TOPIC 3

DESCRIPTION

The experiment compares the performance of three robust linear regression models with traditional linear regression in the presence of outliers.

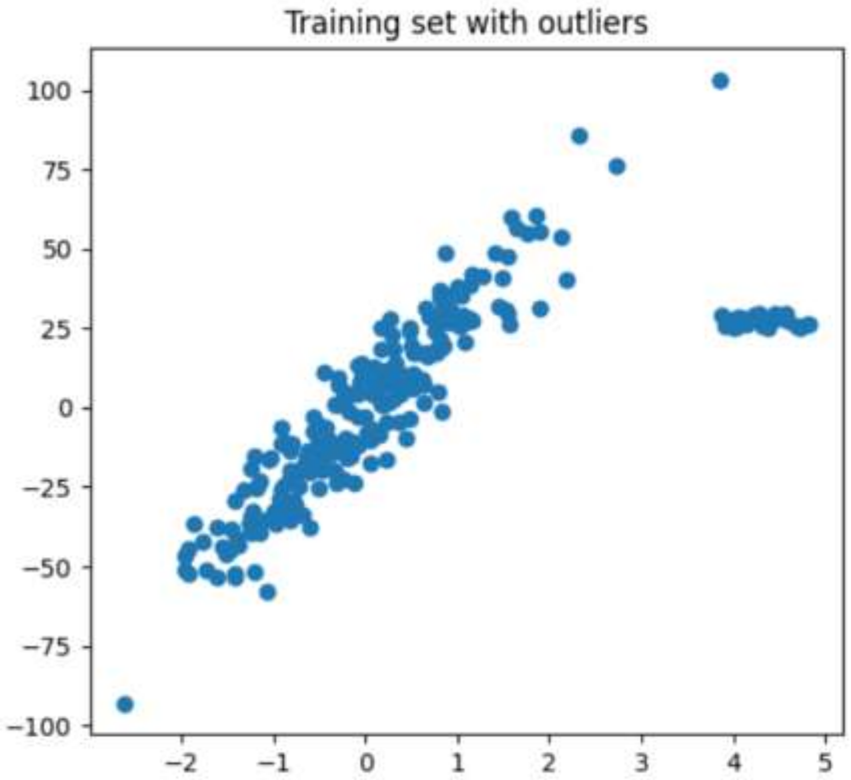
- **RANSAC** (RANdom SAmple Consensus) regression
- **Theil-Sen** regression
- **Huber** regression

INTRODUCTION

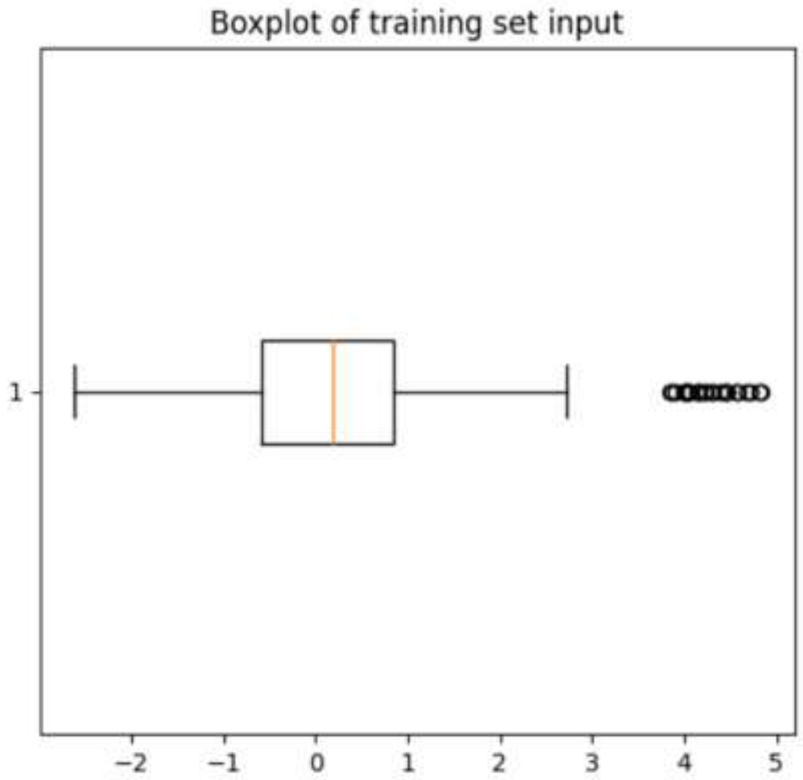
Outliers are objects that significantly deviate from the majority of the dataset. Their presence can lead to longer training times and less accurate models. Therefore, when outliers are present, more robust models are required. Robust linear regression is a regression analysis method that is less affected by outliers compared to traditional linear regression.

