Advance Regression Term Project

Gerard Palomo & Juan Pablo Uphoff

Abstract
Linear quantile regression (QR) extends ordinary least squares (OLS) by modeling conditional quantiles, offering a richer view of the response variable's distribution beyond the conditional mean provided by OLS. This paper highlights two key advantages of QR over OLS. Firstly, QR provides robustness to outliers in the response variable, a significant limitation for OLS which relies on minimizing squared errors. Secondly, QR allows for modeling distributional heterogeneity, such as heteroscedasticity, which OLS inherently overlooks by focusing solely on the mean. We compare the performance of OLS and linear QR estimators through a simulation study in a univariate setting. We examine performance under three conditions: (1) a baseline scenario with homoscedastic Gaussian errors, (2) contamination where outliers are introduced at high-leverage points to assess robustness, and (3) heteroscedastic errors to assess distributional modeling capabilities. We specifically compare OLS with QR at the 0.1, 0.5 (median), and 0.9 quantiles. Key performance metrics, including Mean Absolute Error (MAE) and coefficient stability, are assessed. Results demonstrate that QR estimates, particularly the median ($\tau = 0.5$), remain reliable under outlier contamination where OLS estimates become significantly biased. Furthermore, QR effectively captures distributional effects like heteroscedasticity, providing quantile-specific insights that OLS cannot. All simulation code is provided in R for reproducibility.

Keywords
Quantile Regression, Robustness, Heteroscedasticity, OLS, Outliers

C15, C21

This report was completed for the course **Advanced Regression and Prediction**, as part of the **MSc in Statistics for Data Science** at **University Carlos III of Madrid**.

*:

Contents

1. Introduction

Classical linear regression, estimated via ordinary least squares (OLS), focuses on modeling the conditional mean of a response variable by minimizing the sum of squared residuals. While powerful under ideal assumptions, OLS faces significant limitations in practice. Its reliance on squared errors renders estimates highly sensitive to outliers, potentially leading to biased results. Furthermore, OLS provides only a partial view of the conditional distribution by focusing solely on the central tendency and typically assumes homoscedastic errors, limiting its ability to describe relationships where the variability of the response changes with predictors.

Quantile regression (QR), introduced by Koenker and Bassett (1978), offers a more comprehensive and robust alternative. By modeling conditional quantiles (e.g., the median, quartiles, deciles), QR addresses the shortcomings of OLS in two crucial ways. First, its estimation, based on minimizing an asymmetrically weighted sum of absolute errors (the check loss function), provides inherent robustness against outliers in the response variable; the influence of extreme observations is bounded, unlike in OLS. Prior studies, such as John (2015), have confirmed the superior performance of quantile regression over OLS in the presence of outliers. Second, QR provides a mechanism to characterize the entire conditional distribution of the response variable, not just its mean. This allows researchers to understand how predictors affect different parts of the distribution, making it particularly well-suited for analyzing data with heteroscedasticity or other forms of distributional heterogeneity.

This report aims to compare the performance of OLS and linear quantile regression estimators, focusing on these two key advantages of QR. We will demonstrate the robustness advantage of QR, particularly the conditional median ($\tau = 0.5$), under controlled outlier contamination in a simple linear model, contrasting it with the sensitivity of OLS. Additionally, we will illustrate QR's ability to capture distributional heterogeneity by examining its performance under heteroscedastic errors, showcasing

how it provides insights beyond the conditional mean estimated by OLS. We conduct a simulation study designed to highlight these differences visually and quantitatively, examining estimator behavior under three distinct scenarios: (1) a baseline homoscedastic Gaussian setting, (2) the same setting contaminated with high-leverage outliers, and (3) a setting with heteroscedastic errors. Coefficient stability and prediction accuracy (using Mean Absolute Error) are assessed across these scenarios.

2. Linear Quantile Regression: Theory and Methods

2.1 Definition and Estimation

For a random variable Y, the τ-th quantile is the value q_τ such that $P(Y \le q_\tau) = \tau$. In a regression setting, quantile regression (QR) estimates:

$$Q_Y(\tau \mid X = x) = x^\top \beta(\tau),$$

where $\beta(\tau)$ is a vector of coefficients specific to quantile level τ. For example, $\beta_1(0.5)$ represents the effect of X_1 on the median of Y.

Koenker and Bassett (1978) proposed estimating $\beta(\tau)$ by minimizing the check loss:

$$\hat{\beta}(\tau) = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta),$$

with $\rho_\tau(u) = u(\tau - \mathbb{I}\{u < 0\})$. For $\tau = 0.5$, this reduces to least absolute deviations (LAD) regression.

