Quantile Regression

18 - WISE - Mengyuan Zhang

2021/4/15

1. Introduction to quantile regression

In previous study, we use linear regression to do estimation. The traditional linear regression model describes the process in which the conditional mean distribution of the dependent variable Y is affected by the independent variable X. The least square method is the most basic method to estimate the regression coefficient. Moreover, if the random error term of the model comes from a distribution with zero mean and the same variance (homoscedasticity), then the least square estimation of the regression coefficient is the best linear unbiased estimation (BLUE). If the random error term is normally distributed, then the least squares estimator of the regression coefficient is consistent with the maximum likelihood estimator, both being the least variance unbiased estimator (MVUL). Under these assumptions and conditions, it has the fine properties of unbiasedness, efficientcy and so on.

But in real life, these assumptions are usually not satisfied. For example, when there are serious heteroscedasticity and skewed conditions in the data, the estimation of least square method will not have the above good properties. And the behavior of the conditional mean fails to fully capture the patterns in the data. In order to compensate for the shortcomings of ordinary least square method (OLS) in regression analysis, Laplace proposed median regression (minimum absolute deviation estimation) in 1818, whcih is more robust to outliers and extreme values in the data set. On this basis, Koenker and Bassett extended median Regression to the general Quantile Regression in 1978. In the following sections, the function and empirical application of quantile regression will be introduced respectively.

1.1 Recap: conditional mean and conditional median

To understand the intuition of quantile regression, let's start with the intuition of L_1 and L_2 loss. For a classical linear regression model given below:

$$y_i = \boldsymbol{\beta} \boldsymbol{X_i} + e_i$$

where $\boldsymbol{\beta}$ is the vector form of a set of β_1, β_2,...,β_n. Similarly, $\boldsymbol{X_i}$ is the vector form of a set of x_{1i}, x_{2i},...,x_{ni}.

For the linear regression, we usually choose to use ordinary least squares (OLS) to estimate $\boldsymbol{\beta}$, which is:

$$\min_{\hat{\boldsymbol{\beta}}} \quad \sum_{i=1}^{n} (y_i - \hat{\boldsymbol{\beta}} \boldsymbol{X_i})^2$$

$$\Longleftrightarrow \quad min \quad \sum_{i=1}^{n} \hat{e_i}^2$$

where $\hat{e_i}$ is the residual.

By using OLS, we are actually minimizing the $\boldsymbol{L_2}$ **loss**, and we get the **conditional mean** in the end:

$$E(y|x) = \boldsymbol{\beta} \boldsymbol{X_i}$$

But we don't have to always estimate the conditional mean. We could also estimate the conditional median, which is in the absolute form:

$$\begin{aligned} \underset{\hat{\boldsymbol{\beta}}}{min} \quad & \sum_{i=1}^{n} |y_i - \hat{\boldsymbol{\beta}} \boldsymbol{X_i}| \\ \iff \quad & min \quad \sum_{i=1}^{n} |\hat{e_i}| \end{aligned}$$

This is also called the $\boldsymbol{L_1}$ **loss**, in which we can get the **conditional median** in the end:

$$Median(y|x) = \boldsymbol{\beta X_i}$$

Comparatively, quantile regression is a more general form of the conditional median mentioned above. Conditional median is just a special case of quantile regression when τ=0.5, namely, the 0.5 quantile. Next, let's take a close look at the concept of quantile regression.

1.2 Objective function of quantile regression

Quantile regression is an extension of standard linear regression, which estimates the conditional median of the outcome variable and can be used when assumptions of linear regression do not meet.

First of all, it should be noted that for quantile regression, dependent variable y need to be continuous with no zeroes or too many repeated values. For the τth quantile regression, the objective function we are minimizing shows in equation (1):

$$\underset{\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}}{min} \quad \tau \sum_{i:y_i \geq \hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} \boldsymbol{X_i}}^{n} |y_i - \hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} \boldsymbol{X_i}| + (1-\tau) \sum_{i:y_i < \hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} \boldsymbol{X_i}}^{n} |y_i - \hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} \boldsymbol{X_i}| \tag{1}$$

where τ represents τth quantile, and $\tau \in (0, 1)$.

Equivalently, the quantile regression is always expressed in a way of minimizing a weighted sum of the positive and negative error terms. We define a function $\rho_\tau(\hat{e})$:

$$\rho_\tau(\hat{e}) = \hat{e}(\tau - I(\hat{e} < 0)) \tag{2}$$

where $\hat{e}$ is the residual $(y - \hat{y})$.

$I(\cdot)$ in the function above represents the indicator function:

$$I(\hat{e} < 0) = \begin{cases} 0 & if \quad \hat{e} > 0 \\ 1 & if \quad \hat{e} < 0 \end{cases}$$

To visualize the equation (2), it can be illustrated graphically in Figure 1.

Furthermore, we can show that the equation (2) is equivalent to:

$$\rho_\tau(\hat{e}) = \begin{cases} \tau\hat{e} = \tau|\hat{e}| & if \quad \hat{e} > 0 \\ (\tau - 1)\hat{e} = (1-\tau)|\hat{e}| & if \quad \hat{e} < 0 \end{cases}$$

So the minimization in equation (1) is equivalent to the minimization in equation (3):

$$\iff \quad min \quad \sum_{i=1}^{n} \rho_\tau(\hat{e}) \tag{3}$$

We can see from the above equation (2) that the quantile regression actually minimizes a sum that gives asymmetric penalties $\tau|\hat{e_i}|$ for underprediction and $(1-\tau)|\hat{e_i}|$ for overprediction.

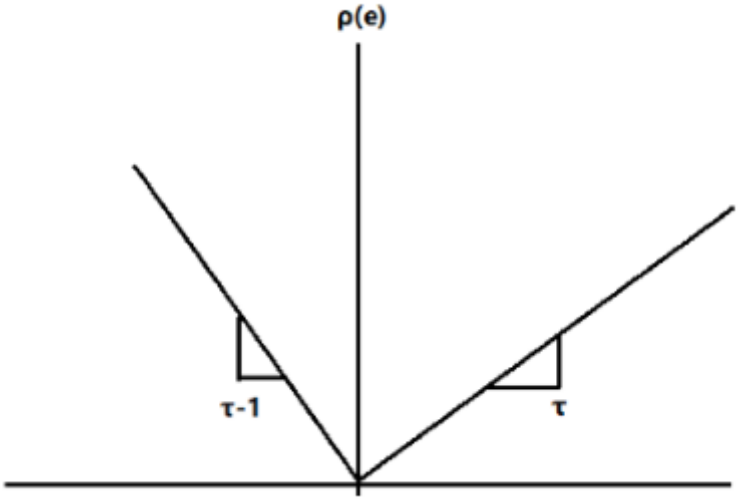


Figure 1: Quantile regression function

To better understand the calculation of the quantile function, let's see an example with the already fitted quantile line and three observational data points. In Figure 2, each orange circle represents an observation while the blue line represents the fitted quantile regression line. The black lines illustrate the distance between the regression line and each observation, which are labelled d_1, d_2 and d_3.

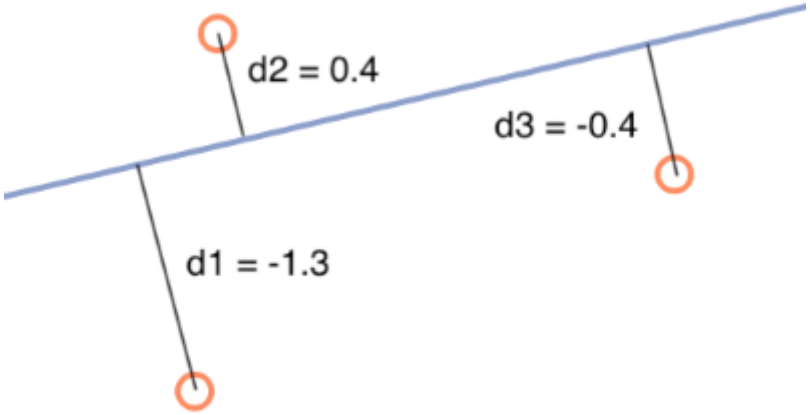


Figure 2: Distances

If we assume τ is equal to 0.9, and the fitted quantile regression line and corresponding best $\hat{\beta}$ have already been generated and given. we can compute the overall regression loss for the data in Figure 2:

$$\begin{aligned} loss &= \tau(d_2) + (1-\tau)(|d_1 + d_3|) \\ &= 0.9 * 0.4 + 0.1 * (|-1.3 + (-0.4)|) \\ &= 0.53 \end{aligned}$$

Optimization gives us an estimated linear relationship between y_i and x_i, where τ portion of the data

lies below the line and the remaining $1-\tau$ portion lies above the line. The situation is shown in Figure 3:

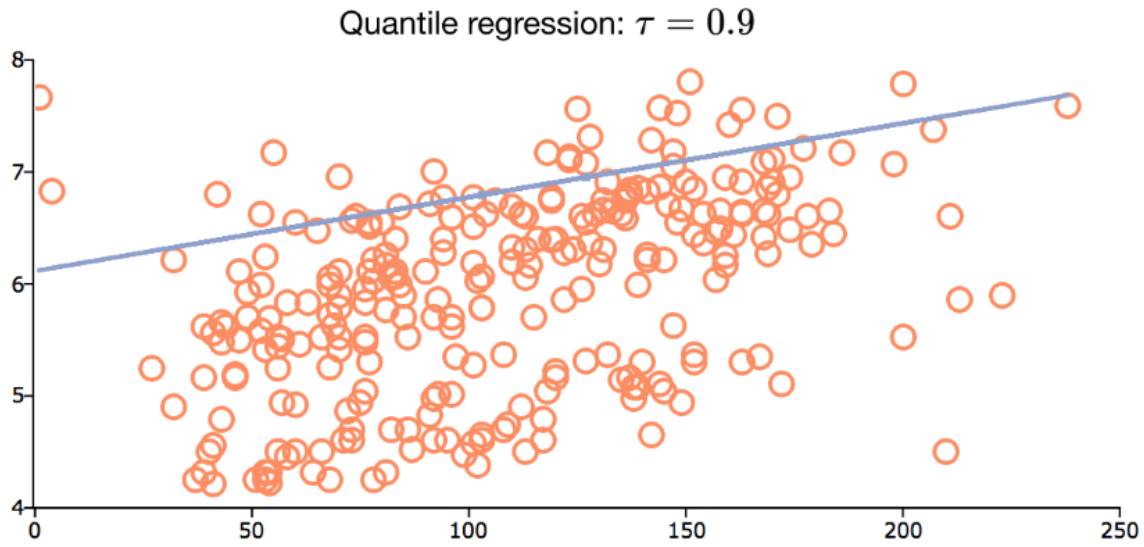


Figure 3: Quantile regression

1.3 Advantages of quantile regression

Apart from the definition and function of quantile regression, there are some advantages when using quantile regression:

(1) Quantile regression is more robust to non-normal errors and outliers than least squares regression, and is semiparametric as it avoids assumptions about the parametric distribution of the error process;

(2) Quantile regression can be used to study the distributional relationships of variables, and thus provides a complete picture of the relationship between y and x. It provides a richer characterization of the data, allowing us to consider the impact of a covariate on the entire distribution of y, not merely its conditional mean;

(3) Furthermore, quantile regression is invariant to monotonic transformations, such as $\log(\cdot)$.

1.4 Application by using empirical data

Let's use the empirical income and expenditure data in the quantreg package in R as an example. Here we use the data set engel, which contains data on the household income (y) and food expenditure (x). We can get the summary table of the data set and the scatter plot.

1.4.1 Scatter plot of data set

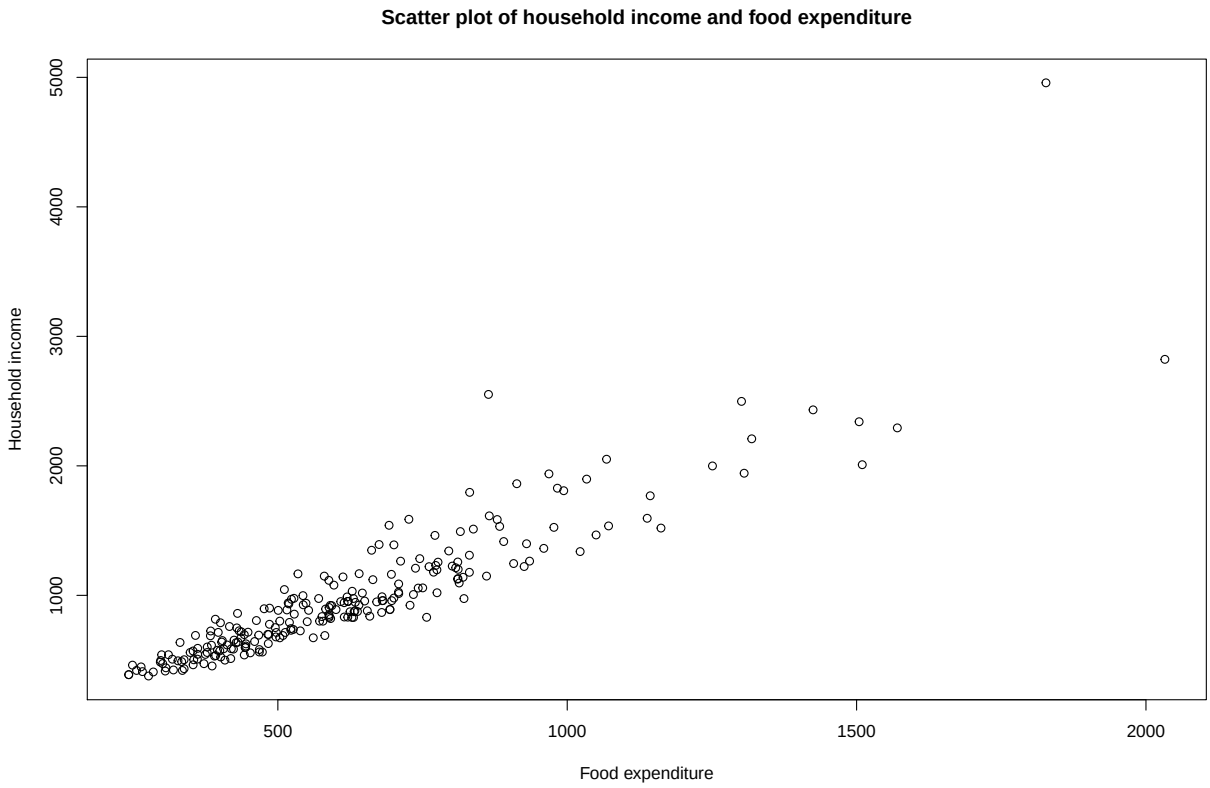
```
library(quantreg)  # Load the quantreg package
library(ggplot2)   # Load the ggplot2 package
data(engel)        # Load the data set that comes with the quantreg package
```

```
summary(engel)  #A summary table of the data set
```

```
##      income          foodexp
```

```
##  Min.   : 377.1   Min.   : 242.3
##  1st Qu.: 638.9   1st Qu.: 429.7
##  Median : 884.0   Median : 582.5
##  Mean   : 982.5   Mean   : 624.2
##  3rd Qu.:1164.0   3rd Qu.: 743.9
##  Max.   :4957.8   Max.   :2032.7
```

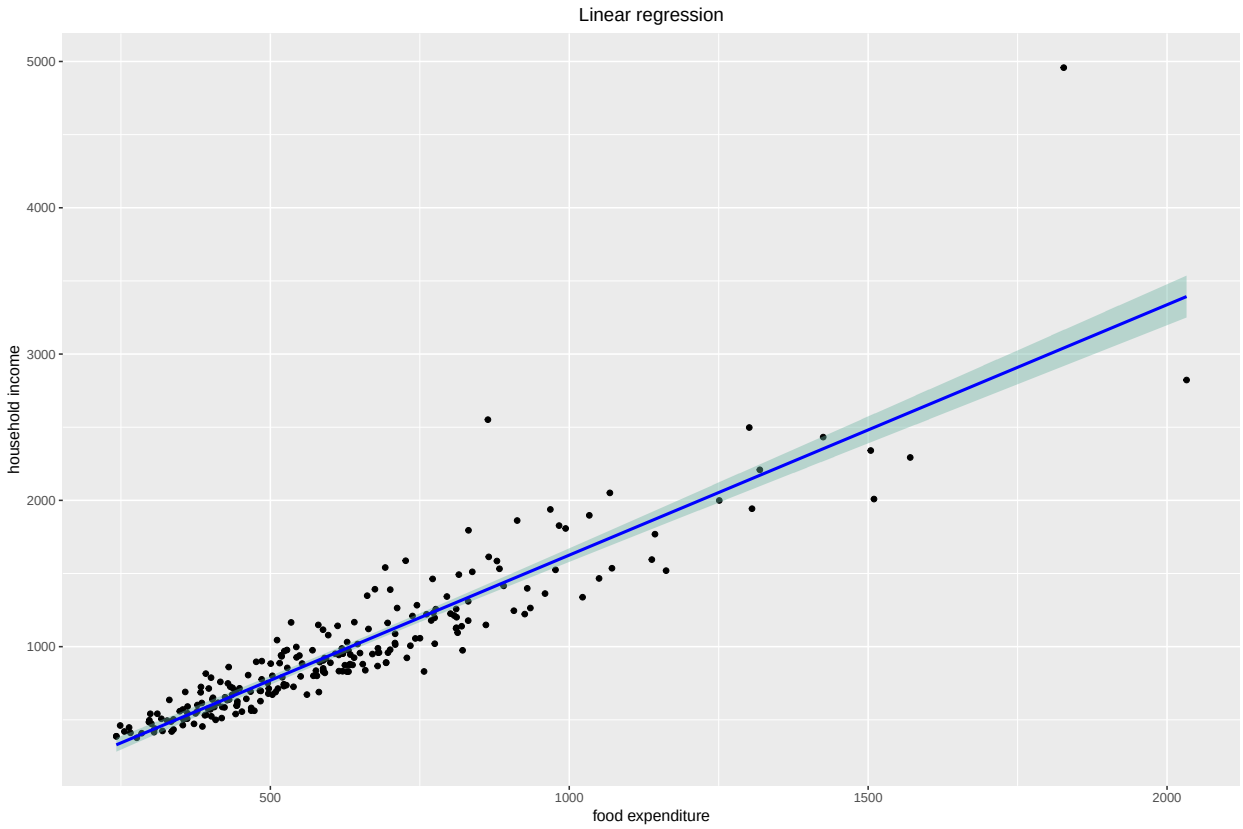
```
plot(engel$foodexp , engel$income,main = "Scatter plot of household income and food expenditure",
     xlab="Food expenditure",ylab="Household income")
```



