Advance Regression Term Project

Gerard Palomo & Juan Pablo Uphoff

Abstract
Linear quantile regression extends ordinary least squares (OLS) by modeling conditional quantiles of a response variable as linear functions of predictors. This offers a more complete view of the conditional distribution, revealing heterogeneous effects not captured by mean regression. Unlike OLS, quantile regression makes no strict distributional assumptions and is robust to outliers. We review the formulation, estimation, and inference for linear quantile regression, contrasting it with OLS. A simulation study compares OLS and quantile regression at various quantile levels (0.1 to 0.9) in both univariate and multivariate settings. We examine performance under normal and heavy-tailed error distributions, including scenarios with outliers, and assess the impact of sample size. The results illustrate that quantile regression estimates remain reliable under outlier contamination and uncover distributional effects (e.g. heteroscedasticity) that OLS misses. All simulation code is provided in R for reproducibility.

Keywords
OLS, Quantile Regression, Machine Learning

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 (ordinary least squares, OLS) fits a line by minimizing the sum of squared residuals, which targets the conditional mean of the response variable. However, because OLS puts heavy weight on large deviations, its estimates can be strongly affected by even a few outliers. In contrast, quantile regression extends median regression to any conditional quantile (e.g., the 50% quantile) and minimizes an asymmetrically weighted absolute loss. A key advantage of quantile regression is its robustness: the influence function of the median is bounded, so regression quantiles inherit a high breakdown point. In practical terms, estimates like the conditional median are not pulled as far by extreme values.

Prior studies confirm these properties: for example, Herawati (2020) showed in simulations that median regression provided smaller mean-squared error than OLS when outliers are present. In this article, we compare OLS and quantile regression for a linear model under controlled outlier contamination.

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.

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 only two error structures: a baseline homoscedastic Gaussian case and a contaminated version with outliers at high-leverage points. 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 realistic and adversarial conditions, we simulate two distinct error structures for ε. First, we consider homoscedastic Gaussian errors: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 2$, which fulfill all Gauss-Markov conditions and provide a benchmark for both estimators. Second, we introduce contaminated Gaussian errors to test robustness: 2% of the residuals are perturbed by a fixed additive shift (+50), applied to observations with the largest values of X_1, thereby combining vertical outliers with high-leverage covariate values.

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 unaffected, producing a slope estimate that better reflects the central structure of the data. This illustrates the robustness property of quantile regression

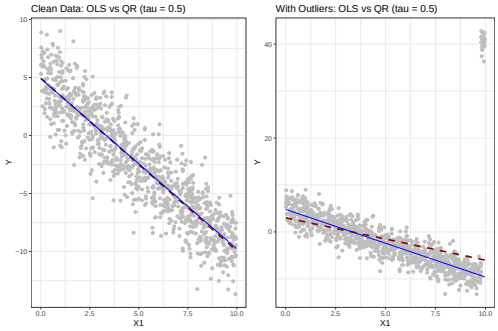

Figure 1. Comparison of slope estimates across settings
(#fig:slope_plot)

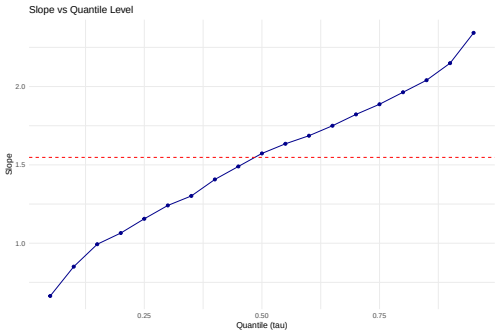

Figure 2. Slope vs Quantile for heteroskedastic data (#fig:slope_quantile)

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

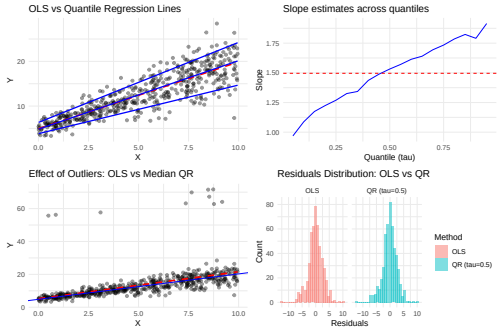

Figure 3. Slope vs Quantile for heteroskedastic data (#fig:other_plots)

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 demonstrated the advantages of quantile regression over ordinary least squares (OLS) when estimating linear relationships under deviations from classical assumptions. Through controlled simulations, we showed that QR remains robust to outliers and captures distributional heterogeneity—features that OLS fails to address. In clean settings, both methods yield similar estimates, but under contamination, OLS suffers severe bias while QR maintains

Table 1. (#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

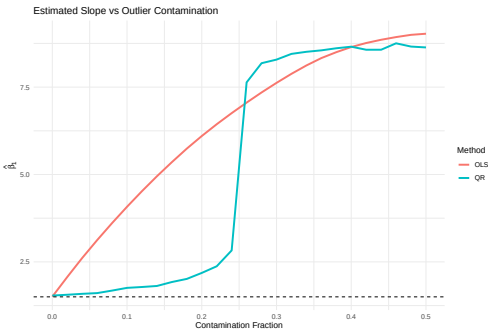

Figure 4. Estimated slope vs outlier contamination (#fig:slope_outliers)

stable performance. Moreover, the quantile-specific slopes reveal structural changes across the distribution that are invisible to mean regression. These results highlight quantile regression as a flexible and resilient alternative to OLS for modeling complex data.

Considerations: ***This report has been made with a template in R Markdown. CHATGPT has been used for cleaning the code and debugging.***