QUANTILE REGRESSION

Pranava Priyanshu and Ankan Kar

Indian Statistical Institute, Bangalore

Presentation on Quantile Regression
February 1st, 2023

Instructor : Prof. Mohan Delampady

Introduction

Standard linear regression focuses only on the expectation of a variable Y conditional on a set of regressors X which gives us only a partial description of the conditional distribution $Y|X$. But sometimes we need to describe the relationship at different points in the conditional distribution of Y for which Quantile Regression is needed.

Quantile regression extends this approach, allowing one to study the conditional distribution of Y on X at different locations and thus offering a global view on the interrelations between Y and X. Using an analogy, we can say that for regression problems, QR is to classical regression what quantiles are to mean in terms of describing locations of a distribution.

Quantile-Based Measures of Location and Shape

We should get familiar with quantile-based measure of central location. Like instead of the mean (the first moment of a density function), the median (the 0.5^{th} quantile is used to indicate the centre of skewed distribution.

Using quantile-based location allows us to get a more general notion of location of a distribution beyond just the centre; like, we can find the location of the lower tail or upper tail for for specific sub-populations.

Two basic properties describe the shape of a distribution: scale and skewness.

To get the spread of the distribution without relying on standard deviation, we measure spread using following quantile-based scale measure(QSC) at a selected θ:

$$QSC(\theta) = Q(1-\theta) - Q(\theta) \qquad for\ \theta < 0.5$$

This way we can get the spread of any desirable middle $(100-2\theta)\%$ of population.

A model-based approach that separates out a predictor's effect in terms of a change in scale as measured by the standard deviation limits the possible patterns that could be discovered. In contrast, the QSC measure not only offers a direct and straightforward measure of scale but also facilitates the development of a rich class of model-based scale-shift measures.

We many times use the terms upper spread which refer to the spread above the median and the lower spread which refer to the spread below the median.

The second measure of distributional shape is skewness. This property focusses on inequality of distribution. Skewness is generally measured using a cubic function of data points' deviations from the mean. When the data points are symmetrically distributed about the sample data mean, the value of skewness is 0. We quantify the measure of quantile-based skewness (QSK) as a ratio of the upper spread to the lower spread minus one at selected θ:

$$QSK(\theta) = \frac{Q(1-\theta) - Q(0.5)}{Q(0.5) - Q(\theta)} - 1 \qquad for\ \theta < 0.5$$

The quantity $QSK(\theta)$ is recentered using subtraction of one, so that it takes the value zero for a symmetric distribution. A value greater than zero indicates right-skewness and a value less than 0 indicates left-skewness. Skewness can be interpreted as saying that there is an imbalance between the spread below and above the median.
This definition of $QSK(\theta)$ is simple and straightforward and can be extended to measure the skewness shift caused by a covariate.

Quantiles as solutions of a minimization problem

Comparison of mean and quantiles and their objective functions : Let Y be a random variable then its mean say μ appears as the solution to the following minimization problem :

$$\mu = \underset{a}{\text{argmin}} E[(Y-a)^2]$$

Similarily the median minimizes the absolute sum of deviations. In terms of a minimization problem, the median is thus :

$$\text{Me} = \underset{a}{\text{argmin}} E[|Y-a|]$$

Quantile Functions : Let Y be a univariate random variable with commulative distribution function F_Y then its quantile function at $\theta \in [0, 1]$ is defined as :

$$Q_Y(\theta) = \inf(y : F_Y(y) \geq \theta)$$

If F_Y is strictly increasing and continuous, then $F^{-1}(\theta)$ is the unique real number y such that $F_Y(y) = \theta$.

It turns out that the $\theta - th$ quantile is the solution of the minimization problem :

$$\min_a E[\rho_\tau(Y - a)], \text{ where } \rho_\tau(x) = (\tau - 1(x < 0))x \text{ and } \tau \in (0, 1)$$

Let $f_Y(y)$ be the pdf of Y and assume that f is a continuous function and Y has a unique $\tau - th$ quantile, then :

$$E[\rho_\tau(Y-a)] = \tau \int_a^\infty (y-a) f_Y(y) dy + (\tau - 1) \int_{-\infty}^a (y-a) f_Y(y) dy$$

$$= \tau \left(\int_a^\infty y f_Y(y) dy - a \int_a^\infty f_Y(y) dy \right) + (\tau - 1) \left(\int_{-\infty}^a y f_Y(y) dy - a \int_{-\infty}^a f_Y(y) dy \right.$$

Differentiationg wrt a to obtain the first order condition we get :

$$\tau(-a f_Y(a) + a f_Y(a) + \int_a^\infty f_y(y) dy) + (\tau - 1)(a f_Y(a) - a f_Y(a) - \int_{-\infty}^a f_Y(y) dy)$$

$$= 0$$

$$\implies F_y(a) - \tau = 0 \implies a = F_Y^{-1}(\tau)$$

Taking $\tau = 0.5$, the above problem boils down to minimizing the absolute sum of deviations the solution of which is median.

Conditional Means and Quantiles

Suppose that Y is the response variable and $\mathbf{X}$ is the set of predictor variables the idea of the unconditional mean as the minimizer of $E[(Y-a)^2]$ can be extended to the estimation of the conditional mean function :

$$\hat{\mu}(x_i,\beta) = \underset{\mu}{\text{argmin}} E[(Y-\mu(x_i,\beta))^2], \text{where} \quad \mu(x_i,\beta) = E[Y|\mathbf{X}=x_i]$$

When the conditional mean function is linear i.e $\mu(x_i,\beta) = x_i^T\beta$, then the previous equation becomes :

$$\hat{\beta} = \underset{\beta}{\text{argmin}} E[Y - x_i^T\beta]$$

which is the least squares linear regression model.

We can extend this approach to the generic $\tau - th$ quantile in which case we obtain :

$$\hat{q}_Y(\tau, \mathbf{X}) = \underset{Q_Y}{\text{argmin}} E[\rho_\tau(Y - Q_Y(\tau, \mathbf{X})], \text{where} \quad Q_Y(\tau, \mathbf{X}) = Q_\tau[Y|\mathbf{X} = x]$$

is the conditional quantile function. Likewise, for the linear model case the previous equation becomes:

$$\hat{\beta}_\tau = \underset{\beta}{\text{argmin}} E[\rho_\tau(Y - \mathbf{X}\beta)]$$

The sample version is :

$$\hat{\beta}_\tau = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^t \beta)$$

QR Model with Dummy Regressor

The simplest form of linear model is a model with a quantative response variable and a dummy predictor variable. The estimation of the QR model:

$$\hat{Y} = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)X$$

for different values of $\theta \in [0, 1]$ permits us to an estimation of the Y quantiles for the groups of X. Like when X is a dichotomous predictor variable then we get two groups of X as 0 or 1. Then we get;

$$\hat{Y} = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \times 0 = \hat{\beta}_0(\theta) \quad for\ X = 0\ and;$$

$$\hat{Y} = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \times 1 = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta) \quad for\ X = 1$$

QR Model with Nominal Regressor

In this linear model we will have the predictor variable as a multilevel categorical variable with g categories. Here QR allows us to compare the different quantiles among the different g groups.

To deal with a g level nominal variable we need to introduce g - 1 dummy variables. We then get this QR model:

$$\hat{Y} = \hat{\beta}_0(\theta) + \sum_{i=1}^{g-1} \hat{\beta}_i(\theta) I(x_i)$$

where $I(.)$ is the indicator function returning 1 if the particular unit assumes the value in the paranthesis 0 otherwise. In this model for a given quantile the combination of intercept with different slopes gives us the conditional quantiles of the response variable.
The estimated effect of a particular group can be obtained by using the dummy variable associated with the particular slope.

A typical QR model

As already mentioned, QR is an extension of the classical estimation of conditional mean models to conditional quantile functions; that is an approach allowing us to estimate the conditional quantiles of the distribution of a response variable Y in function of a set $\mathbf{X}$ of predictor variables. In the framework of a linear regression, the QR model for a given conditional quantile τ can be formulated as follows :

$$Q_\tau(Y|\mathbf{X}) = \mathbf{X}\beta(\tau), \tau \in (0,1)$$

The parameter estimates in QR linear models have the same interpretation as those of any other linear model, as rates of change. Therefore, in a similar way to the OLS model, the $\beta_i(\tau)$ coefficient of the QR model can be interpreted as the rate of change of the τ-th quantile of the dependent variable distribution per unit change in the value of the i-th regressor

The linear programming formulation for the QR problem

Suppose we have a linear QR model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, 2, ...n$ and we want to estimate the conditional median of Y, then we have the following minimization problem :

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} |\beta_0 + \beta_1 x_i - y_i|$$

The equivalent linear programme can be written as :

$$\begin{aligned} & \text{minimize} \sum_{i=1}^{n} e_i \\ \text{subject to} \quad & \mathrm{e}_i \geq \beta_0 + \beta_1 x_i - y_i, i = 1, 2, ...n \\ & \mathrm{e}_i \geq -(\beta_0 + \beta_1 x_i - y_i), i = 1, 2, ...n \end{aligned}$$

The constraints guarantee that $e_i \geq |\beta_0 + \beta_1 x_i - y_i|$. In an optimal solution, e_i has to satisfy $e_i = |\beta_0 + \beta_1 x_i - y_i|$ otherwise, we can decrease the corresponding e_i.