From the scatter plot, we can see that, with the increase of x, the variance of y is gradually increasing, which indicates that our data is heteroscedasticity to some extent. Therefore, it is not complete to obtain the conditional mean only by using the OLS. To serve a a comparison, we first run the linear regression in R by fitting the linear line on the data set.

1.4.2 Linear regression

```
ggplot(engel, aes(engel$foodexp,engel$income)) + geom_point() +
  geom_smooth(method="lm",color="blue", fill="#69b3a2", se=TRUE)+
  labs(x="food expenditure",y="household income",title = "Linear regression")+
  theme(plot.title = element_text(hjust = 0.5))
```



Looking again at the linear fit, we see that linear regression provides a good estimate of y when x is in the range of 0-500. But as x increases, the mean of y given x becomes less meaningful. Now we can do quantile regression. Let's try the quantile regression with tau value 0.5 based on the rq() function in quantreg package.

1.4.3 Quantile regression with tau = 0.5

```
#quantile regression (tau = 0.5)
fit1 = rq(foodexp ~ income, tau = 0.5, data = engel)   #Rregress food expenditure on income
r1 = resid(fit1)    # The residual sequence is obtained and assigned to the variable r1
c1 = coef(fit1)     # The coefficients are obtained and assigned to the variable c1

summary(fit1)       # Show the model and coefficients of quantile regression
```

```
##
## Call: rq(formula = foodexp ~ income, tau = 0.5, data = engel)
##
## tau: [1] 0.5
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept)  81.48225     53.25915 114.01156
## income        0.56018      0.48702   0.60199
```

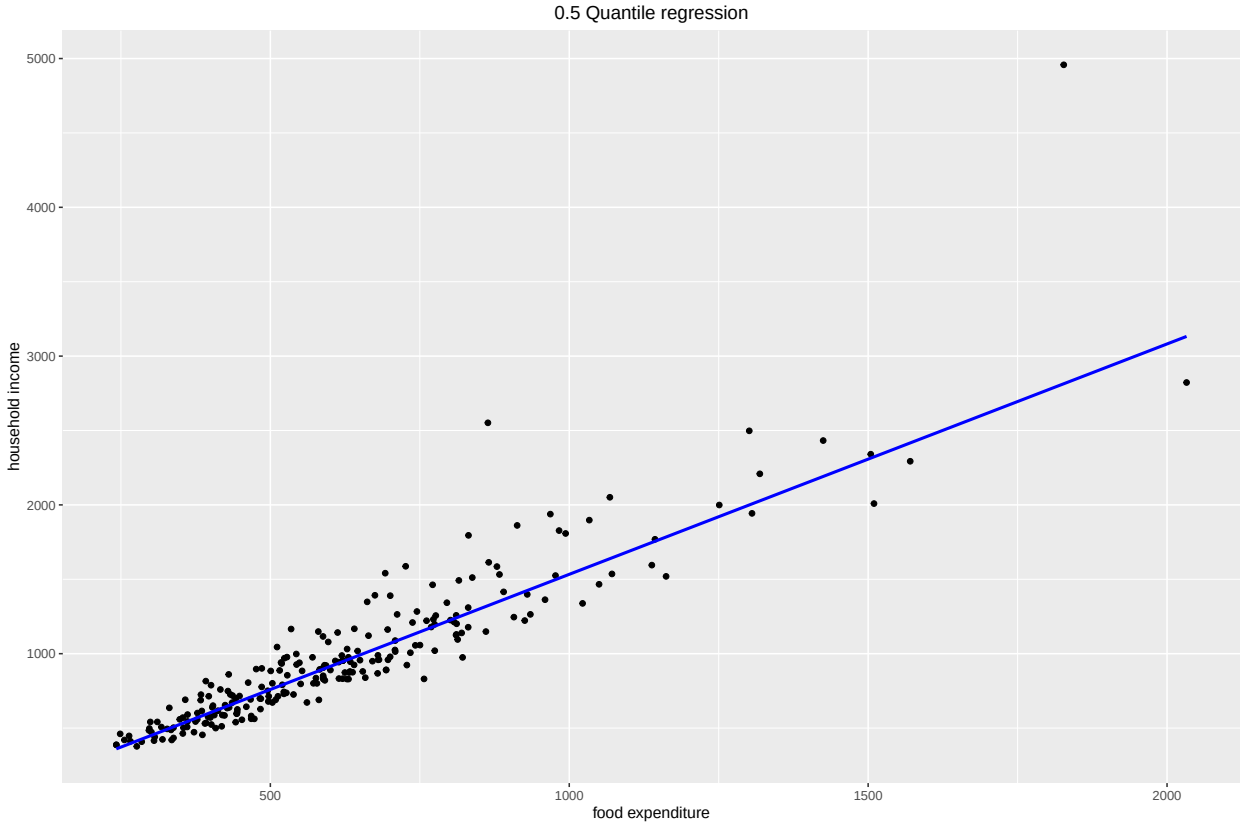
```
summary(fit1, se = "boot")   #The hypothesis test of the coefficients
```

```
##
## Call: rq(formula = foodexp ~ income, tau = 0.5, data = engel)
```

```
##
## tau: [1] 0.5
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept) 81.48225 26.70078    3.05168  0.00254
## income       0.56018  0.03473   16.12728  0.00000
```

We get a summary table of the coefficients and condifence interval for the 0.5 quantile regression. And the slope coefficient is significant at 1% level. To visualize, we plot the regression line:

```
ggplot(engel, aes(foodexp,income)) + geom_point() +
   geom_quantile(quantiles = 0.5,col="blue", size=1) +
labs(x="food expenditure",y="household income",title = "0.5 Quantile regression")+
  theme(plot.title = element_text(hjust = 0.5))
```



More broadly, we can now plot the quantile regressions with tau being 0.1, 0.2,...,0.9. We can get the summary table as well.

1.4.4 Quantile regression with different tau values

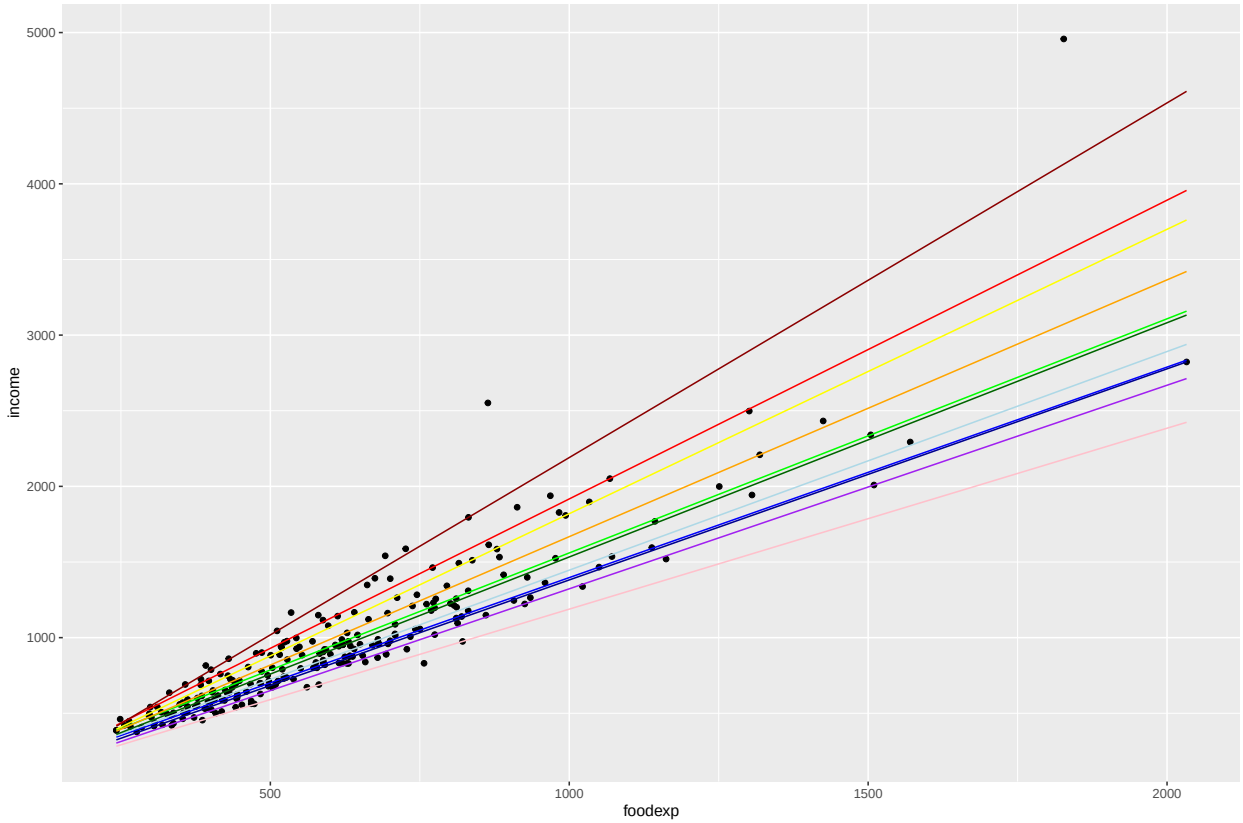
```
qs <- 1:9/10   #Perform quantile regression for the 0.10, 0.20, ...,0.90 quantiles
qr2 <- rq(engel$foodexp ~ engel$income, data=engel, tau = qs)
coef(qr2)
```

```
##                 tau= 0.1    tau= 0.2 tau= 0.3    tau= 0.4   tau= 0.5   tau= 0.6
## (Intercept)  110.1415742 102.3138823 99.11058 101.9598824 81.4822474 79.7022726
## engel$income   0.4017658   0.4468995  0.48124   0.5098965  0.5601806  0.5858492
##               tau= 0.7   tau= 0.8   tau= 0.9
```

```
## (Intercept)  79.283617 58.0066635 67.3508721
## engel$income  0.608851  0.6595106  0.6862995
```

The table shows that the slope of each quantile from 0.1 to 0.9 steadily increases. We can visualize the fitted lines by using ggplot2 package:

```
ggplot(engel, aes(x=foodexp,y=income)) + geom_point() +
  geom_quantile(quantiles = 0.01,col="pink", size=0.5) +
  geom_quantile(quantiles = 0.1,col="purple", size=0.5) +
  geom_quantile(quantiles = 0.2,col="dark blue", size=0.5) +
  geom_quantile(quantiles = 0.3,col="blue", size=0.5) +
  geom_quantile(quantiles = 0.4,col="light blue", size=0.5) +
  geom_quantile(quantiles = 0.5,col="dark green", size=0.5) +
  geom_quantile(quantiles = 0.6,col="green", size=0.5) +
  geom_quantile(quantiles = 0.7,col="orange", size=0.5) +
  geom_quantile(quantiles = 0.8,col="yellow", size=0.5) +
  geom_quantile(quantiles = 0.9,col="red", size=0.5) +
  geom_quantile(quantiles = 0.95,col="dark red", size=0.5)
```




```
  labs(x="food expenditure",y="household income",title = "Quantile regressions")+
  theme(plot.title = element_text(hjust = 0.5))
```

