# Regression Model Metric

- In my Previous notebook i am focussed on classification metrics like precision, recall, AUC etc.
- For a change, I wanted to explore all kinds of metrics including those used in regression as well. MAE and RMSE are the two most popular metrics for continuous variables. Let's start with the more popular one.
- All Types Of Metric Used In Regression

    1. Mean Squared Error(MSE)
    2. Root-Mean-Squared-Error(RMSE).
    3. Mean-Absolute-Error(MAE).
    4. Root Mean Squared Logarithmic Error (RMSLE).
    5. R² or Coefficient of Determination.
    6. Adjusted R²

**1. Mean Squared Error(MSE) :**

- MSE or Mean Squared Error is one of the most preferred metrics for regression tasks.
- It is simply the average of the squared difference between the target value and the value predicted by the regression model.
- As it squares the differences, it penalizes even a small error which leads to over-estimation of how bad the model is. It is preferred more than other metrics because it is differentiable and hence can be optimized better.

$$\underbrace{\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}}_{\text{test set}} \underbrace{(y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i)^2}_{\text{actual value}}$$

**2. Root Mean Squared Error (RMSE) :**

- It is simple root of MSE.
- RMSE is the most widely used metric for regression tasks and is the square root of the averaged squared difference between the target value and the value predicted by the model.
- It is preferred more in some cases because the errors are first squared before averaging which poses a high penalty on large errors. This implies that RMSE is useful when large errors are undesired.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

**3. Mean-Absolute-Error(MAE) :**

- MAE is the absolute difference between the target value and the value predicted by the model.
- The MAE is more robust to outliers and does not penalize the errors as extremely as mse. MAE is a linear score which means all the individual differences are weighted equally. It is not suitable for applications where you want to pay more attention to the outliers.

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

**4. Root Mean Squared Logarithmic Error (RMSLE) :**

- In case of Root mean squared logarithmic error, we take the log of the predictions and actual values. So basically, what changes are the variance that we are measuring.
- RMSLE is usually used when we don't want to penalize huge differences in the predicted and the actual values when both predicted and true values are huge numbers.

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(y_i+1) - \log(\hat{y}_i+1))^2}$$

***All Four Metric***

$$\text{MSE} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{test set}}\underbrace{(y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i)^2}_{\text{actual value}}$$

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(y_i+1) - \log(\hat{y}_i+1))^2}$$

**5. $R^2$ or Coefficient of Determination:**

- Coefficient of Determination or $R^2$ is another metric used for evaluating the performance of a regression model.
- The metric helps us to compare our current model with a constant baseline and tells us how much our model is better.
- The constant baseline is chosen by taking the mean of the data and drawing a line at the mean. $R^2$ is a scale-free score that implies it doesn't matter whether the values are too large or too small, the $R^2$ will always be less than or equal to 1.

Sum Squared Regression Error

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$

Sum Squared Total Error

**6. Adjusted $R^2$:**

- it is suitable for multiple linear regrssion where more than one independent variable.
- Adjusted $R^2$ depicts the same meaning as $R^2$ but is an improvement of it.
- $R^2$ suffers from the problem that the scores improve on increasing terms even though the model is not improving which may misguide the researcher.
- Adjusted $R^2$ is always lower than $R^2$ as it adjusts for the increasing predictors and only shows improvement if there is a real improvement.

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

$R^2$ = sample R-square

p = Number of predictors

N = Total sample size.

## Q. Why is $R^2$ Negative ?

- There is a misconception among people that $R^2$ score ranges from 0 to 1 but actually it ranges from $-\infty$ to 1. Due to this misconception, they are sometimes scared why the $R^2$ is negative which is not a possibility according to them.

The main reasons for $R^2$ to be negative are the following:

- $R^2$ compares the fit of the chosen model with that of a horizontal straight line (the null hypothesis). If the chosen model fits worse than a horizontal line, then $R^2$ is negative.

- Note that $R^2$ is not always the square of anything, so it can have a negative value without violating any rules of math.
- Mainly $R^2$ is negative only when the chosen model does not follow the trend of the data, so fits worse than a horizontal line.
- Maybe their area a large number of outliers in the data that causes the mse of the model to be more than mse of the baseline causing the $R^2$ to be negative(i.e the numerator is greater than the denominator).

In this Notebook, i discovered about the various metrics used in regression analysis and also tried to answer the question of why $R^2$ is negative?