Comparison of Trimmed Mean and Least Trimmed Squares based estimators for linear regression

Akhilesh Chauhan(170070) and Apoorva Singh(16817144)

November 21,2020

Problem Statement

Define trimmed mean in the simple linear regression model. Compare the performances between the LTS estimator and your proposed trimmed mean.

Trimmed Mean

- Trimmed Means are robust estimators of the central tendency
- Let $x_{(1)}, x_{(1)}, ..., x_{(n)}$ denote the ordered observations, then the trimmed mean is defined by,

$$T(x_1, x_2, ..., x_n; \alpha) = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}$$

  where α is the trimming parameter

- For calculating trimmed mean in linear regression, we use regression quantiles

Regression Quantiles

- Consider the linear model

$$\vec{y} = \vec{\beta}X + \vec{Z}$$

where $\vec{y}^{\mathsf{T}} = [y_1, y_2, ..., y_n]$, X is a nxp matrix of known constants whose i^{th} row is $\vec{x_i}$, $\vec{\beta}^{\mathsf{T}} = [\beta_1, \beta_2, ..., \beta_n]$ is a vector of unknown parameters, and $\vec{Z}^{\mathsf{T}} = [Z_1, Z_2, ..., Z_n]$ is a vector of independent and identically distributed random variables.

- Koenker and Bassett(1978) extended the concept of quantiles to the linear model. Let $0 < \theta < 1$. Define

$$\psi_\theta(x) = \theta - \mathbb{I}(x < 0)$$

$$\rho_\theta = x\psi_\theta(x)$$

Then the θ^{th} regression Quantile is given as

$$\hat{\beta}(\theta) = \underset{\vec{b}}{\text{argmin}} \sum_{i=1}^{n} \rho_\theta(y_i - \vec{x_i}^{\mathsf{T}}\vec{b})$$

Regression Quantiles

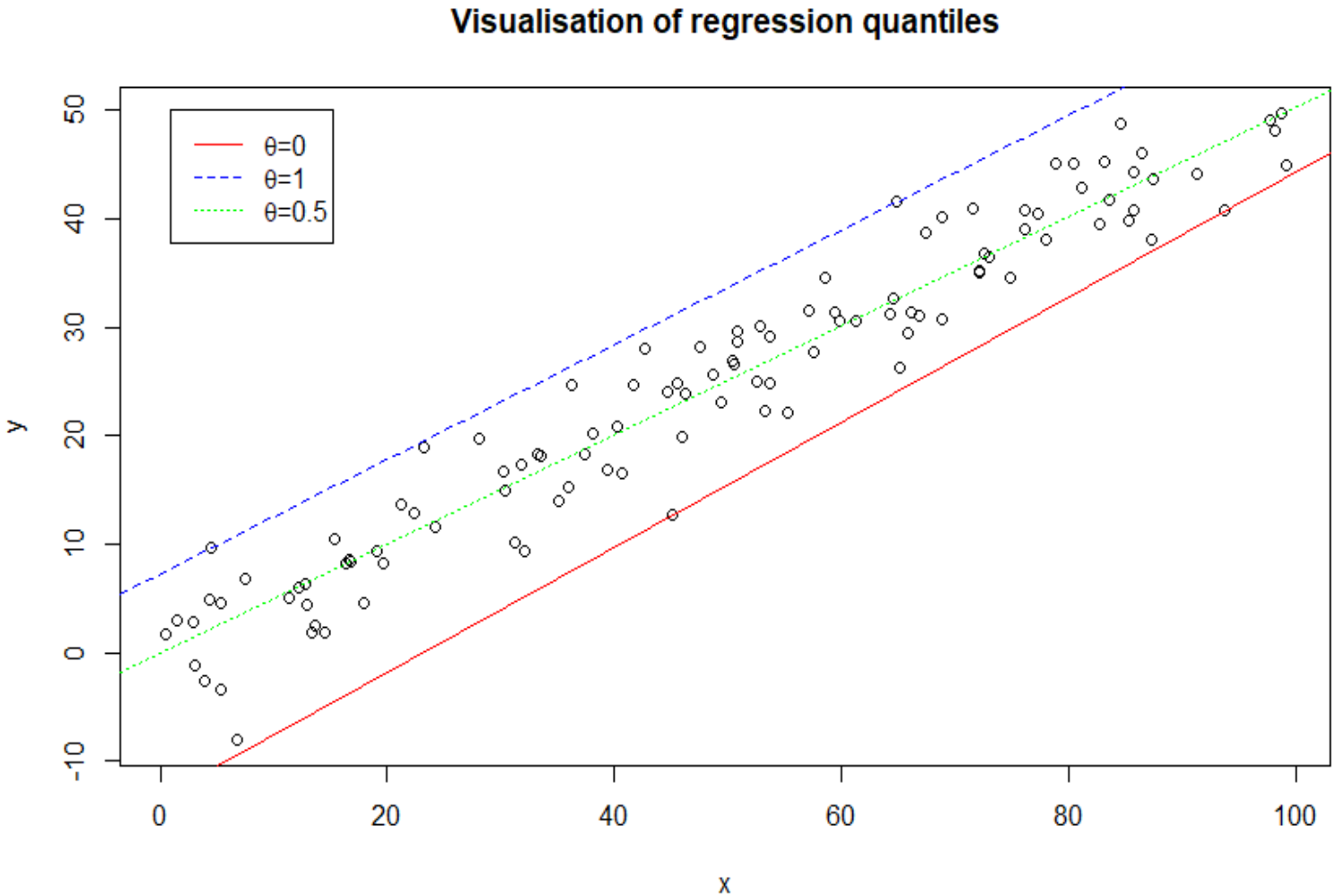
- Note that the above definition can be reduced to a form similar to the definition of multivariate quantiles we saw earlier.i.e

$$\hat{\beta}(\theta) = \underset{\vec{b}}{\text{argmin}} \sum_{i=1}^{n} [|y_i - \vec{x_i}^{\mathsf{T}} \vec{b}| + u(y_i - \vec{x_i}^{\mathsf{T}} \vec{b})]$$

  where $u = (2\theta - 1)$

- $\theta = 0$ and $\theta = 1$ covers the spread of our data
- For $\theta = 0.5$, $\hat{\beta}(\theta)$ gives the LAD regression estimator.

Visualisation of Regression Quantiles



Trimmed Mean for Linear Regression

- Koenker and Bassett proposed the following definition of α trimmed mean for regression and they call it $\hat{\beta}_{KB}$: Remove from the sample any observations whose residual from $\hat{\beta}(\alpha)$ is negative or whose residual from $\hat{\beta}(1-\alpha)$ is positive and calculate the least squares estimator using the remaining observations.

- We are using the above definition of trimmed mean for our purpose.

Computation of Trimmed Mean

- Consider the set of equations

$$y_i - \vec{x_i}^T \vec{\beta}(1-\alpha) \geq 0 \text{ or } y_i - \vec{x_i}^T \vec{\beta}(\alpha) \leq 0$$

- Let $b_i = 0$ or 1 according as i satisfies the above set of equations or not, and let B be the nXn diagonal matrix with $B_{ii} = b_i$. Then

$$\hat{\vec{\beta}}_{KB}(\alpha) = (X^T B X)^{-1}(X^T B \vec{y})$$

where $(X^T B X)^{-1}$ is the inverse of $(X^T B X)$

Least Trimmed Squares

- A method for robust regression that minimises the sum of squared residuals over a subset of data points.
- Let $(y_1, x_1), (y_2, x_2), ..., (y_n, x_n)$ denote the data points with the linear regression model defined as $y_i = \beta x_i + \epsilon_i$ for $i = 1, ..., n$
- Then, $\hat{\beta}^{(LTS)}$, the LTS estimator of β, is defined as

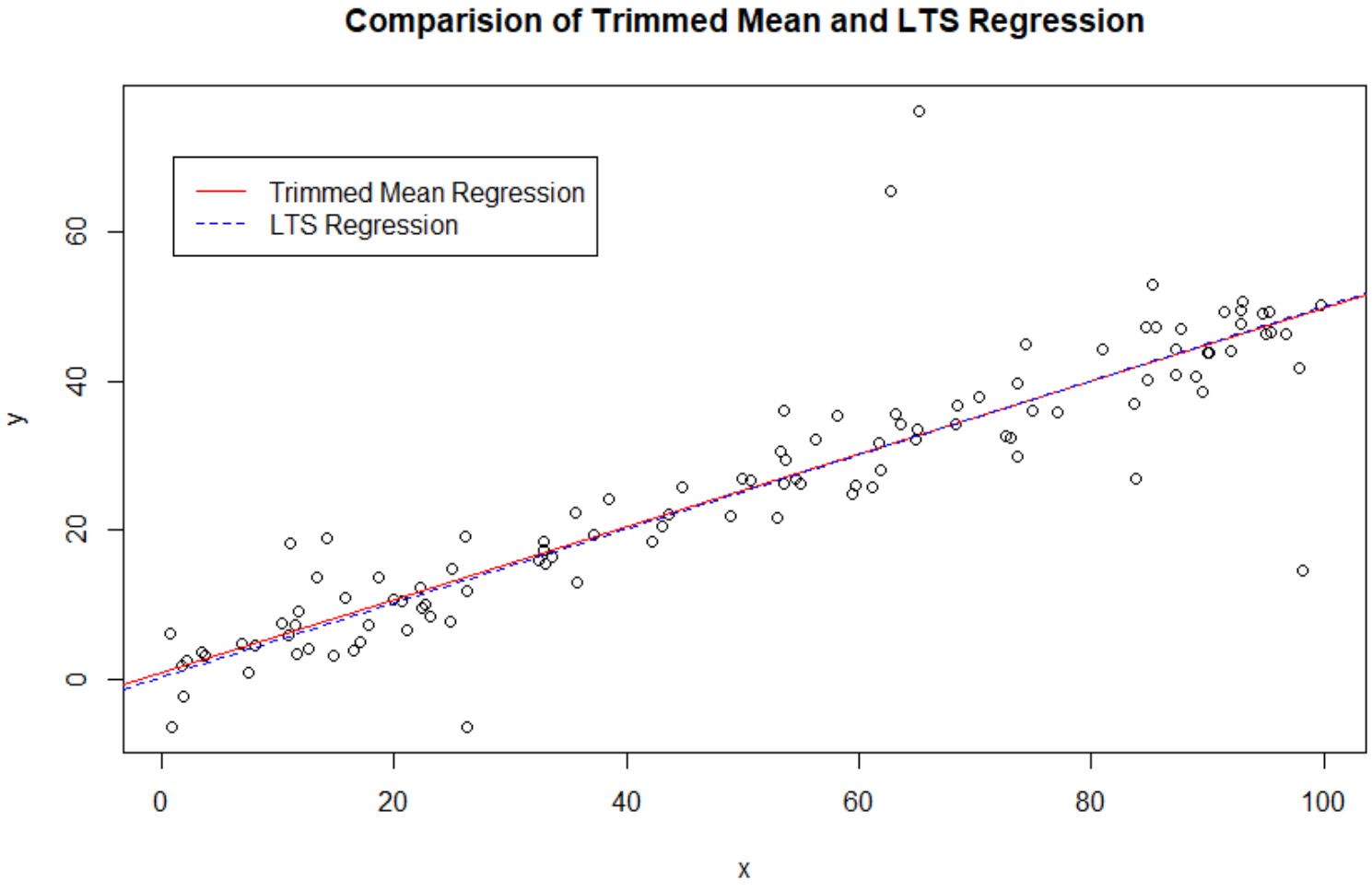
$$\hat{\beta}^{(LTS)} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{h} \epsilon_i^2$$

where $n/2 \leq h \leq n$

Methodology for simulated data

- We chose the Noise distribution to be a mixture of a low(9) and a high variance(400) Gaussians. We added a small proportion of high variance Gaussians to get some outliers in our data.
- We generated the data from the from the process $y_i = 0.5x_i + \epsilon_i$, where x_i is generated from $U[0, 100]$ distribution.
- We fitted a Trimmed Mean and an LTS estimator to the data.
- The parameters of Trimmed Mean and LTS estimator are chosen such that the Mean Squared Error(MSE) is minimised(Best Possible Fit).
- We repeated this process 100 times and calculated the MSE for all the replications.

Results for Simulated Data



- For this particular replication
  MSE(Trimmed Mean) = 58.50234 MSE(LTS) = 58.72007

Results for Simulated Data

- For around 95% replications, the MSE for the Trimmed Mean is better than LTS.
- Although keep in mind that the MSEs for all the replications are very close to each other.
- From the above exercise, we may infer that Trimmed Mean based regression performs slightly better than LTS regression.

Methodology for real data

- **Data:** Our data contains the price series for the closing price of Dow Jones Industrial Average(DJIA) in the period 1980-1995.
- Tried to get the best fit Trimmed Mean and LTS estimators of $\vec{\beta} = [\beta_0, \beta_1]^{\mathsf{T}}$(Let it be $\hat{\vec{\beta}}$). (The one having the minimum MSE)
- Compared the MSEs of both the estimators.
- Also tried to compare their performance using bootstrapping. Generated 100 replications of the data.
- For each replication, we tried fitting both the trimmed mean as well as the LTS estimator. Lets denote them by $\vec{\beta}_i^{1*}$ and $\vec{\beta}_i^{2*}$ respectively.(where i denotes the replication index)

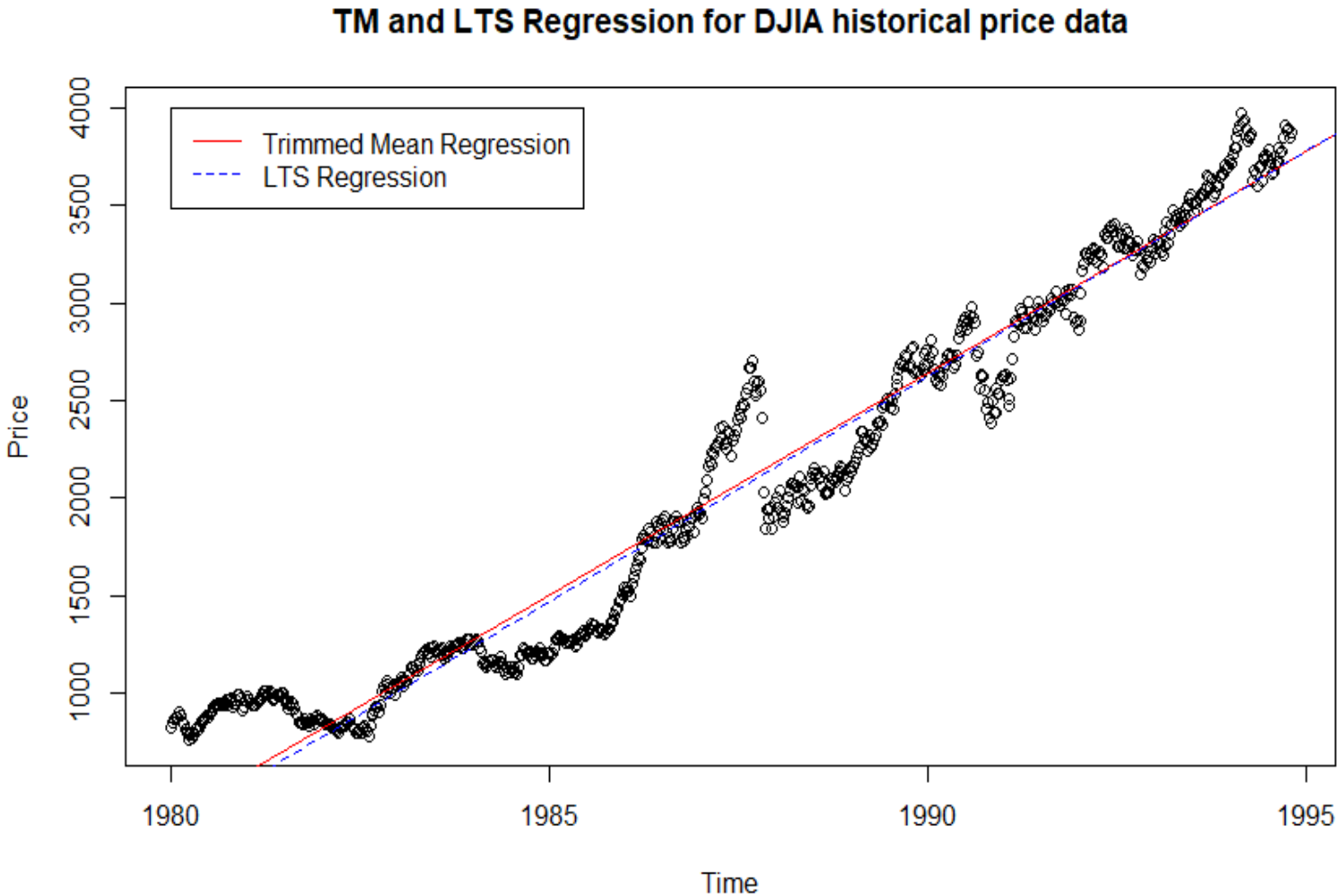
Methodology for real data

- **Method 1 :** We calculate the MSE for both the estimators in all the replications. In the end, we calculate the fraction of times Trimmed Mean MSE is better than the LTS MSE.
- **Method 2 :** We get the values of $\vec{\beta}_i^{1*}$ and $\vec{\beta}_i^{2*}$ for all the replications. Using these values, we calculate the Bootstrap Mean Squared Error.
- The Bootstrap Mean Squared Error is defined as :

$$E_j = \frac{1}{nboot} \sum_{i=1}^{nboot} \|(\hat{\vec{\beta}} - \vec{\beta}_i^{j*})\|^2$$

where nboot = number of replications in bootstrap and j =1,2 represents the trimmed mean and the LTS Bootstrap MSE respectively.

Results for real data



Results for real data

- MSE(Trimmed Mean) =44719.23
- MSE(LTS) = 44715.09
- $MSE(TrimmedMean) > MSE(LTS)$
- For around 70% of the replications, Trimmed Mean Regression had a better MSE than LTS Regression.
- Bootstrap MSE(Trimmed Mean) = 16247482
- Bootstrap MSE(LTS) = 33333897
- $BootstrapMSE(TrimmedMean) < BootstrapMSE(LTS)$
- We may conclude from the above that Trimmed Mean Estimator is better than LTS Regression.

References

- A.H. Welsh.The Trimmed Mean in the Linear Model(1987)
- David Ruppert and Raymond J. Carroll.Trimmed Least Squares Estimation in the Linear Model(1980)
- Roger Koenker, Gilbert Bassett, Jr. Regression Quantiles(1978)
- Peter J. Rousseeuw. Least Median of Squares Regression
- (Data source) https://vincentarelbundock.github.io/Rdatasets/datasets.html

Thank You!