MTH516A : Nonparametric Inference

Quantile Regression

Supervised by : Dr. Dootika Vats

Koyel Pramanick[*] Vijay Kumar[†] Harshit Garg[‡]

January 4, 2025

Our journey of accomplishing this project really involves many ones to whom we are highly obliged. We would like to express our deepest appreciation to all those who have provided us the possibility to complete this project. We give a special gratitude to our respected instructor Dr. Dootika Vats, Associate professor at Department of Mathematics and Statistics, IIT KANPUR, whose contribution in stimulating suggestion, valuable guidance, constructive criticism and encouragement help us to coordinate our project.

[*] 201333,koyel20@iitk.ac.in

[†] 201454,kumarvijay20@iitk.ac.in

[‡] 201321,harshitg20@iitk.ac.in

Abstract

Quantile regression offers a more complete statistical model than mean regression and now has widespread applications. Consequently, we provide a review of this technique. We begin with an introduction to and motivation for quantile regression.The main purpose of this project is to analyze the advantages of using Quantile Regression over Linear Regression. We have done various simulations to infer that the quantile regression works better when the assumptions of linear regression are not met. We have defined the tests and optimizing technique we are using to obtain the estimates of the coefficients of the quantile regression model. We then discuss some typical application areas. Next we outline various approaches to estimation. Some applications of quantile regression over simualted and real dataset are also done in this project.

Contents

1 Introduction

Quantile Regression is statistical analysis which is not restricted to the conditional mean. It approximates whole conditional distribution of a response variable.

In section 2, we will discuss about our objective; in section 2, we will discuss brief about Quantile Regression; in section 3 and 4 we will discuus implementation of Quantile Regression over simulated and real life dataset respectively.

We have used R software for necessary computations.

2 Objective of Our Project

Our objective is to give brief idea about:

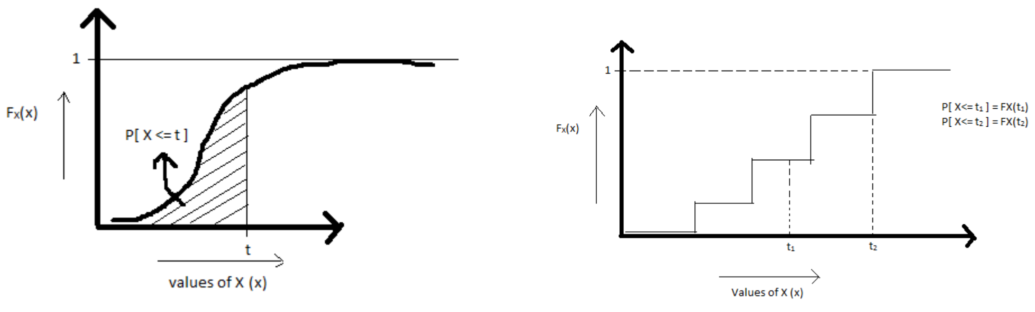
- Quantile Regression
- Why traditional Linear Regression is not appropriate in some cases and
- How we have to use Quantile regression in those cases.
- Will show some results from simulated as well as real life dataset.

3 Methodology

3.1 What is Quantile?

A quantile is where a sample is divided into equal-sized adjacent subgroups i.e. **a quantile determines how many values in a distribution are above or below a certain limit**. Quartiles are also quantiles; they divide the distribution into four equal parts.The first quantile include all values that are smaller than a quarter of all values.Median is placed in a probability distribution so that exactly half of the data is lower than the median and half of the data is above the median.The interquantile range between the first and third quartile equals the range in which 50% of all values lie that are distributed around the mean.
Let Y be a real valued random variable with cumulative distribution function $F_Y(y) = P(Y \leq y)$. The τ th quantile of Y is given by

$q_Y(\tau) = F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$



(a) Quantile(Continuous)

(b) FQuantile(Discrete)

Figure 1: Diagram to Show Quantiles

3.2 Theory of Linear and Quantile Regression

3.2.1 Linear Regression and its assumptions

Linear regression is used to study the linear relationship between a dependent variable (y) and one or more independent variables (X). The linearity of the relationship between the dependent and independent variables is an assumption of the model. The relationship is modeled through a random disturbance term (or, error variable) ϵ. The disturbance is primarily important because we are not able to capture every possible influential factor on the dependent variable of the model. To capture all the other factors, not included as independent variable, that affect the dependent variable, the disturbance term is added to the linear regression model. In this way, the linear regression model takes the following form:

$$y = X\beta + \epsilon$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where $\beta_j \in \mathbf{R}, j = 1, \cdots, p$

are the regression coefficients of the model (which we want to estimate!), and p is the number of independent variables included. The equation is called the regression equation.
Assumptions:

- The relationship between the dependent variable, independent variable, and the disturbance is linear.
- We have a random sample of size n $\{(x_i, y_i) : i = 1, \cdots, n\}$, where the observations are independent of each other.
- None of the independent variables is constant, and there are no exact linear relationships among the indeepndent variables.
- The disturbance term has an expected value of zero given any value of the independent variable. In other words $E(\epsilon|x_i) = 0$.
- The disturbance term has the same variance given any value of the indeepndent variable. In other words $Var(\epsilon|x_i) = \sigma^2$.

3.3 Quantile regression

Quantile regression models the relationship between a set of predictor (independent) variables and specific percentiles (or ”quantiles”) of a target (dependent) variable, most often the median.
It has two main advantages over Ordinary Least Squares regression:

- Quantile regression makes no assumptions about the distribution of the target variable.
- Quantile regression tends to resist the influence of outlying observations.

