# QUANTILE REGRESSION WITH Q1/Q3 ANCHORING: A ROBUST ALTERNATIVE FOR OUTLIER-RESISTANT MODELING

Xiaoying Zeng [1] and Eugene Pinsky [2]

[1] Faculty of Computing & Data Sciences, Boston University
[2] Department of Computer Science, Metropolitan College, Boston University

### ABSTRACT

*Outliers introduce considerable difficulties in statistical modeling and regression analysis by skewing parameter estimates and reducing model reliability. To mitigate these effects, we introduce an enhanced Quantile Regression (QR) framework that strategically incorporates the 25th (Q1) and 75th (Q3) percentiles of the target variable. By emphasizing these robust statistical markers, our approach effectively minimizes the influence of extreme values while preserving the underlying data structure. Through comprehensive evaluations across multiple datasets, including Iris, Fish, Advertising Budget and Sales, and Geyser, we demonstrate that this method consistently delivers stable and accurate predictions. The experimental results further highlight the superior resilience of QR compared to conventional Linear Regression (LR), particularly in handling datasets affected by noise and outliers.*

### KEYWORDS

*Quantile Regression, Linear Regression, Outliers, Statistical Modeling, Comparative analyses*

## 1. INTRODUCTION

Linear regression is a fundamental tool in statistical modeling, widely applied across disciplines such as economics, biology, and machine learning. However, traditional least squares (L2) regression is highly sensitive to outliers, as it minimizes the sum of squared residuals. In real-world datasets, the presence of noise and extreme values can lead to biased parameter estimates and degraded model performance. This limitation has motivated extensive research into robust regression techniques that mitigate the influence of outliers while preserving the core structure of the data.

To address this issue, Koenker and Bassett [1] introduced *Quantile Regression (QR)*, which estimates conditional quantiles instead of the conditional mean. Unlike classical regression models that assume normally distributed residuals with constant variance, Quantile Regression provides a more flexible framework that accounts for heteroske dasticity and skewed distributions. Later, Buchinsky [2] demonstrated that Quantile Regression remains stable under heavy-tailed distributions and extreme values, reinforcing its effectiveness in robust statistical modeling.

Over the years, Quantile Regression has been enhanced by integrating machine learning techniques. Meinshausen [3] developed *Quantile Regression Forests (QRF)*, which extends Quantile Regression by leveraging decision trees to model conditional quantiles non-parametrically. Similarly, Takeuchi [4] introduced *Quantile Regression Support Vector Machines (QR-SVM)*, which improves Quantile Regression's ability to capture nonlinear dependencies in high-dimensional spaces. More recently, Delcroix et al. [5] demonstrated the effectiveness of

*Quantile Regression Gradient Boosting Decision Trees (QRGBDT)* in modeling heteroskedastic and non-normal datasets. In another significant advancement, Gu and Zuo [6] proposed *Sparse Composite Quan- tile Regression (SCQR)*, which optimizes Quantile Regression for high-dimensional datasets by incorporating sparsity constraints and feature selection.

Despite these advances, most QR-based methods estimate multiple quantiles independently, leading to potential inconsistencies such as quantile crossing. Moreover, many existing QR techniques require fitting multiple models across different quantiles, making them computationally expensive for large datasets.

In this paper, we introduce a *Quantile Regression (QR)* technique that prioritizes key data points around the 25th (Q1) and 75th (Q3) percentiles of the response variable. By selectively leveraging these representative subsets, Quantile Regression provides a robust alternative to traditional regression methods, offering enhanced stability and reliability in noisy and messy datasets.

Our contributions are as follows:

1. Novel Regression Technique: We propose a Quantile Regression method that selectively uses data near Q1 and Q3 to minimize the impact of outliers.
2. Robustness and Stability: We demonstrate that Quantile Regression consistently outperforms Linear regression across multiple datasets and varying levels of noise and outlier fractions.
3. Extensive Analysis: We evaluate Quantile Regression on diverse datasets (*Iris, Fish, Advertising Budget and Sales, and Geyser*) and provide empirical evidence of its effectiveness in real-world applications.

Building on the theoretical foundations of Quantile Regression while addressing its computational and interpretational limitations, our approach provides an efficient and robust alternative to existing regression techniques.

# 2. METHODOLOGY

## 2.1. Overview

The Quantile Regression (QR) method is designed to achieve robust modeling by focusing on representative subsets of data. Unlike traditional Linear Regression, which uses all data points equally, Quantile Regression employs a selective approach by identifying and incorporating data points closest to the Q1 and Q3 percentiles of the response variable.

## 2.2. Steps

### 2.2.1. Data Loading and Splitting

Select two features, one as the independent variable □ and the other as the dependent variable □. The dataset is split into a training set (80%) and a test set (20%).

### 2.2.2. Outlier Generation

To introduce controlled levels of outliers into the training data, we employ a systematic approach based on the inter quartile range (IQR). The first step involves computing the quartiles (□1 and □3) and the IQR for both the independent variable □ (Width) and the dependent variable □ (Weight). Observations that exceed □3 + 1.5 × IQR or fall below □1 − 1.5 × IQR are classified as potential outliers.

For each predefined outlier fraction, a corresponding number of outliers is generated and added to the training set. These outliers are created through three distinct mechanisms to simulate different types of contamination in the data. The first type consists of upper-bound outliers, which are drawn from a uniform distribution beyond the threshold $\square 3 + 1.5 \times \mathrm{IQR}$, ensuring that these points lie outside the natural range of the data. Similarly, lower-bound outliers are sampled from a range below $\square 1 - 1.5 \times \mathrm{IQR}$, mimicking extreme low-value deviations. In addition to these structured anomalies, we introduce randomly distributed outliers that span the entire range of $\square$ and $\square$ in the dataset, simulating unpredictable noise that does not necessarily follow the IQR-based pattern.

Once generated, these outliers are combined with the original training data and shuffled to avoid order bias. The dataset, now containing both original observations and injected outliers, serves as the foundation for training both Quantile Regression (QR) and Linear Regression (LR). To maintain reproducibility and ensure controlled variations in outlier placement, different random seeds are assigned to different outlier fractions.

For the Fish dataset in Horizontal Comparison, different random seeds were assigned to each outlier fraction to introduce controlled variations while ensuring reproducibility. The mapping between outlier fractions and their respective random seeds is as follows: outlier fraction of 0.0 used seed 0, 0.1 used seed 3, 0.2 used seed 2, 0.3 used seed 0, 0.4 used seed 2, 0.5 used seed 2, 0.6 used seed 6, 0.7 used seed 1, and 0.8 used seed 0. In the Vertical Comparison,

In the Vertical Comparison, different random seeds were assigned to each dataset to ensure controlled variations while maintaining reproducibility. For an outlier fraction of 0.3, the Iris dataset used a random seed of 4, the Fish dataset used seed 0, the Advertising Budget and Sales dataset was assigned seed 4, and the Geyser dataset utilized seed 3. For an outlier fraction of 0.0, a similar approach was followed to ensure consistency in the baseline comparisons. The Iris dataset maintained a random seed of 4, while the Fish dataset, Advertising Budget and Sales dataset, and Geyser dataset all used seed 0.

#### 2.2.3. Data Sorting and Percentile Calculation

The training data, including outliers, is first sorted based on $\square$. Then, the quartiles ($\square 1$, median, and $\square 3$) for $\square$ are computed. Finally, the corresponding $\square$ values are identified based on the $\square$ values closest to these quartiles.

#### 2.2.4. Key Point Selection

The $\square$ values closest to $\square 1$, the median, and $\square 3$ are identified, along with their corresponding $\square$ values ($\square_1$, $\square_2$, and $\square_3$). If multiple points are equally close to a quartile, the mode function is used to select the most frequent $\square$ value, ensuring consistency and avoiding ambiguity.

#### 2.2.5. Linear Model Fitting

A line is fitted between the points ($\square_1$, $\square 1$) and ($\square_3$, $\square 3$). The slope $\square$ and intercept $\square$ of this line are then calculated.

#### 2.2.6. Model Evaluation

The $\square$ values are predicted on the test set, and the model's performance is evaluated using three key metrics: the Mean Squared Error (MSE), which quantifies the average squared differences between predicted and actual values; the Mean Absolute Error (MAE), which measures the average absolute differences; and the Coefficient of Determination ($\square^2$), which assesses the explanatory power of the model.

#### 2.2.7. Comparison with Traditional Linear Regression

A traditional linear regression model is fitted to the test set, and the same evaluation metrics (□ □□ and $\square^2$) are computed. The performance of Quantile Regression (QR) is then compared with that of traditional regression, with a particular focus on robustness to outliers.

#### 2.2.8. Data Visualization

The data distribution, identified outliers, and the fitted regression line are plotted for visualization. Key points are highlighted using different colors and annotations to emphasize their significance in the regression model.

## 3. EXPERIMENTAL DESIGN

### 3.1. Datasets

We evaluated the performance of Quantile Regression using multiple datasets:

#### 3.1.1. Iris Dataset [7]

This dataset consists of 150 observations across three species of iris flowers (setosa, versicolor, and virginica). Each observation includes four features: sepal length, sepal width, petal length, and petal width. For this study, we selected petal length as the independent variable (X) and petal width as the dependent variable (y).

#### 3.1.2. Fish Dataset [8]

This dataset includes 159 observations of fish across 7 species, with measurements capturing physical traits such as weight, length, height, and width. In our analysis, Width was used as the independent variable (X) and Weightas the dependent variable (y).

#### 3.1.3. Advertising Budget and Sales Dataset [9]

This dataset contains 200 observations of advertising budgets allocated to three media channels (*TV*, *Radio*, and *Newspaper*) and their impact on sales revenue. For the analysis, Radio Ad Budget ($) served as the independent variable (X), while Sales ($) was the dependent variable (y).

#### 3.1.4. Geyser Dataset [10]

The Geyser dataset records 272 observations of waiting times between eruptions and the corresponding eruption durations of the Old Faithful geyser. For this study, waiting (time between eruptions) was used as the independent variable (X) and duration (eruption duration) as the dependent variable (y).

### 3.2. Evaluation Metrics

We compared Quantile Regression against Linear Regression using metrics such as: Mean Absolute

Error (MAE) and $R^2$.

### 3.3. Experimental Setup

We simulated varying levels of outlier fractions (0.1 to 0.8). Regression lines were fitted using both QR and LR, and the results were analyzed under different scenarios.

## 4. RESULTS

### 4.1. Horizontal Comparison: Fish Dataset Across Outlier Fractions

Table 1: Comparison of Quantile Regression and Linear Regression performance across varying outlier fractions for the Fish dataset.

| Outlier Fraction | MAE (QR) | $R^2$ (QR) | MAE (LR) | $R^2$ (LR) |
|---|---|---|---|---|
| 0.0 | 240.82 | 0.60 | 250.62 | 0.57 |
| 0.1 | 207.70 | 0.71 | 231.52 | 0.59 |
| 0.2 | 300.50 | 0.46 | 236.70 | 0.59 |
| 0.3 | 195.84 | 0.67 | 211.35 | 0.63 |
| 0.4 | 196.45 | 0.68 | 229.60 | 0.60 |
| 0.5 | 195.18 | 0.67 | 208.93 | 0.63 |
| 0.6 | 207.36 | 0.70 | 213.29 | 0.63 |
| 0.7 | 198.92 | 0.66 | 200.71 | 0.66 |
| 0.8 | 169.20 | 0.76 | 198.83 | 0.66 |

The following figures demonstrate the Quantile Regression results using the Fish dataset, with outlier fractions varying from 0.0 to 0.8. These visualizations highlight the model's robustness against outliers and its ability to focus on key quantile points.

The dataset's primary points, represented as blue dots, show natural variability but do not include artificial noise or outliers when the fraction is 0.0. Artificial outliers, shown as pink dots, are incrementally added with fractions ranging from 0.1 to 0.8 to evaluate the model's performance under varying levels of data contamination.

The Quantile Regression line, displayed in green, captures the trend of the data by leveraging key quantile points, specifically the lower quartile ($Q1$), the median, and the upper quartile ($Q3$). These points are marked respectively as purple, brown, and cyan dots in the visualizations. By focusing on these quantiles, the model effectively ignores the influence of extreme values introduced as outliers, ensuring a robust regression line that represents the core data trend.

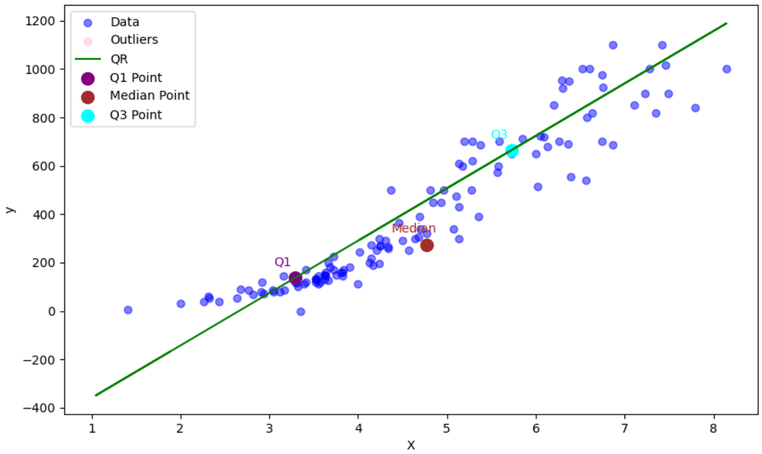
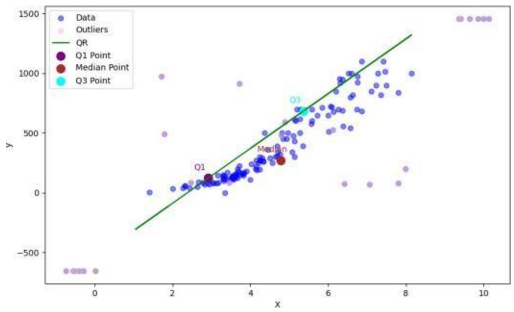
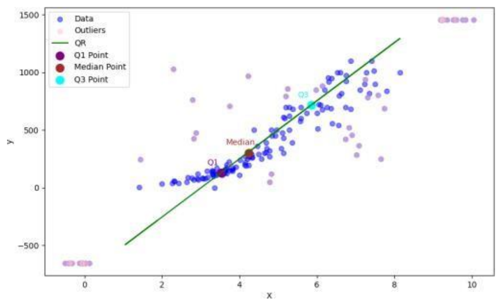



Figure 1: Regression results for the Fish dataset with outlier fraction = 0.0.

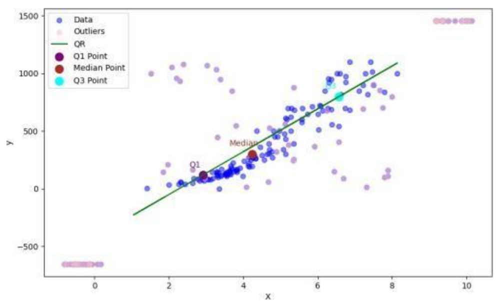
As shown in Figure 1, without any outliers, the Quantile Regression (QR) line closely aligns with the core data trend. This highlights the standard performance of Quantile Regression, effectively capturing $Q1$, Median, and $Q3$ points in a noise-free scenario. From Table 1, for an outlier fraction of 0.0, QR achieves a MAE of 240.82 and an $R^2$ value of 0.60, showing slightly better performance in terms of MAE compared to Linear Regression (LR), which has MAE = 250.62 and $R^2$ = 0.57.



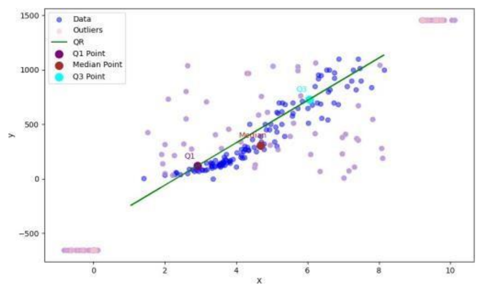
(a) Outlier Fraction = 0.1



(b) Outlier Fraction = 0.2



(c) Outlier Fraction = 0.3

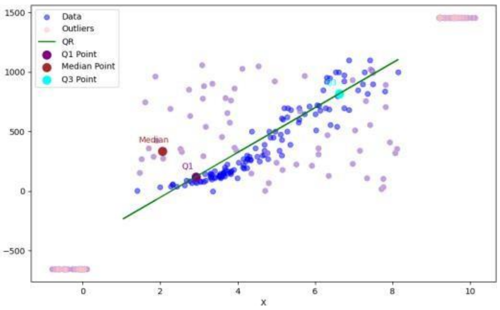


(d) Outlier Fraction = 0.4

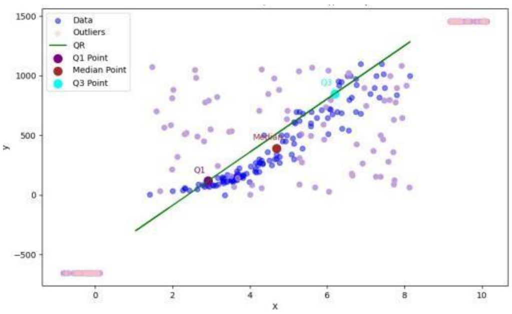
Figure 2: Regression results for the Fish dataset under outlier fractions from 0.1 to 0.4.

Figure 2 illustrates QR's performance as outlier fractions increase from 0.1 to 0.4. At lower fractions, such as 0.1 (Figure 2a) and 0.2 (Figure 2b), QR remains robust, with its regression line closely aligned to the quantile points ($Q1$, Median, $Q3$). This stability is reflected in Table 1, where QR maintains relatively low MAE values (207.70 at 0.1, 300.50 at 0.2) and high $R^2$ values (0.71 at 0.1, 0.46 at 0.2). QR performs comparably to LR, which has fluctuating MAE values (231.52 at 0.1, 236.70 at 0.2) and slightly lower $R^2$ at 0.1 (0.59) but matches QR at 0.2 (0.59).

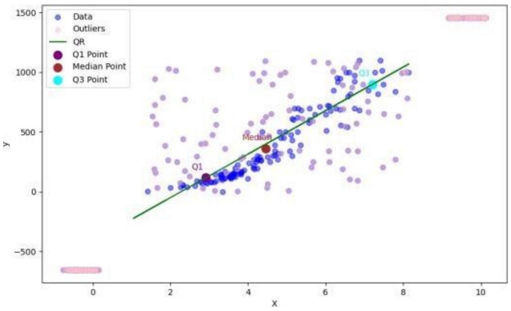
As outlier fractions rise to 0.3 (Figure 2c) and 0.4 (Figure 2d), QR continues to mitigate the influence of extreme values, preserving the underlying data trend effectively. Table 1 supports this observation, with QR showing MAE values of 195.84 at 0.3 and 196.45 at 0.4 while maintaining strong $R^2$ values (0.67 at 0.3, 0.68 at 0.4). Notably, QR outperforms LR in both cases, as LR shows higher MAE (211.35 at 0.3, 229.60 at 0.4) and slightly lower $R^2$ values (0.63 and 0.60, respectively).



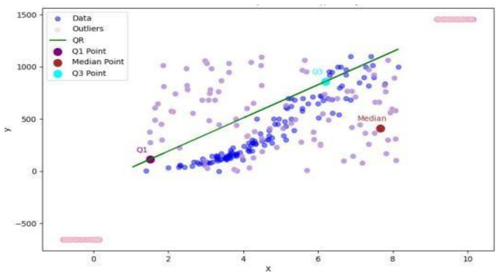

(a) Outlier Fraction = 0.5




(b) Outlier Fraction = 0.6




(c) Outlier Fraction = 0.7




(d) Outlier Fraction = 0.8

Figure 3: Regression results for the Fish dataset under outlier fractions from 0.5 to 0.8.

As shown in Figure 3, QR's performance is analyzed under higher outlier fractions (0.5 to 0.8). For fractions of 0.5 (Figure 3a) and 0.6 (Figure 3b), QR retains its robustness by anchoring the regression line to key quantile points. Table 1 reflects this resilience, with QR maintaining a stable MAE of 195.18 at 0.5 and 207.36 at 0.6, alongside $R^2$ values of 0.67 and 0.70, respectively. Meanwhile, LR struggles more with outliers, showing higher MAE (208.93 at 0.5, 213.29 at 0.6) and consistently lower $R^2$ values (0.63 in both cases).

At fractions of 0.7 (Figure 3c) and 0.8 (Figure 3d), outliers dominate the data, creating significant variability. Despite minor deviations, QR continues to align with the core data trend, as shown in Table 1. QR achieves an MAE of 198.92 and 169.20, and $R^2$ values of 0.66 and 0.76, respectively. Notably, at 0.8, QR slightly outperforms LR in both metrics, with LR showing MAE = 198.83 and $R^2$ = 0.66.

The analysis of the Fish dataset across varying outlier fractions demonstrates the robustness and effectiveness of QR in maintaining alignment with the core data trend. As the fraction of outliers increases, QR consistently focuses on key quantile points ($Q1$, Median, $Q3$), thereby mitigating the impact of extreme values. This resilience is evident in its stable performance metrics, with minimal deviation in MAE compared to LR, which exhibits more sensitivity to outliers. While QR begins to show slight degradation at higher outlier fractions (e.g., 0.7 and 0.8), its performance remains superior to LR, highlighting its utility in scenarios involving outlier-dominated datasets. These findings emphasize the reliability of Quantile Regression as a robust alternative to traditional regression techniques in real-world applications.

## 4.2. Vertical Comparison: Different Datasets

The performance of Quantile Regression is further evaluated across multiple datasets (Iris, Fish, ABS, and Geyser). **Table 2** summarizes the results of Quantile Regression and traditional linear regression models.

Table 2: Comparison of Quantile Regression and Linear Regression across datasets at different outlier fractions (rounded to two decimal places).

| Outlier | Dataset | MAE (QR) | MAE (LR) | $R^2$ (QR) | $R^2$ (LR) |
|---|---|---|---|---|---|
| 0.0 | IRIS | 0.28 | 0.30 | -0.33 | -0.46 |
| | Fish | 240.82 | 250.62 | 0.60 | 0.57 |
| | ABS | 3.43 | 3.26 | 0.40 | 0.43 |
| | Geyser | 0.38 | 0.37 | 0.82 | 0.82 |
| 0.3 | IRIS | 0.56 | 0.29 | -4.09 | -0.37 |
| | Fish | 195.84 | 211.35 | 0.67 | 0.63 |
| | ABS | 3.43 | 3.26 | 0.40 | 0.43 |
| | Geyser | 0.51 | 0.45 | 0.68 | 0.72 |

### 4.2.1. Iris Dataset

In the absence of outliers, as shown in Figure 4, both QR and LR perform similarly in the absence of outliers. Quantile Regression achieves slightly lower MAE (0.28 vs. 0.30) and higher $R^2$ (-0.33 vs. -0.46). However, since both $R^2$ values are negative, it indicates that neither model predicts the data well and simply using the mean would be more accurate. This suggests a weak relationship between the chosen features, making it difficult for either model to capture a meaningful trend.

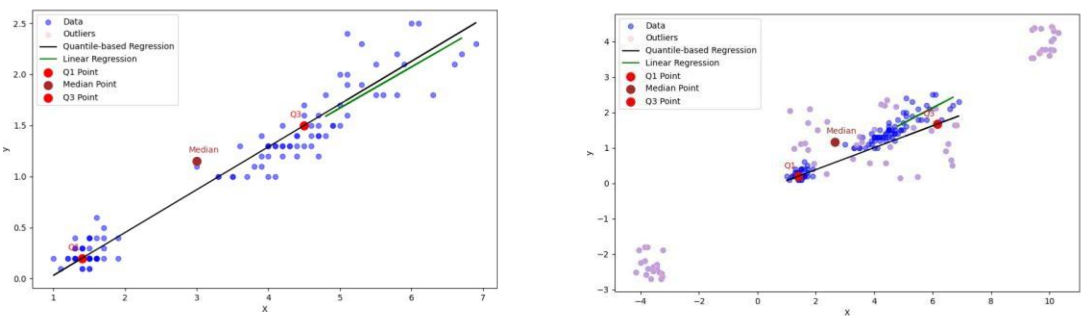



Figure 4: Regression results for Iris dataset: (Left) No Outliers (Outlier Fraction = 0); (Right) With Outliers (Outlier Fraction = 0.3).

When outliers are present at an outlier fraction of 0.3, Figure 4 (Right) shows that Quantile Regression remains aligned with the core data trend, whereas Linear Regression is significantly distorted by the presence of outliers. Table 2 further highlights this difference. While Linear Regression achieves a lower MAE (0.29 vs. 0.56), its $R^2$ value of -0.37 indicates poor predictive power, as the model fits worse than simply using the mean. Quantile Regression, however, has an even lower $R^2$ of -4.09, suggesting that the outliers have severely impacted its ability to generalize in this case. This result implies that although QR is designed to be more robust, extreme contamination in the data can still degrade its performance, particularly when the underlying relationship is already weak.

#### 4.2.2. Fish Dataset

Without any outliers, Figure 5 (Left) illustrates that Quantile Regression achieves slightly better than Linear Regression in terms of MAE (240.82 vs. 250.62) and $R^2$ (0.60 vs. 0.57). Both regression lines closely follow the trend, demonstrating good performance in the absence of noise.

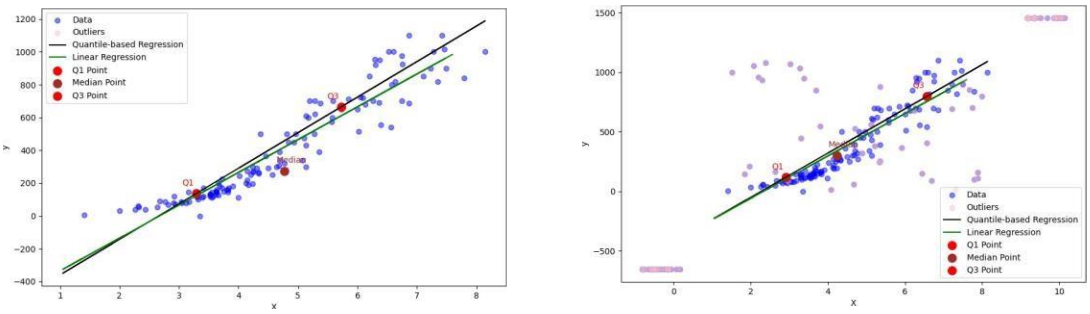



Figure 5: Regression results for Fish dataset: (Left) No Outliers (Outlier Fraction = 0); (Right) With Outliers (Outlier Fraction = 0.3).

When the outlier fraction increases to 0.3, Figure 5 (Right) shows that Quantile Regression remains robust against outliers, whereas Linear Regression line is skewed by extreme values. Table 2 confirms this observation, with Quantile Regression achieving better MAE (195.84 vs. 211.35) and $R^2$ (0.67 vs. 0.63), highlighting its superior performance.

#### 4.2.3. Advertising Budget and Sales (ABS) Dataset

Outlier-Free Data Figure 6 (Left) shows both Quantile Regression and Linear Regression performing comparably, with Linear Regression having a slight edge in MAE (3.26 vs. 3.43) and $R^2$ (0.43 vs. 0.40). This similarity arises from the linear relationship in the dataset.

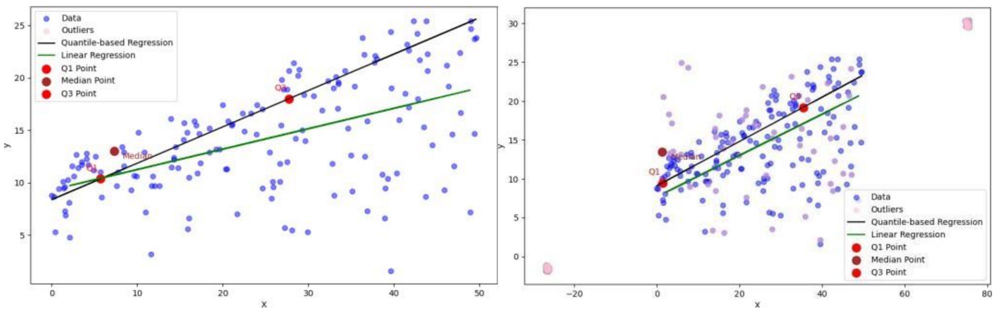



Figure 6: Regression results for ABS dataset: (Left) No Outliers (Outlier Fraction = 0); (Right) With Outliers (Outlier Fraction = 0.3).

Presence of Outliers (Outlier Fraction = 0.3) Figure 6 (Right) shows that Linear Regression performs better in this case, achieving a lower MAE (3.26 vs. 3.43) and a higher $R^2$ (0.43 vs. 0.40) compared to Quantile Regression. This suggests that outliers have a limited impact on Linear Regression for this dataset, making it the better-performing model under these conditions.

#### 4.2.4. Geyser Dataset

Outlier-Free Condition Figure 7 (Left) demonstrates similar performance for Quantile Regression and Linear Regression, with Linear Regression slightly outperforming QR in MAE (0.38 vs. 0.37) and $R^2$ (0.82 vs. 0.82).

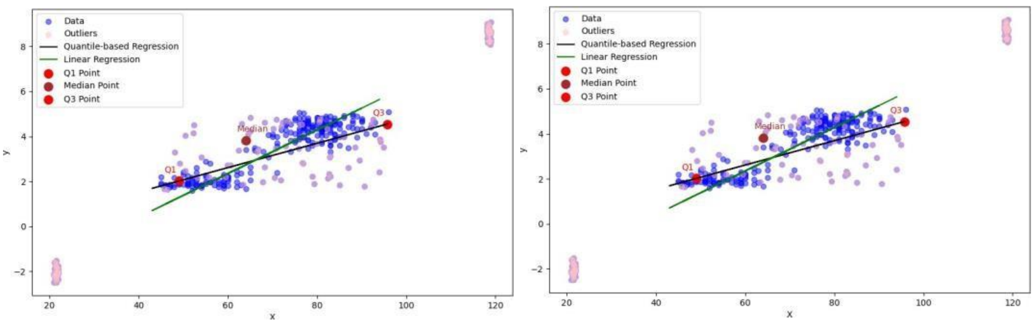



Figure 7: Regression results for Geyser dataset: (Left) No Outliers (Outlier Fraction = 0); (Right) With Outliers (Outlier Fraction = 0.3).

With a 0.3 outlier fraction, Figure 7 (Right) shows that Linear Regression performs slightly better, achieving a lower MAE (0.45 vs. 0.51) and a higher $R^2$ (0.72 vs. 0.68) compared to Quantile Regression. While QR still aligns well with the core data trend, these results indicate that in this case, Linear Regression handles the outliers a little more effectively.

Quantile Regression is more effective in scenarios where extreme outliers significantly impact the data, as seen in the Fish dataset, where it maintains a more reliable trend. In contrast, Linear Regression performs better when outliers have a limited impact, making it the preferred choice for datasets with strong linear relationships, such as ABS and Geyser. However, in datasets with weak predictive relationships, like Iris, neither model performs well, and Quantile Regression may even suffer more under heavy outlier contamination.

## 5. CONCLUSION

Quantile Regression (QR) proves to be a valuable alternative to Linear Regression (LR) in handling datasets with substantial outlier influence. Its ability to focus on key quantile points allows it to maintain robustness in scenarios where extreme values distort traditional regression methods. However, its advantages are context-dependent. When data follows a strong linear structure with minimal outlier impact, as seen in the ABS and Geyser datasets, LR tends to perform better. Conversely, in datasets with weak predictive relationships, like Iris, neither model is particularly effective, and QR may even exhibit greater performance degradation under severe contamination.

Overall, while QR offers significant resilience in outlier-heavy conditions, its effectiveness is reduced when applied to well-structured or weakly correlated data. These findings suggest that selecting between QR and LR should be based on the nature of the dataset, particularly considering the extent of outliers and the underlying relationship between variables.
While the current study focused primarily on low dimensional applications, the underlying

principles of Quantile Regression can be extended to higher-dimensional data. In multidimensional scenarios, quantile-based methods can be adapted to define key points along each dimension or through multivariate analogs such as the Mahalanobis distance for detecting outliers. This opens up opportunities for Quantile Regression to be applied in fields requiring multidimensional modeling, such as image processing, genomics, and multi-factor financial analysis.

## ACKNOWLEDGMENTS

We thank the open data repositories for providing datasets and acknowledge the valuable feedback from peers during the development of this work.

## DECLARATIONS

**Conflict of Interest:** We declare that there are no conflicts of interest regarding the publication of this paper.

**Author Contributions:** All authors contributed equally to the effort.

**Funding:** This research was conducted without any external funding. All aspects of the study, including design, data collection, analysis, and interpretation, were carried out using the resources available within the authors' institution.

**Data Availability (including Appendices):** All the relevant data, Python code for analysis, detailed annual tables and graphs are available via: https://github.com/ArielZeng/ Quantile-Regression

## REFERENCES

[1] R. Koenker and G. Bassett, ”Regression quantiles,” Econometrica, vol. 46, no. 1, pp. 33–50, 1978.
[2] M. Buchinsky, ”Estimating the asymptotic covariance matrix for quantile regression models: A Monte Carlo study,” Journal of Econometrics, vol. 68, no. 2, pp. 303–338, 1995.
[3] N. Meinshausen, ”Quantile regression forests,” Journal of Machine Learning Research, vol. 7, pp. 983–999, 2006.
[4] I. Takeuchi, ”Non-crossing quantile regressions by SVM,” in Proc. IEEE Int. Joint Conf. Neural Netw., vol. 1, pp. 401–406, 2004.
[5] B. Delcroix, S. Sansregret, G. L. Martin, and A. Daoud, ”Quantile regression using gradient boosted decision trees for daily residential energy load disaggregation,” Journal of Physics: Conference Series, vol. 2069, no. 1, pp. 012107, 2021.
[6] Y. Gu and H. Zuo, ”Sparse composite quantile regression in ultrahigh dimensions with tuning parameter calibration,” IEEE Transactions on Information Theory, vol. 66, no. 11, pp. 7132– 7154, 2020.
[7] UCI Machine Learning Repository, ”Iris dataset,” 2024. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Iris.
[8] Kaggle, ”Fish Market Dataset,” 2024. [Online]. Available: https://www.kaggle.com/datasets/vipullrathod/fish-market.
[9] Kaggle, ”Advertising Dataset,” 2024. [Online]. Available: https://www.kaggle.com/datasets/yasserh/advertising-sales-dataset.
[10] National Park Service, ”Old Faithful Geyser,” 2024. [Online]. Available: https://www.kaggle.com/datasets/emmanueldjegou/old-faithful-dataset.

## AUTHORS

**Xiaoying Zeng** is a Learning Facilitator at the Faculty of Computing & Data Sciences, Boston University. Her research focuses on statistical modeling and machine learning.

**Eugene Pinsky** is an associate professor at the Department of Computer Science, Metropolitan College, Boston University. His research interests include pattern recognition, clustering, regression, prediction, factor models, support vector machines, and other machine learning algorithms and methods for data analysis.