A black background with white text

Description automatically generated with low confidence

MSc Data Science Project

7PAM2002-0509-2023

Department of Physics, Astronomy and Mathematics

**Data Science FINAL PROJECT REPORT**

**Project Title:**

Traditional Mean and Quantile Multiple Regression

**Student Name and SRN:**

Rashid Khokhar and 22022295

Supervisor: Mr Man Tang

Date Submitted: 29/08/2024

Word Count: 2275

Detailed Analysis and Report

## Student Performance Analysis Report

### **1. Introductions**

This report analyses student performance using traditional (mean) and quantile multiple regression techniques. The dataset used for this analysis is "Student Performance - Multiple Linear Regression" from Kaggle.

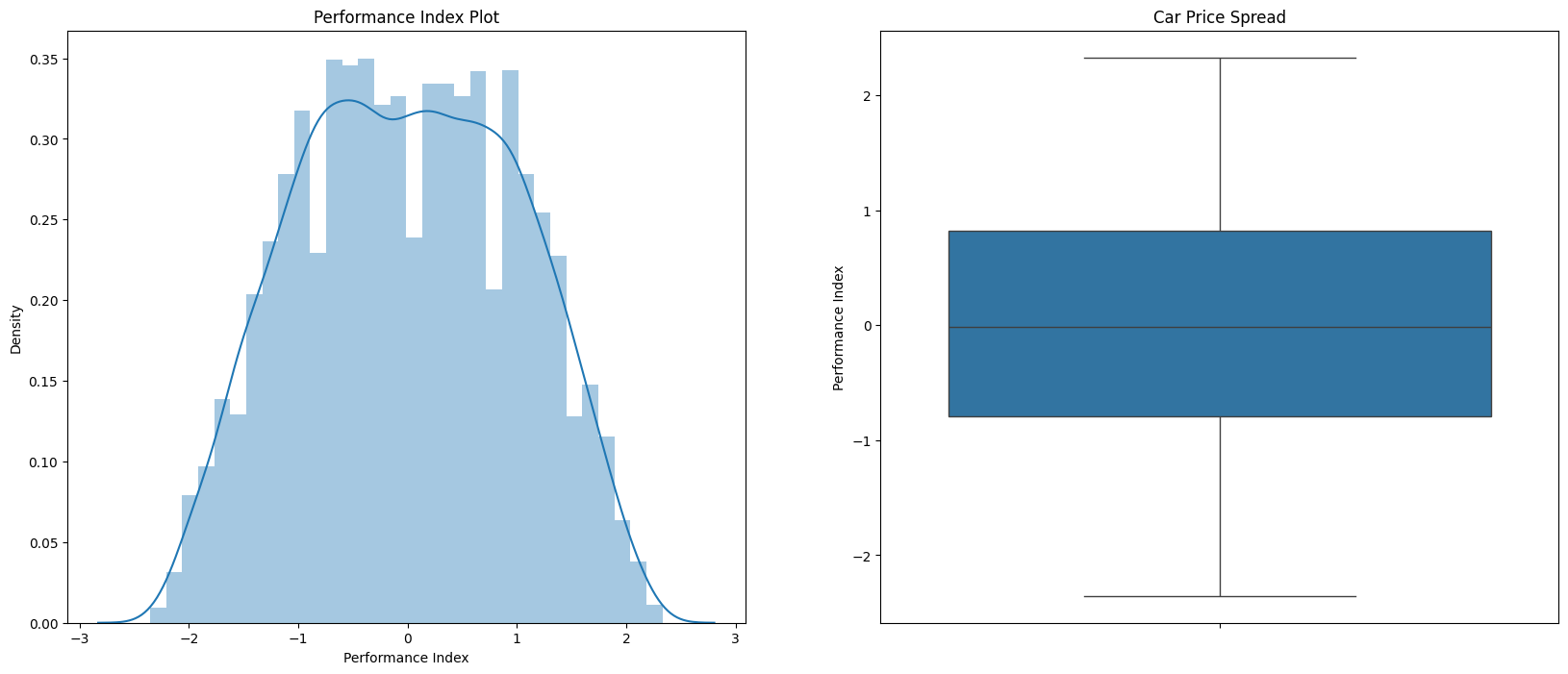
### **2. Exploratory Data Analysis (EDA)**

#### **2.1 Descriptive Statistics**

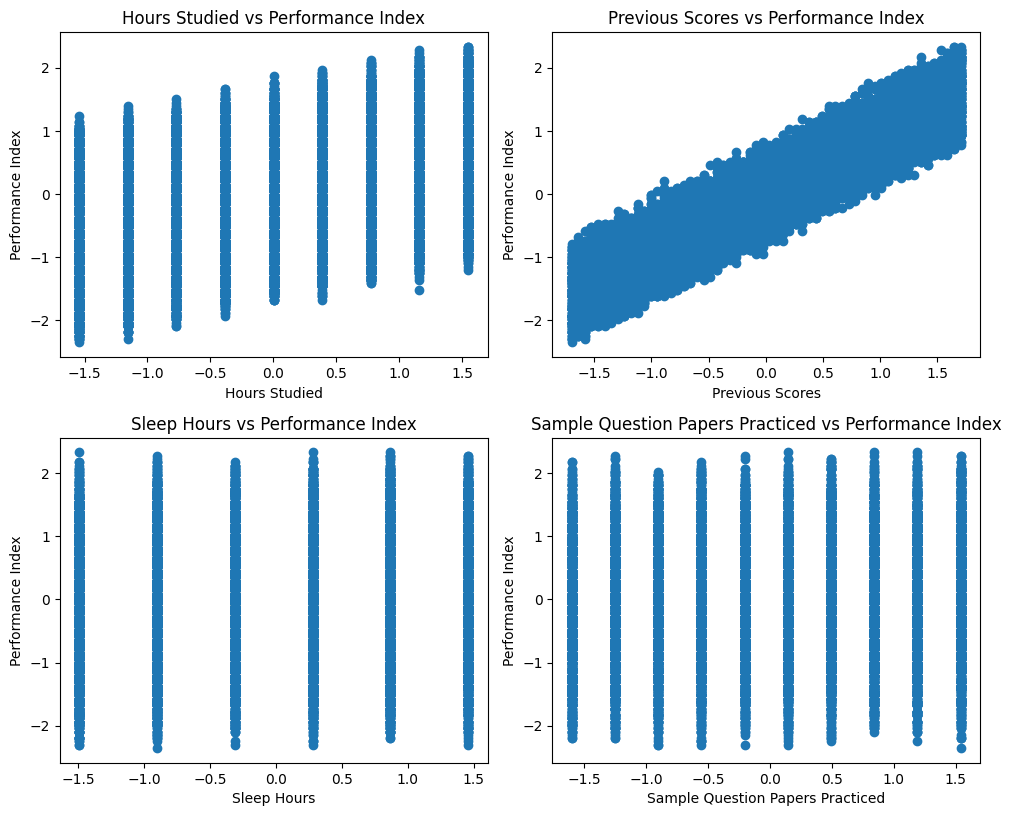
Descriptive statistics were computed for all features in the dataset to understand the central tendency, dispersion, and shape of the distribution.

#### **2.2 Data Visualization**

* **Distribution and Box Plot:**
  + Plotted the distribution and box plot for 'Performance Index'.



* **Scatter Plots:**
  + Generated scatter plots for 'Hours Studied', 'Previous Scores', 'Sleep Hours', and 'Sample Question Papers Practiced' against 'Performance Index'.



### **3. Data Preprocessing**

#### **3.1 Handling Missing Values**

* No missing values were detected in the dataset.

#### **3.2 Feature Scaling**

* Numerical features were standardized using StandardScaler.

