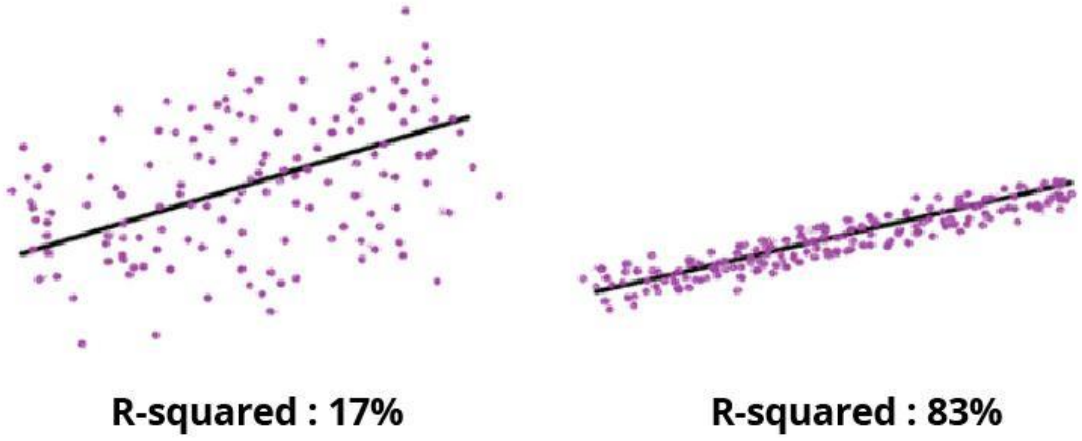
## An Overview of R-squared ($R^2$)

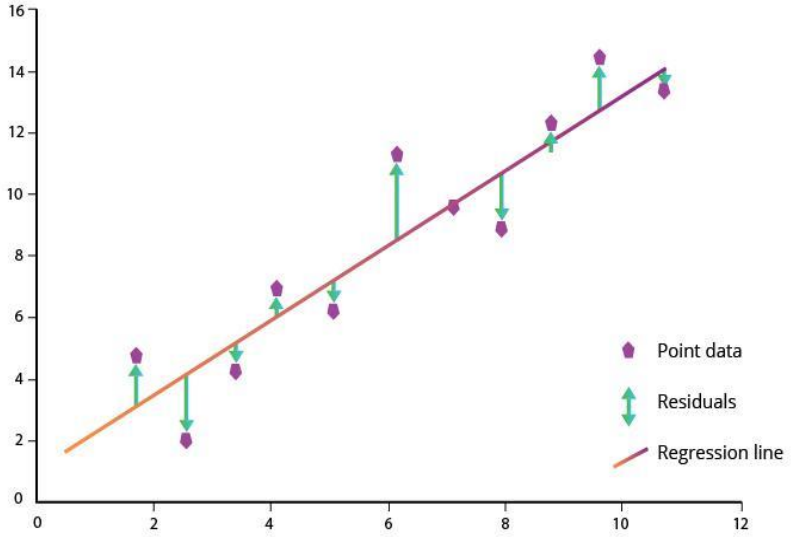
- Regression analysis is a statistical method used to study the relationship between a dependent variable and one or more independent variables.
- R-squared ($R^2$) is a statistical measure, one of the most commonly used methods for linear regression analysis, to determine the proportion of variance in a dependent variable that can be predicted or explained by an independent variable in the model. Alternatively, R-squared is a statistical measure of how close the data are to the fitted regression line. An R-squared value shows how well the model predicts the outcome of the dependent variable.
- In linear regression models, $R^2$ shows how well a regression model (independent variable) predicts the outcome of observed data (dependent variable).
- R-squared is also known as a goodness-fit-measure. It takes into account the strength of the relationship between the model and the dependent variable. Additionally, $R^2$ is commonly known as the coefficient of determination. It is a goodness-of-fit model for linear regression analysis. In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.
- The formula below is mostly used to find the value of R-squared:

  $$R^2 = 1 - \frac{RSS}{TSS}$$

  where,
  $R^2$ = Coefficient of Determination
  RSS = Residuals Sum of Squares (RSS)
  TSS = Total Sum of Squares (TSS)
- In general, R-squared = Explained variation / Total variation. The higher the R-squared, the better the model fits your data.
- R-squared values range from 0 to 1 (R-squared is always between 0 and 100%)
- An R-squared value of 0 means that the model explains or predicts 0% of the relationship between the dependent and independent variables (0% indicates that the model explains none of the variability of the response data around its mean)
- 100% indicates that the model explains all the variability of the response data around its mean. A value of 1 indicates that the model predicts 100% of the relationship, and a value of 0.5 indicates that the model predicts 50%, and so on.
- Before we look at the statistical measures for goodness-of-fit, it is recommended to check the residual plots. Residual plots can reveal unwanted residual patterns that indicate biased results more effectively than numbers. When your residual plots pass muster, you can trust your numerical results and check the goodness-of-fit statistics.
- Plotting fitted values by observed values graphically illustrates different R-squared values for regression models. The regression model on the left accounts for 17.0% of the variance while the one on the right accounts for 83.0%. The more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line. Theoretically, if a model could explain 100% of the variance, the fitted values would always equal the observed values and, therefore, all the data points would fall on the fitted regression line.



**R-squared : 17%** **R-squared : 83%**




- Low R-squared values are not always Bad. There are two major reasons why it can be just fine to have low R-squared values.
- First, In some fields, it is entirely expected that R-squared values will be low. For example, any field that attempts to predict human behavior, such as psychology, typically has R-squared values lower than 50%. Humans are simply harder to predict than, say, physical processes.
- Second, if the R-squared value is low but the coefficients are statistically significant, one can still draw important conclusions about how changes in the predictor values are associated with changes in the response value. Regardless of the R-squared, the significant coefficients still represent the mean change in the response for one unit of change in the predictor while holding other predictors in the model constant. This type of information can be extremely valuable.
- Higher R-squared values suggest a better fit, but it doesn't necessarily mean the model is a good predictor in an absolute sense. The addition of a new variable will lead to a decline in RSS, resulting in higher $R^2$.

- R-squared doesn't tell the entire story. To get the full picture, you must consider $R^2$ values in combination with residual plots, other statistics, and in-depth knowledge of the subject area.
- R-squared only works as intended in a simple linear regression model with one explanatory variable. With a multiple regression made up of several independent variables, the R-squared must be adjusted.

## An Overview of Adjusted R-squared

- Adjusted R-squared is a statistical measure used to evaluate the goodness of fit of a regression model. It provides insights into how well the model explains the variability in the data. Adjusted R-squared addresses a limitation of R-squared, especially in multiple regression (models with more than one independent variable).
- Unlike the standard R-squared, which simply tells you the proportion of variance explained by the model, Adjusted R-squared takes into account the number of predictors (independent variables) in the model.
- The advantage of Adjusted R-squared is that it penalizes the inclusion of unnecessary variables. This means that as you add more predictors to the model, the Adjusted R-squared value will only increase if the new variables significantly improve the model's performance.
- Alternatively, while R-squared tends to increase as more variables (K in our case) are added to the model (even if they don't improve the model significantly), Adjusted R-squared penalizes the addition of unnecessary variables.
- It considers the number of predictors in the model and adjusts R-squared accordingly. This adjustment helps to avoid overfitting, providing a more accurate measure of the model's goodness of fit, given a sample size of N.

$$Adj.R^2 = 1 - \frac{(1-R^2)(N-1)}{(N-K-1)}$$

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.

- As K increases, the value of the fraction will increase, resulting in lower $Adj.R^2$.
- So, if R-squared does not increase significantly on the addition of a new independent variable, then the value of Adjusted R-squared will decrease.

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

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

- Adding a random independent variable does not help in explaining the variation in the target variable always, leading to a false indication that this variable might help predict the output. However, the Adjusted R-squared value decreased which indicated that this new variable is 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.
- In summary, a higher Adjusted R-squared value indicates that more of the variation in the dependent variable is explained by the model, while also considering the model's simplicity. It's a valuable tool for model selection, helping you strike a balance between explanatory power and complexity.