LOGISTIC REGRESSION

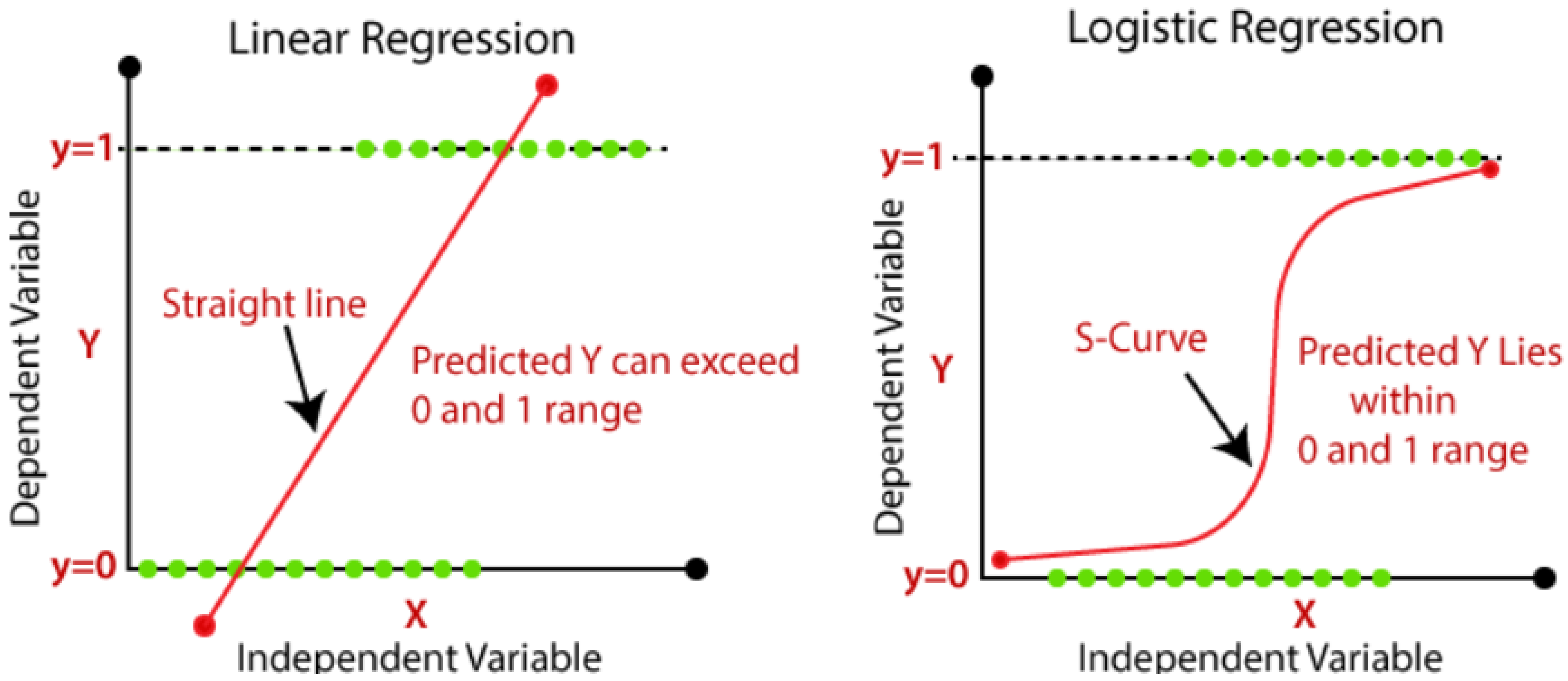
What is Logistic regression?

- **Logistic Regression** is a statistical method for analyzing a dataset with one or more independent variables that determine an outcome.

There will be **two** possible outcomes in **Binary Logistic Regression** and multiple possible outcomes in Multinomial Logistic Regression.

What is Logistic regression?

ML Labs Pvt Ltd



Logistic regression

- The goal of logistic regression is to find the best fit model to describe the relationship between the dependent Variable and a set of independent variables.

Assumptions of Logistic Regression:

- The dependent variable must be of 2 categories for binary Logistic and ordinal for multinomial Logistic Regression.

- Assumes a linear relationship between the logit of the Independent Variables and Dependent Variables.

- Absence of multi-collinearity.

Assumptions of Logistic Regression:

- More samples are needed when compared to linear regression.

- Normal distribution is **not** assumed either for the dependent variable or for errors.

- The independent variables need **not** be in **intervals**, nor normally distributed, nor of equal variance within each group.

Assumption of Appropriate Outcome Structure:

Binary logistic regression requires dependent variable to be **binary** and

Multinomial logistic regression requires the dependent variable to be **ordinal**.

Assumption of Linearity

Although Logistic regression does not require the dependent and independent variables to be related linearly, it requires independent variables to be linearly related to the **log odds**.

Assumption of Absence of Multi-collinearity

Logistic regression requires **little or no multicollinearity** among the independent variables, in otherwards independent variables should not be too highly correlated with each other.

Assumption of Large Sample Size

Logistic regression typically requires a **large sample size**.

A general guideline is that a minimum of 10 samples are needed with the least frequent outcome for each independent variable in our model.

Assumption of Large Sample Size

For example, if we have 3 independent variables and the expected probability of our least frequent outcome is .10, then we would need a minimum sample size of 300 (10*3 / .10).

Assumption of Observation Independence

Logistic regression requires the observations to be independent of each other.

The observations should not come from repeated measurements or matched data.

Math behind Logistic Regression

Model Development- Binary Logistic Regression

ML Labs Pvt Ltd

As Logistic Regression gives the formula to predict a logit transformation of probability of presence of character of interest, so, the model is,

$$logit(p) = b_0 + b_1x_1 + ...+ b_kx_k$$

In logistic regression, the dependent variable is in fact a logit, which is a log of odds,

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

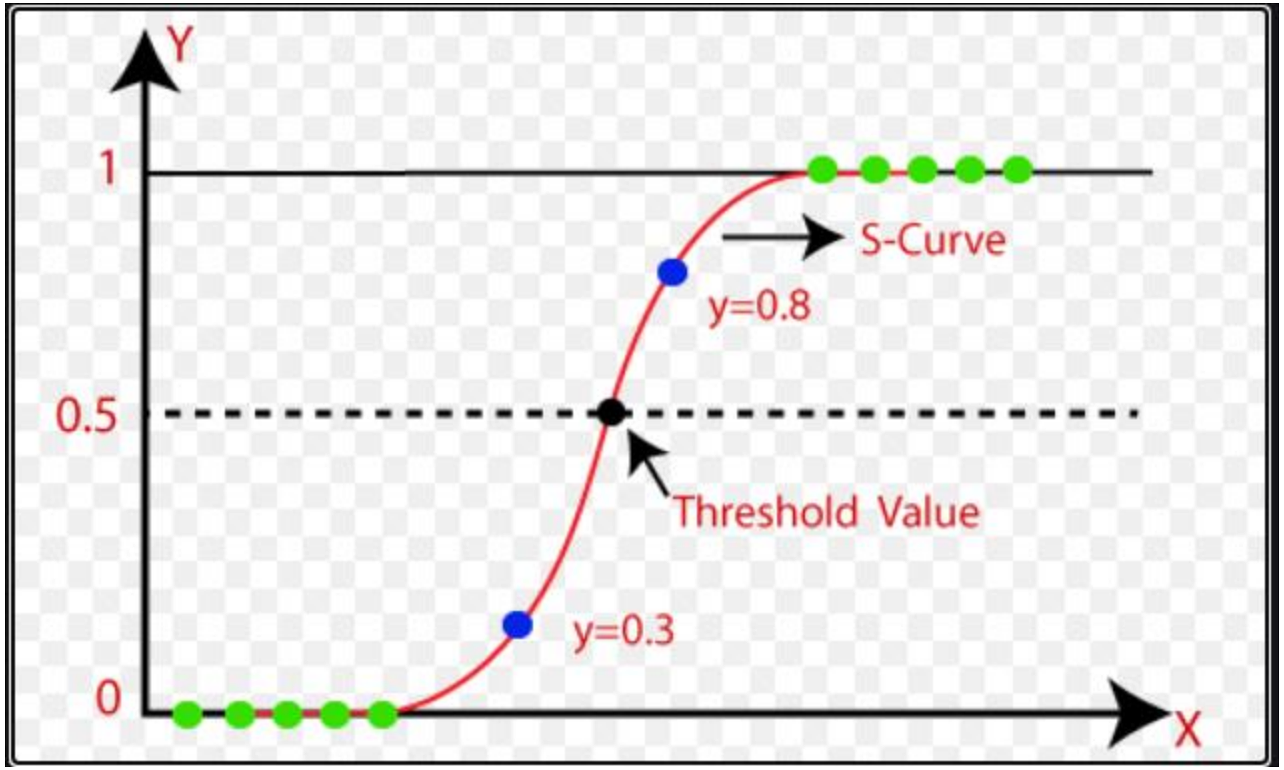
So, the required probability is-

$$p = \frac{e^{logit(p)}}{1 + e^{logit(p)}}$$

Model Development- Binary Logistic Regression

probability value needs to be converted to class value which is “0” or “1”.

If p < 0.5----------------□ Class 0
p>=0.5---------------□ Class 1



Multinomial Logistic Regression

Multinomial logit regression is used when the dependent variable in question is nominal and for which there are more than two categories.

Assumptions of Multinomial Logistic Regression

The multinomial logit model assumes that data are case specific, that is, each independent variable has a single value for each case.

There is no need for the independent variables to be statistically independent from each other.

Model:

In multinomial logistic regression there are more than two categories for dependent variable, so the probability of belonging to category ‘j’ is given by:

$$pr(y_i\text{=j})=\frac{\exp(x_iB_j)}{\sum_i^j(\exp(x_iB_j)}$$