Logistic regression:

 is used to predict the class (or category) of individuals based on one or multiple predictor variables (x). It is used to model a binary outcome, that is a variable, which can have only two possible values: 0 or 1, yes or no, diseased or non-diseased.

Logistic regression belongs to a family, named Generalized Linear Model (GLM), developed for extending the linear regression model (Chapter @ref(linear-regression)) to other situations. Other synonyms are binary logistic regression, binomial logistic regression and logit model.

Logistic regression does not return directly the class of observations. It allows us to estimate the probability (p) of class membership. The probability will range between 0 and 1. You need to decide the threshold probability at which the category flips from one to the other. By default, this is set to p = 0.5, but in reality, it should be settled based on the analysis purpose.

Is it Supervised/Unsupervised/Reinforcement learning?

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

What does the algorithm do?

The standard logistic regression function, for predicting the outcome of an observation given a predictor variable (x), is an s-shaped curve defined as p = exp(y) / [1 + exp(y)] (James et al. 2014). This can be also simply written as p = 1/[1 + exp(-y)], where:

* y = b0 + b1\*x,
* exp() is the exponential and
* p is the probability of event to occur (1) given x. Mathematically, this is written as p(event=1|x) and abbreviated asp(x), sopx = 1/[1 + exp(-(b0 + b1\*x))]`

By a bit of manipulation, it can be demonstrated that p/(1-p) = exp(b0 + b1\*x). By taking the logarithm of both sides, the formula becomes a linear combination of predictors: log[p/(1-p)] = b0 + b1\*x.

When you have multiple predictor variables, the logistic function looks like: log[p/(1-p)] = b0 + b1\*x1 + b2\*x2 + ... + bn\*xn

b0 and b1 are the regression beta coefficients. A positive b1 indicates that increasing x will be associated with increasing p. Conversely, a negative b1 indicates that increasing x will be associated with decreasing p.

The quantity log[p/(1-p)] is called the logarithm of the odd, also known as log-odd or logit.

The odds reflect the likelihood that the event will occur. It can be seen as the ratio of “successes” to “non-successes”. Technically, odds are the probability of an event divided by the probability that the event will not take place (P. Bruce and Bruce 2017). For example, if the probability of being diabetes-positive is 0.5, the probability of “won’t be” is 1-0.5 = 0.5, and the odds are 1.0.

Note that, the probability can be calculated from the odds as p = Odds/(1 + Odds).

In which situations will it be most useful?

Logistic regression is used to predict a discrete outcome based on variables which may be discrete, continuous or mixed. Thus, when the dependent variable has two or more discrete outcomes, logistic regression is a commonly used technique. The outcome could be in the form of Yes / No, 1 / 0, True / False, High/Low, given a set of independent variables.

Logistic regression has an array of applications. Here are a few applications used in real-world situations.

Marketing: A marketing consultant wants to predict if the subsidiary of his company will make profit, loss or just break even depending on the characteristic of the subsidiary operations.

Human Resources: The HR manager of a company wants to predict the absenteeism pattern of his employees based on their individual characteristic.

Finance: A bank wants to predict if his customers would default based on the previous transactions and history.