Chapter 14
Logistic Regression Models

In the linear regression model $X\beta+\varepsilon$, there are two types of variables – explanatory variables $X_1, X_2,.., X_k$ and study variable y. These variables can be measured on a continuous scale as well as like an indicator variables. When the explanatory variables are qualitative, then their values are expressed as indicator variables and then dummy variable models are used.

When the study variable is qualitative variable, then its values can be expressed using an indicator variable taking only two possible values 0 and 1. In such a case, the logistic regression is used. For example, y can denotes the values like success or failure, yes or no, like or dislike which can be denoted by two values 0 and 1.

Consider the model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \\ &= x_i^{'}\beta + \varepsilon_i,\ i = 1, 2, ..., n \end{aligned}$$

where $x_i^{'} = \left[1, x_{i1}, x_{i2}, ..., x_{ik}\right],\ \beta' = \left[\beta_0, \beta_1, \beta_2, ..., \beta_k\right].$

The study variable takes two values as $y_i = 0$ or 1. Assume that y_i follows a Bernoulli distribution with parameter π_i, so its probability distribution is

$$y_i = \begin{cases} 1 & \text{with } P(y_i = 1) = \pi_i \\ 0 & \text{with } P(y_i = 0) = 1 - \pi_i. \end{cases}$$

Assuming $E(\varepsilon_i) = 0,$

$$E(y_i) = 1.\pi_i + 0.(1-\pi_i) = \pi_i.$$

From the model $y_i = x_i^{'}\beta + \varepsilon_i$, we have

$$\begin{aligned} & E(y_i) = x_i^{'}\beta \\ \Rightarrow\ & E(y_i) = x_i^{'}\beta = \pi_i \\ \Rightarrow\ & E(y_i) = P(y_i = 1). \end{aligned}$$

Thus response function $E(y_i)$ is simply the probability that $y_i = 1.$

Note that $\varepsilon_i = y_i - x_i'\beta$, so

- when $y_i = 1$, then $\varepsilon_i = 1 - x_i'\beta$
- $y_i = 0$, then $\varepsilon_i = -x_i'\beta$.

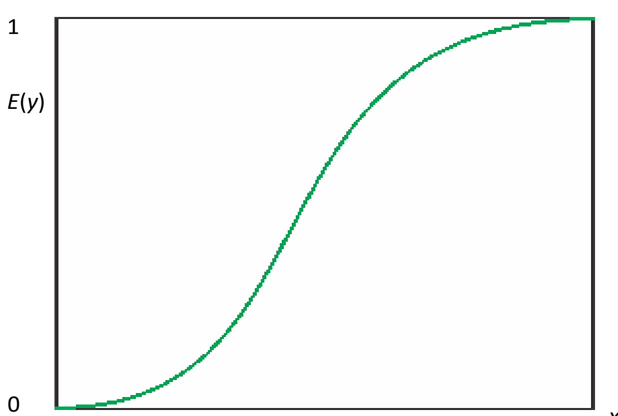
Recall that earlier ε_i was assumed to follow a normal distribution when y was not an indicator variable. When y is an indicator variable, then ε_i takes only two values, so it cannot be assumed to follow a normal distribution.

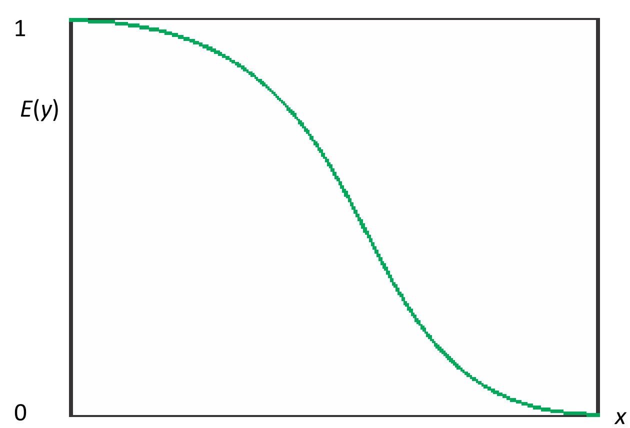
In usual regression model, the errors are homoskedastic, i.e., $Var(\varepsilon_i) = \sigma^2$ and so $Var(y_i) = \sigma^2$. When y is an indicator variable, then

$$\begin{aligned} Var(y_i) &= E\left[y_i - E(y_i)\right]^2 \\ &= (1-\pi_i)^2 \pi_i + (0-\pi_i)^2 (1-\pi_i) \\ &= \pi_i(1-\pi_i)\left[1-\pi_i+\pi_i\right] \\ &= \pi_i(1-\pi_i) \\ &= E(y_i)\left[1-E(y_i)\right] \\ &= \sigma^2_{y_i} . \end{aligned}$$

Thus $Var(y_i)$ depends on y_i and is a function mean of y_i. Moreover, since $E(y_i) = \pi_i$ and π_i is the probability, so $0 \le \pi_i \le 1$ and thus there is a constraint on $E(y_i)$ that $0 \le E(y_i) \le 1$. This puts a big constraint on the choice of linear response function. One cannot fit a model in which the predicted values lie outside the interval of 0 and 1.

