Machine learning and statistical learning

2.2 Quantile Regression

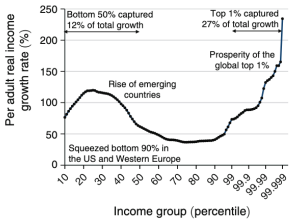
Ewen Gallic
ewen.gallic@gmail.com

MASTER in Economics - Track EBDS - 2nd Year

2020-2021

1. Quantiles

Quantiles



Source: Alvaredo et al. (2018)

Figure 1: Total Income Growth by Percentile Across All World Regions, 1980-2016

- Quantiles are used to **describe a distribution**
- Very useful for highly **dispersed distributions**
  - *e.g.*, with income
- Useful if we are interested in **relative variations** (rather than **absolute variations**)

Some references

- Arellano, M (2009). *Quantile methods*, Class notes
- Charpentier, A (2020). *Introduction à la science des données et à l'intelligence artificielle*, https://github.com/freakonometrics/INF7100
- Charpentier, A. (2018). *Big Data for Economics*, Lecture 3
- Davino et al. (2014). *Quantile Regression*. John Wiley & Sons, Ltd.
- Givord, P., D'Haultfoeuillle, X. (2013). *La régression quantile en pratique*, INSEE
- He, X., Wang, H. J. (2015). *A Short Course on Quantile Regression*.
- Koenker and Bassett Jr (1978). *Regression quantiles*. Econometrica: journal of the Econometric Society (46), 33–50
- Koenker (2005). *Quantile regression*. 38. Cambridge university press.

1.1 Median

Median: a specific quantile

- The **median** m of a real-valued probability distribution separates this probability distribution into two halves
  - 50% of the values below m, 50% of the values above m
- For a random variable Y with cumulative distribution F, the median is such that:

$$\mathbb{P}(Y \leq m) \geq \frac{1}{2} \text{ and } \mathbb{P}(Y \geq m) \geq \frac{1}{2}$$

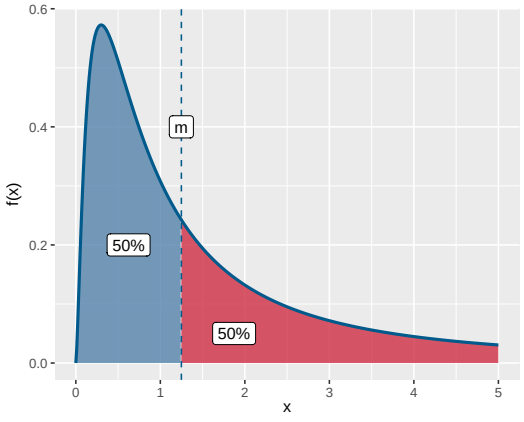


Figure 2: Probability density function of a $\mathcal{F}(5, 2)$.

Median: a specific quantile

- From the cumulative distribution function F, that gives $F(y) = \mathbb{P}(Y \leq y)$
  - find the antecedent of .5
  - if the cumulative distribution is strictly increasin, the antecedent is unique
  - if not, all antecedents of 0.5 are possible values for the median

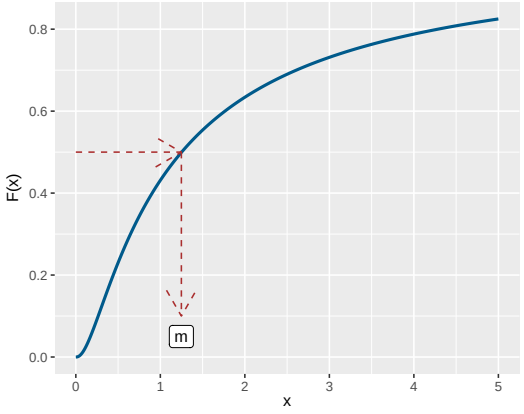


Figure 3: Cumulative distribution function of a $\mathcal{F}(5, 2)$.

Median and symmetric distribution

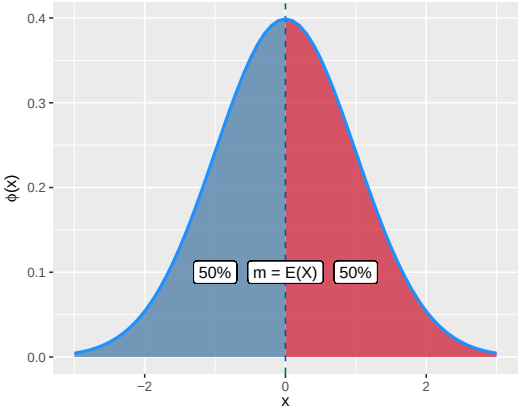


Figure 4: PDF of the $\mathcal{N}(0, 1)$ distribution.

