3. GENERALISED LINEAR MODELS

3.1 Introduction :

In classical linear model, least square theory has been used to analyse the data, where the responses are assumed to have constant variance and the systematic effects are linearly related to the mean of response variate. Then to test different hypotheses about the model parameters β normality is assumed. In section (2.7) it has pointed out that, classical linear models are not suitable in all situations. This is because, there are many real life situations where at least one of the assumptions of constant variance and linearity of systematic effects may not hold good. As an example consider the following real life case.

Example 3.1 : An experiment was carried out on the group of 20 students, to compare three teaching methods. The students are divided into three groups and different teaching methods are applied to the various groups. At the end of the course, a test is conducted to check the performance of these students. The marks (out of 20) obtained by them are given below.

TABLE-3.1

Teaching method	1	2	3	4	5	6	7
I	19	0	5	10	4	17	13
II	6	12	11	9	10	13	12
III	9	8	9	7	8	7	

Below the graph of marks against teaching methods is plotted.

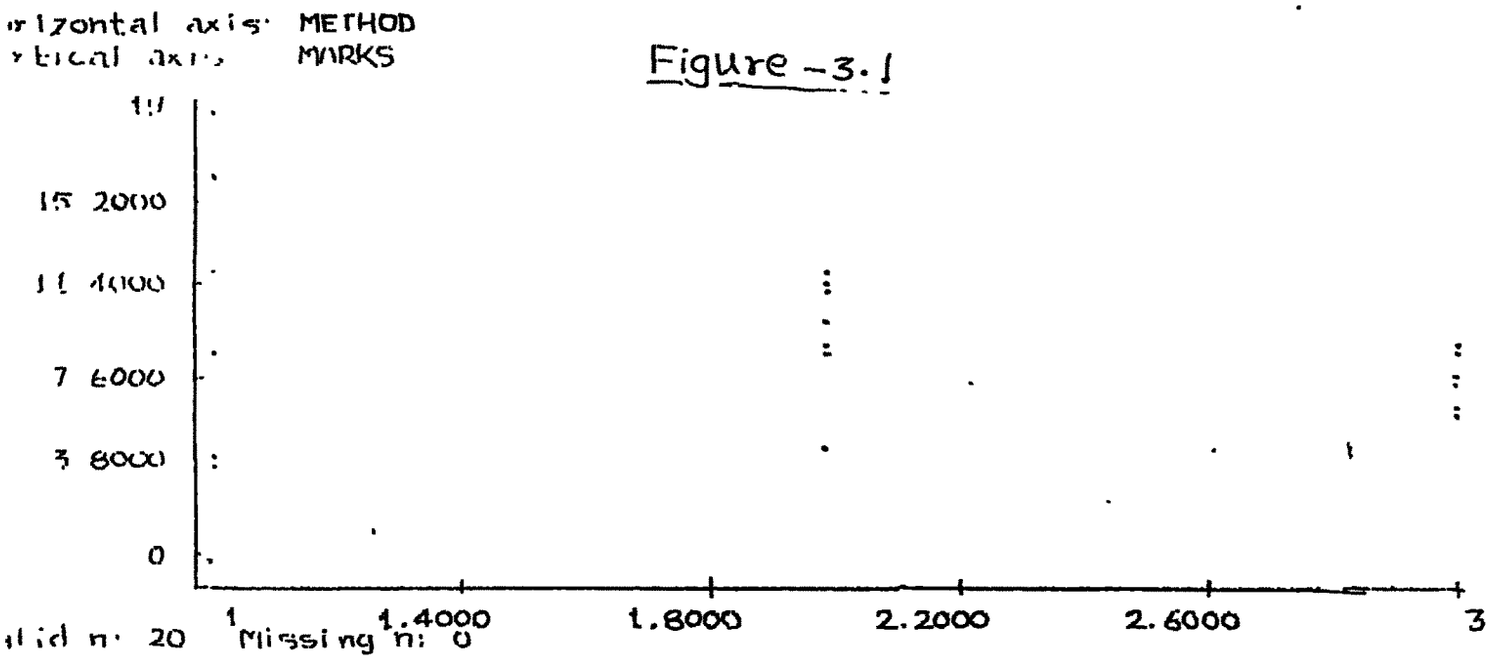

Figure -3.1

The graph in Fig(3.1) shows that, the variance is not constant. Hence some variance stabilising transformation is required. Thus the first assumption of constant variance required by least square method fails. Therefore least square method can not be used to the untransformed data. In other words, classical linear model is not proper to this data. Hence it becomes necessary to search for some other models appropriate for such type of data. Alternative techniques have been developed to analyse such type of data. They are probit analysis (see e.g. Finney (1947), Prentice (1976)), logistic regression analysis (see e.g. Prentice (1976), Hosmer & Lemeshow (1989), Collett (1991)) for Binomial data; fitting log-linear models (e.g. Bishop & others (1975), Good & Kurskal (1964), Gokhale & Kullback (1978)) for the data in the form of contingency tables.

By studying similarities of many model fitting methods involving linear combinations of the parameters Nelder and Wedderburn (1972) developed the new class of models, namely, 'generalised linear models'(GLMs). This class of generalised linear models includes all the above types of models. McCullagh & Nelder (1983), Dobson (1989) have discussed the theory associated with generalised linear model. Breslow & Day (1980)

described the applications of generalised linear models in 'cancer research'. In this chapter we discuss explicitly the theory associated with generalised linear model and the illustrations are given in the subsequent chaptersr.

Here discussion is made in the following direction.

(1) Description of a generalised linear model with k explanatory variates,

(2) link function,

(3) fitting of generalised linear model,

(4) measures of adequacy of the fitted model,

(5) analysis of deviance (ANODEV),

(6) model checking for generalised linear model,

(7) method of obtaining 'robust' estimates of the model parameters in generalised linear model,

(8) generalised linear model with varying dispersion,

(9) fitting of generalised linear model with varying dispersion.

3.2: Generalised linear model with k explanatory variates :

The definition of generalised linear model requires the term 'one parameter natural exponential family'. But before describing this family of distributions, it is essential to explain in brief, the meaning of 'One parameter exponential family'. For this suppose,

(i) S ($\subseteq \mathbb{R}$) is the range of response variate Y, giving positive p.d.f.,

(ii) θ^* is parameter of the distribution of Y,

and

(iii) Θ^* ($\subseteq \mathbb{R}$) is the parameter space of θ^*.

Definition-1 : One parameter exponential family :A class of p.d.f.s(or p.m.f.s) depending on a real valued parameter θ^* of

the following form

$$f(y;\theta^*,\phi)=\exp\left\{\alpha(\phi)[a(y)b(\theta^*)-g(b(\theta^*))+h(y)]+\xi(\phi;y)\right\}I_S(y),$$
$$\theta^* \in \Theta^*;$$

or euivalently,

$$f(y;\theta,\phi) = \exp\left\{\alpha(\phi)[a(y)\theta - g(\theta) +h(y)] +\xi(\phi;y)\right\}I_S(y), \quad (1)$$
$$\theta^* \in \Theta^*;$$

is known as one parameter exponential family, if S does not depend on θ^*.

Here

(i) $I_S(Y) = \begin{cases} 1; \text{ if } Y \in S \\ 0; \text{ otherwise} \end{cases}$,

(ii) $\theta = b(\theta^*)$ is a function of θ^* only,

(iii) $\alpha(\phi)$ is the function of ϕ alone and having positive value,

(iv) $a(Y)$ and $h(Y)$ are the functions of Y only,

(v) $g(\theta)$ is the function of θ only,

(vi) $\xi(\phi;Y)$ is the function free from θ.

Note : (1) If $\Theta^* \subseteq \mathbb{R}^k$, distribution of Y is said to be a member of k-parameter exponential family.

(2) If the distribution of response variate Y, belongs to one parameter exponential family, the log likelihood of θ is given by,

$$\ell(\theta,\phi;y) = \{\alpha(\phi)[a(y)\theta - g(\theta) + h(y)] + \xi(\phi;y)\}. \quad (2)$$

Now, we give few illustrations for understanding purpose.

Illustration 1 : Normal distribution :

Let the response variate Y has normal distribution $N(\mu,\sigma^2)$

($\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$). Then the p.d.f. of Y is given by,

$$f(y;\mu,\sigma^2) = (2\pi\sigma^2)^{-(1/2)} \exp\{(y-\mu)^2/(2\sigma^2)\}, \quad y \in \mathbb{R}, \tag{3}$$

which gives the log likelihood function based on single observation as

$$\ell(\mu,\sigma^2;y) = -[\ln(2\pi\sigma^2)]/2 - (y^2 - 2\mu y + \mu^2)/(2\sigma^2). \tag{4}$$

The equation (4) can be rewritten as

$$\ell(\mu,\sigma^2;y) = [\mu y - (\mu^2/2) - (y^2/2)]/\sigma^2 - \ln(2\pi\sigma^2)/2. \tag{5}$$

On comparing the equation (5) with the equation (2) the following findings can be obtained.

$$\left.\begin{aligned} &\phi = \sigma^2, \quad \alpha(\phi) = (1/\phi), \quad \theta = \mu, \quad g(\theta) = \theta^2/2, \\ &a(y) = y, \quad h(y) = -y^2/2, \quad \xi(\phi;y) = -[\ln(2\Pi/\alpha(\phi))]/2. \end{aligned}\right\} \tag{6}$$

Since the range of Y does not depend on the distribution parameter, the $N(\mu,\sigma^2)$ distribution belongs to one parameter exponential family. Here σ^2 is the nuisance parameter.

Illustration 2. Beta distribution of first kind :

Suppose the response variate Y has beta distribution of first kind with parameters ($\mu,\nu>0$). Therefore, the p.d.f. of Y is given by,

$$f(y;\mu;\nu) = \{\Gamma(\mu)\Gamma(\nu)/\Gamma(\mu+\nu)\}\, y^{(\mu-1)}(1-y)^{(\nu-1)} I_{[0,1]}(y),$$

where,

$$I_{[0,1]}(Y) = \begin{cases} 0 \ ; \text{ if } Y \in [0,1] \\ 1 \ ; \text{ otherwise.} \end{cases}$$

Hence, the log likelihood of θ is,

$$\ell(\theta,\phi;y) = \ln(c) + (\mu-1)\ln(y) + (\nu-1)\ln(1-y), \tag{7}$$

where,

$$c = \{\Gamma(\mu)\Gamma(\nu)/\Gamma(\mu+\nu)\}.$$

From the equations (7) and (2), it can be seen that,

$$\left.\begin{aligned} \phi = \alpha(\phi) = 1,\ \underline{\theta} = [(\mu-1)\ (\nu-1)]',\ g(\underline{\theta}) = c \\ \underline{a}(y) = [\ln(y)\ \ \ln(1-y)]',\ h(y) = \xi(\phi;y) = 0 \end{aligned}\right\} \quad (8)$$

As the range of Y is free from parameters of the distribution, equation (8) indicate that, beta distribution of first kind belongs to two parameter exponential family.

