LOGISTIC REGRESSION

Ramasubramanian V.
I.A.S.R.I., Library Avenue, Pusa New Delhi - 110 012
ramsub@iasri.res.in

1. Introduction

Regression analysis is a method for investigating functional relationships among variables. The relationship is expressed in the form of an equation or a model connecting the response or dependent variable and one or more explanatory or predictor variables. Most of the variables in this model are quantitative in nature. Estimation of parameters in this regression model is based on four basic assumptions. First, response or dependent variable is linearly related with explanatory variables. Second, model errors are independently and identically distributed as normal variable with mean zero and common variance. Third, independent or explanatory variables are measured without errors. The last assumption is about equal reliability of observations.

In case, our response variable in model is qualitative in nature, then probabilities of falling this response variable in various categories can be modeled in place of response variable itself, using same model but there are number of constraints in terms of assumptions of multiple regression model. First, since the range of probability is between 0 and 1, whereas, right hand side function in case of multiple regression models is unbounded. Second, error term of the model can take only limited values and error variance are not constants but depends on probability of falling response variable in a particular category.

Generally, conventional theory of multiple linear regression (MLR) analysis has been applied for a quantitative response variable, while for the qualitative response variable or more specifically for binary response variable it is better to consider alternative models. As for example, considering following scenarios:

- A pathologist may be interested whether the probability of a particular disease can be predicted using tillage practice, soil texture, date of sowing, weather variables etc. as predictor or independent variables.
- An economist may be interested in determining the probability that an agro-based industry will fail given a number of financial ratios and the size of the firm (i.e. large or small).

Usually discriminant analysis could be used for addressing each of the above problems. However, because the independent variables are mixture of categorical and continuous variables, the multivariate normality assumption may not hold. Structural relationship among various qualitative variables in the population can be quantified using number of alternative techniques. In these techniques, primary interest lies on dependent factor which is dependent on other independent factors. In these cases the most preferable technique is either probit or logistic regression analysis as it does not make any assumptions about the distribution of the independent variables. The dependent factor is known as response factor. In this model building process, various log odds related to response factors are modelled. As a special case, if response factor has only two categories with probabilities p_1 and p_2 respectively then the odds of getting category one is (p_1 / p_2). If log (p_1 / p_2) is modelled using ANalysis Of VAriance (ANOVA) type of model, it is called logit model. Again, if the same model is being treated as regression type model then it is called logistic regression model. In a real sense, logit and logistic are names of transformations. In case of logit transformation, a number p between values 0 and 1 is transformed with log {p/(1-p)}, whereas in case of logistic transformation a number x between - ∞ to + ∞ is transformed with $\{e^x /(1 + e^x)\}$ function. It can be seen that these two transformation are reverse of each other i.e. if logit transformation is applied on logistic transformation function, it provides value x and similarly, if logistic transformation is applied to logit transformation function it provides value p. Apart from logit or logistic regression models, other techniques such as CART i.e. Classification and Regression Trees can also be used to address such classification problems. A good account of literature on logistic regression are available, to cite a few, Fox(1984), Klienbaum (1994) etc.

2. Violation of Assumptions of Linear Regression Model when Response is Qualitative

Linear regression is considered in order to explain the constraints in using such model when the response variable is qualitative. Consider the following simple linear regression model with single predictor variable and a binary response variable:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \ , i = 1, 2, \ldots, n$$

where the outcome Y_i is binary (taking values 0,1), $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and are independent and n is the number of observations.

Let π_i denote the probability that Y_i =1 when X_i = x, i.e.

$$\pi_i = P(Y_i = 1 | X_i = x) = P(Y_i = 1)$$

thus $P(Y_i = 0) = 1 - \pi_i$.

Under the assumption $E(\varepsilon_i) = 0$, the expected value of the response variable is

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

If the response is binary, then the error terms can take on two values, namely,

$$\varepsilon_i = 1 - \pi_i \quad \text{when } Y_i = 1$$

$$\varepsilon_i = -\pi_i \quad \text{when } Y_i = 0$$

Because the error is dichotomous (discrete), normality assumption is violated. Moreover, the error variance is given by:

$$V(\varepsilon_i) = \pi_i(1-\pi_i)^2 + (1-\pi_i)(-\pi_i)^2$$
$$= \pi_i(1-\pi_i)$$

It can be seen that variance is a function of π_i's and it is not constant. Therefore the assumption of homoscadasticity (equal variance) does not hold.

3. Binary Logistic regression

Logistic regression is normally recommended when the independent variables do not satisfy the multivariate normality assumption and at the same time the response variable is qualitative. Situations where the response variable is qualitative and independent variables are mixture of categorical and continuous variables, are quite common and occur extensively in statistical applications in agriculture, medical science etc. The statistical model preferred for the analysis of such binary (dichotomous) responses is the binary logistic regression model, developed primarily by a researcher named Cox during the late 1950s. Processes producing sigmoidal or elongated S-shaped curves are quite common in agricultural data.

Logistic regression models are more appropriate when response variable is qualitative and a non-linear relationship can be established between the response variable and the qualitative and quantitative factors affecting it. It addresses the same questions that discriminant function analysis and multiple regression do but with no distributional assumptions on the predictors. In logistic regression model, the predictors need not have to be normally distributed, the relationship between response and predictors need not be linear or the observations need not have equal variance in each group etc. A good account on logistic regression can be found in Fox (1984) and Kleinbaum (1994).

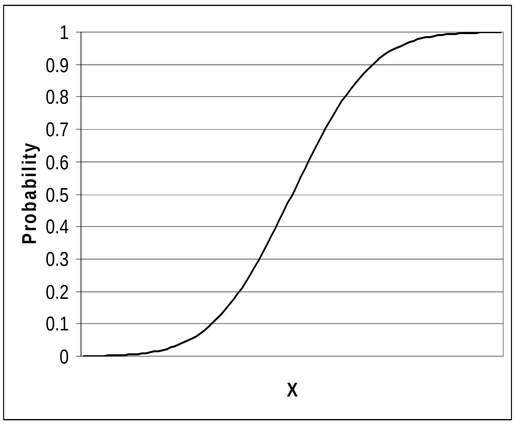
The problem of non-normality and heteroscadasticity (see section 2) leads to the non applicability of least square estimation for the linear probability model. Weighted least square estimation, when used as an alternative, can cause the fitted values not constrained to the interval (0, 1) and therefore cannot be interpreted as probabilities. Moreover, some of the error variance may come out to be negative. One solution to this problem is simply to constrain π to the unit interval while retaining the linear relation between π and regressor X within the interval. Thus

$$\pi = \begin{cases} 0 & ,\beta_0 + \beta_1 X < 0 \\ \beta_0 + \beta_1 X & ,0 \le \beta_0 + \beta_1 X \le 1 \\ 1 & ,\beta_0 + \beta_1 X > 1 \end{cases}$$

However, this constrained linear probability model has certain unattractive features such as abrupt changes in slope at the extremes 0 and 1 making it hard for fitting the same on data. A smoother relation between π and X is generally more sensible. To correct this problem, a positive monotone (i.e. non-decreasing) function is required to transform ($\beta_0 + \beta_1 x_i$) to unit interval. Any cumulative probability distribution function (CDF) P, meets this requirement. That is, respecify the model as $\pi i = P(\beta_0 + \beta_1 x_i)$. Moreover, it is advantageous if P is strictly increasing, for then, the transformation is one-to-one, so that model can be rewritten as $P^{-1}(\pi i) = (\beta 0 + \beta 1 xi)$, where P^{-1} is the inverse of the CDF P. Thus the non-linear model for itself will become both smooth and symmetric, approaching π = 0 and π = 1 as asymptotes. Thereafter maximum likelihood method of estimation can be employed for model fitting.

3.1 Properties of Logistic Regression Model

The Logistic response function resembles an S-shape curve, a sketch of which is given in the following figure. Here the probability π initially increases slowly with increase in X, and then the increase accelerates, finally stabilizes, but does not increase beyond 1.



The shape of the S-curve can be reproduced if the probabilities can be modeled with only one predictor variable as follows:

$$\pi = P(Y=1|X=x) = 1/(1+e^{-Z})$$

where $z = \beta_0 + \beta_1 x$, and e is the base of the natural logarithm. Thus for more than one (say r) explanatory variables, the probability π is modeled as

$$\pi = P(Y=1|X_1 = x_1 ... X_r = x_r)$$
$$=1/(1+e^{-Z})$$

where $z = \beta_0 + \beta_1 x_1 + ... + \beta_r x_r$.

This equation is called the logistic regression equation. It is nonlinear in the parameters β_0, $\beta_1 \ldots \beta_r$. Modeling the response probabilities by the logistic distribution and estimating the parameters of the model constitutes fitting a logistic regression. The method of estimation generally used is the maximum likelihood estimation method.

To explain the popularity of logistic regression, let us consider the mathematical form on which the logistic model is based. This function, called f (z), is given by

$$f(z) = 1/(1+e^{-z}), -\infty < z < \infty$$

Now when $z = -\infty$, $f(z) = 0$ and when $z = \infty$, $f(z) = 1$. Thus the range of f (z) is 0 to1. So the logistic model is popular because the logistic function, on which the model is based, provides

- Estimates that lie in the range between zero and one.
- An appealing S-shaped description of the combined effect of several explanatory variables on the probability of an event.

3.2. Maximum Likelihood Method of Estimation of Logistic Regression

For simplicity, a simple binary logistic regression model with only one explanatory variable is considered. The model is given by

$$\pi_i = P(Y_i=1|X_i = x_i) = 1/(1+e^{-Z})$$

where $z = \beta_0 + \beta_1 x_i$, and e is the base of the natural logarithm. The binary response variable Y_i takes only two values (say 0 and 1). Since each Y_i observation is an ordinary Bernoulli random variable, where:

$$P(Y_i = 1) = \pi_i$$

and $$P(Y_i = 1) = 1 - \pi_i,$$

the probability distribution function is represented as follows:

$$f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \quad, Y_i = 0,1; i=1,2,...,n$$

Since Y_i's are independent, then the joint probability density function is:

$$g(Y_1...Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$\log_e g(Y_1...Y_n) = \log_e \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$= \sum_{i=1}^{n} Y_i \log_e \pi_i/(1-\pi_i)$$

Since E (Y_i) = π_i, for a binary variable it follows that

$$1\text{-}\pi_i = \left[1+e^{-(\beta_0+\beta_1 X_i)}\right]^{-1}$$

Then,

$$\log_e\left[\pi_i/(1\text{-}\pi_i)\right]=\beta_0+\beta_1 X_i$$

Hence the log likelihood function can be expressed as follows:

$$\log_e L(\beta_0,\beta_1)=\sum_{i=1}^{n} Y_i(\beta_0+\beta_1 X_i)\text{-}\sum_{i=1}^{n}\log_e\left[1+e^{-(\beta_0+\beta_1 X_i)}\right]$$

where L (β_0, β_1) replaces g (Y_1… Y_n) to show explicitly that the function can now be viewed as the likelihood function of the parameters to be estimated, given the sample observations.

The maximum likelihood estimates β_0 and β_1 in the simple logistic regression model are those values of β_0 and β_1 that maximize the log-likelihood function. No closed-form solution exists for the values of β_0 and β_1 that maximize the log-likelihood function. Computer intensive numerical search procedures are therefore required to find the maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Standard statistical software programs such as SAS (PROC LOGISTIC), SPSS (Analyze- Regression-Binary Logistic) provide maximum likelihood estimates for logistic regression. Once these estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are found, by substituting these values into the response function the fitted response function, say, $\hat{\pi}_i$, can be obtained. The fitted response function is as follows:

$$\hat{\pi}_i=\left(\frac{1}{1+e^{-\left(\hat{\beta}_0+\hat{\beta}_1 X_i\right)}}\right)$$

When log of the odds of occurrence of any event is considered using a logistic regression model, it becomes a case of logit analysis. Here the thus formed logit model will have its right hand side as a linear regression equation.

4. Model Validation

The model validation can be done by employing various tests on any fitted logistic regression model. The tests related to the significance of the estimated parameters, goodness of fit and predictive ability of the models are discussed subsequently.

4.1 Testing the overall significance of model

Wald, Likelihood ratio and Score tests are three commonly used tests for testing the overall significance of the logistic regression model.

4.1.1 Wald test

Let $\hat{\boldsymbol{\beta}}$ be the vector of parameter estimates obtained. Let a set of restrictions be imposed in the form of a hypothesis H_0: $\beta = 0$. If the restrictions are valid, then at least approximately $\hat{\boldsymbol{\beta}}$ should satisfy them. The Wald statistic is then defined as $W=\hat{\boldsymbol{\beta}}'\left[\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\hat{\boldsymbol{\beta}}$

Under H_0, in large samples, W has a Chi-square distribution with degrees of freedom equal to the number of restrictions imposed.

4.1.2 Likelihood Ratio (LR) Test

The LR statistic is defined as two times the logarithm of the ratio of the likelihood functions of two different models evaluated at their MLEs. The LR statistic is used for testing the overall significance of the model. Assuming that there are r_1 variables in the model under consideration which can be considered as the full model, based on the MLEs of the full model, L (full) is calculated. Beside this, the likelihood function L (reduced) is calculated for the constant only model. The LR statistic is then defined as: $\mathrm{LR} = -2\left[\ln\left\{\mathrm{L}\,(\text{reduced})\right\} - \ln\left\{\mathrm{L}\,(\text{full})\right\}\right]$. LR is asymptotically distributed as Chi-square with degrees of freedom equal to the difference between the number of parameters estimated in the two models.

4.1.3 Goodness of Fit in Logistic Regression

Among various testing problems, goodness of fit is one of the most important aspects in the context of the logistic regression analysis for testing whether the model fitted well or not. Hosmer-Lemeshow goodness-of-fit test is one of the most common tool conveniently used in logistic regression analysis. This test is performed for a binary logistic regression model by first sorting the observations in increasing order of their estimated event probabilities. The observations are then divided into approximately ten groups on the basis of the estimated probabilities. Comparison between the numbers actually in each group (observed) to the numbers predicted by the logistic regression model (predicted) is carried out subsequently. The number of groups may be smaller than 10 if there are fewer than 10 patterns of explanatory variables. There must be at least three groups in order that the Hosmer-Lemeshow statistic can be computed.

The Hosmer-Lemeshow goodness-of-fit statistic is obtained by calculating the Pearson chi-square statistic from the (2×g) table of observed and expected frequencies where g is the number of groups.

The statistic is written as:

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{\left(O_i - N_i\bar{\pi}_i\right)^2}{N_i\bar{\pi}_i\left(1-\bar{\pi}_i\right)} \sim \chi^2_{g-2}$$

where

O_{i} = the observed number of events in the ith group

$\mathrm{N_i}$ = the number of subjects in ith group

and $\bar{\pi}_i$ = the average estimated probability of an event in the ith group.

4.1.4 Predictive ability of the model

Once models are fitted and relevant goodness of fit measures are employed, judging the predictive ability of the model can be done. In logistic regression modeling setup, predictive ability of models can be judged by employing various measures such as Somers'D, Gamma, Kendall's Tau (Tau-a) and c. Here two measures viz. Gamma and Somers'D have been discussed. Gamma statistic is the simplest one. The measures Gamma, and Somers'D are based on concordance and discordance. By observing the ordering of two subjects on each of two variables, one can classify the pair of subjects as concordant or discordant. The pair is concordant if the subject ranking is higher on both the variables. The pair is discordant if the subject ranking is higher on one variable and lower on the other. The pair is tied if the subjects have the same prediction on both of the variables. The Gamma is defined as

$$\frac{\mathrm{N_s} - \mathrm{N_d}}{\mathrm{N_s} + \mathrm{N_d}}$$

where $\mathrm{N_s}$ is the number of same pairs and $\mathrm{N_d}$ the number of different pairs. Gamma ignores all tied pairs of cases. It therefore may exaggerate the "actual" strength of association. Gamma lies between -1 to1.

The Somers'D is a simple modification of gamma. Unlike gamma, the Somers' D includes tied pairs in one way or another. Somers'D is defined as

$$\frac{\mathrm{N_s} - \mathrm{N_d}}{\mathrm{N_s} + \mathrm{N_d} + \mathrm{T_y}}$$

where T_y is the number of pairs tied on the dependent variable, Y. Somers' d ranges from -1.0 (for negative relationships) to 1.0.

4.1.5 Classificatory ability of the models

Comparison between various logistic regression models fitted and with other classification methods such as discriminant function and decision tree methods can be made with respect to their classifying ability with the help of (2 x 2) classification tables in case of a binary response group variable. The columns are the two observed values of the dependent variable, while the rows are the two predicted values of the dependent. In a perfect model, all cases will be on the diagonal and the overall percent correct will be 100%.

Critical terms associated with classification table are as follows:

Hit rate: Number of correct predictions divided by sample size. The hit rate for the model should be compared to the hit rate for the classification table for the constant-only model.
Sensitivity: Percent of correct predictions in the reference category (usually 1) of the dependent. It also refers to the ability of the model to classify an event correctly.
Specificity: Percent of correct predictions in the given category (usually 0) of the dependent. It also refers to ability of the model to classify a non event correctly.
False positive rate: It is the proportion of predicted event responses that were observed as nonevents
False negative rate: It is the proportion of predicted nonevent responses that were observed as events.

Higher the sensitivity and specificity lower the false positive rate and false negative rate, better the classificatory ability.

5. Association between attributes/ variables

An association exists between two variables if the distribution of one variable changes when the level (or value) of the other variable changes. If there is no association, the distribution of the first variable is the same regardless of the level of other variable. Odds ratio is usually used for measuring such associations. For example, consider the following table having two attributes ‘Weather’ and ‘Mood of boss’ each at two levels.

	Mood of boss	
	Good	Bad
Weather		
Rain	82	18
Shine	60	40

Odds of an event is the ratio of the probability of an event occurring to the probability of it not occurring. That is,

$$\text{Odds}=P(\text{event})/\{1-P(\text{event})\} = P(\text{event}=1)/P(\text{event}=0)$$

In the above table, there is 82% probability that the mood of the boss will be ‘Good’ in case of ‘Rain’. The odds of ‘Good mood’ in ‘Rain’ category =0.82/0.18 =4.5. The odds of ‘Good mood' in 'Shine' category =0.60/0.40 =1.5. The odds ratio of 'Rain' to 'Shine' equals (4.5/1.5) =3 indicating that the odds of getting 'Boss in good mood' during 'Rain' is three times those during 'Shine'. Also there is 18% probability that mood of boss will be 'Bad' in case of 'Rain'; the odds of 'Bad mood' in 'Rain' =0.18/0.82 =0.22. Thus, in case the probability is very small (0.18 in this case), there is no appreciable difference in mentioning the same as probability or odds.

The importance of odds ratio is case of logistic regression modeling can be further explained by taking a simple case of influence of an attribute "Gender" X with two levels (Male or Female) on another attribute "opinion towards legalized abortion" Y with two levels (Yes=1, No=0). Logistic regression when written in its linearised form takes the following 'logit' form:

$$logit\ \{Y=1|\ X=x\} = log\ (\pi/\ (1-\pi))=log\ (odds) = \beta 0+\beta 1*x$$

Now,
Odds (Females)= exp($\beta 0+\beta 1$) and Odds (Males)= exp($\beta 0$). Hence

Odds ratio = exp($\beta 0+\beta 1$)/exp($\beta 0$) = $exp(\beta 1)$.
Thus here regression coefficient of Y on X i.e. $\beta 1$ is not directly interpreted but after taking exponentiation of it.

6. Multinomial logistic regression modeling
Let **X** is a vector of explanatory variables and π denotes the probability of binary response variable then logistic model is given by

$$\log it(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \mathbf{X}\beta = g(\pi)$$

where, 'alpha' is the intercept parameter and 'beta' is a vector of slope parameters. In case response variable has ordinal categories say 1,2,3,--------, I, I+1 then generally logistic model is fitted with common slope based on cumulative probabilities of response categories instead of individual probabilities. This provides parallel lines of regression model with following form

g [Prob ($\mathbf{y} \le \mathbf{i}(\mathbf{x})$)] $=\alpha_i + x\beta\ ,\ 1\le i \le I$

where, $\alpha_1, \alpha_2, ------ \alpha_{\mathbf{k}}$, are k intercept parameters and β is the vector of slope parameters.

Multinomial logistic regression (taking qualitative response variable with three categories, for simplicity) is given by

$$\text{logit}[\Pr(Y \le j - 1 / \mathbf{X})] = \alpha_j + \boldsymbol{\beta}^T \mathbf{X}, \quad j = 1,2$$

where α_j are two intercept parameters ($\alpha_1 < \alpha_2$), $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \ldots\ldots, \beta_k)$ is the slope parameter vector not including the intercept terms, $\mathbf{X}^T = (X_1, X_2, \ldots., X_k)$ is vector of explanatory variables. This model fits a common slope cumulative model i.e. 'parallel lines' regression model based on the cumulative probabilities of the response categories.

$$\text{logit}(\pi_1) = \log\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots.. + \beta_k X_k,$$

$$\text{logit}(\pi_1 + \pi_2) = \log\left(\frac{\pi_1 + \pi_2}{1-\pi_1-\pi_2}\right) = \alpha_2 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots.. + \beta_k X_k$$

where

\

$$\pi_1(X) = \frac{e^{\alpha_1+\beta^T X}}{1+e^{\alpha_1+\beta^T X}}$$

$$\pi_1(X) + \pi_2(X) = \frac{e^{\alpha_2+\beta^T X}}{1+e^{\alpha_2+\beta^T X}}$$

$$\pi_1 + \pi_2 + \pi_3 = 1$$

π_j (X) denotes classification probabilities Pr(Y=j-1 / X) of response variable Y, j = 1,2,3, at X^T.
These models can be fitted through maximum likelihood procedure.

7. Application of binary logit models in agriculture and other sciences

Sometimes quantitative information on pests and diseases is not available but is available in qualitative form such as occurrence / non-occurrence, low / high incidence etc. The statistical model preferred for the analysis of such binary (dichotomous) responses is the binary logistic regression model. It can be used to describe the relationship of several independent variables to the binary (say, named 0 & 1) dependent variable. The logistic regression is used for obtaining probabilities of occurrence, say E, of the different categories when the model is of the form: P(E =1) = $\frac{1}{1+\exp(-z)}$ where z is a function of associated variables, if P(E =1) ≥ 0.5 then there is more chance of occurrence of an event and if P(E =1) < 0.5 then probability of occurrence of the event is minimum. If the experimenter wants to be more stringent, then the cutoff value of 0.5 could be increased to, say, 0.7.

Consider the dataset given in the Table given below. Weather data during 1987-97 in Kakori and Malihabad mango (*Mangifera indica* L.) belt (Lucknow) of Uttar Pradesh is used here to develop logistic regression models for forewarning powdery mildew caused by *Oidium mangiferae* Berthet and validated the same using data of recent years. The forewarning system thus obtained satisfactorily forewarns with the results obtained comparing well with the observed year-wise responses. The status of the powdery mildew (its epidemic and spread) during 1987-97 are given in the following table, with the occurrence of the epidemic denoted by 1 and 0 otherwise. The variables used were maximum temperature (X_1) and relative humidity (X_2). The model is given by

$$P\,(Y{=}1) = 1/\,[1{+}\exp\,\{-\,(\beta_0 + \beta_1\, x_1 + \beta_2\, x_2)\}]$$

Table: Epidemic status (Y) of powdery mildew fungal disease in Mango in U.P.

Year	Third week(Y) of March	Average weather data in second week of March	
		X_1	X_2
1987	1	30.14	82.86
1988	0	30.66	79.57
1989	0	26.31	89.14
1990	1	28.43	91.00
1991	0	29.57	80.57
1992	1	31.25	67.82
1993	0	30.35	61.76
1994	1	30.71	81.14
1995	0	30.71	61.57
1996	1	33.07	59.76
1997	1	31.50	68.29

Logistic regression models were developed using the maximum likelihood estimation procedure in SAS. Consider 1987-96 model based on second week of March average weather data using which forewarning probability is obtained for the year 1997. The parameter estimates corresponding to intercept, X1 and X2 are obtained as $\hat{\beta}_0$ = -72.47; $\hat{\beta}_1$ = 1.845; $\hat{\beta}_2$ = 0.22.

Then the model becomes

$$P\ (Y=1) = 1/\{1+ \exp\ (-(-72.47+ (1.845* x_1) + (\ 0.22* x_2))\}$$

Plugging in the values X_1 = 31.50 and X_2 = 68.29, of year 1997 it can be seen that P(Y=1) = 0.66. This is the forewarning probability of occurrence of powdery mildew in mango using logistic regression modeling for 1997. The logistic regression model yielded good results. If P (Y=1) <0.5, then probability that epidemic will occur is minimal, otherwise there is more chance of occurrence of epidemic and this can be taken as objective procedure of forewarning the disease. As we were having the information that there was epidemic during the year 1997, it can be seen that the logistic regression model forewarns the actual status correctly.

The SAS syntax is given as follows:

```
data PMildew;
  input epidemic MaxT RH;
  cards;
1       30.14  82.86
0       30.66  79.57
0       26.31  89.14
1       28.43  91.00
0       29.57  80.57
1       31.25  67.82
0       30.35  61.76
1       30.71  81.14
0       30.71  61.57
1       33.07  59.76
```

```
;
 proc logistic data=PMildew;
        model epidemic(event='1') = MaxT RH / lackfit;
  run;
```

The abridged SAS Output is given as follows:

```
:
    Probability modeled is epidemic=1.
                                              The LOGISTIC Procedure

                                Analysis of Maximum Likelihood Estimates

                                                        Standard          Wald
   Parameter    DF    Estimate        Error     Chi-Square    Pr > ChiSq

Intercept      1    -72.4880     46.8321       2.3958        0.1217
MaxT           1      1.8459      1.2023       2.3573        0.1247
RH             1      0.2199      0.1530       2.0653        0.1507
```

Consider as another example, data from the field of medical sciences relating to Occurrence or Non-occurrence of Coronary Heart Disease (CHD) in human beings as given in the following table.

Group	Age	No. of observations	Presence of CHD
1	25	10	1
2	30	15	2
3	35	12	3
4	40	15	5
5	45	13	6
6	50	8	5
7	55	17	13
8	60	10	8

The SAS syntax is given as follows:

```
data medical;
   input age n CHD;
   cards;
25   10   1
30   15   2
35   12   3
40   15   5
45   13   6
50   8    5
55   17   13
60   10   8
;
 proc logistic data=medical;
```

```
     model CHD/n  = age /lackfit;
   run;
```

The SAS output is given below:

```
                                      The LOGISTIC Procedure
                                        Model Information
                         Data Set                           WORK.MEDICAL
                         Response Variable (Events)         CHD
                         Response Variable (Trials)         n
                         Model                              binary logit
                         Optimization Technique             Fisher's scoring

                             Number of Observations Used           8
                             Sum of Frequencies Used             100

                                         Response Profile
                                Ordered     Binary           Total
                                  Value     Outcome      Frequency
                                      1     Event               43
                                      2     Nonevent            57

                                       Model Fit Statistics
                                                           Intercept
                                            Intercept            and
                              Criterion          Only     Covariates

                              AIC             138.663        112.178
                              SC              141.268        117.388
                              -2 Log L        136.663        108.178

                           Testing Global Null Hypothesis: BETA=0

                   Test                 Chi-Square       DF     Pr > ChiSq

                   Likelihood Ratio        28.4851        1         <.0001
                   Score                   26.0782        1         <.0001
                   Wald                    21.4281        1         <.0001

                            Analysis of Maximum Likelihood Estimates

                                                    Standard          Wald
Parameter    DF     Estimate         Error    Chi-Square    Pr > ChiSq
Intercept     1      -5.1092        1.0852       22.1641        <.0001
age           1       0.1116        0.0241       21.4281        <.0001


                                       Odds Ratio Estimates
                                         Point          95% Wald
                            Effect    Estimate      Confidence Limits
                            age          1.118       1.066       1.172
```

```
Association of Predicted Probabilities and Observed Responses
                        Percent Concordant     74.6    Somers' D    0.588
                        Percent Discordant     15.7    Gamma        0.651
                        Percent Tied            9.7    Tau-a        0.291
                        Pairs                  2451    c            0.794

                            Partition for the Hosmer and Lemeshow Test
                            Event                  Nonevent
Group       Total     Observed     Expected     Observed     Expected
1           10           1          0.90            9         9.10
2           15           2          2.20           13        12.80
3           12           3          2.77            9         9.23
4           15           5          5.16           10         9.84
5           13           6          6.22            7         6.78
6            8           5          4.93            3         3.07
7           17          13         12.53            4         4.47
8           10           8          8.30            2         1.70

                             Hosmer and Lemeshow Goodness-of-Fit Test
                                Chi-Square       DF     Pr > ChiSq
                                    0.2178        6         0.9998
```

The Interpretation of the above output is given subsequently.

The fitted model is given by

$P(CHD=1) = 1/(1+\exp(-z))$ where $z = \beta 0+\beta 1 * (\text{age group})$

Testing of overall Null Hypothesis that BETA = 0 using Likelihood and other tests indicate that they are highly significant and hence there is considerable effect on age on CHD disease.

The Hosmer-Lemeshow Goodness of Fit Test with 6 degrees of freedom suggests that the fitted model is adequate. Here one has to see for a large p-value (>0.05). in order to infer that the model is very well fitted.

Table : Classification Table for Predicted Event frequencies

Correct		Incorrect		Percentages				
Event	Non-event	Event	Non-event	Hit rate	Sensitivity	Specificity	False POS	False NEG
48	26	17	9	74.0	84.2	60.5	26.2	25.7

Here "Correct" columns list the numbers of subjects that are correctly predicted as events and nonevents. Also "Incorrect" columns list both the number of nonevents incorrectly predicted as events and the number of events incorrectly predicted as nonevents.