Equation of Quantile Regression at τ^{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,
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.

3.4 Why should we use QR?

- The main advantage of quantile regression methodology is that the method allows for understanding relationships between variables outside of the mean of the data,making it useful in understanding outcomes that are non-normally distributed and that have nonlinear relationships with predictor variables.
- Quantile regression allows the analyst to drop the assumption that variables operate the same at the upper tails of the distribution as at the mean.
- The quantile regression method weights different portions of the sample to generate coefficient estimates, thus increasing the power to detect differences in the upper and lower tails.
- QR is an invaluable tool for facing heteroskedasticity, and provides a method for modeling the rates of change in the response variable at multiple points of the distribution when such rates of change are different.
- QR is useful in the case of homogeneous regression models outside of the classical normal regression model, and in the case where the error independence assumption is violated, as no parametric distribution assumption is required for the error distribution.

3.5 Statistical Tests Used

3.5.1 Normal Q-Q plot

A normal probability plot, or more specifically a quantile-quantile (Q-Q) plot, shows the distribution of the data against the expected normal distribution.
For normally distributed data, observations should lie approximately on a straight line. If the data is non-normal, the points form a curve that deviates markedly from a straight line. Possible outliers are points at the ends of the line, distanced from the bulk of the observations.

3.5.2 Shapiro-wilk test

The Shapiro–Wilk test is a test of normality in frequentist statistics. The Shapiro-Wilk test tests the null hypothesis that a sample $x_1, x_2, \cdots, x_n$ came from a normally distributed population. The test statistic is:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $x_{(i)}$ is the order statistic and $\bar{x}$ is the sample mean.
The coefficients a_i are given by:

$$(a_1, a_2, \cdots, a_n) = \frac{m^T V^{-1}}{C}$$

where C is a vector norm.

$$C = ||V^{-1}m|| = (m^T V^{-1} V^{-1} m)^{1/2}$$

and the vector m, $m = (m_1, \cdots, m_n)^T$ is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution. Finally, V is the covariance matrix of those normal order statistics.
The null-hypothesis of this test is that the population is normally distributed. Thus, if the p value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. On the other hand, if the p value is greater than the chosen alpha level, then the null hypothesis (that the data came from a normally distributed population) can not be rejected (e.g., for an alpha level of .05, a data set with a p value of less than .05 rejects the null hypothesis that the data are from a normally distributed population)

3.6 Transformations Over Original Variable

3.6.1 Boxcox Transformation

Boxcox transformation is used to make non-normally distributed data to normally distributed data. The function is as below:

$$f(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0, \\ log(x) & \lambda = 0 \end{cases}$$

where λ is the transformation parameter. Below table will show some common transformations for some particular λ.

λ value	Expression
-2	$\frac{1}{x^2}$
-1	$\frac{1}{x}$
-0.5	$\frac{1}{\sqrt{x}}$
0	$log(x)$
0.5	$\sqrt{x}$
1	x
2	x^2

Table 1: Table Showing Common Values of λ Along With Expression

3.7 Optimization Problem

The model for linear quantile regression is

$$y = A'\beta + \epsilon$$

where $y = (y_1, \cdots y_n)$' is the $(n \star 1)$ vector of responses, $A' = (x_1, \cdots x_n)$' is the $(n \star p)$ regressor matrix,$\beta = (\beta_1, \cdots \beta_p)$' is the $(p \star 1)$ vector of unknown parameters, and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)$' is the $(n \star 1)$ vector of unknown errors.

L1 regression, also known as median regression, is a natural extension of the sample median when the response is conditioned on the covariates. In L1 regression, the least absolute residuals estimate $\hat{\beta}_{LAR}$, referred to as the L1-norm estimate, is obtained as the solution of the following minimization problem:

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

More generally, for quantile regression Koenker and Bassett (1978) defined the τ regression quantile, $0 < \tau < 1$, as any solution to the following minimization problem:

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

The solution is denoted as $\hat{\beta}$, and the L1-norm estimate corresponds to $\hat{\beta}(1/2)$. The τ regression quantile is an extension of the τ sample quantile $\hat{\xi}(\tau)$, which can be formulated as the solution of

$$\min_{\xi \in \mathbf{R}} [\sum_{i \in i: y_i \geq \xi} \tau |y_i - \xi| + \sum_{i \in i: y_i < \xi} (1-\tau)|y_i - \xi|]$$

If you specify weights $w_i, i = 1, 2, \cdots, n$, with the WEIGHT statement, weighted quantile regression is carried out by solving

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

Weighted regression quantiles β_w can be used for L-estimation (Koenker and Zhao 1994).

3.8 Detailed derivation of Simplex method

Let $\mu = [y - A'\beta]_+$, $\nu = [A'\beta - y]_+$, $\phi = [\beta]_+$ and $\psi = [-\beta]_+$
where $[z]_+$ is the nonnegative part of z.
Let $D_{LAR}(\beta) = \sum_{i=1}^{n} |y_i - x_i'\beta|$
For the L1 problem, the simplex approach solves $\min_\beta D_{LAR}(\beta)$ by reformulating it as the constrained minimization problem

$$\min_{\beta} \{e'\mu + e'\nu | y = A'\beta + \mu - \nu, \{\mu, \nu\} \in \mathbf{R}_+^n\}$$

where e denotes an $(n \star 1)$ vector of ones.
Let $B = [A' - A' I - I]$, $\theta = (\phi' \psi' \mu' \nu')'$ and $d = (0' 0' e' e')'$ where $0' = (0 \cdots 0)_p$.
The reformulation presents a standard LP problem:

