**logistic regression**

By

* **George Lawton**
* **Ed Burns**
* **Linda Rosencrance**

**What is logistic regression?**

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

A logistic regression model predicts a [dependent data variable](https://www.techtarget.com/whatis/definition/dependent-variable) by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted or not to a particular college. These binary outcomes allow straightforward decisions between two alternatives.

A logistic regression model can take into consideration multiple input criteria. In the case of college acceptance, the logistic function could consider factors such as the student's grade point average, SAT score and number of extracurricular activities. Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into one of two outcome categories.

Logistic regression has become an important tool in the discipline of [machine learning](https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML). It allows algorithms used in machine learning applications to classify incoming data based on historical data. As additional relevant data comes in, the algorithms get better at predicting classifications within data sets.

Logistic regression can also play a role in [data preparation](https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation) activities by allowing data sets to be put into specifically predefined buckets during the extract, transform, load ([ETL](https://www.techtarget.com/searchdatamanagement/definition/Extract-Load-Transform-ELT)) process in order to stage the information for analysis.

**THIS ARTICLE IS PART OF**

[**What is predictive analytics? An enterprise guide**](https://www.techtarget.com/searchbusinessanalytics/definition/predictive-analytics)

* Which also includes:
* [**Predictive analytics vs. machine learning**](https://www.techtarget.com/searchenterpriseai/feature/Machine-learning-and-predictive-analytics-work-better-together)
* [**7 top predictive analytics use cases: Enterprise examples**](https://www.techtarget.com/searchbusinessanalytics/feature/Top-5-predictive-analytics-use-cases-in-enterprises)
* [**Descriptive vs. prescriptive vs. predictive analytics explained**](https://www.techtarget.com/searchbusinessanalytics/tip/Descriptive-vs-prescriptive-vs-predictive-analytics-explained)

**Logistics vs. logistic regression**

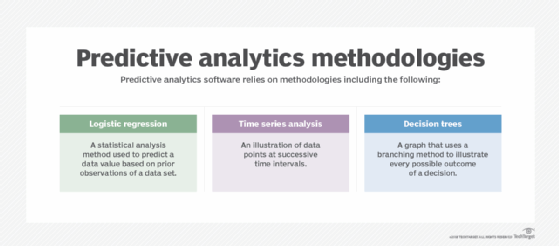
The etymology of logistic regression is a bit confusing. It is not linked to *logistics*, which evolved separately from a French word to describe a process for optimizing complex [supply chain calculations](https://www.techtarget.com/searcherp/Guide-to-supply-chain-management). In contrast, *logistic* (without the *s*) characterizes a mathematical technique for dividing phenomena into two categories.

Francis Galton coined the term *regression* in 1889 to characterize a biological phenomenon in which tall people's descendants regress toward the average heights of the population. Subsequent researchers adopted the term to describe a process for representing the effect of independent variables on probability.

Regression is a cornerstone of modern [predictive analytics](https://www.techtarget.com/searchbusinessanalytics/definition/predictive-analytics) applications.

"Predictive analytics tools can broadly be classified as traditional regression-based tools or machine learning-based tools," said Donncha Carroll, a partner in the revenue growth practice of Axiom Consulting Partners.

Regression models essentially represent or encapsulate a mathematical equation that approximates the interactions between the different variables being modelled. Machine learning models use and train on a combination of input and output data and use new data to predict the output.

Logistic regression is one of various data modelling techniques used to forecast outcomes.

**What is the purpose of logistic regression?**

Logistic regression streamlines the mathematics for measuring the impact of multiple variables (e.g., age, gender, ad placement) with a given outcome (e.g., click-through or ignore). The resulting models can help tease apart the relative effectiveness of various interventions for different categories of people, such as young/old or male/female.

Logistic models can also transform raw data streams to create features for other types of AI and [machine learning techniques](https://www.techtarget.com/searchenterpriseai/feature/Machine-learning-and-predictive-analytics-work-better-together). In fact, logistic regression is one of the commonly used algorithms in machine learning for binary classification problems, which are problems with two class values, including predictions such as "this or that," "yes or no," and "A or B."

Logistic regression can also estimate the probabilities of events, including determining a relationship between features and the probabilities of outcomes. That is, it can be used for classification by creating a model that correlates the hours studied with the likelihood the student passes or fails. On the flip side, the same model could be used for predicting whether a particular student will pass or fail when the number of hours studied is provided as a feature and the variable for the response has two values: pass and fail.

**Logistic regression applications in business**

Organizations use insights from logistic regression outputs to enhance their business strategy for achieving business goals such as reducing expenses or losses and increasing [ROI](https://www.techtarget.com/searchcio/definition/ROI) in marketing campaigns.

An [e-commerce](https://www.techtarget.com/searchcio/definition/e-commerce) company that mails expensive promotional offers to customers, for example, would like to know whether a particular customer is likely to respond to the offers or not: i.e., whether that consumer will be a "responder" or a "non-responder." In marketing, this is called *propensity to respond modelling*.

Likewise, a credit card company will develop a model to help it predict if a customer is going to default on its credit card based on such characteristics as annual income, monthly credit card payments and the number of defaults. In banking parlance, this is known as *default propensity modeling*.

**Why is logistic regression important?**

Logistic regression is important because it transforms complex calculations around probability into a straightforward arithmetic problem. Admittedly, the calculation itself is a bit complex, but modern statistical applications automate much of this grunt work. This dramatically simplifies analyzing the impact of multiple variables and helps to minimize the effect of confounding factors.

As a result, statisticians can quickly model and explore the contribution of various factors to a given outcome.

For example, a medical researcher may want to know the impact of a new drug on treatment outcomes across different age groups. This involves a lot of nested multiplication and division for comparing the outcomes of young and older people who never received a treatment, younger people who received the treatment, older people who received the treatment, and then the whole spontaneous healing rate of the entire group. Logistic regression converts the relative probability of any subgroup into a logarithmic number, called a regression coefficient, that can be added or subtracted to arrive at the desired result.

These more straightforward regression coefficients can also simplify other machine learning and data science algorithms.

**What are key assumptions of logistic regression?**

Statisticians and citizen data scientists must keep a few assumptions in mind when using logistic regression. For starters, the variables must be independent of one another. So, for example, zip code and gender could be used in a model, but zip code and state would not work.

Other less transparent relationships between variables may get lost in the noise when logistic regression is used as a starting point for complex machine learning and data science applications. For example, data scientists may spend considerable effort to ensure that variables associated with discrimination, such as gender and ethnicity, are not included in the algorithm. However, these can sometimes get indirectly woven into the algorithm via variables that were not thought to be correlated, such as zip code, school or hobbies.

Another assumption is that the raw data should represent unrepeated or independent phenomena. For example, a survey of customer satisfaction should represent the opinions of separate people. But these results would be skewed if someone took the survey multiple times from different email addresses to qualify for a reward.

It's also important that the relationship between the variables and the outcome can be linearly related via logarithmic odds, which is a bit more flexible than a linear relationship.

Logistic regression also requires a significant sample size. This can be as small as 10 examples of each variable in a model. But this requirement goes up as the probability of each outcome drops.

Another assumption with logistic regression is that each variable can be represented using binary categories such as male/female, click/no-click. A special trick is required to represent categories with more than two classes. For example, you might transform one category with three age ranges into three separate variables, where each specifies whether an individual is in that age range or not.

**Logistic regression use cases**

Logistic regression has become particularly popular in online advertising, enabling [marketers to predict the likelihood of specific website users](https://www.techtarget.com/searchbusinessanalytics/tip/Predictive-analytics-in-marketing-Achieving-success) who will click on particular advertisements as a yes or no percentage.

Logistic regression can also be used in the following areas:

* in [healthcare to identify risk factors for diseases](https://www.techtarget.com/searchbusinessanalytics/tip/Predictive-analytics-in-healthcare-12-valuable-use-cases) and plan preventive measures;
* in drug research to tease apart the effectiveness of medicines on health outcomes across age, gender and ethnicity;
* in weather forecasting apps to predict snowfall and weather conditions;
* in political polls to determine if voters will vote for a particular candidate;
* in insurance to predict the chances that a policyholder will die before the policy's term expires based on specific criteria, such as gender, age and physical examination; and
* in banking to predict the chances that a loan applicant will default on a loan or not, based on annual income, past defaults and past debts.

**Advantages and disadvantages of logistic regression**

The main advantage of logistic regression is that it is much easier to set up and train than other machine learning and AI applications.

Another advantage is that it is one of the most efficient algorithms when the different outcomes or distinctions represented by the data are linearly separable. This means that you can draw a straight line separating the results of a logistic regression calculation.

One of the biggest attractions of logistic regression for statisticians is that it can help reveal the interrelationships between different variables and their impact on outcomes. This could quickly determine when two variables are positively or negatively correlated, such as the finding cited above that more studying tends to be correlated with higher test outcomes. But it is important to note that other techniques like [causal AI](https://www.techtarget.com/searchenterpriseai/feature/Causal-deep-learning-teaches-AI-to-ask-why) are required to make the leap from correlation to causation.

**Logistic regression tools**

Logistic regression calculations were a laborious and time-consuming task before the advent of modern computers. Now, modern [statistical analytics tools such as SPSS and SAS](https://www.techtarget.com/searchbusinessanalytics/tip/6-top-predictive-analytics-tools) include logistic regression capabilities as an essential feature.

Also, data science programming languages and frameworks built on R and Python include numerous ways of performing logistic regression and weaving the results into other algorithms. There are also various tools and techniques for doing logistic regression analysis on top of Excel.

Managers should also consider other data preparation and management tools as part of significant data science democratization efforts. For example, data warehouses and data lakes can help organize larger data sets for analysis. Data catalogue tools can help surface any quality or usability issues associated with logistic regression. Data science platforms can help analytics leaders create appropriate guardrails to simplify the broader use of logistic regression across the enterprise.

**Logistic regression vs. linear regression**

The main difference between logistic and linear regression is that logistic regression provides a constant output, while linear regression provides a continuous output.

In logistic regression, the outcome, or dependent variable, has only two possible values. However, in linear regression, the outcome is continuous, which means that it can have any one of an infinite number of possible values.

Logistic regression is used when the response variable is categorical, such as yes/no, true/false and pass/fail. Linear regression is used when the response variable is continuous, such as hours, height and weight.

For example, given data on the time a student spent studying and that student's exam scores, logistic regression and linear regression can predict different things.

With logistic regression predictions, only specific values or categories are allowed. Therefore, logistic regression predicts whether the student passed or failed. Since linear regression predictions are continuous, such as numbers in a range, it can predict the student's test score on a scale of 0 to100.