LOGISTIC REGRESSION

- The general problem addressed by logistic regression is that of establishing relationship between certain explanatory variables (can be both numeric and categorical variables) with a categorical response variable.
- Logistic regression addresses the problem of classification. It is also used to estimate/assess risk.

DATA COLLECTION:

Scenario-1: In certain data collection frameworks, the explanatory variables related to a subject are observed at a point of time and the outcomes are observed later. In such a case the subjects being studied may have to be followed-up over a period of time. Such studies are called follow-up studies:

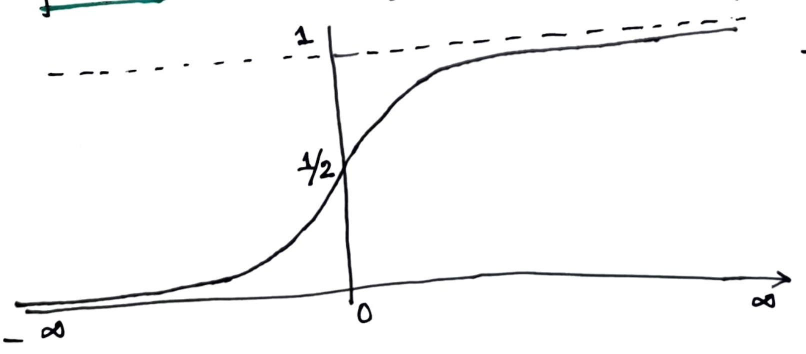
Example 1: We observe a set of people with certain lifestyle habits over a period of time. We then observe how many of these people have developed a particular disease.

Example 2: We observe a set of people who have been recruited. We note their characteristics and follow them up for a period of time to see how long they stay with the company (or how many of them leave within a given time frame).

Scenario-2: In other data collection formats we observe the outcomes of certain subjects. We then find the value of the explanatory variables pertaining to the subject.

CONCEPT OF LOGISTIC REGRESSION:

The function $f(x) = \dfrac{1}{1+e^{-z}}$, $z \in \mathbb{R}$ $(-\infty < z < \infty)$ is called the "logistic function". Note that $f(z)$ has the following graph:



- Note that $0 \leq f(z) \leq 1$.
- Note further that $f(z)$ has an S-shaped curve (often referred to as the sigmoidal curve).

USAGE OF SIGMOIDAL CURVE:

- The dosage of insecticide has an impact of killing insects. The probability is low when dosage is very small. From a threshold, the probability increases fast.
- The probability of a customer returning a loan may depend on factors like value of loan and level of disposable income. In this case, the variable Z may be considered to be a linear combination of these variables.

LOGISTIC MODEL: In general, the logistic model may be considered to be the following function:

$$Z = \beta_0 + \sum_{i=1}^{p} \beta_i X_i \; ; \text{ where } X_1, X_2, \ldots, X_p \text{ are the explanatory variables.}$$

In essense then Z is an index that combines the explanatory variables.

- Consider a binary classification problem with the explanatory variables as $X_1, X_2, \ldots, X_p$ and Y being the response variable.
- Suppose Y takes values 0 and 1.

  Then $P\left(Y=1 \mid X_1, X_2, \ldots, X_p\right) = \dfrac{1}{1+e^{-(\beta_0 + \sum_i \beta_i X_i)}}$.
- The coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the unknown parameters.

LOGIT TRANSFORMATION: $\quad \text{Logit}\left(P(\underset{\sim}{X})\right) = \ln\left(\dfrac{P(Y=1 \mid \underset{\sim}{X})}{1-P(Y=1 \mid \underset{\sim}{X})}\right)$.

Note that $P(Y=1 \mid \underset{\sim}{X}) = \dfrac{1}{1+e^{-(\beta_0 + \sum \beta_i X_i)}}$.

$$\Rightarrow 1 - P(Y=1 \mid \underset{\sim}{X}) = \frac{e^{-(\beta_0 + \sum \beta_i X_i)}}{1+e^{-(\beta_0 + \sum \beta_i X_i)}}.$$

$$\Rightarrow \ln\left(\frac{P(Y=1 \mid \underset{\sim}{X})}{1-P(Y=1 \mid \underset{\sim}{X})}\right) = \beta_0 + \sum_i \beta_i X_i$$

Note further that $\dfrac{P(Y=1 \mid \underset{\sim}{X})}{P(Y=0 \mid \underset{\sim}{X})}$ gives the odds of $P(Y=1)$ vs. $P(Y=0)$ for a given explanatory set up.

BASELINE ODDS: Note that β_0 gives the baseline odds. This refers to the odds that would result for a logistic model without any odds at all.

INTERPRETATION OF β_j: Suppose X_j is a variable measured in the ratio scale. Then

$$\ln\left(\text{odds}\left(Y=1 \mid X_1=x_1, X_2=x_2, \ldots, X_j=x_j, \ldots, X_p=x_p\right)\right) = \beta_0 + \sum \beta_i X_i .$$

$$\ln\left(\text{odds}\left(Y=1 \mid X_1=x_1, X_2=x_2, \ldots, X_j=x_j+1, \ldots, X_p=x_p\right)\right) = \beta_0 + \sum_{i=1}^{j-1} \beta_i x_i + \beta_j(x_j+1) + \sum_{i=j+1}^{p} \beta_i x_i$$

$$\Rightarrow \ln\left(\text{odds}\left(Y=1 \mid X_j=x_j+1\right)\right) - \ln\left(\text{odds}\left(Y=1 \mid X_j=x_j\right)\right) = \beta_j$$

$$\Rightarrow \frac{\text{odds}\left(Y=1 \mid X_j=x_j+1\right)}{\text{odds}\left(Y=1 \mid X_j=x_j\right)} = e^{\beta_j}$$

Thus, logistic regression model is one of 'constant odds ratio'.

MAXIMUM LIKELIHOOD ESTIMATES:

Note that $\pi(\underset{\sim}{x}_i) = P\left(Y=1 \mid X_1=x_{i1}, X_2=x_{i2}, \ldots, X_p=x_{ip}\right)$

$$= \frac{1}{1-e^{-\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}}$$

gives the probability that the response takes the value 1 for a given setting of explanatory variables. Likelihood function is:

$$l(\hat{\beta}) = \prod_{i=1}^{n} \pi(\underset{\sim}{x}_i)^{y_i}\left(1-\pi(\underset{\sim}{x}_i)\right)^{1-y_i}$$ follows directly from the Bernoulli PMF.

- Likelihood of the null model: $L_0 = \hat{p}^{\sum y_i}(1-\hat{p})^{\sum(1-y_i)}$, where $\hat{p}$ is the estimated proportion of the response variable taking value 1.

  Saturated model: $L_S = \prod y_i^{y_i}(1-y_i)^{(1-y_i)} = 1$

- Deviance: $D = -2\ln\left[\dfrac{\text{Likelihood of the fitted model}}{\text{Likelihood of the estimated model}}\right]$

- Likelihood Ratio (LR): $LR = -2\ln\left[\dfrac{\text{Likelihood of the fitted model}}{\text{Likelihood of the null model}}\right]$

Logit transformation: The transformation

$$g(\underset{\sim}{x}) = \ln\left(\frac{\pi(\underset{\sim}{x})}{1-\pi(\underset{\sim}{x})}\right); \text{ where } \pi(\underset{\sim}{x}) = P(Y=1 \mid X = \underset{\sim}{x})$$

$g(\underset{\sim}{x})$ has many desirable properties. The properties are given below:

(a) The logit $g(\underset{\sim}{x}) = \beta_0 + \sum \beta_i x_i$ are linear in its parameters.

(b) The logit $g(\underset{\sim}{x})$ is a continuous function.

(c) $-\infty < g(\underset{\sim}{x}) < \infty$.

Errors in Logistic Regression (Binary): We estimate Y by $\pi(\underset{\sim}{x}) = P(Y=1 \mid \underset{\sim}{x})$.

If $Y=1$ then $\varepsilon = 1-\pi(\underset{\sim}{x})$ with probability $\pi(\underset{\sim}{x})$.

If $Y=0$ then $\varepsilon = -\pi(\underset{\sim}{x})$ with probability $(1-\pi(\underset{\sim}{x}))$.

Then $E(\varepsilon) = \pi(\underset{\sim}{x})(1-\pi(\underset{\sim}{x})) - \pi(\underset{\sim}{x})(1-\pi(\underset{\sim}{x})) = 0$.

Note that each ε_i may be considered to be a Bernoulli trial.

The variance is not constant.

[Since, if $X \sim \text{Bernoulli}(p)$, $P(X=1)=p$, $P(X=0) = 1-p$;

$\Rightarrow E(X) = p$, $V(X) = E(p^2) - p^2 = p - p^2 = p(1-p)$.]

Evaluation of a screening test:

Let B = Risk event

B^c = Risk event does not happen

Also let T = Test result is positive

T^c = Test result is negative

- Prob$(T \mid B)$ is called sensitivity. This is the probability of the test showing positive result given that the risk event turns out to be true.

Examples:

i. Suppose on the basis of a logistic regression model, a transaction is classified to be fradulant. Sensitivity is the probability that the model identifies a transaction to be fraudulant when it actually is fraudulant.

ii. Similar logic is applicable when a model is used to classify a loan application.

- Prob$(T \mid \bar{B})$ is called specificity. This is the probability of a false alarm, i.e., the model identifies a transaction to be fraudulant when in reality it is NOT.

Goodness of Fit: Basic criteria for goodness-of-fit is that the distances between the observed and estimated values be unsystematic and within the variation of the model. This criteria is not satisfied in classification matrix.

Drawbacks of classification Table:

(a) Classification is sensitive to the relative size of the component groups and always favours classification into the larger group (i.e., probability of correctly classifying when a subject belongs to the larger group is high).

(b) The classification matrix converts a probability — an outcome measured on a continuum into a dichotomous variable leading to substantial loss of information.

(c) The sensitivity and specificity measured from a 2X2 classification table depends entirely on the distribution of the subjects rather than superiority of a model.

Example of a classification Table:

Consider the following hypothetical case:

Classification through model	Observed values		Total
	1	0	
1	16	11	27
0	131	417	548
Total	147	428	575

[letting disease: 1]

$$\text{Sensitivity} = \text{Prob}(\text{Predicted disease} \mid \text{Disease})$$
$$= \frac{16}{147}$$
$$= 10.90\%$$

$$\text{Specificity} = \text{Prob}(\text{Predicted disease-free} \mid \text{No disease})$$
$$= \frac{417}{428}$$
$$= 97.4\%$$

$$\text{Overall correct classification} = \frac{16+417}{575} = 0.753.$$

From the above table, the distribution of the subjects with disease probability >0.50 actually had about 40% of the subjects without disease. This implies that the estimated probabilities were >0.50 but sufficiently close to 0.5.

NOTE: Suppose among n subjects, the probability of a disease is a constant, say $\hat{\pi}$. Then $n\hat{\pi}$ subjects are expected to actually have the disease and $n(1-\hat{\pi})$ would not develop the disease. Thus, when $\hat{\pi} > 0.50$, $n(1-\hat{\pi})$ subjects are expected to be misclassified.

- In the last example, if we slightly modify the table as follows:

Classification	Observation		Total
	1	0	
1	26	1	27
0	27	521	548
Total	53	522	575

$$\text{Sensitivity} = \frac{26}{53} = 49.1\%$$

$$\text{Specificity} = \frac{521}{522} = 99\%$$

– Thus, the sensitivity and specificity depend heavily on the subject matrix.

- Another measure of "goodness-of-fit" for classification model is ROC-AUC.

AREA UNDER RECEIVER OPERATING CHARACTERISTIC CURVE:

Sensitivity = Pr (Model predicts disease | disease)

Specificity = Pr (Model predicts no disease | no disease)

1 − Specificity = Pr (Model predicts disease | no disease)

- Higher the sensitivity than (1 − specificity); better is the ability of the model to discriminate true positives and false positives.
- The ROC is the graph of sensitivity Vs. (1 − specificity) drawn over all possible cut points.
- When the ROC is on the diagonal line (area = 0.5) there is no discrimination.