$$\min_{\theta} d'\theta;$$

$$\text{subject to } B\theta = y, \theta \geq 0$$

This problem has the following dual formulation:

$$\max_z y'z;$$

$$\text{subject to } B'z \leq d$$

This formulation can be simplified as

$$\max_z y'z;$$

$$\text{subject to } Az = 0, z \in [-1, 1]^n$$

By setting $\eta = \frac{1}{2} + \frac{1}{2}e$, $b = \frac{1}{2}Ae$, the problem becomes

$$\max_\eta y'\eta;$$

$$\text{subject to } A\eta = b, \eta \in [0, 1]^n$$

For quantile regression, the minimization problem is $\min_\beta \sum \rho_\tau(y_i - x_i'\beta)$, and a similar set of steps leads to the dual formulation

$$\max_z y'z;$$

$$\text{subject to } Az = (1-\tau)Ae, z \in [0, 1]^n$$

The QUANTREG procedure solves this LP problem by using the simplex algorithm of Barrodale and Roberts (1973). This algorithm exploits the special structure of the coefficient matrix B by solving the primary LP problem (P) in two stages: The first stage chooses the columns in A' or -A' as pivotal columns. The second stage interchanges the columns in I or –I as basis or nonbasis columns, respectively. The algorithm obtains an optimal solution by executing these two stages interactively. Moreover, because of the special structure of B, only the main data matrix A is stored in the current memory.

3.9 Evaluation Metric

$$R_\tau^2 = 1 - \frac{RASW_\tau}{TASW_\tau}$$

where,

RASW$_\tau$ = Residual Absolute Sum of Weighted Differences
TASW$_\tau$ = Total Absolute Sum of Weighted Differences
Here, $0 \leq {R^2}_\tau \leq 1$

We have used R software and it uses 'br' method by default to estimate coefficients of quantile regression. This method is modified over simplex method of linear programming problem. This method was first introduced by Barrodale and Roberts in 1973.

3.10 Distribution of the coefficients of quantile regression

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

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

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

- For example, 0.99 sparsity means 99% of the values are zero. Similarly, a sparsity of 0 means the matrix is fully dense.
- $D = lim_{n\to\infty}\frac{1}{n}\sum_i x_i^T x_i$, a positive definite matrix

3.11 Tests for checking significance of coefficients of quantile from zero

3.11.1 Student t-test

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}:$\beta_i(\tau) = 0$ (The coefficients are not significant from zero)
H_{1i} : $\beta_i(\tau) \neq 0$ (The coefficients are significant from zero)

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

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

3.11.2 Wald Test

The exclusion of a single explanatory variable can be decided on the basis of the Student-t test.The test to verify the hypothesis on more than one coefficient at a time is the Wald test. The Wald (W) test, is asymptotically equivalent and asymptotically distributed as a χ^2.The degrees of freedom of the χ^2 are equal to the number of coefficients under test.
It considers the gradient g of the model excluding the variables under test.The gradient is a function of the sign of the errors, of their position above or below the QR line. Their numerical value is not relevant, but their position with respect to the fitted line and thus their sign is crucial.
The D matrix, a quadratic form comprising all the explanatory variables $D = X^T X$, is partitioned in blocks
$D = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}$
The vector g together with **the submatrix D_{22} select the variables under test**, and its inverse D^{22} is given by $D^{22} = (D_{22} - D_{21}D_{11}D_{12})^{-1}$
If their quadratic function is close to zero, the variables under test can be safely excluded.

To test the null $H_0 : \beta_i = \beta_j = 0$, the vector $\beta(\tau)$ of the estimated coefficients under test, is given by $\hat{\beta}(\tau) = (\hat{\beta}_i, \hat{\beta}_j)'$
The test function is:

$$W = n\omega^{-2}\hat{\beta}(\tau)^T[D^{22}]^{-1}\hat{\beta}(\tau)$$

4 Analysis Over Simulated Life Data

We have simulated data of different type:
Equation used in Simulation:

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

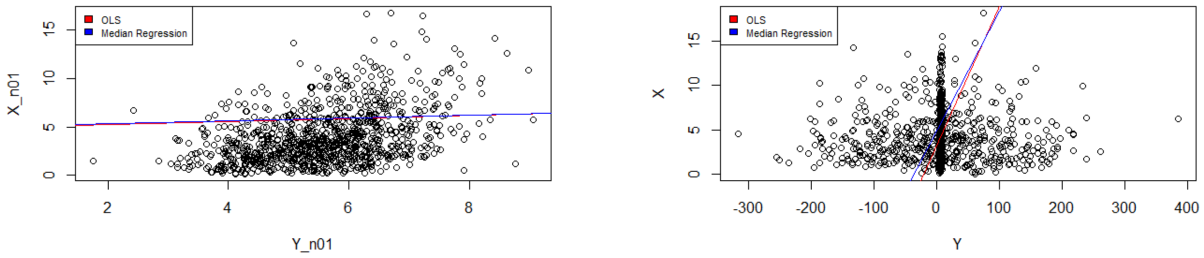
- simulaion contains total of 10000 repetations
- each repetation contains 1000 sample draw
- x_i's are drawn from χ^2_4
- For homoscedastic Error - $e_i \sim N(0,1) \text{for} i = 1, \cdots, 1000$
- For heteroscedastic Error -

$$e_i \sim \left\{ \begin{array}{lll} N(0,1) & \text{for} & i = 1, \cdots, 500 \\ N(0,100) & \text{for} & i = 501, \cdots, , 1000 \end{array} \right.$$

