Quantile Regression

By **Diksha Dilip Patil**

Department Of Statistics,
School Of Mathematical Sciences,
Kavayitri Bahinabai Chaudhari
North Maharashtra University,
Jalgaon.

Content

Introduction

- First proposed by Koenker and Bassett (1978), quantile regression is an extension of the classical least squares estimation.
- OLS regression only enables researchers to approximate the conditional mean and median located at the center of the distribution.
- Quantile regression is applied when an estimate of the various quantiles in a population is desired.
- Classical linear regression estimates the mean response of the dependent variable dependent on the independent variables.
- There are many cases, such as skewed data, multimodal data, or data with outliers.
- when the behavior at the conditional mean fails to fully capture the patterns in the data.

Background

❖ Boscovich proposed median regression in the 18th century.

❖ Laplace and Edgeworth further investigated that idea.

❖ Mosteller and Tukey (1977) first stated that functions could be fitted to describe parts of the response variable (y) distribution aside from simply the mean of the distribution.

❖ Quantile regression (other than median) is the work of Roger Koenker and Gilbert Bassett (1978) – University of Illinois.

What is Quantile?

.

❖ Quantiles are cut points dividing the range of a probability distributions into continuous intervals with equal probabilities, or dividing the observation in a sample in the same way.

❖ In general, $100\tau\%$ of students score less than or equal to q_τ , then q_τ is the τ quantile of the distribution of scores.

❖ Example: If 80 out of 100 students score less than or equal to 85.4, then 85.4 is the 0.80 quantile of the distribution.

What is Quantile Regression?

❖ A form of regression analysis designed to estimate models for the conditional median or other conditional quantile functions of the predictor variable (Y) against the covariates (X's).

❖ Different slopes/rates of change (β's) for different quantiles of the response variable (Y) distribution.

❖ QR is an extension of linear regression. It is an used when the conditions of LR are not met like linearity homoscedasticity and normality. The quantile regression has stronger distributional assumptions.

Basic Model Structure

The τth conditional quantile function is defined as

$$Q(\tau \mid x) = x'\beta(\tau), \quad \text{for } 0 < \tau < 1.$$

Here $\beta(\tau) = (\beta_0(\tau) + \beta_1(\tau) + \beta_2(\tau) + \cdots + \beta_p(\tau))'$ is the quantile coefficient vector.

Thus, $Q(\tau \mid x) = \beta_0(\tau) + \beta_1(\tau)X_1 + \beta_2(\tau)X_2 + \cdots + \beta_p(\tau)X_p$,

$X_1, X_2, \ldots X_p$: predictor variables

Note that the coefficients depend on τ.

Continued...

where $\beta_k(\tau)$, for k = 1,2,...,p measures the change in the τth quantile of Y with respect to x_k.
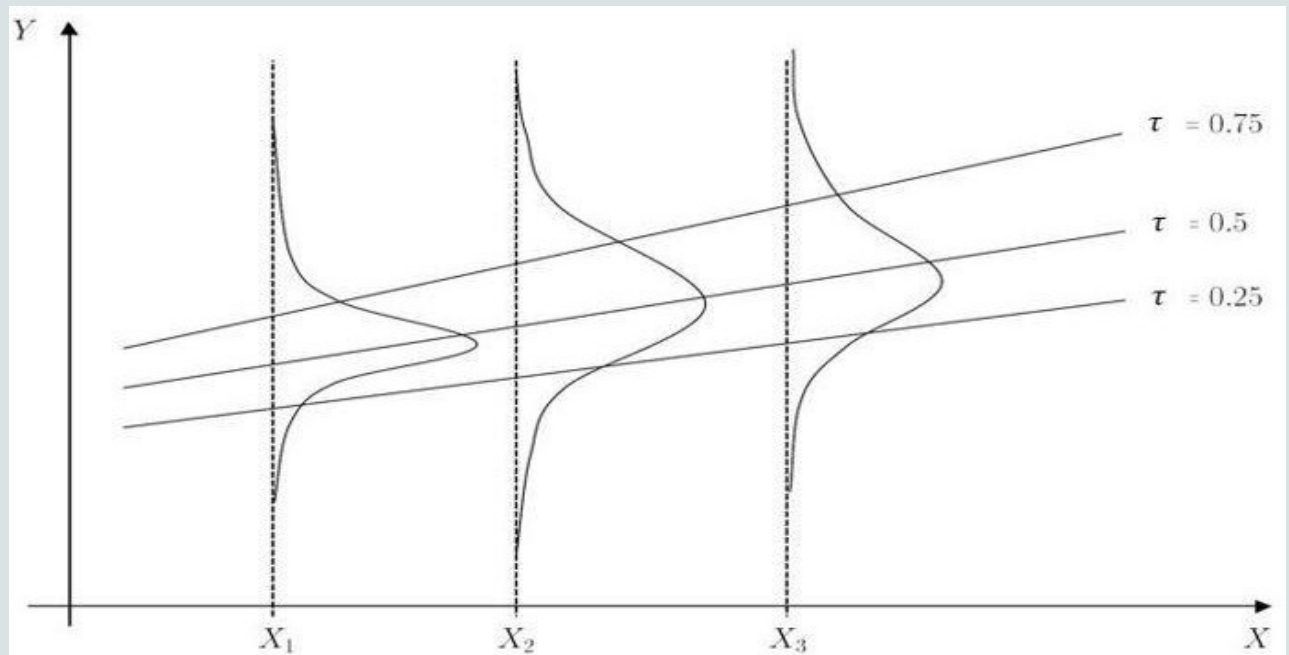
Now, define the quantile regression model

$$Yi=x_i'\beta(\tau)+\epsilon_i(\tau),$$

Where $P(\epsilon_i(\tau)<0) = \tau$, i=1,2,...,n.

Analogous to OLS regression, $\beta(\tau)$ can be estimated by solving the optimization problem.

Quantile Regression

Example of Quantile Regression Benefits : $q_\tau(Y|Xi) - \beta^\tau_0 + \beta^\tau_1 X_i$

Graph adapted from Fitzenberger (2012)

Quantile Regression

- Parameters vary with quantile τ.

- A different equation for every quantile estimated.

- Parameter estimates in linear quantile regression models have the same interpretation as those in any other linear model.

- For a given covariate (x variable), the beta can be interpreted as the change in the quantile of the conditional distribution of the y variable given a 1 unit change in the x variable keeping all other variables in the model constant.

OLS vs. Quantile Regression

- OLS Regression :

1. Minimizes the sum of squared residuals.

2. $E(Y|X=x)=x'\beta$

3. $\widehat{\beta} = \text{argmin}_{\text{All }\beta}\Sigma(y_i - x_i'\beta)^2$

- Quantile Regression :

1. Median: Minimizes the sum of absolute residuals.

2. Other Quantiles: Minimizes the sum of asymmetrically weighted absolute residuals.

3. $q_\tau(Y|X=x)= x'\beta(\tau)$

4. $\widehat{\beta}(\tau)=\text{argmin}_{\text{All }\beta}\Sigma\rho_\tau(y_i - x_i'\beta)$, where $\rho_\tau(.)$ is the quantile loss function.

OLS vs. Quantile Regression

| Characteristics	OLS Regression	Quantile Regression
Assumed Distribution for Errors	Normal	No Distribution Assumption
Variance Assumption	Constant Variance (Homoscedasticity)	Non-Constant Variance (Heteroscedasticity)
Linearity Assumption	Mean is a linear Function of X	Quantile is a linear function of X
Uncorrelated Errors Assumption	Assumptions are imp but adjustments available.	Assumptions are imp but adjustments available.

Why/When Quantile Analysis?

- Compared with conventional mean regression, QR can characterize the entire heterogeneous conditional distribution of the outcome variable.

- It may be more robust to outliers and misspecification of error distribution.

- It provides more comprehensive statistical modelling than traditional mean regression.

- QR models are used to detect heterogeneous effects of covariates at different quantiles of the outcome variable.

- When asymmetries and heavy tails exists, the sample median(the 50th quartile or percentile), one of the best known example of quantiles, provides a better summary of centrality than the mean.

Continued...

- It offers more robust and complete estimates compared to the mean regression, when the normality assumption is violated or outliers and long tails exist.

- These make QR attractive and are applicable to different types of data.

- Quantiles provides more information and describe the distribution of the dependent variable.

- Quantiles and percentiles are synonymous such that the 0.99 quantile=99th percentile.

- The median=0.50 quantile, quartile, quintile or percentile.

Example :
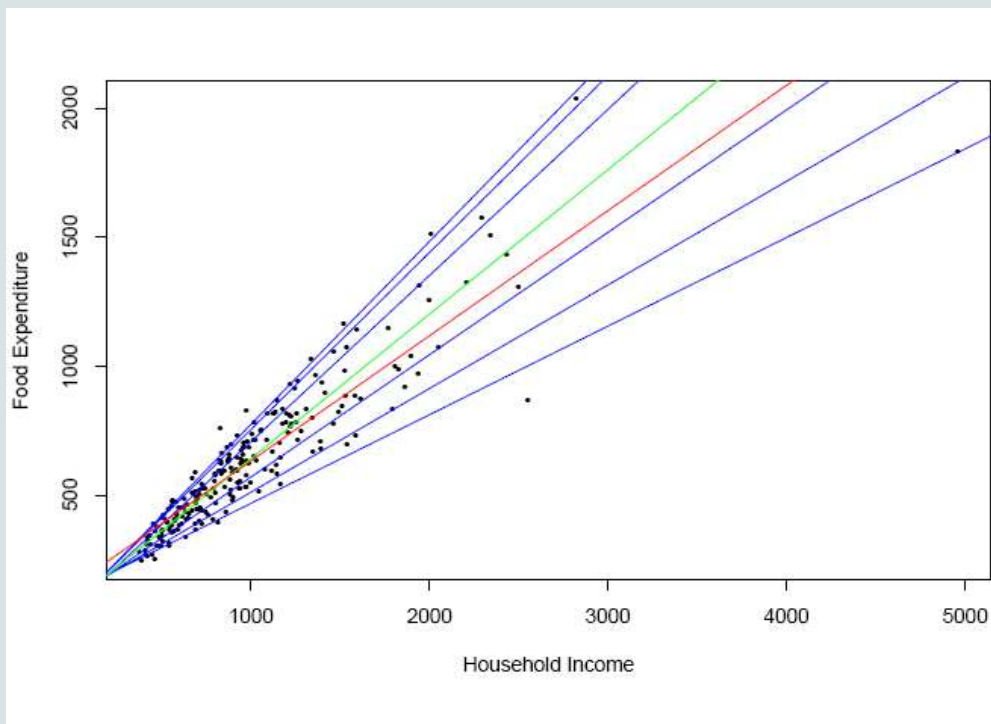
we briefly consider a classical empirical application in economics, Engel's (1857) analysis of the relationship between household food expenditure and household income. In Figure, we plot Engel's data taken from 235 European working-class households. Superimposed on the plot are seven estimated quantile regression lines corresponding to the quantiles {0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95}. The median τ =0.5 fit is indicated by the Green line; the least squares estimate of the conditional mean function is plotted as the Red line.

Significance of Quantile Regression in Real World Scenarios

Here are some of the significant applications of quantile regression in real-world scenarios:

1. <u>Financial Risk Management:</u> Quantile regression is widely used in finance to estimate Value at Risk (VaR) and Expected Shortfall (ES) of different assets or portfolios. These measures are essential for financial risk management and help in estimating the risk of large losses in investment portfolios.

2. <u>Environmental Science:</u> Quantile regression is used in environmental science to estimate the impact of climate change on different quantiles of ecological responses. It helps in understanding the effects of climate variables, such as temperature or precipitation, on the distribution of species abundance, and how this distribution may change in the future.

Continued…

3. <u>Healthcare:</u> Quantile regression is used in healthcare to estimate the treatment effect of drugs on different quantiles of patient outcomes. This helps in identifying patients who are likely to benefit more from a specific treatment and in personalizing treatment decisions.

4. <u>Economics:</u> Quantile regression is used in economics to analyze the relationship between economic variables and different quantiles of income or wealth distributions. This helps in understanding the distributional effects of different policies on income or wealth and identifying the groups of individuals who may be more or less affected by these policies.

Continued...

5. <u>Education:</u> Quantile regression is used in education to estimate the impact of different factors, such as family income or parental education, on different quantiles of student achievement. This helps in identifying the students who are likely to benefit more from specific interventions and in improving the efficiency of education policies.

Overall, quantile regression is a versatile statistical technique that has significant applications in many real-world scenarios, where traditional regression analysis may not be sufficient. It helps in understanding the distributional effects of different factors on the response variable, which is essential for making informed decisions in various fields.

Hands on in Python

Let's Use A Data Example:

  I will use the **<u>Boston house price dataset</u>**. This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston, Massachusetts. It has been used extensively throughout the literature to benchmark algorithms. It has 506 rows and 14 columns. This popular dataset has many sources and is even included in the **<u>scikit-learn datasets</u>** for practice purposes. The goal is to predict the median house value (MEDV) with the following variables.

- MEDV: (Target) median value of owner-occupied homes in \$1000s.

- CRIM: per capita crime rate by town.

- ZN: proportion of residential land zoned for lots over 25,000 sq. ft.

- INDUS: proportion of non-retail business acres per town.

- CHAS: Charles River dummy variable (= 1 if tract bounds river).

Continued...

- NOX: nitrogen oxides concentration (parts per 10 million).

- RM: average number of rooms per dwelling.

- AGE: proportion of owner-occupied units built before 1940.

- DIS: weighted mean of distances to five Boston employment centers.

- RAD: index of accessibility to radial highways.

- TAX: full-value property-tax rate per \$10,000.

- PTRATIO: pupil-teacher ratio by town.

- BLACK: $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town.

- LSTAT: lower status of the population (percent).

How to fit QR model in python is given in following link:

https://github.com/Diksha289/Quantile-Regression

Summary

Quantile regression is a statistical technique used to model the relationship between a response variable and one or more predictor variables. Unlike traditional regression, it models the conditional distribution of the response variable, allowing for greater flexibility and robustness. Quantile regression has several advantages over traditional regression, including the ability to capture the shape of the distribution and handle outliers more effectively. It is also useful when dealing with heteroscedastic data. The technique can be used in a wide range of fields, including finance, economics, healthcare, and social sciences, and has various advanced applications, such as mixed effects models, penalized quantile regression, and generalized linear models. However, like any statistical technique, quantile regression has assumptions and limitations that must be considered when interpreting the results. Overall, quantile regression is a powerful tool for modeling complex data and has numerous practical applications.

References

- http://www.econ.uiuc.edu/~roger/research/intro/jep.pdf

- https://en.wikipedia.org/wiki/Quantile_regression

- https://medium.com/

- https://chat.openai.com

- https://github.com/

- https://www.youtube.com/

Thank you!