Generalized linear models

Juha Karvanen

April 30, 2009

Contents

Preface

This document contains short lecture notes for the course Generalized linear models, University of Helsinki, spring 2009. A more detailed treatment of the topic can be found from

- P. McCullagh and John A. Nelder, Generalized linear models. Second edition 1989. Chapman & Hall.
- A. J. Dobson, An introduction to generalized linear models. Second edition 2002. Third edition 2008. Chapman & Hall/CRC.
- lecture notes 2008. `http://www.rni.helsinki.fi/~jmh/glm08/`
- lecture notes 2005 (in Finnish). `http://www.rni.helsinki.fi/~jmh/glm05/glm05.pdf`.

1 What is a generalized linear model?

1.1 Model

Mathematical view: A statistical model is a set of probability distributions on the sample space $\mathcal{S}$. A parameterized statistical model is a parameter set Θ together with a function $P : \Theta \to P(\mathcal{S})$, which assigns to each parameter point $\theta \in \Theta$ a probability distribution P_θ on $\mathcal{S}$. A Bayesian model requires an additional component in the form of a prior distribution on Θ. [P. McCullagh (2002). What is a statistical model. The Annals of Statistics. Vol. 30, No. 5, 1225-1310.]

Applied view: Statistical model is a description of the probability distribution of random variables which can be assumed to represent a real world phenomenon.

Which of these are statistical models?

a) $X \sim N(\mu, \sigma^2)$

b) "The height of Finnish men follows a normal distribution."

c)

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) \propto \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(g_i) p_{\boldsymbol{\psi}}(x_i \mid g_i) p_{\boldsymbol{\theta}}(y_i \mid g_i, x_i),$$

d) "The risk of smokers to die to cardiovascular diseases is about twice the risk of non-smokers."

e) `glm(y ~ x, family=binomial(link = "logit"), data=doseresponse)`

1.2 Linear model

A simple linear model that describes the relationship of a single covariate x and a continuous response variable Y can be written as

$$Y_i = \alpha + \beta x_i + \epsilon_i, \tag{1}$$

where α is the intercept term, β is the regression coefficient for X and ϵ_i is an error term. Further assumptions are needed for the error term. For instance, we may assume that the error terms are mutually independent and $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n$. A less restrictive assumption is to specify only the first two moments $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$, i.e. the variance does not depend on x. Note that in model (1), the error term ϵ_i is written explicitly. It is also possible to write the same model without explicitly specifying ϵ_i

$$\mathrm{E}(Y_i \mid x_i) = \mu_i = \alpha + \beta x_i. \tag{2}$$

Model (2) tells on the expected value of Y_i on the condition of x. As a such, model (2) does not specify how the values of Y_i vary around the expected value $\mathrm{E}(Yi \mid x_i)$. Defining $\mathrm{Var}(Y_i) = \sigma^2$ we obtain a model equivalent to model (1). If the variation of Y_i is normally distributed, it can be also written $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$.

The linearity of linear model means linearity respect to the parameters. In other words, the model $\mu_i = \alpha + \beta x_i^3$ is also a linear model.

1.3 Generalized linear model

The linear model (2) can be transformed to a generalized linear model by replacing μ_i by $g(\mu_i)$

$$g(\mu_i) = \alpha + \beta x_i = \eta_i, \tag{3}$$

where g is a real-valued monotonic and differentiable function called link function and the term η_i is called linear predictor. In the other words, μ_i is the expected value of the response, η_i is a linear combination of the covariates and $g()$ defines the relationship between μ_i and η_i. Because $g()$ is monotonic, the relationship of μ_i and η_i is also monotonic. With the inverse of $g()$ we may write

$$\mu_i = g^{-1}(\eta_i), \tag{4}$$

which provides an alternative way to define GLM. Linear model is a special case of GLM where $g(\mu_i) = \mu_i$.

With multiple covariates the GLM is defined as

$$g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}. \tag{5}$$

The assumptions of the GLM are given in Section 3.

Note that GLM is different from applying a nonlinear transformation to response variable. In GLM, the nonlinear transformation is applied to the expected value of the response.