Remark : If a(y) = y, the p.d.f. of Y given in equation (1) is said to be in the canonical form, and 'θ' is known as canonical parameter.

Moris (1982), has named the class of distributions, having p.d.f. in the canonical form, as natural exponential family. Therefore, 'One parameter natural exponential family' can be defined as below.

Definition-2 : One parameter natural exponential family :A class of p.d.f.s(or p.m.f.s) depending on a real valued parameter θ^* of the following form

$$f(y;\theta,\phi) = \exp\left\{\alpha(\phi)[y\theta - g(\theta) + h(y)] + \xi(\phi;y)\right\} I_S(y), \quad (9)$$

$$\theta = b(\theta^*) \text{ and } \theta^* \in \Theta^*;$$

is known as one parameter natural exponential family, if S does not depend on θ^*. The parameter θ is called as 'natural parameter.

Note : If the distribution of response variate Y belongs to one parameter natural exponential family, the log likelihood of θ is given by,

$$\ell(\theta,\phi;y) = \{\alpha(\phi)[\, y\theta - g(\theta) + h(y)] + \xi(\phi;y)\}. \qquad (10)$$

Note that the natural exponential family is a sub class of exponential family. Further, it can be seen that $N(\mu,\sigma^2)$ distribution is a member of one parameter natural exponential family. On the other hand beta distribution of first kind is not a member of this class.

Below some illustrations are given to have the idea of natural exponential family more clear.

Illustration 1 : Gamma distribution :

Let the response variate Y has gamma distribution with parameters (ν,μ), $(\nu,\mu > 0)$ p.d.f.

$$f(y;\mu,\nu) = \{[1/\Gamma(\nu)](\nu/\mu)^{\nu}\}\exp(-\nu y/\mu)y^{\nu-1}I_{[0,\infty)}(y), \qquad (11)$$

where

$$I_{[0,\infty)}(Y) = \begin{cases} 1 \; ; \text{ if } Y \in [0,\infty) \\ 0 \; ; \text{ otherwise.} \end{cases}$$

Thus the log likelihood function based on single observation is

$$\ell(\mu,\nu;y) = -\ln\Gamma(\nu) - (\nu y/\mu) + (\nu-1)\ln(y) + \nu\ln(\nu/\mu). \qquad (12)$$

Rearrangement of the terms in equations (12) gives

$$\ell(\mu,\nu;y) = \nu\,[-(y/\mu)-\ln(\mu)] - \ln(\Gamma(\nu)) + \nu\ln(\nu) + (\nu-1)\ln(y). \qquad (13)$$

Comparison of the equations (13) and (10) gives

$$\left.\begin{aligned} &\phi = (1/\nu),\ \alpha(\phi) = 1/\phi,\ \theta = -1/\mu, \quad g(\theta) = \ln(-1/\theta), \\ &a(y) = y,\ h(y) = 0,\ \xi(\phi,y) = -\ln(\Gamma(\phi))+\phi\ln(\phi)+(\phi-1)\ln(y) \end{aligned}\right\}. \qquad (14)$$

Since the range of Y is free from the parameters (μ,ν), the distribution of Y is a member of one parameter natural exponential family.

Illustration 2 : Poisson distribution :

Let Y be the response variate having Poisson distribution with parameter $(\lambda > 0)$ Then the p.m.f.of Y is given by

$$f(y;\lambda) = \begin{cases} \exp(-\lambda)\lambda^{y}/y!; & y = 0,1,\ldots \\ 0; \text{ otherwise} \end{cases} \qquad (15)$$

Hence the log likelihood function based on single observation y becomes

$$\ell(\lambda;y) = -\lambda + y \ln(\lambda) - \ln(y!) \qquad (16)$$

Comparing the equation (16) with the equation (10) we have,

$$\left.\begin{aligned} \phi = a(\phi) = 1, \quad \theta = \ln(\lambda), \quad g(\theta) = \exp(\theta), \\ a(y) = y, \; h(y) = 0, \; \xi(\phi;y) = -\ln(y!) \end{aligned}\right\} \qquad (17)$$

Therefore, it can be observed that, Poisson distribution belongs to one parameter natural exponential family.

Illustration 3 : Binary distribution for grouped data :

Let response variate Y be such that $Y^{*}=m^{*}Y$ has Binomial distribution $B(m^{*},p)$. Let μ denote the mean. Note that for grouped binary distribution $\mu = p$. Then the p.m.f. of Y is given by

$$f(y;m^{*},\mu) = m^{*}\left\{ {}^{m^{*}}C_{m^{*}y}\, \mu^{m^{*}y}(1-\mu)^{(m^{*}-m^{*}y)} I_{A_1}(y) \right\} \qquad (18)$$

Where

(i) $A_1 = \{0, 1/m^{*}, \ldots, 1\}$

and

(ii) $I_{A_i}(y) = \begin{cases} 1; & y \in A_i, \\ 0; & \text{otherwise.} \end{cases}$

Thus, log-likelihood function based on single observation is

$$\ell(m^*,\mu;y) = \ln\left\{ {}^{m^*}C_{m^*y} \right\} + m^*y\ln(\mu) + m^*(1-y)\ln(1-\mu) + \ln(m^*) \tag{19}$$

The rearrangment of the terms in equation (19) gives

$$\ell(m^*,\mu;y) = m^*[y\ln[\mu/(1-\mu)] + \ln(1-\mu)] + \ln\left\{ {}^{m^*}C_{m^*y} \right\} + \ln(m^*) \tag{20}$$

On comparing the equations (10) and (20), we get,

$$\left.\begin{aligned} &\phi=1/m^*, \alpha(\phi)=1/\phi, \quad \theta = \ln(\mu/(1-\mu)), \quad g(\theta) = \ln(1+\exp(\theta)), \\ &a(y)=y, \ h(y) = 0, \ \xi(\phi,y) = \ln[\,{}^{m^*}C_{m^*y}\,] + \ln(m^*). \end{aligned}\right\} \tag{21}$$

Thus we conclude that this distribution belongs to one parameter natural exponential family.

Results of the above given distributions belonging to one parameter natural exponential family can be summarised as in the following table.

TABLE 3.2

Distribution of Y	Normal	Gamma	Poisson	Grouped binary
Range of Y	$(-\infty,\infty)$	$[0,\infty)$	$0,1,\ldots$	$0,1/m^*,\ldots,1$
ϕ	σ^2	$1/\nu$	1	$1/m^*$
$\alpha(\phi)$	$1/\phi$	$1/\phi$	1	$1/\phi$
θ	μ	$-1/\mu$	$\ln(\mu)$	$\ln(\mu/(1-\mu))$
$g(\theta)$	$\theta^2/2$	$\ln(-1/\theta)$	$\exp(\theta)$	$\ln(1+\exp(\theta))$
$h(y)$	$-y^2/2$	0	0	0
$\xi(\phi,y)$	$-\ln\left[\frac{2\Pi}{\alpha(\phi)}\right]/2$	$-\ln(\Gamma(\phi))$ $+\phi\ln(\phi)$ $+(\phi-1)\ln(y)$	$-\ln(y!)$	$\ln\left\{{}^{m^*}C_{m^*y}\right\}$ $+\ln(m^*)$

A very important result which holds for one parameter natural exponential family is proved below.

Result-1: If the dependent variate Y comes form one parameter natural exponential family with natural parameter 'θ' and nuisance parameter 'ϕ', then

$$\left.\begin{aligned} E(Y) &= g'(\theta) \\ Var(Y) &= g''(\theta)/\alpha(\phi) \end{aligned}\right\} \qquad (22)$$

Proof : Suppose Y is the single observation on the response variate with mean μ. Assume that Y comes from one parameter natural exponential family. Hence, the log likelihood of θ can be written as,

$$\ell = \ell(\theta,\phi;y) = \{\alpha(\phi)[y\theta - g(\theta) + h(y)] + \xi(\phi;y)]\}. \qquad (23)$$

It is well known that (See e.g. Kendall & Stuart (1968),

$$E(\partial \ell / \partial \theta) = 0, \tag{24}$$

and

$$E(\partial \ell / \partial \theta)^2 = -E(\partial^2 \ell / \partial \theta^2). \tag{25}$$

Differentiating equation (23) w.r.t. θ twice, we get,

$$\partial \ell / \partial \theta = \alpha(\phi)[y - g'(\theta)] \tag{26}$$

and

$$\partial^2 \ell / \partial \theta^2 = \alpha(\phi)[-g''(\theta)]. \tag{27}$$

The equations (24) and (26) combinedly imply

$$\mu = g'(\theta); \tag{28}$$

and hence

$$\partial \ell / \partial \theta = \alpha(\phi)[y - \mu]. \tag{29}$$

From equations (25), (27) and (29), it can be seen that

$$g''(\theta) = \alpha(\phi)[\text{var}(Y)] = V. \tag{30}$$

This result is useful in developing algorithm for fitting generalised linear model (GLM), whose definition is given below.

Definition-3 : Generalised linear model :

Let Y be a response variate with p.d.f. (or p.m.f.) $f(.;\theta^*)$ which belongs to the one parameter natural exponential family with θ as a natural parameter.

Suppose Y_i $(i=1,2,\ldots,n)$ are n independent observations on the response variate Y. Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_k$ be the vectors of known values of the covariates X_j $(j=1,2,\ldots,k)$. Let

$$T_i = \beta_0 + \sum_j x_{ij}\beta_j, \text{ for } i=1,2,\ldots,n, \tag{31}$$

where

$\underline{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ is the vector unknown model parameters and

$\underline{T} = (T_1, T_2, \ldots, T_n)'$ is the linear predictor.

Then a class of models of the form,

$$\underline{T} = m(E(\underline{Y})); \qquad (32)$$

where m(.) is a strictly monotoic differentiable function is called the class of 'Generalised Linear Models' (GLMs).

Since '$\underline{T}$' is a linear sum of the effects of explanatory variates, it is called as 'linear predictor'. Further, since the function m(.) gives the relationship between $E(\underline{Y} | \underline{X} = \underline{x})$ and $\underline{x}$, Nelder & Wedderburn (1972) named this function as 'link function'.

We now show that classical linear model is a particular case of generalised linear model. In classical linear model we assume normal distribution for responses. It is already shown that normal distribution is a member of one parameter natural exponential family. Further, for classical linear model, we have

$$E(\underline{Y}) = X\underline{\beta} = \underline{T}.$$

From the definition of generalised linear model, it is clear that, classical linear model is a special case of generalised linear model with 'identity' link function. By the 'identity' link function we mean the link, $\underline{T} = \underline{\mu}$. Thus the generalised linear model allows two extensions, namely, distribution of the response variate may be any distribution from natural exponential family, and the link function m(.) may be any strictly monotonic differtiable function.

As we have seen earlier, the functional relationship between

Y and $\underline{X}$ is known as link function. Along with the likelihood function, this function is also required for fitting generalised linear model to the data. Thus, link function plays a vital role in the theory of generalised linear model. Hence, it is discussed explicitly in the next section.

