Merck, was designed to evaluate the effectiveness of the antiretroviral drug indinavir when used in a triple-drug regimen compared to a standard two-drug treatment for HIV patients.

Variable	Description
`time` (target)	Follow-up time to AIDS progression or death (in days). Represents the time from enrollment to the event (end of study or death).
`age`	Age of the patient at the time of enrollment (in years).
`cd4_cell_count`	Baseline CD4 T-cell count (cells/mL), a key indicator of immune function.
`race_*`	Indicator variables representing the patient's race.
`group_*`	Indicator variables representing the treatment group.

Table 1: ACTG 320 dataset features (simplified)

The associated dataset contains aprox. 1,150 records of HIV-infected patients who were randomized to receive either the novel triple-drug regimen or the conventional two-drug therapy.

Quantile regression: The target variable is `time`, representing the follow-up duration. Linear quantile regression

$$\mathbb{Q}_q[Y|X] = \sum_j \beta_j \cdot \boldsymbol{x}^j, \quad \beta_j \equiv \beta_j(q) \tag{34}$$

was used to estimate the impact of various linear predictors $\boldsymbol{x}^j$ from Table 1 on the time y to AIDS progression or death.

Quantile regression coefficients $\beta_j(q)$ as functions of quantile q are plotted in Figure 8. Low q values represent individuals who progressed to AIDS or died quickly, while high q values correspond to individuals with longer survival times.

Baseline estimate: Baseline survival time is estimated by the model intercept (Figure 8a), *e.g.*, median intercept $\beta_{\text{intercept}}(q = 1/2)$ is approximately 240 days. Note that the intercept would be the median survival time if all other predictors were zero, in our case, they are not.

Reliability of coefficients: For $q \approx 0.5$, the estimates are most reliable and often close to the OLS estimates. Extreme quantiles are estimated at tails where data is sparse, leading to higher variance $\mathbb{D}\big[\hat{\beta}_j\big]$ and less reliable estimates, as seen in the fluctuations in Figure 8e at both tails. Quantiles $q < 0.1$ and $q > 0.9$ were not estimated at all.

Sign of quantile regression coefficients $\beta_j(q)$: The sign of $\beta_j(q)$ reflects the predictor $\boldsymbol{x}^j$'s impact on survival time y at the q-quantile, *i.e.*, $\mathbb{Q}_q[Y] \propto \beta_j(q) \cdot \Delta\boldsymbol{x}^j$ at the q-quantile.

Consistently positive $\beta_j(q)$ across all q suggest that the predictor $\boldsymbol{x}^j$ has only positive contributions to survival time y for all individuals. For the indinavir group (Figure 8b), the positive impact (in days) is greatest for short-survived patients (low q) and decreases for long-lived patients (high q).

Likewise, consistently negative $\beta_j(q)$ across all q suggest that the predictor $\boldsymbol{x}^j$ has only negative contributions to survival time y for all quantiles. AIDS patients generally have lower CD4 cell counts than healthy individuals, and the lower the CD4 cell count, the more pronounced its negative contribution (Figure 8d) to survival time y.

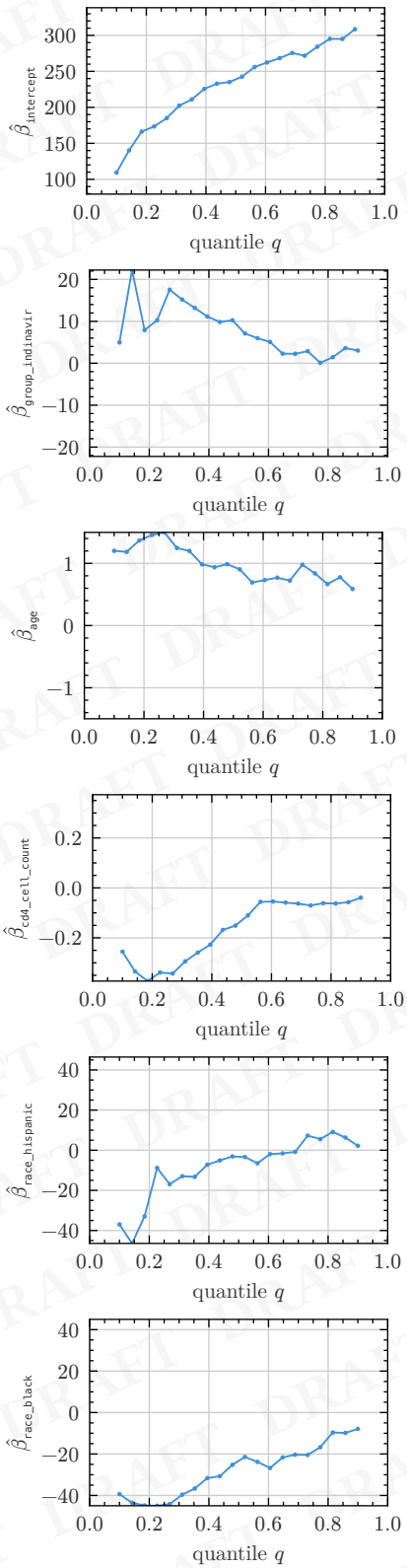


Figure 8: Quantile regression coefficients for ACTG 320 dataset

Check (Koenker & Hallock, 2001) for more examples.

8 Practical considerations

Targets: Median is sometimes more interpretable and a better measure of centrality than the mean, particularly for skewed or multimodal data:

- Median salary or house price characterizes the central tendency of a distribution better than the mean, which can be skewed by extreme values.

Parameters:

- Coefficients β in linear quantile regression are noisier than in OLS and depend on quantile q, making them harder to interpret. The Gauss-Markov theorem ensuring convergence and variance in OLS does not apply to quantile regression.
- Exact values of β in OLS are interpretable, but in quantile regression, they are generally not. In simple cases, they can be close to OLS coefficients and interpretable. However, when quantile regression is applied to transformed data (*e.g.*, $\log(y)$), coefficients remain invariant, but their contribution to $y' = \log(y)$ becomes less obvious. For skewed data where OLS fails, quantile regression coefficients differ significantly from OLS but may still be interpretable.

Computational complexity: Quantile regression lacks a universal analytical solution and is typically solved numerically. The quantile loss function from (8) combines two linear functions separated at $\varepsilon = 0$. Residuals can be decomposed into positive and negative parts:

$$\varepsilon = \varepsilon^+ - \varepsilon^-, \quad \text{where} \quad \begin{cases} \varepsilon^+ := \max\{0, \varepsilon\} \\ \varepsilon^- := -\min\{0, \varepsilon\} \end{cases}, \tag{35}$$

Using this decomposition, the quantile loss can be expressed as:

$$\mathcal{L}_q(\varepsilon) = q \cdot \varepsilon^+ + (1-q) \cdot \varepsilon^-. \tag{36}$$

This formulation leads to a constrained linear programming problem [(Koenker et al., 2018), p.282]:

$$\begin{aligned} & \frac{1}{\ell} \sum_{i=1}^{\ell} \{q \cdot \varepsilon_i^+ + (1-q) \cdot \varepsilon_i^-\} \to \min_{\varepsilon^+, \varepsilon^-} \\ \text{s.t.} \quad & y_i - \hat{y}_i = \varepsilon_i^+ - \varepsilon_i^-, \quad i = 1..\ell, \\ & \varepsilon_i^+ \geq 0, \quad \varepsilon_i^- \geq 0, \quad i = 1..\ell. \end{aligned} \tag{37}$$

Solving this optimization problem is computationally more intensive than OLS's closed-form solution, particularly for large datasets or when estimating multiple quantiles simultaneously.

Extreme quantiles: Estimates for extreme quantiles (*e.g.*, $q = 0.01$ or $q = 0.99$) are often less reliable due to sparse data in distribution tails, resulting in higher variance as shown in the parameter convergence section.

Complete picture of conditional distributions: Quantile regression allows modeling multiple quantiles, providing a comprehensive view of how predictors affect the entire conditional distribution of the response, not just its center. This reveals heterogeneous effects that OLS cannot capture.

9 Goodness-of-fit

Bad metrics: Classical metrics (*e.g.*, MAE, MSE, R^2) evaluate predictions based on their distribution around the mean $\mathbb{E}[Y]$. However, quantile regression focuses on other distribution properties, intentionally ignoring the mean. As a result, classical metrics are not suitable for evaluating quantile regression models.

Mean quantile loss: The simplest approach to evaluate quantile regression models is to use the quantile loss $\mathcal{L}_q$ directly:

$$\langle \mathcal{L}_q \rangle \coloneqq \frac{1}{\ell} \cdot \sum_{(\boldsymbol{x},y)\in(X,Y)^\ell} \mathcal{L}_q\big(y - \hat{y}_q(\boldsymbol{x})\big), \tag{38}$$

where $\hat{y}_q(\boldsymbol{x})$ is the quantile regression model.

For two quantile regression models $\hat{y}_q$ and $\hat{y}'_{q'}$ (*e.g.*, for different quantiles, regularization, or features), the model with the lower quantile loss better fits the data and is preferred. In `sklearn`, this metric is implemented as `sklearn.metrics.mean_pinball_loss`.

In fact, MAE is equivalent to the mean quantile loss for $q = 1/2$, making it suitable for **median regression specifically**

R^1 metric: Another approach involves metrics specifically designed for quantile regression. Classical R^2 measures the proportion of variance explained by the model:

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} \;\rightarrowtail\; 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}, \tag{39}$$

where RSS is the sum of squared residuals between the predicted and actual values, and TSS is the squared difference between the actual values $y(\boldsymbol{x})$ and the mean $\bar{y} = \mathbb{E}[Y]$. TSS can be viewed as RSS for a very simple constant model $\hat{y}(\boldsymbol{x}) = \bar{y}$:

$$R^2 = 1 - \frac{\mathrm{RSS}[\hat{y}]}{\mathrm{RSS}[\bar{y}]}, \tag{40}$$

where $\mathrm{RSS}[\hat{y}]$ and $\mathrm{RSS}[\bar{y}]$ are the residual sum of squares for the actual (proposed) model $\hat{y}$ and the mean constant (baseline) model $\bar{y}$, respectively.

For unbiased models, $\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$, which is used to derive (39).

The choice of the baseline model is **arbitrary**, so a pseudo-R^2 metric can be used to compare RSS of any two arbitrary models $\hat{y}$ and $\hat{y}'$.

For quantile regression, a similar metric can be defined ((Koenker & Machado, 1999), eq. 7). For two quantile regression models $\hat{y}_q$ and $\hat{y}'_{q'}$ and corresponding mean quantile losses $\langle \mathcal{L}_q\big[\hat{y}_q\big] \rangle$ and $\langle \mathcal{L}'_{q'}\big[\hat{y}'_{q'}\big] \rangle$ computed via (38), the analog of R^2 is:

$$R^1 \coloneqq 1 - \frac{\langle \mathcal{L}_q\big[\hat{y}_q\big] \rangle}{\langle \mathcal{L}'_{q'}\big[\hat{y}'_{q'}\big] \rangle}, \tag{41}$$

where $\hat{y}_q$ and $\hat{y}'_{q'}$ are the proposed and baseline models, respectively. Usually, models are compared for the same quantile q.

A straightforward choice for the baseline model $\hat{y}'_q$ is the empirical quantile value $\mathbb{Q}_q[Y]$ calculated from the training set. The difference in the upper index arises because in R^2, quadratic units (TSS, RSS, and ESS) are used, while in R^1, linear units (quantile loss) are used, as seen in (8).

Like the general definition of R^2, which is not bound to $[0, 1]$ and can be negative, the R^1 metric can also be negative if the model $\hat{y}_q$ is worse than the baseline model $\hat{y}'_q$. In `sklearn`, this metric is implemented as `sklearn.metrics.d2_pinball_score`.

Ordered metrics: A quantile model can be evaluated on how well it preserves the order. This is particularly important for risk modeling applications and ranking. For example, if a patient $\boldsymbol{x}$ died at $y(\boldsymbol{x})$ and another at $y(\boldsymbol{x}')$, where $y(\boldsymbol{x}) < y(\boldsymbol{x}')$, the model should predict $\hat{y}_q(\boldsymbol{x}) < \hat{y}_q(\boldsymbol{x}')$.

The proportion of correctly ordered pairs is measured by the *concordance index* (C-index):

In (42), the numerator counts the number of observation pairs where the model predicts the same order as the actual order. The denominator counts the total number of comparable observation pairs. Equation (42) can also be extended to handle ...

$$C := \frac{\sum_{i=1}^{\ell} \sum_{j=1}^{\ell} [\![y(\boldsymbol{x}_i) < y(\boldsymbol{x}_j)]\!] \cdot [\![\hat{y}_q(\boldsymbol{x}_i) < \hat{y}_q(\boldsymbol{x}_j)]\!]}{\sum_{i=1}^{\ell} \sum_{j=1}^{\ell} [\![y(\boldsymbol{x}_i) < y(\boldsymbol{x}_j)]\!]}, \tag{42}$$

The C-index ranges from 0.5 (random predictions) to 1 (perfect predictions).