Quantile Regression: A Study

Arvind Raghavendran
ISI Bangalore
B.Math 3^{rd} year

August 19, 2021

Contents

1 Introduction

1.1 Regression and OLS

Simply defined, regression is a statistical procedure to establish the relationship between a dependent variable some independent variables. The former is called an outcome, and the latter is called a covariate/predictor. Let us call the outcome as Y and the predictors as X_i's. The fundamental concept in regression, is how we predict $Y|X$. We are familiar with the most common form of regression, linear regression. Let us get a quick overview of this.

In linear regression we model $\mathrm{E}[Y|X]$ as a linear model of the covariates, and we solve for the coefficients by applying a specific algorithm. Formally speaking,
Suppose $X_1, X_2, ..., X_n$ are the covariates, and Y is our outcome. Then,

$$\mathrm{E}[Y|X_1, ..., X_n] = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

Suppose y_i's are our actual observed outcomes, and $x_{1i}, ..., x_{ni}$ are the corresponding covariates. Then,

$$y_i = \mathrm{E}[Y|x_{1i}, ..., x_{ni}] = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} + \varepsilon_i$$

where, ε_i is the error in measurement. There are many algorithms that can be used to estimate the coefficients. The most common method is Ordinary Least Squares, i.e., minimizing the function:

$$\sum_i (y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij})^2$$

If we denote the corresponding estimates of β_i as $\hat{\beta}_i$, then the model is given by

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{n} \hat{\beta}_j x_{ij}$$

$\mathrm{E}[Y|X_1, ..., X_n]$ is the best predictor for Y given $X_1, X_2, ..., X_n$; i.e., the minimizer of $\mathrm{E}[(Y-a)^2|X_1, ..., X_n]$ is $\mathrm{a} = \mathrm{E}[Y|X_1, ..., X_n]$ under any distribution for $Y|X_1, ..., X_n$. This is the linear regressor under a linear model for $\mathrm{E}[Y|X_1, ..., X_n]$. That directly implies that the assumption is $\mathrm{E}[Y|X_1, ..., X_n]$ has a linear model based on the predictors $X_1, X_2, ..., X_n$. Since $\mathrm{E}[Y|X_1, ..., X_n]$ is being modeled as a linear function, linear regression is called mean regression (i.e., linear model for the mean). The usage of the least squares is just a convenient mathematical algorithm that provides the best fitting linear model.

1.2 Should the modelling always be done using the mean?

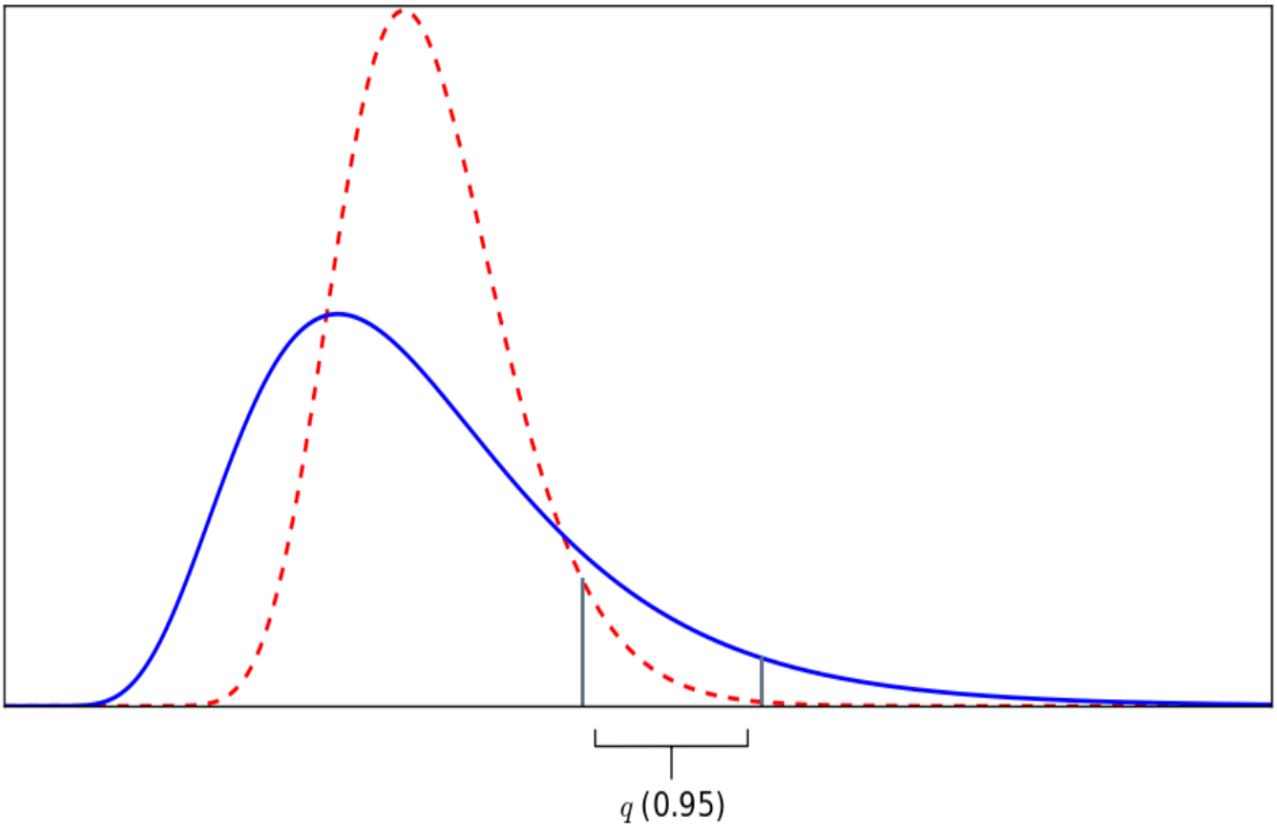
But is mean always the best estimate? Sure, in most of the cases, mean is the best estimate, but that is not the case all the time. For example, consider the following scenario. Suppose we are given the data of household income in an area, and we are asked to report an estimate of the income available to the majority of households. Now one could report the average, but this needn't be accurate all the time. For instance, where there happen to be a small number of extremely high income households and a large number of lower income households, the average income can become inflated and give the false impression that all households on average have more money than they actually have. So in that case, the median is the better estimate, since it won't be affected by high-income outliers.

So this provides a possible scenario where we'd like to use the median as the best estimate, rather than the mean. But what about the other quantiles? Consider the following case. Suppose a taxi company develops a new algorithm to dispatch taxis, say like how Ola/Uber work. One way to check the relative efficacy of the algorithm, is to model the conditional average wait time given the algorithm, that is,

$$\mathrm{E}[\text{Wait Time}|\text{Algorithm}] = \beta_0 + \beta_1 \text{Algorithm}$$

where algorithm is binary, i.e., 1 if new algorithm and 0 if old algorithm. Now the average wait time under the old algorithm(W1) is β_0 and under the new algorithm(W2) is $\beta_0 + \beta_1$. For the model to be better, we'd need

W2<W1, and hence, $\beta_1 < 0$. Now suppose that for a particular sample, $\hat{\beta}_1$ does come out to be negative, then we can infer that on an average, the customers would get taxis dispatched to their location faster than before. But what about the rest? While the average documents a general picture, what if the picture becomes worse for a few people. For example, suppose a particular customer has been getting his taxis late for the last two times and the new algorithm now makes things worse, i.e., his taxi comes even later. That would just make him lose faith in the company resulting in a potential loss of a customer. This would be highly desirable, albeit providing a better overall picture. More concretely speaking, what if the new algorithm improves wait times for 95% of customers by 1 min, but makes the wait times for the remaining 5% longer by 5 min? Overall, we would see a decrease in mean wait time, but things got significantly worse for a segment of the population. What if that 5% whose wait times became 5 minutes longer were already having the longest wait times to begin with? The conditional-mean model would not have picked this up. One way to pick up such situations is to model conditional quantile functions instead. In our example above, instead of trying to estimate the mean wait time, we could estimate the 95^{th} quantile wait time to catch anything going wrong out in the tails of the distribution.



2 Linear Regression vs Quantile Regression

2.1 What are Quantiles?

In the example in 1.2, we saw a that median was a better estimate than the mean. We also saw that with a skewed distribution, the median may become the more appropriate measure of central tendency; therefore, conditional-median regression, rather than conditional-mean regression, should be considered for the purpose of modeling location shifts. To model both location shifts and shape shifts, Koenker and Basset proposed a more general form than the median-regression model, the QRM. The QRM estimates the potential differential effect of a covariate on various quantiles in the conditional distribution, for example, a sequence of 19 equally distanced quantiles from the $.05^{th}$ quantile to the $.95^{th}$ quantile. Before we move on any further, let us recollect a little but about quantiles.