Variance is defined by the variance function V that specifies the variance of Y_i as a function μ_i

$$\mathrm{Var}(Y_i) \propto V(\mu_i). \tag{6}$$

1.4 Motivating examples

Generalized linear models are needed because linear models are not appropriate for all situations. In linear model it is implicitly assumed that the response can be have all real values, which is not the case in many practical situations. Examples:

- The number of hospital visits in a certain year for an individual is a count response that can have values $0, 1, 2, \ldots$.
- Monthly alcohol consumption (liters of absolute alcohol) for an individual is a nonnegative response that has zeroes for some individuals.
- Gamma-glutamyltransferase (GGT) measured from serum blood is a positive response.
- Daily rainfall is a nonnegative response.
- Presence or absence of a voltage peak in switching measurements of superconducting Josephson Junctions is a binary response.
- Fatality (fatal/non-fatal) of myocardial infarction (heart attack) is a binary response.
- Level of education (primary school, secondary school, B.Sc., M.Sc., PhD) is an ordinal response.
- The date of an event of coronary heart disease measured for a cohort of people is a time-to-event (or survival) response.

There are also situations where a linear model may be suitable although strictly speaking the response has an inappropriate distribution.

- Height of an adult is positive but can be modeled by linear model because all values are far from zero.

- The daily number of customers in a big supermarket is actually a count response but could be modeled by linear model because all values are far from zero and the number of possible values of the response is high.

1.5 Link functions

The choice of the link function $g()$ depends on the data, especially on the type of the response variable. If the response is a count, i.e. an integer, log-link $g(\mu_i) = \log(\mu_i)$ may be used. Log-link leads multiplicative model

$$\mu_i = \exp(\eta_i) = e^{\beta_1 x_{i1}} e^{\beta_2 x_{i2}} \cdots e^{\beta_p x_{ip}} \tag{7}$$

If the response Y_i is a binary variable with possible values 0 and 1, it holds

$$\mu_i = \mathrm{E}(Y_i) = 1 \cdot P(Y_i = 1) + 0 \cdot P(Y_i = 0) = P(Y_i = 1). \tag{8}$$

The logit-link

$$g(\mu_i) = \operatorname{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) \tag{9}$$

is maybe the most typical choice for binary response data. For positive continuous responses typical link functions are inverse link

$$\mu_i^{-1} = \eta_i \tag{10}$$

and inverse-squared link

$$\mu_i^{-2} = \eta_i. \tag{11}$$

1.6 Confusing terminology

1.6.1 Generalized linear model (GLM) and general linear model (GLM)

Unfortunately, the acronym GLM is sometimes used for general linear model. General linear model is a linear model. The word ‘general’ is used to indicate that the response $\mathbf{Y}$ may be multivariate and the covariates may include both continuous and categorical variables. In SAS, PROC GLM fits a general linear model, not a generalized linear model.

1.6.2 Names of X and Y

In different applications X and Y have various names that sometimes might be confusing. Examples are given below. Some of the names are synonyms and some have special emphasis in certain applications. Particularly, the terms ‘independent variable’ and ‘dependent variable’ may cause a confusion.

Names of X

- covariate
- explanatory variable
- factor
- risk factor
- exposure (variable)
- design variable
- controlled variable
- carrier variable
- regressor
- predictor
- input
- determinant
- *independent variable

Names of Y

- response
- explained variable
- outcome
- responding variable
- regressand
- experimental variable
- measured variable
- output
- *dependent variable

2 Generalized linear models in statistical software

2.1 Generalized linear models in R

In R (www.r-project.org) generalized linear models can be fitted using function glm. The syntax is

```
glm(formula, family = gaussian, data, weights, subset, na.action,
start = NULL, etastart, mustart, offset, control = glm.control(...),
model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, contrasts
= NULL, ...)
```

Arguments
Some important arguments are

`formula` an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted.

`family` a description of the error distribution and link function to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function.

`data` an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. `weights`] an optional vector of weights to be used in the fitting process.

`subset` an optional vector specifying a subset of observations to be used in the fitting process.

`offset` can be used to specify an a priori known component to be included in the linear predictor during fitting. This should be NULL or a numeric vector of length either one or equal to the number of cases. One or more offset terms can be included in the formula instead or as well, and if both are specified their sum is used. See model.offset.

`control` a list of parameters for controlling the fitting process.

Output
As an output an object of class "glm" is returned. A glm object is a list that contains the following components among the others:

`coefficients` a named vector of coefficients

`fitted.values` the fitted mean values, obtained by transforming the linear predictors by the inverse of the link function.

`deviance` up to a constant, minus twice the maximized log-likelihood. Where sensible, the constant is chosen so that a saturated model has deviance zero.

`aic` Akaike's An Information Criterion, minus twice the maximized log-likelihood plus twice the number of coefficients (so assuming that the dispersion is known).