The p-variables problem: The model for QR is $Y = \mathbf{X}\beta(\tau) + \epsilon$ where Y is a vector of responses, $\mathbf{X}$ is the regression matrix, $\beta(\tau)$ is the vector of unknown parameters for the generic conditional quantile τ and ϵ is the vector of unknown errors. We get the estimate for the τ-th quantile by solving :

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta)$$

We want to get an equivalent linear programming problem in the standard form. Let $\epsilon_i = u_i - v_i$ where $u_i = \max(0, \epsilon_i)$ and $v_i = \max(0, -\epsilon_i)$. Then we can write

$$\sum_{i=1}^{n} \rho_\tau(\epsilon_i) = \sum_{i=1}^{n} \tau u_i + (1-\tau)v_i = \tau \mathbf{1}_n^\top u + (1-\tau)\mathbf{1}_n^\top v$$

where $u = (u_1, ...u_n)'$ and $v = (v_1, ...v_n)'$

Now the residuals must satisfy the n constraints $y_i - \mathbf{x}_i^\top \beta = \epsilon_i = u_i - v_i$ and so we get the following linear programme :

$$\min_{\beta \in \mathbb{R}^p, u \in \mathbb{R}_+^n, v \in \mathbb{R}_+^n} \{\tau \mathbf{1}_n^\top u + (1-\tau)\mathbf{1}_n^\top v | y_i = \mathbf{x}_i^T \beta + u_i - v_i, i = 1, ..., n\}$$

The elements of β are still not positive which is required for a LP in standard form. So let $\beta = \beta^+ - \beta^-$ where $\beta^+ = (\beta_1^+, ...\beta_p^+)$ and $\beta^- = (\beta_1^-, ...\beta_p^-)$ in a similar way as we did for the residuals. Then the n constraints can be written as :

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} (\beta^+ - \beta^-) + \mathbf{I}_n u - \mathbf{I}_n v,$$

where $\mathbf{I}_N = diag\{\mathbf{1}_N\}$.

Now we can write :

$$\mathbf{X}(\beta^+ - \beta^-) + \mathbf{I}_n u - \mathbf{I}_n v = [\mathbf{X}, -\mathbf{X}, \mathbf{I}_n, -\mathbf{I}_n] \begin{bmatrix} \beta^+ \\ \beta^- \\ u \\ v \end{bmatrix} = Az$$

where $A = [\mathbf{X}, -\mathbf{X}, \mathbf{I}_n, -\mathbf{I}_n]$ and $z = (\beta^+, \beta^-, u, v)^T$. Then we get our LP problem as :

$$\underset{z}{\text{minimize}} \quad c^T z$$

$$\text{subject to} \quad Y = Az, z \geq 0.$$

where $c = \begin{bmatrix} \mathbf{0} \\ \tau \mathbf{1}_n \\ (1-\tau)\mathbf{1}_n \end{bmatrix}$

How does Quantile Regression Work

1.General Linear Position : A set of points in a d-dimensional Euclidean space is in general linear position if no k of them lie in a $k-2$-dimensional flat for $k = 2, 3, ..., d+1$.

2.Subgradient : A subgradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ is any vector $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T(y - x), \forall y$$

3.Subgradient Optimality Condition : For any function f, x^* is a minimizer if and only if 0 is a subgradient of f at x^* :

$$f(x^*) = \min_x \Leftrightarrow 0 \in \partial f(x^*)$$

4.Subgradient of the sum is the sum of the subgradients : Suppose that $f = f_1 + f_2 + ...f_m$, where $f_1, ...f_m$ are convex functions, then $\partial f(x) = \partial f_1(x) + ... + \partial f_m(x)$.

5.If f is a differentiable function then $\nabla f = \partial f$, the gradient is the only subgradient.

6.Necessary and sufficient conditions for one half space to contain the other : Suppose we have two half spaces gievn by

$$H_1 = (x \in R^m : a_1^T x \leq b_1) \quad \text{and} \quad H_2 = (x \in R^m : a_2^T x \leq b_2)$$

Then $H_1 \subseteq H_2$ iff there exists a $k > 0$ such that $a_2 = ka_1$ and $kb_1 \leq b_2$

Let Y be the response variable and **X** be the regression matrix whose 1st column has all the elements equal to 1.

Claim : When the regression observations $(Y, \mathbf{X})$ are in general linear position then the solutions to the regression analogue of our elementary problem, $\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta)$ has the property that roughly $(1-\tau)n$ of the residuals, $r_i = y_i - x_i^T \beta, i = 1, 2, ...n$ are positive and τn are negative.

Proof : The subgradient optimality condition implies that if $\hat{\beta}(\tau)$ minimizes $\sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\beta)$, then

$$0 \in \partial_\beta(\sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\beta))|_{\beta=\hat{\beta}}$$

Now $\partial_\beta(\rho_\tau(y_i - x_i^T\beta) = -\phi_\tau(y_i - x_i^T\beta)x_i$ with $\phi(x) = \tau - 1(x < 0)$ whenever $y_i \neq x_i^T\beta$ as at these points $\rho_\tau(y_i - x_i^T\beta)$ is differentiable and its gradient is the only subgradient at these points. At the points where the residuals are zero, the subgradient is set valued.
Fix a $1 \leq i \leq n$. We show that $\partial_\beta\rho_\tau(y_i - x_i^T\beta) = [-\tau, 1-\tau]x_i$.Let g be a sugradient at $\beta = \hat{\beta}$. Then by definition

$$\rho_\tau(y_i - x_i^T\beta) \geq \rho_\tau(y_i - x_i^T\hat{\beta}) + g^T(\beta - \hat{\beta}).$$

$$\implies \rho_\tau(y_i - x_i^T\beta) \geq g^T(\beta - \hat{\beta}) \quad (*)$$

Suppose β satisfies $x_i^T\beta \leq y_i$ then β lies in the half space

$$H = (b \in R^p : x_i^T b \leq y_i)$$

And so from $(*)$ we get that β lies in the half space

$$H_0 = (b \in R^p : (\tau x_i + g)^T b \leq \tau y_i + g^T \hat{\beta})$$

$$\implies H \subseteq H_0$$

So by (6), there exists a $k > 0$ such that $\tau x_i + g = k x_i$ and $k y_i \leq \tau y_i + g^T \hat{\beta} \implies g = x_i(k - \tau)$. Now suppose β satisfies $x_i^T \beta >= y_i$ then

β lies in the half space

$$H_1 = (b \in R^p : -x_i^T b \leq -y_i)$$

Again from $(*)$ we get that β lies in the half space

$$H_2 = (b \in R^p : ((\tau - 1)x_i + g)^T \beta \leq (\tau - 1)y_i + g^T \hat{\beta})$$

$$\implies H_1 \subseteq H_2$$

Again by (6), there exists a k_1 such that $(\tau - 1)x_i + g = -k_1 x_i \implies g = (1 - \tau - k_1)x_1 \implies k - \tau = 1 - \tau - k_1 \implies k + k_1 = 1 \implies 0 \leq k \leq 1$. Also for each such k we can easily check that $g = (k - \tau)x_i$ is a subgradient at $\beta = \hat{\beta}$. So, whenever the residuals are zero we have

$$\partial_\beta \rho_\tau (y_i - x_i^T \beta) = [-\tau, 1 - \tau] x_i$$

Let I, J, K be the sets of indices for which the residuals are positive, negative and zero respectively. Then the subgradient of $\sum_{i=1}^n \rho_\tau (y_i - x_i^T \beta))|_{\beta = \hat{\beta}}$ is of the form

$$\sum_{i \in I} -\tau x_i + \sum_{i \in J} -(\tau - 1)x_i + \sum_{i \in K} a_i x_i$$

where $a_i \in [-\tau, 1 - \tau]$.

Now as $0 \in \partial_\beta(\sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\beta))|_{\beta=\hat{\beta}}$, so there exits a sequence of a_{i_s} such that

$$\sum_{i\in I} -\tau x_i + \sum_{i\in J} -(\tau-1)x_i + \sum_{i\in K} a_i x_i = 0$$

Also as the first coordinate of each x_i is 1 so we get

$$\sum_{i\in I} -\tau + \sum_{i\in J} -(\tau-1) + \sum_{i\in K} a_i = 0 \quad (**)$$

Let $x_{i_1}, \hat{\beta}_1$ denote the vector x and $\hat{\beta}$ with their first coordinate removed and let the first coordinate of $\hat{\beta}$ be $\hat{\beta}_0$.Now for some i if the residual is zero then $y_i = x_i^T\hat{\beta}$ and so we get

$$(\hat{\beta}_1, -1)^T \begin{pmatrix} x_{i_1} \\ y_i \end{pmatrix} = -\hat{\beta}_0$$

This is a hyperplane of dimension p. As the observations $(x_{i_1}, y_i) \in R^p$ and are in general linear position so no more than p of them can lie on a hyperplane of dimension p and hence the number of residuals equal to 0 is atmost p and so $|K| \leq p$. From $(**)$ and using the fact that $a_i \in [-\tau, 1-\tau]$, we get

$$-\tau|I| - (\tau - 1)|J| + \sum_{i \in K} a_i = 0$$

$$\implies |J| - \tau(|I| + |J|) - |K|\tau \leq 0 \leq |J| - \tau(|I| + |J|) + (1-\tau)|K|$$

Now as $|I| + |J| + |K| = n$ so we get

$$\frac{|J|}{n} \leq \tau \leq \frac{|J|+|K|}{n} \leq \frac{|J|+p}{n}$$

$$\implies |I| \leq (1-\tau)n, |J| \leq \tau n$$

So roughly $(1-\tau)n$ of the residuals are positive and roughly τn are negative. Hence solutions $\hat{\beta}(\tau)$ of such minimization problems can be considered analogues of the sample quantiles for the linear model, estimating the parameters of models that specified affine conditional quantile functions for $Y|\mathbf{X}$.