Table 2: Results from Simulated Data

Error of the random variable		Estimator					
		θ=0.10	θ=0.25	θ=0.50	θ=0.75	θ=0.90	OLS
$e_i \sim N(0,1)$	Mean	0.1501	0.1499	0.1501	0.1501	0.1498	0.1500
	SD	0.0189	0.0153	0.0141	0.0152	0.0190	**0.0111**
$e_i \sim N(5,100)$	Mean	0.1454	0.1305	0.1460	0.1404	0.1308	0.1381
	SD	1.9115	1.5268	1.4039	1.5242	1.9342	**1.1161**
$e_i \sim \chi^2_5$	Mean	0.1510	0.1506	0.1505	0.1498	0.1495	0.1501
	SD	**0.0279**	0.0319	0.0406	0.0587	0.0911	0.0353
$e_i \sim t_2$	Mean	0.1479	0.1494	0.1499	0.1498	0.1513	0.1498
	SD	0.0440	0.0212	**0.0158**	0.0212	0.0444	0.0451
$e_i \sim \left\{ \begin{array}{lll} N(0,1) & \text{for} & i = 1, \cdots, 500 \\ N(0,100) & \text{for} & i = 501, \cdots, , 1000 \end{array} \right.$	Mean	0.1759	-0.0955	0.1499	0.4037	0.1380	0.1513
	SD	2.3853	0.7644	**0.0287**	0.7838	2.3939	0.7970

From Table: 2, it can be easily seen that, for cases where errors are generated from Normal distribution, OLS regression is working better. For cases, where errors are coming from heteroscedastic distribution and skewed distribution, there Quantile Regression is working better. **Working Better** is indicated by low precision here.



(a) For Homoscedastic Error

(b) For Heteroscedastic Error

Figure 2: Diagrams on OLS and Median Regression for Simulated Data with Homoscedastic and Heteroscedastic Error

From Fig: 2, it is also clear that, for cases where errors are generated from Normal distribution, OLS regression and median regression (50th quantile regression) are coinciding. For cases, where errors are coming from heteroscedastic distribution, there OLS regression and median regression (50th quantile regression) has different slopes.

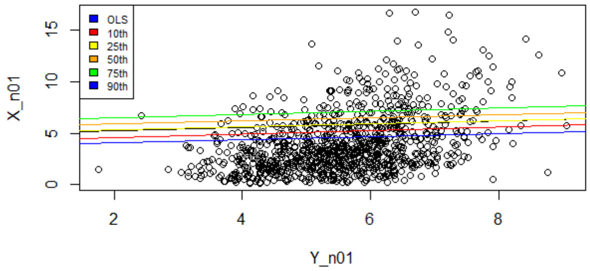


Figure 3: Plot of Quantile Regression of All Quantiles

In above figure Fig: 5, we have plotted Quantile Regression equations (at 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} quantiles) along with OLS Regression.

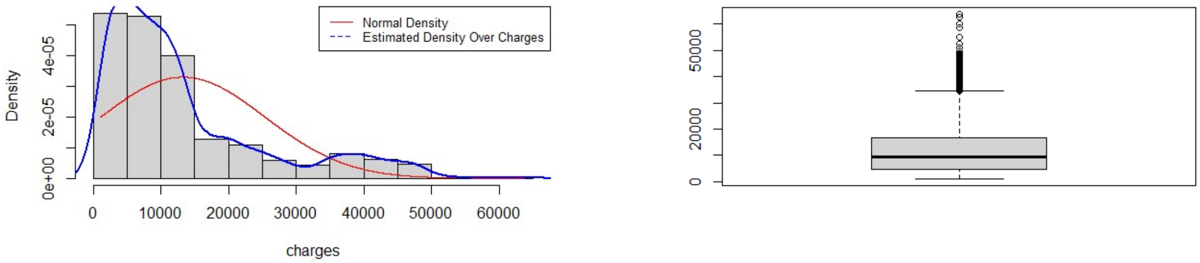
5 Analysis Over Real Life Data

5.1 About Dataset

We are using Medical Cost Personal Dataset containing 6 features (age, sex, bmi, children, smoker, region) along with charges as dependent variable.

5.2 Why Quantile Regression?

As here one question should definitely arise that why we are using Quantile Regression instead of traditional Linear Regression method. To answer this first we have plot histogram and boxplot of the original response variable **charges**.



(a) Histogram of Original Response

(b) Boxplot of Original Response

Figure 4: Diagrams on Original Response Variable (Charges)

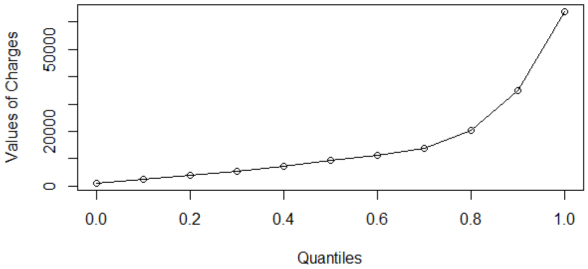


Figure 5: Plot Original Response Values by Quantiles

Now we can see from Fig:5 that starting from around 0.8 quantile till end i.e. at the tail portion there are high values (from Fig:4(b) also this fact can be seen).

Though it is clear from Fig:4 that response variable is not normally distributed and contains outliers also (i.e. our data is skewed enough), we have fitted Linear Regression over data with original response values. Below are some diagrams which will show the assumptions are not getting satisfied here.