TOPIC 3

DATASET GENERATION



Training set with outliers



Boxplot of training set input

Training set:

250 data points in total

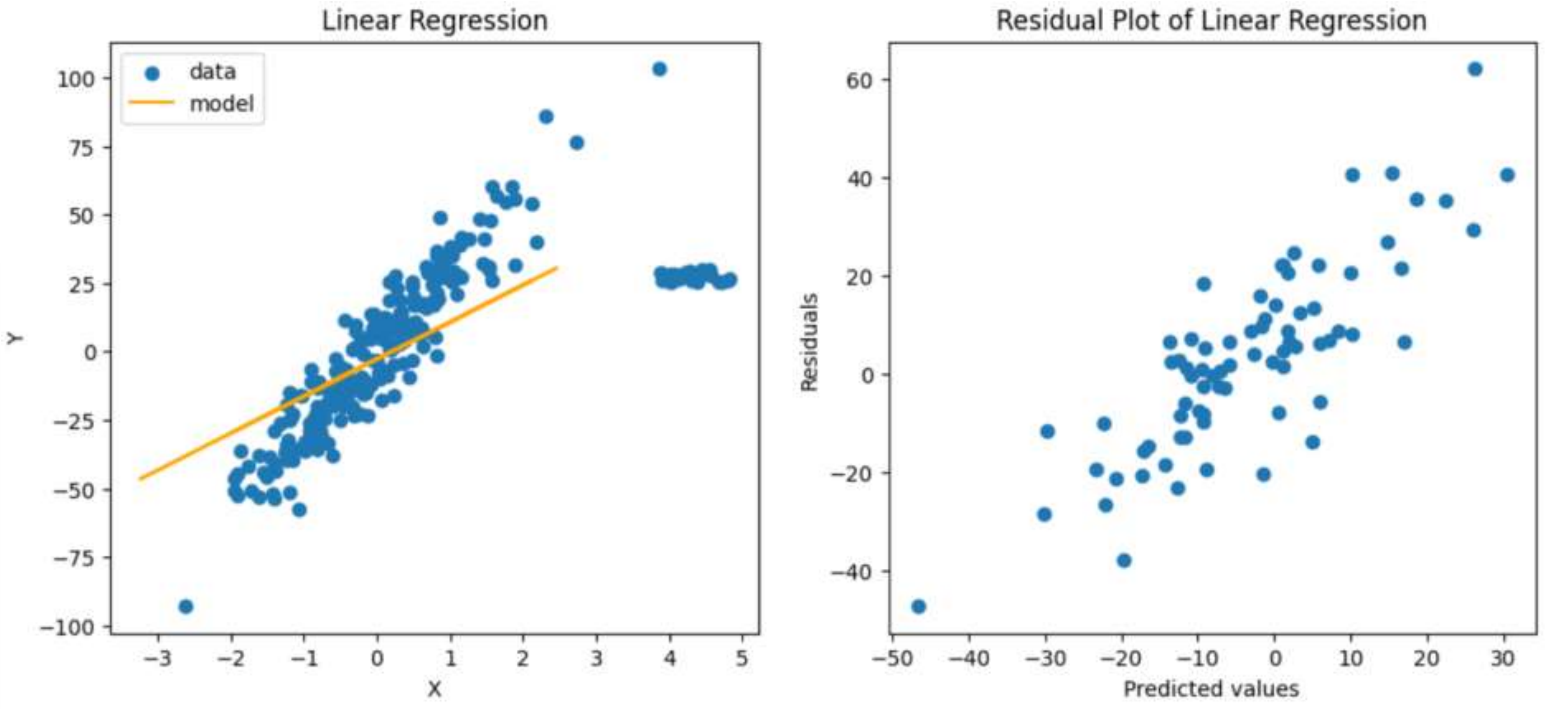
25 points are outliers

Testing set:

75 data points

TOPIC 3

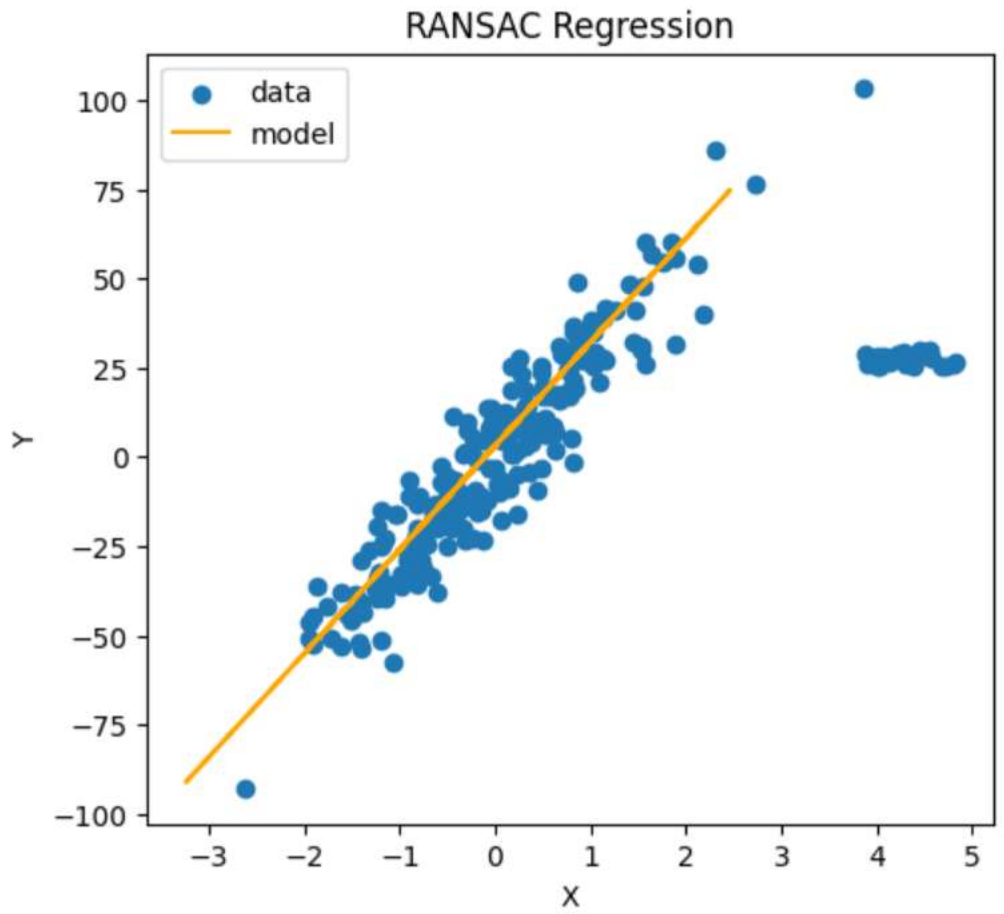
LINEAR REGRESSION



Loss function: $\mathrm{MSE} = (1/n) * \Sigma(y_i - \hat{y}_i)^2$

TOPIC 3

RANSAC REGRESSION



1. Randomly select a subset of examples from the dataset as inliers and train the model.
2. Test the remaining data points against the trained model.
3. Identify data points as inliers if they fall within a specified tolerance, typically measured using the median absolute deviation.

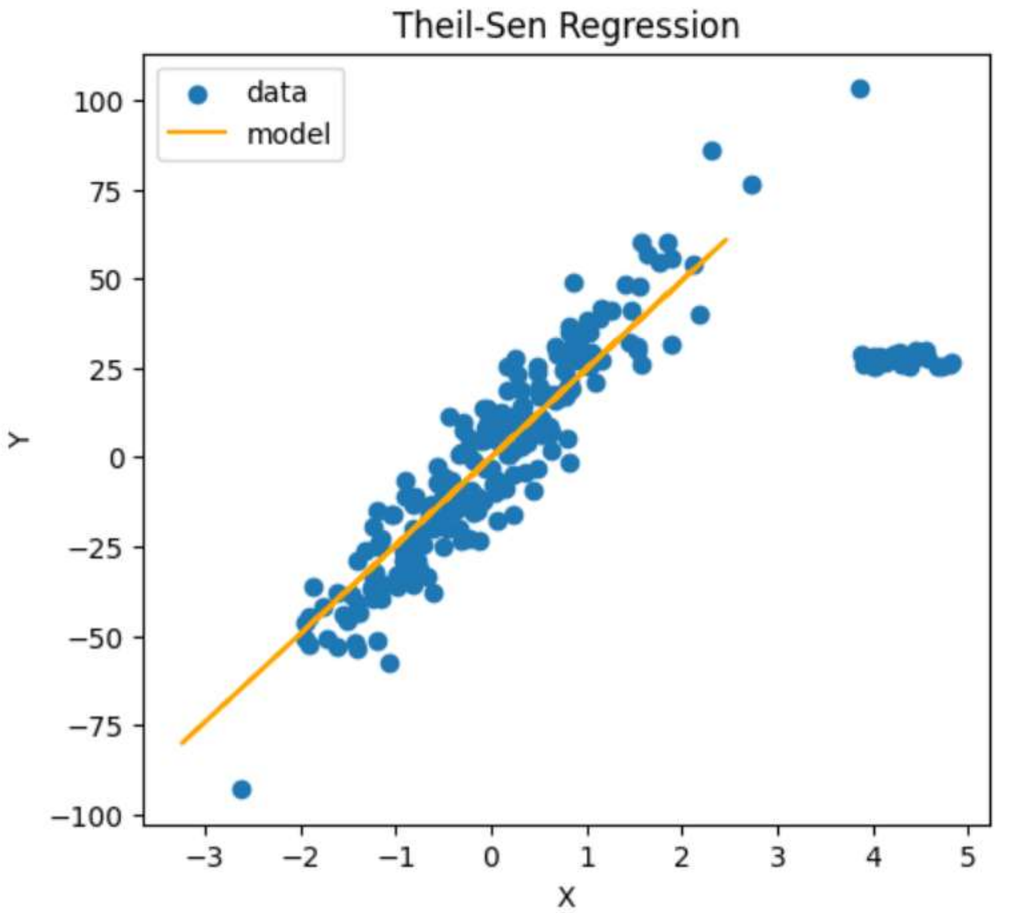
4. Retrain the model using only the inlier data points.