Homogeneous Error Models

The Quantile Regression Model can be expressed as :

$$y_i = x_i^T \beta^\tau + \epsilon_i^\tau, i = 1, 2, ...n, \tau \in (0, 1)$$

When the error terms have constant variance then the model is said to be homogeneous and the error terms are called homoscedastic errors. We will consider special cases where the error terms are independent and identically distributed. In such cases the θ-th quantile of $Y|x_i$ is

$$Q_{Y|x_i}(\theta) = x_i^T \beta^\tau + Q_{\epsilon_i^\tau}(\theta)$$

Now as ϵ_{i_s} are i.i.d so $Q_{\epsilon_i^\tau}(\theta)$ is independent of i and only depends on τ, θ. So we can write

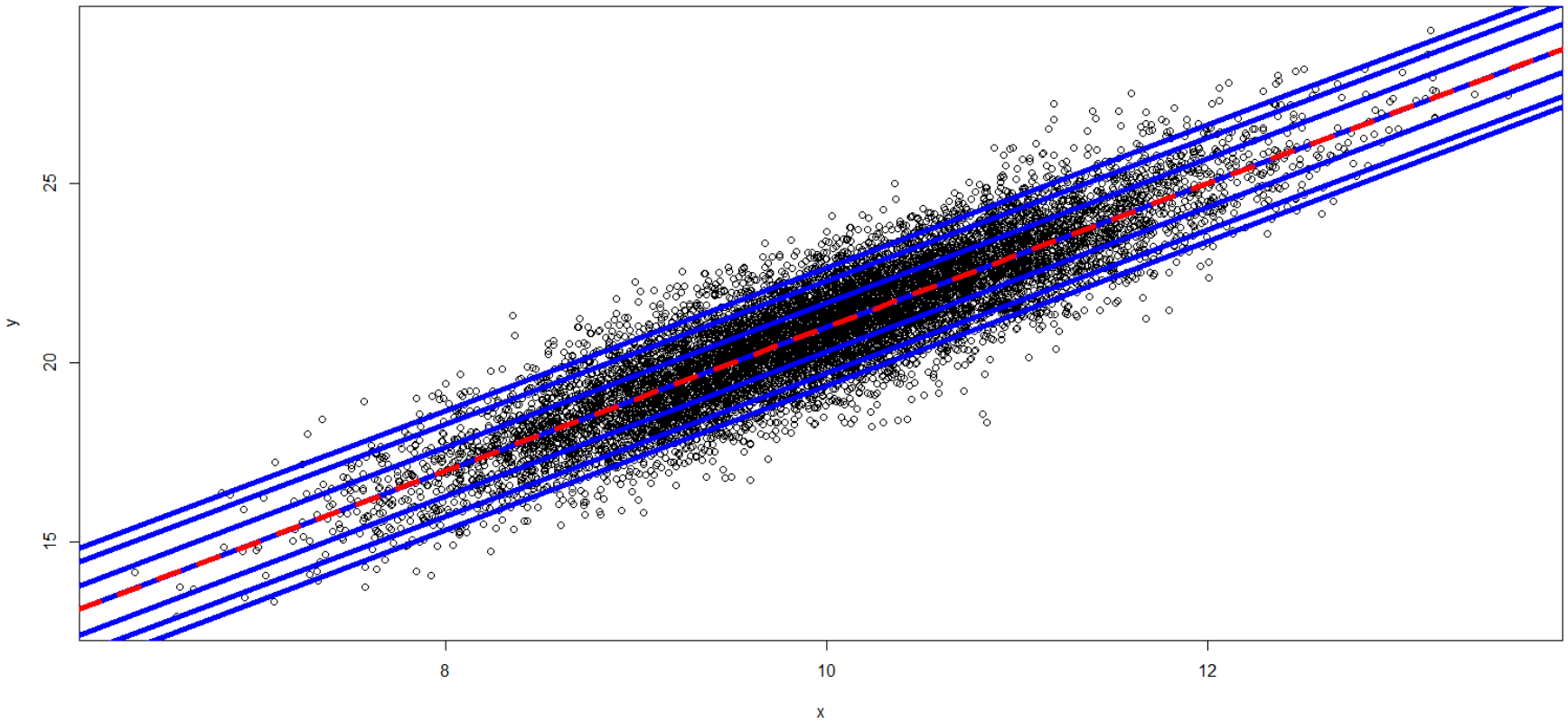
$$Q_{Y|x_i}(\theta) = x_i^T \beta^\tau + c_{\tau,\theta}$$

We conclude that in the i.i.d. case, the conditional quantile functions are simple shifts of one another.

In case on only one regressor our model becomes
$y_i = \beta_0 + x_i\beta_1^\tau + \epsilon_i^\tau, i = 1, 2, ...n, \tau \in (0, 1)$ and the conditional quantile functions are parallel lines with diferent intercepts when the error terms are iid.

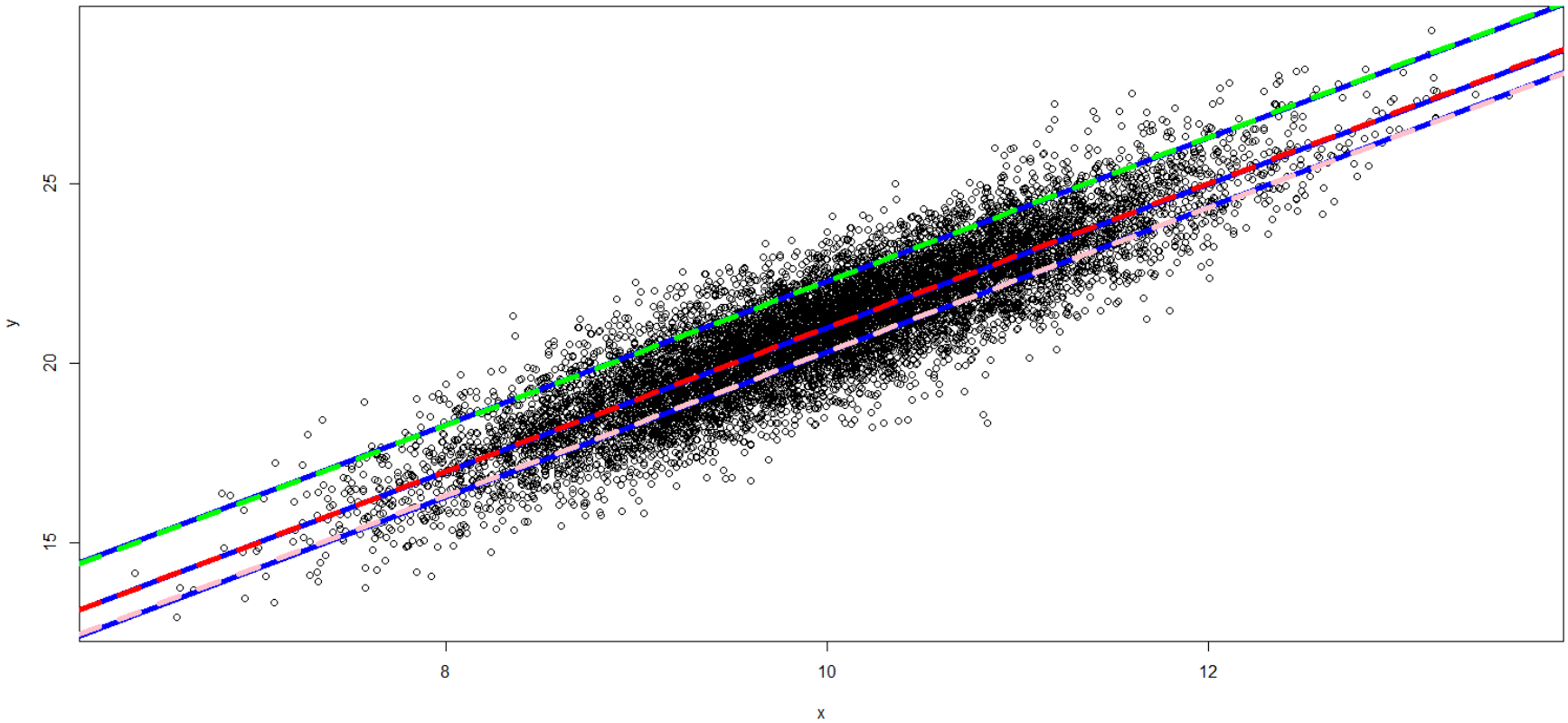
We illustrate this fact by simulating some generated data. 10000 observations of an independent variable x_i are generated from $N(10, 1)$, the dependent variable is computed as $y_i = 1 + 2x_i + \epsilon_i$ where the error terms are generated from $N(0, 1)$.

The OLS estimates and the QR estimates for
$\tau = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$ are computed and the corresponding regression lines are ploted in the picture below :



The middle dashed line is the OLS regression line and the blue lines are the regression lines for $\tau = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$

In the case when the error terms are iid and follow $N(0,1)$ then the QR estimates can be obtained by just shifting the OLS regression line by adding or substracting the corresponding standardized normal quantiles. This is illustrated in the picture below :



The middle dashed line is the OLS regression line the the other dashed lines are obtained by shifting the OLS line by adding or substracting the corresponding $N(0,1)$ quantiles. The blue lines are estimated QR lines

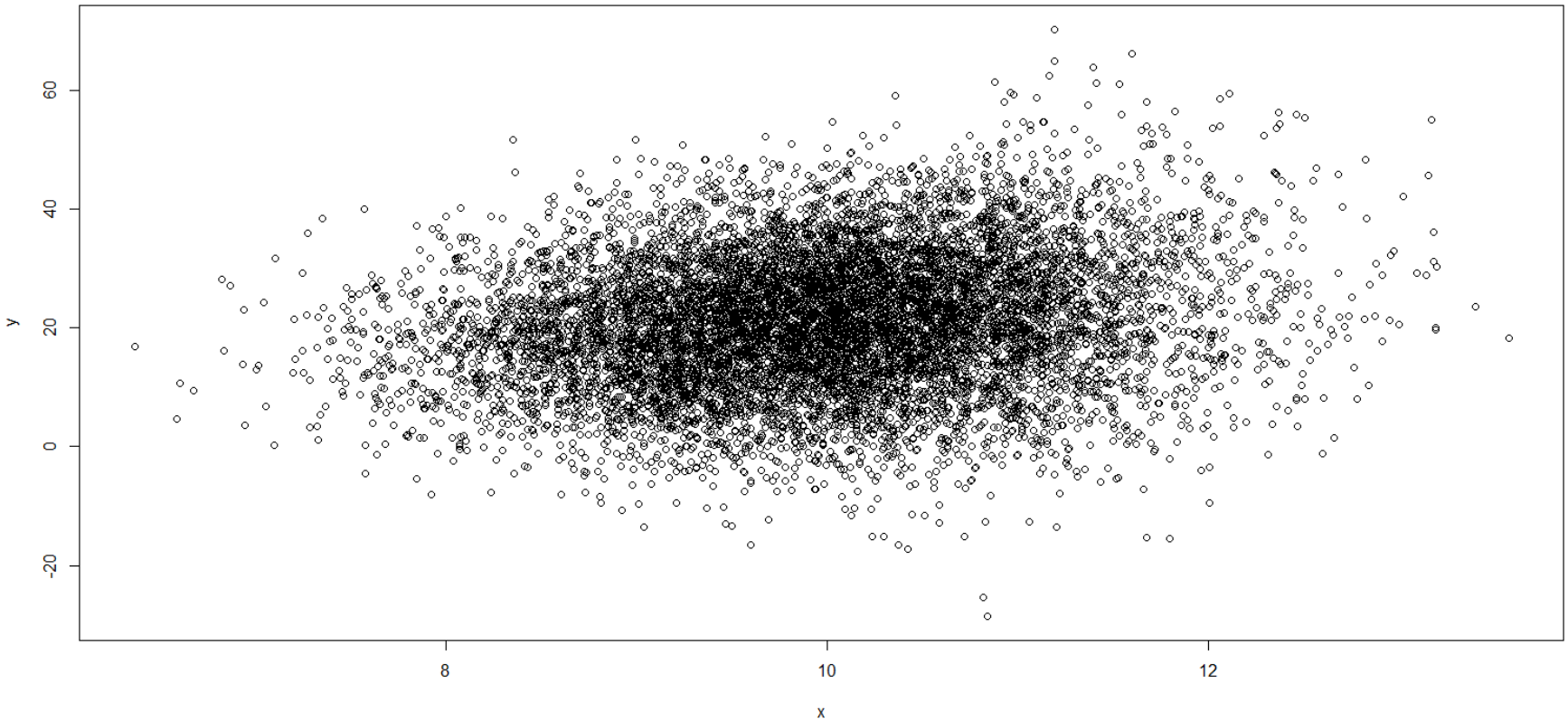
Heterogeneous Error models

These are the models in which the error terms have non constant variance and these errors are called heteroscedastic error terms. In such cases the OLS regression porvides an incomplete picture of the relationship between the variables as it focusses only on the conditional mean. On the other hand QR becomes much more powerful tool in such cases. By estimating different conditional quantiles using QR one can get a complete view of the conditional distribution of $Y|\mathbf{X} = x$. In heterogeneous error models, both the slope and the intercept vary across different quantiles (location-scale model).

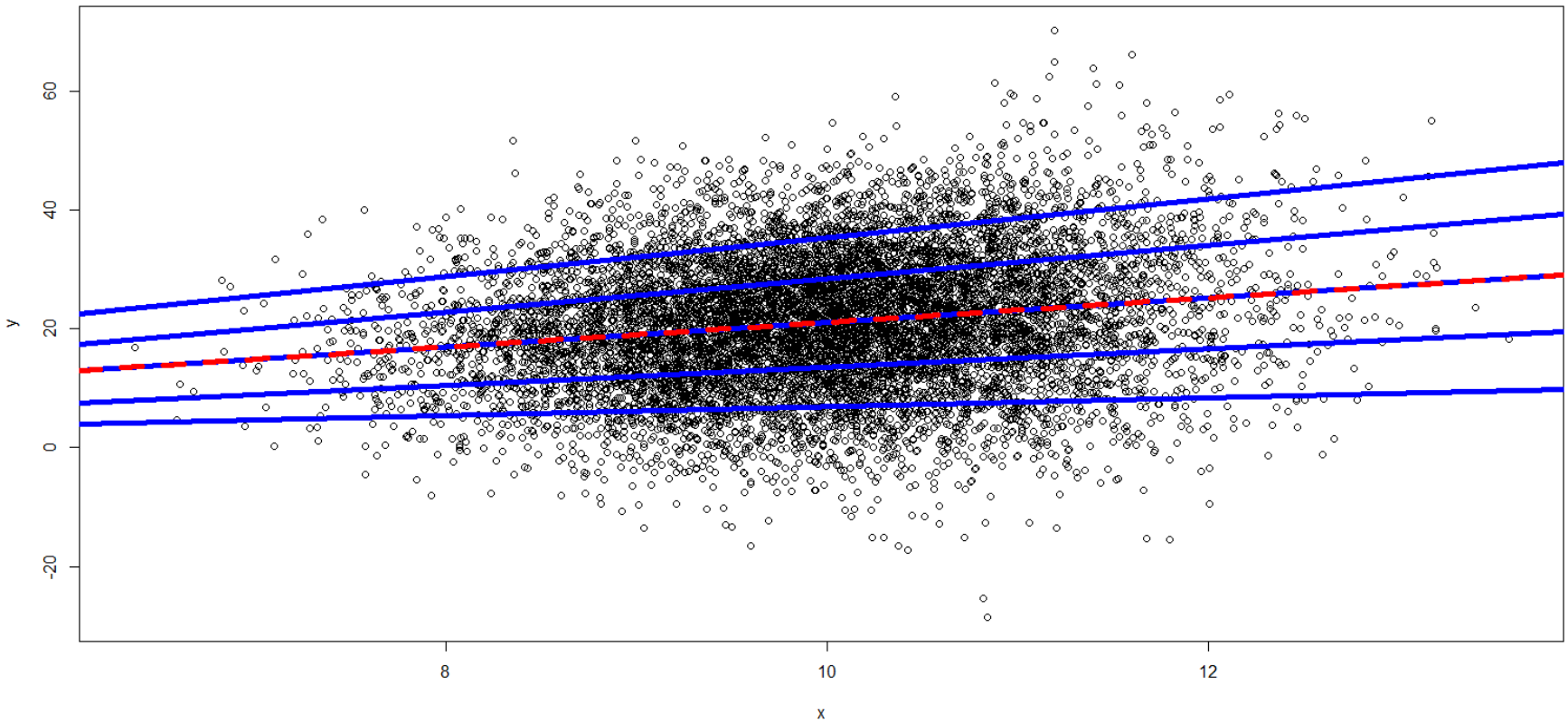
In order to take into account a simple heteroscedastic pattern, we consider the following model, starting from the standard normal error term

$$y_i = 1 + 2x_i + (1 + x_i)\epsilon_i$$

Again the independet variable x_i is generated from $N(0, 1)$ but this time the dependent variable is computed as $y_i = 1 + 2x_i + (1 + x_i)e_i$ where the error terms are generated from $N(0, 1)$.



Scatter plot of the data points generated by the model $y = 1 + 2x + (1 + x)\epsilon$



The middle dashed line is the OLS line and the blue lines are the QR lines for $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$

From the plot above it is clear that the QR lines for different quantiles vary not only in the intercept but also in the slope in the case of heteroscedasticity.

The estimates for both the homegeneous and heterogeneous models are listed below :

- Homogeneous Model :

	OLS	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$
intercept	0.9508	-0.301	0.150	0.959	1.538	2.318
x	2.00	2.00	2.01	2.00	2.01	1.99

- Heterogeneous Model :

	OLS	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$
intercept	0.367	-0.472	-1.756	0.496	0.240	2.567
x	2.07	0.73	1.53	2.05	2.82	3.27

QR prediction intervals

Prediction Intervals : A $100(1-\alpha)\%$ prediction interval for $Y|\mathbf{X}=x$ is an interval $[l,u]$ such that

$$P(l \leq Y \leq u|\mathbf{X}=x) = 1-\alpha$$

If Y is a continuous random variable with invertible CDF F_Y then $[F_Y^{-1}(\tau_1), F_Y^{-1}(\tau_2)]$ is a $100(\tau_2-\tau_1)\%$ prediction interval for Y as

$$\mathrm{P}(\mathrm{F}_Y^{-1}(\tau_1) \leq Y \leq F_Y^{-1}(\tau_2)) = P(Y \leq F_Y^{-1}(\tau_2)) - P(Y \leq F_Y^{-1}(\tau_1)) = F_Y(F_Y^{-1}(\tau_2)) - F_Y(F_Y^{-1}(\tau_1)) = \tau_2 - \tau_1$$

Hence, QR estimates can also be used to obtain prediction intervals for the conditional distribution of the response variable. For a given model, the interval provided by two distinct quantile estimates, $\hat{q_Y}(\tau_1, \mathbf{X}=x)$ and $\hat{q_Y}(\tau_2, \mathbf{X}=x)$, at any specified value of the regressor $\mathbf{X}$, is a $100(\tau_2-\tau_1)\%$ prediction interval for a single future observation.

[$q_Y(\theta = 0.1, X = x)$; $q_Y(\theta = 0.9, X = x)$]					
	x = 8	x = 9	x = 10	x = 11	x = 12
$y_i = 1 + 2x_i + \epsilon_i$ $\epsilon_i \sim N(0,1)$	[15.72; 18.28]	[17.72; 20.28]	[19.72; 22.28]	[21.72; 24.28]	[23.72; 26.28]
$y_i = 1 + 2x_i + \epsilon_i$ $\epsilon_i \sim LN(0,1.25)$	[17.20; 21.96]	[19.20; 23.96]	[21.20; 25.96]	[23.20; 27.96]	[25.20; 29.96]
$y_i = 1 + 2x_i - \epsilon_i$ $\epsilon_i \sim LN(0,1.25)$	[12.04; 16.80]	[14.04; 18.80]	[16.04; 20.80]	[18.04; 22.80]	[20.04; 24.80]
$y_i = 1 + 2x_i + (1 + x_i)\epsilon_i$ $\epsilon_i \sim N(0,1)$	[5.47; 28.53]	[6.18; 31.82]	[6.90; 35.10]	[7.62; 38.38]	[8.34; 41.66]

10-th percentile and 90-th percentile, $[q_Y(\tau = 0.1, X = x), q_Y(\tau = 0.9, X = x)]$, of the population conditional distribution $Y|X = x$ in correspondence with five distinct values of X

$[\hat{q}_Y(\theta = 0.1, X = x); \hat{q}_Y(\theta = 0.9, X = x)]$ **Width**						
		x = 8	x = 9	x = 10	x = 11	x = 12
$y_i = 1 + 2x_i + \epsilon_i$ $\epsilon_i \sim N(0,1)$	LS	[15.82 ; 18.42] **2.60**	[17.79 ; 20.36] **2.57**	[19.75 ; 22.31] **2.56**	[21.69 ; 24.27] **2.57**	[23.63 ; 26.24] **2.60**
	QR	[15.54 ; 18.23] **2.69**	[17.78 ; 20.22] **2.53**	[19.83 ; 22.20] **2.38**	[21.97 ; 24.19] **2.22**	[24.11 ; 26.17] **2.06**
$y_i = 1 + 2x_i + \epsilon_i$ $\epsilon_i \sim LN(0, 1.25)$	LS	[15.95 ; 23.18] **7.24**	[17.74 ; 24.89] **7.16**	[19.50 ; 26.63] **7.13**	[21.23 ; 28.39] **7.16**	[22.94 ; 30.18] **7.24**
	QR	[17.15 ; 21.65] **4.49**	[19.19 ; 23.57] **4.38**	[21.23 ; 25.49] **4.26**	[23.28 ; 27.42] **4.14**	[25.32 ; 29.34] **4.02**
$y_i = 1 + 2x_i - \epsilon_i$ $\epsilon_i \sim LN(0, 1.25)$	LS	[10.82 ; 18.05] **7.24**	[13.11 ; 20.26] **7.16**	[15.37 ; 22.50] **7.13**	[17.61 ; 24.77] **7.16**	[19.82 ; 27.06] **7.24**
	QR	[12.35 ; 16.85] **4.49**	[14.43 ; 18.81] **4.38**	[16.51 ; 20.77] **4.26**	[18.58 ; 22.72] **4.14**	[20.66 ; 24.68] **4.02**
$y_i = 1 + 2x_i + (1 + x_i)\epsilon_i$ $\epsilon_i \sim N(0,1)$	LS	[3.97 ; 32.14] **28.17**	[5.71 ; 33.59] **27.87**	[7.36 ; 35.13] **27.78**	[8.90 ; 36.78] **27.89**	[10.33 ; 38.53] **28.20**
	QR	[3.48 ; 28.13] **24.65**	[5.86 ; 31.17] **25.30**	[8.25 ; 34.20] **25.95**	[10.64 ; 37.24] **26.59**	[13.03 ; 40.27] **27.24**

Prediction intervals for the 10 illustrative models at five distinct values of X

Empirical coverage level [Nominal coverage level $(1-\alpha) = 80\%$]						
		x = 8	x = 9	x = 10	x = 11	x = 12
$y_i = 1 + 2x_i + \epsilon_i$ $\epsilon_i \sim N(0,1)$	LS	81.6	78.7	76.4	78.7	81.4
	QR	80.3	78.7	81.6	78.6	78.1
$y_i = 1 + 2x_i + \epsilon_i$ $\epsilon_i \sim LN(0,1.25)$	LS	18.2	17.8	17.1	17.1	17.9
	QR	80.6	82.1	81.2	79.2	78.4
$y_i = 1 + 2x_i - \epsilon_i$ $\epsilon_i \sim LN(0,1.25)$	LS	18.2	17.8	17.1	17.1	17.9
	QR	80.6	82.1	81.2	79.2	78.4
$y_i = 1 + 2x_i + (1+x_i)\epsilon_i$ $\epsilon_i \sim N(0,1)$	LS	26.5	38.8	74.6	95.3	99.4
	QR	80.0	81.4	81.4	79.4	78.1

Empirical coverage levels for OLS and QR prediction intervals computed using 1000 random samples extracted from each of the 4 considered models (rows of the table). The intervals are computed for five distinct values of X (columns of the table) to cover the whole range of the regressor.

The obtained percentages show how QR prediction intervals offer an empirical coverage level consistent with the nominal one for all the models, in spite of the nature of the error term. The rows for OLS prediction intervals indicate their underperformance in the case of a violation of the normal classical framework

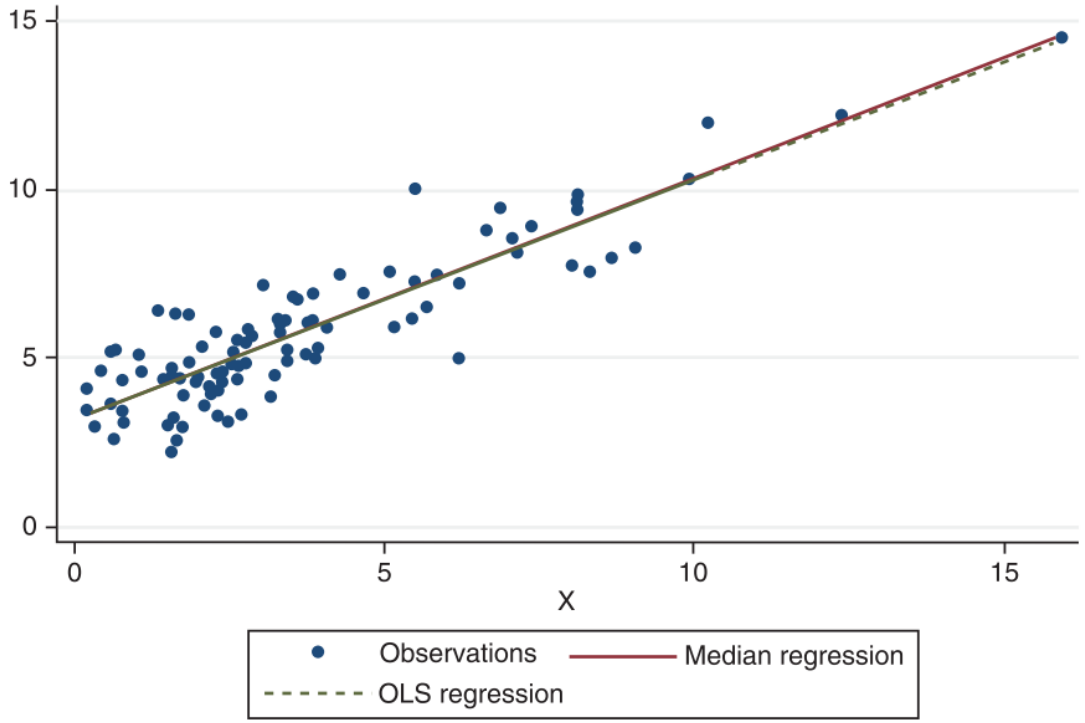
Empirical distribution of the quantile regression estimator

The case of i.i.d. errors : Consider the Linear Quantile Regression model $y_i = x_i^T \beta^\tau + \epsilon_i^\tau, i = 1, ...n$ with i.i.d error terms having a pdf f and cdf F which is strictly positive at a given quantile $f(F^{-1}(\tau))$. Then the quantile regression estimator $\hat{\beta}(\tau)$ is asymptotically distributed as

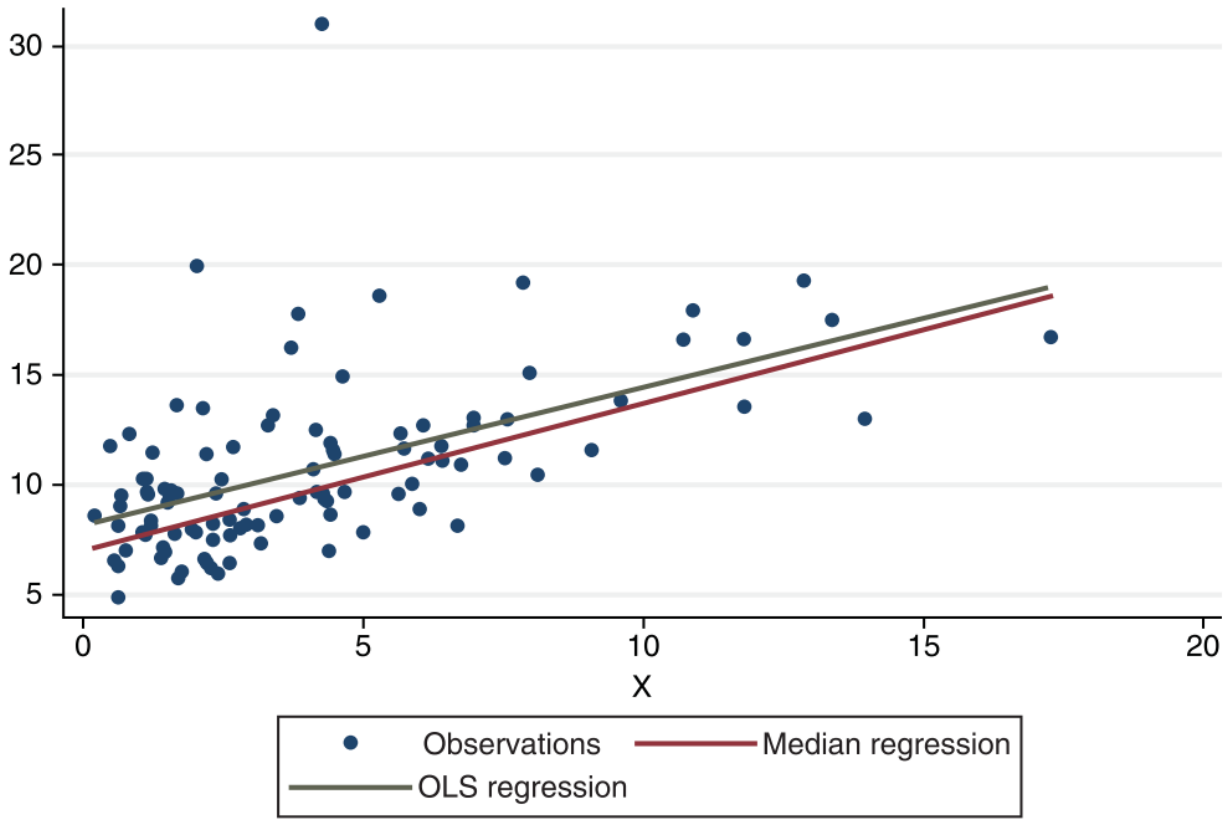
$$\sqrt{(n)}(\hat{\beta}(\tau) - \beta(\tau)) \to N(0, \omega^2(\tau)\mathbf{D}^{-1}) \quad (*)$$

with scale parameter $\omega^2(\tau) = \frac{\tau(1-\tau)}{f(F^{-1}(\tau))^2}$, being a function of $s = f(F^{-1}(\tau))$ called the sparcity function and $\mathbf{D} = \lim_{n\to\infty} \frac{\mathbf{X}^T\mathbf{X}}{n}$ is a assumed to be a positive definite matrix.

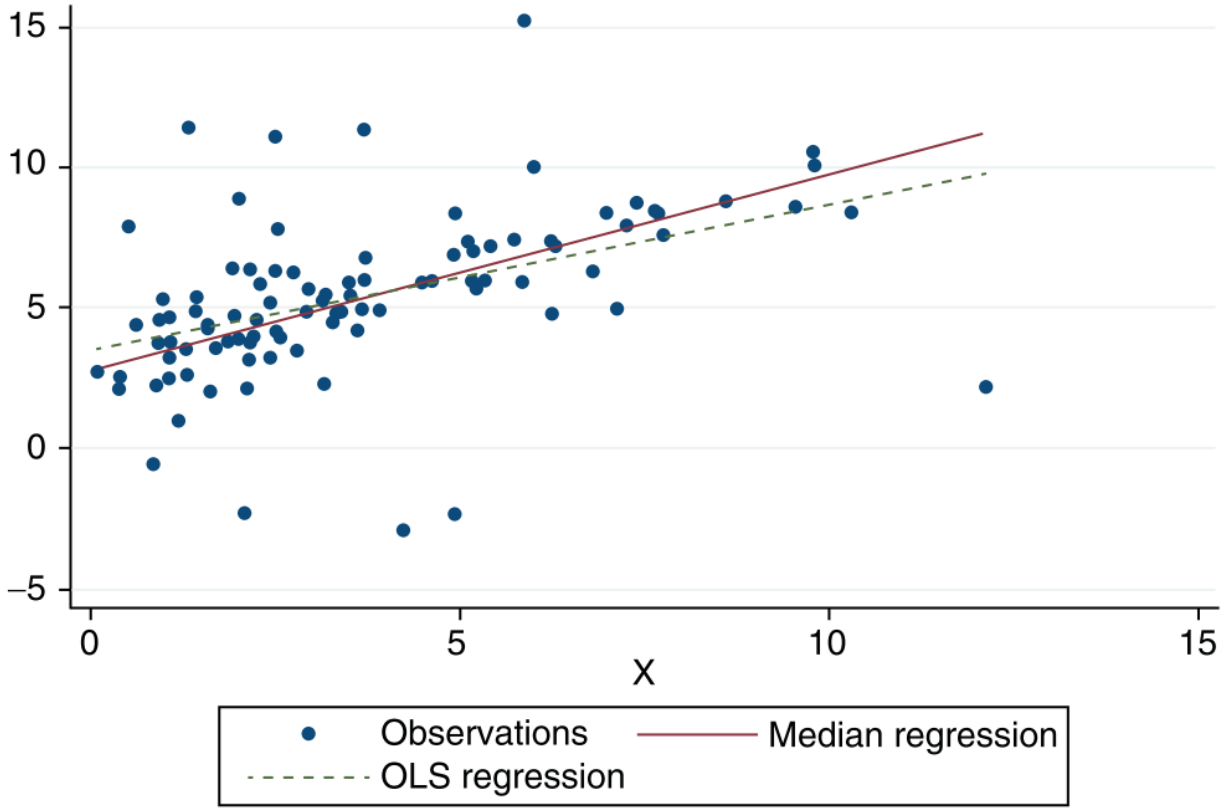
The asymptotic distribution of the QR estimator in $(*)$ is explored by means of a small simulation experiment. One hundred observations of an independent variable, x_i, are drawn from a χ^2_4 distribution and the dependent variable is computed as $y_i = 3 + 0.7x_i + e_i$ and the error term e_i follows, in turn, a standard normal distribution, a χ^2_5 distribution and a Student-t with 2 degrees of freedom.



OLS and median regression when the errors are drawn from a standard normal distribution. The difference between the median and the OLS regression is minor.

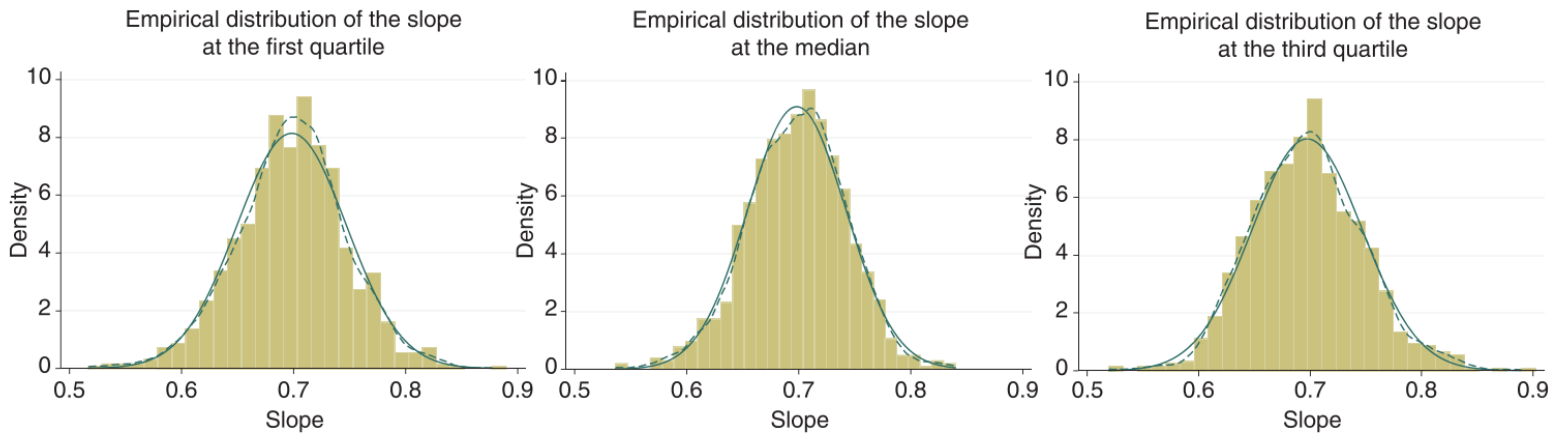


OLS and median regression when the errors are drawn from a χ^2_5 distribution. This error distribution generates five outliers in the dependent variable, as can be seen in the top section ofthe graph. The outliers shift upward the OLS fitted line but not the median regression.

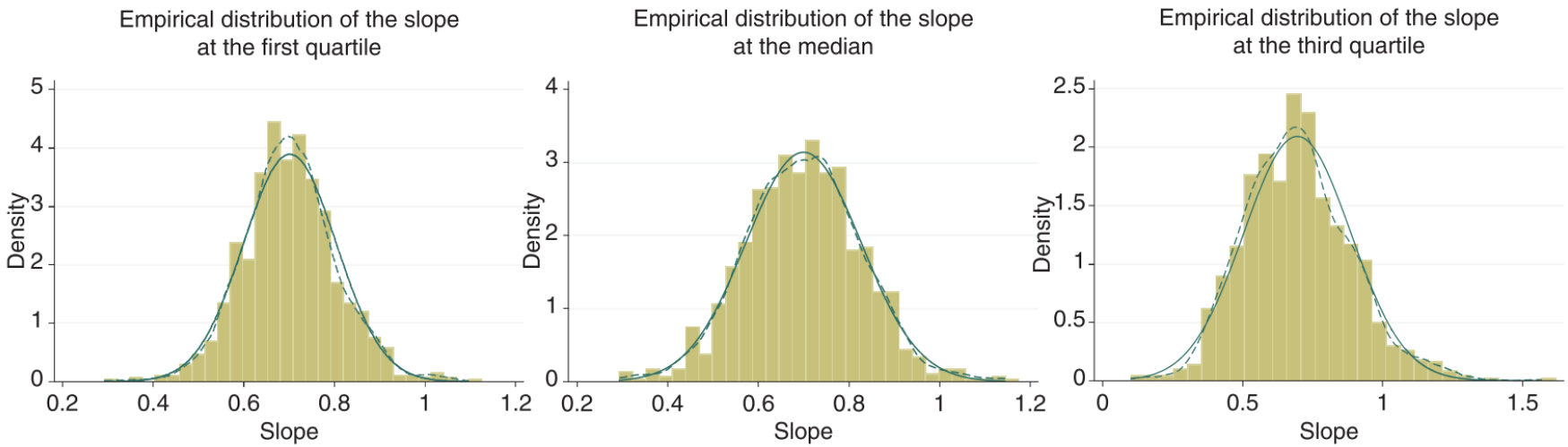


OLS and median regression when the errors are drawn from a Student-t_2 distribution. This distribution generates outliers which tilt the OLS fitted line but not the median regression

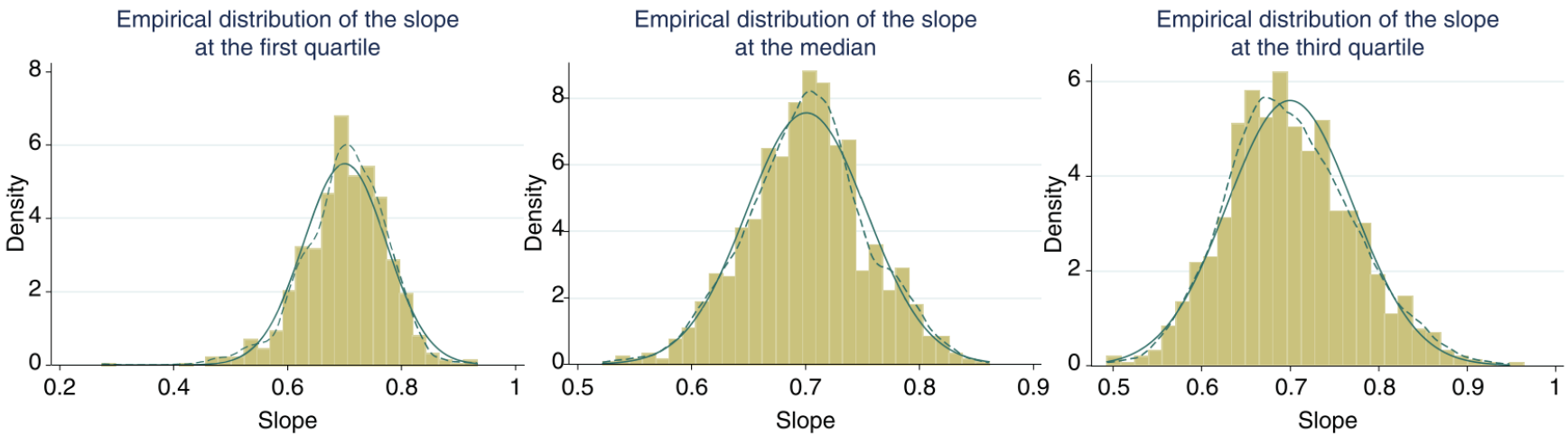
By repeating 1000 times each experiment, the estimated coefficients of each iteration are collected and the empirical distributions of the estimated slope are reported below :



Empirical distributions of the QR estimates of the slope in 1000 replicates in the case of i.i.d. standard normal errors. The solid line is the normal density and the dashed line is the kernel density. The good approximation of the empirical distributions and their kernel smoothed version to the normal density shows the asymptotic normality of the QR estimator.



Empirical distributions of the QR estimates of the slope in 1000 replicates in the case of i.i.d. errors following a χ^2_5 distribution, The solid line is the normal density and the dashed line is the kernel density. The good approximation of the empirical distributions and their kernel smoothed version to the normal density shows the asymptotic normality of the QR estimator.



Empirical distributions of the quantile regression estimated slope in 1000 replicates in the case of i.i.d. errors following a Student-t with 2 degrees of freedom. The solid line is the normal density and the dashed line is the kernel density. The good approximation ofthe empirical distributions and their kernel smoothed version to the normal density shows the asymptotic normality of the QR estimator.

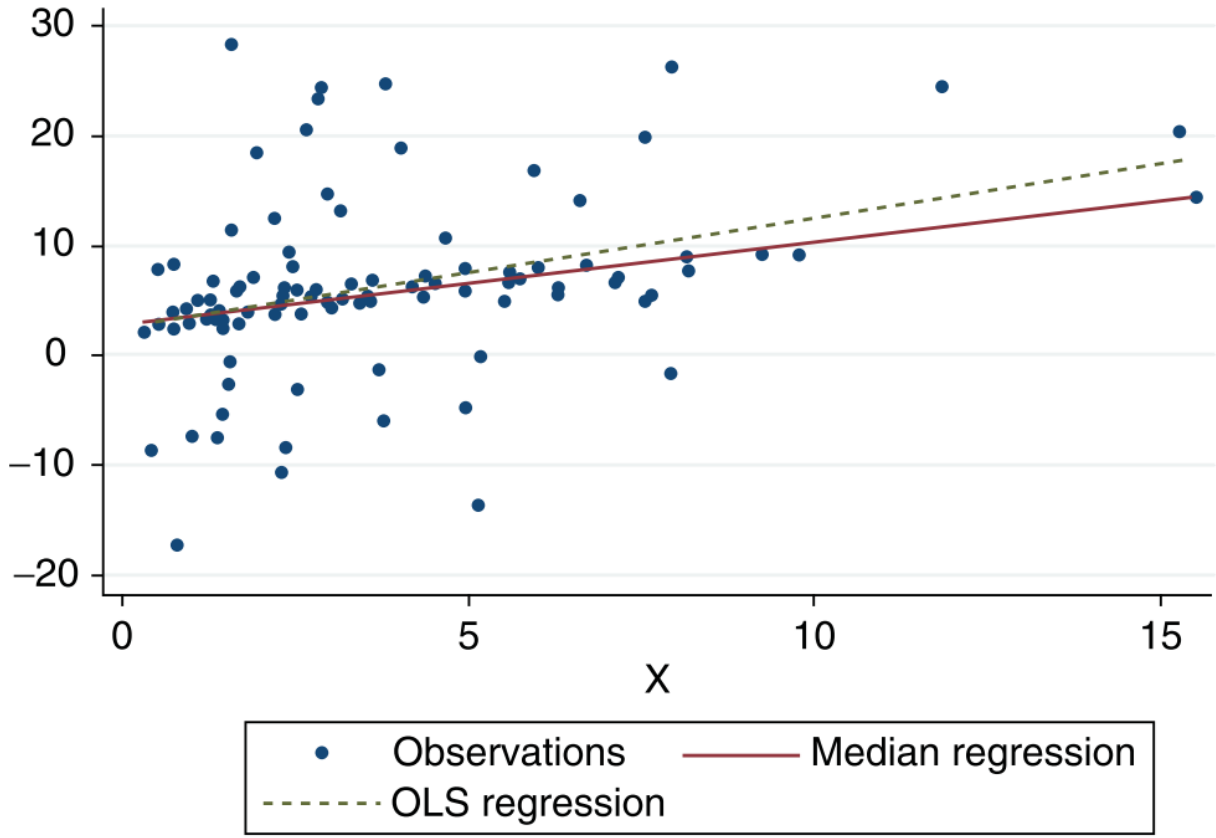
The case of i.ni.d. errors : Non-identically distributed errors are generally characterized by changing variance across the sample, which implies an error density f_i changing in the sample. In the case of non-identically distributed f_i, the asymptotic distribution of the QR estimator is given by :

$$\sqrt{(n)}(\hat{\beta}(\tau) - \beta(\tau)) \to N(0, \tau(1-\tau)\mathbf{D}_1(\tau)^{-1}\mathbf{D}\mathbf{D}_1(\tau)^{-1}),$$