3.3 : Link function

Generalised linear model is fitted to the data through reparameterisation process. It should be done in such a way that the new parameters gives linear relationship between $E(Y|\underline{X}=\underline{x})$ and $\underline{x}$. The function which relates the covariates $(X_1, X_2, \ldots, X_k)'$ and the new parameters β with $E(Y)$ is known as link function. In other words, it is a function which gives relationship between linear predictor T and expected value μ of the response variate Y. Since μ completely depends on the behaviour of the responses, and values of the stimulus variates are fixed, the function m(.) gives link between random component and systematic components.

In classical linear models we have an identity link, $T = \mu$. This type of link is suitable in classical linear models, because both T and μ can take any value on real line. Here it should be noted that such type of link is not always suitable. Now we illustrate this by considering two situations and suggest appropriate links in each case.

Suppose response variate Y has Poisson distribution with parameter μ. Here identity link is not appropriate, since the linear predictor T may be negative, whereas μ is strictly positive. So the link function m(.) must be such that $m(.) : (0,\infty) \rightarrow (-\infty,\infty)$. One such function is $\ln(\mu)$. Thus one appropriate link in this situation is log link given by $T=\ln(\mu)$.

Secondly, assume that the response variate Y is such that

$Y^* = nY$ has binomial distribution $B(n,\mu)$. Here also identity link is not proper, as value of the parameter μ is in between $(0,1)$. Hence the link function $m(.)$ should map from $(0,1)$ to $(-\infty,\infty)$. Three important functions of this type are mentioned below.

(i) Logit function, $T = \ln\left\{\frac{\mu}{1-\mu}\right\}$;

(ii) complementary log-log function, $T = \ln(-\ln(1-\mu))$;

(iii) probit function, $T = \Phi^{-1}(\mu)$; where $\Phi(.)$ is the normal cumulative distribution function.

Remark : The link $T = \theta$, is known as canonical link. Thus, 'identity link' for normal distribution, 'log link' for Poisson responses and 'logistic link' for binomial distribution are some illustrations of canonical links.

After discussing the link function explicitly, we are in a position to fit generalised linear model to the data.

3.4 : Fitting of GLM

Fitting the model means estimating unknown parameters in the model. Parameters in the model can be estimated by using different methods. Some of the important methods of fitting generalised linear model to the data are

(I) weighted least square method;

(II) method of maximum likelihood;

and

(III) method of obtaining robust estimates.

The third method we discuss at later stage of the chapter in section (3.9). Below we discuss the first two methods explicitly.

3.4.1 : Weighted least square method :

In least square method of estimation, estimates of the model parameters are obtained under the following two assumptions.

(i) Error components in the model are independently

distributed with mean zero and constant variance σ^2.

(ii) The identity link holds.

In case of generalised linear models, other than classical linear model, the identity link does not hold. In other words, the systematic effects are not linearly related with the means of original responses $Y_i (i=1,2,\ldots,n)$. To overcome this problem, Nelder & Wedderburn (1972) defined the new variates $Z_i (i=1,2,\ldots,n)$ as,

$$Z_i = T_i + (Y_i - \mu_i)(dT_i/d\mu_i), \qquad (1)$$

so that $E(\underline{Z}) = X\beta$. This implies that, if the new dependent variable Z is considered instead of Y, the identity link is suitable for Z. Hence the generalised linear model in terms of $\underline{Z}$ can be written as,

$$\underline{Z} = X\beta + \underline{e}\ , \qquad (2)$$

with

$$E(\underline{e}) = \underline{0},$$

and

$$Var(\underline{e}) = diag(var(e_1),\ var(e_2),\ldots,var(e_n)). \qquad (3)$$

Consider

$$\begin{aligned} var(e_i) &= var\left[(Y_i-\mu_i)(dT_i/d\mu_i)\right] \\ &= (dT_i/d\mu_i)^2\ var(Y_i) \\ &= \left\{V_{ii}(\mu_i)/[\alpha(\phi)1]\right\}(dT_i/d\mu_i)^2 \\ &= (\alpha(\phi)W_{ii})^{-1}\ (say), \end{aligned}$$

where

$$W_{ii} = \left\{[V_{ii}(\mu_i)]\,(dT_i/d\mu_i)^2\right\}^{-1}. \qquad (4)$$

Define

$$W = \text{diag}(W_{11}, W_{22}, \ldots, W_{nn}). \tag{5}$$

Premultiplying by $W^{1/2}$ to the model (2), we have

$$\underline{Z}^* = X^* \underline{\beta} + \underline{e}^*, \tag{6}$$

with

$$\underline{Z}^* = W^{(1/2)} \underline{Z} ,$$

$$X^* = W^{(1/2)} X ,$$

and

$$\underline{e}^* = W^{(1/2)} \underline{e}.$$

Since $\alpha(\phi)$ is assumed to be constant, for the model (6), one can obtain least square estimates of the model parameters $\underline{\beta}$ by using well known normal equations,

$$(X^{*\prime} X^*) \underline{\beta} = X^{*\prime} \underline{Z}^*,$$

or equivaletly,

$$(X' W X) \underline{\beta} = X' W \underline{Z}. \tag{7}$$

Since in generalised linear model, the newly defined variate Z and the weight matrix W both depend on the estimates of model parameters $\underline{\beta}$, it is necessary to use the weighted least square method iteratively. Here it is essential to note that, as the newly defined dependent variate Z is unobservable, it is not possible to obtain weighted least square estimates of $\underline{\beta}$ without making any assumption about the distribution of response variate. Hence, for fitting generalised linear model, this method can be used only after making a valid distributional assumption for the response variate Y. Once we assume a particular distribution

from one parameter natural exponential family, this method can be used.

Note :- In case of classical linear model we have

$$E(Y) = \mu \text{ and } V(\mu) = 1. \tag{8}$$

Therefore we have from equations (1) and (4)

$$Z = Y \text{ and } W = 1 \tag{9}$$

Hence for n observations equations (7) reduces to

$$(X'X)\underline{\beta} = X'\underline{Y}, \tag{10}$$

which are the normal equations given in equation (2.3-6). Thus in case of classical linear model, weighted least square method is equivalent to the usual least square method.

Now we discuss the second method.

3.4.2 : Method of maximum likelihood :

In the class of generalised linear models, distribution of the response variate is a member of one parameter natural exponential family. Hence, the method of maximum likelihood can be used to estimate the model parameters $\underline{\beta}$.

Suppose $Y_i (i=1,2,\ldots,n)$ are n independent responses with p.d.f. (or p.m.f.) of Y_i is given by,

$$f(y_i;\theta_i,\phi) = \exp\{\alpha(\phi)[y_i\theta_i - g(\theta_i) + h(y_i)] + \xi(\phi;y_i)]\}.$$

Therefore, the log likelihood of $\underline{\theta}$ based on n observations is given by,

$$\ell(\underline{\theta},\phi;\underline{Y}) = \sum_i \{\alpha(\phi)[y_i\theta_i - g(\theta_i) + h(y_i)] + \beta(\phi_i,y_i)]\}. \tag{11}$$

In order to obtain maximum likelihood estimates of $\underline{\beta}$, we differentiate the expression (11) and set the results equal to zero. Differentiation of expression (11) can be obtained by using chain rule. According to this rule, we can write

$$\partial \ell / \partial \beta_j = \sum_i (\partial \ell / \partial \theta_i)(d\theta_i / d\mu_i)(d\mu_i / dT_i)(\partial T_i / \partial \beta_j). \qquad (12)$$

From equation (11) it can be observed that,

$$(\partial \ell / \partial \theta_i) = \alpha(\phi)[\, y_i - g'(\theta_i)] \qquad (13)$$

Further, from the result (3.2-22) we have,

$$\left.\begin{aligned} g'(\theta_i) &= \mu_i \\ g''(\theta_i) &= V_{ii}(\mu_i) \end{aligned}\right\}. \qquad (14)$$

Also, we have

$$T_i = \sum_j x_{ij}\beta_j \,, \quad i=1,2,..,n; \quad x_{io} = 1 \text{; for all } i.$$

Hence,

$$\partial T_i / \partial \beta_j = x_j \text{ ; } j=0,1,\ldots,k. \qquad (15)$$

Using equations (13) to (15), in (12) we get the estimating equations as

$$T_j^* = \sum_i \left\{ \frac{\alpha(\phi)[\, y_i - \mu_i]}{V_{ii}(\mu_i)} \frac{d\mu_i}{dT_i} x_{ij} \right\} = 0 \qquad (16)$$

$$j=0,1\ldots,k.$$

Alternatively, the above equations (16) can be written in the matrix form as,

$$D' C^{-1} (\underline{y}-\underline{\mu}) = \underline{0} \,, \qquad (17)$$

where

(i) $D = (D_{ij})$, with $D_{ij} = (d\mu_i / d\beta_j)$ for every i and j value,

(ii) C is the variance covariance matrix of $\underline{Y}$.

After solving equations (17) we obtain estimates of β_j $(j=0,1,\ldots,k)$.

Since Y_i's $(i=1,2,\ldots,n)$ are independent,

$$C = \text{diag}(\text{var}(Y_1), \text{var}(Y_2), \ldots, \text{var}(Y_n)). \qquad (18)$$

It can be observed that in case of classical linear model, the estimating equations (17) reduces to the normal equations given in (2.3-6). This can be justified as follows.

Justificatin : For classical linear model, we have,

$$\underline{T} = \underline{\mu}, \qquad (19)$$

and,

$$C = \sigma^2 I_n,$$

where $\underline{T} = X\beta$ and $\underline{\mu} = E(\underline{Y})$.

Equation (19) implies, $D = X$, and $\underline{\mu} = X\beta$. Hence for classical linear model, the estimating equations (17) become,

$$X'(\underline{y}-X\beta) = \underline{0}.$$

i.e.

$$(X'X)\beta = X'\underline{y}, \qquad (20)$$

which are the normal equations as given in equations (2.3-6). Since the normal equations in (20) are linear equations in β, the estimates of β can be obtained by solving them.

For non normal distributions, the equations (17) are non linear eqations in β. Therefore, it may not be possible to obtain the maximum likelihood estimates explicitly. However by using appropriate numerical technique, the estimates of β can be obtained.

For the class of generalised linear models defined by Nelder & Wedderburn (1972), the approach of Newton-Raphson method is

used to compute maximum likelihood estimates. Nelder & Wedderburn (1972) introduced this method. Later McCullagh & Nelder (1983) reproduced the same with more explanation. We also use the Newton-Raphson method as a numerical technique to solve the equations (17). According to this method m^{th} approximation of the estimates of $\underline{\beta}$ is given by

$$\hat{\underline{\beta}}^{(m)} = \hat{\underline{\beta}}^{(m-1)} - \left[\underline{T}^{*(m-1)}\right]_{\underline{\beta}=\hat{\underline{\beta}}^{(m-1)}} \cdot \left[-\frac{\partial^2 \ell}{\partial\beta_j \partial\beta_l}\right]^{-1}_{\underline{\beta}=\hat{\underline{\beta}}^{(m-1)}} ; \qquad (21)$$

where $\left[-\frac{\partial^2 \ell}{\partial\beta_j \partial\beta_l}\right]$ is the matrix of second order derivatives of ℓ. Hence to solve the equations (21) it is essential to obtain the expressions for $\left[-\frac{\partial^2 \ell}{\partial\beta_j \partial\beta_l}\right]$. For this consider,

$$\begin{aligned}
-\frac{\partial^2 \ell}{\partial\beta_j \partial\beta_l} &= -\frac{\partial}{\partial\beta_l}(\partial\ell/\partial\beta_j) \\
&= \sum_i -\frac{\partial}{\partial\beta_l}\left[\frac{\partial\ell}{\partial T_i} \cdot \frac{\partial T_i}{\partial\beta_j}\right] \\
&= \sum_i -\frac{\partial}{\partial\beta_l}\left[\frac{\partial\ell}{\partial T_i} \cdot x_{ij}\right] \\
&= \sum_i -\frac{\partial}{\partial T_i}\left[\frac{\partial\ell}{\partial T_i} \cdot \frac{\partial T_i}{\partial\beta_l}\right] x_{ij} \\
&= \sum_i -\frac{\partial^2 \ell}{\partial T_i^2} \cdot (x_{ij} x_{il}). \qquad (22)
\end{aligned}$$

Consider,

$$\begin{aligned}
-\frac{\partial^2 \ell}{\partial T_i^2} &= -\frac{\partial}{\partial T_i}\left[\frac{\partial\ell}{\partial\theta_i} \cdot \frac{d\theta_i}{dT_i}\right] \\
&= \frac{\partial\ell}{\partial\theta_i} \cdot \frac{d^2\theta_i}{dT_i^2} + \frac{d\theta_i}{dT_i} \cdot \frac{\partial}{\partial\theta_i}\left[\frac{\partial\ell}{\partial T_i}\right]
\end{aligned}$$

$$= \frac{\partial \ell}{\partial \theta_i} \cdot \frac{d^2\theta_i}{dT_i^2} + \frac{d\theta_i}{dT_i} \cdot \frac{\partial}{\partial \theta_i}\left[\frac{\partial \ell}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial T_i}\right]$$

$$= \frac{\partial \ell}{\partial \theta_i} \cdot \frac{d^2\theta_i}{dT_i^2} + \left[\frac{d\theta_i}{dT_i}\right]^2 \cdot \left[\frac{\partial^2 \ell}{\partial \theta_i^2}\right]. \qquad (23)$$

Since the response variate is assumed to have a distribution from one parameter natural exponential family, we get from the equations (3.2-27) and (3.2-28),

$$\partial^2 \ell / \partial \theta_i^2 = -\alpha(\phi) V_{ii}(\mu_i). \qquad (24)$$

$$d\mu_i / d\theta_i = V_{ii}(\mu_i). \qquad (25)$$

Using equations (3.2-29) and (24), equation (23) can be rewritten as,

$$-\frac{\partial^2 \ell}{\partial T_i^2} = \alpha(\phi) \left\{(y_i - \mu_i) \frac{d^2\theta_i}{dT_i^2} - V_{ii}(\mu_i)\left[\frac{d\theta_i}{d\mu_i} \cdot \frac{d\mu_i}{dT_i}\right]^2\right\}. \qquad (26)$$

Sometimes, Fisher's scoring method is simpler than Newton-Raphson method. In Fisher's method of scoring, matrix of second order derivatives in equation (21) is replaced by its expectation. Now taking expected value equation (26) can be written as

$$E\left[-\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_l}\right] = -\sum_i \left\{\frac{\alpha(\phi)}{V_{ii}(\mu_i)}\right\} \cdot \left[\frac{d\mu_i}{dT_i}\right]^2 x_{ij} x_{il}$$

$$= -\sum_i \alpha(\phi) W_{ii} (x_{ij} x_{il}). \qquad (27)$$

Hence by proper changes in equation (21), and using equation (27) we get,

$$\hat{\underline{\beta}}^{(m)} = \hat{\underline{\beta}}^{(m-1)} + \left[\underline{T}^{*(m-1)}\right]_{\underline{\beta}=\hat{\underline{\beta}}^{(m-1)}} \cdot \left[\alpha(\phi)\sum_i \{W_{ii} x_{ij} x_{il}\}\right]^{-1}_{\underline{\beta}=\hat{\underline{\beta}}^{(m-1)}}$$

$$= \hat{\underline{\beta}}^{(m-1)} + \left[\sum_i \left\{\frac{\alpha(\phi)(x_{ij}(y_i-\mu_i))}{V_{ii}}\right\} \cdot (d\mu_i/dT_i)\right]_{\underline{\beta}=\hat{\underline{\beta}}^{(m-1)}}$$

$$\cdot \left[\alpha(\phi)\sum_i \{W_{ii} x_{ij} x_{il}\}\right]^{-1}_{\underline{\beta}=\hat{\underline{\beta}}^{(m-1)}}$$

$$\hat{\underline{\beta}}^{(m)} = \hat{\underline{\beta}}^{(m-1)} + \left\{X'W\, F^{-1}(\underline{y}-\underline{\mu})\right\}_{\underline{\beta}=\hat{\underline{\beta}}^{(m-1)}} \cdot \left[(X'W\, X)^{-1}\right]_{\underline{\beta}=\hat{\underline{\beta}}^{(m-1)}}, \quad (28)$$

Equation (28) is same as

$$\left[(X'W\, X)\right]_{\underline{\beta}=\hat{\underline{\beta}}^{(m-1)}} \hat{\underline{\beta}}^{(m)} = \left[(X'W\, X)\right]_{\underline{\beta}=\hat{\underline{\beta}}^{(m-1)}} \hat{\underline{\beta}}^{(m-1)}$$

$$+ \left\{X'W\, F^{-1}(\underline{y}-\underline{\mu})\right\}_{\underline{\beta}=\hat{\underline{\beta}}^{(m-1)}}$$

$$= X'W\, \hat{\underline{z}}^{(m-1)};$$

where

$$\underline{z} = X\underline{\beta} + \left\{F^{-1}(\underline{y} - \underline{\mu})\right\}$$

i.e.

$$\underline{z} = \underline{T} + \left\{F^{-1}(\underline{y} - \underline{\mu})\right\}.$$

Thus m^{th} approximation of estimate of $\underline{\beta}$ can be obtained from the equation

$$(X'W\, X)\hat{\underline{\beta}} = X'W\, \hat{\underline{z}}, \quad (29)$$

where the matrix W is obtained from current estimate of $\underline{\beta}$.

Remark : From equations (7) and (29) it is observed that for generalised linear model with density (or mass) function as in equation (3.2-9), solution of the maximum likelihood equations is equivalent to an iterative weighted least squares procedure with a weight function

$$W = V^{-1}F^2, \tag{30}$$

and a modified dependent variate

$$\underline{Z} = \hat{\underline{T}} + (\underline{Y} - \underline{\mu})F^{-1}; \tag{31}$$

where

(i) $\underline{\mu} = E(\underline{Y})$

(ii) $V = \text{diag}(v_{11}, v_{22}, \ldots, v_{nn})$,

and

(iii) $F = \text{diag}(f_{11}, f_{22}, \ldots, f_{nn})$,

with

$$\left.\begin{aligned} f_{ii} &= (d\mu_i/dT_i) \\ v_{ii} &= \text{var}(Y_i)/\alpha(\phi) \end{aligned}\right\}, \text{ for } i=1,2,\ldots,n.$$

In fitting generalised linear model, we have used Fisher's scoring method as a numerical technique to obtain parameter estimates. This method was introduced by Fisher(1935), in the appendix of a paper by Bliss (1935). Green(1984), Jorgenson (1984), Finney (1987) and McCullagh & Nelder (1989) used the same method to fit generalised linear model. Below we give an algorithm of the same method to compute the maximum likelihood estimates of $\underline{\beta}$.

3.4.3 : Algorithm to obtain maximum likelihood estimates of $\underline{\beta}$:

While fitting generalised linear model, approximate values

of maximum likelihood estimates of $\underline{\beta}$ can be obtained by using weighted least square method as a numerical technique. As stated earlier, since the new dependent variable Z and W both depend on the fitted values, the procedure is iterative. Hence estimates should be obtained iteratively by using N-R-method or Fisher's method of scoring. To obtain estimates of $\underline{\beta}$, equation (29) can be used. This method has the following steps.

(I) Write the incidence matrix X ;

(II) fix a small positive number ε (say,) to get estimate of $\underline{\beta}$ with desired accuracy ;

(III) find expressions for $(d\mu_i/dT_i)$ and V_{ii} ;

(IV) take averages of different samples in the data as initial estimates $\hat{\mu}_i^{(0)}$ of μ_i, $(i=1,2,\ldots,n)$;

(V) obtain $\hat{T}_i^{(0)} = m(\hat{\mu}_i^{(0)})$;

(VI) compute

$$\hat{W}_{ii}^{(0)} = \left[\frac{1}{V_{ii}} \cdot (d\mu_i/dT_i)^2 \right]_{\mu_i = \hat{\mu}_i^{(0)}}$$

and

$$\hat{z}_i^{(0)} = \hat{T}_i^{(0)} + \left[\frac{(y_i - \mu_i)}{V_{ii}} \cdot (d\mu_i/dT_i) \right]_{\mu_i = \hat{\mu}_i^{(0)}} ;$$

(VII) compute $\hat{\underline{\beta}}^{(0)}$ as

$$\hat{\underline{\beta}}^{(0)} = (X'W^{(0)}X)^{-1}X'W^{(0)}\underline{\hat{z}}^{(0)};$$

(VIII) obtain $\underline{\hat{T}}^{(1)} = X\hat{\underline{\beta}}^{(0)}$, and $\hat{\mu}_i^{(1)} = m^{-1}(\hat{T}_i^{(1)})$;

(IX) repeat steps similar to the steps (VI) to (VIII) until

$$\left| \beta_j^{(l)} - \beta_j^{(l-1)} \right| < \varepsilon \ , \quad \text{for all } j = 0,1,\ldots,k; \qquad (32)$$

(X) $\beta_j^{(P)}$, (j = 0,1,...,k) are the final estimates of β_j if conditions (32) are satisfied for $\beta_j^{(P)}$, but are not satisfied for $\beta_j^{(P-1)}$, for atleast one value of j.

Remarks : (i) When any of the $\hat{\mu}_i^{(0)}$ takes extreme value, some adjustment should be done by adding or subtracting proper positive number, so that the corrected initial estimate will not be an extreme value.

(ii) Another important fact to be noted here is about existence of inverse of the matrix (X'W X). Wedderburn (1976) has proved that $(X'W X)^{-1}$ exists for log concave link functions.

After fitting generalised linear model to the data, one may be interestd in testing different hypotheses about the model parameters $\underline{\beta}$. For this purpose it is necessary to obtain sampling distribution of maximum likelihood estimates of $\underline{\beta}$.

