logistic
ion?

regression can help make
ance decision-making

What is logistic regression?

This type of statistical model (also known as *logit model*) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

$$\mathrm{Logit}(pi) = 1/(1+ \exp(-pi))$$

$$\ln(pi/(1-pi)) = Beta_0 + Beta_1*X_1 + \ldots + B_k*K_k$$

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than .5 will predict 0 while a probability greater than 0 will predict 1. After the model has been computed, it's best practice to evaluate the how well the model predicts the dependent variable, which is called goodness of fit. The Hosmer–Lemeshow test is a popular method to assess model fit.

Interpreting logistic regression

Log odds can be difficult to make sense of within a logistic regression data analysis. As a result, exponentiating the beta estimates is common to transform the results into an odds ratio (OR), easing the interpretation of results. The OR represents the odds that an outcome will occur given a particular event, compared to the odds of the outcome occurring in the absence of that event. If the OR is greater than 1, then the event is associated with a higher odds of generating a specific outcome. Conversely, if the OR is less than 1, then the event is associated with a lower odds of that outcome occurring. Based on the equation from above, the interpretation of an odds ratio can be denoted as the following: the odds of a success changes by exp(cB_1) times for every c-unit increase in x. To use an example, let's say that we were to estimate the odds of survival on the Titanic given that the person was male, and the odds ratio for males was .0810. We'd interpret the odds ratio as the odds of survival of males decreased by a factor of .0810 when compared to females, holding all other variables constant.

Read the white paper (776 KB) PDF

Linear regression vs logistic regression

Both linear and logistic regression are among the most popular models within data science, and open-source tools, like Python and R, make the computation for them quick and easy.

Linear regression models are used to identify the relationship between a continuous dependent variable and one or more independent variables. When there is only one independent variable and one dependent variable, it is known as simple linear regression, but as the number of independent variables increases, it is referred to as multiple linear regression. For each type of linear regression, it seeks to plot a line of best fit through a set of data points, which is typically calculated using the least squares method.

Similar to linear regression, logistic regression is also used to estimate the relationship between a dependent variable and one or more independent variables, but it is used to make a prediction about a categorical variable versus a continuous one. A categorical variable can be true or false, yes or no, 1 or 0, et cetera. The unit of measure also differs from linear regression as it produces a probability, but the logit function transforms the S-curve into straight line.

While both models are used in regression analysis to make predictions about future outcomes, linear regression is typically easier to understand. Linear regression also does not require as large of a sample size as logistic regression needs an adequate sample to represent values across all the response categories. Without a larger, representative sample, the model may not have sufficient statistical power to detect a significant effect.

Types of logistic regression

There are three types of logistic regression models, which are defined based on categorical response.

– **Binary logistic regression:** In this approach, the response or dependent variable is dichotomous in nature—i.e. it has only two possible outcomes (e.g. 0 or 1). Some popular examples of its use include predicting if an e-mail is spam or not spam or if a tumor is malignant or not malignant. Within logistic regression, this is the most commonly used approach, and more generally, it is one of the most common classifiers for binary classification.
– **Multinomial logistic regression:** In this type of logistic regression model, the dependent variable has three or more possible outcomes; however, these values have no specified order. For example, movie studios want to predict what genre of film a moviegoer is likely to see to market films more effectively. A multinomial logistic regression model can help the studio to determine the strength of influence a person's age, gender, and dating status may have on the type of film that they prefer. The studio can then orient an advertising campaign of a specific movie toward a group of people likely to go see it.
– **Ordinal logistic regression:** This type of logistic regression model is leveraged when the response variable has three or more possible outcome, but in this case, these values do have a defined order. Examples of ordinal responses include grading scales from A to F or rating scales from 1 to 5.

A glimpse inside the mind of a data scientist
(776 KB) PDF

Logistic regression and machine learning

Within machine learning, logistic regression belongs to the family of supervised machine learning models. It is also considered a discriminative model, which means that it attempts to distinguish between classes (or categories). Unlike a generative algorithm, such as naïve bayes, it cannot, as the name implies, generate information, such as an image, of the class that it is trying to predict (e.g. a picture of a cat).

Previously, we mentioned how logistic regression maximizes the log likelihood function to determine the beta coefficients of the model. This changes slightly under the context of machine learning. Within machine learning, the negative log likelihood used as the loss function, using the process of gradient descent to find the global maximum. This is just another way to arrive at the same estimations discussed above.

Logistic regression can also be prone to overfitting, particularly when there is a high number of predictor variables within the model. Regularization is typically used to penalize parameters large coefficients when the model suffers from high dimensionality.

Scikit-learn (link resides outside IBM) provides valuable documentation to learn more about the logistic regression machine learning model.

Use cases of logistic regression

Logistic regression is commonly used for prediction and classification problems. Some of these use cases include:

- **Fraud detection:** Logistic regression models can help teams identify data anomalies, which are predictive of fraud. Certain behaviors or characteristics may have a higher association with fraudulent activities, which is particularly helpful to banking and other financial institutions in protecting their clients. SaaS-based companies have also started to adopt these practices to eliminate fake user accounts from their datasets when conducting data analysis around business performance.

- **Disease prediction:** In medicine, this analytics approach can be used to predict the likelihood of disease or illness for a given population. Healthcare organizations can set up preventative care for individuals that show higher propensity for specific illnesses.
- **Churn prediction**: Specific behaviors may be indicative of churn in different functions of an organization. For example, human resources and management teams may want to know if there are high performers within the company who are at risk of leaving the organization; this type of insight can prompt conversations to understand problem areas within the company, such as culture or compensation. Alternatively, the sales organization may want to learn which of their clients are at risk of taking their business elsewhere. This can prompt teams to set up a retention strategy to avoid lost revenue.

Examples of logistic regression success

Assess credit risk

Binary logistic regression can help bankers assess credit risk. Imagine that you are a loan officer at a bank and you want to identify characteristics of people who are likely to default

Increase profits in the banking industry

First Tennessee Bank boosted profitability with IBM SPSS software and achieved increases of up to 600 percent in cross-sale campaigns. Leaders at this regional bank in on loans. Then you want to use those characteristics to identify good and bad credit risks. You have data on 850 customers. The first 700 are customers who have already received loans. See how you can use a logistic regression model and classify the 150 remaining customers as good or bad risks.

→

the US wanted to approach the right customers with the right products and services. There is no shortage of data to help, but it was a challenge to bridge the gap from having data to taking action. First Tennessee is using predictive analytics and logistic analytics techniques within an analytics solution to gain greater insight into all of its data. As a result, decision-making is improved to optimize customer interactions. (1 MB)

PDF

Related solutions

IBM SPSS Advanced Statistics

Reach more accurate conclusions when analyzing complex relationships using univariate and multivariate modeling techniques.

Explore SPSS Advanced Statistics →

IBM SPSS Modeler

Drive return on investment with a drag-and-drop data science tool.

Explore SPSS Modeler →

IBM SPSS Regression

Predict categorical outcomes and apply a wide range of nonlinear regression procedures.

Explore SPSS Regression →

IBM Watson Studio

Build and train AI and machine-learning models, prepare and analyze data — all in a flexible, hybrid cloud environment.

Explore Watson Studio →

IBM Watson Discovery

Get a smart, simple way to mine and explore all your unstructured data with cognitive exploration, powerful text analytics and machine-learning capabilities.

Explore Watson Discovery →

Resources

IBM SPSS Statistics 14-day free trial

→

IBM SPSS Statistics statistical analysis demo

→

Learn more about IBM Watson Studio Local

→