If the probability distribution is symmetric, the median m and the mean (if it exists) are the same (the point at which the symmetry occurs).

Numeric Optimization

It may be useful to think of the median as the solution of a **minimization problem**:

- The **median** minimizes the absolute sum of deviations:

$$\text{median}(Y) \in \arg\min_m \mathbb{E} \mid Y - m \mid$$

- An analogy can be made with the **mean** μ of a random variable.
  - It can be obtained using the following numerical optimization, where m is defined as the center of the distribution:

$$\mu = \arg\min_m \mathbb{E}(Y - m)^2$$

1.2 Quantiles : definition

What is a quantile?

> *In statistics and probability, quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way. There is one fewer quantile than the number of groups created (Wikipedia)*

We thus define a **probability threshold** τ and then look for the value $Q(\tau)$ (i.e., the **τth quantile**) such that:

- a proportion τ of the observations have a value lower or equal to τ;
- a proportion $1 - \tau$ of the observations have a value greater or equal to τ.

Formal definition

The **quantile** $Q_Y(\tau)$ of level $\tau \in (0,1)$ of a random variable Y is such that:

$$\mathbb{P}\{Y \leq Q_Y(\tau)\} \geq \tau \text{ and } \mathbb{P}\{Y \geq Q_Y(\tau)\} \geq 1-\tau$$

We can also define it as:

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

The most used quantiles are:

- $\tau = 0.5$: the median
- $\tau = \{0.1, 0.9\}$: the first and last deciles
- $\tau = \{0.25, 0.75\}$: the first and last quartiles

Illustration

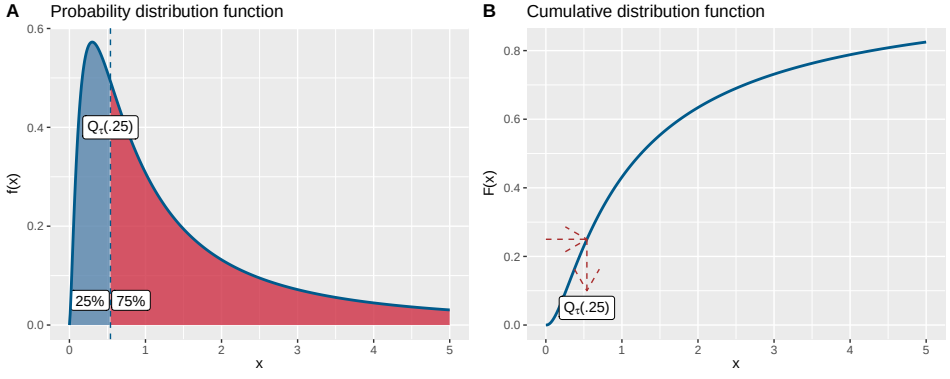


Figure 5: Quantiles of a probability distribution (here, $\mathcal{F}(5, 2)$).

Numerical optimization

Once again, *quantiles* can be viewed as particular centers of the distribution that minimize the **weighted absolute sum of deviation**. For the τth quantile:

$$Q_\tau(Y) \in \arg\min_m \mathbb{E}\left[\rho_\tau(Y-m)\right]$$

where $\rho_\tau(u)$ is a loss function defined as follows:

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

for $0 < \tau < 1$.

Loss function

This loss function is

- a continuous piecewise linear function
- non differentiable at $u = 0$

Then,

- a $(1 - \tau)$ weight is assigned to the negative deviations
- a τ weight is assigned to the positive deviations

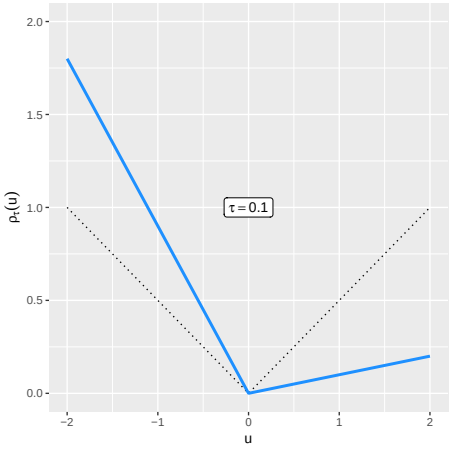


Figure 6: Quantile regression loss function ρ_τ.