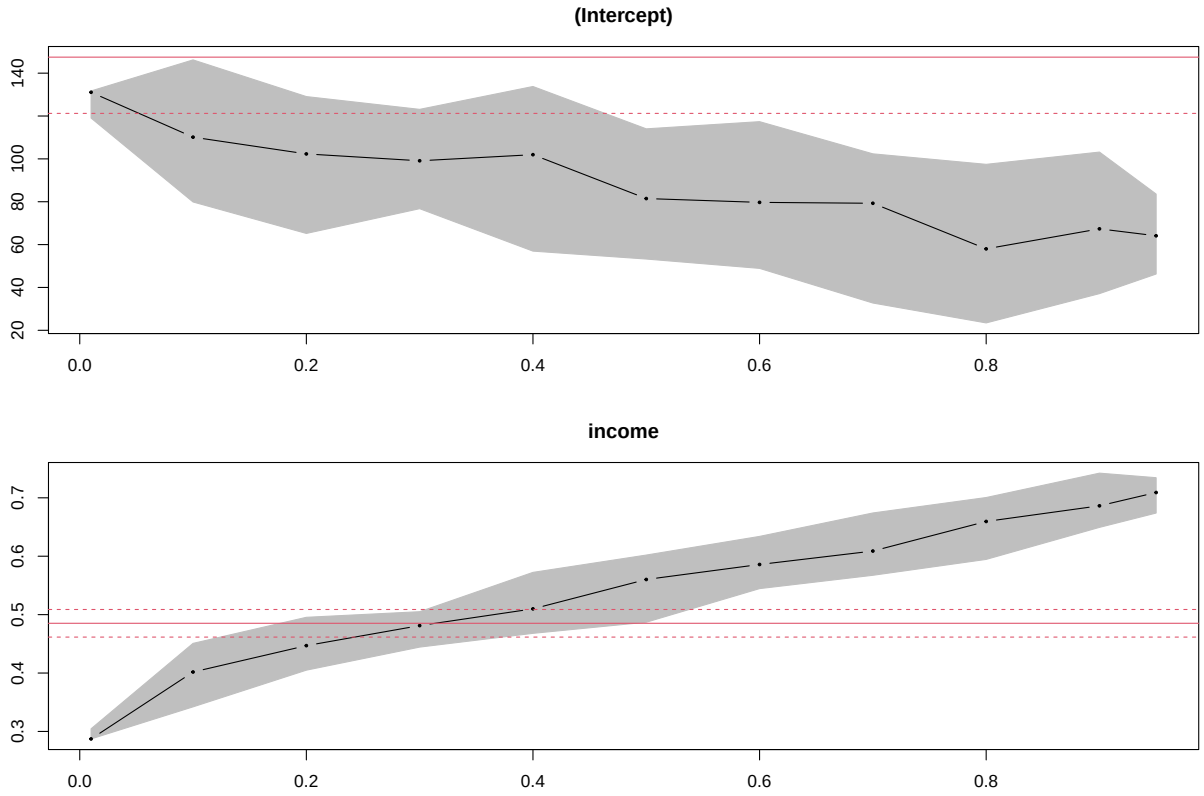
```
## NULL
```

Apart from the tau values from 0.1, 0.2,...,0.9, we add another two tau values which are 0.01 and 0.95. Different color in the chart shows different fitted line of quantile regression. The pink line at the bottom indicates the fitted line with tau = 0.01; the dark red line at the top indicates the one with tau = 0.95.

1.4.5 Quantile coefficients and confidence intervals

Finally, we can also visualize the changes in quantile coefficients at different tau values along with confidence intervals.

```
#  Regression comparison of different tau values
fit = rq(foodexp ~ income,tau = c(0.01,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,0.95), data = engel)
plot( summary(fit))
```



Each black dot is the coefficient for the quantile indicated on the x axis. The plot on the top shows the coeffinicents for intercept at different tau values, while the plot on the bottom shows the slope coefficients. The red lines are the least squares estimate and its confidence interval. You can see how the lower and upper quartiles are well beyond the least squares estimate. It appears that the quantile slope estimates are statistically different from the least squares estimate. And the linear regression's slope is not sufficient to describe the relationship between x and y.

2. Summary

In a nutshell, first, We use L_1 loss and L_2 loss as a introduction to systematically introduce quantile regression, as well as the differences and connections among these three functions. Next, after introducing the objective function of quantile regression, we introduce its empirical application. We make use of the income and expenditure data from the quantreg package of R. Moreover, We plot the linear regression fitted line and quantile regression fitted lines respectively to make a comparison. Finally, we plot the coefficients and confidence intervals of linear regression and quantile regression. Quantile regression is a better tool to deal with outlier and heteroscedasticity problems than OLS. Feel free to make good use of it!

3. References

(1) The Basics of Quantile Regression (ERICA)

https://www.aptech.com/blog/the-basics-of-quantile-regression/

(2) Getting Started with Quantile Regression (University of Virginia Library)

https://data.library.virginia.edu/getting-started-with-quantile-regression/

(3) QUANTILE REGRESSION IN R: A VIGNETTE (ROGER KOENKER)

https://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf