4. GENERALISED LINEAR MODEL FOR DISCRETE DATA

4.1: Introduction

In chapter 3, a wider class of models, namely, generalised linear models is discussed. These models are useful for some non normal distributions. As mentioned at the end of chapter 3, in many practical situations we come across the discrete data.

If the response variate takes one of the fixed set of possible values, it is called as discrete response variate and the corresponding data as discrete data. Suppose there are 'p' possible values of the response variate. These possible values of the discrete response variate are frequently called as response categories. Some of the examples of discrete responses are number of births in a certain period, number of girls marrried before maturity age in the specific geographical area etc.

In many practical situations we come accross the observations which takes one of the two possible forms. Such observations are known as 'quantal' or 'binary' observations. Thus in case of binary responses we have two response categories. Some examples of binary responses are dead-alive, married - unmarried, defective-non defective, male-female etc. Generally, the two possible forms of binary responses are termed as 'success' and 'failure'.

Some techniques are developed for analysing binary data. These techniques are discussed mainly by Finney (1947), Cox (1970), Nelder & Wedderburn (1972), Prentice (1976), McCullagh & Nelder (1989), Hosmer & Lemeshow (1989), Dobson (1990) and Collett (1991).

Most frequently we come across the counted data not in the form of proportions but in the form of table of counts. In such

cases the data are sorted according to one or more factors each having atleast two levels. Some techniques are developed to analyse such type of data. Birch (1963), Bishop (1967), Bishop & others (1975), Good (1956,1963), Gokhale & Kullback (1978) and many others discussed these techniques.

In this chapter, we discuss the theory associated with analysis of binary data through model fitting and that related with the data in the form of table of counts.

Here the discussion proceeds in the following direction.

(1) Types of binary data;

(2) binary distribution and log-likelihood function;

(3) link functions;

(4) fitting of generalised linear models and analysis of deviance;

(5) model checking;

(6) other appropriate models for binary data;

(7) log linear model (Bishop (1969));

and

(8) log linear model (Nelder & Wedderburn (1972)).

4.2 : Types of binary data

Suppose $X = (\underline{1}, X_1)$ is the incidense matrix and $X_1=(\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_K)$. Let for a particular combination $(X_{i1}, X_{i2}, \ldots, X_{ik})$, observations are available on m_i^* items. Depending on the values of m_i^*, there are two types of binary data. These are

(a) ungrouped binary data;

and

(b) grouped binary data.

(a) *Ungrouped binary data* : If $m_i^*=1$, for all i(i=1,2,...,k), then the binary data is known as 'ungrouped' binary data.

Example 4.1: This example is taken from Collett (1991) (Given on page 6). Here the response variate (Y) is taken as a variate directly related to erythrocyte sedimentation rate (ESR). This rate is nothing but the rate at which the red blood cells settle out of suspension in blood plasma, when measured under standard conditions. The ESR increases if the levels of certain proteins in the blood plasma increases. In Malaysia, an experiment was carried out to examine the extent to which the ESR is related to two plasma proteins, fibrinogen and γ-globulin. Since the ESR for healthy person should be less than 20 mm/h and since it is non negative, the ESR is observed to see whether it is less than 20. Therefore the binary response variate Y takes value '1' if ESR exceeds 20 and '0' otherwise. The experiment was carried out on 32 individuals. The data are given in Collett (1991). In this example, since the observations are available on each item, the data is ungrouped binary data.

(b) *Grouped binary data* : If $m_i \neq 1$, for atleast one value of $i (i=1,2,\ldots,k)$, the data is known as 'grouped' binary data.

Example 4.2 :

The data is collected from Kolhapur city in the year 1992, to study the occupational and educational relationship between parents and their children having age more than 27. Proper clubing of collected data gives table 4.1.

TABLE 4.1

Occupational status of fathers	Occupational status of children					
	Sons			Doughters		
	Ser.	Busi	Total	Ser.	Busi.	Total
Service	316	284	600	119	6	125
Business	102	552	654	210	108	318

Foot note : Ser :- service , Busi :- Business.

Remark : The category of house wife is ignored, because it will result in structural zero.

Here the response variate is occupation of children and the two covariats are occupation of father (X_1) and sex of children (X_2). We denote the occupation 'service' by value '0' and 'business' by value '1'. Similarly for X_2, we define the value '0' to 'daughters' and '1' to 'sons'.

Here since more than one observations occurs corresponding to every combination of values of the stimulus variates (X_1, X_2), the values $Y_i = Y_i^*/m_i^*$ $(i=1,2,3,4)$ constitutes grouped binary data.

Most of the times grouped binary data are available as they are condensed forms of ungrouped binary data. The examples (4.1) and (4.2) are sufficient to distinguish between grouped and ungrouped binary data. Now one can proceed towards the distribution of these types of data.

4.3 : Distribution and log-likelihood function

The distribution and log-likelihood function for grouped binary data is discussed in section (3.3). Here we discuss the distribution and log-likelihood function for ungrouped binary data.

Distribution and log-likelihood function for ungrouped binary data : Suppose the response variate Y has binary distribution with mean μ. Then p.m.f. of Y is given by,

$$f(y;\mu) = \mu^{y} (1-\mu)^{1-y} I_{A_i}(y), \qquad (1)$$

where

(i) $A_i = \{0, 1\}$

and

(ii) $I_{A_i}(y)$ is the indicater function.

Therefore, log-likelihood function based on single observation becomes

$$\ell(\mu,y) = y\ln(\mu) + (1-y)\ln(1-\mu)$$

$$= y\ln[\mu/(1-\mu)] + \ln(1-\mu). \qquad (2)$$

Comparison of this equation with the equation (3.3-3) gives.

$$\left.\begin{aligned} \phi = \alpha(\phi) = 1,\ \theta = \ln[\mu/(1-\mu)],\ g(\theta) = \ln(1+\exp(\theta)), \\ h(y) = 0,\ \xi(\phi;y) = 0. \end{aligned}\right\} \quad (3)$$

For fitting generalised linear model to the binary data, it is necessary to obtain link functions.

4.4 : Link function

To obtain the relationship between the parameter μ and the vector of stimulus variates $\underline{X} = (X_1, X_2, \ldots, X_k)'$, it is convenient though not necessary, to construct a model, which is able to describe the effect of different combinations of values of covariates on μ. While constructing a model, certain assumptions are made. Validity of these assumptions should be checked. Further, the model constructed should be consistent in its behaviour.

In section (2.7), we have seen, though classical linear

model is widely used model, it is not suitable for binary data. Thus the identity link $T=\mu$, is not appropriate for binary data. The number of suitable link functions for binary data are available. Four of them which are mostly used in practice are as follows.

(i) Logit or linear logistic function,

$$T = \ln\left\{\frac{\mu}{1-\mu}\right\} . \qquad (1)$$

(ii) Complementory log-log function,

$$T = \ln[-\ln(1-\mu)] . \qquad (2)$$

(iii) Probit or inverse normal function,

$$T = \phi^{-1}(\mu). \qquad (3)$$

(iv) Log-log function,

$$T = -\ln[-\ln(\mu)] . \qquad (4)$$

For $\mu < 0.5$, the behaviour of log-log function is not appropriate and hence the log-log fnction is not much useful when $\mu < 0.5$.

Corresponding to each link function given in equations (1) to (4), we can fit generalised linear model to the binary data.

Note :- While fitting a generalised linear model to the grouped binary data, it is assumed that the observations Y_i^* $(i=1,2,\ldots,n)$ are independent. It happens only when Y_i's are independent with constant probability of success for all items in the same group. In most of the practical situations, observations between the groups are independent or atleast uncorrelated, but observations within the group are likely to be correlated. The situations with positive and negative correlation are named as 'over' and 'under' dispersion

respectively. Thus if

$$V_{ii} = \phi^* \left\{ \frac{\mu_i (1-\mu_i)}{n_i} \right\} ,$$

then in over dispersion case ϕ^* is greater than unity and in under dispersion case it is less than unity. Here, since ϕ^* is an unknown parameter, it is to be estimated from the available data. These two situations can be handled with the help of 'quasi likelihood functions', and will be discussed thoroughly later in chapter 5. For a time being it is assumed that ϕ^* is unity.

Before fitting any model, it is necessary to describe the model explicitly. In the next section, we discuss definition and fitting of different generalised linear models to the two types of binary data.

4.5 : Fitting of generalised linear model to the binary data

Let Y be the response variate and $X_j (j=1,2,\ldots,k)$ be the stimulus variates. First we consider fitting of four different models, corresponding to the four link functions, for grouped binary data. Fitting of each of these four models is followed by a note about fitting the same model for ungrouped binary data. A generalised linear model to the binary data with the four types of link functions given in equations (4.4-1) to (4.4-4) are respectively known as, 'linear logistic model', 'complementory log-log model', 'probit model' and 'log-log model'.

4.5.1 : Fitting of linear logistic model for grouped binary data

Logistic regression model was first suggested by Berkson (1944). He pointed out that, the model can be fitted using 'iterative weighted least square' method as a numerical technique.

Definition-1 : Logistic model for grouped binary data : Suppose the response variate Y is such that $Y^* = m^*Y$ has $B(m^*,p)$ distribution. Suppose Y_i $(i=1,2,\ldots,n)$ are n realisations of the response variate Y. Let $\underline{x}_j$ be the vectors of the known values of the covariates X_j $(j=1,2,\ldots,k)$ and $\underline{T} = X\beta$ be the linear predictor. Then the model of the form,

$$T_i = \ln[\mu_i/(1-\mu_i)], \text{ for } i=1,2,\ldots,n, \qquad (1)$$

with $E(Y_i) = \mu_i$, is called linear logistic model.

Fitting a logistic model : Fitting of any generalised linear model to the data consists of the following steps.

(I) Finding expressions for the vector $\underline{z}$ of adjusted dependent variable, and for W;

(II) applying the method of fitting generalised linear model discussed in sub section (3.4.3).

Logistic model is one of the proper models for binary data, having a vast scope. From equation (1) we have,

$$\mu_i = \frac{\exp(T_i)}{(1+\exp(T_i))}, \quad i=1,2,\ldots,n. \qquad (2)$$

Now, the respective values of Z_i and W_{ii} can be obtained as below. Differentiation of equation (2) w.r.t. μ_i gives,

$$(dT_i/d\mu_i) = [\mu_i(1-\mu_i)]^{-1}. \qquad (3)$$

Also note that,

$$\left.\begin{aligned} E(Y_i) &= \mu_i, \\ Var(Y_i) &= [\mu_i(1-\mu_i)/n_i] \end{aligned}\right\} \qquad (4)$$

Equation (3.4-19) gives i^{th} component of the vector $\underline{Z}$ as

$$Z_i = T_i + \left\{ \frac{(Y_i - \mu_i)}{\mu_i (1-\mu_i)} \right\} . \qquad (5)$$

Similarly, from equations (3.4-18),(3) and (4),we have

$$W_{ii} = \mu_i (1-\mu_i). \qquad (6)$$

After obtaining expressions for Z_i and W_{ii}, to fit a linear logistic model, method described in sub section (3.4.3) can be applied easily. As mentioned there, one can start the calculations by taking initial estimates $\hat{\mu}_i^{(0)}$ of μ_i as y_i.

Remark : If any of the samples has either all failures or all successes, the value y_i reaches to its extreme value, zero or one respectively. In such a case add or subtract a number $\frac{1}{2}$ 0.5 from the corresponding value of Y^* accordingly, so that none of $\hat{\mu}_i^{(0)}$ is either zero or one.

Gore & Shanubhogue (1984) have discussed the applications of linear logistic regression models in the field of ecology. Strauss (1992) focused on many faces of logistic regression and showed how it can be used to analyse the data in different fields.

sometimes instead of using linear logistic model, it is preferable to use the generalised logistic model. Johansson (1973) has used generalised logistic model to study the effect of advertising.

For single observation on the response variate, and with the single explanatory variate Y, the linear logistic function given in equation (2) becomes,

$$ln\left\{ \frac{\mu}{1-\mu} \right\} = \beta_0 + \beta_1 X . \qquad (7)$$

Note that the linear logistic function is symmetric about 0.5. The value about which the function is symmetric, is known as point of inflection.

When a saturated level K^* different from 1 is required, the modified linear logistic function can be written as,

$$\ln\left\{\frac{\mu}{K^* - \mu}\right\} = \beta_0 + \beta_1 X \ . \qquad (8)$$

To introduce non symmetry in the model, one can add second order term in X on the right hand side (RHS) of the equation (8). Thus equation (8) becomes,

$$\ln\left\{\frac{\mu}{K^* - \mu}\right\} = \beta_0 + \beta_1 X + \beta_2 X^2. \qquad (9)$$

When the explanatory variable X takes only positive values, the non-symmetry can be allowed in the model by reformulating the model as,

$$\ln\left\{\frac{\mu}{K^* - \mu}\right\} = \ln(\beta_0) + \beta_1 \ln(X). \qquad (10)$$

Applying least square method of estimation to equation (10), it can easily be seen that, intercept $\ln(\beta_0)$ of the model is zero. To have a model with with non-zero intercept, Johansson (1973) suggested the following form of generalised logistic function.

$$\ln\left\{\frac{\mu - I^*}{K^* - \mu}\right\} = \ln(\beta_0) + \beta_1 \ln(X). \qquad (11)$$

When there are K stimulus variates $X_1, X_2, \ldots, X_k$ $(k > 1)$, then

logistic function in equation (11) changes to,

$$\ln\left\{\frac{\mu - I^*}{K^* - \mu}\right\} = \ln(\beta_o) + \sum_j \beta_j \ln(X_j) \quad . \qquad (12)$$

Johansson has used generalised logistic function given in equation (12) to study the effect of advertising.

Analysis of deviance(ANODEV) : Detailed discussion on ANODEV is made in section (3.6). As mentioned ther, ANODEV can be carried out by computing deviance function for the sequence of models, and then taking differences between appropriate deviance functions. Finally, for drawing conclusions, assymptotic results about the distribution of difference between deviances is used. Equation (3.5-35) gives the deviance for grouped binary data. This equation is useful to carry out ANODEV.

Example 4.2 (Cont.):-

Data in table (4.1) is useful to study the relationship between stimulus variates X_1, X_2 and the response variate Y. We can fit four linear logistic models stated below, to the data.

$$\text{Model (1)} : E(Y_i) = \frac{\exp(\beta_o)}{1+\exp(\beta_o)} \quad . \qquad (13)$$

$$\text{Model (2)} : E(Y_i) = \frac{\exp(\beta_o + \beta_1 X_{i1})}{1 + \exp(\beta_o + \beta_1 X_{i1})} \quad . \qquad (14)$$

$$\text{Model (3)} : E(Y_i) = \frac{\exp(\beta_o + \beta_1 X_{i1} + \beta_2 X_{i2})}{1 + \exp(\beta_o + \beta_1 X_{i1} + \beta_2 X_{i2})} \quad . \qquad (15)$$

$$\text{Model (4)} : E(Y_i) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2})} \quad (16)$$

Once the model is fitted, it can be used to estimate the values of Y_i, under the respective model.

Data in the table (4.1) is useful to study the relationship between occupational status of children (Y) and occupational status of fathers (X_1). If the factor sex of the children (X_2) is taken as another explanatory variable, the relationship between Y and X_1, X_2 can be studied. Here both the stimulus variates are at two levels '1' and '0',(say). Let Y_i be the proportion of children joining service in different fields. Then four linear logistic regression models for Y_i are as stated in models (1) to (4).

Executing program in appendix-2(A), to fit the models (1) to (4) to the data from example (4.2), parameter estimates obtained are given in the tables (4.2) to (4.5). At the bottom of table deviance is given.

TABLE 4.2

Sr.No.	Estimate	S.E.	Parameter
1	-0.2403968	0.48901E-01	INTERCEPT

Deviance = 476.1927

TABLE 4.3

Sr.No.	Estimate	S.E.	Parameter
1	0.4054649	0.75810E-01	INTERCEPT
2	-1.1547020	0.10231	OF

Deviance = 343.9368

TABLE 4.4

Sr.No.	Estimate	S.E.	Parameter
1	2.5576600	0.1605	INTERCEPT
2	-1.850483	0.12779	OF
3	-2.430897	0.14714	SEX

Deviance = 1.495048

TABLE 4.5

Sr.No.	Estimate	S.E.	Parameter
1	2.9873710	0.41841	INTERCEPT
2	-2.322401	0.43485	OF
3	-2.880606	0.42633	SEX
4	0.5270557	0.4554	OF.SEX

Deviance = 0

Now the analysis of deviance table can be prepared as below

ANODEV TABLE

TABLE 4.6

Model Description	d.f.	Deviance	First difference
Null	3	476.1927	
OF	2	343.9368	132.2559
OF+SEX	1	1.4950	342.4418
OF+SEX+OF.SEX	0	0	1.4950

Comparing the first difference with the table value of chi-square with appropriate d.f. (here d.f. of the difference is unity), following conclusions are drawn.

(1) Sex (of a child) plays a vital role in the occupation.

(2) Occupation of father is also an important factor showing effect on occupation of their children.

(3) The factors occupation of fathers and sex of their children are likely to be independent.

Table (4.6) is the ANODEV table which has its own importance in deciding which factors shows effect on the response variate.

NOTE :- To fit a linear logistic model for ungrouped binary data following changes are required.

(1) In the definition of linear logistic model, the distribution of response variate Y is bernaulli distribution with parameter μ.

(2) To start the calculations, take $(\sum_i y_i / n)$ as initial estimate $\hat{\mu}_i^{(0)}$ of μ_i.

(3) ANODEV can be carried out as usual in case of grouped binary data.

(4) As the ungrouped binary data occurs very rarely, we will not illustrate its fitting numerically. However the following discussion is very important.

Discussion : McCullagh & Nelder (1989) claimed that in case of ungrouped binary data, deviance is not giving any inference about the goodness of fit of the fitted model. This claim can be justified by showing that deviance depends only on fitted values and not on the observations.

From equation (4.3-2), we get

$$\ell(\underline{y}, \phi; \underline{y}) = \sum_i \left\{ y_i \ln(y_i) + (1-y_i)\ln(1-y_i) \right\} . \qquad (17)$$

Since Y_i takes only two possible values zero and one, it is clear that,

$$\ell(\underline{y}, \phi; \underline{y}) = \underline{0}. \qquad (18)$$

Hence deviance is given by,

$$D(\underline{y}, \hat{\underline{\mu}}) = -2 \{ \ell(\hat{\underline{\mu}}, \phi, \underline{y}) \} \qquad (19)$$

$$= -2 \sum_i [\, y_i \hat{T}_i + \ln(1-\hat{\mu}_i) \,]. \qquad (20)$$

For simplification, consider equation (3.4-22). It gives,

$$(\partial \ell / \partial \beta_j) = \sum_i (y_i - \mu_i) X_{ij}; \qquad (21)$$

which implies,

$$\sum_j \beta_j (\partial \ell / \partial \beta_j) = \sum_i (y_i - \mu_i) T_i. \qquad (22)$$

Further, since estimates $\hat{\beta}$ of β are the solutions of the equations

$$(\partial \ell / \partial \beta_j) = 0; \quad j=0,1,\ldots,k, \qquad (23)$$

equation (21) gives,

$$\sum_i (y_i - \hat{\mu}_i) \hat{T}_i = 0. \qquad (24)$$

Finally, equations (20) and (24) combinedly give,

$$D(\underline{y}, \hat{\underline{\mu}}) = -2 \sum_i \left\{ \hat{\mu}_i \hat{T}_i + \ln(1- \hat{\mu}_i) \right\}. \qquad (25)$$

This proves the claim.

Thus, usual measure (deviance) of goodness of fit is not useful for ungrouped binary data. Hence the other measure, namely, Pearson's χ^2 statistic defined in equation (3.5-2) must be used.

4.5.2 : Fitting of complementory log-log model

This is another important model for grouped binary data. The complementory log-log transformation and its use in dilution assay was described by Fisher (1922). Crowder (1991) gave the application of this model in the field of reliability. This model is useful for extreme values of Y.

Definition-2 : Complementory log-log model : Suppose the

response variate Y is such that $Y^* = m^*Y$ has $B(m^*,p)$ distribution. Suppose Y_i $(i=1,2,\ldots,n)$ are n realisations of the response variate Y. Let $\underline{x}_j$ be the vectors of the known values of the covariates X_j $(j=1,2,\ldots,k)$ and $\underline{T} = X\underline{\beta}$ be the linear predictor. Then the model of the form,

$$T_i = \ln[-\ln(1-\mu_i)], \text{ for } i=1,2,\ldots,n, \qquad (26)$$

with $F(Y_i) = \mu_i$, is called complementory log-log model.

Fitting a complemantory log-log model :

The equation (26) implies,

$$\mu_i = 1 - \exp\{-\exp(T_i)\}. \qquad (27)$$

Differentiating equation (27) w.r.t. μ_i, we have

$$(dT_i / d\mu_i) = -\{(1-\mu_i)\ln(1-\mu_i)\}^{-1}. \qquad (28)$$

Now equations (3.4-19) and (28) gives i^{th} component of $\underline{Z}$ as

$$\hat{Z}_i = \hat{T}_i - \left\{ \frac{(y_i - \hat{\mu}_i)}{(1-\hat{\mu}_i)\ln(1-\hat{\mu}_i)} \right\}. \qquad (29)$$

Similarly, from equations (3.4-18), (6) and (28) it is clear that,

$$\hat{W}_{ii} = \frac{(1-\hat{\mu}_i)[\ln(1-\hat{\mu}_i)]^2}{\hat{\mu}_i}. \qquad (30)$$

Once the expressions for Z_i and W_{ii} are obtained, a complementory log-log model can be fitted by using method described in sub section (3.4.3). The initial estimates $\hat{\mu}^{(0)}$ of μ can be taken as y_i. Remarks mentioned in (4.5.1) are also applicable here.

Another important point to be noted here is that in this

model fitting, since

$$\mu_i = 1 - \exp\{-\exp(T_i)\}, \qquad (31)$$

the values of estimates often shoots towards both the extremities, oftenly. Thus it is necessary to keep restriction on values of the estimates so that their values does not go beyond accuracy limit of the computer at any cycle.

We have developed a FORTRAN-77 program and it is given in Appendix-2(A). This is useful to estimate β under this model.

Example 4.2 (cont.) :

Here four complementory log-log models which can be fitted to the corresponding data are given below.

Model (5) : $E(Y_i) = 1-\exp(-\exp(T_{1i})$, (31)

Model (6) : $E(Y_i) = 1-\exp(-\exp(T_{2i})$, (32)

Model (7) : $E(Y_i) = 1-\exp(-\exp(T_{3i})$, (33)

Model (8) : $E(Y_i) = 1-\exp(-\exp(T_{4i})$, (34)

where

(i) $T_{1i} = \beta_0$,

(ii) $T_{2i} = T_{1i} + \beta_1 X_{i1}$,

(iii) $T_{3i} = T_{2i} + \beta_2 X_{i2}$,

and

(iv) $T_{4i} = T_{3i} + \beta_3 X_{i1} . X_{i2}$.

Execution of the program in Appendix-2(A) to fit the models (5) to (8), for the given data, gives the parameter esimates and

deviance under each model as shown in the tables (4.7) to (4.10).

TABLE 4.7

Sr.No.	Estimate	S.E.	Parameter
1	-0.5444595	0.37103E-01	INTERCEPT

Deviance = 476.1927

TABLE 4.8

Sr.No.	Estimate	S.E.	Parameter
1	-0.0874217	0.49641E-01	INTERCEPT
2	-0.8616092	0.75562E-01	OF

Deviance = 343.9369

TABLE 4.9

Sr.No.	Estimate	S.E.	Parameter
1	1.33573	0.10102	INTERCEPT
2	-1.326136	0.93079E-01	OF
3	-1.668342	0.96917E-01	SEX

Deviance = 5.654145

TABLE 4.10

Sr.No.	Estimate	S.E.	Parameter
1	1.110723	0.13118	INTERCEPT
2	-1.033836	0.14984	OF
3	-1.401135	0.14326	SEX
4	-0.4503045	0.18866	OF.SEX

Deviance = 0

To decide which terms in the model are siginificant, table of deviance difference is useful. This table is given below.

TABLE 4.11

Model Description	d.f.	Deviance	First difference
Null	3	476.1927	
OF	2	343.9369	132.2558
OF+SEX	1	5.6541	338.2828
OF+SEX+OF*SEX	0	0	5.6541

Comparison of the deviance differences in table (4.22) with the table chi-square value with one d.f. gives the first two conclusions similar to those based on linear logistic model. Here fitting of complementory log-log model indicates that occupation of father and sex of their children may be related.

Now we will not discuss the 'probit' model as Berkson (1951) pointed out the reasons to support why he prefers 'logits' to 'probits'. However Fisher (1947) used probit model to analyse

binary data. Now we discuss the 'log-log' model.

4.5.3 : Fitting of log-log model

This is another appropriate model for extreme values.

Definition-3 : Log-log model : Suppose the response variate Y is such that $Y^* = m^*Y$ has $B(m^*,p)$ distribution. Suppose Y_i $(i=1,2,\ldots,n)$ are n realisations of the response variate Y. Let $\underline{X}_j$ be the vectors of the known values of the covariates X_j $(j=1,2,\ldots,k)$ and $\underline{T} = X\underline{\beta}$ be the linear predictor. Then the model of the form,

$$T_i = \ln[-\ln(\mu_i)], \text{ for } i=1,2,\ldots,n, \qquad (35)$$

with $E(Y_i) = \mu_i$, is called log-log model.

Fitting a log-log model :

The equation (35) gives

$$\mu_i = \exp\{-\exp(-T_i)\}. \qquad (36)$$

Differentiating equation (36) w.r.t. μ_i, we have

$$(dT_i / d\mu_i) = -\{\mu_i \ln(\mu_i)\}^{-1}. \qquad (37)$$

Now equations (3.4-19) and (37) gives i^{th} component of $\underline{Z}$ as

$$\hat{Z}_i = \hat{T}_i - \left\{ \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i \ln(\hat{\mu}_i)} \right\}. \qquad (38)$$

Similarly, from equations (3.4-18), (7) and (37) it is clear that,

$$\hat{W}_{ii} = \frac{\hat{\mu}_i [\ln(\hat{\mu}_i)]^2}{(1-\hat{\mu}_i)}. \qquad (39)$$

Once the expressions for Z_i and W_{ii} are obtained, a log-log model can be fitted by using method described in sub section (3.4.3). The initial estimates $\hat{\underline{\mu}}^{(0)}$ of $\underline{\mu}$ can be taken as y_i.

Remark mentioned in (4.5.1) is also applicable here.

Another important point to be noted here is that, in this model fitting, since

$$\mu_i = \exp\{-\exp(-T_i)\}, \qquad (40)$$

the values of estimates often shoots towards both the extremities, oftenly. Thus it is necessary to keep restriction on values of the estimates so that their values does not go beyond accuracy limit of the computer at any cycle.

A FORTRAN-77 program is developed and it is given in Appendix-2(A) which is useful to estimate $\underline{\beta}$ under this model.

Example 4.2 (cont.) :

Here four log-log models which can be fitted to the corresponding data are given below.

Model (9) : $E(Y_i) = \exp(-\exp(-T_{1i})$, (41)

Model (10) : $E(Y_i) = \exp(-\exp(-T_{2i})$, (42)

Model (11) : $E(Y_i) = \exp(-\exp(-T_{3i})$, (43)

Model (12) : $E(Y_i) = \exp(-\exp(-T_{4i})$, (44)

where

(i) $T_{1i} = \beta_0$,

(ii) $T_{2i} = T_{1i} + \beta_1 X_{i1}$,

(iii) $T_{3i} = T_{2i} + \beta_2 X_{i2}$,

and

(iv) $T_{4i} = T_{3i} + \beta_3 X_{i1} \cdot X_{i2}$.

Execution of the program in Appendix-2(A) to fit the models (9) to (12), for the given data, gives the parameter esimates and deviance under each model as shown in the tables (4.12) to (4.15).

TABLE 4.12

Sr.No.	Estimate	S.E.	Parameter
1	0.1977779	0.33362E-01	INTERCEPT

Deviance = 476.1927

TABLE 4.13

Sr.No.	Estimate	S.E.	Parameter
1	0.6717269	0.49362E-01	INTERCEPT
2	-0.7995505	0.72175E-01	OF

Deviance = 343.9368

TABLE 4.14

Sr.No.	Estimate	S.E.	Parameter
1	2.088353	0.11453	INTERCEPT
2	-1.119431	0.75430E-01	OF
3	-1.610214	0.1032	SEX

Deviance = 8.4670

TABLE 4.15

Sr.No.	Estimate	S.E.	Parameter
1	3.012016	0.4059	INTERCEPT
2	-2.132401	0.41731	OF
3	-2.567581	0.41036	SEX
4	1.068398	0.42448	OF.SEX

Deviance = 0

To decide which terms in the model are siginificant, table of deviance difference is useful. This table is given below.

TABLE 4.16

Model Description	d.f.	Deviance	First differenc
Null	3	476.1927	
OF	2	343.9368	132.2559
OF+SEX	1	8.4670	335.4698
OF+SEX+OF.SEX	0	0	8.4670

Comparison of the deviance differences in table (4.16) with the table chi-square value with one d.f.gives the conclusions similar to those based on complementory loglog model.

After fitting any model to the data, next part is to check whether the model fits well to the data and assumptions made during fitting the model are not necessarily invalid. Therefore in the next section, we discuss model checking.

4.6 : Model checking :

While checking the model, one has to look towards the model in different angles. It is atleast necessary to check the fitted model for

(1) the form of the linear predictor,

(2) the inclusion of explanatory variate,

(3) the adequacy of link function,

(4) the presence of outliers,

(5) the goodness of link test.

As in section (2.5) one can use the different residual plots to draw the conclusions about adequacy of the fitted model. Before drawing the residual plots, it is necessary to decide, which among the three residuals, discussed in section (3.5) should be used. Williams (1984) and Pierce & Schafer (1986) showed that, when binomial indexes are not very small, a standard normal distribution is a better approximation. Hence, when model is suitable for the data,have their values in between -2 to 2. For a suitable model, if the absolute value of residual exceeds the value 2 for any data point, it indicates that particular data point may be outlier. In case of binary distribution, since computation of Anscombe residual is difficult, the deviance residual is computed.

Further part of the model checking can be carried out as discussed in section (3.8). Since the computer programs in FORTRAN-77 are not developed for model checking, we will not check adequacy of any model numerically. We discuss in chapter 6,the model checking based on residual plots.

At this stage the question in one's mind may be 'Are there some more models suitable for binary data ?'.Hence below we discuss one suitable model for binary data and other methods of analysing the binary data.

4.7 : Other methods of analysing binary data :

Another way of analysing any data is to make the proper transformations on the response variate, so that a classical linear model is suitable for transformed data. sometimes, use of transformed responses for the analysis is helpful because it may be simpler to fit classical linear model to the transformed data than to fit generalised linear model to the untransformed one. Nelder (1968) has discussed these normalising and linearising transformations in general case. As our aim is restircted to generalised linear model, the details of trnsformations are not given here. One can refer Nelder (1968) for further details.

Log-linear model is another proper model for binary data. This is so because the data can be considered as in the form of contingency table. However this model is also suitable for the data presented in the form of cotingency table of any dimension. If the data are classified according to the categories of various variables (factors) related to the response variate, the resulting table is called as contingency table. If the data are classied accodring to the categories of one variable then the contingency table is known as one dimensional contingency table. If for the classification two factors are used, we get two dimensional contingency table and so on. For such type of data the systematic effects are multiplicative in nature. This fact is used to suggest a suitable model in this situation. 'Log linear' model is one of the suitable model for the data where the systematic effects are thought to be of multiplicative nature. Log linear model is nothing but the linear model in terms of log probabilities or in terms of logarithm of the expected cell counts (cell counts are entries in the contingency table). In the next few sections we discuss this model explicitly.

4.8 Log-linear model (Bishop (1969)) :

Bishop (1969) introduced this model for data in the form of contingency table. She defined the log linear model as below.

definition-4 : Log linear model :

Suppose the discrete data are presented in the form of IxJ contingency table. Let Y_{ij} and μ_{ij} ($i \in A_1$, $j \in A_2$) be respectively the observed and expected cell counts for $(i,j)^{th}$ cell. Then the full (saturated) log linear model is given by,

$$\ln[E(Y_{ij})] = U + U_{1(i)} + U_{2(j)} + U_{12(i,j)}, \qquad (1)$$

for $i \in A_1$ and $j \in A_2$

with each of the three subscripted U-terms sums to zero over each lettered subscript. Here $A_1 = \{1,2,\ldots,I\}$ and $A_2 = \{1,2,\ldots,J\}$.

One can define 'unsaturated' log linear model for two dimensional contingency table by deleting any of the four terms on the right hand side of the model (1). Similarly one can easily write log linear model for higher dimension also. The theory related to this model is anologous to that of 'factorial experiments'.

The interpretation of the different U-terms (model parameters) in model (1) is given in detail by Bishop & others (1975). She developed the model and the model fitting procedure anologous to that of factorial experiment. She mentioned three sampling schemes suitable for log linear models. These are

1. independent Poisson sampling (IPS),
2. simple multinomial sampling (SMS)

and

3. product multinomial sampling (PMS).

If the sample size is unrestricted, the observed counts are independently distributed Poisson variates. In this case the probability mass function (p.m.f.) of Y_{ij} ($i \in A_1$, $j \in A_2$) is

given by

$$P(Y_{ij}=y_{ij}) = \left\{(\mu_{ij})^{(y_{ij})} \exp(-\mu_{ij})/(y_{ij})!\right\} I_A(y_{ij}), \quad (2)$$

where

(i) $A = \{0,1,\ldots\}$

(ii) $I_A(y_{ij}) = \begin{cases} 1, & \text{if } y_{ij} \in A \\ 0, & \text{otherwise} \end{cases}$

When the observations are made over a period of time with no prior knowledge about the total number of observations, the distribution of the observed cell counts will be of the above type.

When the sample size N (= IJ) is fixed the restriction on the fixed sample size imposed on a series of independent Poisson distributions give multinomial distribution. Hence the mass function of Y_{ij} ($i \in A_1$, $j \in A_2$) is given by

$$P(Y_{ij}= y_{ij}) = \left[\frac{N!}{\prod_{i,j}(y_{ij})!}\right]\prod_{i,j}\left\{\frac{\mu_{ij}}{N}\right\}^{(y_{ij})} I_B(y_{ij}), \quad (3)$$

where

(i) $B = \{y_{ij} \mid y_{ij} = 0,1,\ldots,N; \sum_{(i,j)} y_{ij} = N\}$,

and

(ii) $I_B(y_{ij}) = \begin{cases} 1; & \text{if } y_{ij} \in B \\ 0; & \text{otherwise} \end{cases}$

sometimes it may happen that though we are examining theoretically a single group, in the experimental situations we frequently have several groups with the total number of observations in each group are determined by the separate sampling scheme. In such situations product multinomial sampling

is a suitable sampling scheme. To describe it explicitly we need to introduce the various terms like 'configuration'. Therefore we will not give the further details.

It is proved that under any of the above three sampling schemes stimates of the expected cell counts are same.This is because, all the three distributions belong to one parameter natural exponential family and for all the distribution cornel of the log likelihood is same. Hence in the further discussion we consider the independent Poisson sampling scheme. Bishop & others (1975) discussed the conditions under which the direct estimates of the cell counts are obtainable. Birch's (1963) iterative procedure of obtaining estimates of the expected cell counts for three dimensional contingency table when the direct estimates are not available can be used to fit a log linear model. As Bishop & others discussed the model fitting procedure explicitly, we will not discuss this method of fitting log linear model. However we have developed independently a PC-based software package in FORTRAN-77 useful for three dimensional contingency table with each factor having at most seven categories (levels). Below we give one numerical example and fit the different log linear models to it.

Example 4.3 : This example is taken from Nelder & Wedderburn (1972). Maxwell (1961) discusses the analysis of a 5x4 contingency table giving the number of boys with four different ratings for distributed dreams in five different age groups. The data are in the following table. The higher the rating the more the boy suffers from disturbed dreams.

TABLE-4.17

Age in years	Rating 4	3	2	1	Total
5-7	7	3	4	7	21
8-9	13	11	15	10	49
10-11	7	11	9	23	50
12-13	10	12	9	28	59
14-15	3	4	5	32	44
Total	40	41	42	100	223

We can fit four different types of log-linear models stated below by considering age as first factor and rating as second factor. The ANODEV table for this problem looks as below.

TABLE 4.18

Model Description	d.f.	Deviance	First differenc
Null	19	94.6068	
A	15	73.7673	20.8395
A+R	12	32.4571	41.3102
A+R+A.R	0	0	32.4571

Foot Note :- 'A' and 'R' respectively denote the factors age and rating.

The conclusions can be drawn by comparing the differences with the table chi square value for suitable degrees of freedom. If the log linear model is viewed in different angle, since the Poisson distribution is a member of one parameter natural exponential family, the thory of generalised linear model will be straight way applicable to the log linear model. In the next section we discuss how this approach can be applied.

4.9 Log-linear model (Nelder & Wedderburn (1972)) :

Suppose the response variate Y has Poisson distribution with mean μ. Here we discuss log linear model for two dimensional contingency table only. Now suppose Y_{ij} ($i \in A_1$, $j \in A_2$) be the n realisations on the response variate Y. Reindex these observations as Y_i ($i=1,2,\ldots,n$) such that first J observations corresponds to the first sample, next J to the second sample and so on. In the same way reindex the parameters as μ_i ($i=1,2,\ldots,n$) so that first J components of the vector $\underline{\mu}$ are μ_1^*, next J components are μ_2^* and so on. Here

$$\mu_i^* = (1/J)\sum_j \mu_{ij} \; ; \text{ for } i=1,\ldots,I. \qquad (1)$$

As we are assuming Poisson distribution to the responses, one of the possible link function is 'log-link' function. This link function is given by,

$$T = ln(\mu). \qquad (2)$$

Now we give below the definition of log linear model.

Definition-5 : Log linear model : Suppose the response variate Y is having Poisson distribution with parameter μ. Suppose Y_i ($i=1,2,\ldots,n$) are n realisations of the response variate Y. Let $\underline{x}_j$ be the vectors of the known values of the covariates X_j ($j=1,2,\ldots,k$) and $\underline{T} = X\underline{\beta}$ be the linear predictor.Then the model of the form,

$$T_i = ln(\mu_i), \text{ for } i=1,2,\ldots,n, \qquad (3)$$

with $E(Y_i) = \mu_i$, is called log-log model.

Fitting a log-log model :

The equation (3) gives

$$\mu_i = \exp(T_i). \qquad (4)$$

Differentiating equation (3) w.r.t. μ_i, we have

$$(dT_i/d\mu_i) = 1/\mu_i \; , \; i = 1,2,\ldots,n \qquad (5)$$

Now equations (3.4-19) and (5) gives i^{th} component of $\underline{Z}$ as

$$\hat{Z}_i = \hat{T}_i + \left\{\frac{(y_i - \hat{\mu}_i)}{\mu_i}\right\}. \qquad (6)$$

Similarly, from equations (3.4-18) and (6) it is clear that,

$$\hat{W}_{ii} = \hat{\mu}_i. \qquad (7)$$

Once the expressions for Z_i and W_{ii} are obtained, a log-linear model can be fitted by using method described in sub section (3.4.3). The initial estimates $\hat{\underline{\mu}}^{(0)}$ of $\underline{\mu}$ can be taken as $(1/J)\sum_j(y_{ij})$. If all the observations in any sample are equal to zero, take the corresponding sample total as 0.5.

A FORTRAN-77 program given in Appendix-2(B) is useful to estimate $\underline{\beta}$ under this model.

Example 4.3 (Cont.):-

Data in table (4.22) is useful to study the relationship between stimulus variates X_1, X_2 and the response variate Y. We can fit four types of log-linear models stated below, to the data.

Model (13) : $E(Y_i) = \exp(\beta_0)$. (8)

Model (14) : $E(Y_i) = \exp(\beta_0 + \beta_1 X_{i1})$. (9)

Model (15) : $E(Y_i) = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})$ (10)

Model (16) : $E(Y_i) = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2})$. (11)

Once the model is fitted, it can be used to estimate the values of Y_i, under the respective model.

Data in the table (4.22) is useful to study the relationship between the number of dreams (Y) and age group of boys(X_1). If the rating factor (X_2) is taken as another explanatory variable, the relationship between Y and X_1, X_2 can be studied. Here none

of the covariate is at two levels. Let Y_i be the number of dreams for boys in different age groups. Then four log linear models for Y_i are as stated in models (13) to (16).

Executing program in appendix-2(B), to fit the models (13) to (16) to the data from example (4.3), parameter estimates obtained are given in the tables (4.19) to (4.22). At the bottom of table deviance is given.

TABLE-4.19

Sr.No.	Estimate	Parameter
1	2.4114	INTERCEPT

Deviance = 94.6075

TABLE-4.20

Sr.No.	Estimate	Parameter
1	2.3955	INTERCEPT
2	0.1264	AGE

Deviance = 87.551

TABLE-4.21

Sr.No.	Estimate	Parameter
1	3.1637	INTERCEPT
2	-0.1264	AGE
3	-0.3349	RATING

Deviance = 57.7147

TABLE-4.22

Sr.No.	Estimate	Parameter
1	3.0643	INTERCEPT
2	0.5155	AGE
3	-0.3061	RATING
4	-0.1836	AGE.RATING

Deviance = 40.7375

Now the analysis of deviance table can be prepared as below

ANODEV TABLE

TABLE 4.23

Model Description	d.f.	Deviance	First difference
Null	19	94.6075	
AGE	18	87.5510	7.0565
AGE+RATING	17	57.7147	29.8363
AGE+RATING+A.R	16	40.7375	16.9772

Comparing the first difference with the table value of chi-square with appropriate d.f. conclusions can be drawn.

Difference between the procedures of fitting log linear model by the two different approaches is Bishop's approach do not require the 'dummy' covariates where as in Nelder's approach the dummy covariates should be defined and their values though nominal should be provided.

While fitting generalised linear model to the data we are assuming that the response variate has a particular distribution from one parameter natural exponential family. Many times it is not possible to specify the underlying distribution completely. In such cases generalised linear model cannot be fitted. In the next chapter we discuss the models useful in such situations.