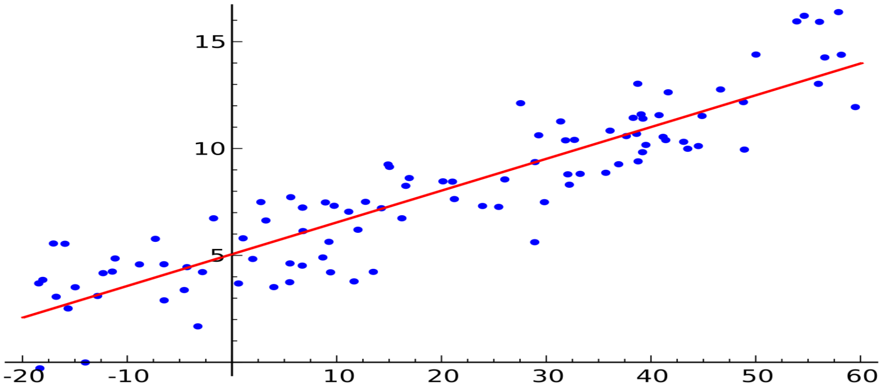
Regression analysis

In statistical modeling, **regression analysis** is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. For example, the method of ordinary least squares computes the unique line (or hyperplane) that minimizes the sum of squared differences between the true data and that line (or hyperplane). For specific mathematical reasons (see linear regression), this allows the researcher to estimate the conditional expectation (or population average value) of the dependent variable when the independent variables take on a given set of values. Less common forms of regression use slightly different procedures to estimate alternative location parameters (e.g., quantile regression or Necessary Condition Analysis[1]) or estimate the conditional expectation across a broader collection of non-linear models (e.g., nonparametric regression).



Regression analysis is primarily used for two conceptually distinct purposes. First, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Second, in some situations regression analysis can be used to infer causal relationships between the independent and dependent variables. Importantly, regressions by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed dataset. To use regressions for prediction or to infer causal relationships, respectively, a researcher must carefully justify why existing relationships have predictive power for a new context or why a relationship between two variables has a causal interpretation. The latter is especially important when researchers hope to estimate causal relationships using observational data.

Linear regression:

In statistics, **linear regression** is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called **multiple linear regression**.[1] This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.[2]

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.[3] Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.[4] This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, forecasting, or error reduction,[*clarification needed*] linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
- If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L^2-norm penalty) and lasso (L^1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked,

Simple linear regression

In statistics, simple linear regression is a linear regression model with a single explanatory variable.That is, it concerns two-dimensional sample points with one independent variable and one dependent variable (conventionally, the x and y coordinates in a Cartesian coordinate system) and finds a linear function (a non-vertical straight line) that, as accurately as possible, predicts the dependent variable values as a function of the independent variable. The adjective simple refers to the fact that the outcome variable is related to a single predictor.

It is common to make the additional stipulation that the ordinary least squares (OLS) method should be used: the accuracy of each predicted value is measured by its squared residual (vertical distance between the point of the data set and the fitted line), and the goal is to make the sum of these squared deviations as small as possible. Other regression methods that can be used in place of ordinary least squares include least absolute deviations (minimizing the sum of absolute values of residuals) and the Theil–Sen estimator (which chooses a line whose slope is the median of the slopes determined by pairs of sample points). Deming regression (total least squares) also finds a line that fits a set of two-dimensional sample points, but (unlike ordinary least squares, least absolute deviations, and median slope regression) it is not really an instance of simple linear regression, because it does not separate the coordinates into one dependent and one independent variable and could potentially return a vertical line as its fit.

Correlation Analysis

Correlation analysis is applied in quantifying the association between two continuous variables, for example, an dependent and independent variable or among two independent variables.

