Quantile Regression

Koyel Pramanick, Vijay Kumar, Harshit Garg

MTH516A: Non-Parametric Inference Course project
Supervised by Dr. Dootika Vats

16th April 2022

Today's Topic

Introduction

Objective of Our Project

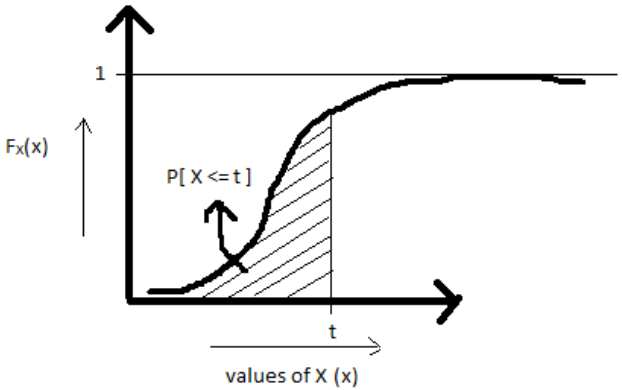
To discuss:

- What is Quantile Regression?
- Where to use?
- Application

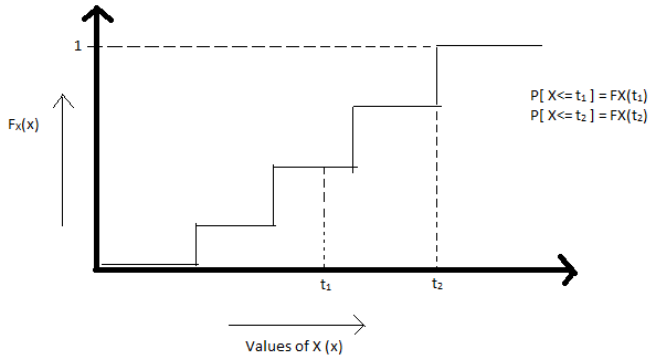
Brief About Our Work

- Discussion About Quantile Regression
- Use of Quantile Regression in Some Simulated Cases
- Use of Quantile Regression in a Real Life Dataset

Brief About Quantile Regression



(a) Quantile(Continuous)



(b) Quantile(Discrete)

Figure: Diagram to Show Quantiles

- If Q_τ is τ^{th} quantile of a distribution, then:

$$P[X \leq Q_\tau] = \tau$$

Equation of quantile regression

Equation of Quantile Regression at $\tau^{th}(0 < \tau < 1)$ quantile:

$$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \cdots + \beta_p(\tau)x_{ip} + \epsilon_i$$

where,
$y_i = i^{th}$ response variable
$x_{i1}, \cdots x_{ip}$ = set of p independent variables
ϵ_i =error term in model
$i = 1, \cdots, n$,
n= Total number of observations,
p=total number of independent variables
$\beta_i(\tau)'s$:coefficient 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.

Optimization of β coeficients

The least absolute estimates $\hat{\beta}$ for the conditional median is obtained as the solution of the minimization problem:

- To minimize:

$$\min_{\beta \in \mathbf{R}^p} \sum_{i=1}^{n} |y_i - x_i' \beta|$$

- Minimization problem reduces to:

$$\min_{\beta \in \mathbf{R}^p} [\sum_{y_i \geq x_i' \beta} \tau |y_i - x_i' \beta| + \sum_{y_i < x_i' \beta} (1 - \tau) |y_i - x_i' \beta|]$$

- The above expression can be formulated in linear programming and can be optimize using simplex method

- Check Function:

$$\rho_\tau(u) = \tau max(u, 0) + (1 - \tau)max(-u, 0)$$

where,

$$u = y_i - (\beta_0(\tau) + \beta_1(\tau)x_{i1} + \cdots + +\beta_p(\tau)x_{ip})$$

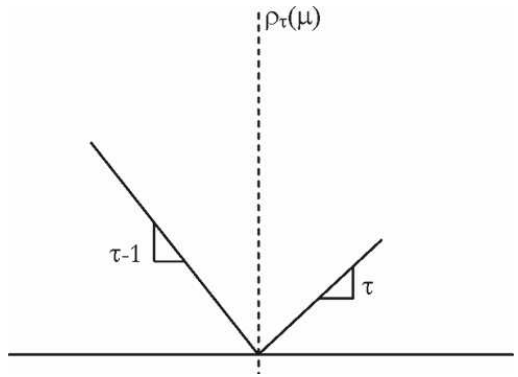


Figure: Visual Representation of Check Function

The check function is a loss function that retrieves the τ-th sample quantile.

Distribution of β coefficents

The quantile regression estimator $\hat{\beta}(\tau)$ is asymptotically distributed as:

$$\sqrt{n}[\hat{\beta}(\tau) - \beta(\tau)] \rightarrow N(0, w^2(\tau)D^{-1})$$

Here,

- scale parameter $w^2(\tau) = \frac{\tau(1-\tau)}{f(F^{-1}(\tau))^2}$, a function of $s = \frac{1}{f(F^{-1}(\tau))}$, the so-called sparsity function
- sparsity (as a count or proportion) of a matrix. For example, 0.99 sparsity means 99% of the values are zero. Similarly, a sparsity of 0 means the matrix is fully dense.
- $D = lim_{n\rightarrow\infty} \frac{1}{n} \sum_i x_i^T x_i$, a positive definite matrix

Hypothesis Testing For β coefficents

Testing whether the β coefficients are significantly different from zero or not:

The statistical relevance of the median(quantile) estimated regression is assessed through the usual Student-t statistic since QR estimator is asymptotically normal, standardizing it with estimated error in place of the unknown true standard error yields a Student-t distribution

$H_{0i}: \beta_i(\tau)=0$ vs $H_{1i}: \beta_i(\tau) \neq 0$

Test Statistic under $H_{0i} = \hat{\beta}_i(\tau)/\hat{SE}(\hat{\beta}_i(\tau),\ i = 1, \cdots, p$

If p-value$<$0.05, we will reject the null hypothesis (i.e. coefficient will be significant) at 5% level of significance.

For exclusion of more than one coefficient at a time

The exclusion of a single explanatory variable can be decided on the basis of the Student-t test.
The Wald (W), Lagrange multiplier (LM) and likelihood ratio (LR) tests, which are asymptotically equivalent and asymptotically distributed as a χ^2, allow to verify the exclusion of more than one coefficient at a time. The degrees of freedom of the χ^2 are equal to the number of coefficients under test.

- **Wald Test**: If their quadratic function is close to zero, the variables under test can be safely excluded. The test function is:
  $W = n\omega^{-2}\hat{\beta}(\tau)^T[D^{22}]^{-1}\hat{\beta}(\tau)$
  To test the null $H_0 : \beta_3 = \beta_5 = 0$, the vector $\beta(\tau)$ of the estimated coefficients under test, is given by $\hat{\beta}(\tau) = (\hat{\beta}_3, \hat{\beta}_5)'$

Some cases where Quantile Regression is appropriate:

- If errors in Linear Regression Model is not normally distributed
- If errors in Linear Regression Model are heteroscedastic
- If data contains outliers
- In case of dependent errors

In next few slides we will show use of Quantile Regression instead of Linear Regression (OLS) with help of simulated and real life data.

Doing some simulation

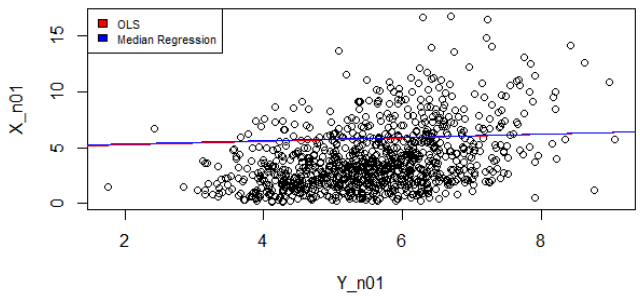
- Simulation Equation:

$$y_i = 5 + (0.15 * x_i + e_i)$$

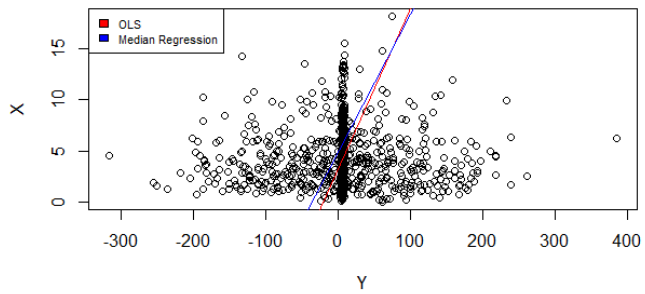
- simulaion contains total of 10000 repetations
- each repetation contains 1000 sample draw
- x_i's are drawn from χ^2_4 distribution

- For homoscedastic Error - $e_i \sim N(0,1) \text{for} i = 1, \cdots, 1000$
- For heteroscedastic Error -

$$e_i \sim \begin{cases} N(0,1) & \text{for} \quad i = 1, \cdots, 500 \\ N(0,100) & \text{for} \quad i = 501, \cdots, , 1000 \end{cases}$$



(a) For Homoscedastic Error



(b) For Heteroscedastic Error

Figure: Diagrams on OLS and Median Regression for Simulated Data with Homoscedastic and Heteroscedastic Error

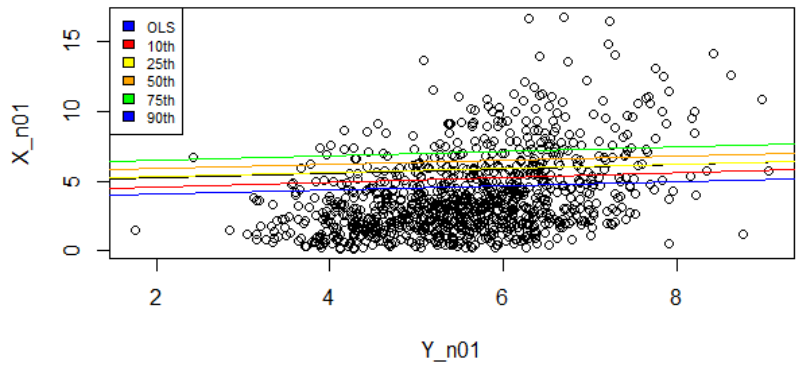


Figure: Plot of Quantile Regression of All Quantiles along with OLS Regression Over