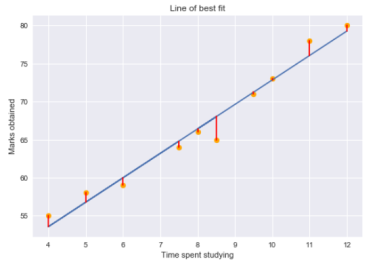
EVALUATION PARAMETERS FOR LINEAR REGRESSION MODEL:-

* R2 (using TSS and RSS)
* Adjusted R2 (using R2, m, n)
* MAE
* MSE
* RMSE
* **RSS**

Residual for a point in the data is the difference between the actual value and the value predicted by our linear regression model.

Residual

Residual plots tell us whether the regression model is the right fit for the data or not. It is actually an assumption of the regression model that there is no trend in residual plots.

Using the residual values, we can determine the sum of squares of the residuals also known as Residual sum of squares or RSS.

Residual Sum of Squares

The lower the value of RSS, the better is the model predictions. Or we can say that – a regression line is a line of best fit if it minimizes the RSS value. But there is a flaw in this – RSS is a scale variant statistic. Since RSS is the sum of the squared difference between the actual and predicted value, the value depends on the scale of the target variable.

Even though the data does not change, the value of RSS varies according to the scale of the target. This makes it difficult to judge what might be a good RSS value. So, can we come up with a better statistic that is scale-invariant? This is where R-squared comes into the picture.

* **R-SQUARE**

R-squared statistic or coefficient of determination is a scale invariant statistic that gives the proportion of variation in target variable explained by the linear regression model.

* **TSS**

Total variation in target variable is the sum of squares of the difference between the actual values and their mean.

Total Sum of Squares

TSS or Total sum of squares gives the total variation in Y. We can see that it is very similar to the variance of Y. While the variance is the average of the squared sums of difference between actual values and data points, TSS is the total of the squared sums.

Now that we know the total variation in the target variable, how do we determine the proportion of this variation explained by our model? We go back to RSS.

RSS gives us the total square of the distance of actual points from the regression line. But if we focus on a single residual, we can say that it is the distance that is not captured by the regression line. Therefore, RSS as a whole gives us the variation in the target variable that is not explained by our model.

* **R-SQUARE**

Now, if TSS gives us the total variation in Y, and RSS gives us the variation in Y not explained by X, then TSS-RSS gives us the variation in Y that is explained by our model! We can simply divide this value by TSS to get the proportion of variation in Y that is explained by the model. And this our R-squared statistic!

R-squared = (TSS-RSS)/TSS

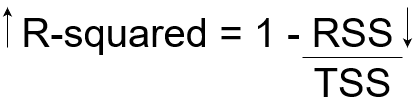
= Explained variation/ Total variation

= 1 – Unexplained variation/ Total variation

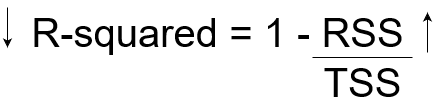
So R-squared gives the degree of variability in the target variable that is explained by the model or the independent variables. If this value is 0.7, then it means that the independent variables explain 70% of the variation in the target variable.

R-squared value always lies between 0 and 1. A higher R-squared value indicates a higher amount of variability being explained by our model and vice-versa.

If we had a really low RSS value, it would mean that the regression line was very close to the actual points. This means the independent variables explain the majority of variation in the target variable. In such a case, we would have a really high R-squared value.



On the contrary, if we had a really high RSS value, it would mean that the regression line was far away from the actual points. Thus, independent variables fail to explain the majority of variation in the target variable. This would give us a really low R-squared value.



Thus,R-squared value gives us the variation in the target variable given by the variation in independent variables.

* **ADJUSTED R-SQUARE**

The R-squared statistic isn’t perfect. In fact, it suffers from a major flaw. Its value never decreases no matter the number of variables we add to our regression model. That is, even if we are adding redundant variables to the data, the value of R-squared does not decrease. It either remains the same or increases with the addition of new independent variables. This clearly does not make sense because some of the independent variables might not be useful in determining the target variable. Adjusted R-squared deals with this issue.

The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.

Adj.R2 = 1 – [(1-R2)(m-1) /(m-n-1) ]

So, if R-squared does not increase significantly on the addition of a new independent variable, then the value of Adjusted R-squared will actually decrease.

On the other hand, if on adding the new independent variable we see a significant increase in R-squared value, then the Adjusted R-squared value will also increase.

If, adding a random independent variable did not help in explaining the variation in the target variable. Our R-squared value remains the same. Thus, giving us a false indication that this variable might be helpful in predicting the output. However, the Adjusted R-squared value decreased which indicated that this new variable is actually not capturing the trend in the target variable.

Clearly, it is better to use Adjusted R-squared when there are multiple variables in the regression model. This would allow us to compare models with differing numbers of independent variables.