MAT555E: Statistical Data Analysis for Computational Sciences

Spring24-Presentation: Quantile Regression

BERİA AYŞENUR CAN

090200705

ISTANBUL TECHNICAL UNIVERSITY

Learning Objectives

1. **Quantile Regression**
   a. What are Quantiles? Some Definitions
   b. Why Use Quantile Analysis? Advantages of Quantile Regression.
   c. QR Model Specification
   d. QR Objective Function
   e. The differences between linear regression and quantile regression.
   f. The limitations and alternatives of quantile regression models.

2. **References**

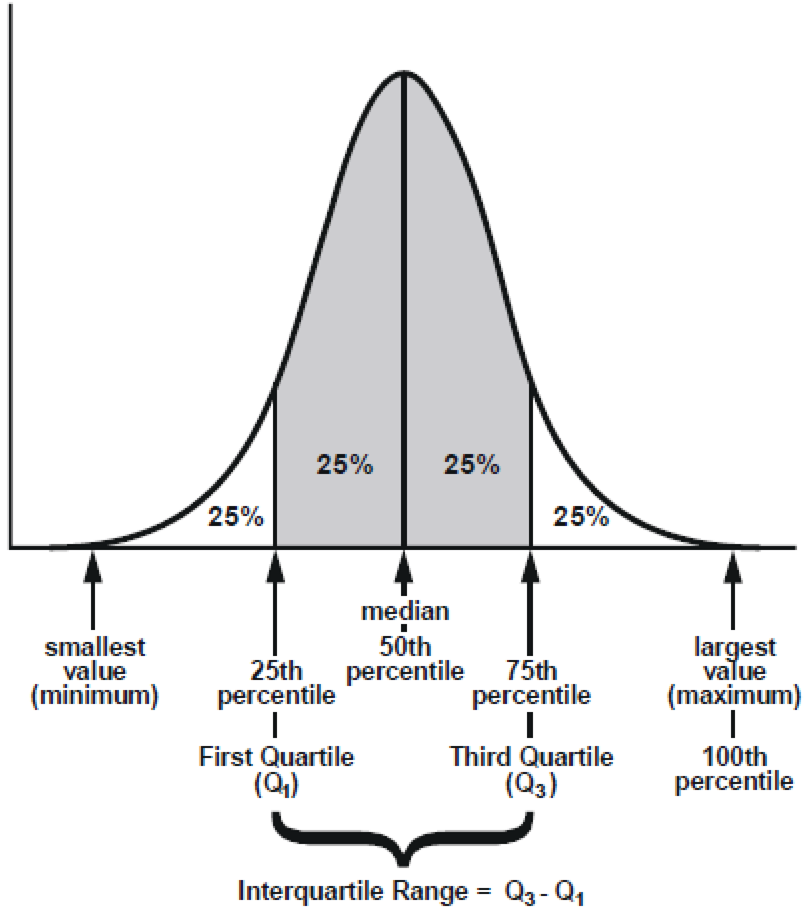
a. What are Quantiles?

• Quantiles are the inverse of CDF (Cumulative Distribution Function), mathematically they are defined as;

$$q_\alpha(Y) = F_Y^{-1}(\alpha) = \inf\{y : F(y) \geq \alpha\}$$

- A quantile defines a particular part of data set.
- We use quantiles to describe the distribution of dependent variable.
- Quantiles and percentiles are synonyms –the 0.99 quantile is 99th percentile.
- The best-known quantile is the median. The median is the 0.50 quantile.
- Recall that ordinary least squares (OLS) models examine the relationship between the independent variables x and the conditional mean of a dependent variable y.
- Remember this interpretation "a percentage change in x causes y to change by %k on average ."
- But quantile regression (QR) models examine the relationship between x and the conditional quantiles of y rather than just the conditional mean of y.

a. What are Quantiles?



$$Q_{0.5}(Y) = Median = \textit{Half the values will be below it} \quad (1)$$

$$Q_{0.25}(Y) = 1^{st} quartile = \textit{1/4 of the values will be below it} \quad (2)$$

$$Q_{0.75}(Y) = 3^{rd} quartile = \textit{3/4 of the values will be below it} \quad (3)$$

- 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 modeling than traditional mean regression.

b. Why Use Quantile Analysis? Advantages of Quantile Regression.

- Robustness to Outliers:

Quantile regression **reduces the excessive influence of outliers on the predictions**. The model is less affected by extreme values in the dataset, leading to more robust and reliable estimates.

- Modeling Relationships at Different Points in the Distribution:

Quantile regression can examine relationships at various points in the conditional distribution of the dependent variable.

This provides researchers with the opportunity to explore not only the median (0.5 quantile) but also other quantiles (for example, 0.25 or 0.75).

- Handling Heteroscedasticity:

The classical linear regression model (OLS) is based on the assumption that the error terms are homoscedastic (equal variance);

That is, the error terms are expected to have the same variance for the values of all independent variables. However, in real-world data this assumption does not always hold.

Quantile regression lends itself naturally to heteroscedasticity because it estimates a separate regression line for each quantile, and the slope of each line reflects the relationship in that part of the distribution of the independent variable.

b. Why Use Quantile Analysis? Advantages of Quantile Regression.

- Focusing on the Tails of the Distribution:

The OLS (Ordinary Least Squares) method focuses on modeling the average relationship between the independent variables and the dependent variable, using all data points.
However, OLS ignores behavior at the extremes of the distribution and cannot capture specific dynamics at these segments.

Quantile regression can examine relationships at these extremes by focusing on a specific quantile (for example, 10%, 90%). **This method reveals in detail how the effect of independent variables on the dependent variable may differ in the lower and upper tails of the data distribution.** Thus, the heterogeneity in distribution can be better understood.

c. QR Model Specification

In quantile regression, we aim to model the relationship between independent variables and the dependent variable at a specific α quantile.

The general formula is as follows:

$$q_\alpha(Y) = \beta_0(\alpha) + \beta_1(\alpha)x_1 + \beta_2(\alpha)x_2 + \ldots + \beta_k(\alpha)x_k$$

Where:

- $q_\alpha(Y)$ is the α-quantile of Y,
- $x_1, x_2, \ldots, x_k$ are the independent variables,
- $\beta_0(\alpha)$ is the quantile-specific intercept term,
- $\beta_1(\alpha), \beta_2(\alpha), \ldots, \beta_k(\alpha)$ are the quantile-specific coefficients of the independent variables, measuring the effect of each x_i variable on the changes in the α-quantile of Y.

Quantile regression can have different β coefficients for different α values, allowing for different regression lines for different quantiles.

d. QR Objective Function

In quantile regression, the objective function used is the quantile loss function.

The quantile loss function is different from the squared loss function used in OLS regression.

The quantile loss function is used to calculate the total loss by specifically weighing the differences between the quantile values predicted by the regression model at each point and the actual values:

$$\text{Minimize} \sum_{i=1}^{n} \rho_\tau(y_i - \hat{q}_i(\tau))$$

$$\rho_\tau(u) = u \cdot (\tau - I(u < 0))$$

d. QR Objective Function

The quantile loss function ρ_τ is defined for the difference u (the residual) between the predicted quantile and the actual value as:

$$\rho_\tau(u) = u \cdot (\tau - I(u < 0)) \tag{15}$$

where:

- τ represents the quantile level (e.g., 0.25, 0.50, 0.75).
- I is the indicator function, which takes the value 1 when u is negative and 0 otherwise.
- u is the difference between the predicted quantile and the actual value and is computed as $y_i - \hat{q}_i(\tau)$.

d. QR Objective Function

How the Quantile Loss Function Works:

- If the predicted value is less than the actual value (u is positive), the function weights the loss by τ. This situation implies the model places more weight on errors below the τ quantile level.
- If the predicted value is greater than the actual value (u is negative), the function weights the loss by $1 - \tau$. This implies the model places more weight on errors above the τ quantile level.

This weighting allows the model to optimize the accuracy of predictions specifically for the chosen quantile level. For example, when $\tau = 0.5$, both positive and negative errors are equally weighted, known as median regression. For other values like $\tau = 0.25$ or $\tau = 0.75$, the model optimizes for those specific quantiles by weighting their respective errors less.

e. The differences between linear regression and quantile regression.

	Linear Regression	Quantile Regression
Model	Mean $= E[Y] = \beta_0 + x\beta_1$	$q_\alpha(Y) = \beta_0 + x\beta_1$ e.g., Median $= q_{0.5}(Y) = \beta_0 + x\beta_1$ 1st Quartile $= q_{0.25}(Y) = \beta_0 + x\beta_1$
Objective	Squared Error / Log-Likelihood	Quantile Loss
Solve by	Ordinary Least Squares (OLS) / Maximum Likelihood (ML) *[Closed form solution]*	Linear Programming methods (Simplex, IPM) Iterated Reweighted Least Squares (IRLS) Gradient based methods *[Optimization, iterative methods]*

f. The advantages and limitations of quantile regression models.

Limitations of Quantile Regression:

- The computation of quantile regression **can be more complex and time-consuming than OLS**, especially for large datasets, because it typically requires linear programming to solve.

- The coefficients in quantile regression can be **more difficult to interpret** than OLS coefficients, particularly in the presence of heterogeneous effects.

- Standard errors and confidence intervals for quantile regression are not as straightforward to compute as in OLS. While bootstrap methods can be used, they add computational burden.

- Ideally, quantile regression lines for different quantiles should not cross each other; however, the basic quantile regression algorithm does not guarantee this. Advanced methods are required to enforce non-crossing conditions.

- Quantile regression, particularly for extreme quantiles, may not perform well for extrapolation outside the range of the data.

f. The advantages and limitations of quantile regression models.

Alternatives to Quantile Regression:

- **Robust Regression:**

To mitigate the influence of outliers, robust regression methods such as M-estimators can be employed. These methods provide more reliable estimates by reducing the excessive influence of outliers. If a dataset contains a substantial number of outliers that are suspected to distort the predictions of the regression model, robust regression may be the method of choice.

- **Transformation Models:**

When the dependent variable or error terms do not follow a normal distribution, applying an appropriate transformation (log, square root, etc.) before employing OLS can be effective. This approach helps to meet OLS assumptions such as homoscedasticity and normal distribution of error terms. Transformations can enhance the interpretability and statistical properties of the model.

References:

https://scikit-learn.org/stable/auto_examples/linear_model/plot_quantile_regression.html

https://www.statsmodels.org/devel/examples/notebooks/generated/quantile_regression.html

Gilchrist, W. (2000). *Statistical Modelling with Quantile Functions*. ISBN 1-58488-174-7.

https://sites.google.com/site/econometricsacademy/econometrics-models/quantile-regression