A Research Project Report On

Quantile Regression
A Robust Alternative to OLS

As a part of the requirements for the 3rd semester of the

M.Sc. in Statistics Program

DEPARTMENT OF STATISTICS

SCHOOL OF MATHEMATICAL SCIENCES

KAVAYITRI BAHINABAI CHAUDHARI NORTH MAHARASHTRA UNIVERSITY

Submitted By:

Ms. Devyani Bhosle *(Seat No. 366340)*

Mr. Harshal Tilawane *(Seat No. 366366)*

Ms. Karishma Bhamare *(Seat No. 366339)*

Under the Guidance of:

Prof. K. K. Kamalja

(2024 - 25)

CERTIFICATE

This is to certify that **Miss. Devyani Bhosle, Mr. Harshal Tilawane, Miss. Karishma Bhamare** students of M.Sc. (Statistics) with specialization in Industrial Statistics,at Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon, have successfully completed their research project work entitled **Quantile Regression: A Robust Alternative to OLS** as a part of M.Sc. (Statistics) program under my guidance and supervision during the academic year 2024-25. Their dedication, effort, and in-depth analysis contributed significantly to the project, showcasing their aptitude for statistical research.

(Prof. K.K. Kamalja)
KBCNMU, Jalgaon
Project Guide

Acknowledgement

We would like to express our heartfelt gratitude to Professor K. K. Kamalja, Department of Statistics, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon, for her invaluable guidance, expertise, and unwavering support throughout the duration of this project. Her insightful feedback, constructive suggestions, and meticulous attention to detail played a pivotal role in shaping the quality of this work. Her encouragement and mentorship have not only enhanced our knowledge but also inspired us to strive for excellence.

We also extend our appreciation to the faculty and staff of the Department of Statistics for creating a conducive academic environment and providing resources that facilitated the successful completion of this project.

Finally, we are deeply thankful to all those who, directly or indirectly, supported and encouraged us during this endeavor. Your contributions and encouragement are greatly valued and deeply appreciated..

Contents

Chapter 1

Introduction

Regression analysis aims to uncover the relationship between a response variable and predictor variables. In practical scenarios, the response variable cannot be precisely predicted from the predictors because it behaves as a random variable for any given set of predictor values. Therefore, we often describe the response variable's behavior using measures of central tendency. Common measures include the mean (average value), the median (middle value), and the mode (most frequent value). These summaries help us understand the typical behavior of the response variable given specific values of the predictors.

1.1 Conditional mean model and it's limitations

Traditional regression analysis centers on the mean, summarizing the relationship between the response variable and predictor variables by describing the average response for each fixed set of predictor values. This average response is known as the conditional mean. Modeling and fitting this conditional mean is fundamental to various regression approaches, including simple linear regression, multiple regression, weighted least squares for heteroscedastic errors, and nonlinear regression models.

The conditional-mean framework has certain limitations. It focuses on the central location of the response variable for fixed predictor values, making it difficult to extend to noncentral locations, which are often of interest in social-science research. For example, studies on economic inequality and mobility are concerned with the lower and upper tails of the distribution, such as the poor and the rich. Similarly, educational researchers

aim to understand and address gaps at specific achievement levels, like basic, proficient, and advanced. By concentrating on the mean, researchers may overlook or inadequately address questions about these noncentral locations, leading to inefficient or irrelevant Conclusionss for their studies.

Additionally, model assumptions often do not hold true in real-world scenarios. Specifically, the assumption of homoscedasticity (constant variance of errors) frequently fails. By concentrating only on central tendencies, important trends within the response distribution can be missed. Social phenomena often exhibit heavy-tailed distributions, resulting in numerous outliers. In such cases, the conditional mean becomes a misleading measure of central tendency, as it is significantly affected by these outliers.

Conditional-mean models are inadequate for capturing the full relationship between a response distribution and predictor variables. Issues like economic inequality (wages, income, wealth), educational inequality (academic achievement), and health inequality (height, weight, disease incidence, drug addiction, treatment, life expectancy) are often analyzed using these models. However, focusing solely on the mean overlooks other significant distributional properties that are more relevant to understanding these inequalities.

1.2 Quantile Regression

An alternative to conditional-mean modeling, known as conditional-median modeling or median regression, has origins dating back to the mid-18th century. This approach addresses issues related to choosing a measure of central tendency by replacing least-squares estimation with least-absolute-distance estimation. Although the least-squares method is straightforward and requires minimal computing power, least-absolute-distance estimation is more computationally demanding. It wasn't until the late 1970s, with advancements in computing technology and algorithms like linear programming, that median regression using least-absolute-distance estimation became feasible.

The median is a specific quantile representing the central location of a distribution. Conditional-median regression, which models the conditional 0.5th quantile as a function of covariates, is a particular form of quantile regression. Quantile regression, however, can model other quantiles to describe different positions within a distribution. Quantiles encompass terms like quartile, quintile, decile, and percentile, with the τ-th quantile rep-

resenting the value below which τ proportion of the population lies. For instance, the 0.025th quantile indicates the point below which 2.5% of the population falls. Koenker and Bassett [16] introduced quantile regression in 1978, providing a method to model conditional quantiles as functions of predictors. This approach extends the linear regression model to offer a more comprehensive analysis of the data distribution.

The linear regression model predicts how the average value of the dependent variable changes with changes in the independent variables. In contrast, the quantile regression model predicts how specific percentiles (or quantiles) of the dependent variable change with changes in the independent variables. This allows for modeling various points in the distribution, not just the average, such as the median or any other percentile of interest.

Chapter 2

Quantiles and Quantile functions

Quantiles are points that divide a dataset into equal-sized intervals, providing a way to understand the spread and distribution of the data. The quantile function is a mathematical tool that links probabilities to their corresponding quantile values, offering insights into the distribution's characteristics.

2.1 Quantiles and CDF

Quantiles are values that divide a dataset into intervals with equal probabilities. For a given quantile level τ (where $0 < \tau < 1$), the quantile ξ_τ is the value below which a fraction τ of the data falls.

Cumulative Distribution Function (CDF) is a function $F(x)$ that gives the probability that a random variable X is less than or equal to a given value x:

$$F_X(x) = P(X \leq x)$$

Quantiles as Inverse CDF

Quantiles can be expressed in terms of the inverse of the CDF. The τ-th quantile, ξ_τ, is the value of x such that the CDF $F_X(x)$ equals τ:

$$\xi_\tau = F_X^{-1}(\tau)$$

In other words, the τ-th quantile ξ_τ is the value at which the CDF reaches τ:

$$\tau = F_X(\xi_\tau)$$

To summarize:

- The **CDF** $F_X(x)$ gives the probability that X is less than or equal to x.
- The **inverse CDF** (or quantile function) $F_X^{-1}(\tau)$ gives the value x such that the probability of X being less than or equal to x is τ.

Thus, quantiles can be derived by applying the inverse CDF to a given quantile level τ.

2.2 Properties of Quantiles

Quantiles are essential statistical tools used to divide a dataset into equal-sized intervals, providing a clear understanding of the distribution's structure. They are particularly valuable due to their robustness against outliers and their invariance to the order and scaling of data. Quantiles are bounded by the minimum and maximum values of a dataset, and their corresponding quantile function, which maps probabilities to these values, is inherently monotonic. These properties make quantiles a powerful tool for analyzing and interpreting data distributions in various fields.

2.2.1 Monotone Equivariance Property of Quantiles

The monotone equivariance property of quantiles states that if a monotonic (i.e., non-decreasing or non-increasing) transformation is applied to a dataset, the quantiles of the transformed dataset will be the same transformation applied to the quantiles of the original dataset.

In more formal terms, let f be a monotonic function and X be a random variable with quantile $\xi_\tau(X)$. Then the quantile of the transformed variable $f(X)$, denoted as $\xi_\tau(f(X))$, is given by:

$$\xi_\tau(f(X)) = f(\xi_\tau(X))$$

Suppose you have a dataset X and you apply a monotonic function f (such as $f(x) = x + c$ or $f(x) = ax$ with $a > 0$). If ξ_τ is the τ-th quantile of X, then $f(\xi_\tau)$ will be the τ-th quantile of the transformed dataset $f(X)$.

- If the τ-th quantile of X is ξ_τ, then the τ-th quantile of $X + c$ is $\xi_\tau + c$.
- If the τ-th quantile of X is ξ_τ, then the τ-th quantile of aX is $a\xi_\tau$.

This property holds because the order of the data points is preserved under monotonic transformations, ensuring that the quantiles transform accordingly.

2.2.2 Robustness Property of Quantiles

The robustness property of quantiles refers to their resilience to the influence of outliers or extreme values in a dataset. Unlike the mean, which can be significantly affected by outliers, quantiles provide a measure of central tendency or spread that is less sensitive to extreme values.

- **Insensitivity to Outliers**: Quantiles, such as the median (50th percentile) and other percentiles, are less affected by outliers or extreme values compared to the mean. This is because quantiles are based on the relative positions of data points rather than their magnitudes. For example, the median is the middle value in a sorted list and does not change significantly with the addition of extreme values.
- **General Quantiles**: For any quantile ξ_τ (where τ is the quantile level), the quantile value is determined by the position in the ordered dataset, not by the magnitude of individual values. Therefore, a small proportion of extreme values has a minimal effect on quantiles.
  - **Less Sensitive**: Quantiles are less sensitive to individual extreme values because they depend on the distribution of data rather than the specific values.
  - **Representative**: They provide a better summary of the central location or spread of the data in the presence of outliers.

This robustness makes quantiles valuable in statistical analysis, particularly when dealing with data that may contain outliers or is not symmetrically distributed.

2.3 Quantiles as a solution to a minimization problem

The median m has a property similar to the mean. Instead of using squared distances, we can measure how far Y is from m by the absolute distance $|Y - m|$, and the average distance in the population from m is given by the mean absolute distance $E[|Y - m|]$.

Mathematically, the median m minimizes the expected absolute deviation:

$$\frac{\partial}{\partial m} E[|Y - m|] = \frac{\partial}{\partial m} \int_{-\infty}^{\infty} |y - m| f_Y(y)\, dy = 0$$

Differentiating with respect to m and setting the partial derivative to zero will lead to the solution for the minimum,

$$F(m) = P(Y \leq m) = \frac{1}{2}$$

where $F(m)$ is the CDF of Y.

This means that among all possible values for m, the median is the one that results in the smallest average absolute deviation from Y.

This representation of the median generalizes to other quantiles as follows. For any $p \in (0, 1)$, the distance from Y to a given quantile ξ_τ is measured by the absolute distance, but we apply a different weight depending on whether Y is to the left or to the right of ξ_τ.

Thus, we define the distance from Y to a given quantile ξ_τ as:

The distance function $d_\tau(Y, \xi_\tau)$ for a given quantile ξ_τ and probability level τ is defined as follows:

$$d_\tau(Y, \xi_\tau) = \begin{cases} (1 - \tau) \cdot |Y - \xi_\tau| & \text{if } Y < \xi_\tau \\ \tau \cdot |Y - \xi_\tau| & \text{if } Y \geq \xi_\tau \end{cases}$$

where ξ_τ is the quantile at level τ, and $|Y - \xi_\tau|$ denotes the absolute distance between Y and ξ_τ.

This function applies a different weight to the absolute deviation depending on whether

Y is less than or greater than or equal to the quantile ξ_τ.

where ξ_τ is the τ-th quantile, which is the value such that:

$$P(Y \leq \xi_\tau) = \tau$$

This generalization reflects that different weights are applied based on whether Y is below or above the quantile ξ_τ, thus allowing for the definition of any quantile ξ_τ in terms of minimizing the weighted absolute deviations.

