Working Paper, MSc in Statistics for Data Science, Carlos III University of Madrid
DOI: N/A (internal report)

Advance Regression Term Project

Gerard Palomo & Juan Pablo Uphoff

Abstract
Ordinary least squares (OLS) regression fits the mean and is sensitive to extreme observations, whereas quantile regression (e.g. median regression) fits conditional quantiles and tends to be more robust to outliers. We conduct a simulation study comparing OLS and median (50% quantile) regression under increasing levels of outlier contamination in the response variable. Synthetic data are generated from a simple linear model and a fraction of responses are replaced by large aberrant values. We fit both OLS and quantile regression and compare the resulting coefficient estimates. As expected, OLS estimates become severely biased in the presence of outliers, while the median regression estimates remain closer to the true values. These findings confirm that quantile regression can better resist outliers than OLS

Keywords
OLS, Quantile Regression, Machine Learning

JEL
C45, C55

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**.

***Corresponding author**: , https://www.uc3m.es

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. We focus on interpreting simulation results (accessible to students) rather than deep theory.

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 $P(Y \leq q_\tau) = \tau$. 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:

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

Table 1. Comparison of OLS and Quantile Regression Coefficients (Clean Data) - Manual

Coefficient	OLS	QR 0.50
(Intercept)	5.082	5.16
X1	-1.512	-1.52

Table 2. Comparison of OLS and Quantile Regression Coefficients (Clean Data) - Manual

Coefficient	OLS	QR 0.50
(Intercept)	0.390	4.723
X1	-0.065	-1.391

- **Homoscedastic Gaussian**: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\sigma = 2$. This setting satisfies OLS assumptions and serves as a benchmark.
- **Contaminated Gaussian (outliers)**: $\varepsilon_i = Z_i + O_i$ with $Z_i \sim \mathcal{N}(0, 2^2)$ and $O_i = 50$ with probability $p = 0.05$. This mimics a mixture contamination model, testing robustness. Outliers are introduced conditionally (to the largest X_1) to induce high-leverage effects.
- **Heteroskedastic**: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2(X_i))$ with $\sigma(X_i) = 1 + 0.2X_1$. This setting creates increasing spread in Y conditional on X_1, violating OLS homoscedasticity.

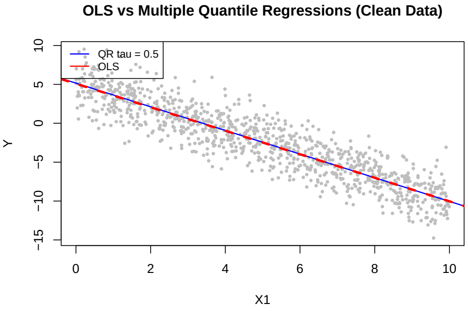
3.3 Evaluation Metrics

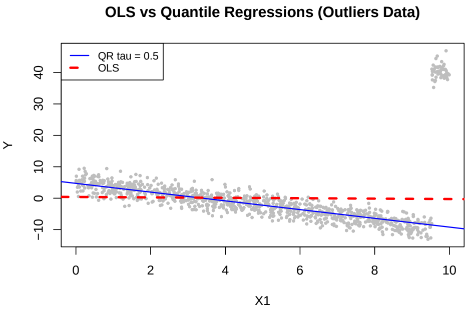
*****PENDING***** Along with some graphical support, these evaluations provide insight into estimator consistency, robustness, and efficiency under both ideal and adversarial conditions.

The theoretical basis for QR (Section 2) suggests that: - OLS is optimal under homoscedastic Gaussian errors (Gauss-Markov assumptions), - QR provides robust, distribution-sensitive estimation that remains valid when these assumptions are violated.

This simulation seeks to verify those expectations empirically.

3.4 Simulation





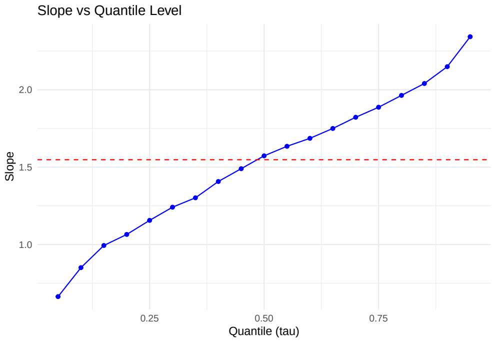


Figure 1. Slope vs Quantile for heteroskedastic case

```
set.seed(666)
data_heterosk <- simulate_data(n = 1000, heterosk = TRUE)
fits_heterosk <- lapply(seq(0.05, 0.95, 0.05), function(tau) rq(y ~
    x1, tau = tau, data = data_heterosk))
slopes_heterosk <- sapply(fits_heterosk, function(fit) coef(fit)[2])
ols_slope <- coef(lm(y ~ x1, data = data_heterosk))[2]
plot_df <- tibble(tau = seq(0.05, 0.95, 0.05), slope = slopes_heterosk)
ggplot(plot_df, aes(x = tau, y = slope)) + geom_line(color = "blue") +
    geom_point(color = "blue") + geom_hline(yintercept = ols_slope,
    color = "red", linetype = "dashed") + labs(x = "Quantile (tau)",
    y = "Slope", title = "Slope vs Quantile Level") +
    theme_minimal()
```

```
fit25 <- rq(y ~ x1, tau = 0.25, data = scenarios$clean_multi)
fit50 <- rq(y ~ x1, tau = 0.5, data = scenarios$clean_multi)
fit75 <- rq(y ~ x1, tau = 0.75, data = scenarios$clean_multi)
anova(fit25, fit50, fit75)
```

