Adjusted R^2

For regression models we use R^2 or Adjusted R^2 to compare and evaluate models. In this article I have tried to explain what the Adjusted R^2 is, why is it considered more reliable than R^2. For this, let us first understand what R and R^2 are.

We build linear regression model when we observe strong geometrically linear relationship between the predictor and the target variables (Fig 1.b and Fig 1.c). The stronger the relationship between the predictor and the target variable, better will the model be. When there is no relationship between the predictor and target variable, the scatter plot will be more similar to Fig 1.a. We use a metric called R (Coefficient of Correlation) to quantify the strength of relation between the target and predictor variable.

1. The close R is to 0, the weaker is the relationship between target and the predictor variables. In fact, R values greater than 0 but close to 0 may be due to statistical fluke
2. R can range from -1 to + 1. The close R is to -1 or +1, the stronger is the relationship between the target and the predictor variables.

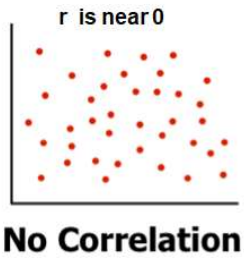

Fig 1.a

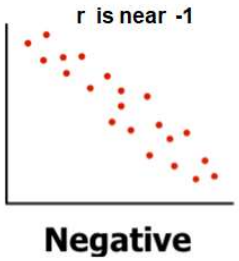

Fig 1.b

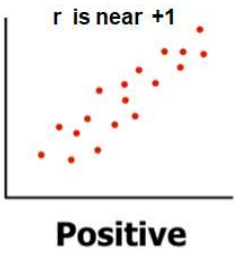

Fig 1.c

Deciphering R (Coefficient of Correlation) Formula

The R value is calculated using the following formula

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

1. This is covariance of X,Y divided by Standard Deviation of X, Y
2. The numerator is measuring how much the X and Y variables influence each other and whether the variables have inverse relationship (negative) or direct relationship (positive)
3. The Covariance for every point is calculated and summed up for the entire data.

To understand what is happening here, let us look at the Figure 2 below where the white lines represent Xbar(vertical) and Ybar(horizontal). These lines split the figure into four quadrants starting with quadrant 1 on top left, quadrant 2 on top right and quadrant 4 on bottom right.

The expression $(X_i - Xbar)*(Y_i - Ybar)$ for a single data point is a simple multiplication of the distances of the point from the Xbar and Ybar respectively. This is formula for area! But…

1. For all points in Q1, the area will be negative because Xi < Xbar for all of them while Yi > Ybar.
2. For points in Q2, the area will be positive because both Xi, and Yi are larger than Xbar and Ybar for all the points
3. For points in Q3, the area will be positive (why?) and for points in Q4 the area will be negative

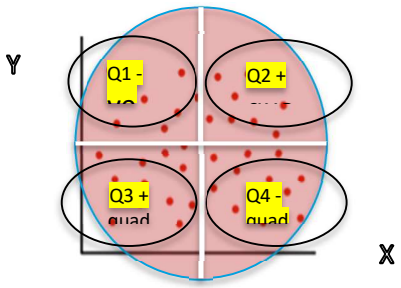

Fig 2

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \quad -$$

4. When we sum up all these areas for all the points from the four quadrants (numerator of covariance), we will get three possible outputs - a positive value or a negative value or a zero
    a. Negative result indicates there are more points in Q1 and Q4 than in Q2 and Q3 indicating a negatively relation

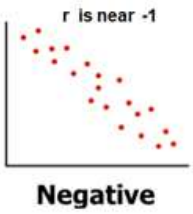

Fig 3

    b. Positive result indicates there are more points in Q2 and Q3 than in Q1 and Q4. This means the two variables are positively related

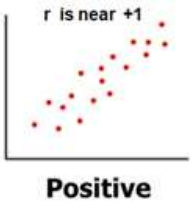

Fig 4

c. Absolute Zero means the number of points in positive quadrants (Q2+Q3) is same as number of points in negative quadrants (Q1+Q4). Indicating no correlation between the two variables.

We need to look at point 4.c a bit more closely. Assuming there is no relation between the two variables X, Y in the population / universe and we take a random sample from the population. When we do a scatter plot between X, Y in the sample, what is the probability that the scatter plot will come absolutely symmetric.

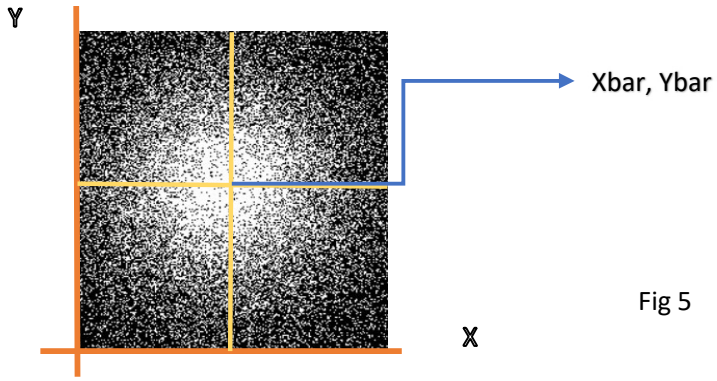

Fig 5

(Image Source : https://www.kaggle.com/mariopasquato/star-cluster-simulations)

Only when the data points are symmetrically distributed around the point (Xbar, Ybar) in the feature space will the correlation come to zero. The question is, what is the probability of getting such a symmetric distribution between X and Y in the sample scatter plot given that there is no correlation between X and Y in the population. The probability is practically zero!

Given that, the sample data will always have asymmetric scatter plot even when the variables X and Y are uncorrelated in the population. Asymmetry in the scatter plot means the summation of the areas of the data points in the four quadrants will result in either the positive values or the negative values. For example, in the figure6 below

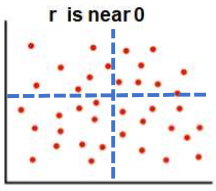

Fig 6

The R value will come close to zero (negative or positive) but not exactly zero. The value of R in this case is not a true correlation (as none really exists). It is the result of statistical chance due to sampling.

Which means R^2 (Coefficient of Determination) also is impacted by statistical chance! During regression model building when attributes are added, the R^2 will keep increasing towards 1. Good attributes with strong correlation with target will increase the R^2 significantly while poor attributes with no true correlation with target will also increase the R^2 but insignificantly. The R^2 metric does not distinguish between true correlation and correlation by chance. To understand why this happens, we need to know what R^2 is.

What is R^2

R^2 is also known as Coefficient of Determination. It is a metric that represents the ratio between

1. the amount of correlation between X and Y in the feature space that is explained / captured by the model and the total correlation in the feature space between X and Y
2. Being a ratio, this will range between 0 and 1 (usually)
    a. R^2 value of zero means the model completely fails to capture any of the available information (this will never happen)
    b. R^2 value of 1 means the model captures all the information in the feature space (this too can never happen unless the model is overfit)
3. Obviously, closer R^2 is to 1 the better the model is. But, the problem with R^2 is, it does not distinguish between true correlation and correlation by chance due to sampling.
4. Which means with every additional feature included in a regression model, R^2 will either remain the same or increase giving an impression that the model is becoming better and better even if the feature added has no true correlation with the target variable in the population

Can R^2 be negative? Yes, it can be:

1. The mathematical formula for R^2 is (1 – (RSS / TSS)). In this expression TSS is total sum of square i.e. total information and RSS is Regression Sum of Squared Residuals.
2. It can be rewritten as (TSS – RSS) / TSS. Here, TSS - RSS is the amount of TSS explained by the model. (TSS – RSS)/ TSS gives us in %age, the amount of information captured by the model.
3. There nothing in the formula that prevents RSS from becoming larger than TSS in magnitude. Thus, R^2 can be negative.
4. R^2 will become negative when there is over regularization of the linear model or when the data in training set and test set have completely different distributions.
5. R^2 is not a simple R X R. This is true only in simple linear regression where we have only one X variable and Y variable.
6. In multi-variate models, R^2 is only a label representing the coefficient of determination of the linear regression. It represents a ratio

What is Adjusted R^2

Adjusted R^2 is a metric which is derived from R^2 after reducing the correlation between X and Y caused by sampling (statistical chance). Every independent variable contributes both true and chance correlation to the model. The good ones have relatively very high true correlation and miniscule chance correlation compared to poor attributes which have higher degree of chance correlation than any true correlation (which actually does not exist). Adjusted R^2 suppresses the chance correlation across all the attributes using the mathematical formula –

$$\bar{R}^2 = 1-\left(1-R^2\right)\left[\frac{n-1}{n-\left(k+1\right)}\right]$$

- R^2 → Coefficient of Determination / Amount of variance captured by the model
- $1 – R^2$ → unexplained / Amount of variance not captured by the model
- n → number of data points
- K → number of attributes (predictor variables in the model)

1. What this formula is doing is jacking up the unexplained variance (1 – R2) using the ratio n-1 / n – (k + 1). Given n and hence n -1 , when we build model sequentially adding one attribute at a time and for every attributed added, calculate the adjusted R2
2. When K =1 (simple linear regression of only one predictor variable), the jack up factor becomes (n-1) /n+2. Suppose n = 10, the ration becomes 9/8 = 1.125. This multiplied by $(1 – R^2)$ has the effect of jacking up unexplained variance by .125 (which is supposed to represent the correlation by chance).
3. If we build the linear regression model with two attributes then the jack up factor becomes 9 / 7 which is 1.285. This will jack up the unexplained variance by .285.
4. The jack up factor will increase with every addition of a new feature as more the features more is the statistically chance correlation in the model.

The net result of this formula is, if we add useless features to a model, R^2 will increase but the Adjusted R^2 will fall. On the other hand if useful features are included in the model, the true correlation is much higher than the chance correlation and the jack up factor will have a very small impact and as a result Adjusted R^2 will increase and so will R^2 .

Summary –

1. R^2 is not as reliable a metric to evaluate and compare linear regression models as adjusted R^2 is
2. R^2 is a symbol. It is not mathematical multiplication (R * R) except in the simple linear regression case
3. R^2 can be negative in rare cases where a model is severely regularized or the data distribution in training is very different from that in the testing
4. All attributes will correlate with target variable but
    a. Good ones will have large true correlation and relatively miniscule chance correlation
    b. Poor attributes will contribute only to the chance correlation

5. Attributes with low correlation need further research to establish whether correlation are real or chance
6. Including attributes with chance correlation will add noise to the model

==========XXXXXXXXXXXXXXXXXXX==========