

**What is Logistic Regression? A Beginner's Guide**

BY ANAMIKA THANDA, UPDATED ON DECEMBER 19, 202215 mins read

If you’re new to the field of [**data analytics**](https://careerfoundry.com/en/blog/data-analytics/what-is-data-analytics/), you’re probably trying to get to grips with all the various techniques and tools of the trade. One particular type of analysis that [**data analysts**](https://careerfoundry.com/en/blog/data-analytics/how-to-become-a-data-analyst/) use is logistic regression—but what exactly is it, and what is it used for?

This guide will help you to understand **what logistic regression is**, together with some of the key concepts related to regression analysis in general. By the end of this post, you will have a clear idea of what logistic regression entails, and you’ll be familiar with the different types of logistic regression. We’ll also provide examples of when this type of analysis is used, and finally, go over some of the pros and cons of logistic regression.

We’ve divided our guide as follows:

1. [**An introduction to regression analysis**](https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/#what-is-regression-analysis)
2. [**What is logistic regression?**](https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/#what-is-logistic-regression)
3. [**Assumptions of logistic regression**](https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/#logistic-regression-assumptions)
4. [**What is logistic regression used for?**](https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/#what-is-logistic-regression-used-for)
5. [**What are the different types of logistic regression?**](https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/#what-are-the-different-types-of-logistic-regression)
6. [**What are the advantages and disadvantages of using logistic regression?**](https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/#what-are-the-advantages-and-disadvantages-of-using-logistic-regression)
7. [**Key takeaways and next steps**](https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/#final-thoughts)

**1. What is regression analysis?**

Logistic regression is a type of regression analysis. So, before we delve into logistic regression, let us first introduce the general concept of regression analysis.

Regression analysis is a type of predictive modelling technique which is used to find the relationship between a dependent variable (usually known as the “Y” variable) and either one independent variable (the “X” variable) or a series of independent variables. When two or more independent variables are used to predict or explain the outcome of the dependent variable, this is known as multiple regression.

Regression analysis can be used for three things:

1. **Forecasting the effects or impact of specific changes**. For example, if a manufacturing company wants to forecast how many units of a particular product, they need to produce in order to meet the current demand.
2. **Forecasting trends and future values.** For example, how much will the stock price of Lufthansa be in 6 months from now?
3. **Determining the strength of different predictors**—or, in other words, assessing how much of an impact the independent variable(s) has on a dependent variable. For example, if a soft drinks company is sponsoring a football match, they might want to determine if the ads being displayed during the match have accounted for any increase in sales.

Regression analysis can be broadly classified into two types: [**Linear regression**](https://careerfoundry.com/en/blog/data-analytics/what-is-linear-regression/) and **logistic regression**.

In statistics, linear regression is usually used for predictive analysis. It essentially determines the extent to which there is a linear relationship between a dependent variable and one or more independent variables. In terms of output, linear regression will give you a trend line plotted amongst a set of data points. You might use linear regression if you wanted to predict the sales of a company based on the cost spent on online advertisements, or if you wanted to see how the change in the GDP might affect the stock price of a company.

The second type of regression analysis is logistic regression, and that’s what we’ll be focusing on in this post. Logistic regression is essentially used to calculate (or predict) the probability of a binary (yes/no) event occurring. We’ll explain what exactly logistic regression is and how it’s used in the next section.

**2. What is logistic regression?**

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables.

Ok, so what does this mean? A **binary outcome** is one where there are only two possible scenarios—either the event happens (1) or it does not happen (0). **Independent variables** are those variables or factors which may influence the outcome (or dependent variable).

So: Logistic regression is the correct type of analysis to use when you’re working with binary data. You know you’re dealing with binary data when the output or dependent variable is dichotomous or categorical in nature; in other words, if it fits into one of two categories (such as “yes” or “no”, “pass” or “fail”, and so on).

However, the independent variables can fall into any of the following categories:

1. **Continuous**—such as temperature in degrees Celsius or weight in grams. In technical terms, continuous data is categorized as either [**interval data**](https://careerfoundry.com/en/blog/data-analytics/what-is-interval-data/), where the intervals between each value are equally split, or [**ratio data**](https://careerfoundry.com/en/blog/data-analytics/what-is-ratio-data/), where the intervals are equally split **and** there is a true or meaningful “zero”. For example, temperature in degrees Celsius would be classified as interval data; the difference between 10 and 11 degrees C is equal to the difference between 30 and 31 degrees C, but there is no true zero—a temperature of zero degrees does not mean there is “no temperature”. On the other hand, weight in grams would be classified as ratio data; it has the equal intervals **and** a true zero. In other words, if something weighs zero grams, it truly weighs nothing.
2. **Discrete, ordinal**—that is, data which can be placed into some kind of order on a scale. For example, if you are asked to state how happy you are on a scale of 1-5, the points on the scale represent ordinal data. A score of 1 indicates a lower degree of happiness than a score of 5, but there is no way of determining the numerical value between each of the points on the scale. Ordinal data is the kind of data you might get from a customer satisfaction survey.
3. **Discrete, nominal**—that is, data which fits into named groups which do not represent any kind of order or scale. For example, eye colour may fit into the categories “blue”, “brown”, or “green”, but there is no hierarchy to these categories.

So, in order to determine if logistic regression is the correct type of analysis to use, ask yourself the following:

* **Is the dependent variable dichotomous?** In other words, does it fit into one of two set categories? Remember: The dependent variable is the outcome; the thing that you’re measuring or predicting.
* **Are the independent variables either interval, ratio, or ordinal?** See the examples above for a reminder of what these terms mean. Remember: The independent variables are those which may impact, or be used to predict, the outcome.

In addition to the two criteria mentioned above, there are some further requirements that must be met in order to correctly use logistic regression. These requirements are known as “assumptions”; in other words, when conducting logistic regression, you’re assuming that these criteria have been met. Let’s take a look at those now.

**3. Logistic regression assumptions**

* **The dependent variable is binary or dichotomous**—i.e. It fits into one of two clear-cut categories. This applies to binary logistic regression, which is the type of logistic regression we’ve discussed so far. We’ll explore some other types of logistic regression in section five.
* **There should be no, or very little, multicollinearity between the predictor variables**—in other words, the predictor variables (or the independent variables) should be independent of each other. This means that there should **not** be a high correlation between the independent variables. In statistics, certain tests can be used to calculate the correlation between the predictor variables; if you’re interested in learning more about those, just search “Spearman’s rank correlation coefficient” or “the Pearson correlation coefficient.”
* **The independent variables should be linearly related to the log odds.** If you’re not familiar with log odds, we’ve included a brief explanation below.
* **Logistic regression requires fairly large sample sizes**—the larger the sample size, the more reliable (and powerful) you can expect the results of your analysis to be.

**What are log odds?**

In very simplistic terms, log odds are an alternate way of expressing probabilities. In order to understand log odds, it’s important to understand a key difference between odds and probabilities: odds are the ratio of something happening to something not happening, while probability is the ratio of something happening to everything that could possibly happen.

For example: if you and your friend play ten games of tennis, and you win four out of ten games, the odds of you winning are 4 to 6 (or, as a fraction, 4/6). The probability of you winning, however, is 4 to 10 (as there were ten games played in total). As we can see, odds essentially describe the ratio of success to the ratio of failure. In logistic regression, every probability or possible outcome of the dependent variable can be converted into log odds by finding the odds ratio. The log odds logarithm (otherwise known as the logit function) uses a certain formula to make the conversion. We won’t go into the details here, but if you’re keen to learn more, you’ll find [**a good explanation with examples in this guide**](https://towardsdatascience.com/https-towardsdatascience-com-what-and-why-of-log-odds-64ba988bf704).

**4. What is logistic regression used for?**

Now we know, in theory, what logistic regression is—but what kinds of real-world scenarios can it be applied to? Why is it useful?

Logistic regression is used to calculate the probability of a binary event occurring, and to deal with issues of classification. For example, predicting if an incoming email is spam or not spam, or predicting if a credit card transaction is fraudulent or not fraudulent. In a medical context, logistic regression may be used to predict whether a tumour is benign or malignant. In marketing, it may be used to predict if a given user (or group of users) will buy a certain product or not. An online education company might use logistic regression to predict whether a student will complete their course on time or not.

As you can see, logistic regression is used to predict the likelihood of all kinds of “yes” or “no” outcomes. By predicting such outcomes, logistic regression helps [**data analysts**](https://careerfoundry.com/en/blog/data-analytics/difference-between-data-scientist-and-data-analyst/) (and the companies they work for) to make informed decisions. In the grand scheme of things, this helps to both minimize the risk of loss and to optimize spending in order to maximize profits. And that’s what every company wants, right?

For example, it wouldn’t make good business sense for a credit card company to issue a credit card to every single person who applies for one. They need some kind of method or model to work out, or predict, whether or not a given customer will default on their payments. The two possible outcomes, “will default” or “will not default”, comprise binary data—making this an ideal use-case for logistic regression. Based on what category the customer falls into, the credit card company can quickly assess who might be a good candidate for a credit card and who might not be.

Similarly, a cosmetics company might want to determine whether a certain customer is likely to respond positively to a promotional 2-for-1 offer on their skincare range. In which case, they may use logistic regression to devise a model which predicts whether the customer will be a “responder” or a “non-responder.” Based on these insights, they’ll then have a better idea of where to focus their marketing efforts.

**[](https://careerfoundry.com/en/short-courses/become-a-data-analyst/?popup-tracking=blog-inline-DAT)**

**5. What are the different types of logistic regression?**

In this post, we’ve focused on just one type of logistic regression—the type where there are only two possible outcomes or categories (otherwise known as binary regression). In fact, there are three different types of logistic regression, including the one we’re now familiar with.

The three types of logistic regression are:

1. **Binary logistic regression** is the statistical technique used to predict the relationship between the dependent variable (Y) and the independent variable (X), where the dependent variable is binary in nature. For example, the output can be Success/Failure,  0/1 , True/False, or Yes/No. This is the type of logistic regression that we’ve been focusing on in this post.
2. **Multinomial logistic regression** is used when you have one categorical dependent variable with two or more unordered levels (i.e two or more discrete outcomes). It is very similar to logistic regression except that here you can have more than two possible outcomes. For example, let’s imagine that you want to predict what will be the most-used transportation type in the year 2030. The transport type will be the dependent variable, with possible outputs of train, bus, tram, and bike (for example).
3. **Ordinal logistic regression** is used when the dependent variable (Y) is ordered (i.e., ordinal). The dependent variable has a meaningful order and more than two categories or levels. Examples of such variables might be t-shirt size (XS/S/M/L/XL), answers on an opinion poll (Agree/Disagree/Neutral), or scores on a test (Poor/Average/Good).

**6. What are the advantages and disadvantages of using logistic regression?**

By now, you hopefully have a much clearer idea of what logistic regression is and the kinds of scenarios it can be used for. Now let’s consider some of the advantages and disadvantages of this type of regression analysis.

**Advantages of logistic regression**

* **Logistic regression is much easier to implement than other methods, especially in the context of machine learning:** A machine learning model can be described as a mathematical depiction of a real-world process. The process of setting up a machine learning model requires training and testing the model. Training is the process of finding patterns in the input data, so that the model can map a particular input (say, an image) to some kind of output, like a label. Logistic regression is easier to train and implement as compared to other methods.
* **Logistic regression works well for cases where the dataset is linearly separable:** A dataset is said to be linearly separable if it is possible to draw a straight line that can separate the two classes of data from each other. Logistic regression is used when your Y variable can take only two values, and if the data is linearly separable, it is more efficient to classify it into two separate classes.
* **Logistic regression provides useful insights:** Logistic regression not only gives a measure of how relevant an independent variable is (i.e. the (coefficient size), but also tells us about the direction of the relationship (positive or negative). Two variables are said to have a positive association when an increase in the value of one variable also increases the value of the other variable. For example, the more hours you spend training, the better you become at a particular sport. **However:** It is important to be aware that correlation does not necessarily indicate causation! In other words, logistic regression may show you that there is a positive correlation between outdoor temperature and sales, but this doesn’t necessarily mean that sales are rising because of the temperature. [**If you want to learn more about the difference between correlation and causation, take a look at this post**](https://towardsdatascience.com/correlation-vs-causation-a-real-world-example-9e939c85581e)

**Disadvantages of logistic regression**

* **Logistic regression fails to predict a continuous outcome.** Let’s consider an example to better understand this limitation. In medical applications, logistic regression cannot be used to predict how high a pneumonia patient’s temperature will rise. This is because the scale of measurement is continuous (logistic regression only works when the dependent or outcome variable is dichotomous).
* **Logistic regression assumes linearity between the predicted (dependent) variable and the predictor (independent) variables.** Why is this a limitation? In the real world, it is highly unlikely that the observations are linearly separable. Let’s imagine you want to classify the iris plant into one of two families: sentosa or versicolor. In order to distinguish between the two categories, you’re going by petal size and sepal size. You want to create an algorithm to classify the iris plant, but there’s actually no clear distinction—a petal size of 2cm could qualify the plant for both the sentosa and versicolor categories. So, while linearly separable data is the assumption for logistic regression, in reality, it’s not always truly possible.
* **Logistic regression may not be accurate if the sample size is too small.** If the sample size is on the small side, the model produced by logistic regression is based on a smaller number of actual observations. This can result in overfitting. In statistics, overfitting is a modelling error which occurs when the model is too closely fit to a limited set of data because of a lack of training data. Or, in other words, there is not enough input data available for the model to find patterns in it. In this case, the model is not able to accurately predict the outcomes of a new or future dataset.



**7. Final thoughts**

So there you have it: A complete introduction to logistic regression. Here are a few takeaways to summarize what we’ve covered:

* Logistic regression is used for classification problems when the output or dependent variable is dichotomous or categorical.
* There are some key assumptions which should be kept in mind while implementing logistic regressions (see section three).
* There are different types of regression analysis, and different types of logistic regression. It is important to choose the right model of regression based on the dependent and independent variables of your data.

Hopefully this post has been useful! If so, you might also enjoy this [**introductory guide to Bernoulli distribution**](https://careerfoundry.com/en/blog/data-analytics/what-is-bernoulli-distribution/)—a type of discrete probability distribution. And, if you’d like to learn more about [**forging a career as a data analyst**](https://careerfoundry.com/en/blog/data-analytics/data-analyst-career-path/), why not [**try out a free, introductory data analytics short course**](https://careerfoundry.com/en/short-courses/become-a-data-analyst/). From there, check out some of the [**best online data analytics courses**](https://careerfoundry.com/en/blog/data-analytics/data-analytics-courses/), then check out the following articles:

* [**A guide to the best data analytics bootcamps**](https://careerfoundry.com/en/blog/data-analytics/best-data-analytics-bootcamps/)
* [**What Is the Monte Carlo Method?**](https://careerfoundry.com/en/blog/data-analytics/monte-carlo-method/)
* [**What’s the difference between classification and regression?**](https://careerfoundry.com/en/blog/data-analytics/regression-vs-classification/)