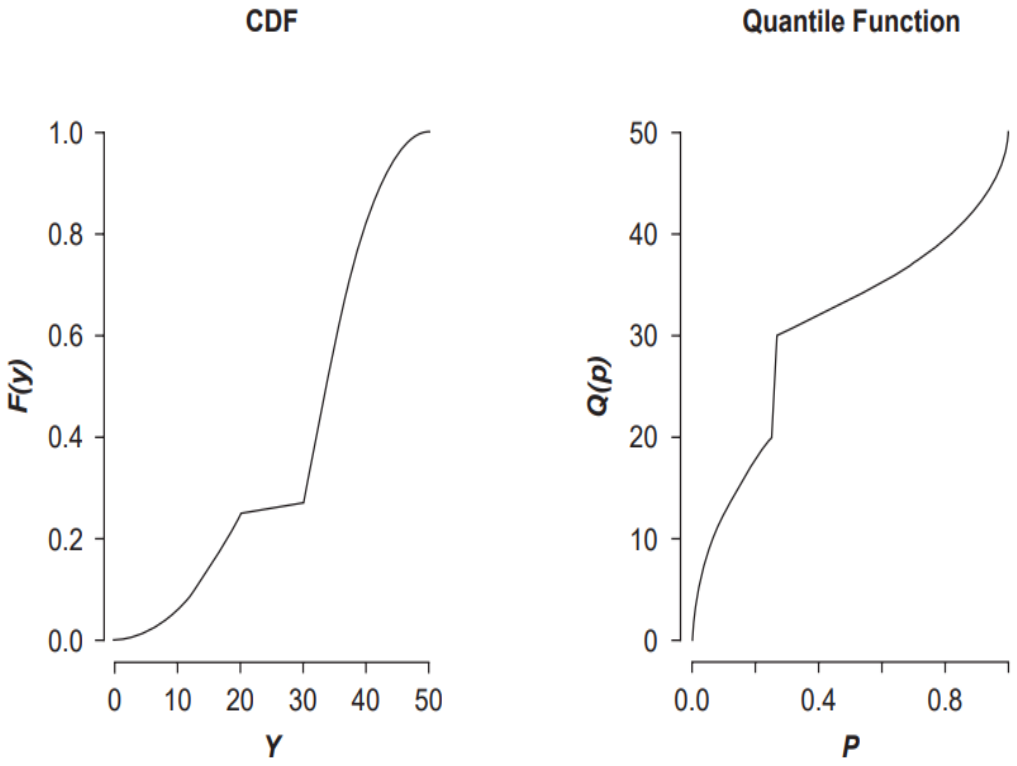
Recall that quantiles are points in a distribution that relate to the rank order of values in that distribution. In other words, quantiles divide the observations in a sample in the same way. In general, in a sample, q-quantiles are values that partition a finite set of values into q subsets of (nearly) equal sizes. There are q-1 q-quantiles, one for each integer k satisfying $0 < k < q$. In some cases the value of a quantile may not be uniquely determined, for example, 2-quantile (median) of a sample of even size. Some quantiles have special names:

- median (2-quantile)
- quartile(4-quantile)
- decile(10-quantile)
- percentile(100-quantile).

2.2 Sample Quantile Distribution

Often we work with data that are assumed to belong to a certain distribution, so it is important to know how quantiles are defined w.r.t. to the CDF of the data.

The pth quantile $Q^{(p)}(F)$ of a cdf F is defined as $\min\{y|F(y) \geq p\}$. The function $Q^{(p)}$ (as a function of p) is referred to as the quantile function of F. Given below is a small example of a CDF vs a Quantile function plot:



Given a sample $y_1.y_2, ..., y_n$, we define its p^{th} sample quantile $\hat{Q}^{(p)}$ to be the p^{th} quantile of the corresponding empirical CDF $\hat{F}$, i.e., $\hat{Q}^{(p)} = Q^{(p)}(\hat{F})$. The corresponding quantile function is called as the sample quantile function.
We can see the connection between the order statistics of the data and the sample quantile. For a sample of size n, the $(k/n)^{th}$ sample quantile is given by $y_{(k)}$, the k^{th} order statistic.

It is important to note how sample quantiles behave in large samples. For a large sample $y_1.y_2, ..., y_n$ drawn from a distribution with quantile function $Q^{(p)}$ and probability density function $f = F'$, the distribution of $\hat{Q}^{(p)}$ is approximately normal with mean $Q^{(p)}$ and variance $\frac{p(1-p)}{n}.\frac{1}{f(Q^{(p)})^2}$. The dependence on the density at the quantile has a simple intuitive explanation: If there are more data nearby (higher density), the sample quantile is less variable; conversely, if there are fewer data nearby (low density), the sample quantile is more variable.

In general for a sample, the pdf is unknown to us. So we need a method to estimate it. We can use the fact that $\frac{d}{dp}Q^{(p)} = 1/f(Q^{(p)}$. The slope can be approximated by $\frac{1}{2h}(\hat{Q}^{(p+h)} - \hat{Q}^{(p-h)})$, for some small value of h. The choice of h differs, and Koenker (cite) suggests a couple of methods to choose one.

2.3 Necessity of QR

The Linear Regression Model (LRM) is a standard statistical method widely used in social-science research, but it focuses on modeling the conditional mean of a response variable without accounting for the full conditional distributional properties of the response variable. In order to see why the topic of QR became a subject of interest, we must know the shortcomings of the LRM. In order to know these, we have to look at the assumptions inherent in an LRM. While these assumptions ensure that a relatively easy algorithm (OLS) can solve the LRM, it might not convey the entire picture. Let us see why. Assume that we have just 1 covariate, i.e., our LRM looks like:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

1. **Residuals have mean 0**: This is the assumption that defines the LRM. As a consequence of the mean being 0, the equation $\beta_0 + \beta_1 x_i$ corresponds to $\mathrm{E}[Y|x_i]$. Suppose we are modeling Height vs Weight for some students, and for a particular value of weight, say 50 kg,the estimated height from the LRM is 145 cm. This means that the average height of students who are 50 kg is 145 cm. Therefore, a fundamental aspect of LR is that it utilizes the mean of the distribution as it's central tendency and attempts to describe the conditional distribution of the data. In contrast, the Quanitle Regression Model (QRM) facilitates analysis of the full conditional distributional properties of the response variable. What it means that, unlike the case of LRM, one could describe the entire distribution of the data if all the quantiles can be modeled using a QRM.

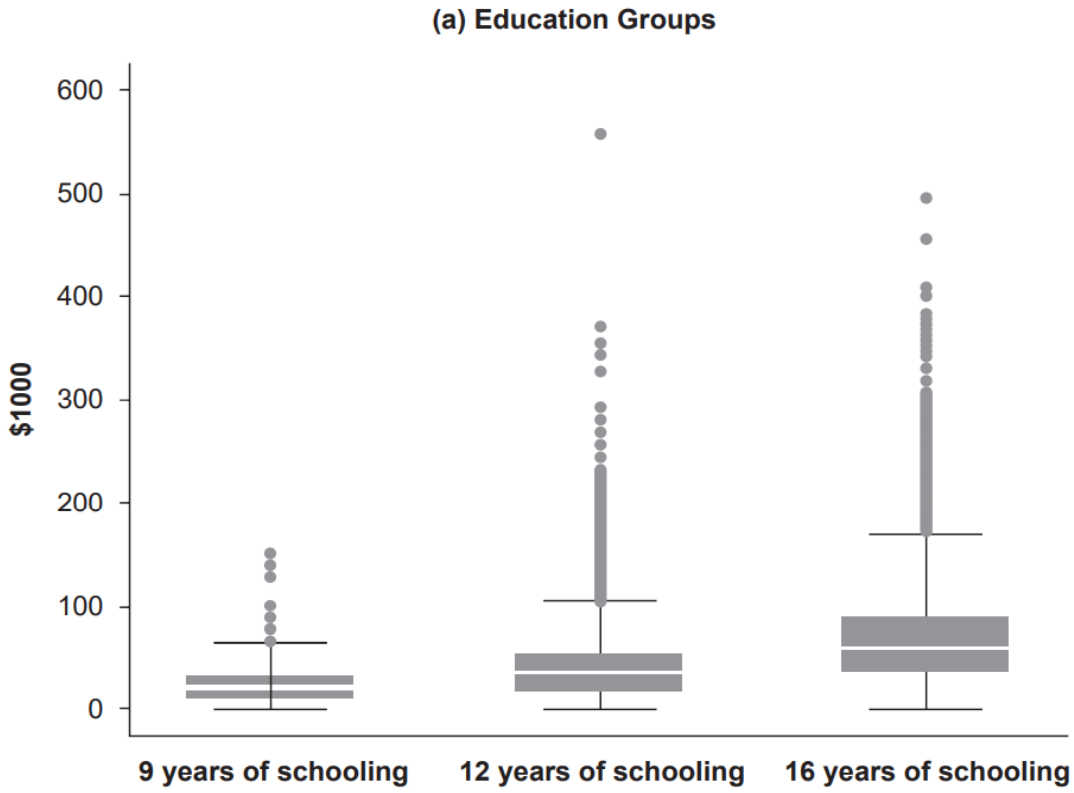
2. **Homoscedasticity**: Homoscedasticity means that the residuals have constant variance. that is, the error term does not vary much as the value of the predictor variable changes. Another way of saying this is that the variance of the data points are roughly the same for all data points. However, the lack of homoscedasticity, or heterescedasticity may suggest that the regression model may need to include additional predictor variables to explain the performance of the dependent variable. Now, since the variance of the residual ε_i is in fact $\mathrm{Var}[Y|x_i]$, the scedasticity of the model plays an important role if scaling is involved. The most practical reason why one would like to scale their variables, is if one is working with a variable having a very large scale, for e.g., salt in the ocean. This value would be in billions, so the coefficients come out to be very small, hence making it hard to use the values in a computer simulation. So we would re-scale the variable into a more convenient unit, like parts per billion (ppb). Sometimes, we might not learn anything from the data using LRM, even if it's homoscedastic. For example suppose we have a response Y and a covariate X and suppose it is found that

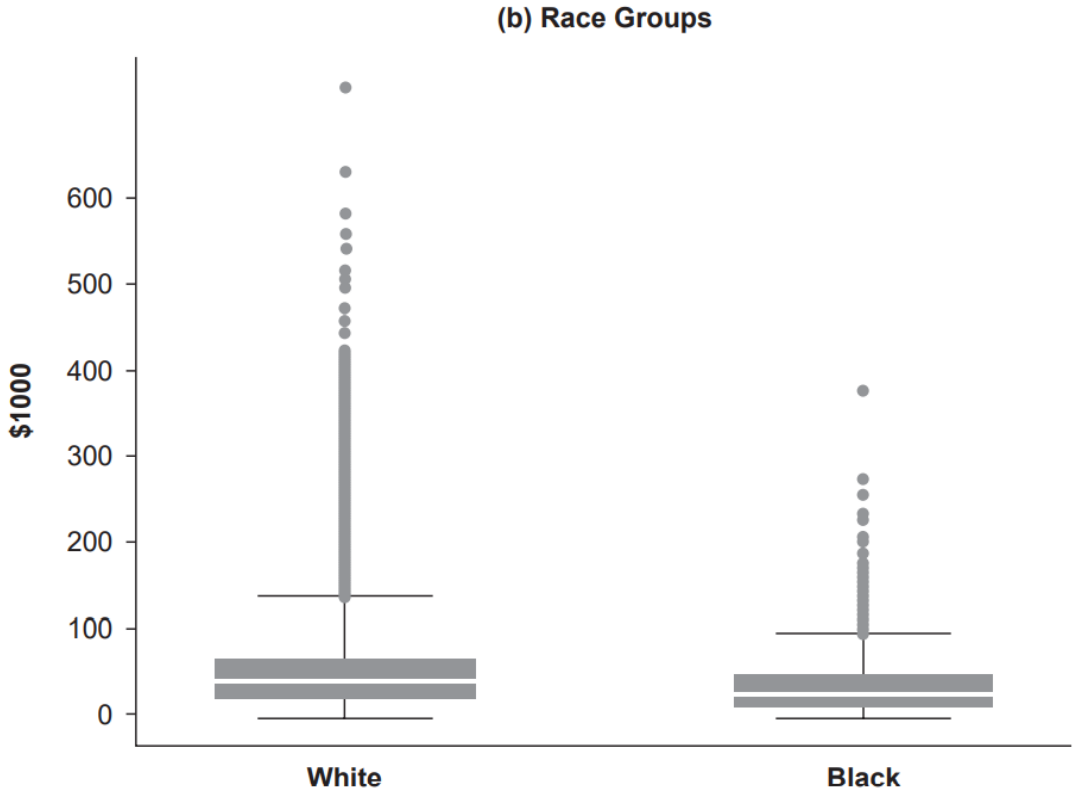
$$\begin{aligned}\mathrm{E}[Y|x_1] = \mathrm{E}[Y|x_2] = \mathrm{E}[Y|x_3]\\ \mathrm{Var}[Y|x_1] = \mathrm{Var}[Y|x_2] = \mathrm{Var}[Y|x_3]\end{aligned}$$

   Since the conditional mean and scale for the response variable Y do not vary with X, there is no information to be gleaned by fitting a LRM to samples from these populations. In order to understand how the covariate affects the response variable, a new tool is required. Quantile regression is an appropriate tool for accomplishing this task.

3. **One-Model Assumption**: A related assumption made in the LRM is that the regression model used is appropriate for all data, which we call the one-model assumption. Outliers in the LRM tend to have undue influence on the fitted regression line, since the mean is affected heavily by any outliers. The usual practice used in the LRM is to identify outliers and eliminate them. Now sometimes the outliers might be an important aspect of the data, for example, in social-science research, particularly studies on social stratification and inequality, the outliers and their relative positions to those of the majority are important aspects of inquiry. So if we were to model this, one would simultaneously need to model the relationship for the majority cases and for the outlier cases, a task the LRM cannot accomplish.

Consider the following plots depicting the distribution of Household Income vs Race and Education:





- The location shifts among the three education groups and between blacks and whites are obvious, and their shape shifts are substantial. Therefore, the conditional mean from the LRM would fail to capture the shape shifts caused by changes in the covariate (education or race). Moreover, all box graphs are right-skewed. Conditional-mean and conditional-scale models are not able to detect these kinds of shape changes.

- Since the spreads differ substantially among the education groups and between the two racial groups, the homoscedasticity assumption is violated, and the standard errors are not estimated precisely.
- In the first set of plots, there are 7 outliers. If we were to add a dummy variable to model these outliers into the LRM, we would see that the contribution by these outliers amount to $483,544. This would have clearly been missed by the LRM under usual assumptions.

So, the LRM approach need not be the best approach in this case. In fact, there are a variety of related examples where one cannot simply model using LR. Now naturally it would strike that if not the mean, maybe the median could be a good fit. As it turns out, the conditional-mean models also do not always correctly model central location shifts if the response distribution is asymmetric. For a symmetric distribution, the mean and median coincide, but the mean of a skewed distribution is no longer the same as the median. Observe the following table:

Household Income Distribution:
Total, Education Groups, and Racial Groups

	Total	*ED = 9*	*ED = 12*	*ED = 16*	*WHITE*	*BLACK*
Mean	50,334	27,841	40,233	71,833	53,466	35,198
Quantile						
Median (.50th Quantile)	39,165	22,146	32,803	60,545	41,997	26,763
.10th Quantile	11,022	8,001	10,510	21,654	12,486	6,837
.25th Quantile	20,940	12,329	18,730	36,802	23,198	13,412
.75th Quantile	65,793	36,850	53,075	90,448	69,680	47,798
.90th Quantile	98,313	54,370	77,506	130,981	102,981	73,030
Quantile-Based Scale						
$(Q_{.75}-Q_{.25})$	44,853	24,521	34,344	53,646	46,482	34,386
$(Q_{.90}-Q_{.10})$	87,291	46,369	66,996	109,327	90,495	66,193
Quantile-Based Skewness						
$\frac{(Q_{.75}-Q_{.50})}{(Q_{.50}-Q_{.25})}-1$	.46	.50	.44	.26	.47	.58
$\frac{(Q_{.90}-Q_{.50})}{(Q_{.50}-Q_{.10})}-1$	1.10	1.28	1.01	.81	1.07	1.32

Simply by observing from the box-plots, or as shown above, the distributions of Income|Education and Income|Race are right-skewed. The right-skewness makes the mean considerably larger than the median for both the total sample and for education and racial groups. When the mean and the median of a distribution do not coincide, the median may be more appropriate to capture the central tendency of the distribution. The location shifts among the three education groups and between blacks and whites are considerably smaller when we examine the median rather than the mean. This leads us to think whether it is possible to use the median as an estimator, and if not, a general quantile.

3 Quantile Regression

3.1 Regression using quantiles

In the case on mean regression, we saw that we could use least squares to find the best estimate for a linear model under the assumption that mean is the best estimate. Is there a mathematical algorithm that does the same when the median is the best estimate? The answer is yes, this can be done. We model median$(Y|X)$ as a linear model of the covariates, and we solve for the coefficients by minimizing the sum of absolute value of the residuals.

3.1.1 Mean Regression Model and extension to a QRM

Suppose $X_1, X_2, ..., X_n$ are the covariates, and Y is our outcome. Then,

$$\text{median}[Y|X_1, ..., X_n] = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

Suppose y_i's are our actual observed outcomes, and $x_{1i}, ..., x_{ni}$ are the corresponding covariates. Then,

$$y_i = \text{median}[Y|x_{1i}, ..., x_{ni}] = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} + \varepsilon_i$$

where, ε_i is the error in measurement.
We can extend this in general for any quantile as well. If we wish to model the p^{th} quantile as the best estimate, then the expression would be similar to as the case of median, except we would model y_i as $Q^{(p)}[Y|x_{1i}, ..., x_{ni}]$, i.e.,

$$y_i = Q^{(p)}[Y|x_{1i}, ..., x_{ni}] = \beta_0^{(p)} + \sum_{j=1}^{n} \beta_j^{(p)} x_{ij} + \varepsilon_i$$

where $Q^{(p)}[Y]$ is the p^{th} quantile of Y.

From here on, our discussion will pertain to the general p^{th} QRM, unless specified otherwise.

3.1.2 Residuals

The assumptions on $\varepsilon_i^{(p)}s$ are different than those in the case of linear regression. Recall that in an LRM, we assume that $\text{E}[\varepsilon_i] = 0$. This comes from the fact that we model $\text{E}[Y|x_i]$ to be linear. Analogously, we can say that in a QRM, since we model $Q^{(p)}[Y|x_i]$ to be linear, we need $Q^{(p)}[\varepsilon_i] = 0$.

Given some data, it is possible to construct just one LRM, since there can be only one possible mean. This is not the case with QRM, since there are multiple choices of p in (0,1). For example, if the QRM specifies 10 quantiles, the 10 equations yield 10 coefficients for x_i ,one at each of the 10 conditional quantiles
($\beta_1^{0.1}, \beta_1^{0.2}, ..., \beta_1^{0.9}$). The quantiles do not have to be equidistant, but in practice, having them at equal intervals makes them easier to interpret.

Suppose, in the simple QRM, we choose $p, q \in$ (0,1) where $p \neq q$. We would have the equations,

$$y_i = \beta_0^{(p)} + \beta_1^{(p)} x_i + \varepsilon_i^{(p)}$$
$$y_i = \beta_0^{(q)} + \beta_1^{(q)} x_i + \varepsilon_i^{(q)}$$

Equating and rearranging would give us,

$$\varepsilon_i^{(q)} - \varepsilon_i^{(p)} = \beta_0^{(p)} - \beta_0^{(q)} + x_i(\beta_1^{(p)} - \beta_1^{(q)})$$

Therefore, given x_i, we have that the distributions of $\varepsilon_i^{(p)}$ and $\varepsilon_i^{(q)}$ are shifts of each other. Under the assumption that $\forall\, i$, $\varepsilon_i^{(p)}$ are i.i.d, the q^{th} quantile of $\varepsilon_i^{(p)}$ is a constant depending on p, q but not i, say, $c_{p,q}$. Then, we get that,

$$Q^{(q)}[Y|x_i] = Q^{(p)}[Y|x_i] + c_{p,q}$$

So we can conclude that if the residuals are i.i.d distributed, then we get that the conditional quantiles of $Y|x_i$ are just shifts of one another, and that the slope β_1^p all take a common value, say β_1; or equivalently, there is not shape shift in the response variable.

3.2 LRM vs QRM : An Example

Previously, we discussed how a QRM could improve the shortcomings of an LRM. Let us take an example and see this. Recall the previous example of Household Income vs Race and Education. A 1000 random data points were chosen and treated as the sample. For the QRM, we choose p as 0.1,0.2,....,0.19. The scatterplot of the data is plotted, upon which the regression lines are drawn. Before we display the plots, let's look at the estimated coefficients for the 19 quantiles:

Quantile-Regression Estimates for Household Income on Education

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
ED	1,019	1,617	2,023	2,434	2,750	3,107	3,397	3,657	3,948	4,208	4,418	4,676	4,905	5,214	5,557	5,870	6,373	6,885	8,385
	(28)	(31)	(40)	(39)	(44)	(51)	(57)	(64)	(66)	(72)	(81)	(92)	(88)	(102)	(127)	(138)	(195)	(274)	(463)
Constant	−4,252	−7,648	−9,170	−11,160	−12,056	−13,308	−13,783	−13,726	−14,026	−13,769	−12,546	−11,557	−9,914	−8,760	−7,371	−4,227	−1,748	4,755	10,648
	(380)	(424)	(547)	(527)	(593)	(693)	(764)	(866)	(884)	(969)	(1,084)	(1,226)	(1,169)	(1,358)	(1,690)	(1,828)	(2,582)	(3,619)	(6,101)

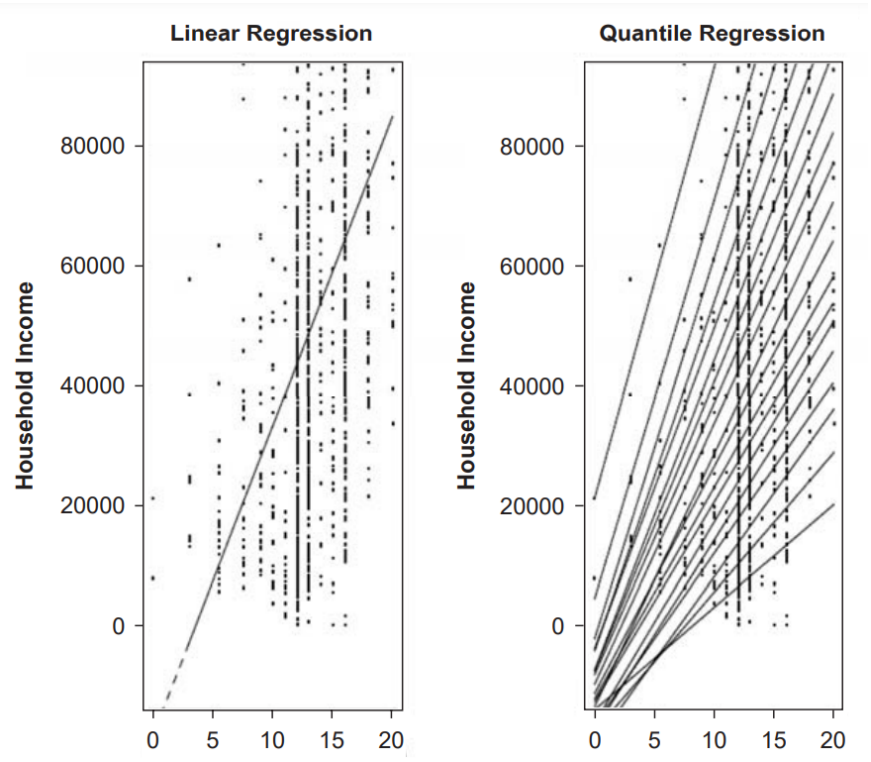
NOTE: Standard errors in parentheses.

Quantile-Regression Estimates for Household Income on Race

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
BLACK	−3,124	−5,649	−7,376	−8,848	−9,767	−11,232	−12,344	−13,349	−14,655	−15,233	−16,459	−17,417	−19,053	−20,314	−21,879	−22,914	−26,063	−29,951	−40,639
	(304)	(306)	(421)	(485)	(584)	(536)	(609)	(708)	(781)	(765)	(847)	(887)	(1,050)	(1,038)	(1,191)	(1,221)	(1,435)	(1,993)	(3,573)
Constant	8,556	12,486	16,088	19,718	23,198	26,832	30,354	34,024	38,047	41,997	46,635	51,515	56,613	62,738	69,680	77,870	87,996	102,981	132,400
	(115)	(116)	(159)	(183)	(220)	(202)	(230)	(268)	(295)	(289)	(320)	(335)	(397)	(392)	(450)	(461)	(542)	(753)	(1,350)

NOTE: Standard errors in parentheses.

Now let's look at the plots with the regression fits.



On the left we have the LRM. The regression line indicates the mean shifts, for example, there is a mean shift of \$22,532 in the household income from 12 years of schooling to 16 years of schooling,i.e., the slope of the line is \$5,633. However, as we noted before, this line cannot capture the shape shifts. On the right, we have the QRM, with the 19 lines depicting (from the left) the 19 quantiles chosen in ascending order. Observe the fit for the 0.5^{th} quantile, i.e., the median. It captures the central location shifts, indicationg a positive relationship between the conditional median-income and education. The slope of this line is \$4,208, a lower shift than the case ofthe LRM. In addition to the estimated location shifts, the other 18 quantile-regression lines provide information about shape shifts. These regression lines are positive, but with different slopes. The regression lines cluster tightly at low levels of education (e.g., $0-5$ years of schooling) but deviate from each other more widely at higher levels of education (e.g., $16-20$ years of schooling). A shape shift is described by the tight cluster of the slopes at lower levels of education and the scattering of slopes at higher levels of education. For instance, the spread of the conditional income on 16 years of schooling (from \$12,052 for the .05th conditional quantile to \$144,808 for the .95th conditional quantile) is much wider than that on 12 years of schooling (from \$7,976 for the .05th conditional quantile to \$111,268 for the .95th conditional quantile). Thus, the off-median conditional quantiles isolate the location shift from the shape shift. This feature is crucial for determining the impact of a covariate on the location and shape shifts of the conditional distribution of the response. We will discuss this in detail later on. The takeaway from this example, was that the information that can be obtained from a QRM is much more detailed than what we can obtain from a simple LRM.

3.3 Model Estimation

Now that we have seen how good QR can be, it is important to understand how to build a QRM. This involves estimating the coefficients of the hypothesized regression equations, i.e., $\hat{\beta}_i$. Let us recall what happens in OLS. As usual, we will pertain to discussing about a simple LRM and a simple QRM. In OLS, one would estimate the coefficients β_0 and β_1 by minimizing the expression,

$$\sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

The estimates are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$. If the LRM assumptions are correct, the fitted response function $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ approaches the population conditional mean $\mathrm{E}(Y|X)$ as the sample size goes to infinity. We can solve the above equation using partial derivates wrt β_1 and $\beta 0$ and solving it, upon which we'll get:

$$\begin{aligned}
\beta_0 &= \bar{y} - \beta_1 \bar{x} \\
\beta_1 &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}
\end{aligned}$$

Analogously, this can be done for a QRM too. A significant departure of the QR estimator from the LR estimator is that in the QR, the distance of points from a line is measured using a weighted sum of vertical distances (without squaring), where the weight is $1-p$ for points below the fitted line and p for points above the line, assuming that we are modeling using p^{th} quantile. Formally speaking, the coefficients $\beta_i^{(p)}$ can be estimated by minimizing the expression

$$\sum_i \rho_p (y_i - \beta_0^{(p)} - \sum_{j=1}^{n} \beta_j^{(p)} x_{ij})$$

where ρ is defined as

$$\rho_p(t) = \frac{|t| + (2p-1)t}{2} \qquad p \in (0,1)$$

ρ is called as the check function corresponding to the p^{th} quantile. Intuitively put, ρ is defined by assigning asymmetric weights to the residuals. If the weights are symmetric, then the answer would be the median, and this can be seen if $p = 0.5$ Since ρ is not differentiable at 0, it not easy to obtain a closed form for the estimators $\hat{\beta}_i^{(p)}$. However, it is solvable using linear programming, specifically, an algorithm proposed by Koenker and D'Orey in 1987. Once we estimate the coefficients given the sample, then under the appropriate model assumptions, we theorize that as the sample size tends to infinity, we obtain the conditional quantile of $Y|X$ at the population level.

3.4 Transformations and Equivariance

Often while analyzing a response variable, we wish to transform the scale to aid in interpretation and/or to obtain a better model fit. Knowledge of equivariance properties helps us to reinterpret fitted models when we transform the response variable.
After any linear transformation of the response variable, either adding a constant and/or multiplying by a constant, the conditional mean of the LRM can be exactly transformed. Formally speaking, $\forall a, c \in \mathbb{R}$,

$$\mathrm{E}[c + ay|x] = c + a(\mathrm{E}[y|x])$$

This property is termed linear equivariance because the linear transformation is the same for the dependent variable and the conditional mean.
The QRM also has this property. However, it depends on the the sign of a. If a is positive, then the conditional quantile of the QRM can be exactly transformed, as in the case of an LRM. However, if a is negative, then the order is reversed, and instead we'd be retrieving the complementary quantile. Formally speaking,

$$Q^{(p)}[c + ay|x] = \begin{cases} c + a(Q^{(p)}[y|x]) & \forall a \geq 0 \\ c + a(Q^{(1-p)}[y|x]) & \forall a \leq 0 \end{cases}$$

However, not all transformations desired are linear in nature. The most common example of such a transformation is a log transformation. Log transformations are also introduced in order to model a covariate's effect in relative terms (e.g., percentage changes). In other words, the effect of a covariate is viewed on a multiplicative scale rather than on an additive one. For example, suppose we have a model that expresses the effects of education on household income, as seen earlier, where the the units was per $. Now suppose we ask the following question: "What is the percentage change in conditional-mean income brought about by one more year of schooling?"; a log model would then help us. The coefficient for education in a log-income equation (multiplied by 100) approximates the percentage change in conditional-mean income brought about by one more year of schooling. However, under the LRM, the conditional mean of log income is not the same as the log of conditional-mean income, i.e., in general,

$$\mathrm{E}[\log(y)|x] \neq \log(\mathrm{E}[y|x])$$

In fact, for any monotonic increasing function h, the above is true. Albeit misleading, the transformation involving such functions are called monotone transformations, and if the equality had held above, then the model is said to show monotone equivariance. Log transformations were a specific case of such transformations. Generally speaking, the monotone equivariance property fails to hold for conditional means, so that LRMs do not possess monotone equivariance. Another example of a monotone transformation is a power transformation. In stark contrast, QRMs possess monotone equivariance. For any monotonous function h we have that,

$$Q^{(p)}[h(y)|x] = h(Q^{(p)}[y|x])$$

Specifically, this means that we can convert to and fro using a log transformation, while working with a QRM. Given the the log QRM model, we can reinterpret the QRM model by the following relation

$$Q^{(p)}[y|x] = \mathrm{e}^{(Q^{(p)}[\log(y)|x])}$$

s. In other words, assuming a perfect fit for the p^{th} quantile function of the form $Q^{(p)}[y|x] = \beta_0 + \beta_1 x$, we have $Q^{(p)}[\log(y)|x] = \log(\beta_0 + \beta_1 x)$, so that we can use the impact of a covariate expressed in absolute terms to describe the impact of a covariate in relative terms and vice versa.
The QRM's monotone equivariance is particularly important for research involving skewed distributions. While the original distribution is distorted by the reverse transformation of log-scale estimates if the LRM is used, the original distribution is preserved if the QRM is used. Hence, the monotone equivariance property allows researchers to achieve both goals: measuring percentage change caused by a unit change in the covariate and measuring the impact of this change on the location and shape of the raw-scale conditional distribution.

4 QR Inference

So far we saw how to estimate the coefficients of the QMR model. But no estimation is complete without a proper inference of the coefficients. We now turn to the topic of inferential statistics, specifically standard errors and confidence intervals for coefficient estimates from the QRM. As we have been doing frequently, we will first look at how inference is done in LRM, and draw an analogue to QRM.

4.1 Standard Errors and Confidence Intervals for the LRM

Suppose we have a general multiple LRM, i.e., of the form,

$$y_i = \sum_{j=i}^{k} \beta_j x_{ij} + \varepsilon_i$$

with the standard assumptions, i.e., ε_i are distributed as i.i.d. $N(0, \sigma^2)$. Here, x_{ij} is the value of the j^{th} covariate corresponding to the i^{th} response. The natural estimator of the error variance is given by,

$$\hat{\sigma}^2 = \frac{RSS}{n-k}$$

where, the RSS is Residual Sum of Squares, given by $\sum_{i=1}^{k}(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the value of the LR equation after obtaining the estimates $\hat{\beta}_i$ corresponding to each β_i. Note that since we need to estimate k coefficients before obtaining the RSS, we have lost k DOF, and hence the expression above.

Now suppose we have n such responses $y_1, y_2, ..., y_n$. Then we can consider a $n \times k$ matrix, whose entries are x_{ij} as in the LRM equation, i.e., the i^{th} row corresponds to the i^{th} response. Let β be $(\beta_1, \beta_2, ..., \beta_k)$ and $\hat{\beta}$ be the corresponding vector of estimates. Then, $\hat{\beta}$ is distributed as multivariate normal with mean as β and covariance matrix given by $\sigma^2(X^tX)^{-1}$. Alternatively, each $\hat{\beta}_j$ is distributed as $N(\beta_j, \delta_j\sigma^2)$, where δ_j is the j^{th} diagonal entry of $(X^tX)^{-1}$. So, we naturally estimate the variance of $\hat{\beta}_j$ using $\delta_j\hat{\sigma}^2$.

We can define the standard error of the LRM as the square-root of the variance estimator of β_j.Let us call this as $s_{\hat{\beta}_j}$. As a consequence of the assumptions on ε_i, the quantity $(\hat{\beta}_j - \beta_j)/s_{\hat{\beta}_j}$ is distributed as a Student's t-distribution with $n-k$ DOF. Thus, we can perform a t-test for the following hypotheses:

$$H_0 : \beta_j = 0$$
$$H_A : \beta_j \neq 0$$

We reject H_0 if $|\hat{\beta}_j / s_{\hat{\beta}_j}| \leq t_{\alpha/2}$. We can also obtain a $100(1-\alpha)\%$ CI around β_j as $\hat{\beta}_j \pm t_{\alpha/2} s_{\hat{\beta}_j}$. When n is very large, we can relax the assumptions on the the error distributions. In that case,$(\hat{\beta}_j - \beta_j)/s_{\hat{\beta}_j}$ will be approximately standard normal, and we can obtain the CI using $z_{\alpha/2}$ instead of $t_{\alpha/2}$.

Now that we have seen the process for a LRM, we can study the analogue for a QRM.

4.2 Standard Errors and Confidence Intervals for the QRM

Let us assume a univariate QRM, i.e., of the form,

$$y_i = \sum_{j=1}^{k} \beta_j^{(p)} x_{ij} + \varepsilon_i^{(p)}$$

Recall that $\varepsilon_j^{(p)}$ have the property that the p^{th} quantile of it's distribution is zero. Inference will be done as in the case of LRM, i.e., finding the standard error and then testing the significance at some level of confidence, and also reporting a CI around the coefficients. Analogous to what was discussed above, we will denote the standard error in finding $\hat{\beta}_j^{(p)}$ by $s_{\hat{\beta}_j^{(p)}}$. The standard error will have the property that asymptotically, $(\hat{\beta}_j^{(p)} - \beta_j^{(p)})/s_{\hat{\beta}_j^{(p)}}$ is distributed as standard normal.
Assume that the residuals are i.i.d. Then the covariance matrix looks like,

$$\Sigma_{\hat{\beta}^{(p)}} = \frac{p(1-p)}{n} . \frac{1}{f_{\varepsilon^{(p)}}(0)^2} (X^tX)^{-1}$$

Similar to an LRM, it is a scalar multiple of $(X^tX)^{-1}$. The scalar used in LRM was σ^2, the variance of the distribution of the ε_i. So, it would make sense that the scalar in a QRM is also the variance of the distribution of the $\varepsilon_i^{(p)}$. Recall the discussion on sample quantiles and their asymptotic behaviour. We will use the same idea here. Suppose our sample is the residuals $\varepsilon_1^{(p)}, \varepsilon_2^{(p)}, ..., \varepsilon_n^{(p)}$. Then for this sample, we know that the sample quantile $\hat{Q}^{(p)}$ is asymptotically distributed as normal with mean as the p^{th} quantile of $F_{\varepsilon^{(p)}}$, the CDF of the true distribution of $\varepsilon^{(p)}$, and variance $\frac{p(1-p)}{n} . \frac{1}{f_{\varepsilon^{(p)}}} (Q^{(p)})^2$, where $f_{\varepsilon^{(p)}} = F'_{\varepsilon^{(p)}}$. But we know that for

a QRM of the p^{th} quantile, the residuals have the property that $Q^{(p)}$ is 0 for their distribution. Hence, the variance is $\frac{p(1-p)}{n}.\frac{1}{f_{\varepsilon^{(p)}}(0)^2}$, and hence the matrix above. Once again, we have to estimate $f_{\varepsilon^{(p)}}$. This can be done as mentioned earlier; one uses the fact that $\frac{d}{dp}Q^{(p)(\varepsilon^{(p)})} = 1/f_{\varepsilon^{(p)}}$. The slope can be approximated by $\frac{1}{2h}(\hat{Q}^{(p+h)} - \hat{Q}^{(p-h)})$, for some chosen (cite) small value of h. Note that the sample quantiles are based on the residuals $\varepsilon_i^{(p)} = y_i - \sum_{j=1}^{k} \hat{\beta}_j^{(p)} x_{ij}$.

An estimated standard error for an individual coefficient estimator $\hat{\beta}_j^{(p)}$ is obtained by taking the square root of the corresponding diagonal element of the estimated covariance matrix $\Sigma_{\hat{\beta}^{(p)}}$. As in the case of LRM, we can test hypotheses and find CI around $\beta_j^{(p)}$.