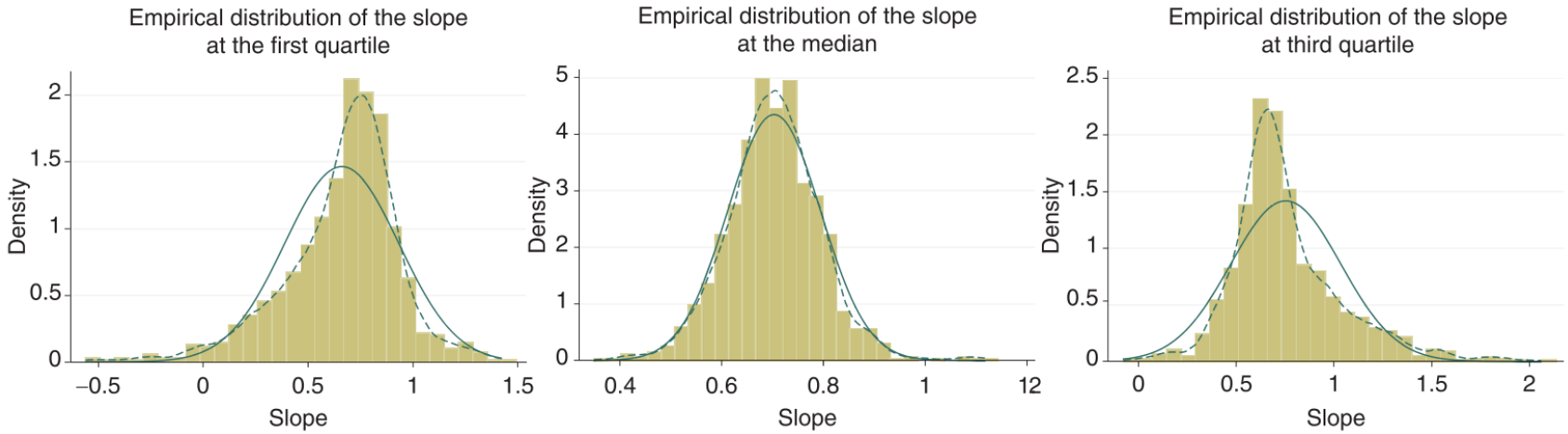
where $\mathbf{D}_1 = \lim_{n\to\infty} \frac{\sum_i f_i(F^{-1}(\tau))x_i^T x_i}{n}$

The simulations for i.ni.d. errors consider the same explanatory variable x_i following χ^2_4 but a different definition of the error term e_i. In each iteration, the first 50 observations of e_i are drawn from a standard normal while the second half of the sample follows a zero mean normal distribution having variance 100. In this subset the errors, and thus the y_{i_s}, are more dispersed than in the first half of the sample. Once again the dependent variable is defined as $y_i = 3 + 0.7x_i + e_i$.

The plot below shows the 1st iteration of the experiment in the case of i.ni.d errors.



OLS and median regression in the case of i.ni.d. errors. This error distribution generates half the observations by a standard normal, and these are the points closer to the fitted line. The remaining half of the data are generated by a $N(0, 100)$, and these are the farthest observations in the graph. The mean and median regression have differing slopes.



Empirical distributions of the QR estimated slope in 1000 replicates in the case of i.ni.d. errors. The solid line is the normal density and the dashed line is the kernel density. Away from the median the distributions become skewed, left skewed at the lower quartile and right skewed at the higher one. At the median the approximation of the empirical distribution and its kernel smoothed version to the normal density is good.

Error		Estimator			
		$\theta = 0.25$	$\theta = 0.50$	$\theta = 0.75$	OLS
$e_i \sim N(0,1)$	Mean	0.698	0.698	0.697	0.698
	S.D	0.049	0.043	0.049	**0.036**
$e_i \sim \chi^2_5$	Mean	0.702	0.699	0.695	0.699
	S.D	**0.102**	0.127	0.190	0.117
$e_i \sim t_2$	Mean	0.700	0.700	0.699	0.698
	S.D	0.073	**0.052**	0.071	0.111

Error		Estimator			
		$\theta = 0.25$	$\theta = 0.50$	$\theta = 0.75$	OLS
$e_i \sim N(0,1)$ $if\ i = 1, \ldots, 50$ $e_i \sim N(0,100)$ $if\ i = 51, \ldots, 100$	Mean	0.661	0.703	0.754	0.707
	S.D	0.272	**0.092**	0.282	0.256

Empirical distribution of the estimated slope in 1000 replicates, with $N(0,1)$, χ^2_5, t_2, non-identically distributed, and dependent errors, in the model $y_i = 3 + 0.7x_i + e_i$.

These results confirm that the QR estimator is indeed unbiased, since the sample mean is equal to the true coefficient or is very close to it. In the case of normality OLS provides the smallest standard deviations.

Wald, Likelihood Ratio and Lagrange Multiplier testss

Wald Test : Suppose we have the quantile regression model $Y = \mathbf{X}\beta + \epsilon$ and we want to test the usefulness of some of the regressors i.e we want to test

$$H_0 : \beta_{i_1} = \beta_{i_2} = ...\beta_{i_k} = 0$$

Let H be a $k \times p$ matrix such that $H(j, i_j) = 1, j = 1, ...k$ and rest all other elements are equal to 0. So our H_0 is $H\beta = 0$ where $\beta = (\beta_1, ..., \beta_p)^T$.

Now we know that $\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \to N(0, \omega^2(\tau)\mathbf{D}^{-1})$ where $D = \lim_{n\to\infty} \frac{\mathbf{X}^T\mathbf{X}}{n}$. Then $\sqrt{n}(H\hat{\beta}(\tau) - H\beta(\tau)) \to N(0, \omega^2(\tau)H\mathbf{D}^{-1}H^T)$. Under the null hypothesis $H\beta(\tau) = 0$, so we get

$$\sqrt{n}(H\hat{\beta}(\tau)) \to N(0, \omega^2(\tau)H\mathbf{D}^{-1}H^T)$$

under the null hypothesis.

We know that if $X \sim N_p(\mu, \sum)$ then $(X-\mu)^T \sum^{-1} (X-\mu) \sim \chi^2_p$. So,

$$\sqrt{n}(H\hat{\beta}(\tau))(\omega^2(\tau) H \mathbf{D}^{-1} H^T)^{-1} \sqrt{n}(H\hat{\beta}(\tau)) \to \chi^2_k$$

$$\implies n(H\hat{\beta}(\tau))(\omega^2(\tau) H \mathbf{D}^{-1} H^T)^{-1} (H\hat{\beta}(\tau)) \to \chi^2_k \quad -(*)$$

So when n is large we can use the LHS of $(*)$ as our test statistic and if

$$n(H\hat{\beta}(\tau))(\omega^2(\tau) H \mathbf{D}^{-1} H^T)^{-1} (H\hat{\beta}(\tau)) < \chi^2_k(1-\alpha)$$

, then we can safely exclude these regressors with $100(1-\alpha)\%$ confidence.

Likelihood Ratio Test : To test the same hypothesis we have another method with the test function

$$LR = 2\omega^{-1}(\tilde{V}(\tau) - \hat{V}(\tau))$$

, where $\omega^2(\tau) = \frac{\tau(1-\tau)}{f(F^{-1}(\tau))^2}$. Suppose that $y_i = x_{i,\text{restricted}}\beta_{\text{restricted}} + \epsilon$ be the restricted model then $\tilde{V}(\tau)$ and $\hat{V}(\tau)$ are defined as :

$$\tilde{V}(\tau) = \sum_{i=1}^{n} \rho_\tau(y_i - x_{i,\text{restricted}}\hat{\beta}_{\text{restricted}})$$

$$\hat{V}(\tau) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\hat{\beta} + \epsilon)$$

Under the null hypothesis LR follows χ^2_k distribution where k is the number of regressors under test. If $LR < \chi^2_k(1-\alpha)$ then we can safely exclude these regressors.

Lagrange Multiplier Test : The LM test is implemented by estimating an auxiliary regression. The residuals of the constrained model become the dependent variable of an additional regression having as explanatory variables those regressors excluded from the model. The term nR^2 is asymptotically χ^2 with degrees of freedom equal to the number of variables under test. The auxiliary regression checks if the excluded regressors have any explanatory content that would be lost once they were eliminated from the main equation. If the variables are erroneously excluded, they will explain at least part of the residuals from the main equation, the auxiliary regression will have a large nR^2 and the null on the validity of the constraints will be rejected. Coversely, if the regressors under test are superfluous, in the auxiliary equation the nR^2 term is small and the null is not rejected.

Estimating the variance of quantile regressions

In the Quantile regression model $Y = \mathbf{X}\beta + \epsilon$, when the error terms are iid we know that $\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \to N(0, \omega^2(\tau)\mathbf{D})$ where the scale parameter of the model $\omega^2(\tau)$, at the selected quantile τ,is defined as $\omega^2(\tau) = \frac{\tau(1-\tau)}{f(F^{-1}(\tau))^2}$. The term $s(\tau) = f(F^{-1}(\tau))$ is unknown and has to be estimated. Many different estimators have been proposed and one of them is :

$$\hat{s}(\tau) = \frac{F^{-1}(\tau+h) - F^{-1}(\tau-h)}{2h}$$

The sparsity function $s(\tau)$ can be computed by differentiating the quantile function $F^{-1}(\tau), s(\tau) = \frac{d}{d\tau}(F^{-1}(\tau)) = \frac{1}{f(F^{-1}(\tau))}$ and thus it represents the slope of the tangent to the quantile function at point τ. This slope can be approximated by the slope of the secant to the quantile function at points $t + h$ and $t - h$.

The value of h as suggested by Koenker is

$$h = n^{-0.2}\left[\frac{4.5\phi^4(\Phi^{-1}((\tau))}{(2\Phi^{-1}(\tau)^2+1)^2}\right]^{0.2}$$

Now we need to estimate the quantile function $F(\tau \pm h)$. We can use the residuals $y_i - x_i^T\hat{\beta}$ to estimate $F^{-1}(\tau)$.

thank you!