Key Difference between R-squared and Adjusted R-squared for Regression Analysis

What is R?

Pearson r correlation: Pearson *r* correlation is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables. For example, in the stock market, if we want to measure how two stocks are related to each other, Pearson *r* correlation is used to measure the degree of relationship between the two. The point-biserial correlation is conducted with the Pearson correlation formula except that one of the variables is dichotomous. The following formula is used to calculate the Pearson *r* correlation:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{}$$

$$\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}$$

r_{xy} = Pearson r correlation coefficient between x and y

n = number of observations

x_i = value of x (for ith observation)

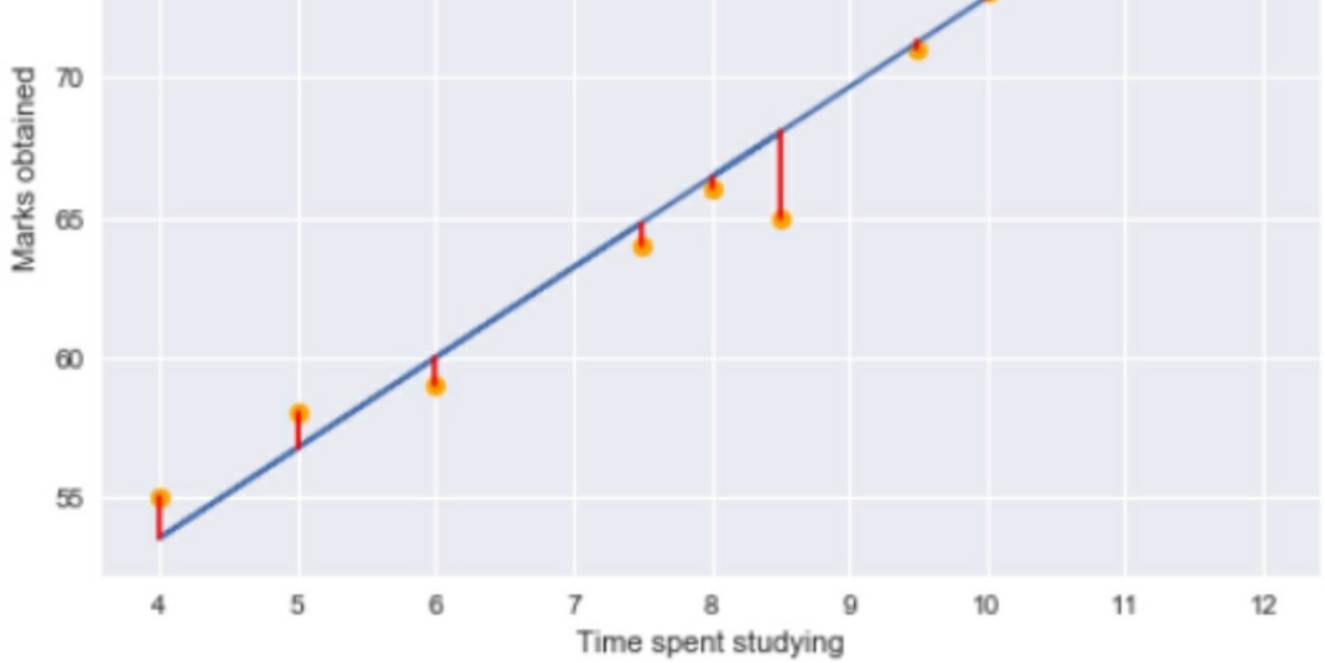
y_i = value of y (for ith observation)

Residual Sum of Squares

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

$$Residual = actual - predicted = y\_\hat{y}$$

$$Residual = actual - predicted = y\_\hat{y}$$



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. To study the assumptions of linear regression in detail, I suggest going through [this great article](#)!

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

$$RSS = \sum (y_i - \hat{y}_i)^2$$

Understanding R-squared statistic

> *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.*

This might seem a little complicated, so let me break this down here. In order to determine the proportion of target variation explained by the model, we need to first determine the following-

1. **Total Sum of Squares**

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

$$TSS = \sum (y_i - \bar{y})^2$$

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.

> *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

variation in the target variable.

R-squared value always lies between 0 and 1.

$$\uparrow R\text{-squared} = 1 - \frac{RSS}{TSS}\downarrow$$

Problems with R-squared statistic

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.

Adjusted R-squared statistic

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.

Let's have a look at the formula for adjusted R-squared to better understand its working.

$$Adjusted\ R^2 = \{1 - [\frac{(1 - R^2)(n-1)}{(n-k-1)}]\}$$

Here,

- **n** represents the number of data points in our dataset
- **k** represents the number of independent variables, and
- **R** represents the R-squared values determined by the model.

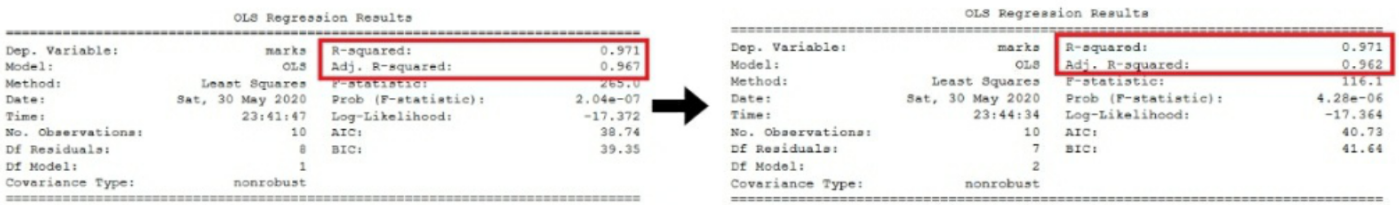
So, if R-squared does not increase significantly on the addition of a new independent variable, then the value of Adjusted R-squared will actually decrease.

$$Adjusted\ R^2 = \{1 - [\frac{(1-R^2)(n-1)}{(n-k-1)}]\}$$

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.

$$Adjusted\ R^2 = \{1 - [\frac{(1-R^2)(n-1)}{(n-k-1)}]\}$$

We can see the difference between R-squared and Adjusted R-squared values if we add a random independent variable to our model.



OLS Regression Results

Dep. Variable:	marks	R-squared:	0.971
Model:	OLS	Adj. R-squared:	0.967
Method:	Least Squares	F-statistic:	265.0
Date:	Sat, 30 May 2020	Prob (F-statistic):	2.04e-07
Time:	23:41:47	Log-Likelihood:	-17.372
No. Observations:	10	AIC:	38.74
Df Residuals:	8	BIC:	39.35
Df Model:	1		
Covariance Type:	nonrobust		

OLS Regression Results

Dep. Variable:	marks	R-squared:	0.971
Model:	OLS	Adj. R-squared:	0.962
Method:	Least Squares	F-statistic:	116.1
Date:	Sat, 30 May 2020	Prob (F-statistic):	4.28e-06
Time:	23:44:34	Log-Likelihood:	-17.364
No. Observations:	10	AIC:	40.73
Df Residuals:	7	BIC:	41.64
Df Model:	2		
Covariance Type:	nonrobust		