`null.deviance` The deviance for the null model, comparable with deviance.

`iter` the number of iterations of IWLS used.

`df.residual` the residual degrees of freedom.

`df.null` the residual degrees of freedom for the null model.

`converged` logical. Was the IWLS algorithm judged to have converged?

Example: binomial family with logit-link (logistic regression)

```
set.seed(3000)
b<-3;
n<-500;
x<-rnorm(n);
y<-runif(n)<exp(b*x)/(1+exp(b*x))
m1<-glm(y~x,binomial(link = "logit"))
print(summary(m1))
```

Summary:

```
Call:
glm(formula = y ~ x, family = binomial(link = "logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.66224  -0.53516   0.01267   0.45869   2.62460
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.01346    0.13572  -0.099    0.921
x            3.27787    0.29793  11.002   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 693.12  on 499  degrees of freedom
Residual deviance: 342.09  on 498  degrees of freedom
AIC: 346.09

Number of Fisher Scoring iterations: 6
```

2.2 Generalized linear models in SAS, Matlab and SPSS

There are several procedures in SAS for generalized linear models. PROC GLM (where G stands for 'general' not for 'generalized') can be used to fit and test linear models. Binary and categorical response data can be handled with PROC LOGISTIC, PROC PROBIT, PROC CATMOD and PROC GENMOD. PROC GENMOD is based on the philosophy of generalized linear models and allows user-defined link functions in addition to the commonly used link functions.

In Matlab, Statistics toolbox has function `glmfit` and `glmval`. SPSS Advanced Statistics contains the module GENLIN.

3 Theory of generalized linear models

3.1 Notation

The observed data set $(\mathbf{y}, \mathbf{X})$ contains n observations of $1 + p$ variables

$$\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \dots y_n \end{pmatrix}^T \tag{12}$$

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots x_{1p} \\ x_{21} & x_{22} & \dots x_{2p} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots x_{np} \end{pmatrix}. \tag{13}$$

Variable y is the response variable and variables $x_1, x_2, \dots x_p$ are explanatory variables or covariates. The observed value y_i is treated as a realization of a random variable Y_i. In experimental setup, the explanatory variables have fixed values set by the experimenter. In observational setup, the value x_{ij} can be understood to be a realization of a random variable X_{ij} but when distribution of Y_i is considered x_{ij} is taken as fixed.

The parameters include the regression coefficients

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 & \beta_2 & \dots & \beta_p \end{pmatrix}^T, \tag{14}$$

the linear predictors

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_1 & \eta_2 & \dots & \eta_n \end{pmatrix}^T, \tag{15}$$

the expected responses

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_n \end{pmatrix}^T, \tag{16}$$

and the canonical parameters

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 & \theta_2 & \dots & \theta_n \end{pmatrix}^T. \tag{17}$$

3.2 Model assumptions

1. The distribution of Y_i belongs to the exponential family. For the exponential family, the density function can be presented in the form

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left(\frac{a_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi/a_i) \right), \tag{18}$$

   where

- θ_i, $i = 1, \ldots, n$ are unknown parameters (canonical parameters),
- ϕ is the dispersion parameter (scale parameter) that can be known or unknown,
- a_i, $i = 1, \ldots, n$ are known prior weights of each observation and
- $b()$ and $c()$ are known functions. The first derivative $b'()$ is monotonic and differentiable.

2. Random variables $Y_1, Y_2, \ldots, Y_n$ are mutually independent.

3. The expected value $\mu_i = \mathrm{E}(Y_i)$ depends on linear predictor $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$ through monotonic and differentiable link function g

$$g(\mu_i) = \eta_i. \tag{19}$$

For instance, normal, binomial, Poisson and gamma distributions belong to the exponential family. For exponential family (18) it holds

$$\mathrm{E}(Y_i) = b'(\theta_i) = \mu_i \tag{20}$$

and

$$\mathrm{Var}(Y_i) = \frac{b''(\theta_i)\phi}{a_i} = \frac{V(\mu_i)\phi}{a_i}. \tag{21}$$

As shown in section 3.8, the assumption on the exponential family can be relaxed.

3.3 Likelihood

The log-likelihood of $y_1, \ldots, y_n$ from an exponential family with known dispersion parameter ϕ can be written

$$l(\theta_1, \ldots, \theta_n; \phi, \mathbf{a}, \mathbf{y}) = \sum_{i=1}^{n} \left(\frac{a_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi/a_i) \right) \tag{22}$$

If there are no restrictions for parameters $\theta_1, \ldots, \theta_n$, the model is saturated, i.e. it has as many parameters as there are observations. In a GLM, the parameters $\theta_1, \ldots, \theta_n$ depend on $\mathbf{X}$ and the parameters $\beta_1, \ldots, \beta_p$ through functions $b()$ and $g()$

$$\sum_{j=1}^{p} \beta_j x_{ij} = \eta_i = g(\mu_i) = g(b'(\theta_i)). \tag{23}$$

Therefore, the log-likelihood can be written also a function of the parameters $\mu_1, \ldots, \mu_n$ or as a function of the parameters $\beta_1, \ldots, \beta_p$

$$l(\mu_1, \ldots, \mu_n; \phi, \mathbf{a}, \mathbf{y}) = \sum_{i=1}^{n} \left(\frac{a_i(y_i(b')^{-1}(\mu_i) - b((b')^{-1}(\mu_i)))}{\phi} + c(y_i, \phi/a_i) \right), \tag{24}$$

$$l(\beta_1, \ldots, \beta_p; \phi, \mathbf{a}, \mathbf{y}) = \sum_{i=1}^{n} \left(\frac{a_i(y_i(b')^{-1}(g^{-1}(\sum_{j=1}^{p} \beta_j x_{ij})) - b((b')^{-1}(g^{-1}(\sum_{j=1}^{p} \beta_j x_{ij}))))}{\phi} + c(y_i, \phi/a_i) \right). \tag{25}$$

3.4 Canonical link

The link function for which it holds $\eta_i = g(\mu_i) = \theta_i$ is called canonical link. Because $\mu_i = b'(\theta)$, it follows $g = (b')^{-1}$. The use of canonical link function simplifies calculations but this alone does not justify the use of canonical link. The link function should be selected on the basis of the data and prior knowledge on the problem.

3.5 Score function, observed information and expected information (Fisher information)

The partial derivative of log-likelihood with respect to some parameter is called score or score function. In the case of the exponential family (22) we obtain

$$\frac{\partial l}{\partial \theta_i} = \frac{a_i(y_i - b'(\theta_i))}{\phi}, \tag{26}$$

$$\frac{\partial l}{\partial \mu_i} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i} = \frac{a_i(y_i - b'(\theta_i))}{\phi}\frac{1}{V(\mu_i)}, \tag{27}$$

$$\frac{\partial l}{\partial \eta_i} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i} = \frac{a_i(y_i - b'(\theta_i))}{\phi}\frac{1}{V(\mu_i)}(g^{-1})'(\eta_i), \tag{28}$$

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}\frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n}\frac{a_i(y_i - b'(\theta_i))}{\phi}\frac{1}{V(\mu_i)}(g^{-1})'(\eta_i)x_{ij} =$$

$$\frac{1}{\phi}\sum_{i=1}^{n}\frac{a_i(y_i - \mu_i(\boldsymbol{\beta}))x_{ij}}{V(\mu_i(\boldsymbol{\beta}))g'(\mu_i(\boldsymbol{\beta}))} \tag{29}$$

where the notation $\mu_i(\boldsymbol{\beta})$ emphasizes the fact that μ_i depends on $\boldsymbol{\beta}$.

The observed information is the negative of the matrix of second order partial derivatives of log-likelihood

$$J(\boldsymbol{\beta}, \mathbf{y}) = -\frac{\partial^2 l(\boldsymbol{\beta}, \mathbf{y})}{\partial \boldsymbol{\beta}^2} = \begin{pmatrix} -\sum_{i=1}^{n}\frac{\partial^2 l(\boldsymbol{\beta}, y_i)}{\partial \beta_1^2} & \dots & -\sum_{i=1}^{n}\frac{\partial^2 l(\boldsymbol{\beta}, y_i)}{\partial \beta_1 \partial \beta_p} \\ \vdots & \ddots & \vdots \\ -\sum_{i=1}^{n}\frac{\partial^2 l(\boldsymbol{\beta}, y_i)}{\partial \beta_p \partial \beta_1} & \dots & -\sum_{i=1}^{n}\frac{\partial^2 l(\boldsymbol{\beta}, y_i)}{\partial \beta_p \partial \beta_p} \end{pmatrix} \tag{30}$$

and the Fisher information or expected information is the expected value of observed information

$$I(\boldsymbol{\beta}) = \mathrm{E}_{\mathbf{Y}}(J(\boldsymbol{\beta}, \mathbf{Y})) = \sum_{i=1}^{n}\mathrm{E}_{Y_i}(J(\boldsymbol{\beta}, Y_i)) = -\sum_{i=1}^{n}\mathrm{E}\left(\frac{\partial^2 l(\boldsymbol{\beta}, Y_i)}{\partial \boldsymbol{\beta}^2}\right). \tag{31}$$