Regression Analysis

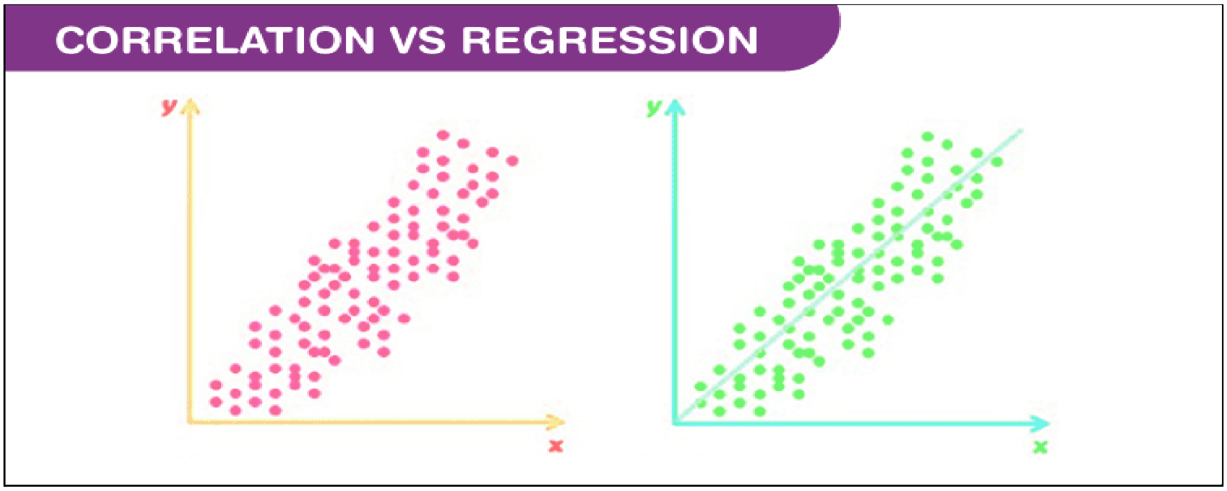
Regression analysis refers to assessing the relationship between the outcome variable and one or more variables. The outcome variable is known as the dependent or response variable and the risk elements, and cofounders are known as predictors or independent variables. The dependent variable is shown by “y” and independent variables are shown by “x” in regression analysis.

The sample of a correlation coefficient is estimated in the correlation analysis. It ranges between -1 and +1, denoted by r and quantifies the strength and direction of the linear association among two variables. The correlation among two variables can either be positive, i.e. a higher level of one variable is related to a higher level of another or negative, i.e. a higher level of one variable is related to a lower level of the other.

The sign of the coefficient of correlation shows the direction of the association. The magnitude of the coefficient shows the strength of the association.

For example, a correlation of r = 0.8 indicates a positive and strong association among two variables, while a correlation of r = -0.3 shows a negative and weak association. A correlation near to zero shows the non-existence of linear association among two continuous variables.

Correlation and Regression Differences



There are some differences between Correlation and regression.

- Correlation shows the quantity of the degree to which two variables are associated. It does not fix a line through the data points. You compute a correlation that shows how much one variable changes when the other remains constant. When r is 0.0, the relationship does not exist. When r is positive, one variable goes high as the other goes up. When r is negative, one variable goes high as the other goes down.
- Linear regression finds the best line that predicts y from x, but Correlation does not fit a line.
- Correlation is used when you measure both variables, while linear regression is mostly applied when x is a variable that is manipulated.

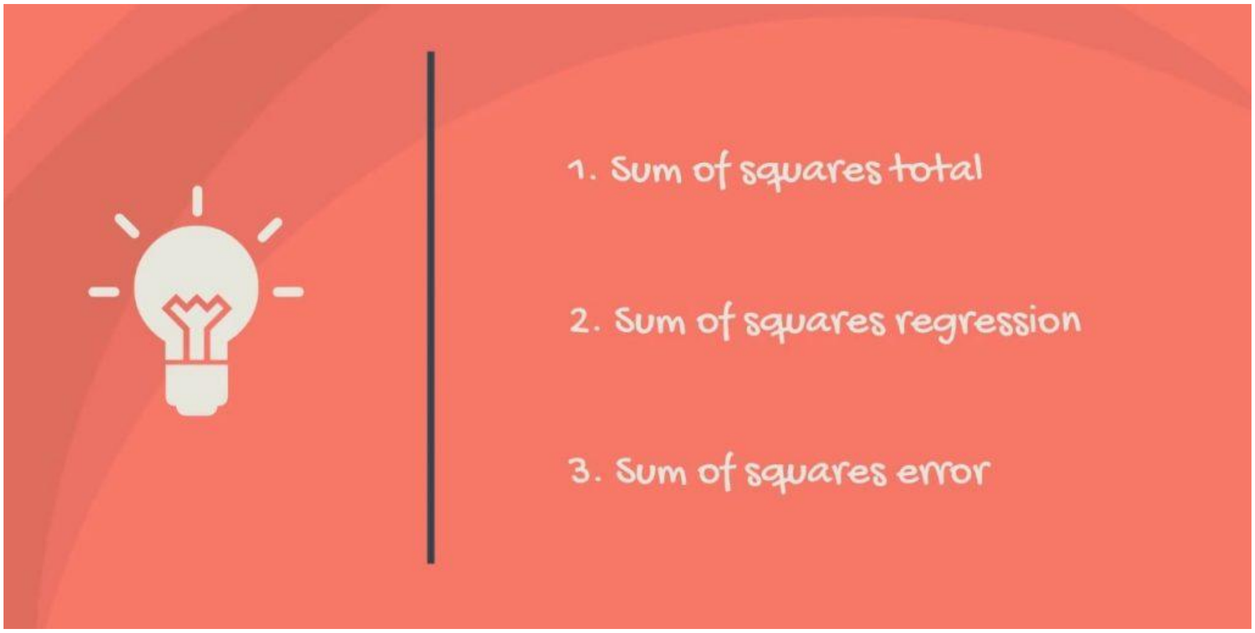
Comparison Between Correlation and Regression

Basis	Correlation	Regression
Meaning	**A statistical measure that defines co-relationship or association of two variables.**	**Describes how an independent variable is associated with the dependent variable.**
Dependent and Independent variables	**No difference**	**Both variables are different.**

Usage	To describe a linear relationship between two variables.	To fit the best line and estimate one variable based on another variable.
Objective	To find a value expressing the relationship between variables.	To estimate values of a random variable based on the values of a fixed variable.

SST, SSR, SSE: Definition and Formulas

There are three terms we must define. The sum of squares total, the sum of squares regression, and the sum of squares error.



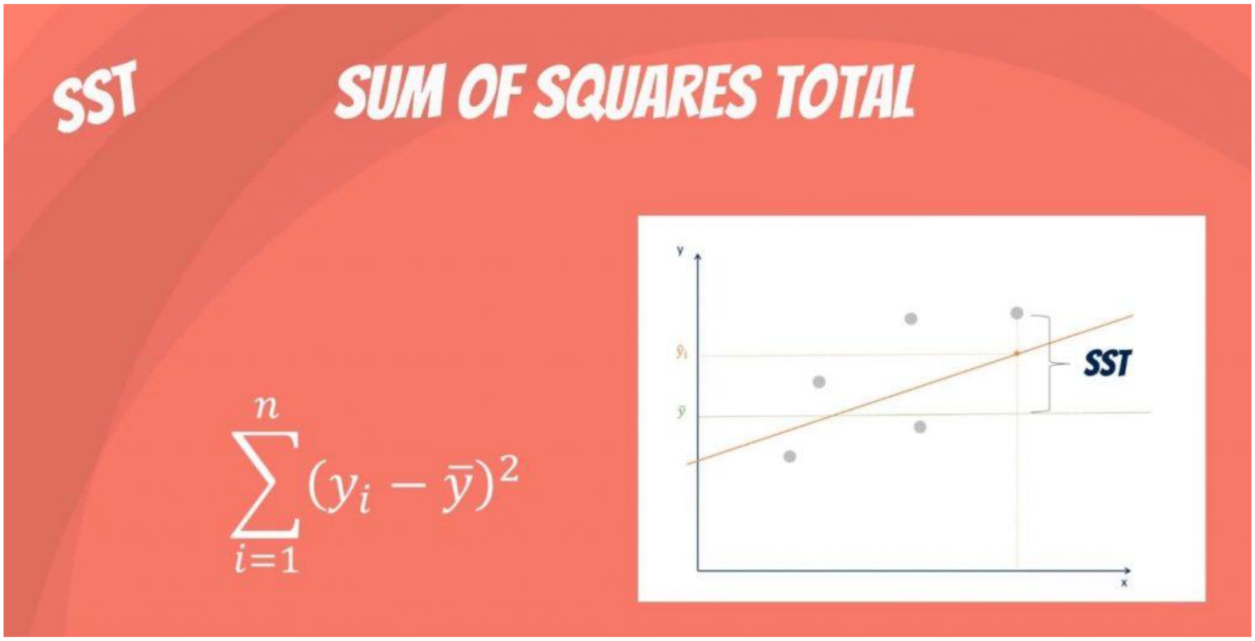
1. Sum of squares total

2. Sum of squares regression

3. Sum of squares error

What is the SST?

The sum of squares total, denoted SST, is the squared differences between the observed dependent variable and its mean. You can think of this as the dispersion of the observed variables around the mean – much like the variance in descriptive statistics.



SST

SUM OF SQUARES TOTAL

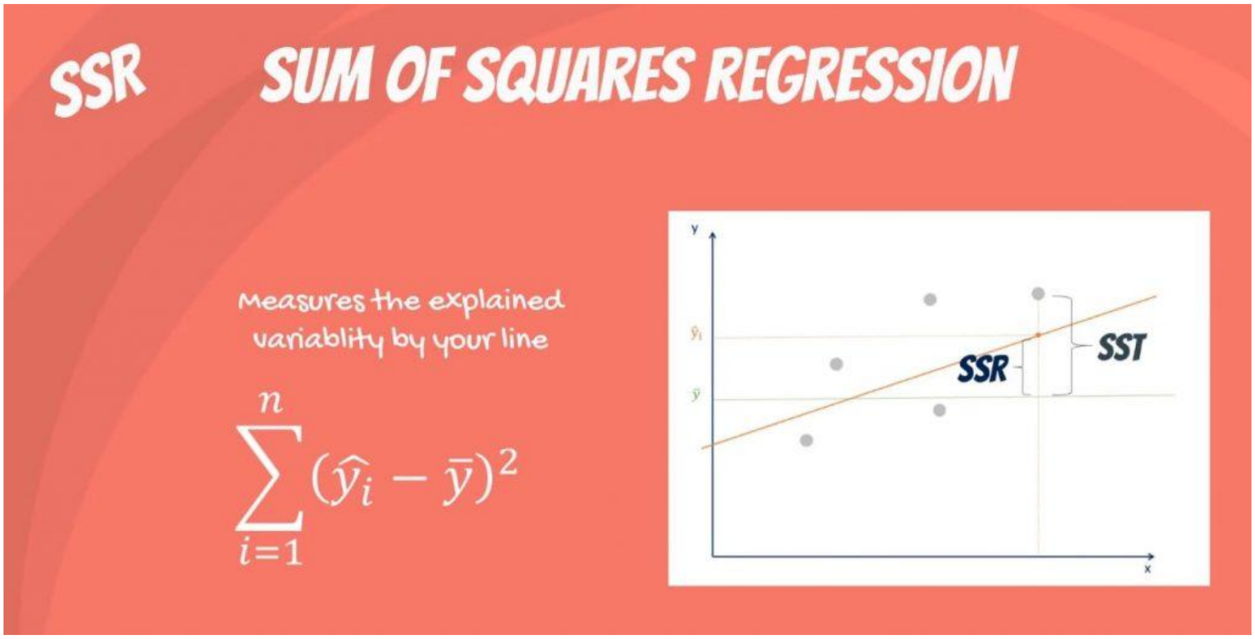
$$\sum_{i=1}^{n}(y_i-\bar{y})^2$$

It is a measure of the total variability of the dataset.

Side note: There is another notation for the SST. It is TSS or total sum of squares.

What is the SSR?

The second term is the sum of squares due to regression, or SSR. It is the sum of the differences between the predicted value and the mean of the dependent variable. Think of it as a measure that describes how well our line fits the data.



SSR

SUM OF SQUARES REGRESSION

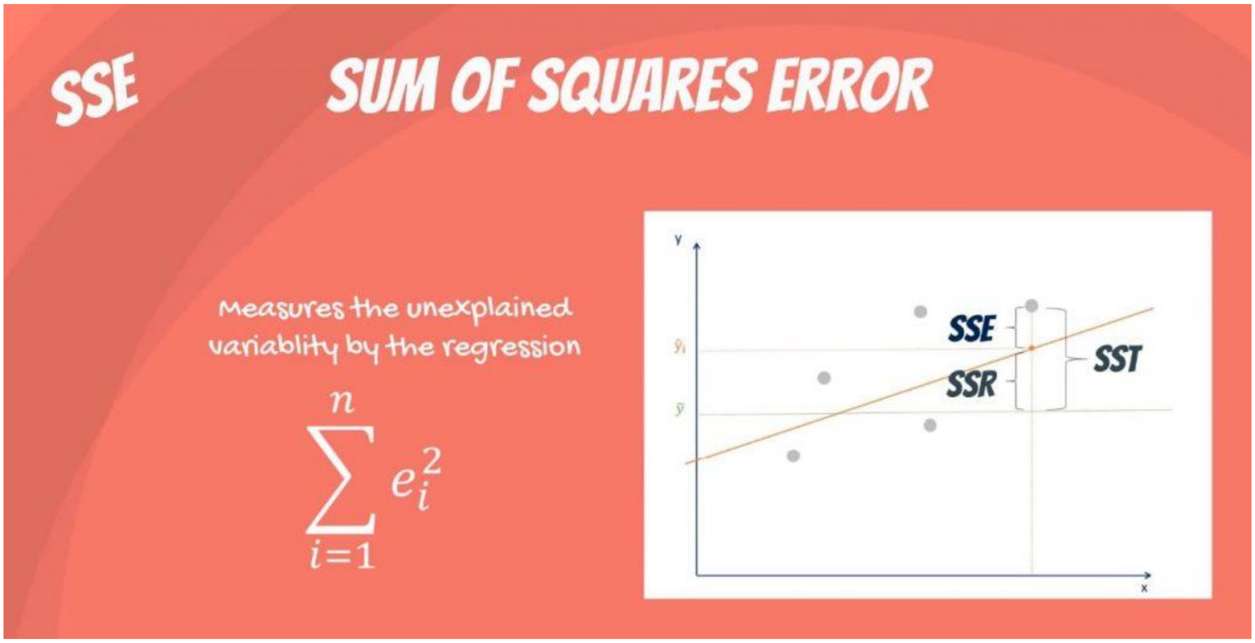
Measures the explained variablity by your line

$$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

If this value of SSR is equal to the sum of squares total, it means our regression model captures all the observed variability and is perfect. Once again, we have to mention that another common notation is ESS or explained sum of squares.

What is the SSE?

The last term is the sum of squares error, or SSE. The error is the difference between the observed value and the predicted value.



SSE

SUM OF SQUARES ERROR

Measures the unexplained variablity by the regression

$$\sum_{i=1}^{n} e_i^2$$

We usually want to minimize the error. The smaller the error, the better the estimation power of the regression. Finally, I should add that it is also known as RSS or residual sum of squares. Residual as in: remaining or unexplained