Name: Ahmed Nabil Ibrahim Awaad

Coefficient of determination

Introduction

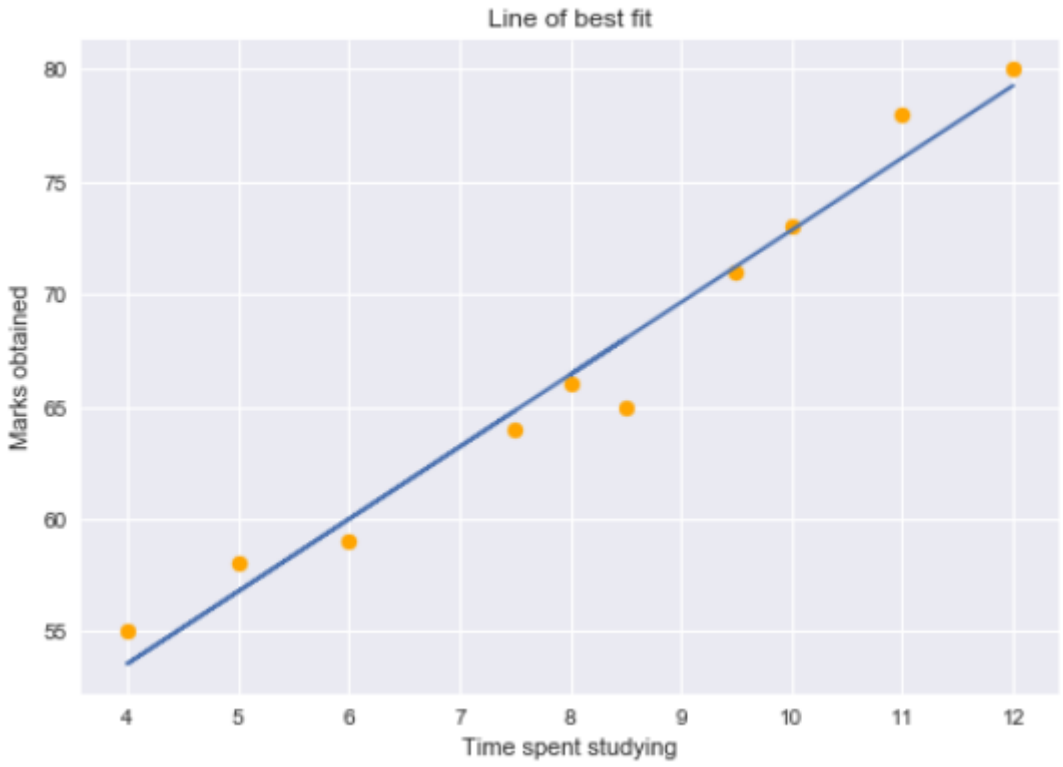
One of the most widely used evaluation metrics for the linear regression models is **R squared** aka **Coefficient of determination**. R squared is considered as a **goodness of fit** metric which in most of the time ranges around 0 to 1. Higher the value of R Squared examined as higher the coherence and predictive ability of the model.

Table of Contents

Residual Sum of Squares

To understand the concepts clearly, we are going to take up a simple regression problem. Here, we are trying to predict the 'Marks Obtained' based on the amount of 'Time Spent Studying'. The **time** spent studying will be our **independent variable** and the **marks achieved** in the test is our **dependent** or **target variable**.

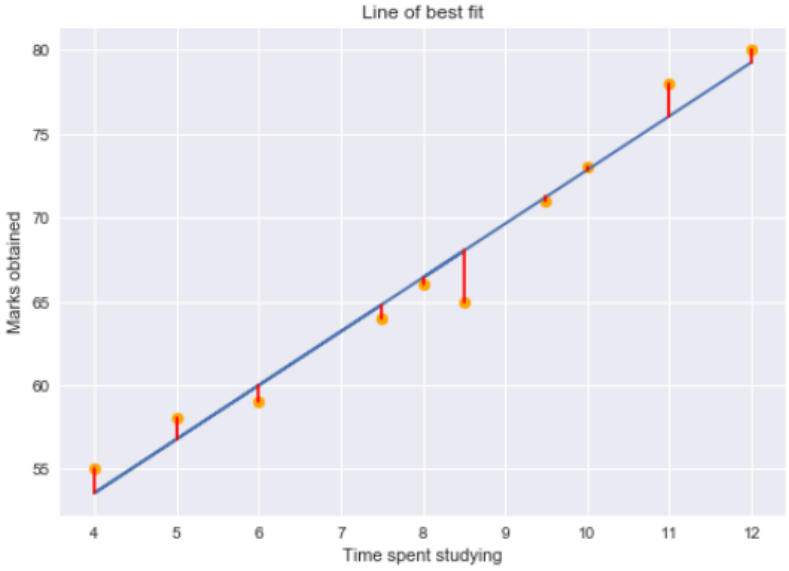
We can plot a simple regression graph to visualize this data.



The yellow dots represent the data points and the blue line is our predicted regression line. As you can see, our regression model does not perfectly predict all the data points. So how do we evaluate the predictions from the regression line using the data? Well, we could start by determining the residual values for the data points.

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

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



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

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

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.

Example:

Consider your target variable is the revenue generated by selling a product. The residuals would depend on the scale of this target. If the revenue scale was taken in “Hundreds of Egyptian Pound”

(i.e. target would be 1, 2, 3, etc.) then we might get an RSS of about 0.54 (hypothetically speaking).

But if the revenue target variable was taken in “Egyptian Pound” (i.e. target would be 100, 200, 300, etc.), then we might get a larger RSS as 5400. 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.

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.

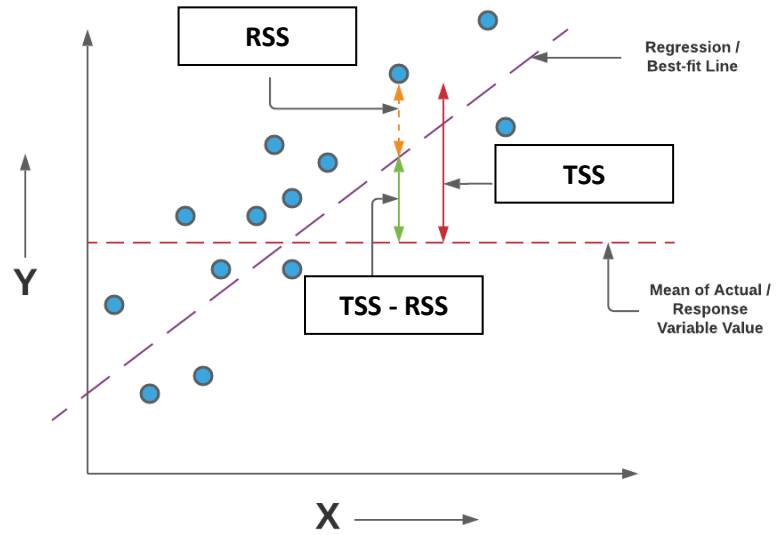
   Now that we know the total variation in the target variable, how do we determine the proportion of this variation explained by our model? We go back to RSS.

2. **Residual Sum of Squares**

As we discussed before, 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.

3. **Calculate R-Squared**

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.

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

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.

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

So, this explains why the R-squared value gives us the variation in the target variable given by the variation in independent variables.

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.

In the example, we had only one independent variable and one target variable but in the real case, we will have 100's of independent variables for a single dependent variable. The actual problem is that, out of 100's of independent variables-

Some variables will have a very high correlation with the target variable.

Some variables will have a very small correlation with the target variable.

Also, some independent variables will not correlate at all.

If there is no correlation then what happens is that — “ Our model will automatically try to establish a relationship with dependent and independent variables and proceed with mathematical calculations assuming that the researcher has already eliminated the unwanted independent variables.”

For example,

For predicting the height of a person, we will have the following independent variables

Weight (High correlation)

Phone number(No correlation)

Location (Low correlation)

Age (High correlation)

Gender (Low correlation)

Here, only weight and age are enough to build an accurate model but the model will assume that the phone number will also influence the height and represent it in a multidimensional space. When a regression plane is built through these 5 independent variables, it’s gradient, intercept, cost and residual will automatically adjust to increase the accuracy. When the accuracy gets increases artificially, obviously R squared will also increase.

In such scenarios, the regression plane will touch all the edges of the original data points in the multidimensional space. It will make the SSR a very small number and that will eventually make the R Squared as a very high number but when test data is introduced, such models will fail miserably.

That is the reason why a high R Squared value does not guarantee an accurate model.

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.

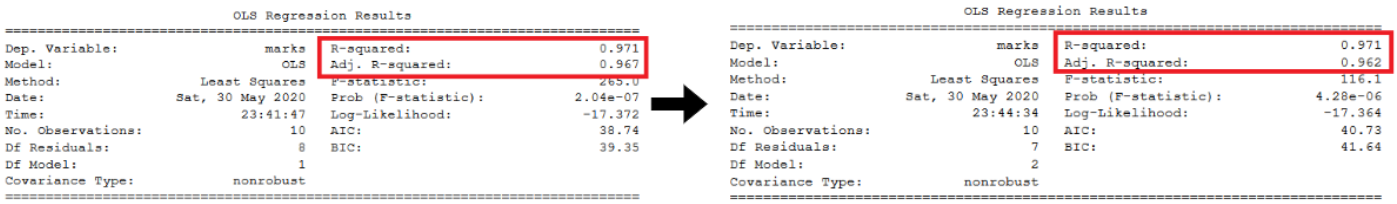
So, if R-squared does not increase significantly on the addition of a new independent variable, then the value of Adjusted R-squared will actually decrease.

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

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.

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

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		

As you can see, 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.

References

1 - https://www.analyticsvidhya.com/blog/2020/07/difference-between-r-squared-and-adjusted-r-squared/

2- https://towardsdatascience.com/the-enigma-of-adjusted-r-squared-57b01edac9f

3- https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp