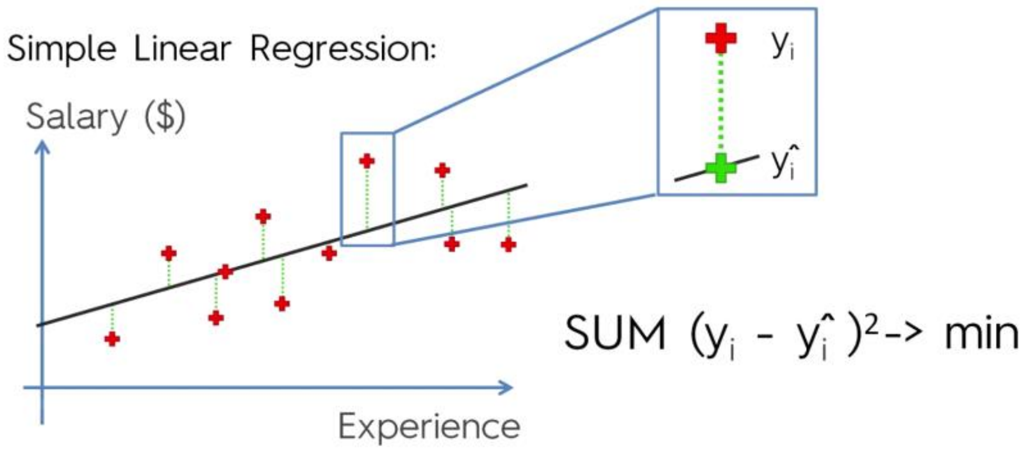
R Squared and Adjusted R Squared Intuition

