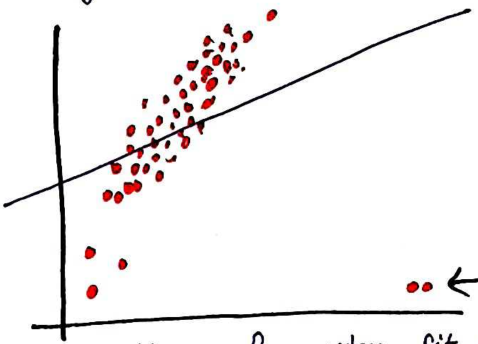
A Simple Technique to Robustify Linear Regression to Outliers

The biggest problem with most regression model is than they are sensitive to outliers.

Consider linear Regression , for instance.

Huber Regression

Even a few outliers can significantly impact Linear Regression performance as show below.

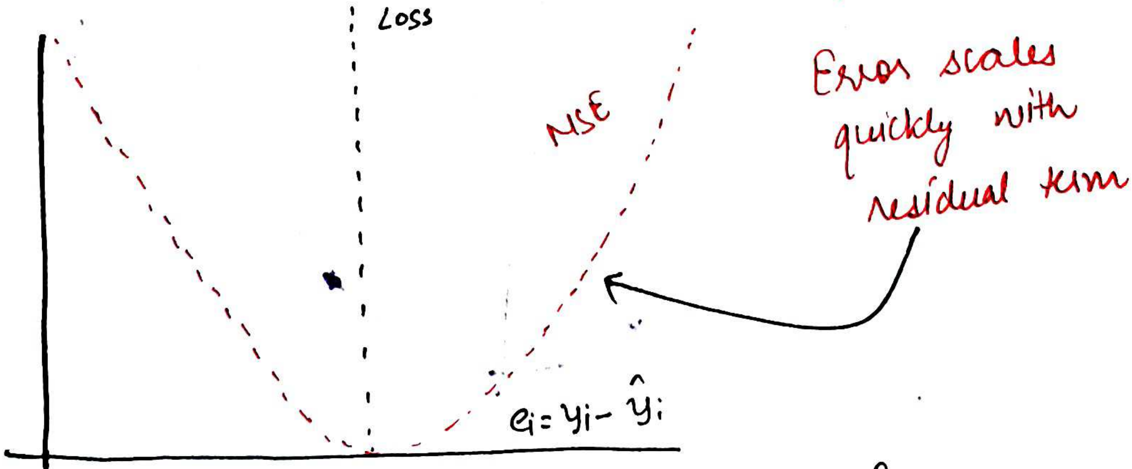


Just two outliers disort the regression fit

Linear Regression fit affected by outliers

And it isn't hard to identify the cause of this problem.

Essentially, the loss function (MSE) scales quickly with the residual term (true- predicted).



Error scales quickly with residual term

Mean squared Error (MSE) loss function

Thus, even a few data points with a large residual can impact parameter estimation.

Huber loss (or Huber Regression) precisely address this problem

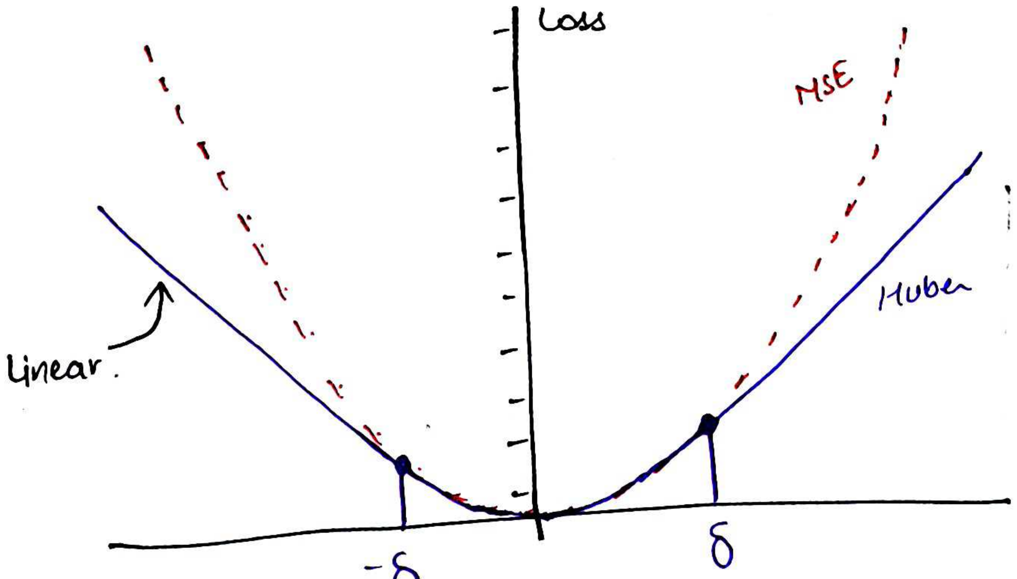
In a gist, it attempts to reduce the error contribution of data points with large residuals.

How?

One simple, intuitive and obviously way to do this is by applying a threshold (δ) on the residual term:

- If the residual is smaller than the threshold, use MSE (no change here).
- Otherwise, use a loss function which has a smaller output than MSE linear, for instance.

This is depicted below:



Huber vs MSE loss function

- For residuals smaller than the threshold (δ), we use MSE.
- Otherwise, we use a linear loss function which has a smaller output than MSE.

Mathematically, Huber loss is defined as follows:

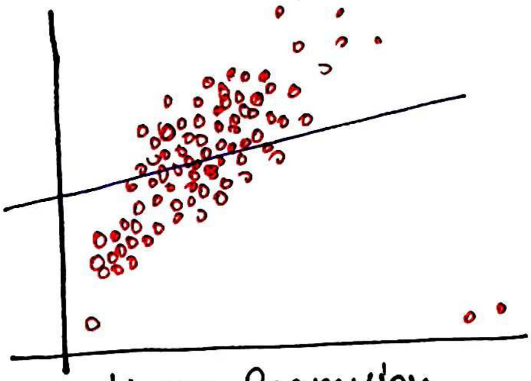
$$\boxed{e_i = y_i - \hat{y}_i}$$

MSE

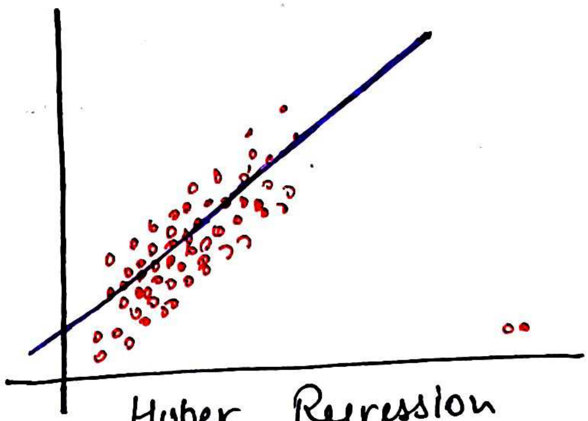
$$L_{Huber}(e_i) = \begin{cases} \frac{1}{2} e_i^2 & : e_i \le \delta \\ \delta (|e_i| - \frac{1}{2}\delta) & : \text{otherwise} \end{cases}$$

Linear

Its effectiveness is evident from the image below:



Linear Regression



Huber Regression

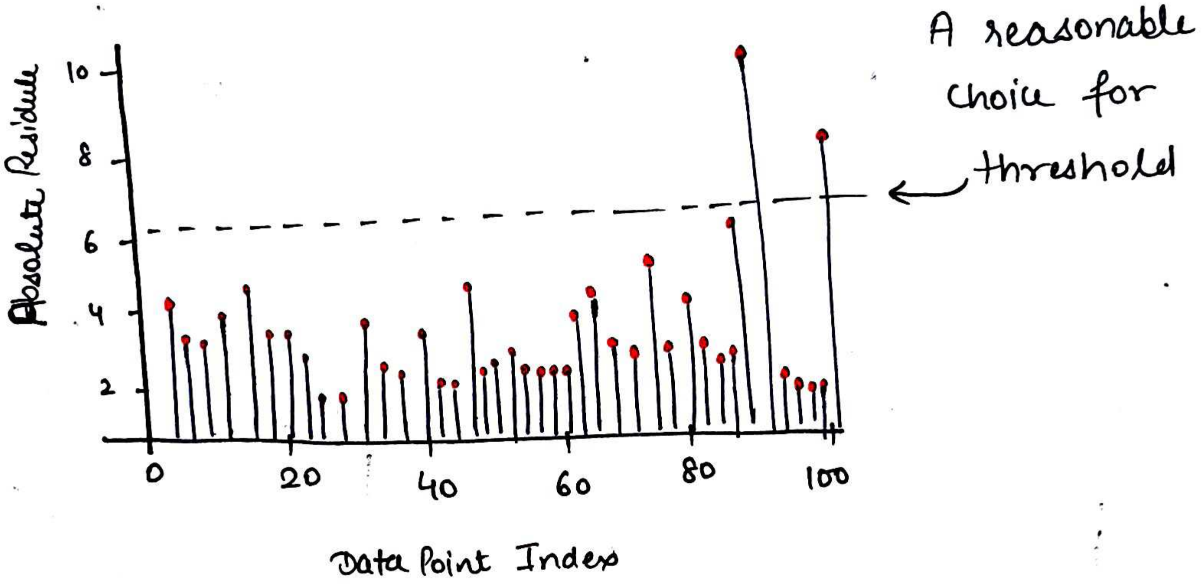
- Linear Regression is affected by outliers
- Huber Regression is more robust.

Now, I know what you are thinking

How do we determine the threshold (δ)?

While trial and errors is one way, I often like to create a residual plot. This a depicted below:

The below plot is generally called a lollipop plot because of its appearance.



- Train a linear regression model as you usually would.
- Compute the residuals (= true - predicted) on the training data.
- Plot the absolute residuals for every data point.

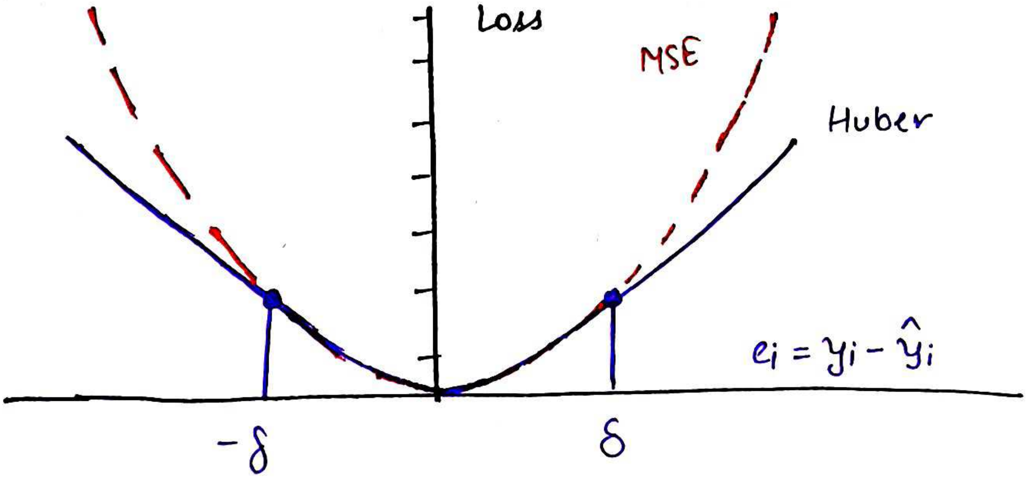
One good thing is that we can create this plot for any dimensional dataset. The objective is just to plot (true - predicted) values, which will always be 1D.

Next, you can subjectively decide a reasonable threshold value δ.

In fact, here's another interesting idea.

By using a linear loss function in Huber regressor, we intended to reduce the large error contributions that would have happened otherwise by using MSE.

Thus, we can further reduce that error contribution by using, say, a square root loss function, as shown below:


Loss
MSE
Huber
$e_i = y_i - \hat{y}_i$
$-\delta$
δ