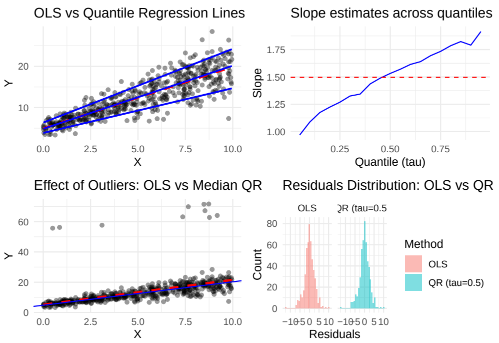
```
## Quantile Regression Analysis of Deviance Table
##
## Model: y ~ x1
## Joint Test of Equality of Slopes: tau in {  0.25 0.5
##
##   Df Resid Df F value Pr(>F)
## 1  2     1498  6.4657 0.0016 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# kbl(anova_table, caption = 'Joint Test of
# Equality of Slopes at Different Quantiles') %>%
# kable_styling(latex_options = 'striped')
fit25_h <- rq(y ~ x1, tau = 0.25, data = scenarios$outlier_multi)
fit50_h <- rq(y ~ x1, tau = 0.5, data = scenarios$outlier_multi)
fit75_h <- rq(y ~ x1, tau = 0.75, data = scenarios$outlier_multi)
anova(fit25_h, fit50_h, fit75_h)
```

```
## Quantile Regression Analysis of Deviance Table
##
## Model: y ~ x1
## Joint Test of Equality of Slopes: tau in {  0.25 0.5
##
##   Df Resid Df F value   Pr(>F)
## 1  2     1501   6.048 0.002421 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(quantreg)
library(gridExtra)
library(dplyr)
# Simulate basic heteroskedastic data
```

```r
set.seed(666)
n <- 500
X <- runif(n, 0, 10)
epsilon <- rnorm(n, mean = 0, sd = 1 + 0.3 * X)
Y <- 5 + 1.5 * X + epsilon
data <- data.frame(X = X, Y = Y)
# Fit models
ols_fit <- lm(Y ~ X, data = data)
rq_10 <- rq(Y ~ X, tau = 0.1, data = data)
rq_50 <- rq(Y ~ X, tau = 0.5, data = data)
rq_90 <- rq(Y ~ X, tau = 0.9, data = data)
# Plot 1: Scatterplot + regression lines
p1 <- ggplot(data, aes(x = X, y = Y)) + geom_point(alpha = 0.4) +
    geom_smooth(method = "lm", se = FALSE, color = "red",
        linetype = "dashed") + geom_quantile(quantiles = c(0.1,
    0.5, 0.9), color = "blue", size = 0.8) + labs(title = "OLS vs Quantile Regression Lines",
    y = "Y", x = "X") + theme_minimal()
# Plot 2: Slope across quantiles
taus <- seq(0.05, 0.95, by = 0.05)
slopes <- sapply(taus, function(tau) coef(rq(Y ~ X,
    tau = tau))[2])
p2 <- ggplot(data.frame(tau = taus, slope = slopes),
    aes(x = tau, y = slope)) + geom_line(color = "blue") +
    geom_hline(yintercept = coef(ols_fit)[2], linetype = "dashed",
        color = "red") + labs(title = "Slope estimates across quantiles",
    y = "Slope", x = "Quantile (tau)") + theme_minimal()
# Plot 3: Robustness to outliers
data_outlier <- data
idx <- sample(1:n, 10)
data_outlier$Y[idx] <- data_outlier$Y[idx] + 50  # Add strong outliers
ols_fit_outlier <- lm(Y ~ X, data = data_outlier)
rq_50_outlier <- rq(Y ~ X, tau = 0.5, data = data_outlier)
p3 <- ggplot(data_outlier, aes(x = X, y = Y)) + geom_point(alpha = 0.4) +
    geom_smooth(method = "lm", se = FALSE, color = "red",
        linetype = "dashed") + geom_abline(intercept = coef(rq_50_outlier)[1],
    slope = coef(rq_50_outlier)[2], color = "blue") +
    labs(title = "Effect of Outliers: OLS vs Median QR",
        y = "Y", x = "X") + theme_minimal()
# Plot 4: Residuals histograms
residuals_ols <- resid(ols_fit)
residuals_rq50 <- resid(rq_50)
residuals_df <- data.frame(residuals = c(residuals_ols,
    residuals_rq50), Method = rep(c("OLS", "QR (tau=0.5)"),
    each = n))
p4 <- ggplot(residuals_df, aes(x = residuals, fill = Method)) +
    geom_histogram(alpha = 0.5, position = "identity",
        bins = 30) + facet_wrap(~Method) + labs(title = "Residuals Distribution: OLS vs QR",
    x = "Residuals", y = "Count") + theme_minimal()
# Arrange in grid (2x2 layout)
grid.arrange(p1, p2, p3, p4, nrow = 2)
```



4. Discussion & Conclusions.

xxxxx

Extras and considerations: This report has been made with a template in R Markdown, following the guidelines from (**?**), and the most famous resource for the creation of reports in R Markdown, the bookdown package (**?**).

References