Thus, we see that the i.i.d. case is easier to handle. The problem is when the residuals are not assumed to be i.i.d. The $\varepsilon_i^{(p)}$ no longer share a common distribution, but they still satisfy the property that the p^{th} quantile of their distribution is 0. In such a case, one proceeds to introduce a weighted covariance matrix, the discussion of which is beyond the scope of this article.
An important concern about the asymptotic standard error is that the i.i.d. assumption of errors is unlikely to hold. The often-observed skewness and outliers make the error distribution depart from being i.i.d. Standard large sample approximations have been found to be highly sensitive to minor deviations from the i.i.d. error assumption. Thus, asymptotic procedures based on strong parametric assumptions may be inappropriate for performing hypothesis testing and for estimating the confidence intervals. One such method will be discussed by us, the bootstrap method.

4.2.1 The Bootstrap Method

Bootstrapping is a very common Monte-Carlo simulation used to estimate population parameters from a sample. A Monte-Carlo simulation involves drawing samples of size n from a population of known parameters, and each sample is used to get an estimate of the parameters. In particular, the standard error of the estimate can be estimated using standard deviation of the sample of parameter estimates. Bootstrapping is slightly different. Given an observed data set S, we sample n of these data points *with replacement*(called as a resample). The number of resamples M, is often very large; typically 50-200 for estimating a standard deviation, and 500-2000 for estimating a CI. Although each resample will have the same number of elements as the original sample, it could include some of the original data points more than once while excluding others. Therefore, each of these resamples will randomly depart from the original sample.
Let us illustrate how this would work for a QRM. Suppose we have QRM involving k covariates, and our observed data set is $y_1, y_2, ..., y_n$. The asymptotic approach involves finding the covariance matrix $\Sigma_{\hat{\beta}^{(p)}}$, and setting the standard error involved in estimating the coefficient $\hat{\beta}_j^{(p)}$ as the square root of the j^{th} diagonal element of $\Sigma_{\hat{\beta}^{(p)}}$. This, of course, inherently involves the approximation of the distribution function using the slope of the secant passing through $(p-h, \hat{Q}^{(p-h)})$ and $(p+h, \hat{Q}^{(p+h)})$, for some appropriate choice of h.

The bootstrap method is quite direct. Before we see how to estimate coefficients, let use first see how it could be used to estimate the distributions of the quantiles of the sample. We draw a large number (M) of samples from $y_1, ...y_n$ and call them as boostrapped sample, or simply, bootstrap. The m^{th} bootstrap would be denoted by $\tilde{y}_1^{(m)}, \tilde{y}_2^{(m)}, ..., \tilde{y}_n^{(m)}$. For this bootstrap, we will compute a sample quantile $\hat{Q}_m^{(p)}$. A similar process is done for all the M resamples and we will do this for a large M, preferrably $50 \leq M \leq 200$. We have now obtained a sample $\hat{Q}_m^{(p)}; m = 1, 2, .., M$, which we can all as S_M This sample can be treated as being drawn from the distribution of $\hat{Q}^{(p)}$, and we can set the standard deviation of S_M as the estimate for the desired standard deviation. Note how the whole process can be finished without the need to estimate the distribution functions as in the case of the asymptotic method.
The bootstrap estimates can also be used to form an approximate confidence interval for the desired population quantile. We could do this in 2 possible ways:

- We could use the original estimate $\hat{Q}^{(p)}$ from the sample, and the standard deviation as s_{boot}, the standard deviation of S_M. We can use a normal approximation and set the $100(1-\alpha)\%$ CI as $\hat{Q}^{(p)} \pm z_{\alpha/2} s_{boot}$.
- We could use the empirical quantiles, i.e., values from S_M. If we can order the data in S_M as order statistics $\hat{Q}_{(1)}^{(p)}, \hat{Q}_{(2)}^{(p)}, ..., \hat{Q}_{(M)}^{(p)}$, then a $100(1-\alpha)\%$ CI would be $(\hat{Q}_{(M\alpha)}^{(p)}, \hat{Q}_{(M(1-\alpha)+1)}^{(p)})$.

Now that we have seen how we could bootstrap information about the quantiles themselves, it is not hard to understand how we could use this to estimate CI intervals around the coefficients themselves. As usual, we denote the vector of coefficients $(\beta_1^{(p)}, \beta_1^{(2)}, ..., \beta_1^{(k)})$ as $\beta^{(p)}$. We wish to estimate standard errors of $\beta^{(p)}$. If we treat the data as $(k+1)$-tuples of the form $(x_{i1}, x_{i2}, ..., x_{ik}, y_i)$, then we can perform

a bootstrap process by sampling n such tuples with replacement. Each bootstrap sample gives rise to a parameter estimate, call it $\hat{\beta}_{mj}^{(p)}$ for the m^{th} bootstrap and we estimate the standard error s_{boot} of a particular coefficient estimate $\hat{\beta}_j^{(p)}$ by taking the standard deviation of all $M\hat{\beta}_{mj}^{(p)}$. As seen earlier, we can compute a $100(1-\alpha)\%$ CI for the $\beta_j^{(p)}$ one of two ways, either by normal approximation or using empirical sample quantiles.

A situation unique to QRM is that given the data, we can have multiple choices of the quantile, i.e., there are a lot of choices of p that could be used to model the data; whereas in an LRM, there is only one such possibility, the mean. For example, suppose we have 4 covariates(including the intercept) and we want to model 9 equally spaced quantiles (0.1,0.2,...,0.9) as the best estimate of the response. We can now consider multiple QRMs. We can estimate the covariance between all possible coeffecients over all the 9 models. So, we would have $9 \times 4 = 36$ coefficients, and hence a 36×36 covariance matrix Σ. Now a natural question arises, are the coefficients the same for the same covariate but for different quantiles? Formally speaking, we would test the following hypothesis:

$$H_0 : \hat{\beta}_i^{(p)} = \hat{\beta}_i^{(q)}$$
$$H_A : \hat{\beta}_i^{(p)} \neq \hat{\beta}_i^{(q)}$$

for some $1 \leq i \leq 4$ and $0.1 \leq p \neq q \leq 0.9$. Note that these numbers are for the above example. One can of course do it for more covariates and quantiles. We can test the null using the Wald statistic which is given by

$$W_{pq} = \frac{(\hat{\beta}_i^{(p)} - \hat{\beta}_i^{(q)})^2}{\hat{\sigma}^2_{\hat{\beta}_i^{(p)} - \hat{\beta}_i^{(q)}}}$$

The denominator is the variance of the difference between the estimators. We can obtain this value using the following relation

$$\hat{\sigma}^2_{\hat{\beta}_i^{(p)} - \hat{\beta}_i^{(q)}} = \hat{\sigma}^2_{\hat{\beta}_i^{(p)}} + \hat{\sigma}^2_{\hat{\beta}_i^{(q)}} - 2Cov(\hat{\beta}_i^{(p)}, \hat{\beta}_i^{(q)})$$

We can obtain the values on the RHS from Σ. Under the null hypothesis, $W_{pq} \sim \chi^2$ with 1 DOF, i.e., χ^2_1. We reject the null at $100(1-\alpha)\%$ significance level if $|W_{pq}| > \chi^2_{\alpha/2}$.

We can do this for multiple covariates at the same time. The Wald statistic would be described using a covariance matrix rather than variance of the difference between the coefficients. Suppose the hypotheses are

$$H_0 : \hat{\beta}_1^{(p)} = \hat{\beta}_1^{(q)} \text{ and } \hat{\beta}_2^{(p)} = \hat{\beta}_2^{(q)}$$
$$H_A : \hat{\beta}_1^{(p)} \neq \hat{\beta}_1^{(q)} \text{ or } \hat{\beta}_2^{(p)} \neq \hat{\beta}_2^{(q)}$$

we first define the matrix $\Sigma_{\hat{\beta}_i^{(p)} - \hat{\beta}_i^{(q)}}$. Σ_{ij} is calculated as $Cov(\hat{\beta}_i^{(p)}, \hat{\beta}_j^{(p)}) + Cov(\hat{\beta}_i^{(q)}, \hat{\beta}_j^{(q)}) - 2Cov(\hat{\beta}_i^{(q)}, \hat{\beta}_j^{(p)})$. Let A be the matrix $\begin{bmatrix} \hat{\beta}_1^{(p)} - \hat{\beta}_1^{(q)} \\ \hat{\beta}_2^{(p)} - \hat{\beta}_2^{(q)} \end{bmatrix}$ Then we define the Wald statistic as $A^t \Sigma^{-1} A$. Under the null, it is distributed as χ^2_2. As done earlier we can construct confidence intervals and test the null.

Thus we have seen how pivotal bootstrapping is to estimate sample parameters. The only con of bootstrapping is the time. If you'd use a 64-bit processor, bootstrapping 2000 times takes about a minute, which is much slower than most computations. However, it is still widely used because it doesn't require us to estimate any distributions before hand. It is the perfect way to obtain CI and standard errors of the coefficients of a QRM.

4.3 Goodness of Fit

Finding the confidence intervals and standard errors tell us about the coefficients we have obtained. But sometimes we desire the answer to the simple question: "Is the obtained model a good fit for the data?". Let us study this.
In LRMs, the goodness of fit for the model is given the R^2-coefficient, also known as coefficient of determination. It is defined as:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2}$$

where, $\bar{y}$ is the sample mean of the response variables and $\hat{y}$ is the predicted value from the LRM. Recall that the numerator is called as the RSS, which can be interpreted as the sum of squared distances between observed

data and data predicted using the LRM. The denominator can be interpreted as the sum of squared distances between observed data and data predicted by using just the intercept term. One can now see why R^2 explains the variability in the response variable expalined by the covariates. $R^2 \in [0, 1]$, and a larger value indicates a better fit.

- $R^2 \to 0$ implies that the fraction $\to 1$. Hence, the best fit for the model is the sample mean itself, implying that there is so significant role played by the covariates in explaining the response.
- $R^2 \to 1$ implies that the numerator $\to 0$. This means that the model fit by us can explain the data with a very high accuracy, and hence there is a signifcant role played by the covariates in explaining the response.

It is not hard to define an analog to the above for a QRM. The quadratic terms in R^2 come about from the fact that LRMs are based on the OLS criterion. Since QRMs are based on minimizing the sum of weighted distances, it would be natural to expect the analog to R^2 to have the same.

The weight we assign would be either p or $1 - p$, depending upon whether $y_i \geq \hat{y}_i$ or $y_i < \hat{y}_i$ respectively. Koenker and Machado (1999) suggest measuring goodness of fit by comparing the sum of weighted distances for the model of interest with the sum in which only the intercept parameter appears. Suppose this weighted distance is denoted by $d_p(y_i, \hat{y}_i)$. Define the restricted weighted sums $V^k(p)$ as,

$$\begin{aligned} V^k(p) &= \sum_i^n d_p(y_i, \hat{y}_i) \\ &= \sum_{y_i \geq \hat{y}_i} p|y_i - \hat{y}_i| + \sum_{y_i < \hat{y}_i} (1-p)|y_i - \hat{y}_i| \end{aligned}$$

and $V^0(p)$ as

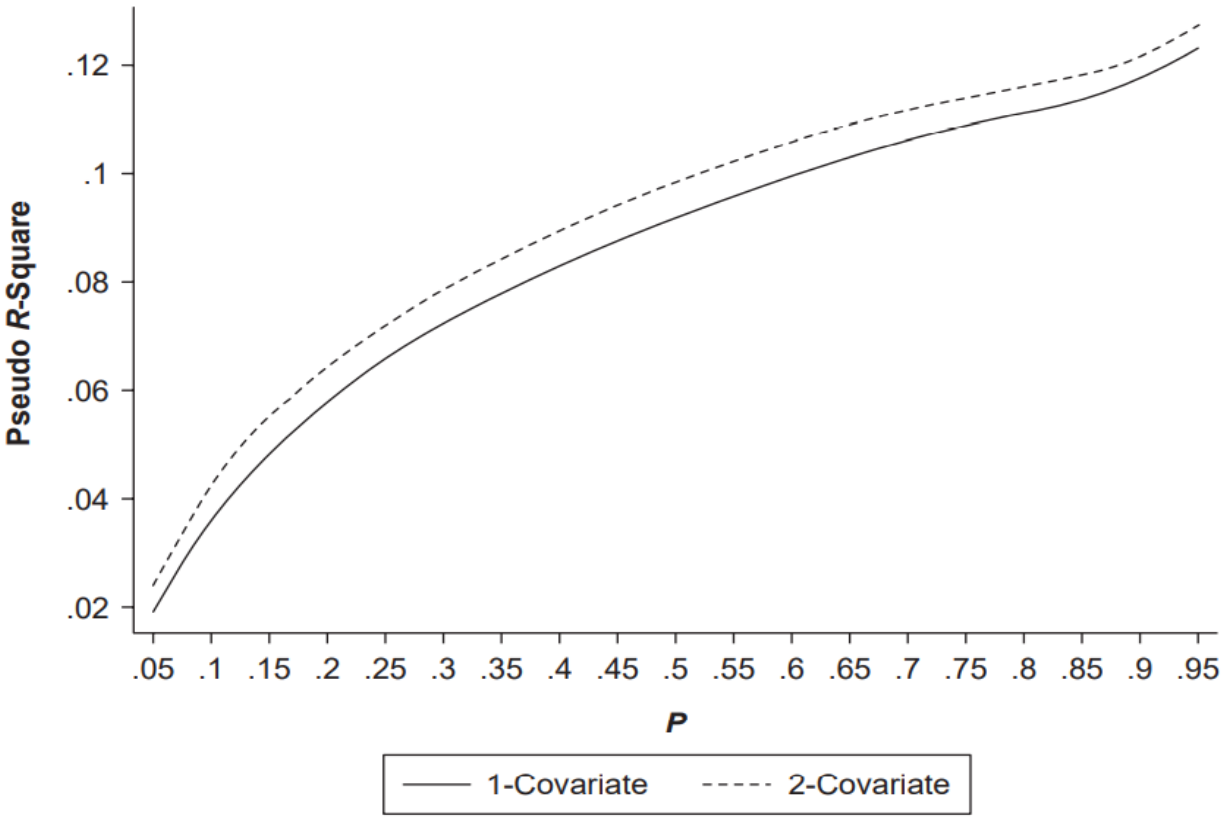
$$\begin{aligned} V^k(p) &= \sum_i^n d_p(y_i, \hat{Q}^{(p)}) \\ &= \sum_{y_i \geq \hat{Q}^{(p)}} p|y_i - \hat{Q}^{(p)}| + \sum_{y_i < \hat{Q}^{(p)}} (1-p)|y_i - \hat{Q}^{(p)}| \end{aligned}$$

$V^k(p)$ is the weighted sum when we consider the entire QRM model. $V^0(p)$ is the weighted sum when we consider only the intercept term in the model. Analogous to LRM, the fitted constant is the sample p^{th} quantile $Q^{(p)}$. Now we can define the goodness of fit $R(p)$ as,

$$R(p) = 1 - \frac{V^k(p)}{V^0(p)}$$

Since the $V^0(p)$ and $V^k(p)$ are nonnegative, $R(p)$ is at most 1. Also, because the sum of weighted distances is minimized for the full-fitted model, $V^k(p)$ is never greater than $V^0(p)$, so $R(p)$ is greater than or equal to zero. So, like $R^2, R(p) \in [0, 1]$. $R(p)$ is often called as "pseudo-Rsquared" to avoid confusion with the actual R^2.

Recall the income example we had explored in detail. Now suppose there is a dispute as to whether both education and race significantly affect the household income or just education affects the income. We can answer this question by using two models, one bivariate and the other univariate. Now if this were to be compared using a LRM, then it could be answered straight away by computing the R^2 of both the models. If we are to build QRMs, then we not only have to consider the psuedo-R^2, but also compare the value of the psuedo-R^2 across different quantiles. So, we will then plot $R(p)$ for these 2 models for different values of p and see how the fits behave. Given below is the graph of pseudo-R^2 values for the dataset that we used before.



The goodness of fit for income is poorer at the lower tail than the upper tail. The mean $R(p)$ over the 19 quantiles for income is 0.0913. The one-covariate model is nested in the two-covariate model, with mean $R(p)$ over the 19 quantiles for income being .0857. These models' $R(p)$s indicate that using race as an explanatory variable improves the model fit; however, there is only a slight increment in the value of $R(p)$ from the one-covariate model to the two-covariate model, so we can conclude that the major explanatory power lies in education.

5 Quantile Regression in R

Now we will see how to build QRMs in R, and proceed to apply and analyze different datasets.

5.1 Synatx and Plots

Quantile regression can be performed using the *rq* function in the *quantreg* package in R. The functions and plots will be explained using an inbuilt dataset *mtcars* present in R itself.

```
library(quantreg)
data("mtcars")
```

The parameters of *rq* are the data. dataset and τ. The first two are understood namesake, and $\tau \in [0, 1]$ refers to the quantile we want to model. By default, $\tau = 0.5$, i.e., median regression is performed.

```
rqfit = rq(mpg ~ wt, data = mtcars)
rqfit

## Call:
## rq(formula = mpg ~ wt, data = mtcars)
##
## Coefficients:
## (Intercept)          wt
##   34.232237   -4.539474
##
## Degrees of freedom: 32 total; 30 residual
```

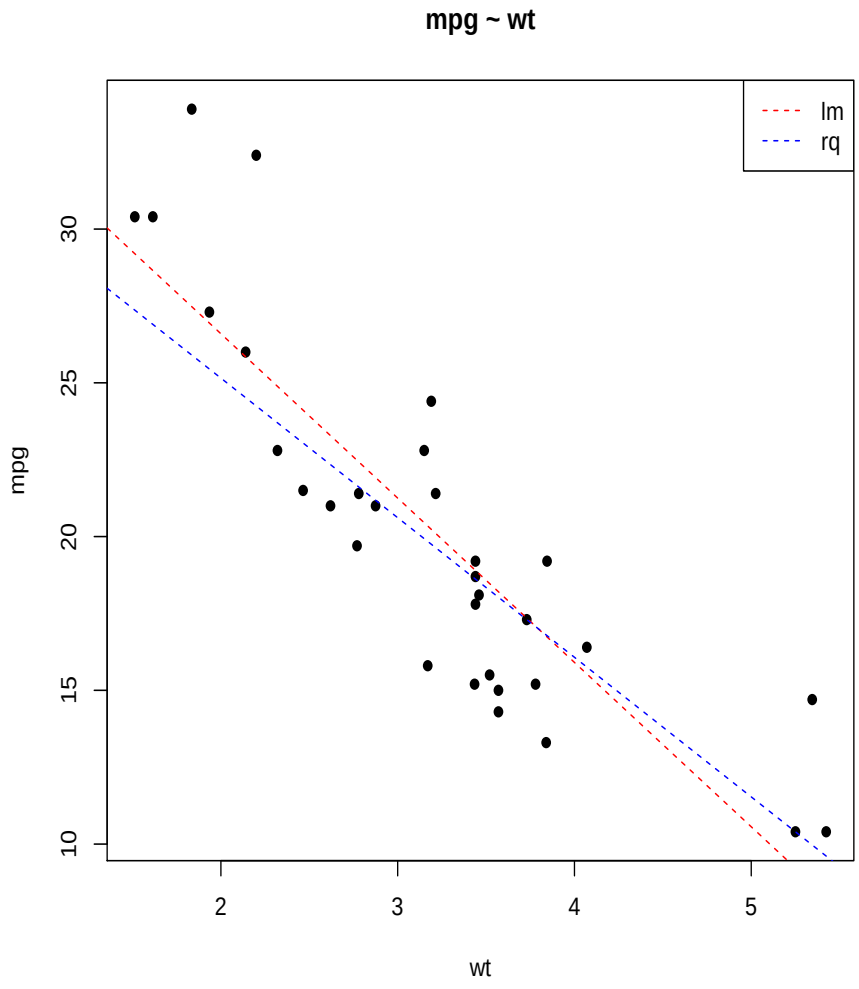
- Coefficients are the estimates $\hat{\beta}_0^{(0.5)}, \hat{\beta}_1^{(0.5)}$ for the model
- If n is the sample size, k is the number of predictors (excluding the intercept), then
  - DF(Regression) $= k + 1$
  - DF(Residuals) $= n - k - 1$
  - DF(Total) $= n$

We can obtain more information if we use call *summary* on the model.

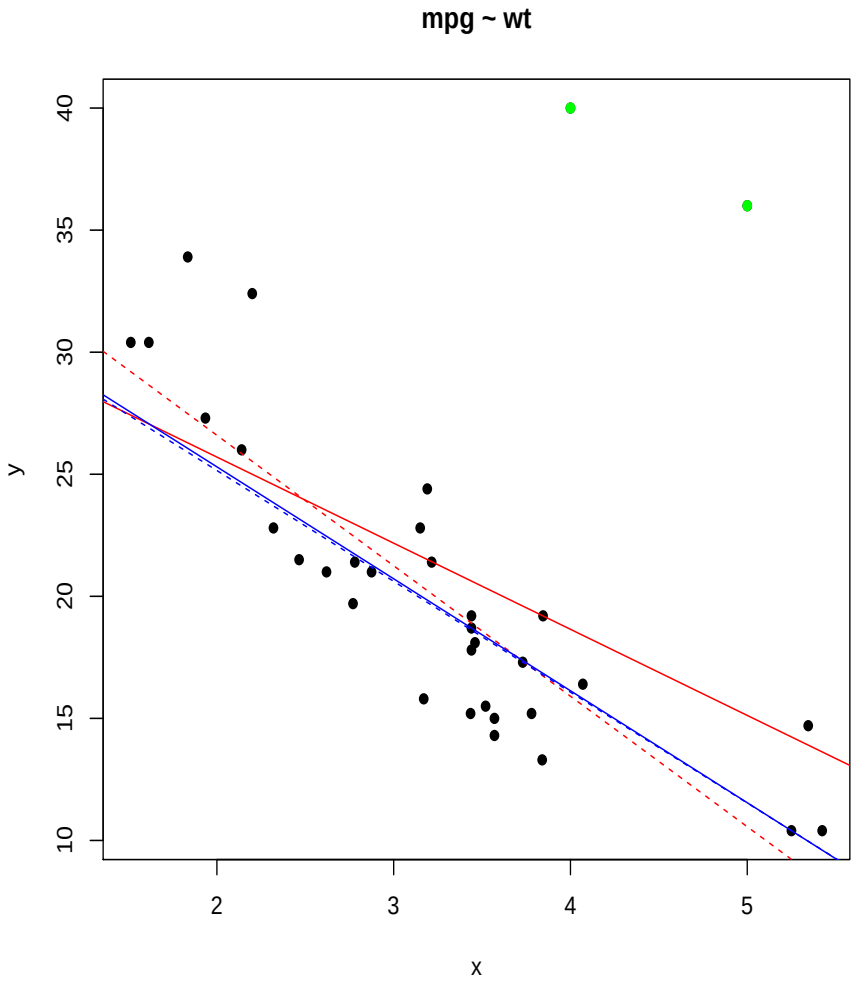
```
summary(rqfit)

##
## Call: rq(formula = mpg ~ wt, data = mtcars)
##
## tau: [1] 0.5
##
## Coefficients:
##             coefficients lower bd upper bd
## (Intercept) 34.23224     32.25029 39.74085
## wt          -4.53947     -6.47553 -4.16390
```

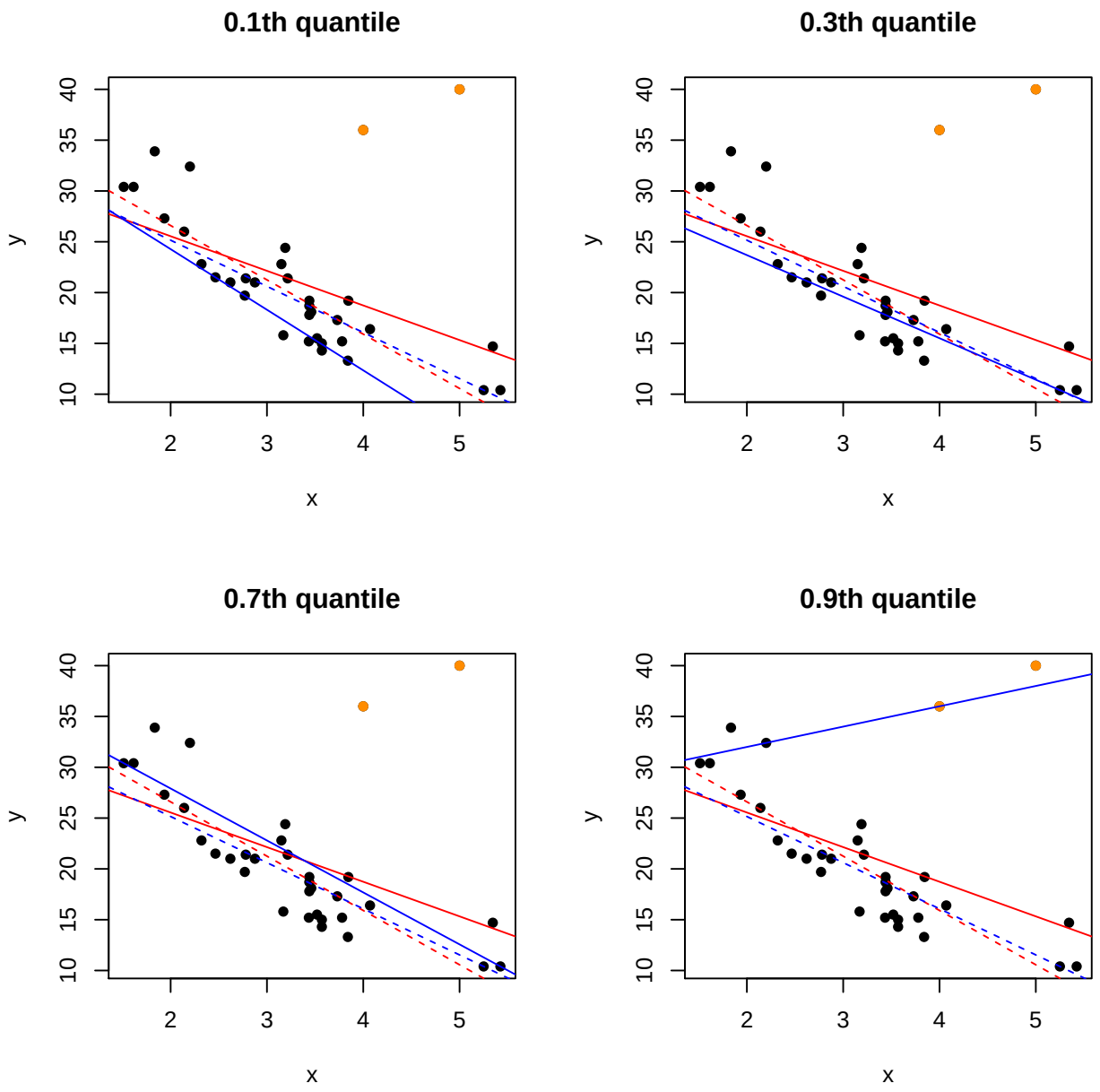
The value of τ and coefficients are explained as above. An additional information provided is the CI around the coefficients, in the form of (lower bd, upper bd). There are a number of ways for these confidence intervals to be computed; this can be specified using the *se* option when invoking the *summary* function. The default value is *se*="rank", with the other options being "iid","nid","ker","boot" and "BLB".
Now let's plot the data as a scatterplot, and fit the regression lines over it to see how appropriate the model is.



Both seem like decent fits to the given data. Recall that we had discussed how a conditional-median model is more robust to outliers than a LRM. Let's see why this is true, visually.



The outliers are marked in green. The dotted curves were the same curves as in the previous plot, i.e., before the outliers were added. Observe the deviation of the curves. The red curve (LRM) deviates quite significantly, while the blue curve (MRM) hardly deviates at all. So, median regression is indeed more robust to outliers than a LRM.
One must not confuse the robustness of the median model and an arbitrary QRM. Let us see this for 4 quantiles: 0.1,0.3,0.7,0 and 0.9.



The line significantly deviates for extreme quantiles, i.e., for 0.1 and 0.9, and much lesser for 0.3 and 0.7. So, we can say that, as we move away from the median, this robustness decreases.

```
rqfit = rq(mpg ~ wt, data = mtcars)
rqfit

## Call:
## rq(formula = mpg ~ wt, data = mtcars)
##
## Coefficients:
## (Intercept)          wt
##   34.232237   -4.539474
##
## Degrees of freedom: 32 total; 30 residual
```