5. Estimate the error of the retrained model compared to the inliers.

6. Repeat steps 1 to 5 iteratively until a termination condition is met.

TOPIC 3

THEIL-SEN REGRESSION



1. Calculate the slopes between each pair of points in the data.

   For a pair of points, (xi, yi) and (xj, yj),

   s = (yj−yi)/(xj−xi) if xi ≠ xj

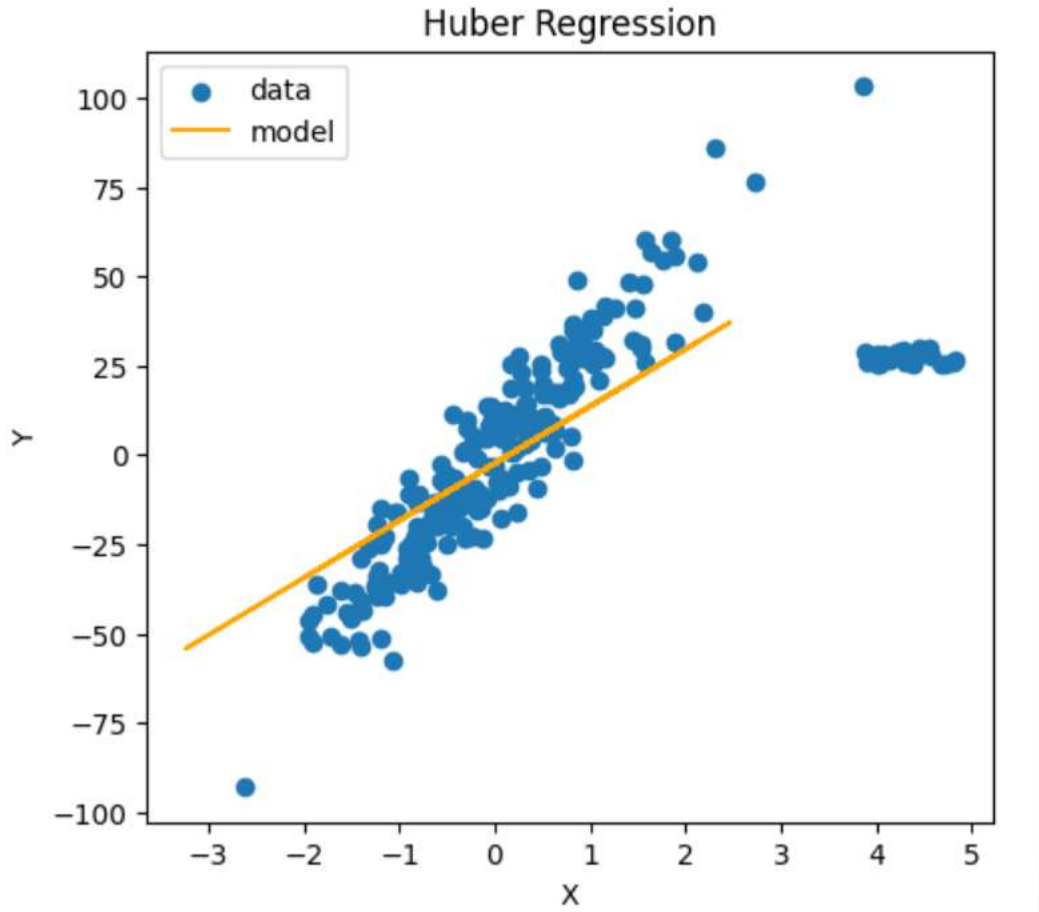
1. Sort the calculated slopes in ascending order.

2. The final slope is then defined as the spatial median of these slopes .

VRIJE UNIVERSITEIT BRUSSEL

TOPIC 3

HUBER REGRESSION



Optimized loss function:

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for} |y - f(x)| \le \delta, \\ \delta \left(|y - f(x)| - \frac{1}{2}\delta\right), & \text{otherwise.} \end{cases}$$

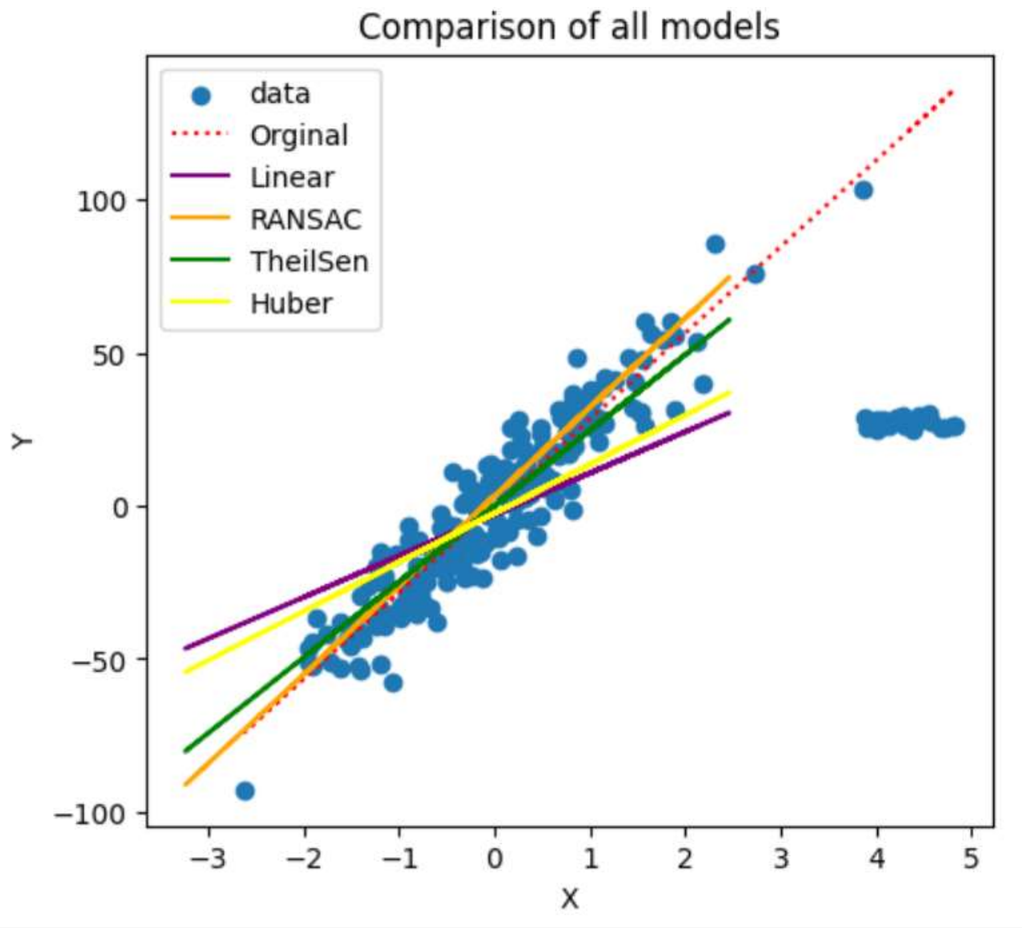
The loss is the half of the square of the usual residual (y-f(x)) only when the absolute value of the residual is smaller than the threshold called Huber parameter.

When the residual is larger than the threshold, the loss is a function regarding the absolute value of the residual and the Huber parameter.

VRIJE UNIVERSITEIT BRUSSEL

TOPIC 3

RESULTS AND CONCLUSION



	Original Dataset	Linear	RANSAC	Theil-Sen	Huber
Coefficient	28.20	13.51	29.08	24.72	16.01

RANSAC: takes care of removing outliers from the training data set while fitting the model.

Theil-Sen: estimates the slope by calculating the median of all possible pairwise slopes between the data points.

Huber: combines squared error loss for small errors and absolute error loss for large errors, using a specified threshold to balance between the two.