Each τ is estimated independently using linear programming, and the family $\{\beta(\tau)\}$ forms a quantile process describing the full conditional distribution of Y.

2.2 Inference

Under regularity conditions, $\hat{\beta}(\tau)$ is asymptotically normal:

$$\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \xrightarrow{d} N(0, \Sigma(\tau)),$$

with variance estimators obtained via bootstrapping or sandwich estimators (Koenker, 2005). Practical inference is available via `summary()` in the R package `quantreg`.

2.3 Comparison to OLS

OLS estimates the conditional mean:

$$\mathbb{E}[Y \mid X = x] = x^\top \beta,$$

whereas QR estimates conditional quantiles. When errors are symmetric and homoscedastic, QR and OLS give similar results. Otherwise, QR captures distributional heterogeneity, such as increasing variance or skewness.

Moreover, QR is robust to outliers in Y, unlike OLS which minimizes squared error and is sensitive to extreme values. QR also allows different slopes across quantiles, offering richer interpretation.

In the next section, we illustrate these theoretical advantages through a simulation study.

3. Simulation Study

This section presents a simulation study designed to compare the performance of **Ordinary Least Squares (OLS)** and **Quantile Regression (QR)** estimators under controlled, interpretable scenarios. Our aim is to assess how both methods behave, particularly under outlier contamination and additionally under heteroscedasticity.

To focus on the theoretical properties discussed in Section 2, we restrict attention to a simple univariate linear model with a single predictor. This allows for clean interpretation and visual representation of the results. Additionally, we consider three error structures: a baseline homoscedastic Gaussian case, a contaminated version with outliers at high-leverage points and a heteroscedastic case. This controlled setup isolates the impact of extreme observations on both methods, and is helps to reveal the key differences in robustness and sensitivity.

3.1 Data Generating Process (DGP)

We consider the simple linear model $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, with fixed parameters $\beta_0 = 5$ and $\beta_1 = -1.5$. These values are chosen to induce a moderate negative slope and an interpretable intercept, ensuring that both OLS and QR coefficients remain in a tractable range for interpretation and graphical analysis. The predictor X_1 is generated from a uniform distribution on $[0, 10]$, which provides a constant density across its support and avoids introducing implicit bias or skewness into the covariate structure. The sample size $n = 1000$ is selected to approximate asymptotic behavior while remaining computationally feasible. While we compute QR estimates for multiple quantiles ($\tau \in \{0.1, 0.5, 0.9\}$), the primary focus is on $\tau = 0.5$, which corresponds to the conditional median and allows direct comparison with the OLS estimator of the conditional mean. As highlighted in Section 2, QR estimates $\beta(\tau)$ independently for each τ, thus providing a richer description of the conditional distribution of Y than OLS.

3.2 Simulation Setup

To evaluate the estimators under different conditions relevant to their theoretical properties, we simulate data using the DGP described above under three distinct error structures for ε:

1. **Baseline Homoscedastic Gaussian Errors:** We first consider $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with a constant variance $\sigma = 2$. This scenario satisfies the classical OLS assumptions and serves as a benchmark

for comparing OLS and QR under ideal conditions.

2. **Contaminated Gaussian Errors (Outlier Robustness Test):** To assess robustness, we start with the baseline homoscedastic Gaussian errors and introduce contamination. Specifically, 2% of the observations ($n \times 0.02$) with the largest values of the predictor X_1 (high-leverage points) have their corresponding error terms perturbed by a large positive constant (+50). This creates vertical outliers combined with leverage, designed to maximally challenge the estimators' stability.
3. **Heteroscedastic Gaussian Errors (Distributional Modeling Test):** To illustrate the ability of QR to model distributional heterogeneity, we simulate errors whose variance increases linearly with the predictor: $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, where $\sigma_i = \sigma_0(1 + \gamma X_{1i})$. We use parameters like $\sigma_0 = 1$ and $\gamma = 0.3$ to generate noticeable heteroscedasticity. This scenario violates the OLS assumption of constant variance.

For each scenario, we fit both the OLS model and QR models at $\tau \in \{0.1, 0.5, 0.9\}$.

3.3 Evaluation Metrics

To quantify the behavior of the estimators, we combine visual inspection with numerical performance metrics. Given that QR minimizes absolute deviations, particularly for $\tau = 0.5$, we employ the **Mean Absolute Error (MAE)** as the primary metric. MAE is defined as $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$, and aligns directly with the objective function of the LAD estimator. It provides a robust measure of predictive accuracy and is less sensitive to outliers than the **RMSE**, which disproportionately penalizes large deviations. This makes MAE more appropriate when comparing methods under contamination or heavy-tailed noise, as discussed by Koenker (2005).

3.4 Simulation

We now illustrate the theoretical differences between OLS and quantile regression through a basic univariate simulation. Using the data-generating process defined in Section 3.1, we compare both estimators under clean and contaminated conditions. We focus on the conditional median estimate (QR at $\tau = 0.5$) and examine how each method responds to the presence of vertical outliers in high-leverage positions.

As shown in the figure @ref(fig:slope_plot), both OLS (dashed red line) and quantile regression (blue line) yield nearly identical slope estimates when applied to clean, homoscedastic data. This is in line with the theoretical results discussed in Section 2.3, where OLS and QR coincide under symmetric and homoscedastic error distributions. However, under contamination, even with as little as 2% of vertical outliers placed at high-leverage points, the OLS slope is visibly distorted—flattening as it attempts to minimize squared error. In contrast, the QR line remains largely

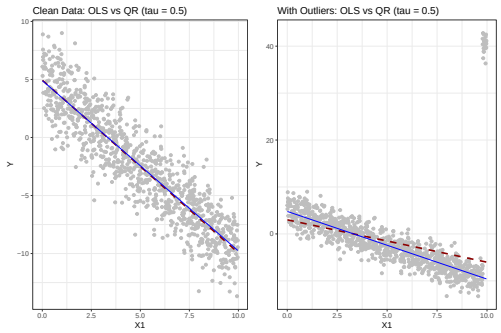

Figure 1. Comparison of slope estimates across settings
(#fig:slope_plot)

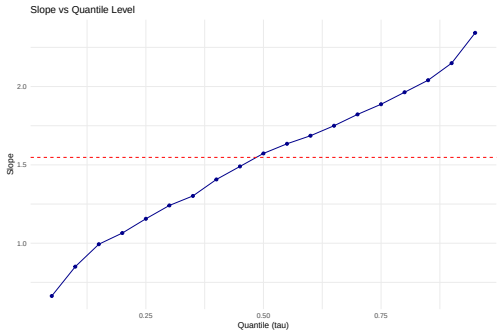

Figure 2. Slope vs Quantile for heteroskedastic data
(#fig:slope_quantile)

unaffected, producing a slope estimate that better reflects the central structure of the data. This illustrates the robustness property of quantile regression highlighted earlier: since QR minimizes a weighted absolute loss, it is less sensitive to large deviations in the response, and more resilient to local anomalies.

Figure @ref(fig:slope_quantile) illustrates how the estimated slope coefficient in quantile regression varies across quantile levels $\tau \in (0.05, 0.95)$ in a heteroskedastic setting. The increasing trend in slope estimates as τ increases reflects the presence of conditional heteroskedasticity: higher quantiles are associated with greater dispersion in the response variable Y, which alters the marginal effect of X_1 across the conditional distribution.

The dashed red line represents the OLS slope, which remains constant because it targets the conditional mean and assumes homoscedasticity. In contrast, quantile regression estimates $\beta_1(\tau)$ independently for each τ, capturing variation in the conditional distribution of Y.

Figure @ref(fig:other_plots) presents a multifaceted comparison between OLS and quantile regression under heteroskedasticity and contamination. In the top-left panel, QR lines at $\tau = 0.1, 0.5, 0.9$ capture the conditional distributional spread of Y, while OLS fits a single conditional mean. The top-right panel plots the estimated slope $\beta_1(\tau)$ across quantiles. The non-constant pattern confirms that

Table 1. (#tab:sum_table)OLS vs Quantile Regression Coefficient Comparison

Coefficient	OLS_Clean	QR_Clean	OLS_Outliers	QR_Outliers
(Intercept)	4.931	4.903	3.004	4.791
X1	-1.487	-1.468	-0.903	-1.442

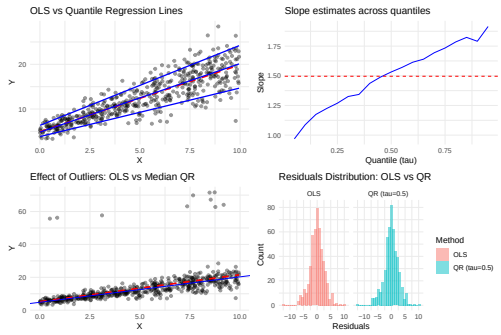

Figure 3. Slope vs Quantile for heteroskedastic data (#fig:other_plots)

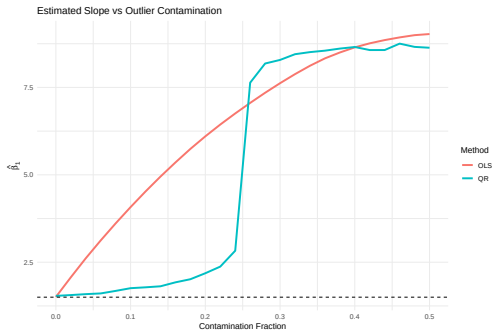

Figure 4. Estimated slope vs outlier contamination (#fig:slope_outliers)

the marginal effect of X varies with τ, violating the constant-slope assumption implicit in OLS, as discussed in Section 2.3.

The bottom-left panel illustrates robustness: under contamination, OLS is pulled downward due to extreme values in Y, while the median QR line remains stable. Finally, the residual histograms (bottom-right) show that OLS residuals are more dispersed and asymmetric, reflecting its sensitivity to outliers, whereas QR residuals remain more concentrated.

Table @ref(tab:eval_table) illustrates the comparative robustness of OLS and quantile regression (QR at $\tau = 0.5$ under clean and contaminated settings. Both estimators perform similarly in the Gaussian case. However, under contamination, OLS coefficients exhibit severe bias, particularly in the slope, while QR remains stable. This behavior is theoretically consistent with QR's bounded influence function and its minimization of the check loss, which limits sensitivity to extreme values. MAE, aligned with the LAD objective, remains nearly unchanged for QR, whereas OLS shows substantial deterioration. RMSE, although not optimized by QR, is included for standard comparison and further confirms OLS's vulnerability under outlier contamination.

Figure @ref(fig:slope_outliers) shows the evolution of the estimated slope $\hat{\beta}_1$ as a function of outlier contamination. As predicted by the robustness theory in Section 2.3, QR remains stable under moderate contamination, exhibiting high resistance to leverage-induced distortion. However, once the proportion of outliers exceeds a critical threshold (~25%), their influence becomes structurally dominant—shifting the conditional median itself and leading to breakdown. In contrast, OLS degrades continuously, with no resistance to contamination at any level.

4. Discussion & Conclusions

This report has compared ordinary least squares (OLS) and quantile regression (QR) for modeling linear relationships, focusing on scenarios where classical OLS assumptions are violated. Through controlled simulations, we demonstrated two primary advantages of QR. Firstly, QR, particularly median regression ($\tau = 0.5$\$, exhibits significant **robustness to outliers**, providing stable and reliable coefficient estimates even under contamination with high-leverage points, a condition where OLS estimates suffered severe bias (as shown in Figure @ref(fig:slope_plot) and Table @ref(tab:eval_table)). This resilience stems from QR's use of the check loss function, which minimizes absolute deviations rather than squared deviations.

Secondly, QR offers a powerful tool for characterizing **distributional heterogeneity**. Our simulations under heteroscedastic errors (Figures @ref(fig:slope_quantile) and @ref(fig:other_plots)) showed that QR can effectively capture how the effect of predictors varies across different quantiles of the response distribution, revealing patterns (like increasing variance) that are entirely missed by the single conditional mean estimate provided by OLS. The quantile-specific slopes provide a richer, more nuanced understanding of the underlying data generating process.

In conclusion, while OLS remains a cornerstone of regression analysis, quantile regression provides a flexible, robust, and more informative alternative, particularly valuable when dealing with non-Gaussian errors, outliers, or heteroscedasticity. These results highlight its utility as a complementary, and often superior, tool for modeling complex data relationships.

Considerations: ***This report has been made with a template in R Markdown. CHATGPT***

Table 2. (#tab:metrics_table, echoFALSE)OLS vs Quantile Regression: Coefficients and Error Metrics under Different Error Structures

	Model	Method	Intercept	Slope	Slope_Bias	MAE	RMSE
(Intercept)...1	Gaussian	QR	4.903	-1.468	0.032	1.533	1.925
(Intercept)...2	Contaminated	QR	4.791	-1.442	0.058	1.534	1.930
(Intercept)...3	Gaussian	OLS	4.931	-1.487	0.013	1.534	1.924
(Intercept)...4	Contaminated	OLS	3.004	-0.903	0.597	2.229	2.746

has been used for cleaning the code and debugging.