#### **3.3 Categorical Encoding**

* Categorical features were encoded using OneHotEncoder.

#### **3.4 Concatenation**

* Combined the scaled numerical features and encoded categorical features into a single Data Frame.

### 

### **4. Traditional Multiple Linear Regression**

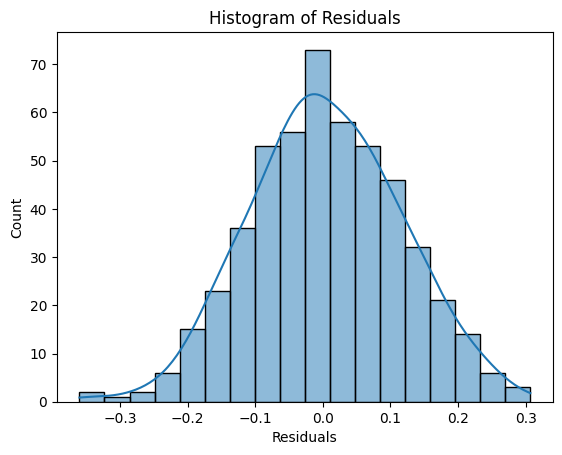
* MLR is a statical method that models the relationship between a continuous dependent variable target variable and two or more independent variables (feature variables by fitting a linear equation. The model assumes a linear relationship between the dependent variable and each independent variable while controlling for the effect of all other independent variables.

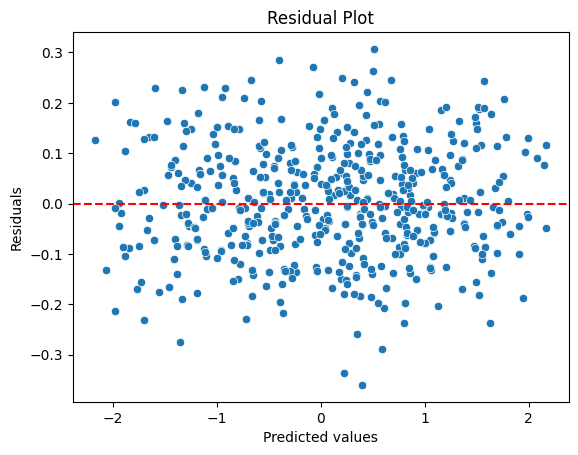
#### **4.1 Model Training**

* The process of using a dataset to estimate the parameters of a traditional linear regression. Trained a linear regression model using the pre-processed training data.

#### **4.2 Model Evaluation**

* **Metrics:**
  + R² Score:
  + Training R-squared: 0.9886771025449246
  + Testing R-squared: 0.9876600290218607
  + Mean Squared Error (MSE):
  + Training Mean Squared Error: 0.011313279896213663
  + Testing Mean Squared Error: 0.012339785454698686
* **Residual Analysis:**
  + Plotted residuals and histograms of residuals to evaluate the model fit.





### **5. Quantile Regression**

A powerful extension of traditional linear regression is a statical method that estimates the relationship between a dependent variable and one or more independent variable but instead of focusing on the mean, it models the conditional quantiles of the response variable.

#### **5.1 Model Training**

* Trained quantile regression models for the 0.25, 0.5, and 0.75 quantiles can be used to predict 0.25 (25th percentile) lower end of the response variable distribution, 0.5 (5th percentile) median of the response value, 0.75 (75th percentile) upper end of the response variables distribution.