3.4.4 :Sampling dustribution of maximum likelihood estimates :

Suppose (X'W X) is non-singular, so that there are unique maximum likelihood estimates of the parameters $\underline{\beta}$ and are close to the true value $\underline{\beta}$. Suppose I is the information matrix. Then using Taylor's first order approximation about $\hat{\underline{\beta}}$ for $\underline{T}^*(\underline{\beta})$, we get,

$$\underline{T}^*(\underline{\beta}) = \underline{T}^*(\hat{\underline{\beta}}) + H(\hat{\underline{\beta}})(\underline{\beta} - \hat{\underline{\beta}}). \qquad (33)$$

For simplicity, let

$$\underline{U} = \partial \ell / \partial \underline{\beta}, \qquad (34)$$

and H is the matrix whose $(j,l)^{th}$ component is $(\partial^2 \ell / \partial \beta_j \partial \beta_l)$. Then result similar to (3.2-25) gives

$$I = E(\underline{U}\,\underline{U}'),$$

$$= -E(H). \qquad (35)$$

As H is asymptotically equal to its expected value and since $\underline{T}^*(\hat{\underline{\beta}}) = 0$, for large samples equation (33) becomes,

$$\underline{T}^*(\underline{\beta}) = -I(\underline{\beta} - \hat{\underline{\beta}}),$$

which implies,

$$(\hat{\underline{\beta}} - \underline{\beta}) = I^{-1}\{\underline{T}^*(\underline{\beta})\}. \qquad (36)$$

If I is the matrix of constants, taking expectation of the equation (36), we have,

$$E(\hat{\underline{\beta}} - \underline{\beta}) = \underline{0}, \qquad (37)$$

and

$$E[(\hat{\underline{\beta}} - \underline{\beta}).(\hat{\underline{\beta}} - \underline{\beta})'] = E\{I^{-1}\underline{T}^*(\underline{\beta}).\underline{T}^{*\prime}(\underline{\beta})I^{-1}\}. \qquad (38)$$

Hence by using equation (36) and since it is assumed that I is the matrix of constants, equation (38) can be rewritten as

$$E[(\hat{\underline{\beta}} - \underline{\beta}).(\hat{\underline{\beta}} - \underline{\beta})'] = I^{-1}. \qquad (39)$$

Thus for large samples,

$$(\hat{\underline{\beta}} - \underline{\beta}) \text{ has } N_{(k+1)}(\underline{0}\ ,\ I^{-1}) \text{ distribution.}$$

This shows that for large samples,

$$\hat{\underline{\beta}} \text{ has } N_{(k+1)}(\underline{\beta}\ ,\ I^{-1}) \text{ distribution.}$$

i.e. $(\hat{\underline{\beta}} - \underline{\beta})'I(\hat{\underline{\beta}} - \underline{\beta})$ has χ^2 distribution with (k+1) d.f.

3.5 : Measures of adequacy of the fitted model :

Once the model is proposed, the next part is to see whether

it fits 'well' to the data; i.e. to check how 'good' fitted model describes the data.

A 'good' model has to balance two requirements.

1. The model should be as complex enough to approximate the real world phenomenon it describes.
2. The model should be as simple as possible for the reason that, simpler it is, the more comprehensible it is.

Thus if there are two models that give approximately the same degree of agreement with reality, we should prefer the 'simpler' model. A model is simpler, if it contains fewer number of parameters. The 'full' model describes the data in the 'best' possible way, but it does not reduces the data as it has the number of parameters equal to the number of observations. This model describes the data in the best possible way because, it assigns complete variation in values of the response variate to the systematic components. On the other hand, 'null' model is the simplest model containing a single parameter. Thus it considers all the variation between values of Y due to the random component. Hence it describes the data in the 'worst' manner. This shows that model should be intermediate model containing p ($1 < p < n$) parameters, and describing the data in sufficiently better way. Some well known techniques used to check the adequacy of the fitted model are as follows.

1. usual chi-square statistic for testing goodness of fit,
2. Pearson's chi-square statistic (X^2),
3. Deviance (D).

3.5.1 : <u>Chi-square statistic for goodness of fit</u> : The usual χ^2-test of goodness of fit can be used to test the goodness fit of the model. This test uses the χ^2 statistic given by,

$$\chi^2 = \sum_i \left[\frac{(O_i - E_i)^2}{E_i} \right] , \quad i=1,2,\ldots,n , \qquad (1)$$

which has chi square distribution with (n-t-1) d.f.

Here

(i) t = number of d.f. lost in pooling;

(ii) O_i = observed frequency of i^{th} class;

(iii) E_i = expected frequency of i^{th} class.

3.5.2 : Pearson's X^2 statistic : Other important measure of goodness of fit is 'generalised Pearson X^2 statistic'. This statistic is obtained by using the formula,

$$X^2 = \sum_i \left\{ \frac{(y_i - \hat{\mu}_i)^2}{V_{ii}(\hat{\mu}_i)} \right\} . \qquad (2)$$

Formulae for generalised Pearson X^2 statistic for distributions in table (3.2) are summarised in table (3.3) given below. These formulae can be obtained easily.

TABLE 3.3

Distribution	Pearson X^2 statistic
Normal	$\sum_i (y_i - \hat{\mu}_i)^2$
Gamma	$\sum_i (y_i - \hat{\mu}_i)^2 / (\hat{\mu}_i)^2$
Exponential	$\sum_i (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i^2$
Poisson	$\sum (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i$
Grouped binary	$\sum (y_i - \hat{\mu}_i)^2 / (\hat{\mu}_i (1 - \hat{\mu}_i))$

3.5.3 : Deviance : To assess the goodness of fit of a generalised linear model, the following statistic namely, 'deviance' (D,say) is suggested by Nelder & Wedderburn (1972), which is based on likelihood ratio statistic. Bishop & others (1975) named the same statistic as G^2 statistic. The full model is useful in defining this test statistic.

Let $\ell(\hat{\underline{\mu}},\phi;\underline{y})$ and $\ell(\underline{y},\phi;\underline{y})$ be the maximised log likelihoods corresponding to the model under study and the full model respectively. Then the deviance (D) is given by

$$D(\underline{y},\hat{\underline{\mu}}) = \phi\{-2[\ell(\hat{\underline{\mu}},\phi;\underline{y}) - \ell(\underline{y},\phi;\underline{y})]\}. \qquad (3)$$

In generalised linear models deviance plays a role similar to the role of residual sum of squares in classical linear models.

To make meaning of the term deviance more clear, below formulae for deviance corresponding to the distributions in table(3.3) are obtained by assuming that there are n observations on Y.

Illustration 1 :Normal distribution : Let observations (Y_i) on the response variate Y are independently Normally distributed random variables with parameters (μ_i,σ^2). Then from equation (3.2-4) we have

$$\ell(\hat{\underline{\mu}},\sigma^2;\underline{y}) = (-n/2)\ln(2\Pi\sigma^2) - (1/2\sigma^2)\sum_i (y_i-\hat{\mu}_i)^2. \qquad (4)$$

Similarly we obtain the formula for $\ell(\underline{y},\phi;\underline{y})$. Hence from equation (3),deviance becomes

$$D(\underline{y},\hat{\underline{\mu}}) = \sum_i (y_i-\hat{\mu}_i)^2. \qquad (5)$$

Remark : As already pointed in chapter 2, the goodness of fit of classical linear model with error components distributed as

normal, is based on residual sum of squares. It can be seen that the deviance (D) in equation (5) is nothing but residual sum of squares (E), given in equation (2.4-2).

Illustration 2 : Gamma distribution : Suppose we have n observations (Y_i) on the response variate Y having gamma distribution with p.d.f. as in equation (3.2-11). Then equation (3.2-13) gives

$$\ell(\hat{\underline{\mu}},\nu;\underline{y}) = \nu\sum_i [-y_i/\hat{\mu}_i - \ln(\hat{\mu}_i)] - n\ln(\Gamma(\nu)) + n\nu\ln(\nu) + (\nu-1)\sum_i \ln(y_i) \qquad (6)$$

Hence after getting $\ell(\underline{y},\phi;\underline{y})$, from equation (3) we have,

$$D(\underline{y};\hat{\underline{\mu}}) = 2\{\sum_i [-\ln(y_i/\hat{\mu}_i) + (y_i-\hat{\mu}_i)/\hat{\mu}_i]\} \qquad (7)$$

Nelder & Wedderburn (1972) have shown that the second term in equation (7) is identically equal to zero. Proof is made available in Appendix-1. Thus equation (7) is equivalent to

$$D(\underline{y};\hat{\underline{\mu}}) = -2\sum_i \ln(y_i/\hat{\mu}_i). \qquad (8)$$

Illustration 3 : Exponential distribution :

Suppose Y_i (i=1,2,...,n) are n independent observations on Y having exponential distribution with mean μ. Then the log likelihood of Y can be written as

$$\ell(\underline{\mu};\underline{y}) = -\sum_i [\ln(\mu_i) + (y_i/\mu_i)]. \qquad (9)$$

Hence from equation (3), we obtain

$$D(\underline{y};\hat{\underline{\mu}}) = -2\{\sum_i [\ln(y_i/\hat{\mu}_i) + (y_i-\hat{\mu}_i)/\hat{\mu}_i]\}. \qquad (10)$$

According to the theorem proved in Appendix-1, it can be seen that the second term in equation (10) is zero. Hence, equation (10) becomes

$$D(\underline{y};\hat{\underline{\mu}}) = -2\{\sum_i \ln(y_i/\hat{\mu}_i)\}. \qquad (11)$$

Illustration 4 : Poisson distribution :

If we assume Poisson distribution with parameter μ, then from equation (3.2-16) it can be seen that

$$\ell(\hat{\underline{\mu}},\underline{y}) = \sum_i [-\hat{\mu}_i + y_i \ln(\hat{\mu}_i) - \ln(y_i!)]. \qquad (12)$$

This implies,

$$D(\underline{y};\hat{\underline{\mu}}) = 2\{\sum_i [y_i \ln(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)]\}. \qquad (13)$$

As per theorem given in Appendix-1, it can be seen that the second term in equation (13) is identically equal to zero. Thus equation (13) is same as

$$D(\underline{y};\hat{\underline{\mu}}) = 2\sum_i y_i \ln(y_i/\hat{\mu}_i). \qquad (14)$$

Illustration 5: Binary distribution for grouped data :

Considering Binomial distribution with parameters (m_i^*,μ_i) for $Y_i^*=m_i^*Y_i$, where Y_i is the i^{th} observation on Y. Then assuming independence, we have,

$$\ell(\hat{\underline{\mu}},\underline{m}^*;\underline{y}) = \sum m_i^*\{y_i \ln(\hat{\mu}_i) + (1-y_i)\ln(1-\hat{\mu}_i)\} + c(\underline{m}^*,\underline{y}), \qquad (15)$$

where $c(\underline{m}^*,\underline{y})$ is a function free from $\hat{\mu}_i$. Thus from equation (3), deviance is

$$D(\underline{y},\hat{\underline{\mu}}) = 2\sum m_i^*[y_i \ln(y_i/\hat{\mu}_i) + (1-y_i)\ln(1-y_i)/(1-\hat{\mu}_i)]. \qquad (16)$$

Deviances discussed in this section, for different distributions, are summarised in table 3.4.

TABLE 3.4

Distribution	Deviance
Normal	$\sum_i (y_i - \hat{\mu}_i)^2$
Gamma	$-2 \sum_i [\ln(y_i/\hat{\mu}_i)]$
Exponential	$-2 \sum_i \{\ln(y_i/\hat{\mu}_i)\}$
Poisson	$2 \sum_i \{[y_i \ln(y_i/\hat{\mu}_i)]\}$
Grouped binary	$2\sum_i m_i \{[y_i \ln(y_i/\hat{\mu}_i)] + (1-y_i)\ln[(1-y_i)/(1-\hat{\mu}_i)]\}$

3.5.4 : Advantages of the deviance (D) statistic :

(i) *D statistic is appropriate for maximum likelihood estimates.*

Explanation : From equation (1), it is clear that since $\ell(\hat{\mu}, \phi, y)$ is maximised log likelihood function for intermediate model, and since D is non negative, maximum likelihood estimates give the minimum value of D. Thus D is appropriate for maximum likelihood estimates.

(ii) *Conditional break down of D is possible.*

Explanation : When there are two models, model (1) and model (2), they are said to be nested if one of them (say e.g. model (2)) contains only a subset of terms contained in other model (i.e. model (1)). For nested models conditional break down of D is possible. In simplest conditional break down of D statistic D(2) corresponding to model (2) can be broken down into two parts;

(i) a measure of distance of the estimates of parameters in the model (2) from those obtained under model (1);

(ii) a D statistic D(1) for model (1).

For this, rewrite equation (1) as,

$$D(\underline{y};\hat{\underline{\mu}}) = \phi\{-2([\ell(\hat{\underline{\mu}}_2,\phi;\underline{y}) - \ell(\hat{\underline{\mu}}_1,\phi;\underline{y})] + [\ell(\hat{\underline{\mu}}_1,\phi;\underline{y}) - \ell(\underline{y},\phi;\underline{y})])\}$$

$$= D[(2)|(1)] + D(1), \qquad (17)$$

where $\ell(\hat{\underline{\mu}}_i,\phi;\underline{y})$ denotes the maximised log likelihood under model (1) and $D[(2)|(1)]$ is the conditional D statistic for model (2) given model (1). If model (1) is a full model then $D[(2)|(1)]$ is equal to D(2). Such type of conditional break down does not exist for χ^2 statistic given in equation (16).

(iii) *Structural breakdown of D is possible.*

The structural break down which is possible with D, is not possible with χ^2 statistic in equation (16).

To use the deviance(D) as a test statistic for testing goodness of fit of the fitted model, the distribution of 'D' must be known. In classical linear models, as normal distribution is assumed for error components, exact distribution of the deviance can be computed easily. But when we depart from normal distribution and from linearity of different effects, generally exact distribution of the deviance is not obtainable. In some situations like exponential distriution, exact sampling distribution can be achieved. When exact distribution can not be obtained, a chi-square distribution is a better approximation for the difference in deviances.

Sampling distribution of the difference between deviances can be obtained by using that for maximum likelihood estimates of the model parameters $\underline{\beta}$. We have already shown that, for large samples,

$(\hat{\underline{\beta}} - \underline{\beta})'I(\hat{\underline{\beta}} - \underline{\beta})$ has χ^2 distribution with (k+1) d.f.

Now, on expanding log likelihood function $\ell(\underline{\beta};\underline{y})$ about the maximum likelihood estimates $\hat{\underline{\beta}}$ of $\underline{\beta}$, we get,

$$\ell(\underline{\beta};\underline{y}) = \ell(\hat{\underline{\beta}};\underline{y}) + (\underline{\beta} - \hat{\underline{\beta}})'\underline{U}(\hat{\underline{\beta}}) + (1/2)(\underline{\beta} - \hat{\underline{\beta}})'H(\hat{\underline{\beta}})(\underline{\beta} - \hat{\underline{\beta}}), \quad (18)$$

where $\underline{U}(\hat{\underline{\beta}})$ and $H(\hat{\underline{\beta}})$ are values of $\underline{U}$ and H evaluated at $\hat{\underline{\beta}}$.

Since $\hat{\underline{\beta}}$ are maximum likelihood estimates of $\underline{\beta}$, these are the solutions of the equations $\underline{U}(\underline{\beta}) = \underline{0}$. Hence,

$$\underline{U}(\hat{\underline{\beta}}) = \underline{0}. \quad (19)$$

Using equation (19) and approximating $H(\hat{\underline{\beta}})$ by E(H), equation (18) becomes

$$2[\ell(\hat{\underline{\beta}};\underline{y}) - \ell(\underline{\beta};\underline{y})] = (\underline{\beta} - \hat{\underline{\beta}})'I(\underline{\beta} - \hat{\underline{\beta}}).$$

Therefore, from the distribution of $\hat{\underline{\beta}}$, it is clear that for large n, the distribution of $2[\ell(\hat{\underline{\beta}};\underline{y}) - \ell(\underline{\beta};\underline{y})]$ has χ^2 distribution with (k+1) d.f. Thus the distribution of deviance is χ^2 distribution with (k+1) d.f.

Though this chi-square approximation is not adequate in all the cases, a better approximation is yet to be suggested. The table of deviance differences has its importance to select the terms showing significant effect on value of the response variate.

3.6 : Analysis of deviance

For orthogonal data with normal errors, analysis of variance (ANOVA) is a very useful statistical tool for separating the effects due to systematic components from those due to random components. Nelder & Wedderburn (1972) suggested a generalisation needed, so that it is applicable for analysing generalised linear models. While making generalisation, two problems should be considered. First is, terms in the model are

generally non orthogonal and the other is, for non normal distributions, different sums of squares (SS) are not measuring properly the effects due to different components.

Before going to the generalisation, we look towards the usual ANOVA table from different angle. ANOVA can be considered as the first difference of the measures of descripancy for a sequence of models each including one term more than the previous model; e.g. in factorial model with two factors A and B (say), we have ANOVA with three terms namely main effects A and B, and the interaction effect A.B. SS for these factors are the first differences of the SSEs corresponding to the sequence of four models 1, A, A+B and A+B+A.B respectively. Note that measure of descripancy for model 1 is just TSS, and that for full model A+B+A.B is 0. It is also clear that for the full model d.f. are equal to 0.

Now the generalisation is very much clear. Nelder & Wedderburn (1972) have used D statistic as a measure of descripancy for a given sequence of nested generalised linear models, and taking the first differences formed the analysis of deviance (ANODEV) table. Since in generalised linear model, mostly the data is non orthogonal, the interpretation is a bit complicated. Each number in the ANODEV table represents variation due to that after eliminating effects of the terms above it and ignoring effects of the terms below it.

In model fitting problem interpretation of the fitted model is also a very important part. In the next section this part is introduced in brief.

3.7 : Interpretation of the fitted model :

Interpretation of the fitted model is also a part and parcel of the model fitting procedure. In other words a model should be

such that the practical conclusions can be drawn from estimated parameters of the model. Thus the question is 'what do the estimated coefficients of the model tells us about research qestion that motivated the study ?' Many times coefficients associated with the explanatory variates are of interest. Very rarely intercept β_0 is of interest. Estimated coefficients of the stimulus variates represents rate of change in the value of a function of the response variate corresponding to per unit change in the value of stimulus variate. Thus interpretation of data involves two parts.

(1) Determining functional relationship between the response variate and the stimulus variates $\underline{X} = (X_1, X_2, \ldots, X_k)'$.

(2) Defining appropriately, unit of change for the independent variables.

As we have seen earlier, the functional relationship between Y and $\underline{X}$ is known as link function. Along with the likelihood function, this function is also required for fitting generalised linear model to the data.

Another important segment of the model building is model checking. When some proposed model is fitted to the data, it is necessary to check whether the fitted model is appropriate. This checking is needed, because some times model fits well to the data, though the assumptions made while fitting are invalid. In the next section, discussion is made on model checking for generalised linear models. Model checking includes checking goodness of fit of the fitted model and checking for validity of the assumptions made while formulating the model.

3.8 : Model checking for generalised linear models :

The main problem in data analysis through fitting a model is to select a proper class of the models, so that conclusions drawn

from the analysis are not far away from the truth. A statistician can choose the model class carefully by taking into account the type and structure of the data. Many times it may happen that, though the class is selected carefully, the data themselves indicate that the model selected is unappropriate. This situation occurs because of two reasons. First, the complete data indicates departure from the fitted model, and secondly, there may be few data points away from the rest which are known as outliers. Below we club the various methods of model checking into three groups, namely,

(I) visual display;

(II) tests of deviations in particular direction;

(III) searching for influential points (outliers).

The methods in the group (I) are similar to the methods of drawing residual plots as described in section (2.5). First we discuss these methods.

As in classical linear models, here also the raw material for model checking is, fitted values of the linear predictor $\underline{T}$, new adjusted dependent variable Z and the projection matrix H. The residuals can also be used as a raw material for model checking. The various types of residuals in generalised linear model are discussed below.

3.8.1 : Residuals : As stated in section (2.5), the residual analysis is essential to test adequacy of the fitted model. Residuals are useful to examine thoroughly the appropriateness of the fitted model. These are also useful to check whether the outliers are present. Hence in the next Part we study the residuals related to generalised linear models.

In case of generalised linear models, the generalisation of residuals is to be done so that it is useful for non Normal

distributions also. Below we discuss three different residuals, suitable for generalised linear models.

Definition-4 : Pearson residual :- It is defined as the signed square-root of the component of the Pearson's chi square statistic of goodness of fit. Thus it is given by,

$$r_p = (y-\hat{\mu})/[V(\hat{\mu})]^{-1}. \qquad (1)$$

An important disadvantage of Pearson residual is that its distribution for non-Normal distribution is skewed. So the properties of Normal distribution turns out to be invalid. Hence this residual is not much useful. The residual which is more appropriate is Anscombe residual. This residual was introduced by Anscombe (1953).

Definition-5 :Anscombe residual:- Anscombe defined the residual by replacing y in Pearson residual by the function A(y) and $\hat{\mu}$ by $A(\hat{\mu})$. This function A(.) is selected such that the distribution of A(y) is very close to the Normal distribution. Wedderburn proved that, for generalised linear model, the function A(.) is given by,

$$A(.) = \int \frac{d\mu}{[V(\mu)]^{(1/3)}} . \qquad (2)$$

This replacement normalises the probability function, but to stabilize the variance it is necessary to scale it by estimated standard deviation of A(Y). Since it is not easy to compute exact variance of A(Y), we take its first order approximated value,

$$[A'(\hat{\mu})]\{V(\hat{\mu})\}^{(1/2)}. \qquad (3)$$

Thus Anscombe residual becomes,

$$r_A = \frac{(A(y) - A(\hat{\mu}))}{[A'(\hat{\mu})]\{V(\hat{\mu})\}^{(1/2)}} \quad (4)$$

Below we obtain expression for 'Anscombe residual' corresponding to Poisson distribution.

Illustration 1. Poisson distribution :

Let the single observation Y be having Poisson distribution with mean μ. From table (3.2) we have,

$$V(\mu) = \mu,$$

which gives,

$$A(\mu) = \int \frac{d\mu}{\mu^{(1/2)}}$$

$$= (3/2)\mu^{(2/3)}, \quad (5)$$

and

$$A'(\mu)\{V(\mu)\}^{(1/2)} = \mu^{(1/6)}. \quad (6)$$

Using equations (5) and (6) in (4), the Anscombe residual for Poisson distribution becomes,

$$r_A = (3/2)(y-\hat{\mu})/(\hat{\mu}^{(1/6)}) \quad (7)$$

Similarly, we can easily obtain formulae for Anscombe residual corresponding to normal, gamma and exponential distributions. To obtain an expression for Anscombe residual corresponding to binomial responses is complicated. Cox & Snell (1968) obtained this expression.

When deviance is used as a measure of goodness of fit, then it is better to use deviance residual.

Definition-6 : Deviance residual :- It is defined as the product of positive square root of a quantity (d_i) contributed by each

unit for deviance and sign of the difference between y and $\hat{\mu}$. Therefore deviance residual (r_D) is given by,

$$r_D = \text{sign}(y_i - \mu_i)[d_i]^{(1/2)}. \qquad (8)$$

The three residuals for some well known distributions from one parameter natural exponential family are tabulated below in table (3.5). These formulae are quite clear from expressions for μ_i, V_{ii}, equations (1) to (4) and equation (8).

TABLE 3.5

Distribution	Pearson Residual (r_P)	Anscombe Residual (r_A)	Deviance Residual (r_D)
Normal	$(y_i - \hat{\mu}_i)$	$(y_i - \hat{\mu}_i)$	$\{\text{Sign}(y_i - \hat{\mu}_i)\}.$ $\lvert y_i - \hat{\mu}_i \rvert$
Gamma	$(y_i - \hat{\mu}_i)/\hat{\mu}_i$	$\dfrac{3(y_i^{1/3} - \hat{\mu}_i^{1/3})}{\hat{\mu}_i^{1/3}}$	$\{\text{Sign}(y_i - \hat{\mu}_i)\}.$ $\{-2[\ln(y_i/\hat{\mu}_i)] + (y_i/\hat{\mu}_i) - 1\}^{1/2}$
Exponential	$(y_i - \hat{\mu}_i)/\hat{\mu}_i$	$\dfrac{3(y_i^{1/3} - \hat{\mu}_i^{1/3})}{\hat{\mu}_i^{1/3}}$	$\{\text{Sign}(y_i - \hat{\mu}_i)\}.$ $\{-2[\ln(y_i/\hat{\mu}_i)] - (y_i/\hat{\mu}_i) + 1\}^{1/2}$
Poisson	$(y_i - \hat{\mu}_i)/\hat{\mu}_i^{1/2}$	$\dfrac{3(y_i^{2/3} - \hat{\mu}_i^{2/3})}{2(\hat{\mu}_i^{1/6})}$	$\{\text{Sign}(y_i - \mu_i)\}.$ $\{2[y_i \ln(y_i/\hat{\mu}_i)] - y_i + \mu_i\}^{1/2}$

Though the formulae for Anscombe residual and deviance

residual are looking different for non-Normal distributions, for given values of y_i and μ_i, these two residuals are having very similar values. By using any of these residuals one can carry out the residual analysis. These residuals are in terms of μ. The replacement of μ by $\hat{\mu}$ requires a standardisation factor. The standardisation of Pearson's residual is discussed by McCullagh & Nelder (1983) and that of deviance residual by Cox & Snell (1968). If absolute value of the residual for some data point is greater than '2', the corresponding data point should be checked for outlier's test. We will not discuss here the methods for group (III) due to their vastness. Now we discuss the methods in group (II) briefly.

(II) Tests of deviations in particular directions :

After fitting the generalised linear model, the most important question comes in the mind is that can the value of deviance function decrease significantly by

(a) including an extra stimulus variate,

(b) changing scale of the stimulus variates,

(c) changing link function in a particular direction,

(d) changing the variance function,

with the help of available information about the stimulus variates, link function and variance function.

(a) Selection of covariates : Generally the experiment contains information on large number of stimulus variates, but a model should contain less number of parameters and should fit sufficiently well to the data. So it is necessary to select a set of useful covariates. Then sequence of nested generalised linear models is fitted and using the deviance function it is decided, which covariate shows significant effect on the response variate. Covariates showing significant effect will be included

in the model.

(b) Checking scale of stimulus variate : Method of checking scale of the stimulus variate has following steps.

(I) Suppose X and β denotes respectively the covariate whose scale is to be checked and coefficient of the covariate.

(II) Replace X by

$$g^*(\theta;X) = \begin{cases} X^{\theta}; & \text{for } \theta \neq 0, \\ \ln(X); & \text{otherwise.} \end{cases} \qquad (9)$$

(III) Fix $\varepsilon > 0$, a small positive number, to obtain estimate of θ with desired accuracy.

(IV) To start calculations take initial value ($\theta^{(0)}$) of θ as unity.

(V) As Taylor series expansion for $g^*(\theta,X)$, with first order approximation gives

$$g^*(\theta,X) = g^*(\theta^{(0)},X) + (\theta-\theta^{(0)})[\partial g^*/\partial\theta]_{\theta=\theta^{(0)}},$$

replace $\beta g^*(\theta,X)$ by two linear terms $\beta U + \gamma V$; where,

$$U^{(0)} = g^*(\theta^{(0)};X),$$

$$V^{(0)} = \left.\frac{\partial g^*(\theta;X)}{\partial\theta}\right|_{\theta=\theta^{(0)}},$$

and

$$\gamma^{(0)} = \beta(\theta-\theta^{(0)}).$$

(VI) Fit generalised linear model with U and V as covariates.

(VII) Compute $\theta^{(1)}$, an improved estimate of θ by

$$\theta^{(1)} = \theta^{(0)} + \gamma^{(0)}/\beta^{(0)}.$$

(VIII) Repeat steps similar to the steps (V) to (VII) until

$$\left| \theta^{(l)} - \theta^{(l-1)} \right| < \varepsilon.$$

(IX) Final estimate of θ is $\theta^{(l)}$.

Remarks : (1) If the initial estimate $\theta^{(0)}$ is far away from the true value of θ, convergence of the process is not sure.

(2) Though this process is very useful, it is not a good technique to include more non linear parameters in the model, when other covariates are highly correlated. This is so because, generally estimates of the non linear parameters have large sampling error and are highly correlated with each other and with corresponding linear parameters.

(c) Checking Link Functions : While fitting generalised linear model, link functions are assumed to be known. Instead of this, it is useful to assume that, link functions come from a class of link functions and particular value of one or more parameters describes elements of that class. Most of the times a class of one parameter link functions is either taken as

$$T = \begin{cases} \mu^{\theta}; & \text{if } \theta \neq 0, \\ \ln(\mu); & \text{otherwise;} \end{cases} \qquad (10)$$

or

$$T = \{[\mu^{\theta} - 1]/\theta\}.$$

Pregibon (1980) proposed linearising technique to get optimum estimate of θ. The procedure described by him has

following steps.

(I) Suppose $T = m^*(\theta,\mu)$ is the link function, as defined in equation (10).

(II) Fix $\varepsilon > 0$, a small number, to estimate θ with desired accuracy.

(III) To start calculations take initial value $(\theta^{(0)})$ of θ as unity.

(IV) As Taylor series expansion for $m^*(\theta,\underline{\mu})$, with first order approximation gives

$$m^*(\theta,\underline{\mu}) = m^*(\theta^{(0)},\underline{\mu}) + (\theta-\theta^{(0)})[\partial m^*/\partial\theta]_{\theta=\theta^{(0)}},$$

$$m^*(\theta,\underline{\mu}) = \underline{T}^{(0)} + (\theta-\theta^{(0)})\,\underline{\mu}^{\theta^{(0)}}\ln(\underline{\mu}). \qquad (11)$$

Equation (11) can be rewritten as,

$$\underline{T}^{(0)} = m^*(\theta,\underline{\mu}) - (\theta-\theta^{(0)})\,\underline{\mu}^{\theta^{(0)}}\ln(\underline{\mu}),$$

which is equivalent to,

$$\underline{T}^{(0)} = \sum_j X_{ij}\beta_j + \gamma^{(0)}X_{i(k+1)}, \qquad (12)$$

where,

$$X_{i(k+1)} = \mu_i^{\theta^{(0)}}\ln(\mu_i). \qquad (13)$$

(V) Fit generalised linear model with linear predictor $\underline{T}^{(0)}$ as given in equation (12).

(VI) Also fit generalised linear model with linear predictor $\underline{T} = \sum_j X_{ij}\beta_j$.

(VII) If difference between deviances of the two fitted

models at steps (V) and (VI) is significant, conclusion is $\theta^{(0)}$ is appropriate value of θ.

(VIII) Take different values of θ in place of $\theta^{(0)}$ and repeat the steps (IV) to (VII). The value of θ for which deviance is minimum, is the maximum likelihood estimate of θ.

(d) Checking Variance Function : Nelder & Pregibon (1987) suggested a method of comparing variance functions for continuous data by using the idea of 'extended quasi likelihood function'. For the distributions discussed in the table (3.2), it can be shown that the log likelihood (ℓ) is close to an extended quasi likelihood discussed in chapter 5. This fact can be used for checking variance function.

3.9 Method of obtaining robust estimates :

As discussed in section (2.3), least absolute deviations approach to estimate the parameters was introduced by Boscovich in the year 1757, about 40 years back to the introduction of least square approach due to Gauss in 1797. Hence it is naturally quite intersesting to see whether least absolute deviations approach can be used in generalised linear models, instead of usual weighted least square approach, to estimate the model parameters. Morgenthaler(1992) has explained haw least absolute deviations principle can be used to find robust estimates of the model parameters in case of generalised linear models.

During the last decade or more, several statisticians worked on robust estimation of the model parameters for generalised linear models. Some of them are Pregibon(1982), Stephanski & others(1986), and Kunsch & others(1989). All the above statisticians cocentrated particularly on logistic model (model

is discussed in chapter 4). Morgenthaler(1992) described robust estimation in case of generalised linear models. Below the method of obtaining robust estimates is given explicitly.

Method of obtaining robust estimates :

Since least absolute deviations resists the gross error significantly, it is of interest to study haw least absolute deviations principle can be used to obtain robust estimates. Suppose $Y_i (i=1,2,\ldots,n)$ are n independent responses with

$$\left.\begin{aligned} E(Y_i) &= \mu_i, \\ Var(Y_i) &= V_{ii}(\mu_i). \end{aligned}\right\} \qquad (1)$$

Then robust estimates for fitting L_q- norm $(q \geq 1)$ can be obtained by minimising the quantity,

$$K(\underline{\mu}, \underline{Y}) = \sum_i \left| \frac{(Y_i - \mu_i)}{[V_{ii}(\mu_i)]^{-(1/2)}} \right|^q.$$

Therfore, i^{th} component of the gradient corresponding to the quantity $K(\underline{\mu}, \underline{Y})$ is

$$q\,[V_{ii}(\mu_i)]^{-(q/2)} |Y_i - \mu_i|^{(q-1)} \operatorname{sgn}(Y_i - \mu_i), \qquad (2)$$

$$i=1,2,\ldots,n.$$

For q = 2, the quantity in (1) is equivalent to the i^{th} component of quasi likelihood (quasi likelihood is introduced in chapter 5). Morgenthaler(1992) mentioned that for any other value of q, gradient in (2) non consistent estimates of β when the responses are not symmetrically distributed around their means. To obtain symmetrically distributed responses, the distributional form of the response variate must be known.

Since in generalised linear models, distributional form of the response variate is known, correction factors for gradient

given in (2), to obtain consistent estimates of β are naturally,

$$C_i = E\left\{|Y_i-\mu_i|^{(q-1)}\,\mathrm{sgn}(Y_i-\mu_i)\right\}, \quad i=1,2,\ldots,n. \tag{3}$$

Hence the i^{th} $(i=1,2,\ldots,n)$ component of the corrected gradient is,

$$q\,[V_{ii}(\mu_i)]^{-(q/2)}\,|Y_i-\mu_i|^{(q-1)}\,\mathrm{sgn}(Y_i-\mu_i) - C_i].$$

Thus estimating equations to obtain estimates of β, corrected for consistency are,

$$q[V_{ii}(\mu_i)]^{-(q/2)}\,|Y_i-\mu_i|^{(q-1)}[\mathrm{sgn}(Y_i-\mu_i) - C_i]\,(d\mu_i/dT_i)X_{ij} = 0, \quad i=1,2,\ldots,n. \tag{4}$$

Hence this method can be used as an alternative method to estimate the model parameters β in case of generalised linear models.

In all the above discussion of generalised linear models, it is assumed, for known distributional form of the responses Y_i $(i=1,2,\ldots,n)$, variance of Y_i is a specific function $V(\mu_i)$ of the mean μ_i of Y_i. Mathematically, the variance of Y_i $(i=1,2,\ldots,n)$ is given by,

$$\mathrm{Var}(Y_i) = \phi\, V(\mu_i). \tag{5}$$

To clearify meaning of the above statement, an illustration is given below.

Illustration 1. Normal distribution :

Suppose $Y_i (i=1,2,\ldots,n)$ are n independent responses having $N(\mu_i,\sigma^2)$ distribution. From the table (3.2) we recall that, for $N(\mu_i,\sigma^2)$ distribution,

$$\phi = \sigma^2 \quad \text{and} \quad V(\mu_i) = 1.$$

Thus in this case, the variance of Y_i (i=1,2,...,n) is a specific function of μ_i upto a muliplicative constant σ^2.

On the other hand, if we assume that responses Y_i (i=1,2,...,n) are independently distributed $N(\mu_i, \sigma_i^2)$ variables, then

$$\phi_i = \sigma_i^2 \quad \text{and} \quad V(\mu_i) = 1.$$

Therefore, in this case $Var(Y_i)$ is not a specific function of the mean μ_i (i=1,2,...,n). This is a case where the factor, related to the dispersion parameter is varying instead of being a constant. In such situations if ϕ_i is unknown, 'generalised linear model' can not be fitted in its original form given by Nelder & Wedderburn (1972). The models which are useful here are 'generalised linear models with varying dispersion.

3.10 Generalised linear models with varying dispersion

Smyth (1989) introduced the 'generalised linear model with varying dispersion'. Thus he generalised 'generalised linear models', by including a 'dispersion model' along with the usual 'mean model'. The term 'mean model' which occurs first time, is nothing but the usual 'generalised linear model' given in definition (3). Before Smyth (1989), Nelder & Pregibon (1987) introduced new class of models, namely, 'extended quasi likelihood models'. The model introduced by Smyth (1989) is very much similar to the extended quasi likelihood model. Below we discuss 'generalised linear model with varying dispersion'.

Generalised linear model with varying dispersion' can be defined in the following two parts.

Definition-7 : Mean model :

Let Y be a response variate with p.d.f. (or p.m.f.) $f(.;\theta^*)$ which belongs to the one parameter natural exponential family.

Suppose Y_i $(i=1,2,\ldots,n)$ are n independent observations on the response variate Y such that $E(Y_i) = \mu_i$ and $Var(Y_i) = \phi_i V_{ii}(\mu_i)$. Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_k$ be the vectors of known values of the covariates X_j $(j=1,2,\ldots,k)$. Suppose

$$T_i = \beta_0 + \sum_j x_{ij}\beta_j, \text{ for } i=1,2,\ldots,n, \qquad (1)$$

where

$\underline{\beta} = (\beta_0,\beta_1,\ldots,\beta_k)'$ is the vector unknown model parameters and

$\underline{T} = (T_1,T_2,\ldots,T_n)'$ is the linear predictor.

Then a class of models of the form,

$$\underline{T} = m(E(\underline{Y})); \qquad (2)$$

where $m(.)$ is a strictly monotoic differentiable function; is called as the mean model.

For the 'dispersion model', deviance residuals (d_i) are generally taken as unobservable responses. Then dispersion model can be defined as below.

Definition-8 : Dispersion model :

Consider the unobservable responses d_i $(i=1,2,\ldots,n)$ with $E(d_i) = \phi_i$ and $Var(d_i) = \theta V_D(\phi_i)$. Let $\underline{z}_j^*$ $(j=1,2,\ldots,k^*)$ be the vectors of known values of the covariates Z_j^* $(j=1,2,\ldots,k^*)$. Suppose

$$\zeta_i = \gamma_0 + \sum_j z_{ij}^*\gamma_j, \text{ for } i=1,2,\ldots,n, \qquad (3)$$

where

$\underline{\gamma} = (\gamma_0,\gamma_1,\ldots,\gamma_{k^*})'$ is the vector unknown model parameters and

$\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_n)'$ is the linear predictor for dispersion model.

Then a class of models of the form,

$$\zeta = h(E(\underline{d})); \qquad (4)$$

where h(.) is a strictly monotoic differentiable function; is called as the dispersion model.

Definition-9 (Smyth 1989) : Generalised linear model with varying dispersion :

The 'generalised linear model with varying dispersion' can be defined in conjunction with the above definitions of mean model and dispersion model as the class of models of the form

$$m(\underline{\mu})=X\underline{\beta},\ h(\underline{\phi})=Z\underline{\gamma},\ Var(Y_i)=\phi_i V_{ii}(\mu_i) \text{ and } Var(d_i)=\theta_i V_D(\phi_i). \qquad (5)$$

for i=1,2,...,n

The following theorem helps in developing the procedure of fitting generalised linear model with varying dispersion.

Theorem 3.1 : Mean and dispersion in the generalised linear model with varying dispersion are orthogonal.

Proof : Suppose Y_i (i=1,2,...,n) are n independent observations on the response variate. Suppose the dispersion parameters ϕ_i (i=1,2,...,n) are not constant for all responses. Then log likelihood of the complete data set is given by,

$$\ell(\underline{\theta},\underline{\phi};\underline{y}) = \sum_i \left\{ -\frac{1}{\phi_i} [y_i\theta_i - g(\theta_i) + h(y_i)] + \xi(\phi_i; y_i) \right\} \qquad (6)$$

Differentiating equation (6) w.r.t. θ_i and ϕ_i (i=1,2,...,n) we have,

$$\partial^2 \ell / (\partial\theta_i \partial\phi_i) = -(y_i - g'(\theta_i))/(\phi_i^2). \qquad (7)$$

Taking expectation on both the sides of equation (7) and using the fact that $g'(\theta_i) = \mu_i$ (for i=1,2,...,n) we get,

$$E(\partial^2 \ell / (\partial\theta_i \partial\phi_i)) = 0. \qquad (8)$$

The equations (8) above imply that mean and dispersion parameters are orthogonal.

In generalised linear models with varying dispersion since the mean and dispersion parameters are orthogonal, it is possible to estimate the parameters β and γ one at a time. We discuss below the procedure of fitting generalised linear model with varying dispersion.

3.11 Fitting of generalised linear model with varying dispersion:

Suppose $\hat{\mu}^{(r)}$ and $\hat{\phi}^{(r)}$ (r=0,1,...) denote the j^{th} approximation to the estimate of μ and ϕ respectively. Due to the interlinking of the two models, fitting procedure is alternating as described below.

While fitting the mean model keep ϕ fixed at $\hat{\phi}$. Similarly, fit dispersion model by fixing μ at $\hat{\mu}$. This method has the following steps.

(1) Decide the 'independent' variates Z_j^* ($j=1,2,\ldots,k^*$) for the 'dispersion' model.

(2) Choose two small positive numbers ϵ_1 and ϵ_2 according to the desired accuracy for $\hat{\beta}$ and $\hat{\gamma}$.

(3) To start the calculations take initial estimates of ϕ and μ as $\hat{\phi}^{(0)} = E_{n1}$ and $\hat{\mu}^{(0)} = \underline{y}$.

(4) Fit the mean model as usual by using algorithm discussed in sub section (3.4.3) for fixed $\hat{\phi}$ (the current estimate of ϕ.

(5) Compute the deviance residuals d_i (i=1,2,...,n) from the fitted mean model.

(6) Fit the dispersion model for d_i (i=1,2,...,n) by

assuming gamma distribution. While fitting the dispersion model, fix the vector of parameters $\underline{\mu}$ at $\hat{\underline{\mu}}$ (the current estimate of $\underline{\mu}$ for the mean model).

(7) Repeat the steps (4) to (6) until,

$$| \beta_j^{(m)} - \beta_j^{(m-1)} | < \varepsilon_1, \text{ for all } j=0,1,\ldots,k \qquad (9)$$

and

$$| \gamma_j^{(m_1)} - \gamma_j^{(m_1-1)} | < \varepsilon_2, \text{ for all } j=0,1,\ldots,k^*. \qquad (10)$$

Then final estimates of $\underline{\beta}$ and $\underline{\gamma}$ are

$$\hat{\underline{\beta}} = \hat{\underline{\beta}}^{(m)} \text{ and } \hat{\underline{\gamma}} = \hat{\underline{\gamma}}^{(m_1)}. \qquad (11)$$

By using the above mentioned procedure one can fit the generalised linear model with varying dispersion.

In many fields like socio-economic field, frequently the data are of discrete type. Hence in the next chapter we discuss the fitting of generalised linear model for various types of discrete data.