3.6 Estimation

The maximum likelihood estimate for $\boldsymbol{\beta}$ is obtained by solving score equations

$$\frac{\partial l(\boldsymbol{\beta}, \mathbf{y})}{\partial \boldsymbol{\beta}} = 0. \tag{32}$$

Usually the estimation requires numerical methods. Traditionally, the maximum likelihood estimation is carried out with Fisher scoring (also called iterative weighted least squares) which is a modification of the Newton-Raphson algorithm.

In Newton-Raphson update rule

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + J^{-1}\frac{\partial l(\boldsymbol{\beta}, \mathbf{y})}{\partial \boldsymbol{\beta}} \tag{33}$$

the observed information J is replaced by the expected information I. After some algebra, this leads to the update formula

$$\hat{\boldsymbol{\beta}}^{(t+1)} = (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{z}^{(t)}, \tag{34}$$

where

$$\mathbf{W}^{(t)} = \begin{pmatrix} w_1^{(t)} & & \\ & \ddots & \\ & & w_1^{(t)} \end{pmatrix}, \tag{35}$$

$$w_i^{(t)} = \frac{a_i}{\left[g'\left(\mu_i(\hat{\boldsymbol{\beta}}^{(t)})\right)\right]^2 V\left(\mu_i(\hat{\boldsymbol{\beta}}^{(t)})\right)}, \tag{36}$$

$$\mathbf{z}^{(t)} = (z_1^{(t)} \dots z_n^{(t)})^T \tag{37}$$

$$z_i^{(t)} = \eta_i(\hat{\boldsymbol{\beta}}^{(t)}) + (y_i - \mu_i(\hat{\boldsymbol{\beta}}^{(t)}))g'\left(\mu_i(\hat{\boldsymbol{\beta}}^{(t)})\right). \tag{38}$$

It can be seen that the updating rule depends on the distribution of Y_i only through the variance function V.

When the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ exists, it is consistent and asymptotically normal with expected value $\boldsymbol{\beta}$ and covariance matrix $\phi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$.

The dispersion parameter ϕ can estimated by the deviance (see Section 3.7) estimator

$$\hat{\phi} = \frac{D}{n-p} \tag{39}$$

or the moment estimator

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{a_i(y_i - \mu_i(\hat{\boldsymbol{\beta}}))^2}{V(\mu_i(\hat{\boldsymbol{\beta}}))}. \tag{40}$$

3.7 Deviance

Deviance is defined as

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\phi(l(\mathbf{y}; \mathbf{y}) - l(\hat{\boldsymbol{\mu}}; \mathbf{y})) \tag{41}$$

where $l(\mathbf{y};\mathbf{y})$ is the log-likelihood of the saturated model (full model). In the saturated model, the number of parameters equals the number of observations and likelihood obtains its maximum for the model class. Scaled deviance is defined as

$$D^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = \frac{D(\mathbf{y};\hat{\boldsymbol{\mu}})}{\phi} \tag{42}$$

As seen in Section 4.5, deviance is closely related to the likelihood ratio test.

3.8 Quasi-likelihood

GLMs allow defining the variance function independently from the link function. The assumption that the distribution of Y_i belongs to the exponential family can be replaced by an assumption that concerns only the variance of Y_i

$$\mathrm{Var}(Y_i) = \frac{\phi V(\mu_i)}{a_i}. \tag{43}$$

Parameters can be estimated maximizing quasilikelihood

$$Q(\boldsymbol{\beta};\mathbf{y}) = \frac{1}{\phi}\sum_{i=1}^{n}\int_{y_i}^{\mu_i}\frac{a(y_i-t)}{V(t)}dt. \tag{44}$$

The form of quasilikelihood function is chosen so that partial derivatives

$$\frac{\partial Q(\boldsymbol{\beta};\mathbf{y})}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\frac{a_i(y_i-\mu_i(\boldsymbol{\beta}))x_{ij}}{V(\mu_i(\boldsymbol{\beta}))g'(\mu_i(\boldsymbol{\beta}))}. \tag{45}$$

are similar to the partial derivatives of likelihood function and consequently the parameters can be estimated by Fisher scoring.