Robustness of OLS vs Quantile Regression: Handling Outliers and Heteroscedasticity

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 simple univariate simulation study under three different settings. Results demonstrate that QR estimates 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

JEL
C15, C21

Acknowledgements
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**.

***Contact**: 100538493@alumnos.uc3m.es and 100508278@alumnos.uc3m.es, www.uc3m.es

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 Onyedikachi (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 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 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. 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, 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 begin with homoscedastic Gaussian errors and introduce contamination. Specifically, 2% of the observations with the largest values of the predictor X_1 (high-leverage points) have their error terms perturbed by a large positive constant (+50), creating vertical outliers with leverage. Additionally, we introduce non high-leverage outliers by randomly selecting 2% of the observations and adding a large positive shift to their response values.
3. **Heteroscedastic Gaussian Errors (Distributional Modeling Test):** To demonstrate QR's ability to handle heteroscedasticity, we simulate errors where the variance increases

linearly with the predictor:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad \text{where} \quad \sigma_i = \sigma_0(1 + \gamma X_{1i}).$$

Here, we use the parameters $\sigma_0 = 1$ and $\gamma = 0.6$ to introduce **heteroscedasticity**—a scenario where the error variance depends on the value of the predictor variable (X_1).

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

3.4.1 Outlier Robustness

We now illustrate the theoretical differences between OLS and quantile regression through a basic univariate simulation. Using the DGP 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.

Figure 1 shows that both OLS (dashed red line) and quantile regression (blue line) yield nearly identical slope estimates when applied to clean, homoscedastic data, in line with the theoretical results. However, under contamination, even with as little as 2% of vertical high-leverage outliers, the OLS slope is distorted—flattening as it attempts to minimize squared error. In contrast, the QR line remains 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 2 illustrates how non high-leverage outliers affect slope estimates. In this specific scenario, both Ordinary Least Squares (OLS) and Quantile Regression (QR) produce similar results, with OLS showing minimal sensitivity to the outliers. This limited impact occurs because the outliers do not occupy high-leverage positions and therefore exert little influence on the slope.

As shown in Table 1, the evaluation of all scenarios reveals that OLS and Quantile Regression (QR) at = 0.5 perform similarly under clean data conditions. However, when outliers are introduced, OLS exhibits

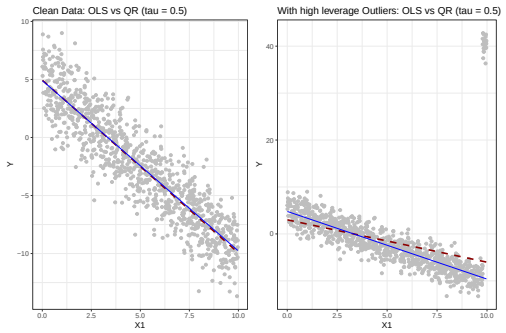

Figure 1. OLS vs QR under high leverage outliers

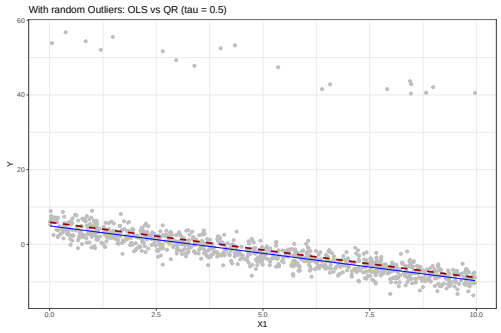

Figure 2. OLS vs QR under random outliers

significant bias, particularly in the slope, and its error metrics (MAE and RMSE), computed on the clean data, deteriorate. In the case of high leverage outliers, where extreme X values exert a large influence on the model, OLS shows substantial bias in the slope, while QR remains stable with minimal bias and consistent error metrics. For random outliers, where extreme Y values are scattered across the data, OLS again shows deterioration in its estimates and error metrics, whereas QR maintains its robustness, demonstrating limited sensitivity to both types of outliers. This behavior aligns with QR's bounded influence function, which limits the impact of extreme values, unlike OLS, which is more vulnerable to contamination.

