Logistic regression

In case of regression approach considered so far, the response variable Y has been regarded as a continuous quantitative variable. In many situations, such as mentioned below where the response variable is qualitative, the least squares driven linear/non-linear regression approaches do not work.

e.g. i) Selection of individuals on the basis of their scores in a series of tests. Since the objective is to check the ability of candidates via the skillfully designed tests to predict the job performance, the candidates are finally classified as 'good' or 'poor'; 'shortlisted' or 'not shortlisted'; 'selected' or 'rejected' etc.

ii) To determine risk factors of cancer, health records of several people are studied. Data were collected on several variables, such as age, gender, smoking, diet, family medical history etc. The outcome in terms of response variable was $Y=1$ for person had cancer or $Y=0$ for person didn't have cancer.

iii) In finance analysis, based on certain parameters i.e. financial characteristics data of various business firms was checked. The response variable Y is the solvency of the firm ($Y=1$ for a solvent firm, $Y=0$ for a bankrupt firm).

Such response variables are called 'dichotomous' variables, which are extensively used in statistical applications. Dichotomy - Process of being divided or split in two distinct (mostly with opposite natures) parts.

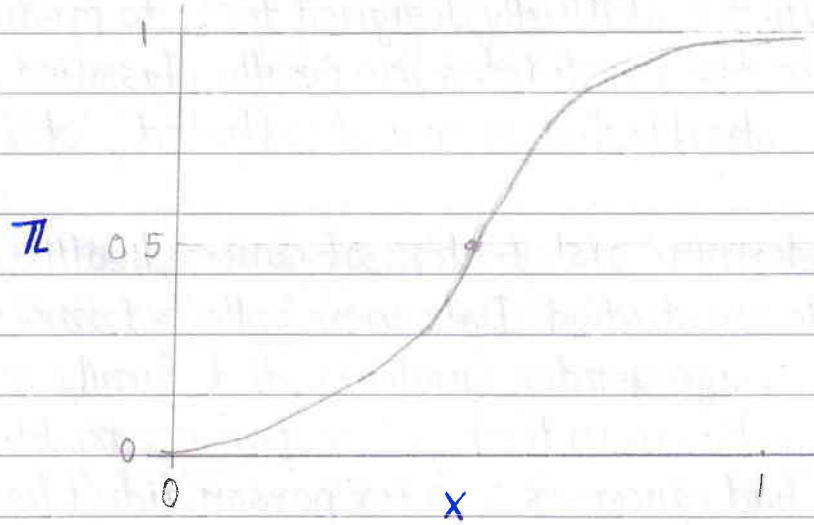
The qualitative data being considered is the binary response variable & can always be coded as having values 0 or 1. Rather than predicting these two values, we try to model the probabilities that the response takes as one of these two values.

Consider a simple regression problem in which we have a single independent variable. The same consideration holds good for the multiple regression case as well.

Let π denote the probability that $Y=1$ when $X=x$. If we use the standard linear model to describe π, our model for probability would be $\pi = P_r(Y=1 \mid X=x) = \beta_0 + \beta_1 x$.

Since π is a probability, it must lie between 0 to 1. The linear function given above is unbounded and hence can not be used to model probability.

The relationship between the probability π and X can often be represented by a logistic response function. It resembles an s-shaped curve and is also called an 'Ogive function'.



The probability π initially increases slowly with increase in X, then the increase accelerates, finally stabilizes but does not increase beyond 1.

The shape of the Ogive function can be reproduced if we model the probabilities as

$$\pi = P_r(Y=1 \mid X=x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad \ldots\ldots (a)$$

(a) is called logistic regression function. It is non-linear in the parameters β_0, β_1. However, it can be linearized by the logit transformation. Here, instead of working directly with π, we work with a transformed value of π.

from equation (a), we can state that

$$1-\pi = 1 - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{\beta_0+\beta_1 x}}$$

Then, $\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x}$

$\therefore \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x$. This is called the logit.

Thus, if π is the probability of an event happening, the ratio $\frac{\pi}{1-\pi}$ is called the 'odds ratio' for the event.

Thus the logarithm of the odds ratio is called the logit. The logit transformation produces a linear function of the parameters β_0, β_1. While the range of values of π is between 0 and 1, the range of values of logit is between $-\infty$ to $+\infty$, which makes it more appropriate for linear regression fitting.

Modeling the response probabilities by the logistic distribution and estimating the parameters of the model constitutes fitting a logistic regression. In logistic regression, the fitting is carried out by working with the logits. The logit transformation produces a model that is linear in the parameters. The method of estimation used is the maximum likelihood method.

For a binary classification, maximum likelihood will try to find values of β_0 and β_1 such that the resultant probabilities are closest to either 1 or 0.

We have logistic regression function (a) given as

$$\pi = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

Case study - A group of 20 customers, having yearly savings between 0 and 6 lakhs are considered. It is to investigate if there is any relationship between the savings of a customer & being a loan defaulter. 0 indicates loan defaulter & 1 indicates non-defaulter.

Amt. in savings	Loan Status	Amt. in Savings	Loan Status
0.5	0	2.75	1
0.75	0	3.00	0
1.00	0	3.25	1
1.25	0	3.50	0
1.5	0	4.00	1
1.75	0	4.25	1
1.75	1	4.5	1
2.00	0	4.75	1
2.25	1	5.00	1
2.5	0	5.5	1

The logistic regression analysis using the maximum likelihood estimate gives the output coefficients as $\beta_0 = -4.07778$, $\beta_1 = 1.5046$.

These coefficients are entered into the logistic regression equation to estimate the probability of being a loan non-defaulter.

$$\pi = \frac{1}{1 + e^{-(-4.07778 + 1.5046 * savings)}}$$

For example, for a customer with 2 lakhs savings per year,

$$\pi = \frac{1}{1 + e^{-(-4.07778 + 1.5046 * 2)}} = 0.2556$$ & so on.

The fitted value > 0.5 is taken as output 1, else output 0.

Amt. in Savings	Loan Status	Fitted Value	Prediction
0.5	0	0.0347	0
0.75	0	0.0497	0
1.00	0	0.0708	0
1.25	0	0.1000	0
1.5	0	0.1393	0
1.75	0	0.1908	0
1.75	1	~~0.2556~~ 0.1908	0
2	0	~~0.3335~~ 0.2556	0
2.25	1	~~0.4216~~ 0.3335	0
2.5	0	~~0.5149~~ 0.4216	0
2.75	1	0.5149	1
3.00	0	0.6073	1
3.25	1	0.6925	1
3.50	0	0.7664	1
4.00	1	0.8744	1
4.25	1	0.9102	1
4.5	1	0.9366	1
4.75	1	0.9556	1
5.00	1	0.9691	1
5.5	1	0.9851	1