Numerical optimization

The minimization program $Q_\tau(Y) = \arg\min_m \mathbb{E}\left[\rho_\tau(Y-m)\right]$ can be written as:

- For a **discrete Y variable** with pdf $f(y) = \mathbb{P}(Y=y)$:

$$Q_\tau(Y) = \arg\min_m \left\{ (1-\tau) \sum_{y \leq m} \mid y-m \mid f(y) + \tau \sum_{y>c} \mid y-m \mid f(y) \right\}$$

- For a **continuous Y variable** with pdf $f(y)$:

$$Q_\tau(Y) = \arg\min_m \left\{ (1-\tau) \int_{-\infty}^{m} \mid y-m \mid f(y)\mathrm{d}y + \tau \int_{m}^{+\infty} \mid y-c \mid f(y)\mathrm{d}y \right\}$$

1.3 Graphical Representation

Boxplots

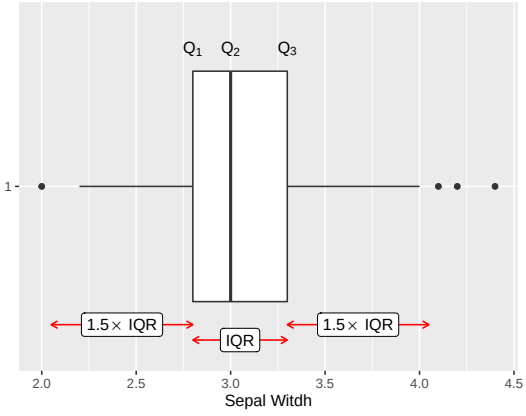
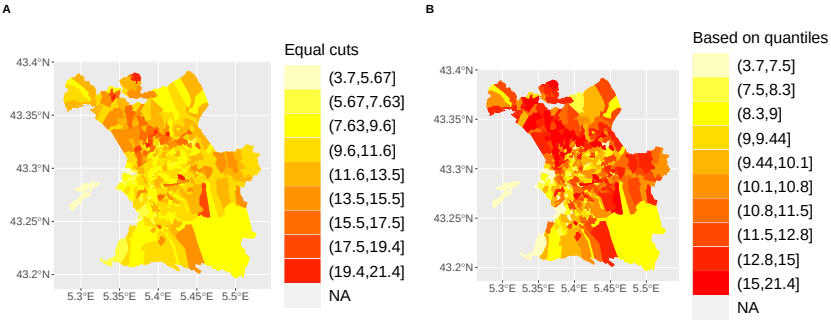


Figure 7: Boxlpot.

A famous graphical representation of the distribution of data relies on quantiles: the **boxplots**.
IQR: interquartile range

Choropleth maps



Note: Example adapted from Arthur Charpentier's slides (INF7100 Ete 2020, slides 224)

Figure 8: Proportion of individuals between the ages of 18 and 25 (inclusive) registered on the electoral lists in Marseilles, by voting center.

2. Quantile regression

2.1 Principles

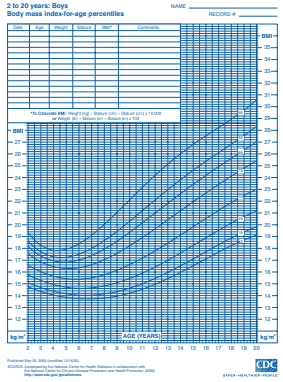
Quantile regression

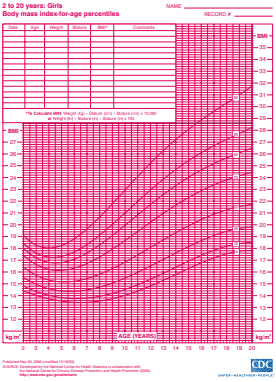
In the linear regression context, we have focused on the conditional distribution of y, but only paid attention to the **mean effect**.

In many situations, we only look at the effects of a predictor on the conditional mean of the response variable. But there might be some **asymetry** in the effects across the quantiles of the response variable:

- the effect of a variable could not be the same for all observations (*e.g.*, if we increase the minimum wage, the effect on wages may affect low wages differently than high wages).

Quantile regression offers a way to account for these possible asymmetries.





