

R-SQUARED vs R-SQ.-ADJUSTED



By Chiranjit Pathak

R-Squared explained:

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.

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

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

Total Sum of Squares

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

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

Calculate R-Squared

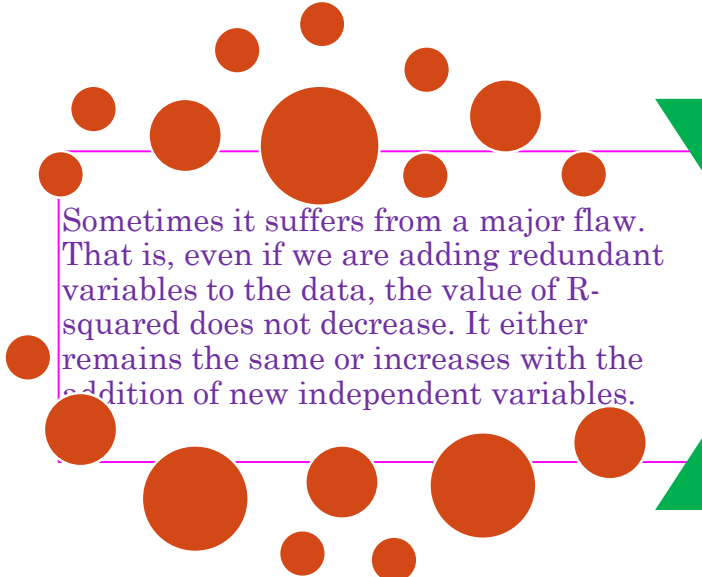
“ $R\text{-squared} = (TSS - RSS) / TSS$
= Explained variation / Total variation
= $1 - \text{Unexplained variation} / \text{Total variation}$

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

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

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.

R-Squared major flaws:



Sometimes it suffers from a major flaw. 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**

Adjusted R-squared deals with this issue.

R-Squared adjusted:

Adjusted R-squared

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.

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

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.

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

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.

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

So adding a random independent variable did not help in explaining the variation in the target variable. R-squared value remains the same.

Thus, giving 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.

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

Thanks for reading!

Lets collaborate and happy to receive any
feedback/suggestion/comment at.....

pathak.chiranjit@gmail.com