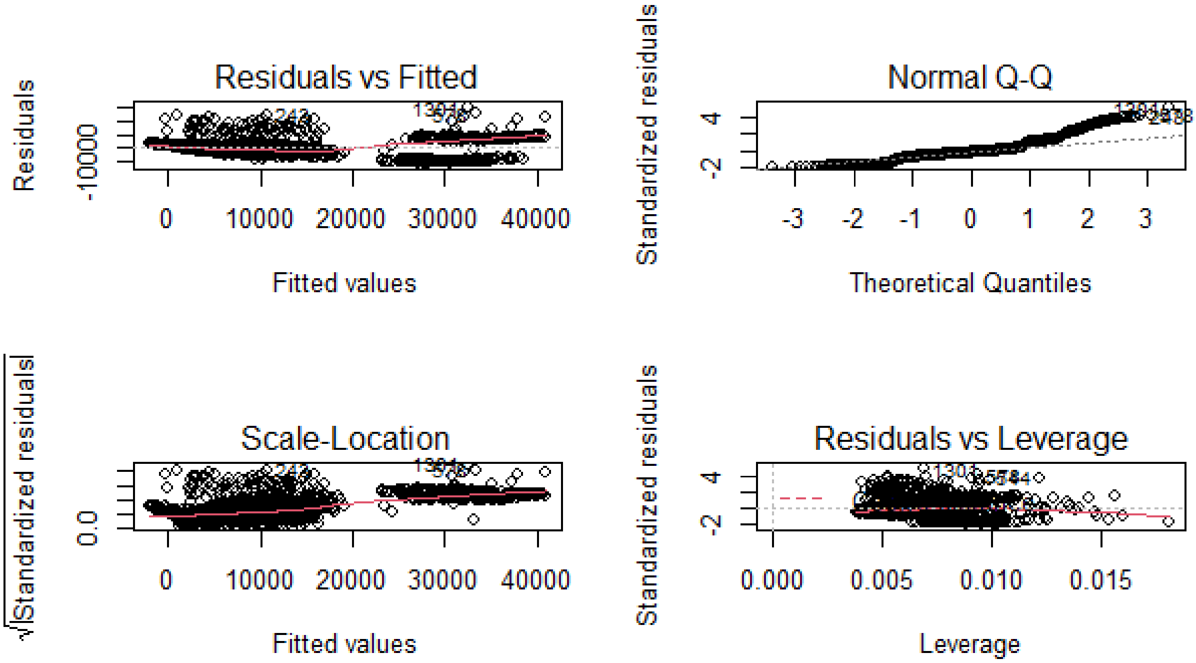


Figure 6: Plot of OLS Regression of Original Response Variable

We have performed shapiro-wilk check over residual distribution.

Then we have performed boxcox transformation over response variable and from Fig:7 below it is clear that optimum value for λ is near 0 which leads us to use $\lambda = 0$ i.e. *log* transformation over response variable. Below the diagrams are shown:

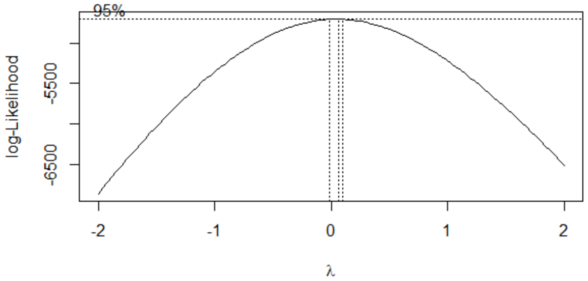
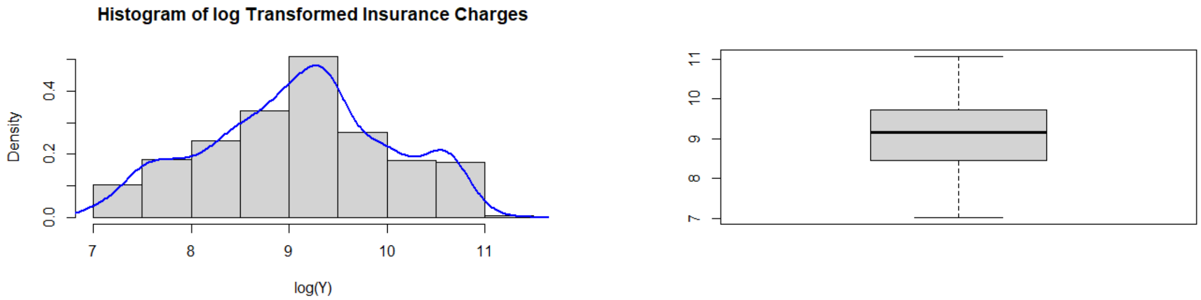


Figure 7: Plot of λ and Log-Liklihood for Boxcox Transformation



(a) Histogram of Log Transformed Response

(b) Boxplot of Log Transformed Response

Figure 8: Diagrams on Log Transformed Response Variable (Charges)

It is clear that log transformed response variable is still not normal, though after Log transformation here is no outlier.

We again perform Linear Regression, plot all diagrams (shown below) and performed Shapiro-wilk test and check on this model as before.

All the results from Shapiro-wilk is shown in tables below:

Type of Linear Regression Model	Test Statistic	P-Value
with Original Response	0.89909	$<$ 2.2e-16
with Log Transformed Response	0.83751	$<$ 2.2e-16

Table 3: Table Showing Results of Shapiro Wilk Test

As shown residual from fitted OLS Linear Regression model are not normally distributed we have performed boxcox transformation and from the diagram in Fig: 7 it is clear that value near 0 will be optimum value. We are taking $\lambda = 0$ for sake of simplicity.

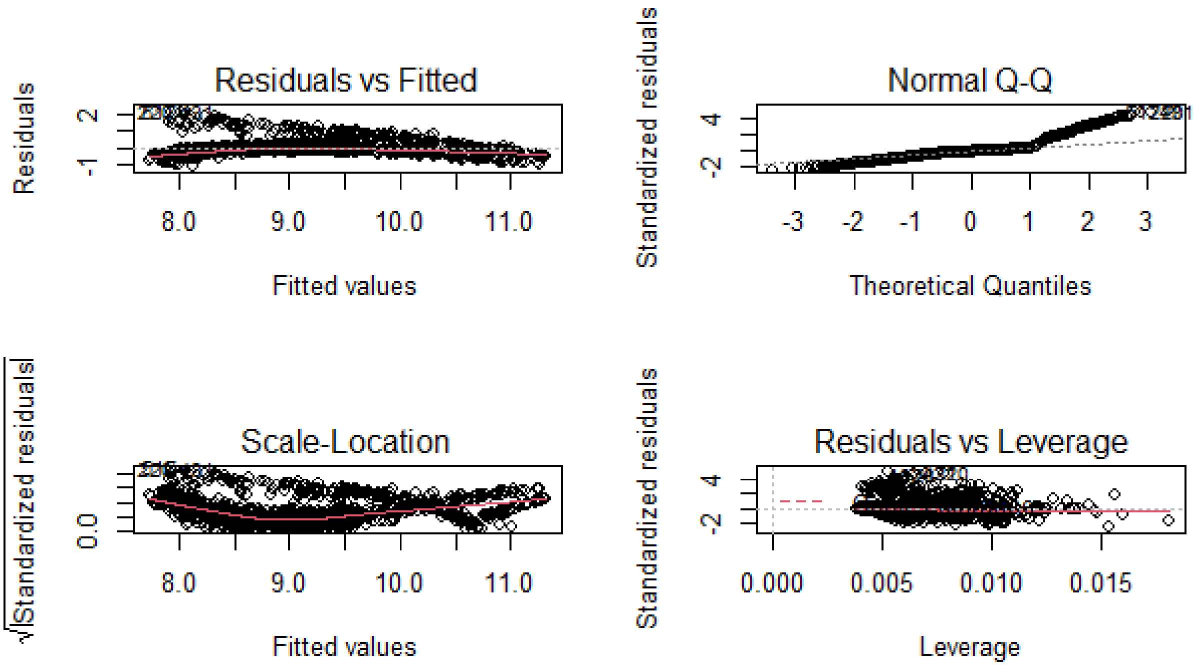


Figure 9: Plot of OLS Regression of log Transformed Response Variable

As already discussed about in theory portion when quantile regression is useful, from previous results and diagrams shown here, it is clear that we should use Quantile Regression in place of traditional OLS Linear Regression.

5.3 Fitted Quantile Equations

We will show fitted Quantile Regression Equations for 10th, 25th, 50th, 75th and 90th quantiles below:

At 10th quantile:

$y = -4250.87314 + 264.95672 * age - 477.13143 * sexmale + 6.87245 * bmi + 568.04284 * children + 13407.36952 * smokeryes - 206.01895 * regionnorthwest - 586.52376 * regionsoutheast - 583.76106 * regionsouthwest$

At 25th quantile:

$y = -4287.23348 + 262.32116 * age - 443.50040 * sexmale + 22.26785 * bmi + 482.63108 * children + 140954.11249 * smokeryes - 251.46096 * regionnorthwest - 706.75379 * regionsoutheast - 645.86575 * regionsouthwest$

At 50th quantile:

$y = -4405.08854 + 263.26995 * age - 464.29432 * sexmale + 42.56613 * bmi + 401.35700 * children + 31228.02056 * smokeryes - 250.66948 * regionnorthwest - 670.40446 * regionsoutheast - 668.59571 * regionsouthwest$

At 75th quantile:

$y = -4151.24679 + 266.63931 * age - 448.12450 * sexmale + 48.54750 * bmi + 434.00336 * children + 32919.79845 * smokeryes - 369.00893 * regionnorthwest - 548.03620 * regionsoutheast - 692.75902 * regionsouthwest$

At 90^{th} quantile:

$y = -10704.06877 + 261.913138 * age - 529.59724 * sexmale + 446.88104 * bmi + 521.48960 * children$
$+27193.15517 * smokeryes - 1117.94259 * regionnorthwest - 1287.09303 * regionsoutheast - 1572.10941 * regionsouthwest$

5.4 Results from Tests of Coefficients

Coefficients are Significantly different at each quantile for each variable separately and jointly

Variable	Value	Std. Error	t value	Pr($>|t|$)
(Intercept)	-4250.87314	37.93514	-112.05634	0.00000
age	264.95672	0.42582	622.22074	0.00000
sexmale	-477.13143	10.70122	-44.58666	0.00000
bmi	6.87245	1.39230	4.93606	0.00000
children	568.04284	5.52643	102.78654	0.00000
smokeryes	13407.36952	254.67764	52.64447	0.00000
regionnorthwest	-206.01895	14.02901	-14.68521	0.00000
regionsoutheast	-586.52376	20.70952	-28.32145	0.00000
regionsouthwest	-583.76106	16.94392	-34.45253	0.00000

Table 4: Coefficients and Significance of Coefficients at 10^{th} Quantile Regression Equation

Variable	Value	Std. Error	t value	Pr($>|t|$)
(Intercept)	-4287.23348	54.00305	-79.38873	0.00000
age	262.32116	1.11522	235.21928	0.00000
sexmale	-443.50040	19.55518	-22.67944	0.00000
bmi	22.26785	2.65538	8.38594	0.00000
children	482.63108	10.43149	46.26672	0.00000
smokeryes	14954.11249	336.47461	44.44351	0.00000
regionnorthwest	-251.46096	31.63621	-7.94852	0.00000
regionsoutheast	-706.75379	35.00185	-20.19190	0.00000
regionsouthwest	-645.86575	34.59252	-18.67068	0.00000

Table 5: Coefficients and Significance of Coefficients at 25^{th} Quantile Regression Equation

Variable	Value	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	-4405.04854	186.91120	-23.56760	0.00000
age	263.26995	2.55394	103.08371	0.00000
sexmale	-464.29432	68.02973	-6.82487	0.00000
bmi	42.56613	6.28591	6.77167	0.00000
children	401.35700	25.30376	15.86155	0.00000
smokeryes	31228.02056	2760.29572	11.31329	0.00000
regionnorthwest	-250.66948	81.76909	-3.06558	0.00222
regionsoutheast	-670.40446	112.53407	-5.95735	0.00000
regionsouthwest	-668.59571	90.79975	-7.36341	0.00000

Table 6: Coefficients and Significance of Coefficients at 50^{th} Quantile Regression Equation

Variable	Value	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	-4151.24679	246.17246	-16.86316	0.00000
age	266.63931	2.81891	94.58941	0.00000
sexmale	-448.12450	84.89501	-5.27857	0.00000
bmi	48.54750	7.47985	6.49043	0.00000
children	434.00336	45.26325	9.58843	0.00000
smokeryes	32919.79845	310.31046	106.08665	0.00000
regionnorthwest	-369.00893	139.71487	-2.64116	0.00836
regionsoutheast	-548.03620	139.29353	-3.93440	0.00009
regionsouthwest	-692.75902	116.44077	-5.94945	0.00000

Table 7: Coefficients and Significance of Coefficients at 75^{th} Quantile Regression Equation

Variable	Value	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	-10704.06877	4663.29718	-2.29539	0.02187
age	261.91318	36.03261	7.26878	0.00000
sexmale	-529.59724	1240.61028	-0.42688	0.66953
bmi	446.88104	157.67766	2.83414	0.00466
children	521.48960	503.83205	1.03505	0.30084
smokeryes	27193.15517	1700.00758	15.99590	0.00000
regionnorthwest	-1117.94259	1822.86140	-0.61329	0.53979
regionsoutheast	-1287.09303	1591.70297	-0.80863	0.41887
regionsouthwest	-1572.10941	1565.04611	-1.00451	0.31531

Table 8: Coefficients and Significance of Coefficients at 90^{th} Quantile Regression Equation

Coefficients are Significantly different at each quantile for each variable separately and jo Jointly

Quantiles	F-value	P-Value
$10^{th}, 25^{th}$	18.773	$<$ 2.2e-16
$25^{th}, 50^{th}$	8.9782	3.171e-12 (*)
$50^{th}, 75^{th}$	1.2187	0.2836
$75^{th}, 90^{th}$	2.7978	0.004366

Table 9: jointly

Quantiles	F-value	P-Value
$10^{th}, 25^{th}, 50^{th}$	12.939	$<$ 2.2e-16
$25^{th}, 50^{th}, 75^{th}$	149.15	$<$ 2.2e-16
$50^{th}, 75^{th}, 90^{th}$	1.8722	0.01855

Table 10: jointly

Quantiles	F-value	P-Value
$10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}$	108.07	$<$ 2.2e-16

Table 11: jointly

Separatetly

Variable	Df	Resid Df	F value	Pr(>F)
age	1	2675	7.5038	0.006198
sexmale	1	2675	4.1811	0.040975
bmi	1	2675	47.4124	7.141e-12
children	1	2675	91.7253	2.2e-16
smokeryes	1	2675	30.2135	4.235e-08
regionnorthwest	1	2675	2.6803	0.101717
regionsoutheast	1	2675	15.6006	8.025e-05
regionsouthwest	1	2675	4.3271	0.037605

Table 12: separately 10 and 25

Variable	Df	Resid Df	F value	Pr(>F)
age	1	2675	0.1888	0.6639194
sexmale	1	2675	0.1223	0.7265817
bmi	1	2675	14.4817	0.0001447
children	1	2675	13.1977	0.0002856
smokeryes	1	2675	39.7575	3.353e-10
regionnorthwest	1	2675	0.0001	0.9909710
regionsoutheast	1	2675	0.1301	0.7183147
regionsouthwest	1	2675	0.0849	0.7708107

Table 13: separately 25 and 50

Variable	Df	Resid Df	F value	Pr(>F)
age	1	2675	1.8377	0.1753
sexmale	1	2675	0.0505	0.8222
bmi	1	2675	0.8549	0.3553
children	1	2675	0.7463	0.3877
smokeryes	1	2675	0.4248	0.5146
regionnorthwest	1	2675	1.0705	0.3009
regionsoutheast	1	2675	1.0287	0.3106
regionsouthwest	1	2675	0.0560	0.8130

Table 14: separately 50 and 75

Variable	Df	Resid Df	F value	Pr(>F) age
1	2675	0.0187	0.8911505	
sexmale	1	2675	0.0046	0.9456541
bmi	1	2675	6.7068	0.0096568
children	1	2675	0.0331	0.8556938
smokeryes	1	2675	12.1784	0.0004913
regionnorthwest	1	2675	0.1822	0.6695564
regionsoutheast	1	2675	0.2358	0.6272989
regionsouthwest	1	2675	0.3407	0.5594650

Table 15: separately 75 and 90

Variable	Df	Resid Df	F value Pr(>F)	
age	2	4012	3.9472	0.0193844
sexmale	2	4012	2.4528	0.0861776
bmi	2	4012	28.5465	4.896e-13
children	2	4012	52.0099	2.2e-16
smokeryes	2	4012	26.4843	3.743e-12
regionnorthwest	2	4012	1.3666	0.2551025
regionsoutheast	2	4012	8.0342	0.0003294
regionsouthwest	2	4012	2.1645	0.1149390

Table 16: separately 10, 25 and 50

Variable	Df	Resid Df	F value	Pr(>F)
age	2	4012	1.3722	0.2536702
sexmale	2	4012	0.0710	0.9314973
bmi	2	4012	9.3606	8.794e-05
children	2	4012	7.0301	0.0008958
smokeryes	2	4012	1160.6713	2.2e-16
*** regionnorthwest	2	4012	0.5352	0.5855723
regionsoutheast	2	4012	0.7421	0.4761984
regionsouthwest	2	4012	0.0946	0.9097199

Table 17: separately 25, 50 and 75

Variable	Df	Resid Df	F value	Pr(>F)
age	2	4012	1.0551	0.348245
sexmale	2	4012	0.0363	0.964353
bmi	2	4012	3.3709	0.034456
children	2	4012	0.3853	0.680263
smokeryes	2	4012	6.2845	0.001883
regionnorthwest	2	4012	0.5366	0.584783
regionsoutheast	2	4012	0.8447	0.429768
regionsouthwest	2	4012	0.1729	0.841213

Table 18: separately 50, 75 and 90

Variable	Df	Resid Df	F value	Pr(>F)
age	4	6686	2.3878	0.04882
sexmale	4	6686	1.2599	0.28339
bmi	4	6686	16.4856	1.896e-13
children	4	6686	27.1017	2.2e-16
smokeryes	4	6686	800.6596	2.2e-16
regionnorthwest	4	6686	0.9690	0.42311
regionsoutheast	4	6686	4.3502	0.00163
regionsouthwest	4	6686	1.1526	0.32978

Table 19: Separately All 5 Quantiles

QR Coefficients are Significant from OLS coefficients or not

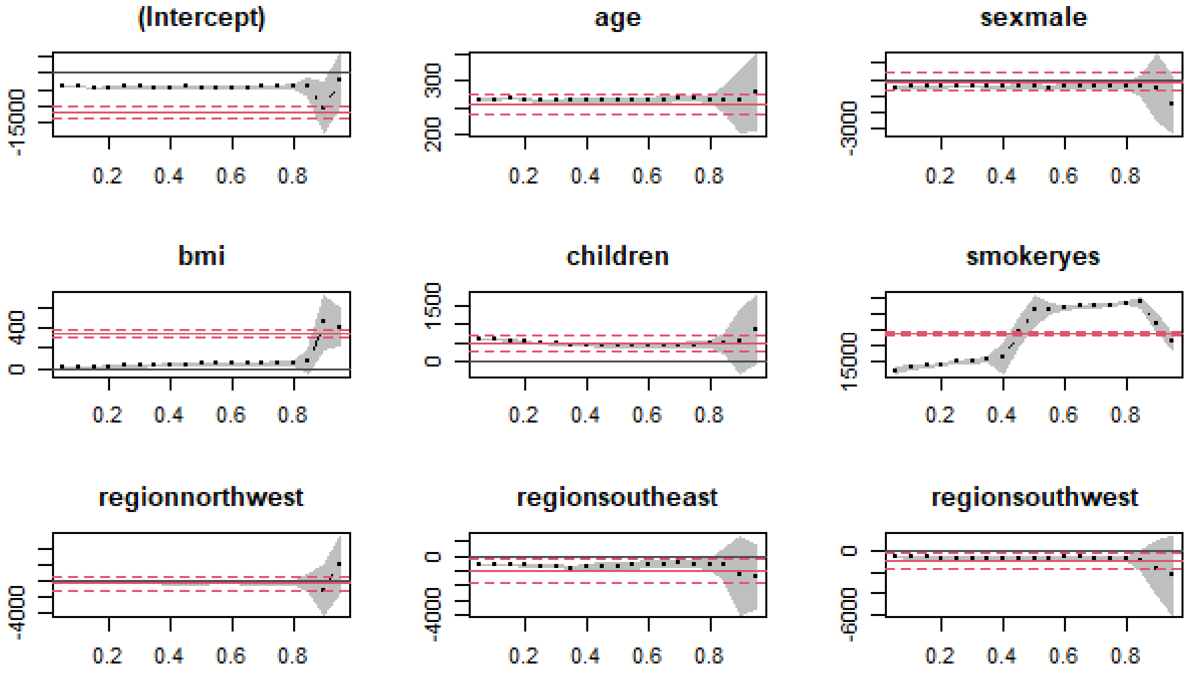


Figure 10: Plot of QR (in Grey) and OLS (in Red) Estimates and Confidence Interval in Both Cases

From this diagram it is easy to conclude that for bmi and smoker along with intercept term QR coefficients are clearly significantly different from OLS coefficients (as for each of these cases, QR coefficients along with QR Confidence Interval is outside of OLS Confidence Interval).

References

1. Quantile Rregression: Applications and Current Research Areas
https://www.jstor.org/stable/4128208
2. Quantile Regression: Theory and Applications by Cristina Davino, Marilena Furno, Domenico Vistocco
3. Quantile Regression Models and their Applications: A Review
https://www.researchgate.net/publication/318557090_Quantile_Regression_Models_and_Their_Application