Source: Centers for Disease Control and Prevention

Figure 9: BMI Percentile Calculator for Child and Teen (boys).

Principles of Quantile Regression

Let Y be the response variable we want to predict using $p+1$ predictors $\boldsymbol{X}$ (including the constant).

In the linear model, using least squares, we write:

$$Y = \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon,$$

with ε a zero mean error term with variance $\sigma^2 I_n$.

Thus, the conditional distribution writes:

$$\mathbb{E}\left(Y \mid X = \mathbf{x}\right) = \mathbf{x}^\top \boldsymbol{\beta}$$

Here, $\boldsymbol{\beta}$ represents the **marginal change** in the **mean** of the response Y to a marginal change in $\mathbf{x}$.

Principles of Quantile Regression

Instead of looking at the mean effect, we can look at the effect at a given quantile. The **conditional quantile** is defined as:

$$Q_\tau(Y \mid X = \mathbf{x}) = \inf\{y : F(y \mid \mathbf{x}) \geq \tau\}$$

The **linear quantile regression model** assumes:

$$Q_\tau(Y \mid X = \mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}_\tau \tag{2}$$

where $\boldsymbol{\beta}_\tau = \begin{bmatrix} \beta_{0,\tau} & \beta_{1,\tau} & \beta_{2,\tau} & \dots & \beta_{p,\tau} \end{bmatrix}^\top$ is the quantile coefficient:

- it corresponds to the **marginal change** in the τth quantile folowing a marginal change in $\mathbf{x}$.

Principles of Quantile Regression

If we assume that $Q_\tau(\varepsilon \mid X) = 0$, then Eq. (2) is equivalent to:

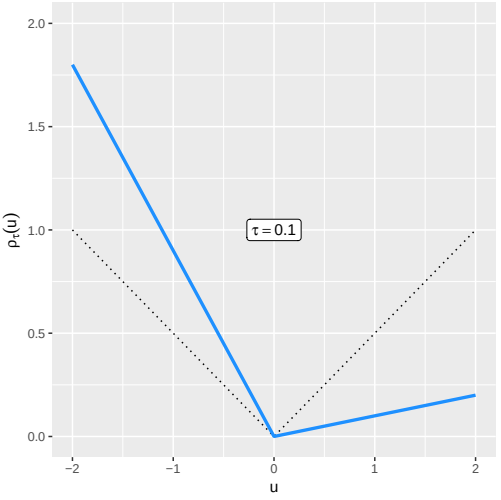
$$Y = \boldsymbol{X}^\top \boldsymbol{\beta}_\tau + \varepsilon_\tau \tag{3}$$

Asymetric absolute loss

Recall the asymetric absolute loss function is defined as:

$$\rho_\tau(u) = (\tau - \mathbb{1}(u < 0)) \times u,$$

for $0 < \tau < 1$.



Asymetric absolute loss

With $\rho_\tau(u)$ used as a loss function, it can be shown that Q_τ **minimizes the expected loss**, *i.e.*:

$$Q_\tau(Y) \in \arg\min_m \{\mathbb{E}[\rho_\tau(Y-m)]\} \tag{4}$$

We can note that in the case in which $\tau = 1/2$, this corresponds to the median.

Quantiles may not be unique:

- any element of $\{x \in \mathbb{R} : F_Y(x) = \tau\}$ minimizes the expected loss
- if the solution is not unique, we have an interval of τth quantiles
  - the smalest element is chosen (this way, the quantile function remains left-continuous).

Empirically

Now, let us turn to the estimation. Let us consider a random sample $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.

In the case of the **least square estimation**, the expectation $\mathbb{E}$ minimizes the risk that corresponds to the quadratic loss function, *i.e.*:

$$\mathbb{E}(Y) = \arg\min_m \left\{ \mathbb{E}\left[(Y - m)^2 \right] \right\}$$

- The sample mean solves $\min_m \sum_{i=1}^n (y_i - m)^2$
- The least squares estimates of the parameters are obtained by minimizing $\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \beta)^2$

Empirically

The **τth quantile** minimizes the risk associated with the asymetric absolute loss function:

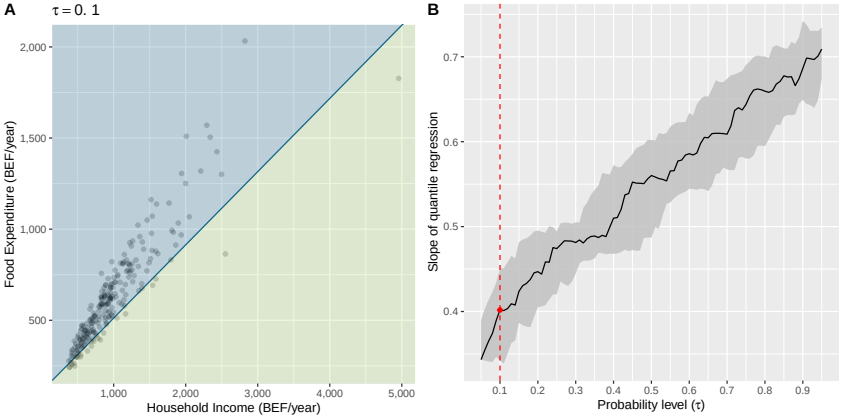
$$Q_\tau(Y) \in \arg\min_m \{\mathbb{E}[\rho_\tau(Y-m)]\}$$

The τth sample quantile of Y solves:

$$\min_m \sum_{i=1}^n \rho_\tau(y_i - m)$$

If we assume that $Q_\tau(Y \mid \boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{\beta}_\tau$, then, the quantile estimator of the parameters is given by

$$\hat{\boldsymbol{\beta}}_\tau \in \arg\min_\beta \left\{ \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) \right\} \tag{5}$$



Source: Koenker and Bassett Jr (1982)

Figure 11: Food Expenditure as a function of Income.

In the previous slide, let us focus on $\tau = .1$.

- *Left graph*: The estimated **slope** and **intercept** for that quantile allows us to show that if the household income is 3~500 Belgian Francs (BEF):
  - there is a 90% chance that food expenditure for the household is lower than 1~500 BEF;
  - there is a 10% chance that foodexpenditure for the household is greater than 1~500 BEF;
- *Right graph*: plotting the slope of the quantile regression of expenditure on income, for different values of τ
  - the slope coefficient is always positive
  - the value increases with the level of the quantile: for relatively richer households, the marginal effect of income on food expenditures is higer.

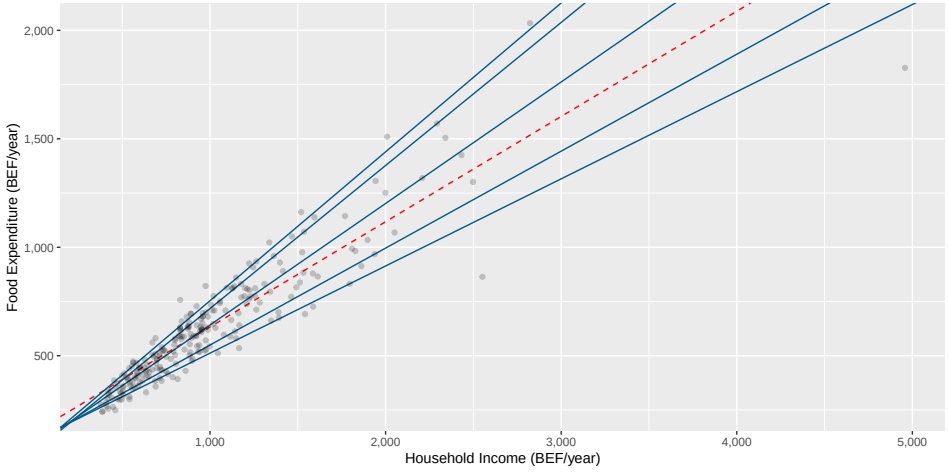


Figure 12: Food expenditure as a function of income: slope of linear quantile regression.

From the previous graph, we observe that the slopes change relatively greatly depending on the level of the quantile :

- for the individuals with the lowest incomes, the slope is relatively small (for $\tau = .1$, it is 0.40)
- for individuals with higher incomes, the slope is relatively higher (for $\tau = .9$, it is 0.69)

Therefore, the higher the income category, the greater the effect of income on consumption.

Location-shift model

Let usconsider a simple model:

$$Y = \boldsymbol{X}^\top \boldsymbol{\gamma} + \varepsilon \tag{6}$$

We have:

$$Q_\tau(Y \mid X = x) = \boldsymbol{X}^\top \gamma + Q_\tau(\varepsilon)$$

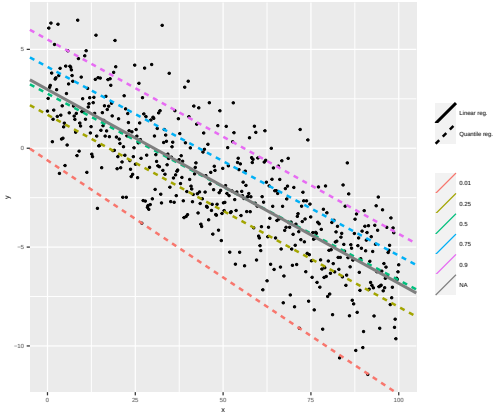
This model known as the **location model**, the only coefficient that varies accordingly with τ is the coefficient associated with the constant: $\beta_0, \tau = \gamma_1 + Q_\tau(\varepsilon)$

Location-shift model

- The conditional distribution $F_{Y|X=\mathbf{x}}$ are parallel when $\mathbf{x}$ varies.
- As a consequence, the conditional quantiles are linearly dependant on X, and the only coefficient that varies with the quantile is $\beta_{0,\tau}$, *i.e.*, the coefficient associated with the constant :
  - $\beta_{0,\tau} = \gamma_1 + Q_\tau(\varepsilon)$
  - $\beta_{j,\tau} = \gamma_j$ for all the coefficients except the constant.

Location-shift model: exemple

Consider the following true process: $Y = \beta_0 + \beta_1 \mathbf{x}_2 + \varepsilon$, with $\beta_0 = 3$ and $\beta_1 = -.1$, and where $\varepsilon \sim \mathcal{N}(0, 4)$.
Let us generate 500 observations from this process.



Location-scale model

Now, let us consider a **location-scale model**. In this model, we assume that the predictors have an impact both on the mean and on the variance of the response:

$$Y = X^\top \boldsymbol{\beta} + (X^\top \boldsymbol{\gamma})\varepsilon,$$

where ε is independant of $\boldsymbol{X}$.

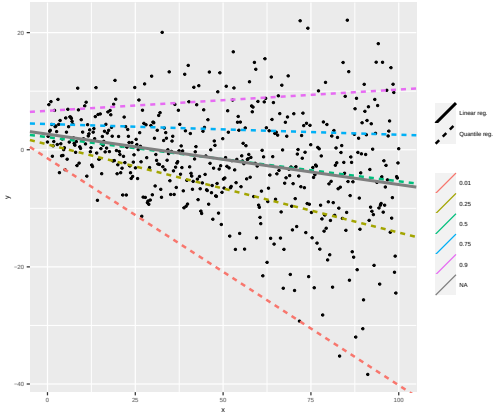
As $Q_\tau(aY + b) = aQ_\tau(Y) + b$, we can write:

$$Q_\tau(Y \mid X = \mathbf{x}) = \mathbf{x}^\top(\boldsymbol{\beta} + \gamma Q_\tau(\varepsilon))$$

- By posing $\boldsymbol{\beta}_\tau = \boldsymbol{\beta} + \gamma Q_\tau(\varepsilon)$, the assumption (2) still holds.
- The impact of the predictors will vary accross quantiles
- The slopes of the lines corresponding to the quantile regressions are not parallel

Location-scale model: example

Consider the following true process: $Y = \beta_1 + \beta_2 \mathbf{x}_2 + \varepsilon$, with $\beta_0 = 3$ and $\beta_1 = -.1$, and where ε is a normally distributed error with zero mean and non-constant variance. Let us generate 500 observations from this process, and then estimate a quantile regression on different quantiles.



2.2 Inference

Asymptotic distribution

As there is no explicit form for the estimate $\hat{\beta}_\tau$, achieving its **consistency** is not as easy as it is with the least squares estimate.

First, we need some assumptions to achieve consistency:

- the observations $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ must be conditionnaly i.i.d.
- the predictors must have a bounded second moment, *i.e.*, $\mathbb{E}\left[|| \boldsymbol{X}_i ||^2\right] < \infty$
- the error terms ε_i must be continuously distributed given $\boldsymbol{X}_i$, and centered, *i.e.*, $\int_{-\infty}^{0} f_\varepsilon(\epsilon) \mathrm{d}\epsilon = 0.5$
- $\left[f\varepsilon(0)\boldsymbol{X}\boldsymbol{X}^\top\right]$ must be positive definite (*local identification* property).

Asymptotic distribution

Under those weak conditions, $\hat{\boldsymbol{\beta}}_\tau$ is asymptotically normal:

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \tau(1-\tau) D_\tau^{-1} \Omega_x D_\tau^{-1}\right) \tag{7}$$

where

$$D_\tau = \mathbb{E}\left[f_\varepsilon(0) \boldsymbol{X}\boldsymbol{X}^\top\right] \text{ and } \Omega_x = \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^\top\right]$$

The asymptotic variance of $\hat{\boldsymbol{\beta}}$ writes, for the location shift model (Eq. 6):

$$\hat{\mathbb{V}}ar\left[\hat{\boldsymbol{\beta}}_\tau\right] = \frac{\tau(1-\tau)}{\left[\hat{f}_\varepsilon(0)\right]^2} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i\right)^{-1} \tag{8}$$

where $\hat{f}_\varepsilon(0)$ can be estimated using an histogram (see Powell 1991).

Confidence intervals

If we have obtained an estimator of $\mathbb{V}ar\left[\hat{\boldsymbol{\beta}}_\tau\right]$, the confidence interval of level $1-\alpha$ is given by:

$$\left[\hat{\beta}_\tau \pm z_{1-\alpha/2}\sqrt{\hat{\mathbb{V}}ar\left[\hat{\boldsymbol{\beta}}_\tau\right]}\right] \tag{9}$$

Otherwise, we can rely on bootstrap or resampling methods (see Emmanuel Flachaire's course):

- Generate a sample $\{(\mathbf{x}_1^{(b)}, y_1^{(b)}), \ldots, (\mathbf{x}_n^{(b)}, y_n^{(b)})\}$ from $\{(\mathbf{x}_1, y_1, \ldots, (\mathbf{x}_n, y_n)\}$
- Then estimate $\boldsymbol{\beta}_\tau^{(b)}$ using $\hat{\boldsymbol{\beta}}_\tau^{(b)} = \arg\min\left\{\rho_\tau\left(y_i^{(b)} - \mathbf{x}_i^{(b)\top}\boldsymbol{\beta}\right)\right\}$
- Do the two previous staps B times
- The bootstrap estimate of the variance of $\hat{\boldsymbol{\beta}}_\tau$ is then computed as:

$$\hat{\mathbb{V}}ar^\star\left(\hat{\boldsymbol{\beta}}_\tau\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\beta}}_\tau^{(b)} - \hat{\boldsymbol{\beta}}_\tau\right)^2$$

2.3 Example: birth data

Birth data

Let us consider an example provided in Koenker (2005), about the impact of demographic characteristics and maternal behaviour on the birth weight of infants botn in the US.

Let us use the Natality Data from the National Vital Statistics System, freely available on the NBER website (https://data.nber.org/data/vital-statistics-natality-data.html)

The raw data for 2018 concerns more than 3M babies born in that year, from american mothers aged bewteen 18 and 50.

For the sake of the exercise, we only use a sample of 20,000 individuals.

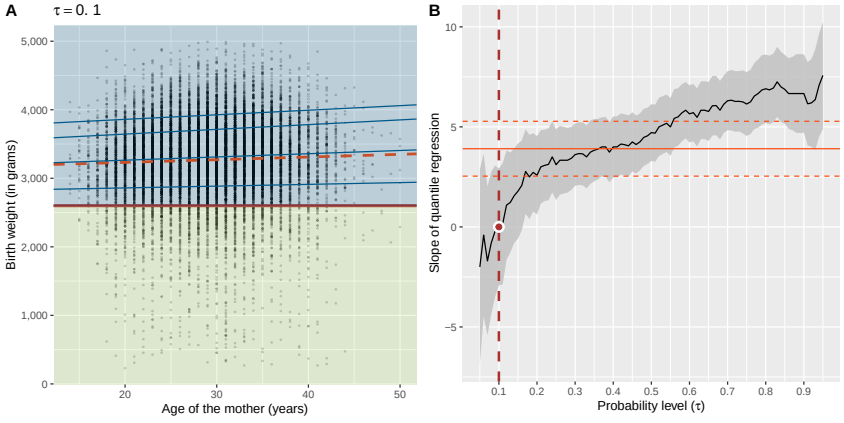
Variables kept

The predictors used are:

- education (less than high school (ref), high school, some college, college graduate)
- prenatal medical care (no prenatal visit, first prenatal visit in the first trimester of pregnancy (ref), first visit in the second trimester, first visit in the last trimester)
- the sex of the infant
- the marital status of the mother
- the ethnicity of the mother
- the **age of the mother**
- whether the **mother smokes**
- the number of cigarette per day
- the gain of weight for the mother

Let us focus only on a couple of them.

Age


A
τ = 0. 1
Birth weight (in grams)
5,000
4,000
3,000
2,000
1,000
0
20
30
40
50
Age of the mother (years)
B
Slope of quantile regression
10
5
0
−5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Probability level (τ)

Age

- If the slopes are not parallel, and the observed differences are significant, then the effect of age on the baby's weight is different depending on where the baby is in the distribution.
- At the **lower part** of the distribution of weights ($\tau = .1$, relatively light babies): effect of age on weight negative but not significant
- At the **upper part** of the distribution of weights ($\tau = .9$, relatively heavy babies): positive and significant effect of age on weight

Age: nonlinear effect?

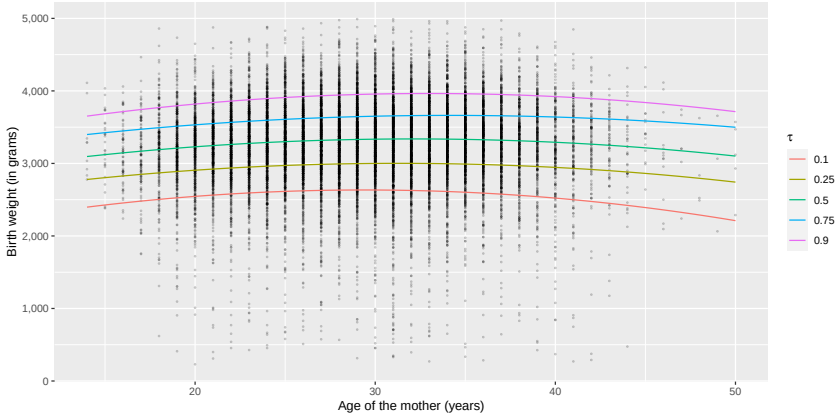


Figure 13: Birth weight as a function of the age of the mother - Non-linear Quantile Regression

Age: nonlinear effect?

The influence of the mother's age may be thought to be **non-linear on birth weight**.

For heavy babies (and also light babies), the weight of the babies is relatively lower on the borders of the mother's age distribution.

Cigarettes

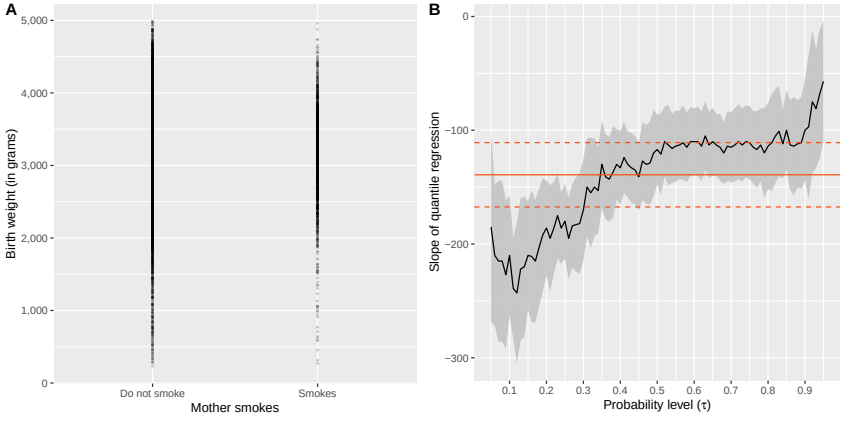


Figure 14: Birth weight depending on whether the mother smokes

Cigarettes

- Smoking during pregnancy is associated with a decrease of birth weight
- For relatively light babies, the coefficient of the slope is relatively dramatically lower than for relatively heavy babies:
  - the fact thet the mother is smoking prior pregnancy has a significant impact on the newborn weight, especially for the relatively lighter babies.

References I

Alvaredo, F., Chancel, L., Piketty, T., Saez, E., and Zucman, G. (2018). The elephant curve of global inequality and growth. *AEA Papers and Proceedings*, 108:103–08.

Davino, C., Furno, M., and Vistocco, D. (2014). *Quantile Regression*. John Wiley & Sons, Ltd.

Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Koenker, R. and Bassett Jr, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 43–61.

Powell, J. L. (1991). Estimation of monotonic regression models under quantile restrictions. *Nonparametric and semiparametric methods in Econometrics*, pages 357–384.