When y is a dichotomous variable, then empirical evidences suggest that the function $E(y)$ on the whole real line that can be mapped to $[0,1]$ has the sigmoid shape. It is a nonlinear $S-$shape like





A natural choice for $E(y)$ would be the cumulative distribution function of a random variable. In particular, the logistic distribution, whose cumulative distribution function is the simplified logistic function yields a good link and is given by

$$
\begin{aligned}
E(y) &= \frac{\exp(y)}{1+\exp(y)} \\
&= \frac{\exp(x'\beta)}{1+\exp(x'\beta)} \\
&= \frac{1}{1+\exp(-x'\beta)}.
\end{aligned}
$$

Linear predictor and link functions:

The systematic component in $E(y)$ is the linear predictor and is denoted as

$$\eta_i = \sum_j \beta_j x_{ij} = x_i'\beta, \;\; i = 1,2,...,n, \;\; j = 0,1,2,...,k.$$

The link function in generalized linear model relates the linear predictor η_i to the mean response μ_i.

Thus

$$
\begin{aligned}
&g(\mu_i) = \eta_i \\
&\text{or } \mu_i = g^{-1}(\eta_i).
\end{aligned}
$$

In the usual linear models based on the normally distributed study variable, the link $g(\mu_i) = \mu_i$ is used and is called as **identity link**. A link function maps the range of μ_i onto the whole real line, provides good empirical approximation and carries meaningful interpretations in real applications.

In case of logistic regression, the link function is defined as

$$\eta = \ln \frac{\pi}{1-\pi}.$$

This transformation is called as the **logit** transformation of probability π and $\dfrac{\pi}{1-\pi}$ is called as **odds**. The link η is also called as **log-odds**. This link function is obtained as follows:

$$\pi = \frac{1}{1+\exp(-\eta)}$$

or $\pi\left[1+\exp(-\eta)\right]=1$

or $e^{-\eta} = \frac{1-\pi}{\pi}$

or $\mu = \ln\frac{\pi}{1-\pi}$.

Note: Similar to logit function, there are other functions also which have same shape as of logistic function. These functions can also be transformed through π. There are two such popular functions – probit transformation and complementary log-log transformation. The probit transformation is based on the transformation of π using the cumulative distribution function of normal distribution and based on this is the **probit regression model.**

The **complementary log-log transformation of** π is $\ln\left[-\ln(1-\pi)\right]$.

Maximum likelihood estimation of parameters:

Consider the general form of the logistic regression model

$$y_i = E(y_i) + \varepsilon_i$$

where y_i's are independent Bernoulli random variable with parameter π_i with

$$\begin{aligned} E(y_i) &= \pi_i \\ &= \frac{\exp(x_i^{'}\beta)}{1+\exp(x_i^{i}\beta)} \end{aligned}$$

The probability density function of y_i is

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, i=1,2,...,n, \ y_i = 0 \text{ or } 1.$$

The likelihood function is

$$\begin{aligned} L(y_1, y_2, ..., y_n, \beta_1, \beta_2, ..., \beta_k) = L &= \prod_{i=1}^{n} f_i(y_i) \\ &= \prod_{i=1}^{n} f_i(y_i)(1-\pi_i)^{1-y_i} \end{aligned}$$

$$\ln L = \sum_{i=1}^{n}\left[\ln \pi_i^{y_i} + \ln(1-\pi_i)^{1-y_i}\right]$$
$$= \sum_{i=1}^{n}\left[y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)\right]$$
$$= \sum_{i=1}^{n}\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n}\left[\ln(1-\pi_i)\right].$$

Since

$$\pi_i = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)},$$
$$1-\pi_i = \frac{1}{1+\exp(x_i'\beta)},$$
$$\frac{\pi_i}{1-\pi_i} = \exp(x_i'\beta),$$
$$\ln\frac{\pi_i}{1-\pi_i} = \exp x_i'\beta \quad ,$$

so

$$\ln L = \sum_{i=1}^{n} y_i x_i'\beta - \sum_{i=1}^{n}\ln\left[1+\exp(x_i'\beta)\right].$$

Suppose repeated observations are available at each level of the x-variables. Let y_i be the numbers of 1's observed for i^{th} observation and n_i be the number of trials at each observation. Then

$$\ln L = \sum_{i=1}^{n} y_i\pi_i + \sum_{i=1}^{n} n_i \ln(1-\pi_i) - \sum_{i=1}^{n} y_i \ln(1-\pi_i).$$

The maximum likelihood estimate $\hat{\beta}$ of β is obtained by the numerical maximization.

If $V(\varepsilon) = \Omega,$ then asymptotically

$$E(\hat{\beta}) = \beta$$
$$V(\hat{\beta}) = (X'\Omega^{-1}X)^{-1}.$$

After obtaining $\hat{\beta}$, the linear predictor is estimated by

$$\hat{\eta}_i = x_i'\beta.$$

The fitted value is

$$\hat{y}_i = \hat{\pi}_i = \frac{\exp(\hat{\eta}_i)}{1+\exp(\hat{\eta}_i)} = \frac{1}{1+\exp(-\hat{\eta}_i)} = \frac{1}{1+\exp(-x_i'\hat{\beta})}.$$

Interpretation of parameters:

To understand the interpretation of the related β's in the logistic regression model, first consider a simple case with only one variable as

$$\eta(x) = \beta_0 + \beta_1 x.$$

After fitting of model, $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained as the estimators of β_0 and β_1 respectively. Then the fitted linear predictor at $x = x_i$ is

$$\hat{\eta}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

which is the log-odds at $x = x_i$. The fitted value at $x = x_i + 1$ is

$$\hat{\eta}(x_i + 1) = \hat{\beta}_0 + \hat{\beta}_1 (x_i + 1)$$

which is the log-odds at $x = x_i + 1.$

Thus

$$\begin{aligned}\hat{\beta}_1 &= \hat{\eta}(x_i + 1) - \hat{\eta}(x_i)\\ &= \ln\left[\text{odds}(x_i + 1)\right] - \ln\left[\text{odds}(x_i)\right]\\ &= \ln\left[\frac{\text{odds}(x_i + 1)}{\text{odds}(x_i)}\right]\\ \Rightarrow \frac{\text{odds}(x_i + 1)}{\text{odds}(x_i)} &= \exp(\hat{\beta}_1).\end{aligned}$$

This is termed as **odd ratio** which is the estimated increase in the probability of success when value of explanatory variable changes by one unit.

When there are more than one explanatory variables in the model, then the interpretation of β_j's is similar as in the case of single explanatory variable case. The odds ratio is $\exp(\hat{\beta}_j)$ associated with explanatory variable x_j keeping other explanatory variables constant. This is similar to the interpretation of β_j in multiple linear regression model.

If there is a m unit change is the explanatory variable, then the estimated increase in odds ratio is $\exp(m\hat{\beta}_j)$.

Test of hypothesis:

The test of hypothesis for the parameters in the logistic regression model is based on asymptotic theory. It is a large sample test based on likelihood ratio test based on a statistic termed as **deviance.**

A model with exactly p parameters that perfectly fits to the sample data is termed as **saturated model.**

The statistic that compares the log-likelihoods of fitted and saturated models is called as **model deviance**. It is defined as

$$\lambda(\beta) = 2\ln L(\text{saturated model}) - 2\ln L(\hat{\beta})$$

where $\ln L(\cdot)$ is the log-likelihood and $\hat{\beta}$ is the maximum likelihood estimate of β.

In case of logistic regression model, $y_i = 0$ or 1 and π_i's are completely unrestricted. So the likelihood will be maximum at $\pi_i = y_i$ and the maximum value of L (saturated modal) is

$$\begin{aligned} &Maximum\ L(\text{saturated model}) = 1 \\ \Rightarrow &\ln Maximum\, L(\text{saturated model}) = 0. \end{aligned}$$

Let $\hat{\beta}$ be the maximum likelihood estimator of β, then log-likelihood is maximum at $\beta = \hat{\beta}$, and

$$\begin{aligned} \ln L(\hat{\beta}) &= \sum_{i=1}^{n} y_i x_i' \hat{\beta}_i - \sum_{i=1}^{n} \ln\left[1 + \exp(x_i'\beta)\right] \\ &\geq \ln L(\text{saturated model}). \end{aligned}$$

Assuming that the logistic regression function is correct, the large sample distribution of likelihood ratio test statistic $\lambda(\beta)$ is approximately distributed as $\chi^2(n-p)$, when n is large.

Large value of $\lambda(\beta)$ implies model is incorrect. Small value of $\lambda(\beta)$ implies that model is well fitted and is as good as the saturated model. Note that generally the fitted model will be having smaller number of parameters than the saturated model that is based on all the parameters. Thus at $\alpha\%$ level of significance.