FALSE positive and FALSE negative rates are low, sensitivity (the ability of the model to predict an event correctly) (84.2%) and specificity (the ability of the model to predict a nonevent correctly) (60.5%) of the model are high enough and hence the fitted model is very effective for prediction/ classification.

8. Application of multinomial logit model in socio-economic studies

Consider the 61st Round. NSSO unit-level data of Household Consumer Survey for the district Vellore of Tamil nadu state. On the basis of MPCE i.e. Monthly Per Capita Expenditure, the available 240 households were divided into three income groups (low, medium or high) i.e. LIG, MIG and HIg denoted numerically by ‘1’, ‘2’ and ‘3’ respectively. If this income group is considered as a response variable taking multinomial values, then let us fit a ordinal logistic regression model. A partial listing of data with variables used is given below for perusal.

HHSize	HHType	Religion	SocialGroup	LandTotalPossess	GroupCode
1	3	1	1	0	1
1	3	1	2	0.004	1
…	…	…	…	…	…
2	2	1	2	2.043	2
…	…	…	…	…	…
2	3	1	2	2.043	3

The syntax code is as follows:

```
Title "Logistic Regression for Vellore District with three levels of dependent variable";
ods html;
proc logistic data=work.vellore outest = betas covout;

Class HHSize  HHType        Religion       SocialGroup          DweelingUnitCode
        CookingCode  LightingCode  RegularSalaryIncome RationCardPossess
        RationCardType        BeneficiaryFoodForWork     BeneficiaryAnnapoorna
        BeneficiaryICDS       BeneficiaryMidDayMeal      CeremonyPerform
        GroupCode    OkStampLevel3       BlankLevel3   ;
model GroupCode=                    HHSize                  HHType        Religion
        SocialGroup           LandTotalPossess
            / lackfit;

   weight WeightLevel2;
output out=work.outputfile/*p=phat lower=lcl upper=ucl*/
              predprob=(individual);
run;
```

The abridged output is as follows:

```
                          Response Profile
         Ordered     Group             Total            Total
           Value     Code          Frequency           Weight

               1         1               117        261942.85
               2         2               111        208105.93
               3         3                12         22104.90

Probabilities modeled are cumulated over the lower Ordered Values.

                       Class Level Information
                                          Design
                Class           Value     Variables

                HHSize          1           1      0
                                2           0      1
                                3          -1     -1

                HHType          1           1      0
                                2           0      1
                                3          -1     -1

                Religion        1           1      0
                                2           0      1
                                3          -1     -1

                SocialGroup     1           1
                                2          -1
```

```
                        The LOGISTIC Procedure
                 Analysis of Maximum Likelihood Estimates
                                          Standard          Wald
Parameter            DF     Estimate         Error    Chi-Square    Pr > ChiSq

Intercept        1    1       0.6758        0.0356      360.0667        <.0001
Intercept        2    1       4.1581        0.0363    13096.4382        <.0001
HHSize           1    1      -0.2642        0.0205      165.6787        <.0001
HHSize           2    1       0.9774        0.0207     2224.6344        <.0001
HHType           1    1       0.6736       0.00593    12918.4494        <.0001
HHType           2    1      -0.0508       0.00769       43.5620        <.0001
Religion         1    1      -0.0711        0.0287        6.1319        0.0133
Religion         2    1       0.5615        0.0312      322.8992        <.0001
SocialGroup      1    1       0.6022       0.00375    25851.2524        <.0001
LandTotalPossess      1      -0.6067       0.00484    15700.3659        <.0001


                          Odds Ratio Estimates

                                      Point          95% Wald
        Effect                     Estimate      Confidence Limits

        HHSize           1 vs 3       1.567         1.390       1.765
        HHSize           2 vs 3       5.423         4.810       6.113
        HHType           1 vs 3       3.656         3.601       3.712
        HHType           2 vs 3       1.772         1.732       1.812
        Religion         1 vs 3       1.521         1.292       1.790
        Religion         2 vs 3       2.863         2.420       3.387
        SocialGroup      1 vs 2       3.335         3.286       3.384
        LandTotalPossess              0.545         0.540       0.550
```

Here we can see that there are two fitted equations of the logistic regression will be obtained if there are three levels of the response variable.

References:

Fox, J. (1984). *Linear statistical models and related methods with application to social research*, Wiley, New York.

Kleinbaum, D.G. (1994). *Logistic regression*: A self learning text, New York: Springer.