1. In a linear equation, what is the difference between a dependent variable and an independent variable?

This is specifically about functions. Think of a function as a machine. When you feed something into it, something comes out. What comes out is called the dependent variable because it depends on what you put in. What you put in is called the independent variable, because you can choose any value you like without having to think of its effect on the other variable.

If you are just interested in the relationship between two variables, it might not be clear that either is independent, they both depend on each other.

For example the graph of the equation x2+y2=25𝑥2+𝑦2=25 is a circle. If that’s what you want to represent, neither should be regarded as independent. Actually neither is a function of the other: if you put x=3𝑥=3 then there are two possible values of y𝑦, y=4𝑦=4 or y=−4𝑦=−4, and: if you put y=4𝑦=4 then there are two possible values of x𝑥, x=3𝑥=3 or x=−3𝑥=−3. A function must return a single value.

On the other hand, the graph of the equation y+x=25𝑦+𝑥=25 is a straight line which passes through the points (0,25)(0,25) and (25,0)(25,0). You can think of y𝑦 as a function of x𝑥, or x𝑥 as a function of y𝑦. If you just want to describe the line, it doesn’t matter which you think of as the independent variable. You are only interested in the relationship between the variables.

The concept of independent and dependent is only applicable if you have reason to think of one variable as causing the other, or you are wondering about how one affects the other. If you do have such a reason, then make the relationship into a function. Once you have a function, not just a relation, then dependent and independent are well defined.

1. What is the concept of simple linear regression? Give a specific example.

**Simple linear regression** is used to estimate the relationship between**two**[**quantitative variables**](https://www.scribbr.com/methodology/types-of-variables/#quantitative-vs-categorical). You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g., the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the [independent variable](https://www.scribbr.com/methodology/independent-and-dependent-variables/#independent) (e.g., the amount of soil erosion at a certain level of rainfall).

**Regression models** describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a [dependent variable](https://www.scribbr.com/methodology/independent-and-dependent-variables/#dependent) changes as the independent variable(s) change.

Simple linear regression exampleYou are a social researcher interested in the relationship between income and happiness. You survey 500 people whose incomes range from 15k to 75k and ask them to rank their happiness on a scale from 1 to 10.

Your independent variable (income) and dependent variable (happiness) are both quantitative, so you can do a regression analysis to see if there is a linear relationship between them.

3. In a linear regression, define the slope.

**Slope (of a Linear Regression)**

Returns the slope of a linear regression line.In a regression line passing through a set of data points in data sets Argument1 and Argument2, the slope is the vertical distance divided by the horizontal distance between any two points on the line. This ratio is also known as the rate of change along the line.

**Syntax**

Slope <FactID> (*Argument1*, *Argument2*)

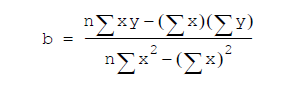
Where:

*Argument1* is a fact or metric containing numerically dependent (y) data points.

*Argument2* is a fact or metric containing numerically independent (x) data points.

FactID is a parameter that forces a calculation to take place on a fact table that contains the selected fact.

**Expression**



**Usage Notes**

The following conditions are invalid:

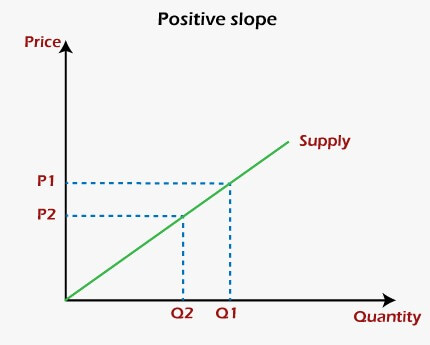
* Either Argument1 or Argument2 contains NULL values.
* Argument1 and Argument2 have a different number of values.

**Example**

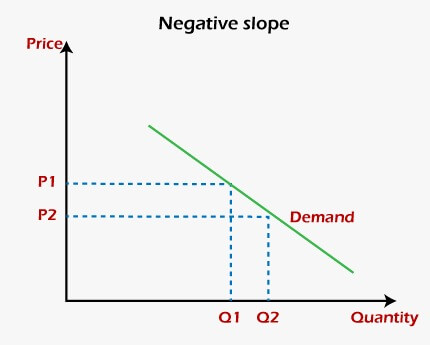
See [Hypothesis Testing example](https://www2.microstrategy.com/producthelp/2021/FunctionsRef/Content/FuncRef/Additional_examples_of_functions_in_expressions.htm) for an example using the Slope function.

4. Determine the graph's slope, where the lower point on the line is represented as (3, 2) and the higher point is represented as (2, 2).

5. In linear regression, what are the conditions for a positive slope?

* **Positive Slope:** A slope in which two variables, i.e., variable at x-axis and variable at the y-axis, shows a **positive relationship** is known as **positive slope.** In simpler words, a positive slope is one in which the variable x increases with the increase in variable y and/or variable y increases with the increase in variable x. Similarly, the variable x decreases with the decrease in variable y, and/or variable y decreases with the decrease in variable x. It means both the variables are **complements** to each other. A positive slope moves in the **upward direction** or is **upward sloping.**  
  In graphical terms, a positive slope is one in which the line on the graph rises when it moves from left to right. The concept of positive slope can be clearly understood with the help of the **supply curve** of a producer or firm in economics. The two variables of the curve are price at the y-axis and quantity of goods at the x-axis. Let us assume the firm is producing the goods for **profit maximization.** Therefore, when the prices of the goods increase, the quantity supplied by the firm of those goods will also increase, while when the prices decrease, the quantity supplied by the firm will decrease. In other words, at higher prices, the firm or producer will increase the quantity supplied to earn more profit, while at lower prices, they will reduce the quantity supplied to reduce the loss. Hence, it shows the prices and quantity supplied are positively related to each other, which can be cleared from the diagram given below:  
  

6. In linear regression, what are the conditions for a negative slope?

* **Negative Slope:** A slope in which two variables, i.e., variable at x-axis and variable at the y-axis, shows a **negative relationship** is known as **negative slope.** In other words, a negative slope is one in which the variable x increases with the decrease in variable y and/or variable y increases with the decrease in variable x. In the same manner, the variable x decreases with the increase in variable y, and/or variable y decreases with the increase in variable x. This represents an **inverse relationship** between these two variables. A negative slope moves in the **downward direction** or is **downward sloping.**  
  Graphically, a negative slope is one in which the line on the graph falls when it moves from left to right. One of the best examples of the negative slope of the graph is the **demand curve** in economics. The two variables of the curve are price at the y-axis and quantity of goods at the x-axis. As we know, the consumers buy a large quantity of a good at a lower price than at a higher price. Therefore, the quantity demanded by the consumers of goods will decrease with an increase in the prices of those goods. On the other hand, when prices of the goods will decrease, the quantity demand will increase. Hence, it shows a **negative relationship** between the prices and quantity supplied of those goods. It can be cleared from the diagram given below:  
  

7. What is multiple linear regression and how does it work?

## What is Multiple Linear Regression?

Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the [dependent variable](https://corporatefinanceinstitute.com/resources/?topics=111064) are known as independent or explanatory variables.

### Multiple Linear Regression Formula

Where:

* **yi​** is the dependent or predicted variable
* **β0** is the y-intercept, i.e., the value of y when both xi and x2 are 0.
* **β1** and **β2** are the regression coefficients representing the change in y relative to a one-unit change in **xi1** and **xi2**, respectively.
* **βp** is the slope coefficient for each independent variable
* **ϵ** is the model’s random error (residual) term.

### Understanding Multiple Linear Regression

Simple linear regression enables statisticians to predict the value of one variable using the available information about another variable. Linear regression attempts to establish the relationship between the two variables along a straight line.

Multiple regression is a type of regression where the dependent variable shows a **linear** relationship with two or more independent variables. It can also be **non-linear**, where the dependent and [independent variables](https://corporatefinanceinstitute.com/resources/financial-modeling/independent-variable/) do not follow a straight line.

Both linear and non-linear regression track a particular response using two or more variables graphically. However, non-linear regression is usually difficult to execute since it is created from assumptions derived from trial and error.

### Assumptions of Multiple Linear Regression

Multiple linear regression is based on the following assumptions:

#### 1. A linear relationship between the dependent and independent variables

The first assumption of multiple linear regression is that there is a linear relationship between the dependent variable and each of the independent variables. The best way to check the linear relationships is to create scatterplots and then visually inspect the scatterplots for linearity. If the relationship displayed in the scatterplot is not linear, then the analyst will need to run a non-linear regression or transform the data using statistical software, such as SPSS.

#### 2. The independent variables are not highly correlated with each other

The data should not show multicollinearity, which occurs when the independent variables (explanatory variables) are highly correlated. When independent variables show multicollinearity, there will be problems figuring out the specific variable that contributes to the variance in the dependent variable. The best method to test for the assumption is the Variance Inflation Factor method.

#### 3. The variance of the residuals is constant

Multiple linear regression assumes that the amount of error in the residuals is similar at each point of the linear model. This scenario is known as homoscedasticity. When analyzing the data, the analyst should plot the standardized residuals against the predicted values to determine if the points are distributed fairly across all the values of independent variables. To test the assumption, the data can be plotted on a scatterplot or by using statistical software to produce a scatterplot that includes the entire model.

#### 4. Independence of observation

The model assumes that the observations should be independent of one another. Simply put, the model assumes that the values of residuals are independent. To test for this assumption, we use the Durbin Watson statistic.

The test will show values from 0 to 4, where a value of 0 to 2 shows positive autocorrelation, and values from 2 to 4 show negative autocorrelation. The mid-point, i.e., a value of 2, shows that there is no autocorrelation.

#### 5. Multivariate normality

Multivariate normality occurs when residuals are normally distributed. To test this assumption, look at how the values of residuals are distributed. It can also be tested using two main methods, i.e., a histogram with a superimposed normal curve or the Normal Probability Plot method.

### More Resources

Thank you for reading CFI’s guide to Multiple Linear Regression. To keep learning and developing your knowledge base, please explore the additional relevant CFI resources below:

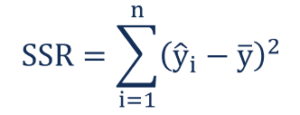
* [High Low Method vs. Regression Analysis](https://corporatefinanceinstitute.com/resources/data-science/high-low-method-vs-regression-analysis/)
* [Forecasting Methods](https://corporatefinanceinstitute.com/resources/financial-modeling/forecasting-methods/)
* [Poisson Distribution](https://corporatefinanceinstitute.com/resources/data-science/poisson-distribution/)
* [Random Variable](https://corporatefinanceinstitute.com/resources/data-science/random-variable/)
* [Regression Analysis](https://corporatefinanceinstitute.com/resources/data-science/regression-analysis/)
* [**See all data science resources**](https://corporatefinanceinstitute.com/topic/data-science/)

8. In multiple linear regression, define the number of squares due to error.

#### Regression sum of squares (also known as the sum of squares due to regression or explained sum of squares)

The regression sum of squares describes how well a regression model represents the modeled data. A higher regression sum of squares indicates that the model does not fit the data well.

The formula for calculating the regression sum of squares is:



Where:

* ŷi– the value estimated by the regression line
* ȳ – the mean value of a sample

9. In multiple linear regression, define the number of squares due to regression.

The sum of squares is a statistical measure of variability. It indicates the dispersion of data points around the mean and how much the dependent variable deviates from the predicted values in regression analysis.

We decompose variability into the **sum of squares total** (SST), the **sum of squares regression** (SSR), and the **sum of squares error** (SSE). The decomposition of variability helps us understand the sources of variation in our data, assess a model’s goodness of fit, and understand the relationship between variables.

## SST, SSR, SSE: Definition and Formulas

This article addresses SST, SSR, and SSE in the context of the [ANOVA framework](https://www.xlstat.com/en/solutions/features/anova-analysis-of-variance), but the sums of squares are frequently used in various statistical analyses.

### ****What Is SST in Statistics?****

The **sum of squares total** **(SST)**or the **total sum of squares (TSS)** is the sum of squared differences between the observed dependent variables and the overall **mean**. Think of it as the dispersion of the observed variables around the [**mean**](https://365datascience.com/tutorials/statistics-tutorials/measures-central-tendency/)—similar to the [variance](https://365datascience.com/tutorials/statistics-tutorials/coefficient-variation-variance-standard-deviation/) in descriptive statistics. But SST measures the total variability of a dataset, commonly used in regression analysis and ANOVA

Mathematically, the difference between variance and SST is that we adjust for the degree of freedom by dividing by n–1 in the variance formula.

SST=n∑i=1(yi−¯y)2𝑆𝑆𝑇=∑𝑖=1𝑛(𝑦𝑖−𝑦¯)2

Where:

yi 𝑦𝑖  – observed dependent variable

¯y 𝑦¯  – mean of the dependent variable

### ****What Is SSR in Statistics?****

The **sum of squares due to regression** **(SSR)** or **explained sum of squares** **(ESS)**is the sum of the differences between the predicted value and the **mean** of the dependent variable. In other words, it describes how well our line fits the data.

The SSR formula is the following:

SSR=n∑i=1(^yi−¯y)2𝑆𝑆𝑅=∑𝑖=1𝑛(𝑦^𝑖−𝑦¯)2

Where:

^yi 𝑦^𝑖  – the predicted value of the dependent variable

¯y 𝑦¯  – mean of the dependent variable

If **SSR** equals **SST**, our **regression** **model** perfectly captures all the observed variability, but that’s rarely the case.

10. In a regression equation, what is multicollinearity?

Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

In general, multicollinearity can lead to wider confidence intervals that produce less reliable probabilities in terms of the effect of independent variables in a model.

In technical analysis, multicollinearity can lead to incorrect assumptions about an investment. It generally occurs because multiple indicators of the same type have been used to analyze a stock.

11. What is heteroskedasticity, and what does it mean?

In statistics, heteroskedasticity (or heteroscedasticity) happens when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant. With heteroskedasticity, the tell-tale sign upon visual inspection of the residual errors is that they will tend to fan out over time, as depicted in the image below.

Heteroskedasticity often arises in two forms: conditional and unconditional. Conditional heteroskedasticity identifies nonconstant [volatility](https://www.investopedia.com/terms/v/volatility.asp) related to prior period's (e.g., daily) volatility. Unconditional heteroskedasticity refers to general structural changes in volatility that are not related to prior period volatility. Unconditional heteroskedasticity is used when future periods of high and low volatility can be identified.

12. Describe the concept of ridge regression.

Ridge regression is a statistical regularization technique. It corrects for overfitting on training data in machine learning models.

Ridge regression—also known as L2 regularization—is one of several types of regularization for [linear regression](https://www.ibm.com/topics/linear-regression) models. [Regularization](https://www.ibm.com/topics/regularization) is a statistical method to reduce errors caused by overfitting on training data. Ridge regression specifically corrects for [multicollinearity](https://www.ibm.com/topics/multicollinearity) in regression analysis. This is useful when developing machine learning models that have a large number of parameters, particularly if those parameters also have high weights. While this article focuses on regularization of linear regression models, note that ridge regression may also be applied in [logistic regression](https://www.ibm.com/topics/logistic-regression#:~:text=Resources-,What%20is%20logistic%20regression%3F,given%20dataset%20of%20independent%20variables.).

13. Describe the concept of lasso regression.

Lasso regression is a regularization technique that applies a penalty to prevent [overfitting](https://www.ibm.com/topics/overfitting) and enhance the accuracy of statistical models.

Lasso regression—also known as L1 regularization—is a form of regularization for [linear regression](https://www.ibm.com/topics/linear-regression) models. [Regularization](https://www.ibm.com/topics/regularization) is a statistical method to reduce errors caused by overfitting on training data. This approach can be reflected with this formula:

w-hat = argminw MSE(W ) + ||w||1

The concepts behind the Lasso technique can be traced to a 1986 [geophysics research paper](https://epubs.siam.org/doi/10.1137/0907087) (link resides outside ibm.com) by Santosa and Symes1, which used the L1 penalty for coefficients. However, in 1996, statistician, Robert Tibshirani, [independently developed and popularized the term](https://webdoc.agsci.colostate.edu/koontz/arec-econ535/papers/Tibshirani%20(JRSS-B%201996).pdf)2(link resides outside ibm.com), "lasso", based on [Breiman's nonnegative garrote work](https://www.tandfonline.com/doi/abs/10.1080/00401706.1995.10484371" \t "_blank)3(link resides outside ibm.com).

Lasso stands for Least Absolute Shrinkage and Selection Operator. It is frequently used in machine learning to handle high dimensional data as it facilitates automatic feature selection with its application. It does this by adding a penalty term to the residual sum of squares (RSS), which is then multiplied by the regularization parameter (lambda or λ). This regularization parameter controls the amount of regularization applied. Larger values of lambda increase the penalty, shrinking more of the coefficients towards zero; this subsequently reduces the importance of (or altogether eliminates) some of the features from the model, resulting in automatic feature selection. Conversely, smaller values of lambda reduce the effect of the penalty, retaining more features within the model.

This penalty promotes sparsity within the model, which can help avoid issues of [multicollinearity](https://www.ibm.com/topics/multicollinearity) and [overfitting](https://www.ibm.com/topics/overfitting) issues within datasets. Multicollinearity occurs when two or more independent variables are highly correlated with one another, which can be problematic for causal modeling. Overfit models will generalize poorly to new data, diminishing their value altogether. By reducing regression coefficients to zero, lasso regression can effectively eliminate independent variables from the model, sidestepping these potential issues within modeling process. Model sparsity can also improve the interpretability of the model compared to other regularization techniques such as [ridge regression](https://www.ibm.com/topics/ridge-regression) (also known as L2 regularization).

As a note, this article focuses on regularization of linear regression models, but it’s worth noting that lasso regression may also be applied in [logistic regression](https://www.ibm.com/topics/logistic-regression#:~:text=Resources-,What%20is%20logistic%20regression%3F,given%20dataset%20of%20independent%20variables.).

14. What is polynomial regression and how does it work?

Polynomial regression is a kind of [linear regression](https://www.simplilearn.com/tutorials/machine-learning-tutorial/linear-regression-in-python) in which the relationship shared between the dependent and independent variables Y and X is modeled as the nth degree of the polynomial. This is done to look for the best way of drawing a line using data points. Keep reading to know more about polynomial regression.

**What Is Polynomial Regression?**

The [algorithm](https://www.simplilearn.com/tutorials/data-structure-tutorial/what-is-an-algorithm) of linear regression works only when the regression in the data is linear. Polynomial regression can be considered one of the exceptional cases of multiple linear regression models. In other words, it is a linear regression type containing dependent and independent variables, and they both share a curvilinear relationship. A polynomial relationship is fitted in the data.

Also, several linear regression equations are converted into polynomial regression equations by including numerous polynomial elements.

In [statistics](https://en.wikipedia.org/wiki/Statistics), **polynomial regression** is a form of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) in which the relationship between the [independent variable](https://en.wikipedia.org/wiki/Independent_variable) *x* and the [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) *y* is modeled as an *n*th degree [polynomial](https://en.wikipedia.org/wiki/Polynomial) in *x*. Polynomial regression fits a nonlinear relationship between the value of *x* and the corresponding [conditional mean](https://en.wikipedia.org/wiki/Conditional_expectation) of *y*, denoted E(*y* |*x*). Although *polynomial regression* fits a nonlinear model to the data, as a [statistical estimation](https://en.wikipedia.org/wiki/Estimation_theory) problem it is linear, in the sense that the regression function E(*y* | *x*) is linear in the unknown [parameters](https://en.wikipedia.org/wiki/Parameter) that are estimated from the [data](https://en.wikipedia.org/wiki/Data). For this reason, polynomial regression is considered to be a special case of [multiple linear regression](https://en.wikipedia.org/wiki/Multiple_linear_regression).

The explanatory (independent) variables resulting from the polynomial expansion of the "baseline" variables are known as higher-degree terms. Such variables are also used in [classification](https://en.wikipedia.org/wiki/Statistical_classification) settings.

15. Describe the basis function.

In [mathematics](https://en.wikipedia.org/wiki/Mathematics), a **basis function** is an element of a particular [basis](https://en.wikipedia.org/wiki/Basis_(linear_algebra)) for a [function space](https://en.wikipedia.org/wiki/Function_space). Every [function](https://en.wikipedia.org/wiki/Function_(mathematics)) in the function space can be represented as a [linear combination](https://en.wikipedia.org/wiki/Linear_combination) of basis functions, just as every vector in a [vector space](https://en.wikipedia.org/wiki/Vector_space) can be represented as a linear combination of [basis vectors](https://en.wikipedia.org/wiki/Basis_vectors).

16. Describe how logistic regression works.

Logistic regression is a data analysis technique that uses mathematics to find the relationships between two data factors. It then uses this relationship to predict the value of one of those factors based on the other. The prediction usually has a finite number of outcomes, like yes or no.

For example, let’s say you want to guess if your website visitor will click the checkout button in their shopping cart or not. Logistic regression analysis looks at past visitor behavior, such as time spent on the website and the number of items in the cart. It determines that, in the past, if visitors spent more than five minutes on the site and added more than three items to the cart, they clicked the checkout button. Using this information, the logistic regression function can then predict the behavior of a new website visitor.