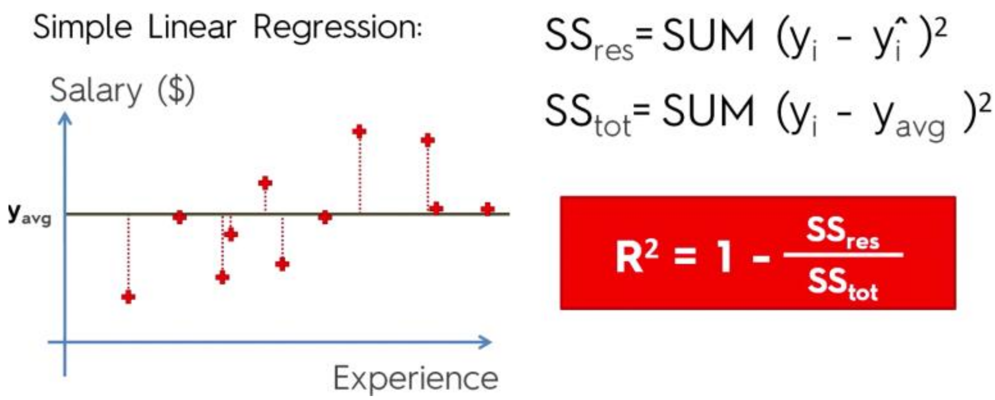
Saturday, January 21, 2017 9:10 AM

R-Squared

Consider a Simple Linear Regression Model



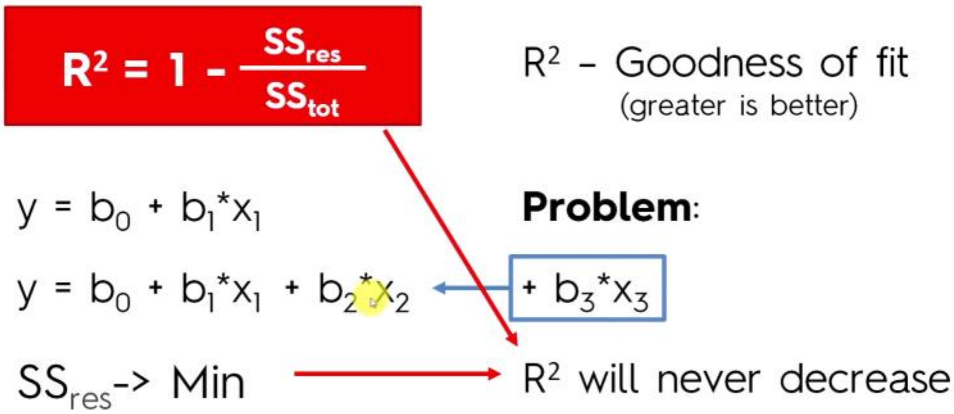
Here, the line represents the Simple Linear Regression Line.
The root mean squared error is given by the formulae stated above and we try to minimize it when predicting parameters.
This is called the Sum of Squares of Residuals. This represents SS(res).
Now, instead let us draw a line which pertains to the average y value. This is a line which will be parallel to the X-axis.
Now, the sum of squares of this value difference is called the total sum of squares.



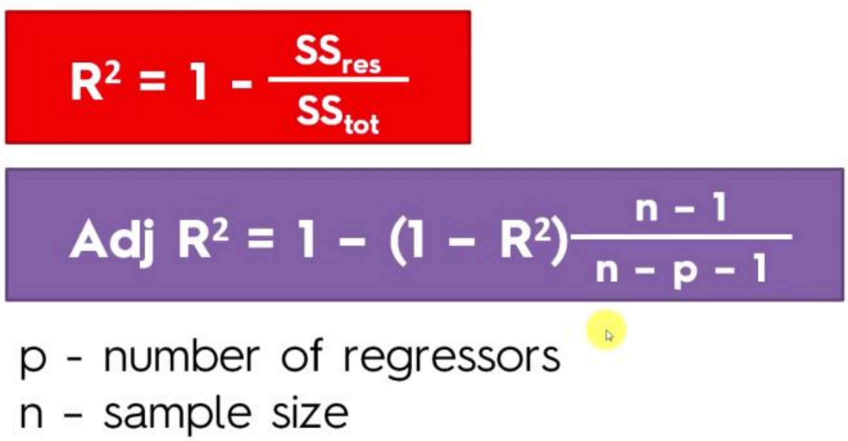
Based on this, what R2 is telling us is that how good our regression line is compared to the average line.
Because, we are trying to minimize SS(res) and SS(tot) is a constant for a given set of values.
So ideally, if our SS(res) is 0, then R2 value = 1 (Best possible line)
R2 can be negative if SS(res) fits the data worse than the average line.
Hence, R2 is an assessment of how good our prediction is.

Adjusted R-Squared

According to R-squared, the best fitting model for prediction is the one which minimizes the Sum of Squares of Residuals.
R2 - goodness of fit - bigger the better



Consider a multi variable regression. Here, suppose we add another variable to the equation x3. Now, the regression function is to decrease SS(res). Hence, the model will give a value to b3 which actually decreases the SS(res) of the model thereby increasing R2. Or, the model will make b3 = 0, making sure that the SS(res) value will not increase, thereby making the value of R2 stay constant.
Note - Most often, there is some sort of basic correlation even between two random entities.
The problem with this is by adding new variables, we will not know if the variables are actually impacting the model or not and to what extent.
For this reason, we use adjusted R-squared.



R2 will never decrease !
The basic conception over here is that, as we add more regressors (variables) to the model, the value of R2 generally tends to increase thereby leading to increase in the value of Adjusted R-squared.
But on the other hand, as we increase the number of regressors, the value of p increases thereby leading to lowering the value of Adjusted R-squared.
Hence, there is a battle on our hands between adding the variables or not adding the variables. Because an insignificant increase in R2 is penalized by a relatively significant increase in p.
Thus, adjusted R2 helps us identify whether we are adding good variables to our model or not. It acts as a tradeoff metric of sorts which strives to capture the essence of using minimum number of features to make the best possible prediction. (All features used by the model have a good impact on the prediction the model makes)