Overall Model Quality

How do we measure the model performance, how do we know that we have a good fit?

There are number of parameters we can check to assess the accuracy of our model. The most knowns are R^2 value and RMSE.

Coefficient of determination: R^2 value

R-Squared determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

It does not indicate the correctness of the regression model. Therefore, it should always be analyzed with the other variables in a statistical model.

R^2 is a proportion with a value between 0 & 1 (R^2 e [0,1]):

– If $R^2 = 0$, predictor accounts for none of the variation in target

– If $R^2 = 1$, predictor accounts for all of the variation in target

The most common interpretation of r-squared is how well the regression model fits the observed data. For example, an r-squared of 60% reveals that 60% of the data fit the regression model. Generally, a higher r-squared indicates a better fit for the model.

However, it is not always the case that a high r-squared is good for the regression model. The quality of the statistical measure depends on many factors, such as the nature of the variables employed in the model, the units of measure of the variables, and the applied data transformation. Thus, sometimes, a high r-squared can indicate the problems with the regression model.

A low r-squared figure is generally a bad sign for predictive models. However, in some cases, a good model may show a small value.

There is no universal rule on how to incorporate the statistical measure in assessing a model. The context of the experiment or forecast is extremely important and, in different scenarios, the insights from the metric can vary.

The formula for calculating R-squared is:

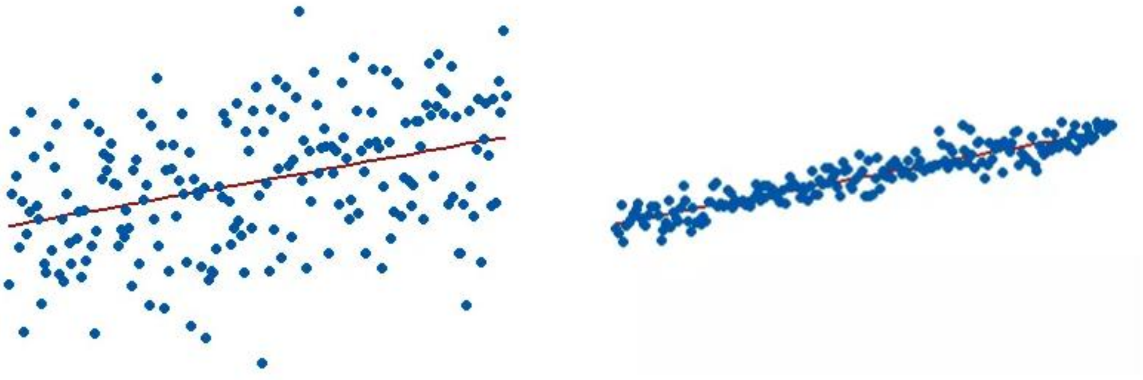
$$\text{R-Squared} = \frac{SS_{regression}}{SS_{total}}$$

SSregression is the sum of squares due to regression (explained sum of squares)

SStotal is the total sum of squares

The sum of squares due to regression measures how well the regression model represents the data that were used for modeling. The total sum of squares measures the variation in the observed data (data used in regression modeling).

Graphical Representation



The R-squared for the regression model on the left is 15%, and for the model on the right it is 85%. When a regression model accounts for more of the variance, the data points are closer to the regression line. In practice, you'll never see a regression model with an R2 of 100%. In that case, the fitted values equal the data values and, consequently, all of the observations fall exactly on the regression line.

RMSE: Root mean squared error

RMSE measures the standard deviation of the residuals (the spread of the points about the fitted regression line). In other words, it tells you how concentrated the data is around the line of best fit.

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

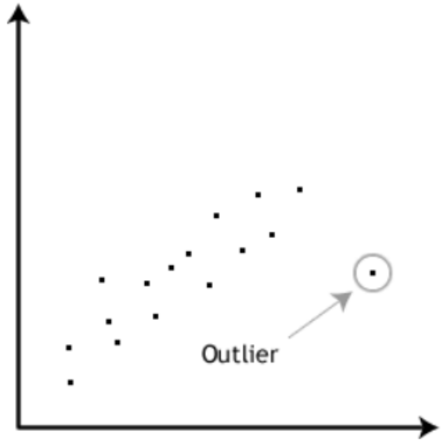
$$\begin{aligned} \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n &\text{ are predicted values} \\ y_1, y_2, \ldots, y_n &\text{ are observed values} \\ n &\text{ is the number of observations} \end{aligned}$$

Lower values of RMSE indicate a better fit as a measure of model accuracy.

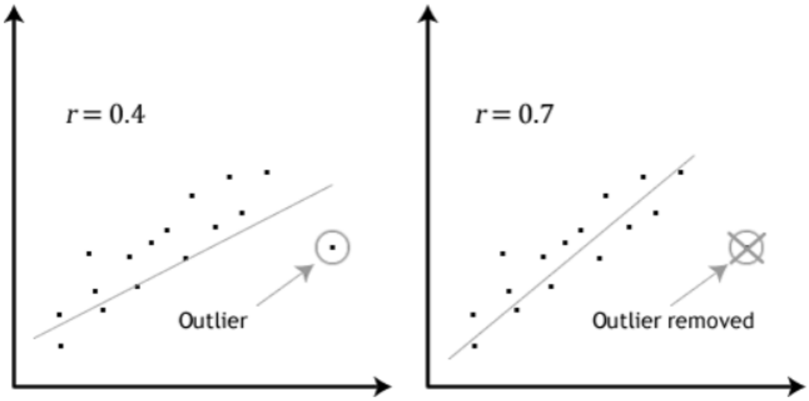
There is no absolute criterion for a good value of RMSE. It depends on the units in which the variable is measured and on the degree of forecasting accuracy.

Outliers

An outlier is an observation of data that does not fit the rest of the data. It is sometimes called an extreme value. When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening.



This point above is an outlier that is included according to both the best fit line and the correlation coefficient. Included in the model and after removal from the model can be seen in the graphs below:


$r = 0.4$
Outlier
$r = 0.7$
Outlier removed