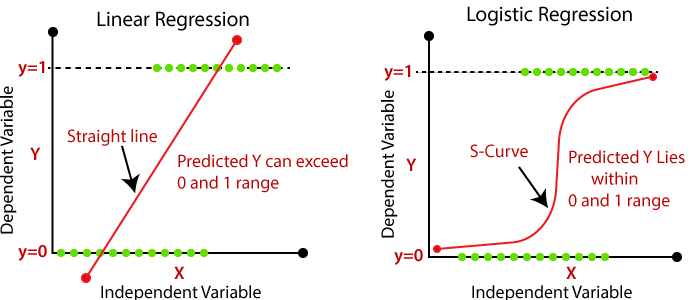
**3.1 Basic Concept**

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a [data set](https://whatis.techtarget.com/definition/data-set). Logistic regression has become an important tool in the discipline of [machine learning](https://searchenterpriseai.techtarget.com/definition/machine-learning-ML). The approach allows an [algorithm](https://whatis.techtarget.com/definition/algorithm) being used in a machine learning application to classify incoming data based on historical data. A logistic regression model predicts a [dependent data variable](https://whatis.techtarget.com/definition/dependent-variable) by analyzing the relationship between one or more existing independent variables. For example, a logistic regression can be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted to a particular college. The purpose of logistic regression is to estimate the probabilities of events, including determining a relationship between features and the probabilities of particular outcomes. Multinomial logistic regression can be used to classify subjects into groups based on a categorical range of variables to predict behavior. Binary logistic regression is most useful when we want to model the event probability for a categorical response variable with two outcomes.



## Figure 3.1 linear vs logistic regression

The regression line can be written as:

Where, a0 and a1 are the coefficients and ε is the error term.

The equation for logistic regression is:

## 3.2 Logistic Function

We must wonder how logistic regression squeezes the output of linear regression between 0 and 1. We start by mentioning the formula of logistic function:

We know the equation of the best fit line in linear regression is:

Where:

* x is the input value
* y is the predicted output
* b0 is the bias or intercept term
* b1 is the coefficient for the single input value (x)

Instead of y we are taking probabilities (P). But there is an issue here, the value of (P) will exceed 1 or go below 0 and we know that range of Probability is (0-1). To overcome this issue we take “odds” of P:

We know that odds can always be positive which means the range will always be (0,+∞ ). Odds are nothing but the ratio of the probability of success and probability of failure. The problem here is that the range is restricted and we don’t want a restricted range because if we do so then correlation will decrease. By restricting the range we are actually decreasing the number of data points and of course, if we decrease our data points, correlation will decrease. It is difficult to model a variable that has a restricted range. To control this we take the *log of odds*which has a range from (-∞, +∞).

Now we just want a function of P because we want to predict probability right? not log of odds. To do so we will multiply by exponent on both sides and then solve for P.

Now we have our logistic function, also called a sigmoid function. The graph of a sigmoid function is as shown below. It squeezes a straight line into an S-curve.

### 3.3 ****Logistic Regression Assumptions****

While logistic regression seems like a fairly simple algorithm to adopt & implement, there are a lot of restrictions around its use. For instance, it can only be applied to large datasets. Similarly, multiple assumptions need to be made in a dataset to be able to apply this machine learning algorithm.

* The dependent variable has to be binary in a binary logistic equation
* The factor level 1 of the dependent variable should represent the desired outcome
* Including non-meaningful variables may throw errors. Only include the variables that are necessary and may show a correlation
* The model should have little or no multicollinearity – the independent variables should be absolutely independent of each other
* The independent variables are linearly related to the log odds

With so many assumptions that need to be made, you may think that the equation is not versatile enough to be implemented across real-life problems but this equation has a lot of applications in the medical field and is helping people across the world with its superpower.

**3.5 Illustrate with an example**

Table 3.1 simple data set for number of hours study and probability of passing

|  |  |
| --- | --- |
| **Hours** | **Pass** |
| 0.5 | 0 |
| 0.75 | 0 |
| 1 | 0 |
| 1.25 | 0 |
| 1.5 | 0 |
| 1.75 | 0 |
| 1.75 | 1 |
| 2 | 0 |
| 2.25 | 1 |
| 2.5 | 0 |
| 2.75 | 1 |
| 3 | 0 |
| 3.25 | 1 |
| 3.5 | 0 |
| 4 | 1 |
| 4.25 | 1 |
| 4.5 | 1 |
| 4.75 | 1 |
| 5 | 1 |
| 5.5 | 1 |

First we need to calculate intercept (b) and slope (m)

Table 3.2 Calculation of mean of hours and pass

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Hours(x)** | **Pass(y)** | ) |  | ) |  |
| 0.5 | 0 | -2.2875 | -0.5 | 1.14375 | 5.23265625 |
| 0.75 | 0 | -2.0375 | -0.5 | 1.01875 | 4.15140625 |
| 1 | 0 | -1.7875 | -0.5 | 0.89375 | 3.19515625 |
| 1.25 | 0 | -1.5375 | -0.5 | 0.76875 | 2.36390625 |
| 1.5 | 0 | -1.2875 | -0.5 | 0.64375 | 1.65765625 |
| 1.75 | 0 | -1.0375 | -0.5 | 0.51875 | 1.07640625 |
| 1.75 | 1 | -1.0375 | 0.5 | -0.51875 | 1.07640625 |
| 2 | 0 | -0.7875 | -0.5 | 0.39375 | 0.62015625 |
| 2.25 | 1 | -0.5375 | 0.5 | -0.26875 | 0.28890625 |
| 2.5 | 0 | -0.2875 | -0.5 | 0.14375 | 0.08265625 |
| 2.75 | 1 | -0.0375 | 0.5 | -0.01875 | 0.00140625 |
| 3 | 0 | 0.2125 | -0.5 | -0.10625 | 0.04515625 |
| 3.25 | 1 | 0.4625 | 0.5 | 0.23125 | 0.21390625 |
| 3.5 | 0 | 0.7125 | -0.5 | -0.35625 | 0.50765625 |
| 4 | 1 | 1.2125 | 0.5 | 0.60625 | 1.47015625 |
| 4.25 | 1 | 1.4625 | 0.5 | 0.73125 | 2.13890625 |
| 4.5 | 1 | 1.7125 | 0.5 | 0.85625 | 2.93265625 |
| 4.75 | 1 | 1.9625 | 0.5 | 0.98125 | 3.85140625 |
| 5 | 1 | 2.2125 | 0.5 | 1.10625 | 4.89515625 |
| 5.5 | 1 | 2.7125 | 0.5 | 1.35625 | 7.35765625 |

) =10.125

= 43.159

Now the equation for Logit is

(-0.1539)

now we calculate logit value for each independent variable

Table 3.3 Calculation logit values

|  |  |  |
| --- | --- | --- |
| Hours | Pass | Logit |
| 0.5 | 0 | -0.0366 |
| 0.75 | 0 | 0.02205 |
| 1 | 0 | 0.0807 |
| 1.25 | 0 | 0.13935 |
| 1.5 | 0 | 0.198 |
| 1.75 | 0 | 0.25665 |
| 1.75 | 1 | 0.25665 |
| 2 | 0 | 0.3153 |
| 2.25 | 1 | 0.37395 |
| 2.5 | 0 | 0.4326 |
| 2.75 | 1 | 0.49125 |
| 3 | 0 | -0.6156 |
| 3.25 | 1 | 0.60855 |
| 3.5 | 0 | 0.6672 |
| 4 | 1 | 0.7845 |
| 4.25 | 1 | 0.84315 |
| 4.5 | 1 | 0.9018 |
| 4.75 | 1 | 0.96045 |
| 5 | 1 | 1.0191 |
| 5.5 | 1 | 1.1364 |

(-0.15366)

Now put x=0.5

(-0.15366)

-0.15366

Now calculate exp(value of logit) also called odds

Table 3.4 Calculation odds values

|  |  |  |  |
| --- | --- | --- | --- |
| Hours | Pass | Logit | Odds |
| 0.5 | 0 | **-0.0366** | 0.96406168 |
| 0.75 | 0 | 0.02205 | 1.0222949 |
| 1 | 0 | 0.0807 | 1.08404563 |
| 1.25 | 0 | 0.13935 | 1.14952636 |
| 1.5 | 0 | 0.198 | 1.21896239 |
| 1.75 | 0 | 0.25665 | 1.29259264 |
| 1.75 | 1 | 0.25665 | 1.29259264 |
| 2 | 0 | 0.3153 | 1.37067045 |
| 2.25 | 1 | 0.37395 | 1.45346448 |
| 2.5 | 0 | 0.4326 | 1.54125959 |
| 2.75 | 1 | 0.49125 | 1.63435789 |
| 3 | 0 | -0.6156 | 0.54031661 |
| 3.25 | 1 | 0.60855 | 1.83776471 |
| 3.5 | 0 | 0.6672 | 1.94877311 |
| 4 | 1 | 0.7845 | 2.19131101 |
| 4.25 | 1 | 0.84315 | 2.32367504 |
| 4.5 | 1 | 0.9018 | 2.46403438 |
| 4.75 | 1 | 0.96045 | 2.612872 |
| 5 | 1 | 1.0191 | 2.77070001 |
| 5.5 | 1 | 1.1364 | 3.11553224 |

Now finally calculate the probability of passing by using the formula

Table 3.5 Calculation of probability of passing

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Hours | Pass | Logit | Odds | pro |
| 0.5 | 0 | -0.0366 | 0.96406168 | **0.490851021** |
| 0.75 | 0 | 0.02205 | 1.0222949 | 0.505512277 |
| 1 | 0 | 0.0807 | 1.08404563 | 0.520164058 |
| 1.25 | 0 | 0.13935 | 1.14952636 | 0.534781235 |
| 1.5 | 0 | 0.198 | 1.21896239 | 0.549338915 |
| 1.75 | 0 | 0.25665 | 1.29259264 | 0.56381261 |
| 1.75 | 1 | 0.25665 | 1.29259264 | 0.56381261 |
| 2 | 0 | 0.3153 | 1.37067045 | 0.578178401 |
| 2.25 | 1 | 0.37395 | 1.45346448 | 0.592413092 |
| 2.5 | 0 | 0.4326 | 1.54125959 | 0.606494353 |
| 2.75 | 1 | 0.49125 | 1.63435789 | 0.620400856 |
| 3 | 0 | -0.6156 | 0.54031661 | 0.350782823 |
| 3.25 | 1 | 0.60855 | 1.83776471 | 0.647609967 |
| 3.5 | 0 | 0.6672 | 1.94877311 | 0.660875909 |
| 4 | 1 | 0.7845 | 2.19131101 | 0.686649156 |
| 4.25 | 1 | 0.84315 | 2.32367504 | 0.699128227 |
| 4.5 | 1 | 0.9018 | 2.46403438 | 0.711319263 |
| 4.75 | 1 | 0.96045 | 2.612872 | 0.723211894 |
| 5 | 1 | 1.0191 | 2.77070001 | 0.734797253 |
| 5.5 | 1 | 1.1364 | 3.11553224 | 0.757018062 |

**3.6 Outline of the proposed approach**

Calculate intercept (b) and slope (m)

Calculate exp(value of logit) or odds values for each Logit value

Now put these value to find the equation Calculate

Calculate Logit value for each independent variable

Finally calculate probability values using the equation

Now put an unknown value to calculate probability of passing and check accuracy

Figure 3.2 Outline of proposed approach