# Error metrics for evaluating Linear Regression Model

## Introduction

For evaluating models, we have various error metrics for us to calculate and compare. The best model is the one which gives us more reliable prediction. To decide that we look at the dependent variable's Observed values and the Predicted values using the model of the test samples.

The common metrics we use for evaluating Linear Regression Model are as given below.

1. Mean Squared Error (MSE)
2. Mean Absolute Error (MAE)
3. Root Mean Square Error ( RMSE)
4. Mean Absolute Percentage Error (MAPE)
5. $R^2$

# Mean Squared Error (MSE)

The Mean Squared Error is calculated by the formula as given below

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

where

y represents the observed value of dependent variable of the test samples

$\hat{y}$ represents the predicted value of the dependent variable of the test samples using the model

n represents the sample size

The MSE, is nothing but the mean of the squared difference between the actual/observed value and the predicted value (nothing but the error in prediction). So if the error is minimal, ie predicted value reflects the observed value then the MSE will be smaller. Here the error is squared and it amplifies the same.

The **smaller the MSE, the better the fit** of the model. In other words, MSE is **nearer to zero** then the model is efficient in prediction. Theoretically the least possible value is zero, but in real life achieving that is near impossible as in most scenarios, the multiple dependent variables influence the dependent variables and capturing all the dependent variables is a resource intensive and sometimes the cost benefit ratio will not support.

# Mean Absolute Error (MAE)

The Mean Absolute Error is calculated by the formula as given below

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where

y represents the observed value of dependent variable of the test samples

$\hat{y}$ represents the predicted value of the dependent variable of the test samples using the model

n represents the sample size

The MAE, is nothing but the absolute difference between the actual/observed value and the predicted value (nothing but the absolute error in prediction). So if the error is minimal, ie predicted value reflects the observed value then the MAE will be smaller. Compared to MSE, MAE will be smaller as the error is not amplified in the case of MAE.

The **smaller the MAE, the better the fit** of the model. In other words, MAE is **nearer to zero** then the model is efficient in prediction. Theoretically the least possible value is zero, but in real life achieving that is near impossible as for the same reason mentioned above under the MSE.

# Root Mean Square Error (RMSE)

The Root Mean Square Error is calculated by the formula as given below

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

where

y represents the observed value of dependent variable of the test samples

$\hat{y}$ represents the predicted value of the dependent variable of the test samples using the model

n represents the sample size

The RMSE, is nothing but the root of the MSE. So as similar to MSE, if the error is minimal, ie predicted value reflects the observed value then the RMSE will be smaller. Compared to MSE, RMSE will be smaller as the amplification of error is removed due to the application of root. Also, to be noted, the RMSE and the MAE will be closer to each other as both are reflecting the error in absolute terms, ie squared and then taken root in case of RMSE and absolute difference in the case of MAE.

The **smaller the RMSE, the better the fit** of the model. In other words, RMSE is **nearer to zero** then the model is efficient in prediction.

# Mean Absolute Percentage Error (MAPE)

The Mean Absolute Percentage Error (MAPE) is calculated by the formula as given below

$$MAPE = \left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|\right) * 100$$

where

y represents the observed value of dependent variable of the test samples

$\hat{y}$ represents the predicted value of the dependent variable of the test samples using the model

n represents the sample size

The MAPE, is the mean of absolute difference of error in prediction divided by the actual value in percentage terms. MAPE works well when there is no outliers and there is no zero values.

In Sklearn.metrics (scikit-learn 0.24), this is being introduced in the latest version in development and the MAPE will be returned as a float and the same needs to be converted to percentage. If the actual value is zero, then instead of returning as ∞, large value of float is returned. So the range of returned value is 0.0 to 1.0 and in percentage terms, it is 0 to 100%.

So as similar to MAE, if the error is minimal, ie predicted value reflects the observed value then the MAPE will be smaller. The best possible value is zero.

The **smaller the MAPE, the better the fit** of the model. In other words, MAPE is **nearer to zero** then the model is efficient in prediction.

# $R^2$ – the coefficient of determination

$R^2$, the coefficient of determination represents the proportion of variance of the dependent variable that has been explained by the independent variables in the model.

The $R^2$ is calculated as given below

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where

y represents the observed value of dependent variable of the test samples

$\hat{y}$ represents the predicted value of the dependent variable of the test samples using the model

$\bar{y}$ represents the mean of the test samples

n represents the sample size

Higher the $R^2$ value for the model, indicates the best fit, however to confirm further, we may need to opt for the residual plot, which should indicate the randomness of residual around 0 for the entire range of the dependent variable.

# Conclusion

The error metrics can give us a guidance to select a best fit model and the most used metrics for the regression model are the RMSE and MAE. R2 can give us the understanding of variability of the dependent variable in terms of percentage. However, to get the best fitted regression model, we need to ensure the regression model is applied when there is a linear relationship between the independent and dependent variables. Also need to ensure there is very less or no multi-collinearity between the independent variables. Variance also should be same of all value of residual. As well as Residual (actual – predicted) should be normally distributed.