To make things simple, we assume that the cdf F has a probability density function f.

$$E[d_\tau(Y, \xi_\tau)] = \int_{-\infty}^{\xi_\tau} (1-\tau) \cdot |y - \xi_\tau| \cdot f(y)\, dy + \int_{\xi_\tau}^{\infty} \tau \cdot |y - \xi_\tau| \cdot f(y)\, dy$$

$$= \int_{-\infty}^{\xi_\tau} (1-\tau) \cdot (\xi_\tau - y) \cdot f(y)\, dy + \int_{\xi_\tau}^{\infty} \tau \cdot (y - \xi_\tau) \cdot f(y)\, dy$$

above equation is a convex function. Differentiating with respect to ξ_τ and setting the partial derivative to zero will lead to the solution for the minimum.

$$\frac{\partial}{\partial \xi_\tau} E[d_\tau(Y, \xi_\tau)] = \frac{\partial}{\partial \xi_\tau} \int_{-\infty}^{\xi_\tau} (1-\tau) \cdot (\xi_\tau - y) \cdot f(y)\, dy + \frac{\partial}{\partial \xi_\tau} \int_{\xi_\tau}^{\infty} \tau \cdot (y - \xi_\tau) \cdot f(y)\, dy$$

The partial derivative of the first term is:

$$\frac{\partial}{\partial \xi_\tau} \int_{-\infty}^{\xi_\tau} (1-\tau) \cdot (\xi_\tau - y) \cdot f(y)\, dy$$

$$= (1-\tau) \cdot (\xi_\tau - y) \cdot f(y)\big|_{y=\xi_\tau} + \int_{-\infty}^{\xi_\tau} \frac{\partial}{\partial \xi_\tau} (1-\tau) \cdot (\xi_\tau - y) \cdot f(y)\, dy$$

$$= (1-\tau) \cdot F(\xi_\tau)$$

The partial derivative of the second term is:

$$\frac{\partial}{\partial \xi_\tau} \int_{\xi_\tau}^{\infty} \tau \cdot (y - \xi_\tau) \cdot f(y)\, dy$$
$$= \tau \cdot (y - \xi_\tau) \cdot f(y)\big|_{y=\xi_\tau} + \int_{\xi_\tau}^{\infty} \frac{\partial}{\partial \xi_\tau} \tau \cdot (y - \xi_\tau) \cdot f(y)\, dy$$

$$= (-\tau) \cdot (1 - F(\xi_\tau))$$

Combining these two partial derivatives leads to:

$$\frac{\partial}{\partial \xi_\tau} E[d_\tau(Y, \xi_\tau)] = (1 - \tau) \cdot F(\xi_\tau) + (-\tau) \cdot (1 - F(\xi_\tau))$$

$$= F(\xi_\tau) - \tau$$

We set the partial derivative $F(\xi_\tau) - \tau = 0$ and solve for the value of $F(\xi_\tau) = \tau$ that satisfies the minimization problem.

Chapter 3

Quantile Regression - Model and Estimation

In the context of skewed distributions, the median often provides a more accurate measure of central tendency compared to the mean. Consequently, when modeling location shifts, it's advisable to use conditional-median regression rather than conditional-mean regression. Conditional-median regression, which was first proposed by Boscovich [21] in the mid-18th century and later explored by Laplace and Edgeworth [8], focuses on the conditional median of the response variable. This approach is particularly useful because it reflects the central location of the distribution even when it is skewed.

For modeling not only location shifts but also changes in the shape of the distribution, the quantile-regression model (QRM), introduced by Koenker and Bassett [15] in 1978, offers a more comprehensive method. Unlike median regression, QRM estimates how a covariate affects different quantiles across the conditional distribution. For example, QRM can estimate the impact of a covariate on a range of quantiles, from the 5th to the 95th, spaced evenly. This results in multiple regression lines—one for the median, and others for various quantiles—allowing for a detailed analysis of location shifts (captured by the median line), as well as scale and more complex changes in the distribution's shape (captured by the lines for off-median quantiles). The QRM thus provides a fuller understanding of how covariates affect the entire distribution and can handle heteroscedasticity effectively.

In the exploration of quantile regression models, we begin with simple linear quantile regression to understand the fundamental principles and methods. This approach focuses

on estimating the relationship between a single predictor variable and a response variable across different quantiles of the response distribution. By fitting linear models to specific quantiles, such as the median or other percentiles, we gain insights into how the predictor influences various points of the response distribution, beyond the mean.

Building on this foundation, we then advance to multiple quantile regression, which involves incorporating multiple predictors into the model. This extension allows for a more comprehensive analysis by examining how a set of predictor variables affects different quantiles of the response variable. It provides a richer understanding of the conditional quantile relationships in complex scenarios where multiple factors are at play.

Finally, we explore quadratic quantile regression, which extends the model to capture non-linear relationships between predictors and the response variable. By including quadratic terms, this approach accommodates scenarios where the effect of predictors changes in a non-linear manner across different quantiles. Together, these methods offer a robust framework for analyzing and interpreting data with varying relationships and distributions.

Following Koenker and Bassett [15] (1978), the simple quantile-regression model (QRM) corresponding to the simple linear regression model (LRM) can be expressed as:

$$y_i = \beta_0(\tau) + \beta_1(\tau)x_i + \varepsilon_i(\tau)$$

where:

- y_i is the response variable for observation i,
- x_i is the covariate for observation i,
- $\beta_0(\tau)$ is the intercept of the regression model at quantile τ,
- $\beta_1(\tau)$ is the slope of the regression model at quantile τ,
- $\varepsilon_i(\tau)$ is the error term for observation i at quantile τ, representing deviations from the quantile regression line.

For $0 < \tau < 1$, the proportion of the population with scores below the quantile at τ is indicated.

3.1 Assumptions of Quantile Regression

- **Quantile-Specific Error Term:** The distribution of the error term ε is such that the τ-th conditional quantile of ε given x is zero:

$$Q_{\varepsilon}(\tau \mid x) = 0$$

  This means that the τ-th quantile of the error term does not depend on x and is zero.

- **Conditional Quantiles of y:** For any given x, the conditional τ-th quantile of y, $Q(\tau)(y \mid x)$, is well-defined. This assumes that the quantiles of y given x exist and are finite.

- **Distribution of Errors:** The error distribution should allow for quantile estimation. Although QR does not assume a specific distribution for the errors, it generally works best when the errors have some form of continuous distribution.

3.2 Simple Quantile Regression Model

Recall that for the Simple Linear Regression Model (LRM), the conditional mean of y_i given x_i is:

$$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i$$

This is equivalent to requiring that the error term ε_i has zero expectation. In contrast, for the corresponding Quantile Regression Model (QRM), we specify that the τ-th conditional quantile given x_i is:

$$Q_{\tau}(y_i \mid x_i) = \beta_0(\tau) + \beta_1(\tau) x_i$$

Thus, the conditional τ-th quantile is determined by the quantile-specific parameters $\beta_0(\tau)$ and $\beta_1(\tau)$, and a specific value of the covariate x_i.

Just like the LRM, the QRM can be formulated equivalently with a statement about the error terms ε_i. Since $\beta_0(\tau) + \beta_1(\tau) x_i$ is a constant, we have:

$$Q_\tau(y_i \mid x_i) = \beta_0(\tau) + \beta_1(\tau)x_i + Q(\tau)(\varepsilon_i)$$

Thus:

$$\boldsymbol{Q_\tau(y_i \mid x_i) = \beta_0(\tau) + \beta_1(\tau)x_i}$$

So an equivalent formulation of the QRM requires that the $\boldsymbol{p}$-th quantile of the error term be zero:

$$Q_{(\tau)}(\varepsilon_i) = 0$$

3.3 Parameter Estimation in QRM

In quantile regression modeling, parameter estimation focuses on determining the coefficients that describe the relationship between predictor variables and the conditional quantiles of the response variable. Unlike traditional regression methods that estimate the conditional mean, quantile regression provides a more comprehensive view by estimating parameters for various quantiles, such as the median or other percentiles. The estimation process involves minimizing a weighted sum of absolute residuals, using the check loss function to handle different quantiles. This approach allows for capturing the effects of predictors across the entire distribution of the response variable, offering insights into how relationships change at different points in the response distribution.

3.3.1 Simple Quantile Regression Model

For a single predictor variable X and a response variable Y, the quantile regression model is:

$$Q_\tau(y_i \mid x_i) = \beta_0(\tau) + \beta_1(\tau)x_i$$

where:

- $Q_\tau(y_i \mid x_i)$ is the τ-th quantile of y_i given x_i.
- $\beta_0(\tau)$ is the intercept.
- $\beta_1(\tau)$ is the slope coefficient for x_i.

3.3.2 Check Function

The check function used in quantile regression is:

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

where:

- $u = y_i - (\beta_0(\tau) + \beta_1(\tau)x_i)$ is the residual for the i-th observation.
- τ is the quantile level (e.g., 0.25, 0.5, 0.75).
- $I(u < 0)$ is an indicator function that equals 1 if $u < 0$ and 0 otherwise.

3.3.3 Loss Function

The goal is to minimize the sum of the check function values over all observations:

$$\text{Minimize } \phi = \sum_{i=1}^{n} \rho_\tau(y_i - (\beta_0(\tau) + \beta_1(\tau)x_i))$$

3.3.4 Minimization Problem

To obtain the estimates of parameters $\hat{\beta}_0(\tau)$ and $\hat{\beta}_1(\tau)$, solve:

$$\hat{\beta}_0(\tau) = \text{argmin}_{\beta_0(\tau)} \sum_{i=1}^{n} \rho_\tau(y_i - (\beta_0(\tau) + \beta_1(\tau)x_i))$$

$$\&$$

$$\hat{\beta}_1(\tau) = \text{argmin}_{\beta_1(\tau)} \sum_{i=1}^{n} \rho_\tau(y_i - (\beta_0(\tau) + \beta_1(\tau)x_i))$$

The minimization problem in quantile regression can be framed as a linear programming problem. Linear programming is a method used to achieve the best outcome (such as minimizing a loss function) in a mathematical model whose requirements are represented by linear relationships.

However, solving a linear programming problem manually is extremely difficult and impractical, especially as the number of data points and the complexity of the model increases. Linear programming problems require specialized algorithms and computational power, which makes manual calculations infeasible for all but the simplest cases.

Because of the complexity of the optimization process in quantile regression, we rely on software to perform these calculations. Software tools like R, Python, Stata, and others are equipped with algorithms that can efficiently solve the linear programming problems inherent in quantile regression.

These tools use sophisticated methods to find the parameter estimates that minimize the loss function, handling the non-smooth nature of the problem and the large number of potential calculations required. By using software, you can quickly and accurately estimate the parameters for different quantiles, allowing for more detailed and nuanced analysis compared to traditional regression methods.

3.3.5 Steps to fit QRM at $\tau = 0.25$

Process of estimating the 0.25th quantile ($\tau = 0.25$) of a dependent variable using quantile regression:

- **Proposed Model**

$$Q_{(0.25)}(Y \mid X) = \beta_0(0.25) + \beta_1(0.25)X$$

- **Residuals**

$$u_i = y_i - (\hat{\beta}_0(0.25) + \hat{\beta}_1(0.25)x_i)$$

- **Check Function**

$$\rho_{0.25}(u) = u \times (0.25 - I(u < 0))$$

- **Objective**

$$\text{Minimize} \sum_{i=1}^{n} \rho_{0.25}(y_i - (\hat{\beta}_0(0.25) + \hat{\beta}_1(0.25)x_i))$$

- **Estimation** Solve for $\hat{\beta}_0^{(0.25)}$ and $\hat{\beta}_1^{(0.25)}$ to minimize the sum of the check function values.

- **Prediction**

$$\hat{y}_i = \hat{\beta}_0^{(0.25)} + \hat{\beta}_1^{(0.25)}x_i$$

3.4 Multiple Quantile Regression

In Multiple Quantile Regression, the model extends the concept of quantile regression to incorporate multiple predictor variables. The goal is to model the conditional quantiles of the response variable Y given a set of predictor variables. The linear form of the multiple quantile regression model can be expressed as:

$$\boldsymbol{Q_\tau(Y \mid X) = \beta_0(\tau) + \beta_1(\tau)X_1 + \beta_2(\tau)X_2 + \cdots + \beta_p(\tau)X_p}$$

where:

- $Q_\tau(Y \mid X)$ represents the τ-th quantile of the response variable Y given the predictors $X = (X_1, X_2, \ldots, X_p)$.
- $\beta_0(\tau)$ is the intercept for the τ-th quantile.
- $\beta_1(\tau), \beta_2(\tau), \ldots, \beta_p(\tau)$ are the coefficients for the predictors $X_1, X_2, \ldots, X_p$ at the τ-th quantile.

In matrix notation, the multiple quantile regression model is formulated as:

$$\boldsymbol{Q_\tau(Y \mid X) = X\beta(\tau)}$$

where:

- Y is an $n \times 1$ vector of response variables.
- X is an $n \times (p+1)$ design matrix that includes n observations and $p+1$ predictors (including a column of ones for the intercept).
- $\beta(\tau)$ is a $(p+1) \times 1$ vector of coefficients corresponding to the τ-th quantile.
- $Q_\tau(Y \mid X)$ denotes the τ-th quantile of Y given X.

The parameter estimation involves minimizing the weighted sum of absolute residuals using the loss function:

$$\rho_\tau(u) = \sum_{i=1}^{n} [u_i(\tau - I(u_i < 0))]$$

where:

- $u = Y - X\beta(\tau)$ is the vector of residuals.
- $I(u_i < 0)$ is an indicator function that equals 1 if $u_i < 0$ and 0 otherwise.

3.5 Simple Quadratic Quantile Regression Model

In the simple quadratic quantile regression model, the relationship between the response variable Y and a single predictor variable X is extended to include a quadratic term. The model is specified as:

$$\boldsymbol{Q_\tau(Y \mid X) = \beta_0(\tau) + \beta_1(\tau)X + \beta_2(\tau)X^2}$$

where:

- $Q_\tau(Y \mid X)$ denotes the τ-th quantile of Y given X.
- $\beta_0(\tau)$ is the intercept for the τ-th quantile.
- $\beta_1(\tau)$ is the coefficient for the linear term X.
- $\beta_2(\tau)$ is the coefficient for the quadratic term X^2.

The parameters $\beta_0(\tau)$, $\beta_1(\tau)$, and $\beta_2(\tau)$ are estimated by minimizing the weighted sum of absolute residuals using the loss function:

$$\rho_\tau(u) = \sum_{i=1}^{n} [u_i(\tau - I(u_i < 0))]$$

where:

- $u = Y - (\beta_0(\tau) + \beta_1(\tau)X_i + \beta_2(\tau)X_i^2)$ is the vector of residuals.
- $I(u_i < 0)$ is an indicator function that equals 1 if $u_i < 0$ and 0 otherwise.

Chapter 4

Accuracy Measures for Quantile Regression Model

4.1 Pseudo R^2

In quantile regression models, which differ from linear regression models by focusing on minimizing a sum of weighted distances based on whether observations are above or below the predicted values, measuring goodness of fit requires a different approach than the traditional R^2 statistic. Koenker and Machado [17] (1999) propose a method for assessing fit by comparing the sum of weighted distances for the full quantile regression model, $V^1(\tau)$, with the sum of distances for a model that includes only an intercept term, $V^0(\tau)$. This comparison helps gauge how well the model with predictors performs relative to a baseline model that does not include any predictors.

For example, using the one-covariate model, we have

$$V^1(\tau) = \sum_{i=1}^{n} d_\tau(y_i, \hat{y}_i)$$

$$V^1(\tau) = \sum_{y_i \geq \hat{\beta}_0(\tau)+\hat{\beta}_1(\tau)x_i} \tau|y_i - \hat{\beta}_0(\tau) - \hat{\beta}_1(\tau)x_i| + \sum_{y_i < \beta_0(\tau)+\beta_1(\tau)} (1-\tau)|y_i - \hat{\beta}_0(\tau) - \hat{\beta}_1(\tau)x_i|$$

$$V^0(\tau) = \sum_{i=1}^{n} d_\tau(y_i, \hat{Q}_{(\tau)}(y_i))$$

$$V^0(\tau) = \sum_{y_i \geq \bar{y}} \tau|y_i - \hat{Q}_{(\tau)}(y_i)| + \sum_{y_i < \bar{y}} (1-\tau)|y_i - \hat{Q}_{(\tau)}(y_i)|$$

For the model that only includes a constant term, the fitted constant is the sample τth quantile $\hat{Q}_{(\tau)}(y_i)$ for the sample $y_1, \ldots, y_n$. The goodness of fit is then defined as

$$\text{Pseudo } R^2 = 1 - \left(\frac{V^1(\tau)}{V^0(\tau)}\right)$$

Since the pseudo R^2 and adjusted R^2 are derived from different underlying concepts, making them fundamentally incomparable. Adjusted R^2, used in OLS regression, measures the proportion of variance explained by the model, adjusted for the number of predictors, and is based on the assumption of a linear relationship between the dependent and independent variables. In contrast, pseudo R^2, often used in models like logistic or quantile regression, does not measure the variance explained in the same way. Instead, it provides a goodness-of-fit measure specific to the type of model being used, often comparing the likelihood of the fitted model to a null model. Since the two metrics are based on different frameworks and assumptions, they cannot be directly compared or interpreted in the same way.

4.2 AIC in Quantile Regression

The Akaike Information Criterion (AIC) was originally introduced by Hirotugu Akaike [1] in 1974 as a measure for model selection based on the likelihood function. For ordinary least squares (OLS) regression and other models with a well-defined likelihood, the AIC is straightforward to apply.

However, for quantile regression, where the traditional likelihood function is not directly applicable, the adaptation of AIC was proposed by different researchers to address this issue. Notably, the adaptation of AIC for quantile regression was formalized by:

Roger Koenker and Kevin F. Hallock [16] in their 2001 paper: "Quantile Regression". They discussed methods to adapt AIC for quantile regression models by approximating the likelihood function with a suitable residual-based approach. In their work, they suggested that AIC for quantile regression could be approximated using the quantile loss function and residual sum, considering that the quantile regression model does not fit the traditional likelihood framework. This adaptation ensures that the AIC can still serve as a useful tool for model comparison in the context of quantile regression.

So, while the original AIC formula was given by Akaike, its adaptation for quantile

regression is attributed to Koenker and Hallock, among others who have contributed to refining the approach for different statistical models. In quantile regression, the Akaike Information Criterion (AIC) is calculated similarly to ordinary least squares (OLS) regression, but with adjustments for the nature of quantile regression.

To calculate AIC for a quantile regression model, start by fitting the quantile regression model to the data. Quantile regression estimates the conditional quantile of the response variable, and the quantile of interest (e.g., the median or 0.5 quantile) must be specified.

The number of parameters, denoted by k, includes all the coefficients and the intercept in the model. In quantile regression, the log-likelihood is based on the quantile loss function, which is different from the squared loss used in OLS regression. Specifically, the quantile loss function, known as the check function, is given by:

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

where τ represents the quantile level (e.g., 0.5 for the median), and $I(u < 0)$ is an indicator function that is 1 if $u < 0$ and 0 otherwise. The check function calculates the absolute deviations weighted by the quantile.

The pseudo-log-likelihood for quantile regression is computed as:

$$\ln(L) = -\sum_{i=1}^{n} \rho_\tau(y_i - \hat{y}_i)$$

where $\hat{y}_i$ are the predicted values from the quantile regression model, and y_i are the observed values.

Finally, the AIC is calculated using the formula:

$$\text{AIC} = 2k - 2\ln(L)$$

where k is the number of parameters in the model, and $\ln(L)$ is the pseudo-log-likelihood. Lower AIC values indicate a better balance between the model's goodness of fit and its complexity.

Chapter 5

Fitting of QRM Using Statistical Softwares

5.1 R software

R is a free, open-source software environment widely used for statistical computing and data visualization. It offers a comprehensive suite of tools for data analysis, including statistical modeling, hypothesis testing, and graphical representation. With its extensive package ecosystem, R enables advanced data manipulation, specialized analyses, and seamless integration with other tools. Its scripting capabilities facilitate reproducible research, making it a trusted choice for statisticians, data scientists, and researchers across diverse fields.

5.1.1 `quantreg` library

The `quantreg` package in R was developed by Roger Koenker to implement quantile regression, a statistical technique introduced by Koenker and Bassett in 1978. Quantile regression provides a comprehensive view of the conditional distribution of the response variable, enabling the analysis of relationships at different quantiles, not just the mean. This approach is particularly valuable for modeling data with heterogeneous variance or outliers, where traditional least squares regression may fail to capture the underlying dynamics effectively. The `quantreg` package has become a cornerstone in statistical computing for robust and flexible regression analysis.

Functions in `quantreg` library

`i) rq(formula, tau=.5, data)`

Arguments

- `formula`: a formula object, with the response on the left of a $a \sim b$ operator, and the terms, separated by + operators, on the right.
- `tau` : the quantile(s) to be estimated, this is generally a number strictly between 0 and 1, but if specified strictly outside this range, it is presumed that the solutions for all values of tau in (0,1) are desired.
- `data` : a data.frame in which to interpret the variables named in the formula, or in the subset and the weights argument. If this is missing, then the variables in the formula should be on the search list.

5.1.2 Functions in base R

`i) lm(formula, data)`

Arguments

- `formula`: an object of class 'formula' (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'. a formula object, with the response on the left of a $a \sim b$ operator, and the terms, separated by + operators, on the right.
- `data` : an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which lm is called.

The output of `lm(object)` in R provides:

- Call: Shows the formula used for the model.
- Coefficients: Displays the estimated coefficients for the intercept and predictor variable x

`ii) summary(object)`

Arguments

- `object`: In the context of regression analysis, the `summary(object)` function, where object refers to the fitted model (e.g., `summary(model)`)

The output of `summary(model)` in R provides:

- Coefficients: Estimates of the regression coefficients, their standard errors, t-values, and p-values, indicating the significance of each predictor.
- Residuals: Summary statistics of the residuals to assess the distribution of errors.
- R-squared & Adjusted R-squared: Measures of the model's explanatory power.
- F-statistic & p-value: Tests the overall significance of the model.
- Residual Standard Error: An estimate of the variability in the response variable that the model does not explain.

`iii) AIC(object)`

Arguments

- `object`: In the context of regression analysis, `AIC(object)`, where object refers to the fitted model (e.g., `AIC(object)`)

The output of `AIC(model)` provides the Akaike Information Criterion (AIC) value for the model. The AIC balances the goodness of fit and model complexity, with lower values suggesting a better model fit relative to complexity. It helps in model selection by penalizing excessive parameters to avoid overfitting.

5.1.3 Model Evaluation using the `performance` Package in R

In the evaluation of regression models, it is essential to assess various diagnostic measures to ensure the robustness and validity of the model. The `performance` package in R provides a suite of tools for this purpose. Below, we describe three key functions used in this research for model evaluation: `check_collinearity`, `check_distribution`, and `check_heteroscedasticity`.

Functions in `performance` Package

`i) check_collinearity(model)`

Purpose: This function assesses the presence of multicollinearity in the regression model by calculating the Variance Inflation Factor (VIF) for each predictor variable. High multicollinearity can inflate the standard errors of the coefficients, leading to unreliable estimates.

Syntax: `check_collinearity(model)`

Interpretation: The function returns VIF values for each predictor in the model. A VIF value greater than 10 (or even 5, depending on the field) suggests high multicollinearity, indicating that the predictor may be highly correlated with other predictors in the model.

`ii) check_distribution(model)`

Purpose: This function checks whether the residuals of the model follow a normal distribution. This is important for the validity of many inferential statistics in linear regression.

Syntax: `check_distribution(model)`

Interpretation: The function provides a graphical representation and a statistical test (such as the Shapiro-Wilk test) to evaluate the normality of residuals. Deviations from normality might suggest that the model does not adequately capture the data's distribution, potentially leading to biased estimates and incorrect Conclusionss.

`iii) check_heteroscedasticity(model)`

Purpose: This function tests for heteroscedasticity, which occurs when the variability of the residuals is not constant across all levels of the predictor variables. Heteroscedasticity violates one of the key assumptions of linear regression and can lead to inefficient estimates.

Syntax:`check_heteroscedasticity(model)`

Interpretation: The function outputs a diagnostic test result, typically using the Breusch-Pagan test. A significant p-value (typically less than 0.05) indicates the presence of heteroscedasticity, suggesting that the model's assumptions have been violated, and alternative modeling techniques or transformations might be necessary.

5.2 SAS: Statistical Analysis System

SAS (Statistical Analysis System) is a software suite used for advanced analytics, statistical modeling, business intelligence, and data management. Its programming language provides an extensive range of statistical and analytical tools, enabling users to perform tasks such as regression analysis, data visualization, and predictive modeling. With its built-in procedures and flexibility, SAS is a popular choice among statisticians and data scientists.

5.2.1 Performing Quantile Regression in SAS

Procedure: PROC QUANTREG

`PROC QUANTREG` in SAS is used for quantile regression analysis. It estimates the conditional quantiles of a response variable and is particularly useful for modeling non-constant variance or skewed distributions.

Syntax

```
PROC QUANTREG DATA=dataset;
   MODEL response = predictor1 predictor2 ... / QUANTILE=quantile_value;
RUN;
```

Arguments

- **DATA**: Specifies the dataset to be used.
- **MODEL**: Defines the regression model.
  - **response**: The dependent variable.
  - **predictor1, predictor2, ...**: Independent variables (predictors).
  - **/ QUANTILE=quantile_value**: Specifies the quantile(s) to be estimated (e.g., `QUANTILE=0.5` for median regression).

Output

- Parameter estimates for the specified quantile.

- Confidence intervals for the regression coefficients.
- Diagnostics such as goodness-of-fit statistics.

5.2.2 Performing Simple Linear Regression in SAS

Procedure: PROC REG

`PROC REG` is used for linear regression analysis, which estimates the relationship between a dependent variable and one or more predictors.

Syntax

```
PROC REG DATA=dataset;
   MODEL response = predictor;
RUN;
```

Arguments

- **DATA**: Specifies the dataset to be used.
- **MODEL**: Defines the regression model.
  - **response**: The dependent variable.
  - **predictor**: The independent variable.

Output

- Parameter estimates (slope and intercept).
- ANOVA table with F-statistic and p-value.
- R-squared value indicating the proportion of variance explained by the model.
- Diagnostic plots for residuals and fitted values (if requested).

Chapter 6

Numerical Study

This study explores the versatility of quantile regression (QR) as a robust statistical tool by applying it to diverse simulated and real-world datasets. The analysis compares QR with traditional Ordinary Least Squares (OLS) regression across various contexts, emphasizing the added insights provided by QR when assumptions such as constant error variance (homoscedasticity) are violated or when outliers and distributional nuances are present.

- **Income Determinants**: A simulation-based analysis evaluates the effects of education level and years of experience on income. While OLS captures average trends, QR reveals how predictors influence lower, median, and upper income levels, providing a nuanced understanding of subpopulations.

- **Vehicle Fuel Efficiency**: A simulated dataset demonstrates QR's ability to model heteroscedastic data. By analyzing the effects of engine size, vehicle weight, and cylinders on fuel efficiency, QR highlights variations across low, median, and high-efficiency vehicles, which are often overlooked by OLS.

- **Student Performance**: In an simulated educational dataset, QR robustly models the impact of study hours, attendance, and prior GPA on test scores, effectively handling outliers and identifying differential effects on low, median, and high performers.

- **Engel Food Expenditure Data**: Using household income and food expenditure data, QR provides a more accurate representation of spending patterns across different income levels, especially in the presence of heteroscedasticity.

- **AirQuality Dataset**: QR uncovers nuanced relationships between Ozone levels and predictors like Wind and Temperature, offering valuable insights for environmental monitoring by examining variations at different pollution levels.

- **Boston Housing Data**: A homoscedastic analysis compares QR and OLS in modeling the relationship between median home values and the average number of rooms. While OLS focuses on central tendencies, QR reveals variations across different housing market segments.

6.1 Demonstration of Quantile Regression using Simulation

Simulating different quantile regression scenarios helps to better understand how Quantile Regression (QR) performs under various conditions, particularly when assumptions like homoscedasticity (constant variance) are violated or when we are intersted in different parts of the outcome distribution. The key advantage of quantile regression is its robustness in handling these complexities, providing insights into the conditional distribution of the response variable.

In this section, we will simulate and apply quantile regression models in the following distinct scenarios:

- **Homoscedasticity**: Where the variance of errors remains constant across different values of the predictor variable.

- **Heteroscedasticity**: Where the variance of errors increases or decreases with the predictor variable, leading to varying impacts of predictors on different parts of the distribution.

Each scenario will demonstrate how quantile regression captures relationships that are missed by standard OLS regression, highlighting QR's capability to model complex data structures. Through these simulations, we will visualize and interpret how predictors influence the response variable across different quantiles, providing a more nuanced understanding of the data in each situation.

6.1.1 A Simulation-Based study with Homoscedasticity

In this scenario, we assume that the variance of errors remains constant across different values of the predictor variable, which means that the spread of residuals around the regression line does not change with the predictor. This is the standard assumption in OLS regression. However, while OLS captures the average effect of predictors, it does not reveal how the relationship may vary at different quantiles (parts) of the outcome distribution.

By applying quantile regression, we can assess whether the impact of predictors, though constant in variance, may vary across lower, median, and upper quantiles of the outcome variable. This offers a more nuanced view of the relationship, especially when we are interested in how predictors influence subpopulations (e.g., low, average, or high outcomes) differently.

The objective of this simulation study is to analyze the effect of a categorical variable, *education level* (with two levels: High School and College), and a continuous variable, *years of experience*, on the dependent variable *income*. We will use quantile regression to model the relationship at different points of the income distribution (i.e., at the 25th, 50th, and 75th percentiles) and compare the results to OLS regression.

Step 1: Simulate the data

A dataset of 500 observations is generated with income modeled as a function of experience and education level:

- **Experience**: Normally distributed with a mean of 10 and a standard deviation of 3.
- **Education Level**: Categorical variable with two levels, *High School* (0) and *College* (1).

The income is computed as:

$$\text{Income} = 30,000 + 1,500 \cdot \text{Experience} + 5,000 \cdot \text{Education} + \varepsilon,$$

where ε represents random noise with a standard deviation of 5,000, simulating variability. The dataset captures both continuous and categorical predictors to analyze income

dynamics effectively.

Step 2: Fit Quantile Regression

We fit quantile regression at three different quantiles: 0.25 (25th percentile), 0.50 (median), and 0.75 (75th percentile). We also fit an OLS regression for comparison.

Step 3: Interpret of the Results

Each regression output provides coefficients for the intercept, years of experience, and education (as a dummy variable).

	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	OLS
Intercept	26976.54	29179.14	33505.78	29506.56
experience	1433.66	1560.62	1507.38	1527.09
educationCollege	5013.60	5262.11	5055.31	5213.90

Table 6.1: Coefficients for Quantile and OLS Models

- Intercept: QR reveals how baseline income varies significantly across income levels, increasing from $\tau = 0.25$ to $\tau = 0.75$, capturing variations missed by OLS, which only reflects the central tendency.
- Experience: QR highlights a stable effect of experience across income levels, with slight variations that are more informative than the OLS estimate tied to the mean.
- Education (College): QR uncovers subtle differences in the returns to education, with diminishing benefits at higher income levels ($\tau = 0.75$), insights that OLS cannot provide.
- AIC Comparison: - While OLS achieves a slightly better AIC (9985.75) in this homoscedastic scenario, quantile regression excels by uncovering predictor effects at different parts of the response distribution, such as the lower ($\tau = 0.25$) and upper ($\tau = 0.75$) quantiles, providing insights into extremes that OLS cannot capture.

	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	OLS
AIC Value	10102.75	10031.22	10081.10	9985.754

Table 6.2: AIC Values for Quantile and OLS Models

Step 4: Visualization of the Quantile Regression Lines

In this step, we plot the quantile regression lines for $\tau = 0.25$, $\tau = 0.5$, and $\tau = 0.75$ alongside the OLS regression line to compare how **experience** affects **income** at different income levels. The quantile regression lines illustrate varying effects of experience on income for lower, median, and higher-income individuals, while the OLS line provides an average effect across the entire distribution. This visualization highlights differences in slopes, revealing how the impact of experience changes across different quantiles of the income distribution.

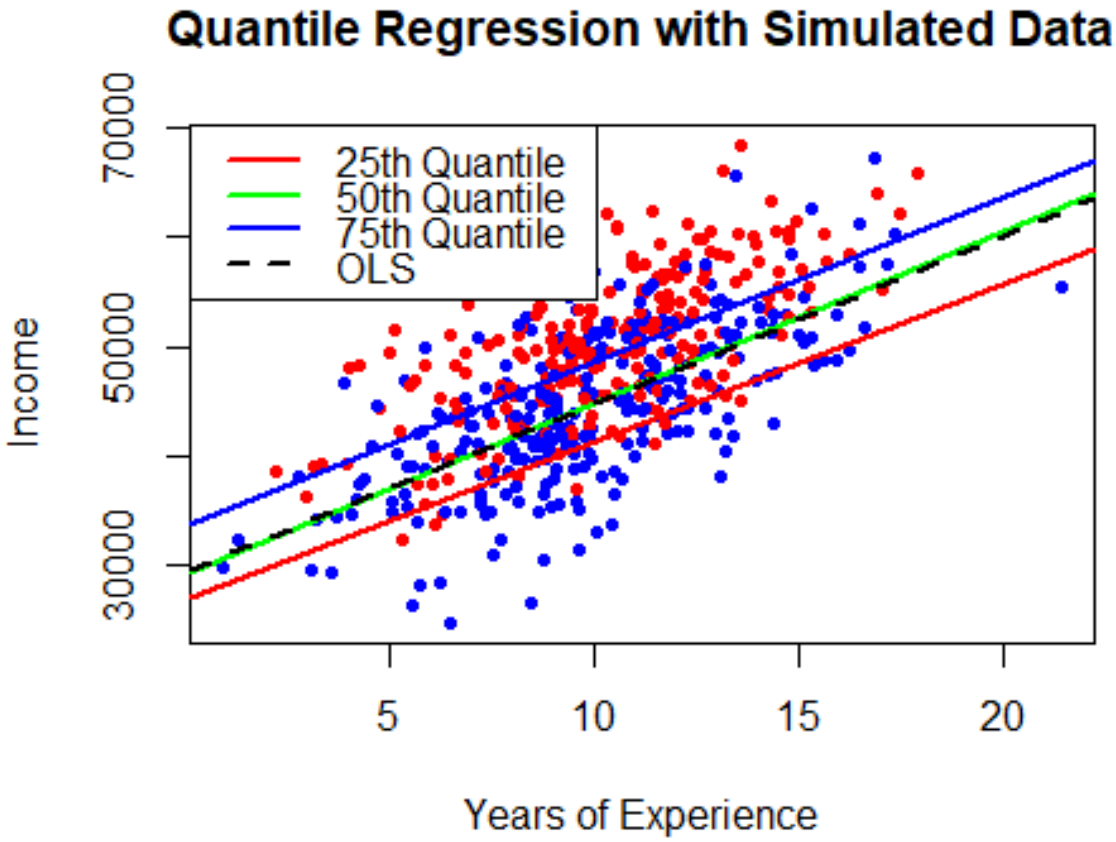

Figure 6.1: Quantile Regression with Homoscedastic Simulated Data

Step 5: Conclusions

- In the homoscedasticity scenario, both Quantile Regression (QR) and Ordinary Least Squares (OLS) models provide consistent estimates at $\tau = 0.5$, where the variance of errors remains constant. However, QR offers a more detailed view of

relationships across different quantiles of the outcome distribution.

- OLS estimates the central tendency (mean), summarizing the overall effect of predictors, while QR captures variations at different quantiles, revealing relationships across the entire income distribution.

- QR allows for understanding how the impact of education and experience may vary across income levels, uncovering differences between lower-income and higher-income individuals that OLS cannot capture.

R Code for Fitting of Simulated data with Homoscedasticity

```
set.seed(1)
n <- 500  # number of observations
# Continuous variable: experience (in years)
experience <- rnorm(n, mean = 10, sd = 3)
# Categorical variable: education (dummy variable, "High School"
   = 0, "College" = 1)
education <- rbinom(n, 1, 0.5)
# Simulate income based on experience and education with some
   random noise
# Base effect for high school education and an additional effect
   for college education
income <- 30000 + 1500 * experience + 5000 * education + rnorm(n,
    0, 5000)
sim_data <- data.frame(income, experience, education = factor(
   education, levels = c(0, 1), labels = c("High School", "
   College")))
# Quantile regression at different quantiles
qr_25 <- rq(income ~ experience + education, data = sim_data, tau
    = 0.25)
qr_50 <- rq(income ~ experience + education, data = sim_data, tau
    = 0.50)  # Median
qr_75 <- rq(income ~ experience + education, data = sim_data, tau
    = 0.75)
# OLS regression for comparison
ols <- lm(income ~ experience + education, data = sim_data)
# Display the results
summary(qr_25,se="nid")
summary(qr_50,se="nid")
summary(qr_75,se="nid")
summary(ols)
#AIC
AIC(qr_25)
AIC(qr_50)
AIC(qr_75)
AIC(ols)
```

```
# Plot the data
plot(sim_data$experience, sim_data$income, col = ifelse(sim_data$
   education == "College", "red", "blue"),
     pch = 16, cex = 0.7, xlab = "Years of Experience", ylab = "
        Income", main = "Quantile Regression with Simulated Data"
        )
# Add quantile regression lines
abline(qr_25, col = "red", lwd = 2)   # 25th percentile
abline(qr_50, col = "green", lwd = 2) # 50th percentile
abline(qr_75, col = "blue", lwd = 2)  # 75th percentile
# Add OLS regression line
abline(ols, col = "black", lwd = 2, lty = 2)  # OLS line (dashed)
# Add legend
legend("topleft", legend = c("25th Quantile", "50th Quantile", "
   75th Quantile", "OLS"),
       col = c("red", "green", "blue", "black"), lwd = 2, lty = c
          (1, 1, 1, 2))
```

6.1.2 Simulation based study with Heteroscedasticity

This study involves a simulation-based analysis to examine the relationship between vehicle fuel efficiency (MPG) and predictors, including engine size, vehicle weight, and the number of cylinders. A simulated dataset is generated to represent realistic vehicle data, incorporating heteroscedastic errors to account for the variability in fuel efficiency that increases with certain predictors, such as vehicle weight. Variables are systematically defined and named to align with real-world automotive contexts. Quantile regression is then applied to estimate the effects of predictors at different points of the MPG distribution (e.g., lower, median, and upper quantiles), providing insights beyond the mean effects captured by traditional regression methods. This approach highlights how these predictors influence vehicles across various performance levels, offering a more nuanced understanding of fuel efficiency patterns.

Step 1: Simulate the data

We simulate a dataset with 500 observations to study the relationship between vehicle fuel efficiency (measured in MPG) and two predictors: engine size and vehicle weight.

- **Engine size:** Simulated as a continuous variable (in liters) from a normal distribution with mean 2.5 and standard deviation 0.5. Larger engines generally lead to lower fuel efficiency.

- **Vehicle weight:** Simulated as a continuous variable (in tons) from a normal distribution with mean 1.5 and standard deviation 0.3. Heavier vehicles typically have lower fuel efficiency.
- **Heteroscedasticity:** The error term's standard deviation increases with engine size and vehicle weight, reflecting greater variability in fuel efficiency for larger and heavier vehicles.

The dependent variable, fuel efficiency (MPG), is computed as:

$$\text{MPG} = 35 - 5 \times \text{engine size} - 7 \times \text{vehicle weight} + \text{errors}$$

where the errors are drawn from a normal distribution with mean 0 and a standard deviation proportional to engine size and vehicle weight. The final dataset, `data_vehicle`, contains fuel efficiency, engine size, and vehicle weight for further analysis.

Step 2: Fit Quantile Regression

We fit quantile regression at three different quantiles: 0.25 (25th percentile), 0.50 (median), and 0.75 (75th percentile). We also fit an OLS regression for comparison.

Step 3: Interpret of the Results

Each regression output provides coefficients for the intercept, years of experience, and education (as a dummy variable).

	$\boldsymbol{\tau = 0.25}$	$\boldsymbol{\tau = 0.50}$	$\boldsymbol{\tau = 0.75}$	**OLS**
(Intercept)	41.6452	41.9676	51.6920	45.959
engine_size	-7.2592	-6.7796	-7.5958	-6.432
vehicle_weight	-13.2765	-8.4035	-6.2359	-10.306

Table 6.3: Coefficients for Quantile and OLS Models

– Intercepts: The intercepts vary across quantiles, being lowest at $\tau = 0.25$ (41.65) and highest at $\tau = 0.75$ (51.69). This indicates that the baseline response variable is higher for observations in the upper quantiles.

– Engine Size: Engine size consistently shows a negative effect across all quantiles, with the strongest impact at $\tau = 0.75$ (-7.60). This suggests that larger engine sizes significantly reduce fuel efficiency for higher-performing outcomes.

– Vehicle Weight: Vehicle weight also has a negative effect, with the strongest impact observed at $\tau = 0.25$ (-13.28). This shows that heavier vehicles cause more reduction in fuel efficiency at the lower end of the response distribution.

– OLS vs. QR: In the presence of heteroscedasticity, where variance varies across quantiles, OLS is less appropriate as it estimates average effects only. Quantile regression highlights how predictors influence different parts of the response distribution, providing more nuanced insights.

– AIC Comparison: Quantile regression exhibits lower AIC values (4225.29 at $\tau = 0.25$, 4230.27 at $\tau = 0.50$) compared to OLS (AIC = 4247.94), further supporting the use of quantile regression in a heteroscedastic context.

	$\boldsymbol{\tau = 0.25}$	$\boldsymbol{\tau = 0.50}$	$\boldsymbol{\tau = 0.75}$	**OLS**
AIC	4225.29	4230.27	4376.84	4247.94

Table 6.4: AIC Comparison for Quantile and OLS Models

Step 4: Conclusions

- Quantile regression reveals how engine size and vehicle weight impact fuel efficiency (MPG) at different quantiles, offering a more detailed analysis than OLS, especially with heteroscedastic errors.

- The effect of engine size and vehicle weight is more pronounced at lower quantiles (poorer MPG) and decreases as MPG improves, highlighting varying impacts across the distribution.

- Heteroscedasticity makes quantile regression more effective than OLS, capturing varying predictor effects and providing deeper insights into fuel efficiency variations across vehicles.

R Code: Simulated Quantile Regression for Fuel Efficiency

```
set.seed(1)
n <- 500
# Independent variables
engine_size <- rnorm(n, mean = 2.5, sd = 0.5)  # Engine size in
   liters
vehicle_weight <- rnorm(n, mean = 1.5, sd = 0.3)  # Vehicle
   weight in tons
# Heteroscedastic errors with stronger scaling for variability
errors <- rnorm(n, mean = 0, sd = (engine_size * 2 + vehicle_
   weight * 5))
# Simulate fuel efficiency (MPG) with added extreme outliers
fuel_efficiency <- 35 - 5 * engine_size - 7 * vehicle_weight +
   errors
# Add outliers at specific indices
fuel_efficiency[c(50, 100, 150)] <- fuel_efficiency[c(50, 100,
   150)] + c(-50, 70, -30)
# Add additional outliers for a range of indices
fuel_efficiency[c(1:10)] <- fuel_efficiency[c(1:10)] - 20  #
   Reduce these by 20
fuel_efficiency[c(200:250)] <- fuel_efficiency[c(200:250)] + 30
   # Increase these by 30
#Combine into a data frame
data_vehicle <- data.frame(fuel_efficiency, engine_size, vehicle_
   weight)
# Quantile regression at different quantiles
library(quantreg)
qr_model <- rq(fuel_efficiency ~ engine_size + vehicle_weight,
   data = data_vehicle, tau = c(0.25, 0.5, 0.75))
# OLS regression for comparison
ols <- lm(fuel_efficiency ~ engine_size + vehicle_weight, data =
   data_vehicle)
# Display the results
print(summary(qr_model, se = "nid"))
print(summary(ols))
# AIC values
print(AIC(qr_model))
print(AIC(ols))
```

6.1.3 Simulation study with Outliers

In statistical analysis, outliers are observations that deviate significantly from most data points, often due to measurement errors or unique circumstances. In educational contexts, outliers may reflect exceptional student performance or anomalies during testing. This study simulates a dataset where variables like study hours, attendance, prior

GPA, and test scores are created. We introduce outliers in test scores to represent unusual student performance, and then use quantile regression to analyze how these factors influence student performance across different quantiles, showcasing QR's ability to account for outliers and provide more detailed insights compared to traditional methods.

Step 1: Simulate the data

We simulate a dataset with 500 observations to study the relationship between student test scores and three predictors: study hours, attendance percentage, and prior GPA.

- **Study Hours:** Total hours studied, drawn from a normal distribution (mean = 10, sd = 3).
- **Attendance:** Percentage of classes attended, drawn from a normal distribution (mean = 85, sd = 10).
- **Prior GPA:** Students' GPA before the course, drawn from a normal distribution (mean = 3.0, sd = 0.5).
- **Test Scores:** Computed as:

$$\text{TestScore} = 50 + 3 \times \text{StudyHours} + 0.5 \times \text{Attendance} + 10 \times \text{PriorGPA} + \text{noise}$$

  Noise is added from a normal distribution (mean = 0, sd = 10). Outliers are introduced by randomly increasing 50 test scores by values drawn from a normal distribution (mean = 50, sd = 5).

The resulting dataset, `data_students`, includes test scores and the predictors for further analysis.

Step 2: Fit Quantile Regression

We perform quantile regression at three different quantiles: 0.25 (25th percentile), 0.50 (median), and 0.75 (75th percentile). We also fit an OLS regression for comparison.

Step 3: Interpret of the Results

Each regression output provides coefficients for the intercept, years of experience, and education (as a dummy variable).

	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	OLS
(Intercept)	30.973	49.978	59.735	65.267
StudyHours	3.223	3.224	2.960	2.385
Attendance	0.606	0.484	0.507	0.445
PriorGPA	10.645	10.153	10.009	9.957
AIC	4181.233	4205.925	4447.700	4348.876

Table 6.5: Coefficients and AIC for Quantile and OLS Models (TestScore Dataset)

- **Intercept:** The intercepts rise from 38.918 ($\tau = 0.25$) to 54.635 ($\tau = 0.75$), indicating higher baseline test scores for higher-performing students. The OLS intercept is 51.878.

- **StudyHours:** The impact of study hours decreases slightly across quantiles, from 3.579 ($\tau = 0.25$) to 3.160 ($\tau = 0.75$), while OLS estimates 3.403. Study hours benefit lower-performing students more.

- **Attendance:** Attendance shows a stronger effect at higher quantiles, increasing from 0.441 ($\tau = 0.25$) to 0.595 ($\tau = 0.75$), with OLS at 0.495. Attendance is more influential for higher-performing students.

- **PriorGPA:** The influence of prior GPA declines from 10.337 ($\tau = 0.25$) to 7.782 ($\tau = 0.75$), with OLS estimating 8.201. Prior GPA is more critical for lower-performing students.

- **AIC:** The AIC values for quantile regression models are 4181.233 ($\tau = 0.25$), 4205.925 ($\tau = 0.50$), and 4447.700 ($\tau = 0.75$), compared to 4348.876 for OLS. This shows that QRis more effective at modeling specific performance levels than OLS, particularly for lower quantiles.

QR highlights varying predictor effects across student performance levels. Study hours boost scores for lower performers, attendance benefits top scorers, and prior GPA has the greatest impact on struggling students. AIC values reinforce the suitability of QR for nuanced analysis.

QR highlights varying predictor effects across student performance levels. Study hours boost scores for lower performers, attendance benefits top scorers, and prior GPA has the greatest impact on struggling students.

Step 4: Conclusions

- QR offers a robust alternative to OLS by capturing predictor effects across different performance levels and providing resilience against outliers. Unlike OLS, which estimates average effects, QR reveals nuanced relationships, making it ideal for datasets with variability or extreme values.
- For lower-performing students ($\tau = 0.25$), study hours and prior GPA have a stronger influence, suggesting targeted interventions in these areas can significantly boost their performance. In contrast, attendance shows a greater impact on higher-performing students ($\tau = 0.75$), highlighting its role in sustaining high achievement.
- AIC values further support QR's utility, showing better model fit at lower quantiles compared to OLS. This underscores QR's ability to provide tailored insights, enabling personalized strategies for improving student outcomes.
- Overall, QR equips educators and policymakers with actionable, reliable insights to address performance disparities across the spectrum, even in the presence of outliers.

R Code: Robust Simulation of Student Performance

```
set.seed(1)
library(quantreg)
# Simulate predictor variables
n <- 500  # Number of students
StudyHours <- rnorm(n, mean = 10, sd = 3)  # Hours studied,
   average of 10 with some variability
Attendance <- rnorm(n, mean = 85, sd = 10) # Attendance
   percentage, average of 85%
PriorGPA <- rnorm(n, mean = 3.0, sd = 0.5) # Prior GPA, average
   of 3.0
# Simulate test scores based on predictors with added noise and
   some outliers
TestScore <- 50 + 3 * StudyHours + 0.5 * Attendance + 10 *
   PriorGPA + rnorm(n, mean = 0, sd = 10)
# Introduce some outliers
TestScore[sample(1:n, 50)] <- TestScore[sample(1:n, 5)] + rnorm
   (5, mean = 50, sd = 5)
data_students <- data.frame(TestScore, StudyHours, Attendance,
   PriorGPA)
# Quantile regression models at different quantiles
```

```
qr_model <- rq(TestScore ~ StudyHours + Attendance + PriorGPA,
   data = data_students, tau = c(0.25,0.5,0.75))
# OLS regression model for comparison
ols <- lm(TestScore ~ StudyHours + Attendance + PriorGPA, data =
   data_students)
# Display model summaries
summary(qr_model, se = "nid")
summary(ols)
#AIC
AIC(qr_model)
AIC(ols)
```

6.2 Simple Quantile Regression Analysis of the Engel Dataset

In this example, we examine the Engel food expenditure data, which includes 235 observations on income and food expenditure for Belgian working-class households, as used in Koenker and Bassett (1982). We aim to compare the performance of a linear regression model and a quantile regression model. Given the presence of heteroscedasticity in the data, the quantile regression model is expected to provide a more accurate representation of the relationship between income and food expenditure across different quantiles of the income distribution. This comparison will highlight the advantages of using quantile regression, particularly in capturing the variability that a linear model might overlook. variable description is as follow:

- *income*: annual household income in Belgian francs
- *foodexp*: annual household food expenditure in Belgian francs

R Code: Simple Quantile Regression on the Engel Dataset

```
# Load necessary libraries
library(quantreg)      # For Quantile Regression
library(performance)   # For model evaluation
library(ggplot2)       # For data visualization
# Load the Engel dataset
data(engel)
# To create scatter plot matrix of engel data
scatter_plot_matrix <- ggpairs(engel,
                    title ='scatter plot matrix of engel data',
```

```
                    lower = list(continuous = "smooth"))
# Fit the OLS regression model
ols_model <- lm(foodexp ~ income, data = engel)
# Fit Quantile Regression models for different percentiles
qr_model_50 <- rq(foodexp ~ income, tau = 0.5, data = engel)  #
   Median (50th percentile)
qr_model_25 <- rq(foodexp ~ income, tau = 0.25, data = engel) #
   25th percentile
qr_model_75 <- rq(foodexp ~ income, tau = 0.75, data = engel) #
   75th percentile
# OR
qr_model <- rq(foodexp ~ income, tau=c(0.25,0.5,0.75), data =
   engel)
# OLS Model evaluation using the performance library
check_model(ols_model)
check_heteroscedasticity(ols_model)
# Display the results
summary(ols_model)
summary(qr_model, se="nid")
# Compare models using AIC
ols_aic <- AIC(ols_model)
qr_aic <- AIC(qr_model)
# Fit an intercept-only model (null model)
null_model <- rq(foodexp ~ 1, tau = c(0.25,0.5,0.75),
                    data = engel)
# calculate Pseudo R^2 for QR models
Pseudo.R2=1-qr_model$rho/null_model$rho
# Create a scatter plot of the data
plot(engel$income, engel$foodexp, pch = 16, col = "blue",
     xlab = "Income", ylab = "Food Expenditure",
     main = "OLS vs Quantile Regression Lines")
# Add the OLS regression line
abline(ols_model, col = "red", lwd = 2, lty = 1)
# Add the Quantile Regression lines
abline(qr_model_50, col = "green", lwd = 2, lty = 1)   # Median
   (50th percentile)
abline(qr_model_25, col = "orange", lwd = 2, lty = 2)  # 25th
   percentile
abline(qr_model_75, col = "purple", lwd = 2, lty = 2)  # 75th
   percentile
# Add a legend
legend("topleft", legend = c("OLS", "QR 50%", "QR 25%", "QR 75%")
   ,
       col = c("red", "green", "orange", "purple"), lty = c(1, 1,
           2, 2), lwd = 2)
```

6.2.1 Output and Interpretations

i) Matrix plot

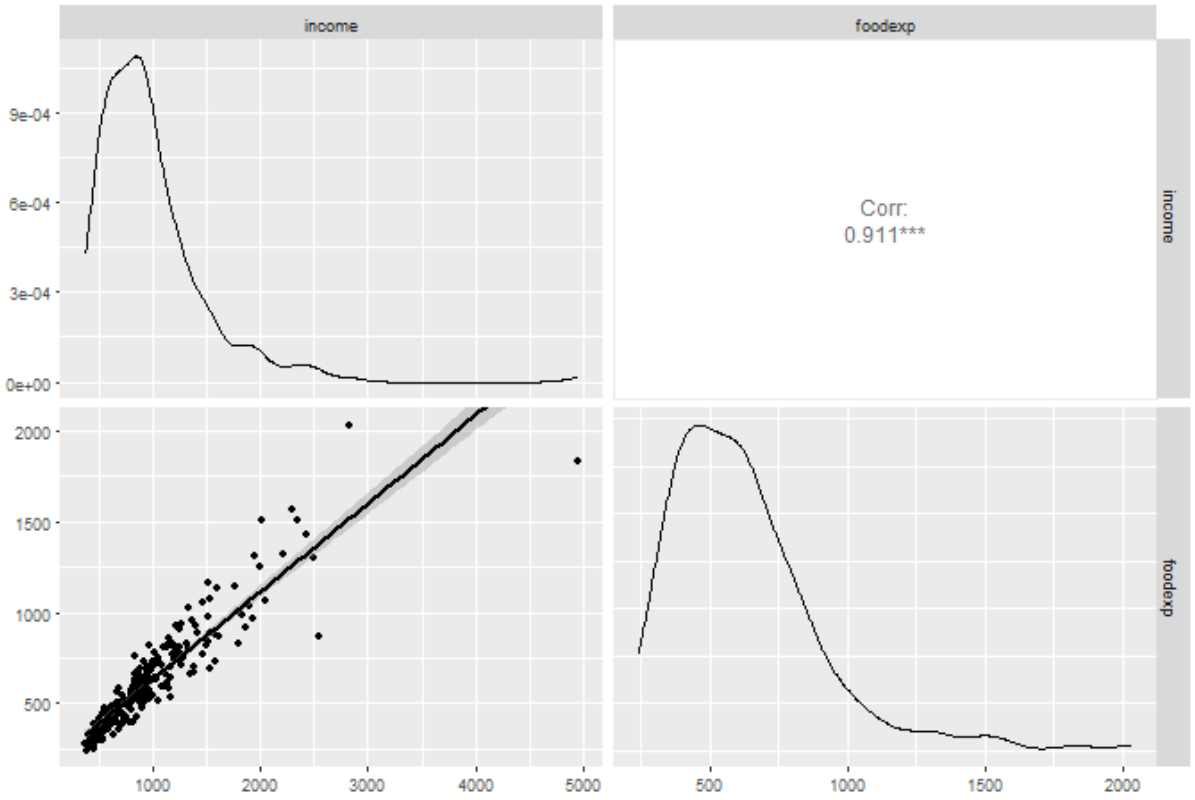

Figure 6.2: scatter plot matrix of engel data

There is a strong positive correlation (0.911) between income and food expenditure, as shown in the upper-right corner. This suggests that as income increases, food expenditure also increases. The lower-left scatter plot demonstrates a clear upward trend between income and food expenditure. However, the relationship appears to slightly deviate from linearity for higher income levels.The density distributions of income and food expenditure are positively skewed, indicating that most households have lower income and food expenditure values, while a smaller proportion of households have higher values. A few points in the scatter plot indicate potential outliers where either the income or food expenditure is unusually high compared to the general trend.

ii) OLS model adequacy checking

The analysis indicates that the Engel dataset exhibits **heteroscedasticity**, as confirmed by the following warning:

Heteroscedasticity (non-constant error variance) detected ($p < 0.001$).

- **Heteroscedasticity:** This occurs when the variance of the error terms (residuals) in a regression model is not constant across all levels of the independent variable(s).
- The **p-value** of less than 0.001 indicates that the test for heteroscedasticity is highly significant. This leads to the rejection of the null hypothesis of homoscedasticity (constant error variance) with a very high level of confidence.
- **Violation of Assumption:** One of the key assumptions of Ordinary Least Squares (OLS) regression is violated. OLS assumes that residuals have constant variance (homoscedasticity).

iii) Model summary

Model Type	**OLS**	$\boldsymbol{\tau = 0.25}$	$\boldsymbol{\tau = 0.50}$	$\boldsymbol{\tau = 0.75}$
Intercept	147.47539	95.48354	81.48225	62.39659
(SE)	(15.95708)	(21.39237)	(19.25066)	(16.30538)
Income	0.48518	0.47410	0.56018	0.64401
(SE)	(0.01437)	(0.02906)	(0.02828)	(0.02324)

Table 6.6: Estimates and Standard Errors for OLS and Quantile Regression Models

The table presents the results of an Ordinary Least Squares (OLS) regression and quantile regression models at three quantiles ($\tau = 0.25, 0.50, 0.75$) for the relationship between income and the dependent variable, with the coefficient estimates and standard errors (SE) provided.

1. **Intercept:**
   - The intercept represents the estimated value of the dependent variable when income is zero.
   - For the OLS model ($\tau = 0.50$), the intercept is 147.47539, which is the average value of the dependent variable across all observations.
   - At the lower quantile ($\tau = 0.25$), the intercept is 95.48354, indicating the predicted value of the dependent variable for the lower 25% of income values.
   - At the higher quantile ($\tau = 0.75$), the intercept is 62.39659, suggesting the predicted value of the dependent variable for the upper 25% of income values.

2. **Income Coefficients:**

   - **OLS Model:** The coefficient for income is 0.48518, suggesting that for every 1-unit increase in income, the dependent variable is expected to increase by 0.48518 units on average. This is the central tendency or median effect.
   - **Quantile Regression:**
     (a) $\tau = 0.25$: The coefficient is 0.47410, indicating that for lower-income individuals (the bottom 25% of the distribution), an increase in income leads to a smaller increase in the dependent variable (0.47410 units).
     (b) $\tau = 0.50$: The coefficient is 0.56018, showing that at the median income level ($\tau = 0.50$), income has a slightly higher effect (0.56018 units increase in the dependent variable) compared to the lower quantile.
     (c) $\tau = 0.75$: The coefficient is 0.64401, which is the highest of the three, indicating that for individuals at the top 25% of the income distribution, income has the greatest effect on the dependent variable (0.64401 units increase).

3. **Standard Errors (SE):**

   - The standard errors represent the variability or uncertainty in the coefficient estimates.
   - As the quantile increases, the standard errors tend to decrease slightly, which suggests more precision in estimating the income effect at higher quantiles.

iv. Comparing OLS and QR models

Model	**OLS**	$\boldsymbol{\tau = 0.25}$	$\boldsymbol{\tau = 0.50}$	$\boldsymbol{\tau = 0.75}$
AIC	2897.351	2861.481	2827.260	2823.266
R^2**/Pseudo** R^2	0.8296	0.5540	0.6206	0.6966

Table 6.7: AIC Values and R^2/Pseudo R^2 for OLS and QR Models at Different Quantiles

- AIC values provide a measure of model fit, where lower values indicate better fit while balancing model complexity.

- In this case, the AIC decreases from the OLS model (2897.351) to the quantile regression models, with the lowest AIC observed at $\tau = 0.75$ (2823.266).
- This suggests that the quantile regression models, particularly at higher quantiles, provide a better fit to the data than OLS.
- For OLS, $R^2 = 0.8296$, indicating that 82.96% of the variation in the dependent variable (e.g., food expenditure) is explained by the independent variables in the model. This reflects a strong fit.
- For the quantile regression models, the Pseudo R^2 values vary by quantile, indicating the goodness of fit of the model at different points of the distribution. At $\tau = 0.25$, the Pseudo $R^2 = 0.5540$, meaning that 55.40% of the goodness of fit of the model at the lower quantile. At $\tau = 0.50$, the Pseudo $R^2 = 0.6206$, showing an improved fit at the median quantile with 62.06% of the goodness of fit. At $\tau = 0.75$, the Pseudo $R^2 = 0.6966$, indicating the best fit among the quantile regression models.
- The pseudo R^2 and adjusted R^2 are based on different concepts and cannot be directly compared. Adjusted R^2 measures the proportion of variance explained in OLS regression, while pseudo R^2 provides a goodness-of-fit measure specific to models like logistic or quantile regression.

The graph compares Ordinary Least Squares regression with Quantile Regression at the 25th, 50th, and 75th percentiles to analyze the relationship between income and food expenditure. The OLS line represents the average trend, showing how food expenditure changes with income on average. However, the quantile regression lines reveal variations across different segments of the data distribution. The QR 25% line indicates lower food expenditures for low-income groups compared to the OLS trend, while the QR 75% line highlights higher food expenditures for high-income groups. This suggests that food expenditure behavior varies significantly across income levels, which the OLS model alone cannot capture.Quantile regression provides a more detailed understanding of this relationship by examining heterogeneity across different income levels.

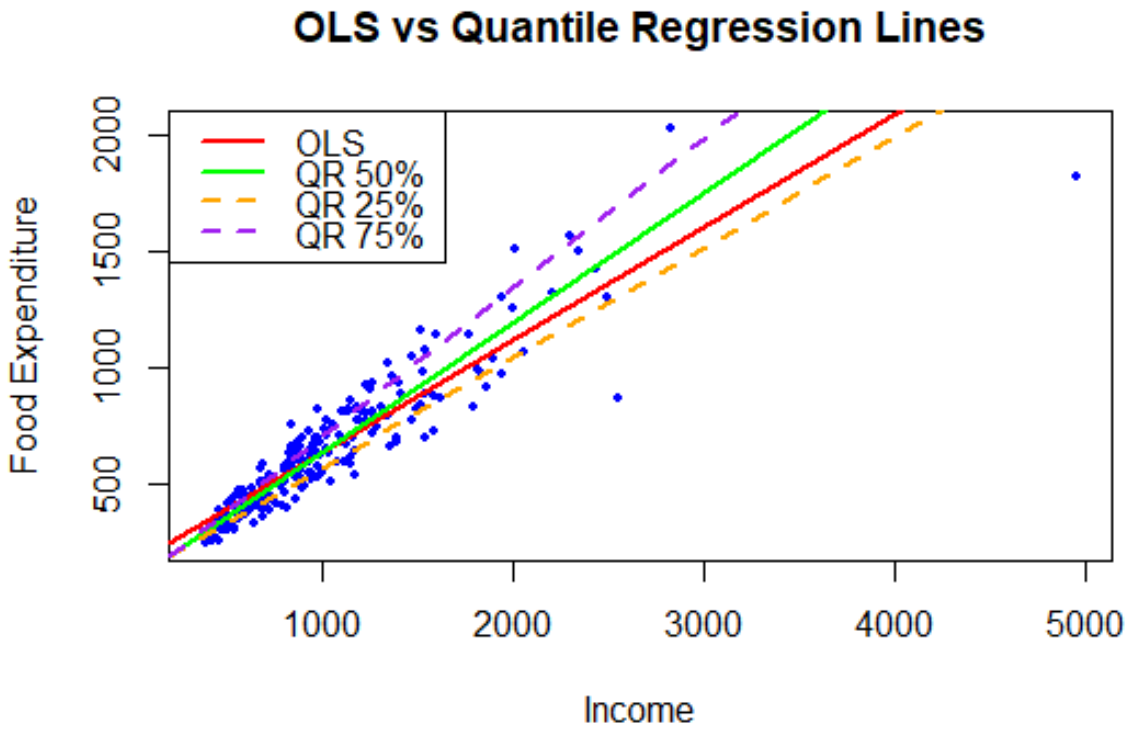

Figure 6.3: QR vs OLS on Engel dataset

6.2.2 Conclusions

- Heteroscedasticity in the Engel dataset suggests that the relationship between *income* and *food expenditure* varies, necessitating adjustments or alternative models.
- Quantile regression reveals that the effect of income on food expenditure increases from lower quantiles ($\tau = 0.25$) to upper quantiles ($\tau = 0.75$), indicating greater sensitivity for higher-income individuals.
- QR is better suited for heteroscedastic or skewed data, providing a more nuanced view of predictor effects at different points in the distribution, and offering a better fit than OLS.
- The decreasing AIC and improving Pseudo R^2 values at higher quantiles suggest that the model performs best at $\tau = 0.75$, capturing the strongest fit for higher food expenditures.

6.3 Quantile Regression of the AirQuality Dataset

In this analysis, we apply quantile regression to the **AirQuality dataset**, a classic dataset that includes measurements of air quality variables. The primary focus is on modeling the *Ozone* levels as the response variable, with *Wind* speed and *Temperature* as the predictor variables. While OLS regression captures the average relationship between Ozone levels and the predictors, quantile regression enables us to explore how these relationships vary across different levels of Ozone concentration, such as the lower (e.g., $\tau = 0.25$), median (e.g., $\tau = 0.50$), and upper (e.g., $\tau = 0.75$) quantiles.

Given the well-documented challenges in predicting Ozone levels due to their variability and sensitivity to environmental conditions, quantile regression offers a robust framework for gaining deeper insights into how Wind and Temperature influence air quality. This analysis aims to uncover these nuanced relationships, providing valuable information for environmental monitoring and decision-making.

The **AirQuality** dataset contains daily air quality measurements collected in New York from May 1, 1973, to September 30, 1973. It provides detailed observations of air pollution and meteorological conditions over 153 days, making it a valuable resource for environmental and statistical analysis.

Variables in the Dataset

- **Ozone** (*numeric*): Represents the mean ozone concentration in parts per billion (ppb) measured from 1300 to 1500 hours at Roosevelt Island. This variable is an indicator of air quality and is often influenced by weather conditions and pollution sources.

- **Solar.R** (*numeric*): Indicates the solar radiation in Langleys, in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park. This variable provides insight into the amount of solar energy received, which can affect ozone formation.

- **Wind** (*numeric*): Denotes the average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport. Wind speed plays a significant role in the dispersion and concentration of air pollutants.

- **Temp** (*numeric*): Represents the maximum daily temperature in degrees Fahrenheit recorded at LaGuardia Airport. Temperature influences atmospheric chemical reactions, including the production of ozone.

- **Month** (*numeric*): Specifies the month of the observation(1 - 12). This variable helps in analyzing seasonal variations in air quality.

- **Day** (*numeric*): Indicates the day of the month when the observation was recorded. This variable is primarily used for identifying specific dates in the dataset.

R code: Multiple Quantile Regression on the Airquality Dataset

```
# Load necessary libraries
library(quantreg)      # For Quantile Regression
library(performance)   # For model evaluation
library(ggplot2)       # For data visualization
library(GGally)
data(airquality)    # Load data
# Remove rows with missing values
airquality <- na.omit(airquality)
#scatter plot matrix of airquality dataset
scatter_plot_matrix <- ggpairs(airquality,
                 title ='scatter plot matrix of airquality data',
                 lower = list(continuous = "smooth"))
# Fit OLS regression model
lm_model <- lm(Ozone ~ Temp + Wind, data = airquality)
# Model adequacy checking of OLS model
check_heteroscedasticity(lm_model)
check_model(lm_model)
check_collinearity(lm_model)
# Fit quantile regression models for multiple quantiles
qr_model <- rq(Ozone ~ Temp + Wind, tau = c(0.25, 0.5, 0.75),
   data = airquality)
# Model Summary
summary(qr_model)
summary(lm_model)
# Model Evaluation with AIC
AIC(lm_model)
AIC(qr_model)
# Fit an intercept-only model (null model)
null_model <- rq(ozone ~ 1, tau = c(0.25,0.5,0.75), data =
   airquality)
# Calculate Pseudo R^2 for QR models
Pseudo.R2=1-qr_model$rho/null_model$rho
```

6.3.1 Outputs and Interpretation

i) Matrix plot

The scatter plot matrix reveals the relationships and correlations among the variables. Specifically focusing on the `Ozone`, `Wind`, and `Temp` variables:

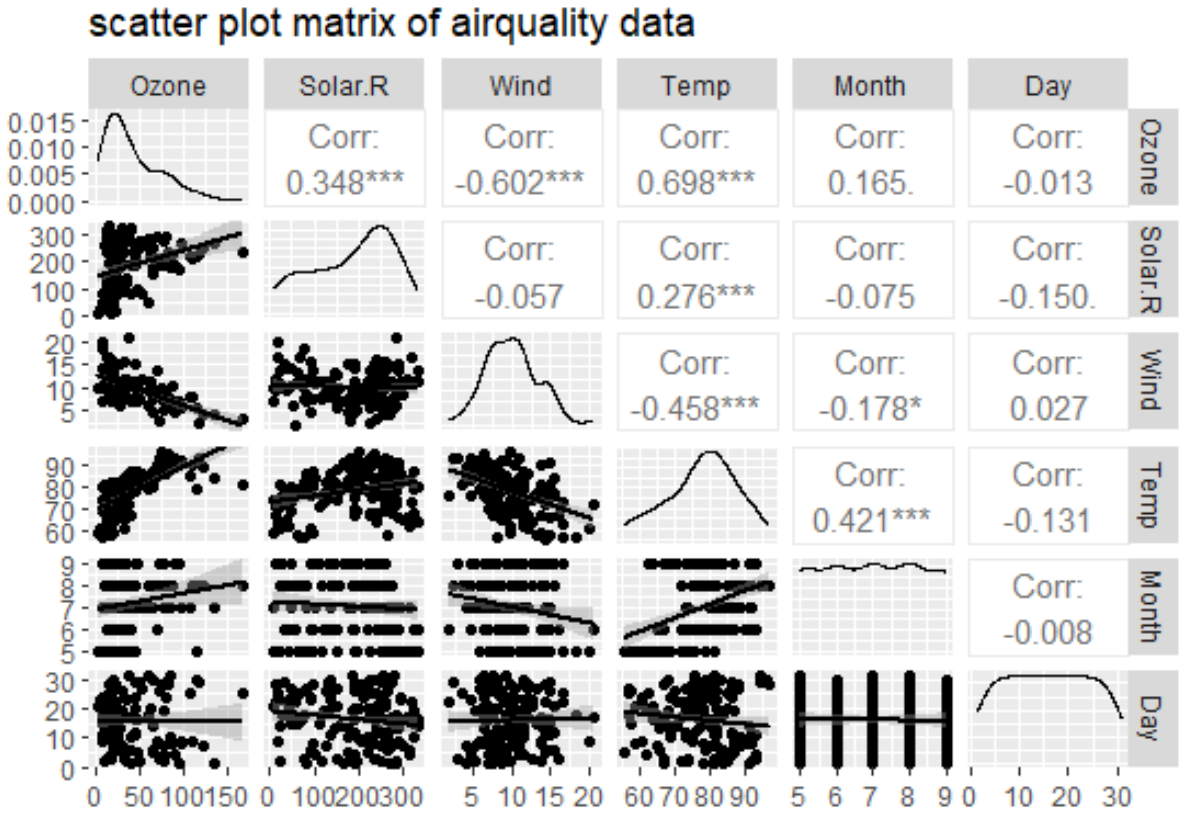

Figure 6.4: Scatter plot matrix of airquality data

- The correlation coefficient between `Ozone` and `Wind` is -0.602, indicating a strong negative correlation. As `Wind` increases, `Ozone` levels tend to decrease.
- The correlation coefficient between `Ozone` and `Temp` is 0.698, indicating a strong positive correlation. As `Temp` increases, `Ozone` levels tend to increase.
- The correlation coefficients between `Ozone` and these variables are much lower (e.g., 0.348 for `Solar.R` and 0.165 for `Month`), indicating weaker relationships.

ii. Model adequacy checking of OLS model

The results indicate that heteroscedasticity is present in the model, as the non-constant error variance test returned a statistically significant p-value ($p = 0.007$). This suggests

that the variability of the residuals is not consistent across levels of the predictor variables, potentially violating one of the key assumptions of linear regression. However, the multicollinearity check shows low correlation among the predictor variables, meaning that multicollinearity is not a concern in this model. Although the predictors are reliable in terms of independence, the presence of heteroscedasticity may impact the efficiency of the estimates and the validity of statistical inferences.

iii. Model summary

Variable	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	**OLS**
Intercept	-72.563	-81.270	-76.876	-67.322
(SE)	(12.800)	(22.926)	(38.342)	(23.621)
Temp	1.580	1.911	1.957	1.828
(SE)	(0.199)	(0.249)	(0.406)	(0.251)
Wind	-2.251	-2.894	-2.298	-3.295
(SE)	(0.475)	(0.744)	(1.118)	(0.671)

Table 6.8: Coefficients and Standard Errors for QR ($\tau = 0.25, 0.5, 0.75$) and OLS

Intercept:

- At all quantile levels ($\tau = 0.25, 0.5, 0.75$) and in the OLS model, the intercept is negative, suggesting that when both `Temp` and `Wind` are 0, the `Ozone` concentration is predicted to be below zero.
- The magnitude of the intercept varies across quantiles, reflecting the differing baseline levels of `Ozone` at different parts of its distribution.

Temperature (`Temp`):

- The coefficient for `Temp` is positive across all quantiles and OLS, indicating a consistent positive relationship between temperature and `Ozone`.
- For instance, at $\tau = 0.25$, a 1-unit increase in `Temp` is associated with a 1.58-unit increase in `Ozone`, whereas at $\tau = 0.75$, the increase is slightly higher at 1.96 units.
- The coefficients increase slightly across quantiles, suggesting that higher temperatures have a stronger effect on the upper levels of `Ozone` concentration.

- The standard errors are relatively small, indicating precise estimates, and the coefficients are statistically significant.

Wind (`Wind`):

- The coefficient for `Wind` is consistently negative across all quantiles and OLS, indicating that higher wind speeds are associated with lower `Ozone` levels.
- For instance, at $\tau = 0.5$, a 1-unit increase in `Wind` is associated with a 2.89-unit decrease in `Ozone`.
- The effect is most pronounced at $\tau = 0.5$, suggesting that wind speed has the largest impact on the median `Ozone` levels. At the upper quantile ($\tau = 0.75$), the effect is smaller.
- Standard errors are larger than for `Temp`, especially at $\tau = 0.75$, but the coefficients remain statistically significant.

iv. Comparing OLS and QR models

Model	$\boldsymbol{\tau = 0.25}$	$\boldsymbol{\tau = 0.50}$	$\boldsymbol{\tau = 0.75}$	**OLS**
AIC	981.30	990.00	1032.37	1003.42
R^2**/Pseudo** R^2	0.2761	0.3807	0.4271	0.5736

Table 6.9: AIC Values and R^2/Pseudo R^2 for OLS and QR Models at Different Quantiles

- The AIC values for the QR models are lower than for the OLS model. The lowest AIC value is observed for $\tau = 0.25$ (981.30), indicating that the QR model at this quantile fits the data better than the others.
- As the quantile increases to 0.50 and 0.75, the AIC values increase (990.00 and 1032.37, respectively), which suggests that the fit of the model deteriorates slightly at these higher quantiles compared to the $\tau = 0.25$ model.
- The OLS model (1003.42) has a higher AIC than all QR models, suggesting that it may not fit the data as well as the QR models, especially at the lower quantiles.

- For the QR models, the pseudo R^2 values increase as the quantile increases, with the highest value of 0.4271 for $\tau = 0.75$. This indicates that the model provides better fit.

- The OLS model has a higher R^2 (0.5736), suggesting it explains more of the variance in the dependent variable compared to the QR models. However, the QR models may provide a better fit at specific quantiles, capturing different aspects of the data that OLS cannot.

6.3.2 Conclusions

- The quantile regression coefficients reveal that the relationships between predictors (`Temp`, `Wind`) and `Ozone` vary across different parts of the distribution of `Ozone`. While the OLS coefficients provide an overall average effect of the predictors on `Ozone`, but they do not capture the variation in effects across different quantiles.

- Higher temperatures consistently increase `Ozone`, with a slightly stronger effect at higher quantiles, while wind reduces `Ozone`, with the strongest effect observed at the median quantile. The effect of `Temp` is stronger at higher quantiles ($\tau = 0.75$), while the effect of `Wind` weakens slightly.

- while the OLS model provides a better overall fit in terms of R^2, the QR models show promise at different quantiles, especially to capture variations in the lower and upper parts of the distribution of the dependent variable (Ozone).

6.4 Comparing QR and OLS in Homoscedastic Context

In the homoscedastic context, where the variance of the errors is constant across all levels of the predictor variable, comparing Quantile Regression and Ordinary Least Squares using the Boston dataset (with `medv` as the response variable and `rm` as the predictor) provides valuable insights into model behavior. The `medv` variable in the Boston dataset represents the median value of owner-occupied homes in $1000s$, and `rm` represents the average number of rooms per dwelling. OLS focuses on estimating the mean of the

dependent variable (`medv`) given the independent variable (`rm`), making it suitable for modeling the central tendency of the data. In contrast, Quantile Regression allows us to examine the relationship between `rm` and `medv` at different quantiles (e.g., 0.25, 0.5, 0.75), offering a more comprehensive view of the conditional distribution of the response variable.

