

## Logistic Regression

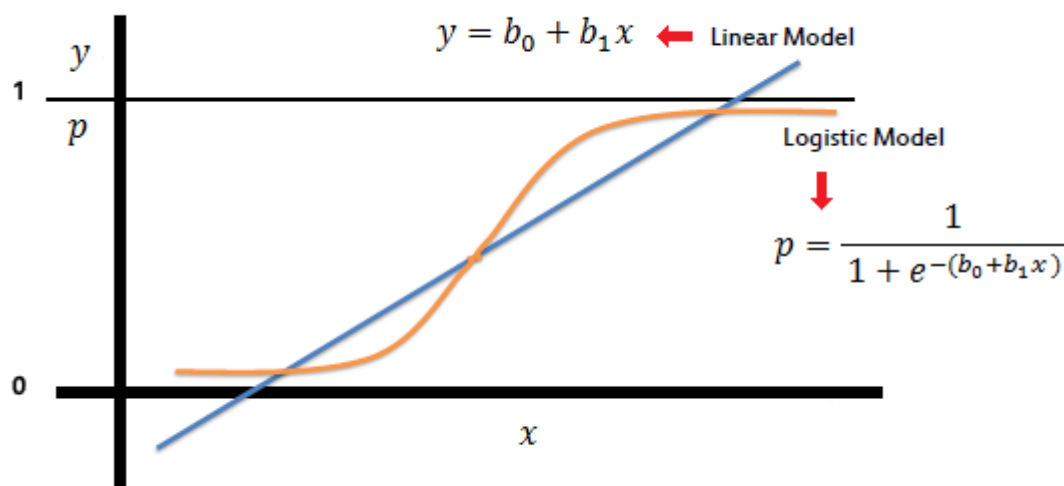
- In statistics, the logistic model is a statistical model that is usually taken to apply to a binary dependent variable. In regression analysis, logistic regression or logit regression is estimating the parameters of a logistic model.
- In Logistic Regression, the dependent variable that is binary (a categorical variable that has two values such as "yes" and "no") rather than continuous and it can also be applied to ordered categories (ordinal data), that is, variables with more than two ordered categories.

## Why use logistic regression rather than ordinary linear regression?

Logistic Regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

- A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)
- Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

On the other hand, logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to linear regression, but the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.



## The Logistic Curve

The logistic curve relates the independent variable,  $X$ , to the rolling mean of the DV (dependent variable),  $P(\bar{Y})$ . The formula to do so may be written either

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

Or

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

where P is the probability of a 1 (the proportion of 1s, the mean of Y), e is the base of the natural logarithm (about 2.718) and 'a' and 'b' are the parameters of the model.

### Derivation

In logistic regression, the dependent variable is a logit, which is the natural log of the odds, that is,

$$odds = \frac{P}{1 - P}$$

$$\log(odds) = \text{logit}(P) = \ln\left(\frac{P}{1 - P}\right)$$

So a logit is a log of odds and odds are a function of P, the probability of a 1. In logistic regression, we find

$$\text{logit}(P) = a + bX,$$

$$\ln\left(\frac{P}{1 - P}\right) = a + bX$$

$$\frac{P}{1 - P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

## Assumptions

- The outcome is a binary or dichotomous variable like yes vs no, positive vs negative, 1 vs 0.
- There is a linear relationship between the logit of the outcome and each predictor variables. Recall that the logit function is  $\text{logit}(p) = \log(p/(1-p))$ , where  $p$  is the probabilities of the outcome.
- There are no influential values (extreme values or outliers) in the continuous predictors
- There is no high intercorrelations (i.e. multicollinearity) among the predictors.





## Confusion Matrix

Well, it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

It is extremely useful for measuring Recall, Precision, Specificity, Accuracy and most importantly AUC-ROC Curve.

Let's understand TP, FP, FN, TN in terms of pregnancy analogy.

		Actual Values	
		1	0
Predicted Values	1	<b>TRUE POSITIVE</b>  You're pregnant	<b>FALSE POSITIVE</b>  You're pregnant <b>TYPE 1 ERROR</b>
	0	<b>FALSE NEGATIVE</b>  You're not pregnant <b>TYPE 2 ERROR</b>	<b>TRUE NEGATIVE</b>  You're not pregnant

**True Positive:**

Interpretation: You predicted positive and it's true.

You predicted that a woman is pregnant and she actually is.

**True Negative:**

Interpretation: You predicted negative and it's true.

You predicted that a woman is not pregnant and she actually is not.

**False Positive: (Type 1 Error)**

Interpretation: You predicted positive and it's false.

You predicted that a woman is pregnant but she actually is not.

**False Negative: (Type 2 Error)**

Interpretation: You predicted negative and it's false.

You predicted that a woman is not pregnant but she actually is.

Let's understand confusion matrix through math.

**Recall**

$$\text{Recall} = \frac{TP}{TP + FN}$$

Out of all the positive classes, how much we predicted correctly. It should be high as possible.

**Precision**

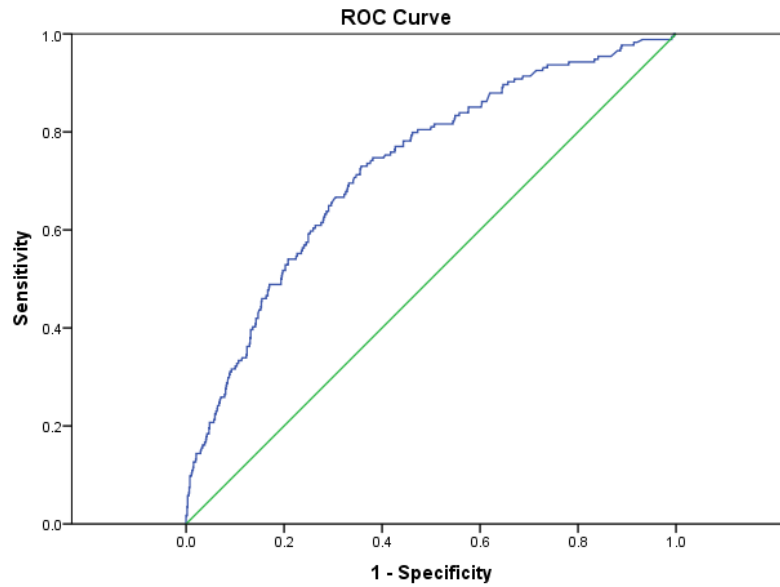
$$\text{Precision} = \frac{TP}{TP + FP}$$

Out of all the classes, how much we predicted correctly. It should be high as possible.

**ROC Curve**

This is a commonly used graph that summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) as you vary the threshold for assigning observations to a given class.

The ROC curve plots out the sensitivity and specificity for every possible decision rule cutoff between 0 and 1 for a model.



Diagonal segments are produced by ties.

This plot tells you a few different things.

A model that predicts at chance will have an ROC curve that looks like the diagonal green line. That is not a discriminating model.

The further the curve is from the diagonal line, the better the model is at discriminating between positives and negatives in general.

There are useful statistics that can be calculated from this curve, like the Area Under the Curve (AUC) and the Youden Index. These tell you how well the model predicts and the optimal cut point for any given model.