Figure 3 shows Ordinary Least Squares (OLS) slope immediately degrading under high leverage outlier contamination due to its sensitivity to squared errors. Quantile Regression (QR) initially resists,

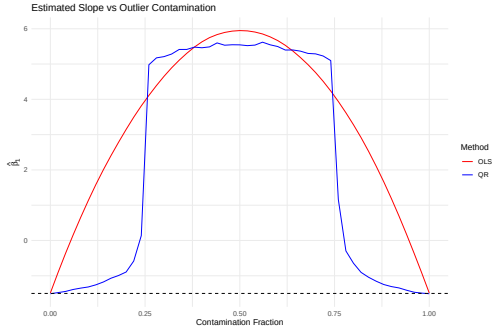

Figure 3. Estimated slope vs outlier contamination

Table 1. OLS vs Quantile Regression: Coefficients and Error Metrics under Different Error Structures

Model	Method	Intercept	Slope	Slope_Bias	MAE	RMSE
Clean	QR	4.903	-1.468	0.032	1.533	1.925
Clean	OLS	4.931	-1.487	0.013	1.534	1.924
Highe Leverage Outliers	QR	4.791	-1.442	0.058	1.534	1.930
Highe Leverage Outliers	OLS	3.004	-0.903	0.597	2.229	2.746
Random Outliers	QR	4.935	-1.468	0.032	1.533	1.927
Random Outliers	OLS	5.917	-1.484	0.016	1.726	2.168

maintaining a stable slope by focusing on the median. However, QR breaks down beyond a critical outlier threshold (around 25%) as outliers then redefine the median.

As contamination approaches 100%, both OLS and QR exhibit a similar pattern: the original data points become the new "outliers." Consequently, the estimated slope for both methods may revert to a value resembling the original slope. Critically, this is not a true recovery for either, as the entire fitted line has significantly shifted, now reflecting the characteristics of the overwhelming majority of (initially outlying) data points.

3.4.2 Heteroscedasticity

Beyond outlier robustness, Quantile Regression also excels in capturing distributional changes. This subsection demonstrates its effectiveness in handling heteroscedasticity, where the response variable's spread is dependent on the predictors.

In Figure 4 illustrates how the estimated slope coefficient in quantile regression varies across quantile levels $\tau \in (0.1, 0.9)$ in a heteroskedastic setting. The increasing slope 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. Quantile regression estimates $\beta_1(\tau)$ independently for each τ, capturing variation in the conditional distribution of Y while the slope of OLS remains constant because it targets the conditional mean and assumes homoscedasticity.

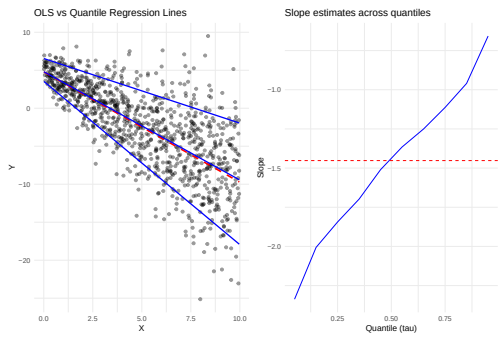

Figure 4. Slope vs Quantile for heteroskedastic data

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 1 and Table 2. This resilience stems from QR's use of the check loss function, which minimizes absolute deviations rather than squared deviations.

Second, QR effectively models distributional heterogeneity, capturing how predictor effects vary across the conditional distribution. Our heteroscedasticity simulations (Figures 2 and 3) revealed trends and spread that OLS, focused solely on the conditional mean, entirely misses. In conclusion, while OLS remains a foundational method for estimating conditional means, quantile regression (QR) provides a more versatile framework—particularly in settings involving outliers, heteroscedasticity, or deviations from Gaussian error structures—by capturing the full conditional distribution and enabling quantile-specific inference on covariate effects.

Acknowledgements: This report has been made with a template in R Markdown, taking as example the paper written by Fan and Li (2001). CHATGPT has been used for cleaning the code and debugging.

References

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, Cambridge, UK.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Onyedikachi, J. (2015). Robustness of quantile regression to outliers. *American Journal of Applied Mathematics and Statistics*, 3(2):86–88.