In the homoscedastic setting, where the assumption of constant error variance holds, OLS typically provides efficient and unbiased estimates. However, QR can still provide added value by highlighting how the relationship between `rm` and `medv` may differ across the distribution of `medv`, which can be useful when the focus is on specific percentiles (such as the lower or upper tails). While OLS may offer a single estimate of the relationship, QR reveals a more nuanced understanding of how the predictor influences different parts of the response distribution. Thus, both methods have their strengths, with QR offering a more detailed exploration at different quantiles and OLS providing an efficient estimate of the mean.

R Code: QR vs OLS on Boston Dataset

```
# Load Boston Housing dataset
data("Boston")
# Use a subset of the data: predict median value of owner-
   occupied homes (medv) using the number of rooms (rm)
boston_data <- Boston[, c("rm", "medv")]
# Fit linear regression model
lm_model <- lm(medv ~ rm, data = boston_data)
# Model adequacy checking of OLS model
check_heteroscedasticity(lm_model)
check_model(lm_model)
# Fit quantile regression model for median (tau = 0.5)
qr_model <- rq(medv ~ rm, tau = c(0.25,0.5,0.75), data = boston_
   data)
# Model Summary
summary(qr_model,se="nid")
summary(lm_model)
# Model Evaluation with AIC
AIC(lm_model)
rbind(tau = c(0.25,0.5,0.75),AIC(qr_model))
# Fit an intercept-only model (null model)
null_model <- rq(medv ~ 1, tau = c(0.25,0.5,0.75), data = Boston)
# Calculate Pseudo R^2 for QR models
Pseudo.R2=1-qr_model$rho/null_model$rho
# Visualization: data points and regression lines
ggplot(boston_data, aes(x = rm, y = medv)) +
```

```
  geom_point(color = "blue", alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "red", linetype
     = "dashed", size = 1.2, fullrange = TRUE) +
  stat_smooth(method = "rq", method.args = list(tau = 0.5), se =
     FALSE, color = "green", size = 1.2, fullrange = TRUE) +
  labs(title = "Linear vs Quantile Regression: Boston Housing",
       x = "Average Number of Rooms (rm)",
       y = "Median Home Price (medv)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_color_manual(values = c("red", "green"), labels = c("
     Linear Regression", "Quantile Regression"))
```

6.4.1 Output and Interpretations

i. OLS model adequacy checking

The results from the heteroscedasticity test for the linear regression model indicate that there is no significant evidence of heteroscedasticity in the data. The test shows a p-value of 0.874, which is well above the common significance threshold of 0.05. This suggests that the error variance is constant across all levels of the predictor variable. Therefore, we can conclude that the assumption of homoscedasticity holds for this linear regression model, and the model does not exhibit any issues with varying error variance.

ii. Model Summary

Model	$\boldsymbol{\tau = 0.25}$	$\boldsymbol{\tau = 0.50}$	$\boldsymbol{\tau = 0.75}$	**OLS**
(Intercept)	−42.1019	−38.2249	−39.6321	−34.671
(SE)	(3.7336)	(1.5413)	(2.6913)	(2.650)
rm	9.9004	9.6998	10.3645	9.1020
(SE)	(0.6386)	(0.2594)	(0.4259)	(0.4190)

Table 6.10: Estimates and Standard Errors

- **Intercepts:** The intercept values for both OLS and QR models are similar across the quantiles, with OLS having a value of -34.671 and QR at $\tau = 0.50$ having -38.2249. Although there is a slight difference in the intercept values, the magnitudes are quite close, indicating that the models estimate a similar baseline value.

- **Slopes:** The slope coefficients for both OLS and QR models are also very similar. OLS has a slope of 9.1020, while QR at $\tau = 0.50$ has a slope of 9.6998. This suggests that both models estimate a very similar relationship between the predictor variable ('rm') and the response variable ('medv').

- **Standard Errors:** The standard errors for the coefficients of both OLS and QR models are comparable. For example, the standard error of the slope in OLS is 0.4190, while for QR at $\tau = 0.50$, it is 0.2594. The slightly smaller standard error for QR at $\tau = 0.50$ may be due to the method's robustness to outliers, although the differences are not large enough to indicate a significant discrepancy.

iii. Comparing OLS and QR models

Model	$\boldsymbol{\tau = 0.25}$	$\boldsymbol{\tau = 0.50}$	$\boldsymbol{\tau = 0.75}$	**OLS**
AIC	3392.938	3225.12	3323.343	3352.151

Table 6.11: AIC Values for OLS and QR Models at Different Quantiles

- At $\tau = 0.50$ (the median), the AIC values for the OLS and QR models are very similar (3225.12 for QR and 3352.151 for OLS). This suggests that, in terms of model fit, both OLS and QR at $\tau = 0.50$ yield almost identical results. Therefore, in a homoscedastic setting, the OLS model can provide a good fit and behave similarly to QR at the median quantile.

- Although OLS and QR at $\tau = 0.50$ behave similarly, QR models at other quantiles ($\tau = 0.25$ and $\tau = 0.75$) provide different insights into the data. At $\tau = 0.25$, the QR model performs worse than OLS, while at $\tau = 0.75$, QR provides a better fit than OLS. This highlights the flexibility of QR to capture the conditional distribution of the response variable at quantiles other than the mean (e.g., lower and upper quantiles).

iv. Visualization of OLS and QR model at ($\tau = 0.5$)

The graph compares the relationship between the average number of rooms (`rm`) and the median home price (`medv`) using **Linear Regression (OLS)** and **Quantile Regression (QR)** at $\tau = 0.5$.

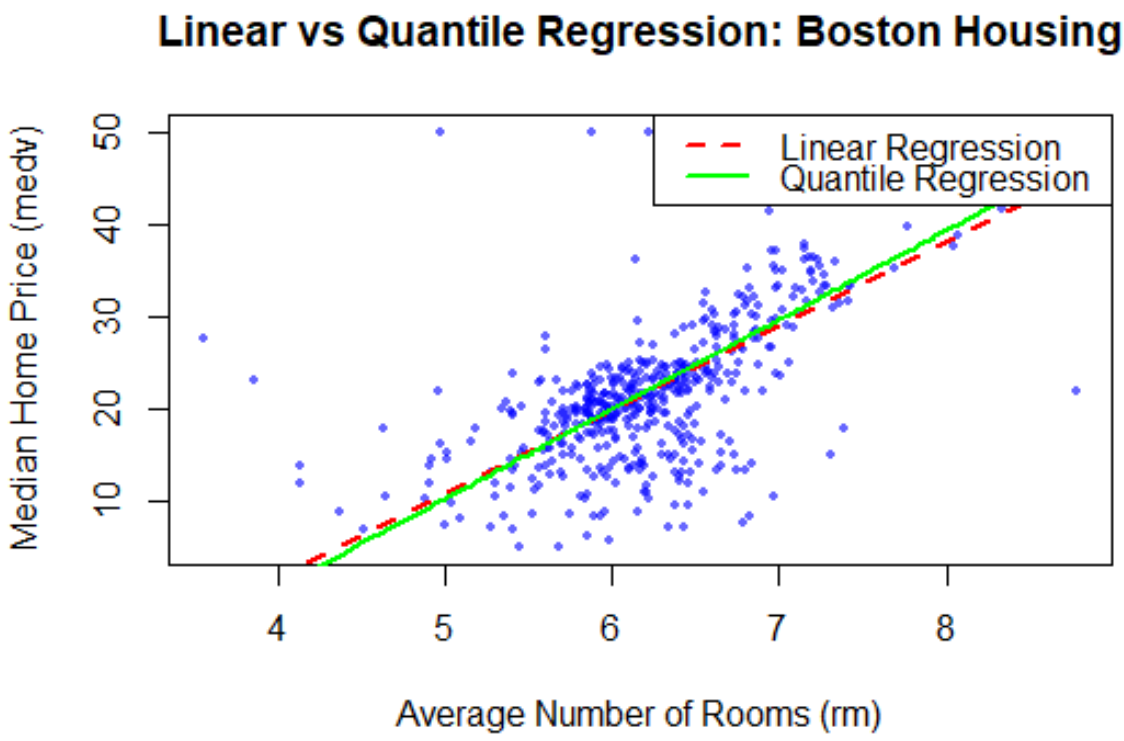

Figure 6.5: Linear vs Quantile Regression at ($\tau = 0.5$)

- Both regression lines demonstrate a positive trend, indicating an increase in the median home price (`medv`) as the average number of rooms (`rm`) increases. The OLS line (red, dashed) and the QR line (green) nearly overlap, particularly in this homoscedastic scenario, indicating similar results at $\tau = 0.5$.

- Although QR at $\tau = 0.5$ provides results similar to OLS, its robustness is evident when analyzing conditional relationships beyond the mean, such as the median or other quantiles.

- The overlap of regression lines further supports the earlier finding that the error variance is homoscedastic in this scenario.

- For this dataset, OLS and QR at $\tau = 0.5$ yield consistent central relationships. However, QR offers the advantage of exploring variations in the response (`medv`) across different conditional quantiles, providing a more nuanced understanding.

6.4.2 Conclusions

- In this homoscedastic scenario, both OLS and QR at $\tau = 0.5$ provide nearly identical results. The slight difference in coefficients and standard errors can be attributed to the fact that QR is a more flexible method, but under homoscedasticity, both methods yield the same Conclusions regarding the relationship between ‘rm‘ and ‘medv‘. Thus, the results from OLS and QR at $\tau = 0.5$ are not significantly different, supporting the idea that OLS is an appropriate model when the residuals have constant variance (homoscedasticity).

- In a homoscedastic scenario, if the focus is on estimating the response at the median (mean) of the distribution, OLS and QR at $\tau = 0.50$ provide virtually the same results. However, if the goal is to understand the response at other parts of the distribution (such as lower or upper quantiles), QR models with different values of τ (e.g., 0.25 or 0.75) become crucial. QR allows for a more nuanced understanding of how the predictors influence the response at different points of the conditional distribution, which OLS cannot capture. Thus, QR models are especially useful when the distribution of the response variable is not symmetric or when the effects of predictors may vary across quantiles.

- In a homoscedastic context, OLS and QR at $\tau = 0.5$ behave similarly, capturing central trends effectively. The true strength of QR lies in its ability to analyze conditional quantiles, making it a valuable tool for studying data patterns beyond mean trends.

Bibliography

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[2] Andreas Beyerlein. Quantile regression—opportunities and challenges from a user's perspective. *American journal of epidemiology*, 180(3):330–331, 2014.

[3] Moshe Buchinsky. Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of human resources*, pages 88–126, 1998.

[4] Sebastian Buhai. Quantile regression: overview and selected applications. *Ad Astra*, 4(4):1–17, 2005.

[5] Victor Chernozhukov, Iván Fernández-Val, and Tetsuya Kaji. Extremal quantile regression. *Handbook of Quantile Regression*, pages 333–362, 2017.

[6] Kiranmoy Das, Martin Krzywinski, and Naomi Altman. Quantile regression. *Nature Methods*, 16(6):0–9, 2019.

[7] Cristina Davino, Marilena Furno, and Domenico Vistocco. *Quantile regression: theory and applications*, volume 988. John Wiley & Sons, 2013.

[8] FY Edgeworth. On the mathematical representation of statistical data. *Journal of the Royal Statistical Society*, 79(4):455–500, 1916.

[9] Bernd Fitzenberger, Roger Koenker, and José AF Machado. *Economic applications of quantile regression*. Springer Science & Business Media, 2013.

[10] Lingxin Hao and Daniel Q Naiman. *Quantile regression*. Number 149. Sage, 2007.

[11] Q Huang, H Zhang, J Chen, and MJJBB He. Quantile regression models and their applications: A review. *Journal of Biometrics & Biostatistics*, 8(3):1–6, 2017.

[12] Roger Koenker. Quantile regression. *Cambridge Univ Pr*, 2005.

[13] Roger Koenker. Censored quantile regression redux. *Journal of Statistical Software*, 27:1–25, 2008.

[14] Roger Koenker. Quantile regression: 40 years on. *Annual review of economics*, 9(1):155–176, 2017.

[15] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[16] Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

[17] Roger Koenker and Jose AF Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94(448):1296–1310, 1999.

[18] Roger Koenker and Zhijie Xiao. Inference on the quantile regression process. *Econometrica*, 70(4):1583–1612, 2002.

[19] Andy H Lee, Wing K Fung, and Bo Fu. Analyzing hospital length of stay: mean or median regression? *Medical care*, 41(5):681–686, 2003.

[20] Quantile Regression. *Handbook of quantile regression*. CRC Press: Boca Raton, FL, USA, 2017.

[21] Pavel Škrabánek, Jaroslav Marek, and Alena Pozdílková. Boscovich fuzzy regression line. *Mathematics*, 9(6):685, 2021.

[22] Elisabeth Waldmann. Quantile regression: a short story on how and why. *Statistical Modelling*, 18(3-4):203–218, 2018.

[23] Keming Yu, Zudi Lu, and Julian Stander. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society Series D: The Statistician*, 52(3):331–350, 2003.