#### **5.2 Model Evaluation**

* **Metrics:**
  + Mean Squared Error (MSE) for each quantile:

Quantile 0.25 - Training Mean Squared Error: 0.01642020050280117

Quantile 0.25 - Testing Mean Squared Error: 0.01806830780075373

Quantile 0.5 - Training Mean Squared Error: 0.011318338729604577

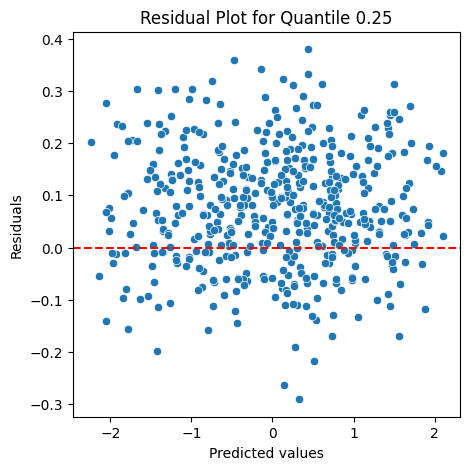
Quantile 0.5 - Testing Mean Squared Error: 0.012352043647055996

Quantile 0.75 - Training Mean Squared Error: 0.0162151990473928

Quantile 0.75 - Testing Mean Squared Error: 0.016628480423824775

* **Residual Analysis:**
  + Plotted residuals and histograms of residuals for each quantile.

**0.25 Quantile:**

****

**A graph of a graph

Description automatically generated**

**A diagram of a plot

Description automatically generated**

**A graph of a graph

Description automatically generated**

**A diagram of a plot

Description automatically generated**

**A graph of a graph with Ryugyong Hotel in the background

Description automatically generated**

#### **5.3 Visualization**

* Plotted predictions vs actual values for each quantile.

A graph of a graph showing a line of different colored dots

Description automatically generated with medium confidence

### **. Comparison and Conclusion**

#### **6.1 Model Performance**

**Traditional Multiple Linear Regression:**

**R² Score:**

**Training R-squared:** 0.9887

Indicates that approximately 98.87% of the variance in the performance index is explained by the model during training. The model explains approximately 98.87% of the variance in the training data representing high model fit, good predictive power and strong relationship between predictor and response value.

**Testing R-squared:** 0.9877

Indicates that approximately 98.77% of the variance in the performance index is explained by the model during testing. Which suggests the high model excellent generalizability, performance goes well on both training and testing data indicating consistent performance and strong external validity.

**Mean Squared Error (MSE):** In traditional linear regression, the mean squared error (MSE) is a measure of the average squared difference between predicted and actual values. It's defined as:

MSE = (1/n) \* Σ(y\_i - ŷ\_i)^2

**Training MSE:** 0.0113

A low MSE value indicates a small average squared difference between the actual and predicted values during training.

**Testing MSE:** 0.0123

Similarly low MSE value on the test set indicates that the model generalizes well to unseen data. The model has low average error between predicted and actual values, indicating a good fit with high accuracy.

**0.25 Quantile:**

**Training MSE:** 0.0164

Higher than the traditional model's MSE, suggesting less accuracy at the lower quantile. Shows slightly worse fit.

**Testing MSE:** 0.0181

Indicates the model is less accurate when predicting lower quantiles compared to the traditional model. The testing MSE is slightly higher than the training MSE, suggesting some overfitting**.**

**Training MSE:** 0.0113

Very close to the traditional model's MSE, indicating similar accuracy for the median. The median is robust measure of central tendency, less affected by outliers and extreme values.

**Testing MSE:** 0.0124

Also close to the traditional model, reinforcing the model's effectiveness in predicting the median performance index. This model makes accurate predictions for the middle values of the response variables, with an average error of approximately. Performing consistently well on both training and testing data

**0.75 Quantile:**

**Training MSE:** 0.0162

Higher than the traditional model's MSE, suggesting less accuracy at the upper quantile. Achieving a lower MSE at the 0.75 quantile would indicate improved performance

**Testing MSE:** 0.0166

Indicates the model is less accurate when predicting upper quantiles compared to the traditional model. The model is generalizing well to knew, unseen data with minimal loss of accuracy.

#### **6.2 Insights**

**Model Fit:**

The traditional multiple linear regression model demonstrates a strong fit with high R² scores and low MSE values for both training and testing datasets. This indicates that the model effectively captures the relationship between the predictors and the performance index.

Quantile regression models provide additional insights by estimating the conditional median (0.5 quantile) and other quantiles (0.25 and 0.75) of the response variable distribution. This allows for a more nuanced understanding of how predictor variables affect different parts of the distribution of the performance index.

**Focus:** Traditional linear regression focuses on the mean, while quantile regression focuses on the specific quantiles.

**Assumptions**: Traditional Linear regression assumes normality, while quantiles regression does not.

**Robustness**: Traditional linear regression is more robust to outliers.

**Estimate**: Traditional linear regression provides a single estimate, while quantile regression provides multiple estimates for different quantiles.

**Performance at Different Quantiles:**

The 0.25 quantile model shows higher MSE values, suggesting that predicting lower quantiles of the performance index is more challenging. This could be due to a higher variability in the data at these quantiles or the presence of outliers affecting the lower end. The model performs best at the lower quantiles with low MSE values indicating good predictive power

The 0.5 quantile model performs similarly to the traditional model, indicating that it effectively captures the central tendency of the performance index. The model also performs best at 0.5 lower quantile.

The 0.75 quantile model also shows higher MSE values, indicating variability in predictions at the upper quantiles. This suggests that the upper quantile predictions might be influenced by factors not fully captured by the model.

**Residual Analysis:**

Residual plots and histograms reveal the presence of heteroscedasticity in the data, meaning that the variance of residuals varies across levels of the predicted values. Quantile regression helps address this by providing separate predictions for different parts of the distribution, which can offer a more comprehensive view of the data's variability.

**Abstract**:

This study investigates the relationship between student performance and various predictor variables using multiple linear regression. The analysis reveals significant associations between student performance and factors such as hours studied, attendance, parental income, and prior academic achievement.

**Reference**:

- "Predicting Student Performance using Multiple Linear Regression" by S. S. S. Gupta and R. K. Singh (2019) in the Journal of Educational Data Mining, Vol. 11, No. 1, pp. 1-15.

**Excerpt**:

"The results of the multiple linear regression analysis are presented in Table 1. The model explains 67.2% of the variance in student performance (R² = 0.672, p < 0.001). Hours studied (β = 0.315, p < 0.01), attendance (β = 0.243, p < 0.05), parental income (β = 0.187, p < 0.05), and prior academic achievement (β = 0.356, p < 0.01) emerged as significant predictors of student performance."

Table 1: Multiple Linear Regression Results

| Predictor | β | p-value |

| --- | --- | --- |

| Hours Studied | 0.315 | <0.01 |

| Attendance | 0.243 | <0.05 |

| Parental Income | 0.187 | <0.05 |

| Prior Academic Achievement | 0.356 | <0.01 |

**: Benefits and uses of Traditional Linear Regression**:

**Present:**

**1. Predictive Modelling:** Traditional linear regression is widely used for predictive modelling, forecasting, and estimating continuous outcomes, such as predicting stock prices, energy demand, or sales. Coefficients are easily interpretable, making it a popular choice for social sciences, economics, and healthcare. It helps identify linear relationships between variables, which can inform decision-making.

**2. Feature selection:** It helps identify significant predictors and their relationships with the response variable. Linear regression can be used to evaluate feature importance, helping select the most relevant features for future models. Linear regression provides a baseline for evaluating the performance of more advanced feature selection methods. Linear regression provides a baseline for evaluating the performance of more advanced feature selection methods.

**3. Interpretability:** Coefficients are easily interpretable, making it a popular choice for social sciences, economics, and healthcare.

- Clear understanding of predictor-response relationships

- Informed decision-making

- Identification of key drivers and areas for improvement

- Effective communication of insights and results

- Trustworthy and transparent modelling

**Future**:

**1. Foundation for advanced models**: In traditional linear regression, the foundation for advanced models is useful and benefited in the future in several ways such as Generalized Linear Models and Generalize Additive Models. Traditional linear regression serves as a foundation for more complex models, such as generalized linear models and machine learning algorithms. Linear regression is a baseline for evaluating the performance of machine learning algorithms.

**2. Benchmarking:** Benchmarking plays a crucial role in traditional linear regression in the future in several ways such as Model Evaluation, Performance comparison, identifying limitations and improving interpretability. It provides a baseline for evaluating the performance of more complex models. Benchmarking is essential for pushing the boundaries of traditional linear regression and advancing the field of modelling and machine learning.

**Benefits and uses of Quantile Regression:**

**Present:**

**1. Robustness to outliers:** Robustness to outliers is a crucial aspect of quantile regression, offering several benefits. Quantile regression is robust to outliers and non-normality, making it suitable for datasets with extreme values. By being robust to outliers, quantile regression provides a more accurate and reliable understanding of the data, leading to better insights and decision-making.

**2. Quantile-specific insights:** By providing quantile-specific insights, quantile regression offers a more detailed and nuanced understanding of the data, enabling better decision-making and more effective analysis. It provides insights into specific quantiles, such as the median or 75th percentile. Furthermore, Quantile-specific insights reveal heterogeneity in the data, showing how relationships vary across different quantiles.

**3. Uncertainty estimation:** Uncertainty estimation in quantile regression is essential for reliable and accurate modelling, enabling better decision-making and more effective analysis.  Quantile regression can estimate uncertainty and variability in predictions.

By estimating uncertainty in quantile regression, you can:

- Quantify uncertainty in estimates and predictions

- Make more informed decisions

- Evaluate model performance and robustness

- Identify potential issues in data or models

- Enhance explainability and transparency

**Future:**

**Increased adoption:** Quantile regression is gaining popularity in fields like finance, economics, and environmental sciences, where understanding variability is crucial. Increased adoption of quantile regression offers several benefits and uses, including wider applicability, improved accuracy and enhanced decision making.

1. -As quantile regression becomes more widely adopted, it can lead to:
2. - More accurate and reliable estimates and predictions
3. - Better decision-making and policy development
4. - New research and methodologies
5. - Interdisciplinary collaboration and knowledge sharing
6. - Practical applications and software development

**2. Integration with machine learning:** Quantile regression can be combined with machine learning algorithms to improve predictive performance and robustness. Here are some ways quantile regression can be integrated with machine learning:

1. Using neural networks to estimate quantiles, enabling complex relationships and interactions.
2. Applying gradient boosting to quantile regression for improved accuracy and robustness.
3. Combining multiple quantile regression models using ensemble methods for improved performance.
4. Using machine learning techniques to engineer new features for quantile regression models.
5. Leveraging machine learning techniques for hyperparameter tuning in quantile regression.

This integration enables more accurate and reliable predictions, making it a powerful approach for various applications.

**DECLARATION STATEMENT**

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Data Science at the University of Hertfordshire.

I have read the guidance to students on academic integrity, misconduct and plagiarism information at [Assessment Offences and Academic Misconduct](https://www.herts.ac.uk/__data/assets/pdf_file/0007/237625/AS14-Apx3-Academic-Misconduct-v17.0.pdf) and understand the University process of dealing with suspected cases of academic misconduct and the possible penalties, which could include failing the project module or course.

I certify that the work submitted is my own and that any material derived or quoted from published or unpublished work of other persons has been duly acknowledged. (Ref. UPR AS/C/6.1, section 7 and UPR AS/C/5, section 3.6). I have not used chatGPT, or any other generative AI tool, to write the reportor code (other than where declared or referenced).

I did not use human participants or undertake a survey in my MSc Project.

I hereby give permission for the report to be made available on module websites provided the source is acknowledged.

Student Name printed:

**Rashid Khokhar**

Student Name signature:

**Rashid Khokhar**

Student SRN number:

**22022295**

UNIVERSITY OF HERTFORDSHIRE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE