**Regression and Correlation**

**Regression** describes how an independent variable is numerically related to the dependent variable. **Correlation** is used to represent the linear **relationship between** two variables. ... In **correlation**, there is no **difference between** dependent and independent variables i.e. **correlation between** x and y is similar to y and x.

In simple **linear regression** a single independent variable is **used** to predict the value of a dependent variable. In multiple **linear regression** two or more independent variables are **used** to predict the value of a dependent variable. The difference between the two is the number of independent variables.

**Correlation** is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect.

In a positive **correlation**, both **variables** move in the same direction. In other words, as one **variable** increases, so **does** the other. ... When **two variables** have a **negative correlation**, they have an inverse **relationship**. This **means that** as one **variable** increases, the other decreases, and vice versa

(The correlation coefficient has a value between +1 and −1. A correlation coefficient of +1 indicates that the variables move in perfect tandem and in the same direction. A correlation coefficient of 0 indicates that there is no relationship between the variables. A correlation coefficient of −1 indicates that the variables move in perfect tandem but in the OPPOSITE direction.)

If we wish to label the strength of the association, for absolute values of r, 0-0.19 is regarded as very weak, 0.2-0.39 as weak, 0.40-0.59 as moderate, **0.6**-0.79 as **strong** and 0.8-1 as very **strong correlation**, but these are rather arbitrary limits, and the context of the results should be considered.

When you perform a hypothesis test in statistics, a **p**-**value** helps you determine the significance of your results. ... The **p**-**value** is a number between 0 and 1 and interpreted in the following way: A small **p**-**value** (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.

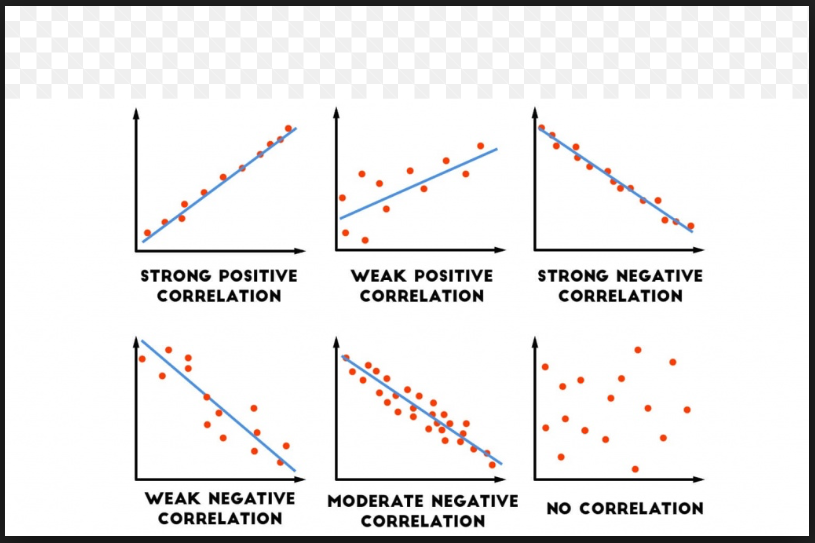
A **null hypothesis** is a type of **hypothesis** used in statistics that proposes that no statistical significance exists in a set of given observations. The **null hypothesis** attempts to show that no variation exists between variables or that a single variable is no different than its **mean**

Both R2 and the adjusted R2 give you an idea of how many data points fall within the line of the [regression equation](https://www.statisticshowto.datasciencecentral.com/what-is-a-regression-equation/). However, there is one main difference between R2 and the adjusted R2: R2 assumes that every single variable explains the variation in the [dependent variable](https://www.statisticshowto.datasciencecentral.com/dependent-variable-definition/). The adjusted R2 tells you the percentage of variation explained by only the [independent variables](https://www.statisticshowto.datasciencecentral.com/independent-variable-definition/) that actually affect the dependent variable.

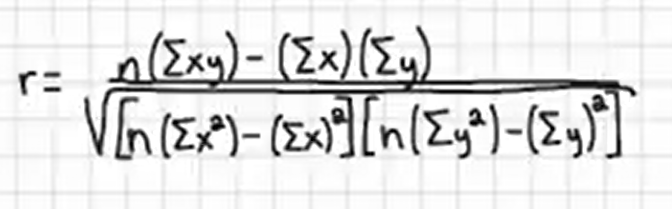
Moving Average

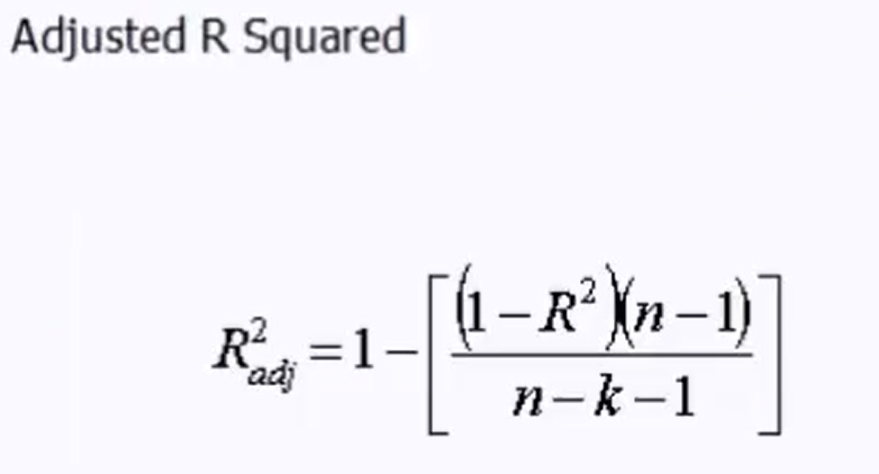
Trend lines

Exponential smoothening



**Correlation coefficient**:





**The Multiple R:** The multiple R is the absolute value of the correlation coefficient of the two variables (X and Y) being evaluated. The correlation coefficient indicates how closely two variables move in tandem with each other. It assumes that the relationship is linear and so measures the linear relationship between the two variables X and Y.

However, since the multiple R is the absolute value of the correlation coefficient, we do not get to know if the correlation is positive or negative! This means that we do not see the direction of the relationship and only know the strength of the relationship.

The correlation coefficient is also referred to as the **Pearson correlation coefficient** or the **Pearson’s r.**

**R-Squared or Multiple R-Squared:** The R-Squared (in Microsoft Excel) or Multiple R-Squared (in R) indicates how well the model or regression line “fits” the data. It indicates the proportion of variance in the dependent variable (Y) that is explained by the independent variable (X).

We know a variable could be impacted by one or more factors. The R-Squared indicates the percentage of variation in the dependent variable that is explained by the independent variables.

**Adjusted R-Squared:** Adjusted R-Squared is used only when analyzing multiple regression output and ignored when analyzing simple linear regression output. When we have more than one independent variable in our analysis, the computation process inflates the R-squared. As the name indicates, the Adjusted R-Squared is the R-Square adjusted for this inflation when performing multiple regression.

The interpretation of the Adjusted R-Squared is similar to the R-square and used only when analyzing multiple regression output.

**The Standard Error:** The standard error in the regression output is a very important number to understand when interpreting regression data. The standard error is a measure of the precision of the model. It reflects the average error of the regression model. In other words, if we were using the regression model to predict or estimate the dependent variable or variable of interest, the standard error shows you how wrong you could be if you used the regression model to make predictions. As the standard error reflects how wrong you could be, we want the standard error to be as small as possible.

The standard error is used to help you get a confidence interval for your predicted values.

**Significance F:** The simplest way to understand the significance F is to think of it as the probability that our regression model is wrong and needs to be discarded!! The significance F gives you the probability that the model is wrong. We want the significance F or the probability of being wrong to be as small as possible.

Significance F: Smaller is better

We can see that the Significance F is very small in given example. We usually establish a significance level and use it as the cut-off point in evaluating the model. Commonly used significance levels are 1%, 5% or 10%.

Statistically speaking, the significance F is the probability that the null hypothesis in our regression model cannot be rejected. In other words, it indicates the probability that all the coefficients in our regression output are actually zero! The significance F is computed from the F value (found to the left of the significance F in Microsoft Excel’s output). The F value is a value similar to the z value, t value, etc. It is a ratio computed by dividing the mean regression sum of squares by the mean error sum of squares. The F value ranges from zero to a very large number.

Note that the significance F is similar in interpretation to the P value discussed. **The key difference is that the significance F applies to the entire model as a whole whereas the P value will be applied only to each corresponding coefficient.**

**What does the intercept indicate?**

**What do the coefficients indicate?**

**What do the signs of coefficients indicate?**

**P-values:** The P value indicates the probability that the estimated coefficient is wrong or unreliable. The best way to understand the P value is as the “probability of an error”. We want the P value to be as small as possible.

How small the P value should be depending on a cut off level that we decide on separately (also called the significance level). The cut-off selected depends on the nature of the data studied and the different error types. The cut-off or significance level is usually 1%, 5% or 10%. Generally, a cut-off point of 5% is used.

Statistically speaking, the P value is the probability of obtaining a result as or more extreme than the one you got in a random distribution. In other words, the P value is the probability that the coefficient of the independent variable in our regression model is not reliable or that the coefficient in our regression output is actually zero! The P value is computed from the t statistic using the Student’s t distribution table.

**The 95% Confidence Interval:** The coefficient of the independent variable is an estimate of the impact this variable has on the variable being studied. This is estimated from a sample that was analyzed in our regression analysis. The 95% confidence interval of your coefficient gives you the range within which the real value of the coefficient you are estimating falls in. The 95% Confidence Interval is also shown as Lower 95% & Upper 95% in many packages.

You can be 95% confident that the real, underlying value of the coefficient that you are estimating falls somewhere in that 95% confidence interval. So, if the interval does not contain 0, your P value will be .05 or less.