Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 71(3) 2017 207–215

Modelling Binary Data by Bayesian Logistic Regression with Random Intercepts

Himadri Shekhar Roy[1], Ranjit Kumar Paul[1], Lal Mohan Bhar[1] and Amrender Kumar[2]
[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi
[2]ICAR-Indian Agricultural Research Institute, New Delhi

Received 28 April 2016; Revised 01 June 2017; Accepted 02 June 2017

SUMMARY

In some cases, occurrence of binary data may vary in different spatio-temporal situations. The logistic regression fails to model binary data from clustered, multi-level and longitudinal studies, because of dependency among the observations. Under such situation, random effects can be included in the linear predictor of the logistic regression model in order to allow for correlated responses. The estimation of parameter of the binary logistic regression with random effects (LRRE) is not straight forward due to the fact that the likelihood involves multiple integrals and the explicit derivation of the integrals is not possible. The Bayesian paradigm provides a natural approach to inference in mixed models. In this paper we modeled the occurrence of aphid populations in two different locations in India using logistic regression with random effects in Bayesian paradigm. We consider that the response variable is binary in nature. The weather variables *i. e.* temperature, relative humidity and their interaction are taken as covariates. We assume different prior distributions for the random intercepts parameters. As we know odds ratio depends on random intercepts, so it is also shown that odds ratio is random. Hence median odds ratio is calculated and it is shown that median odds ratio is robust and better measure than odds ratio. SAS software version 9.4 has been used for present analysis.

Keywords: Aphid population, Bayesian paradigm, LRRE model, Random intercept.

1. INTRODUCTION

Study of binary response data become frequent in several field of research including medical, social, psychological, economics and agricultural research. In many research areas like epidemiologic and biological studies logistic regression is an important tool for analysis of binary and categorical response data. Different designs are adopted for this kind of studies, such as, for longitudinal studies response are measures on same subject repeatedly over time, in genetic studies where family members are closely related and in agricultural studies naturally defined clusters are correlated within clusters. Study of occurrence of pests and diseases is very important in agriculture. Various models are developed for studying various aspects of pests and diseases. Mainly multiple linear regression models are used for the purpose. In these models, pest/disease infestation is taken as dependent variable and some weather variables are taken as independent variables. Off course, some non-linear models are also reported. In these multiple linear regression models, dependent variable, *i.e.*, pest/disease infestation is quantitative. However, there occurs some situations where we do not get quantitative data rather some gradation. For example, pest/disease population is categorized like severe, mild etc. or occurrence or nonoccurrence of pest/disease population instead of having actual quantity. Under such situations, logistic regression model is a good choice. It becomes a standard tool for the analysis of binary data. Moreover, logistic regression has a nice interpretation in terms of odds ratio (OR).

In some cases, occurrence of binary data may vary in different spatio-temporal situations. For example, occurrence or non-occurrence of pest/disease may vary for a prolonged period of time, say several weeks.

Corresponding author: Lal Mohan Bhar
E-mail address: lmbhar@gmail.com

If a logistic model is fitted with whole set of data, then the parameters estimates particularly, the intercept term may be upward biased. As a result the estimated odds ratio will no longer reflects the true scenario of the situation. Actually logistic regression fails to model binary data from clustered, multi-level and longitudinal studies, because of dependency among the observations. Under such situation, inspired by the theory of linear normal models, random effects can be included in the linear predictor of the logistic regression model in order to allow for correlated responses (Larsen *et al.* 2000, Diggle *et al.* 2002, Paulino *et al.* 2005). Here the model accounts for the covariance among the measures in a relatively parsimonious way. Such models are also known as multi-level logistic regression model which comes under the broad category of Linear Mixed Models (LMM). In multi-level logistic regression model, intercept term is no longer fixed. It becomes a random effect in the next higher level making odds ratio a random variable. Thus such logistic regression models with random intercept term also helps to measure the heterogeneity present among the groups or states with the help of random odds ratio.

The estimation of parameter of the binary logistic regression with random effects (LRRE) is not straight forward due to fact that the likelihood involves multiple integrals and the explicit derivation of the integrals is not possible (Zeger and Karim 1991, Breslow and Clayton 1993, Hedeker and Gibbons 1996). The Bayesian paradigm provides a natural approach to inference in mixed models. The random effects are treated as parameters to be estimated. In case of Bayesian analysis, a natural approach to inference in mixed models was proposed by Paulino *et al.* (2005). They estimated the random effects which were treated as parameters in the presence of misclassified data. They also showed that if the posterior distribution was not possible to be obtained analytically, Markov Chain Monte Carlo (MCMC) method could be used to approximate them. Souza and Migon (2010) proposed that inference problem could be solved in easier way if the random effects of the mixed models are distributed as Student-t or finite mixture of normal distributions. Lui and Dey (2008) used prior distributions like skew-normal and non-parametric distribution. Ten Have and Localio (1999) used an empirical Bayes approach for logistic regression with random effects models. Santos *et al.* (2013) provided different prior and posterior interpretations for the parameters in the logistic regression model with random intercepts when skew normal distribution are assumed to model random effects. They obtained the prior distributions for the different parameters and showed odds ratio and median odds ratio under skew normality for the random effects. They showed the misspecifications of the random effects distributions for estimating of odds ratio. They finally concluded that the misspecification of the random effects parameters gives poor estimates.

Zeger *et al.* (1988) showed that when random effects are introduced in the model, fixed effects parameters do not maintain their interpretational features. Parameter interpretation in logistic regression with random effects (LRRE) was first considered by Larsen (2000). Larsen and Marlo (2005) showed that LRRE does not inherit the interpretational features of the standard logistic regression model. Larsen *et al.* (2000) examined the interpretation of both fixed effects and random effects parameters and showed that the random effects parameters are not easily interpreted in LRRE model in case of heterogeneity. They showed that odds ratio depends on the random effects parameters. In that case, Median odds ration which is the function of the original random effects parameters are used as an alternative to odds ratio.

In this paper, we model the occurrence of aphid populations of mustard crop in two different locations in India using logistic regression with random effects in Bayesian paradigm. We consider that the response variable is binary in nature *i.e.* response variable assumes value 1 if the average number of population is greater than the threshold value *i.e.* ≥ 30 (Saunakiya and Tiwari 2014) and otherwise it takes 0. The weather variables *i.e.* temperature, relative humidity and their interaction covariates are taken as covariates. Main focus is to interpret the parameters of this model. We assume different prior distributions for the random intercepts parameters. As we know odds ratio depends on random intercepts, so it is also shown that odds ratio is random. Hence median odds ratio is calculated and it is shown that median odds ratio is robust and better measure than odds ratio.

This paper is organized as follows. Section 2 describes the logistic regression with random effects in general. Section 3 discusses the estimation procedure

of the model parameters and odds ratio with Markov Chain Monte Carlo technique. In Section 4, an illustration is made to interpret the parameters of the models and odds ratio. The paper is concluded with a section on Conclusion.

2. LOGISTIC REGRESSION WITH RANDOM INTERCEPT

Logistic regression with random intercept model can be described by using mixed effect model approach. The linear mixed-effect models are the generalization of linear regression model by including random intercept other than random error component.

For controlling some factors, we generally fit a linear regression model. Single level regression model is given by

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad e_i \sim N(0,\ \sigma_e^2\ ) \tag{2.1}$$

Here both the parameters β_0 and β_1 are fixed. The random part is the only error term which is assumed to follow normal distribution. But when the observations are clustered, it is not possible to identify the variations at each level. Let y_{ij} be the *i*th observation in the *j*th cluster or group and variations can be explained by the following model

$$y_{ij} = \beta_0 + u_j + e_{ij} \qquad e_{ij} \sim N(0,\sigma_e^2)\ ,\ u_j \sim N(0,\sigma_u^2) \tag{2.2}$$

where, the variation at the two levels *i.e.* the between and within is given as u_j and e_{ij}. To see the effects of them, some explanatory variables can also be included in this model. By combining the models (2.1) and (2.2), random intercept model is given as

$$y_i = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \tag{2.3}$$

The random intercept model is mainly divided in two parts. $\beta_0 + \beta_1 x_{ij}$ is called as fixed part *i.e.* the intercept and the coefficient of the explanatory variable times the explanatory variable and $u_j + e_{ij}$ is the random part. So the parameters for the fixed part are the coefficients β_0, β_1 and so on and the parameters for the random part are the variances, σ_u^2 and σ_e^2, both are assumed to be normal distributed generally.

For binary responses, $E(y_{ij}) = \pi_{ij} = \Pr(y_{ij} = 1)$ and $F^{-1}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + u_j$, where F^{-1} is the link function. By using logit link function, it can be written as

$$\log\left[\frac{\Pr\left(y_{ij}=1\right)}{1-\Pr\left(y_{ij}=1\right)}\right] = \beta_0 + \beta_1 x_{ij} + u_j \tag{2.4}$$

In matrix notation, the model can be written as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon} \tag{2.5}$$

where $\mathbf{y}$ is a $n\times1$ vector of responses, $\mathbf{X}$ is a $n\times p$ design matrix of covariates of fixed effects $\boldsymbol{\beta}$, $\mathbf{Z}$ is the $n\times q$ design/covariate matrix for the random effects $\mathbf{u}$, $\boldsymbol{\varepsilon}$ is $n \times 1$ vector of errors assumed to be multivariate normal with mean $\mathbf{0}$ and variance matrix $\sigma_\varepsilon^2\mathbf{R}$.

The equation 2.5 has two parts, $\mathbf{X\beta}$ is the fixed part which is analogous to the linear predictor from a standard OLS regression model where $\boldsymbol{\beta}$ is the regression parameters and $\mathbf{Zu} + \boldsymbol{\varepsilon}$ is the random part. Let $\mathbf{u}$ has a variance covariance matrix $\boldsymbol{\Sigma}$ and it is assumed that $\mathbf{u}$ is orthogonal to $\boldsymbol{\varepsilon}$ such that

$$Var\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\Sigma & \mathbf{0}\\ \mathbf{0} & \sigma_\varepsilon^2\mathbf{R}\end{bmatrix} \tag{2.6}$$

Here, the random effects $\mathbf{u}$ are not directly estimated but although they may be predicted. The variance of random effects is characterized by the variance component $\boldsymbol{\Sigma}$ that is estimated with the overall residual variance σ_ε^2.

For the clustered observations, instead of taking all n observations at once, these observations are organized in k independent groups or clusters.

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i \tag{2.7}$$

For i=1,2,…, k with cluster i consisting of n_i observations, $\mathbf{y}_i$ is the response corresponding to the i^{th} cluster. $\mathbf{u}_i$ is the $q\times1$ random effects which is normally distributed with mean $\mathbf{0}$ and $q\times q$ covariance matrix $\boldsymbol{\Sigma}$. Here the matrix $\mathbf{Z}_i$ is the $n_i\times q$ design matrix for i^{th} cluster random effects.

According to Larsen *et al.* (2000), the response variable $\mathbf{y}_i$ follows Bernoulli distribution with probability distribution

$$P(\mathbf{y}_i = 1 \mid \eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \qquad \eta_i = \mathbf{x}_i\beta + \mathbf{z}_i\mathbf{u} \tag{2.8}$$

where, $\boldsymbol{\beta}$ is the parameters vector of fixed effects, x_i is the i^{th} row of $n\times p$ design matrix $\mathbf{X}$ for the fixed effects, $\mathbf{u}$ is the vector of random effects which are normally distributed with mean $\mathbf{0}$ and variance-covariance

matrix $\boldsymbol{\Sigma}$ and $\mathbf{z}_i$ is the i^{th} row of $n \times q$ design matrix of random effect $\mathbf{Z}$.

Logistic regression with random effects yields the same result and same interpretation by conditioning the random effects in terms of odds ratio as in the case of ordinary logistic regression model. As the random effects are unobservable so in general it is not possible to condition the random effects. It is assumed that for a sample of n individuals where $\boldsymbol{\beta}$, $\mathbf{u}$ are the parameter vector of fixed effects, vector of random effects respectively and $\mathbf{X}$ is $n \times p$ matrix with information related with covariates of all observations, then the response $y_{ij} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{X}$ follows independent $\text{Ber}(\pi_{ij})$. The likelihood function is given by

$$f(y_{ij} | \boldsymbol{\beta},\ \mathbf{u},\ \mathbf{X}) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}\left[\frac{\exp(\boldsymbol{\eta}_{ij})}{1+\exp(\boldsymbol{\eta}_{ij})}\right]^{y_{ij}}\left[\frac{1}{1+\exp(\boldsymbol{\eta}_{ij})}\right]^{1-y_{ij}} \tag{2.9}$$

where, $\mathbf{y} = \left(y_{11},\ldots,y_{1n_1},\ldots,y_{k1,\ldots,}y_{kn_k}\right)'$.

As mentioned earlier that the population is divided in k clusters and a sample of n_i is selected into the i^{th} cluster. Let y_{ij} be the response variable (y_{ij} is 1 if success occurs and 0 otherwise) for ith cluster and jth individual (Larsen *et al.* 2000).

In statistical analysis, the choice of likelihood is very important. Normal likelihood is not the appropriate choice of distribution to model the data, when the researcher is looking at success/failure data or even count data. For discrete data set, binomial likelihood is used to model the counting the number of successes in a sequence of n independent Bernoulli trials with the corresponding probability p, which is the probability of success of each experiment. It is known that the odds for an experiment are found as $\frac{p}{1-p}$. The log of the odds will be set equal to the regression line with an intercept and coefficients for each of the covariates,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{2.10}$$

But here using the random effects it can be represented as

$$\log\left(\frac{p}{1-p}\right) = \exp(\eta_i) \text{ where } \eta_i = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{z}$$

The logit transformation allows for values in the regression equation along the entire real line, but also keeps p in its restricted interval. Thus, p is transformed from the real line to the interval $0 \le p \le 1$ and the parameter space is preserved *i.e.* Logit transformation allows the β's to be in real line, while preserving the parameter space of the parameter p. Hence logit transformation is used because this function keeps things in their proper domain. It links the regression line to the binomial probability p.

In the usual logistic regression model, the fixed effects can be interpreted by odds ratio between the highest and the lowest risk individuals. In the LRRE model Larsen *et al.* (2000) firstly discussed the interpretation of the fixed effects. They showed that the odds ratio depends on both the fixed and random effects. Consequently, several other useful interpretations are drawn from such quantity. Let j_1 and j_2 be two individuals of the different clusters i_1 and i_2, respectively. Then the odds ratio is given by

$$OR = \exp\left\{\left(\mathbf{x}'_{i_1 j_1} - \mathbf{x}'_{i_2 j_2}\right)\boldsymbol{\beta} + \left(\mathbf{z}'_{i_1} - \mathbf{z}'_{i_2}\right)\mathbf{u}\right\} \tag{2.11}$$

If the comparison is between individuals in the same cluster say, $i_1 = i_2 = i$, but having different covariates, then

$$OR = \exp\left\{\left(\mathbf{x}'_{ij_1} - \mathbf{x}'_{ij_2}\right)\boldsymbol{\beta}\right\} \tag{2.12}$$

which depends only on the fixed effects and is exactly the same as for the usual logistic regression model.

To quantify the random effects, the comparison is done assuming that two individuals, j_1 and j_2 belong to different clusters with same covariate vectors *i.e.*, the individual j_k belongs to cluster i_k, $k = 1, 2$. In this case, the odds ratio depends on the random effects only and is given by

$$OR = \exp\left\{\left(\mathbf{z}'_{i_1} - \mathbf{z}'_{i_2}\right)\mathbf{u}\right\} \tag{2.13}$$

The odds ratio in equation 2.11 also permits the comparison between the individuals with the highest risk in two different clusters, among others.

Under the classical approach for inference, the *OR* is a random quantity only when the comparison depends on the random effects. Since it is random, Larsen *et al.* (2000) propose to interpret the *OR* in terms of the median of its distribution. The so-called median odds ratio is named here *MOR*. According to

Larsen *et al.* (2000), the *MOR* quantifies appropriately the heterogeneity among the different clusters. For the general case, whenever we are comparing individuals with different covariates in different clusters, the median odds ratio is defined as

$$MOR = \exp\left\{\left(\mathbf{x}'_{i_1 j_1} - \mathbf{x}'_{i_2 j_2}\right)\boldsymbol{\beta} + \left(\mathbf{z}'_{i_1} - \mathbf{z}'_{i_2}\right)\mathbf{u}, \mathbf{X}\right\} \quad (2.14)$$

Larsen *et al.* (2000) shows that if $\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, then $u_{i_1} - u_{i_2} \sim N(0, \sigma_{12}^2)$, where σ_{12}^2 is the variance of $u_{i_1} - u_{i_2}$.

3. BAYESIAN ESTIMATION AND INFERENCE IN LOGISTIC REGRESSION MODELS WITH RANDOM INTERCEPT

To simplify the inferential process in the logistic regression with random intercept, the random effects are usually assumed to be independent with a common normal distribution. Such an assumption, however, is questionable in some biological data as shown in Liu and Dey (2008), for instance. However, there are instances where other types of distributions are also used like Skewed Normal (SN) distribution. Again, the random effects can be considered as independent as well as correlated.

From the Bayesian point of view, inference for mixed models is simpler since the random effects $\mathbf{u} = (u_1, u_2, ..., u_k)'$ are considered as unknown quantities to be estimated. Assuming the likelihood in 2.9, we should elicit prior distributions for the parameters $\boldsymbol{\beta}$ and $\mathbf{u}$. We assume *i.i.d* random effects follow univariate normal distribution. We also center the normal distribution on zero to avoid non-identifiability. The prior distributions for the fixed effects *i.e.* $\boldsymbol{\beta}$ is assumed as normal with mean zero and constant variance, σ^2. The hyperparameter *i.e.* σ^2 is assumed as inverse Gamma (IG) distribution, $\sigma^2 \sim$IG(*a, d*), where *a* is the scale parameter and *d* is the shape parameter. Thus $u_i \sim N(0, \sigma^2)$. *IG*(*a, d*) distribution will have now $E(\sigma^2) = d(a-1)-1$ and $V(\sigma^2) = d[(a-1)2(a-2)]-1$.

The posterior distribution under the present situation does not have any closed form. So analytically, it is a hard task to obtain the posterior distribution of odds ratio (*OR*) as well as the parameter estimates. Therefore Markov Chain Monte Carlo (MCMC) scheme is considered to sample from it.

Once the parameter estimates are obtained, the distribution of odds ratio can be obtained and hence its estimate. Odds ratio depends on both the parameters $\boldsymbol{\beta}$ and the random effects $\mathbf{u}$. MCMC methods provide good approximation for posterior distribution of *OR* as follows:

Step1: Generation of sample $\left(\boldsymbol{\beta}^{(l)}, \mathbf{u}^{(l)}\right)$ from the posterior distributions at the *l*th iteration:

(i) Consider the 2-level normal response model

$$y_{ij} = \beta_0 + x_{ij}\beta_1 + u_j + e_{ij};$$
$$u_j \sim N(0, \sigma_u^2), e_{ij} \sim N(0, \sigma_e^2)$$

(ii) MCMC algorithms usually work in a Bayesian framework and there is need to add prior distributions for the unknown parameters and the set of unknown parameters is ($\boldsymbol{\beta}$, $\mathbf{u}$, σ_u^2, σ_e^2).

(iii) One possible MCMC algorithm for the model involves simulation from the sets of conditional distributions, *i.e.*, $p(\boldsymbol{\beta})$, $p(\mathbf{u})$, $p(\sigma_u^2)$, $p(\sigma_e^2)$.

(iv) Initialize the values for each group of unknown parameters $\boldsymbol{\beta}^{(0)}, \mathbf{u}^{(0)}, \sigma_u^{2(0)}, \sigma_e^{2(0)}$.

(v) For getting $\boldsymbol{\beta}^{(1)}$, sample from the following conditional distributions

$$p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{u}^{(0)}, \sigma_u^{2(0)}, \sigma_e^{2(0)})$$

Now similarly takes more samples from $p(\mathbf{u} \mid \mathbf{y}, \boldsymbol{\beta}^{(1)}, \sigma_u^{2(0)}, \sigma_e^{2(0)})$, $p(\sigma_u^2 \mid \mathbf{y}, \boldsymbol{\beta}^{(1)}, \mathbf{u}^{(1)}, \sigma_e^{2(0)})$ and $p(\sigma_e^2 \mid \mathbf{y}, \boldsymbol{\beta}^{(1)}, \mathbf{u}^{(1)}, \sigma_u^{2(1)})$ to get $\mathbf{u}^{(1)}$, $\sigma_u^{2(1)}$ and $\sigma_e^{2(1)}$ respectively.

Now updating the all unknown parameters to the model, the process need to be repeated until convergence criterion is met. Each time using the previously generated parameter values to generate the next set.

Step 2: Finally odds ratio is computed

$$OR^l = \exp\left(\left\{\mathbf{x}_{i_1 j_1} - x_{i_2 j_2}\right)' \boldsymbol{\beta}^l + u_{i_1}^l - u_{i_2}^l\right\}$$

Steps 1 must be repeated until all parameters has attained convergence and obtain samples from the posterior distributions of reasonable size. The

expression of Step 2 may vary according to the interest of the researcher as discussed earlier.

By this approach it is easy to compute posterior odds ratio summaries and we can also compute Highest Posterior Density (HPD) interval. For comparing two subjects from two randomly selected clusters Odds ratio is used. To test the significance of *OR* in Bayesian framework HPD interval is frequently used. HPD interval have natural interpretations in terms of probability. Here our main interest is to test the null hypothesis $H_0: \ \mathrm{OR}_{12} = 1$ which will be accepted if the value one is in the HPD interval.

4. ILLUSTRATION

The daily data on aphid population from two locations namely Bharatpur, Rajasthan (L-I) and Mohanpur, West Bengal (L-II) of India was collected. In case of Bharatpur (L-I) farm, the observations were collected in aphid growing season from 11th November to 22nd March, 2001 and at Mohanpur farm, the observations was collected from 25th October to 6th March, 2003. The daily observations are grouped in different weeks. Fig. 1 reveals the graphical representation of aphid population at Bharatpurand Mohanpur locations. The data consists of 16 groups and 19 groups for Bharatpur and Mohanpur respectively and each group has unequal number of observations. The response variable is converted to binary by taking its value 1 if aphid population exceeds the threshold value *i.e.* 30 otherwise 0 (Saunakiya and Tiwari 2014). The dataset on the covariates *i.e.* daily mean temperature and relative humidity were also collected for the same time period in the studied locations. The daily data on weather variables is grouped in different weeks. These covariates *i.e.* temperature, relative humidity and their interaction are considered as fixed effects in the model.

A perusal of the Fig. 1 indicates that there might be heterogeneity present in the data. Therefore grouping of this data is essential and a logistic regression with random intercepts is eminent. Considering the Bernoulli model it was assumed that the random intercepts may follow normal distribution as well as beta distribution. The fitted model differs from each other because two different distributions for the random intercepts are assumed. The other prior and hyperparameter to be selected for the present dataset are discussed earlier.

4.1 Results and Discussion

Table 1 represents the posterior estimates and HPD intervals with 95% probability for the fixed effects and variance for the random effects parameter. A perusal of table 1 indicates that both temperature and relative humidity have negative impact on aphid population for Bharatpur. It also shows that temperature and relative humidity have the negative effects on the aphid population dynamics for Mohanpur. In both the locations, the variance of the random effect distribution are smaller than 1.0 with high posterior probability.

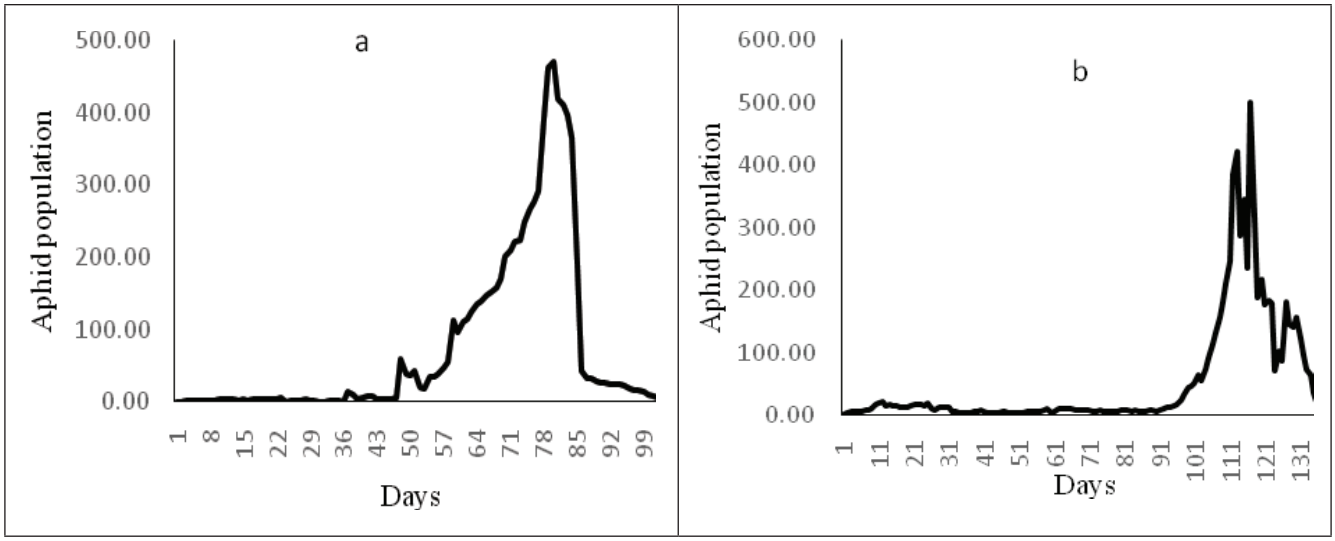

Fig. 1: Daily Aphid population in (a) Bharatpur and (b) Mohanpur

Table 1. Posterior estimates of the parameters

Parameter	Estimates		S.E.		95% HPD intervals	
	L-I	L-II	L-I	L-II	L-I	L-II
β_0	8.47	0.23	0.03	9.75	[8.41, 8.52]	[-18.71, 19.17]
β_1	-0.39	-0.25	0.01	0.47	[-0.41, -0.38]	[-1.19, 0.65]
β_2	-0.08	-0.05	0.001	0.14	[-0.09, -0.09]	[-0.32, 0.22]
β_3	0.001	0.002	0.001	0.01	[0,0.01]	[-0.01, 0.02]
$V(u)$	0.24	0.45	0.38	0.86	[0.01, 0.88]	[0.00, 1.84]

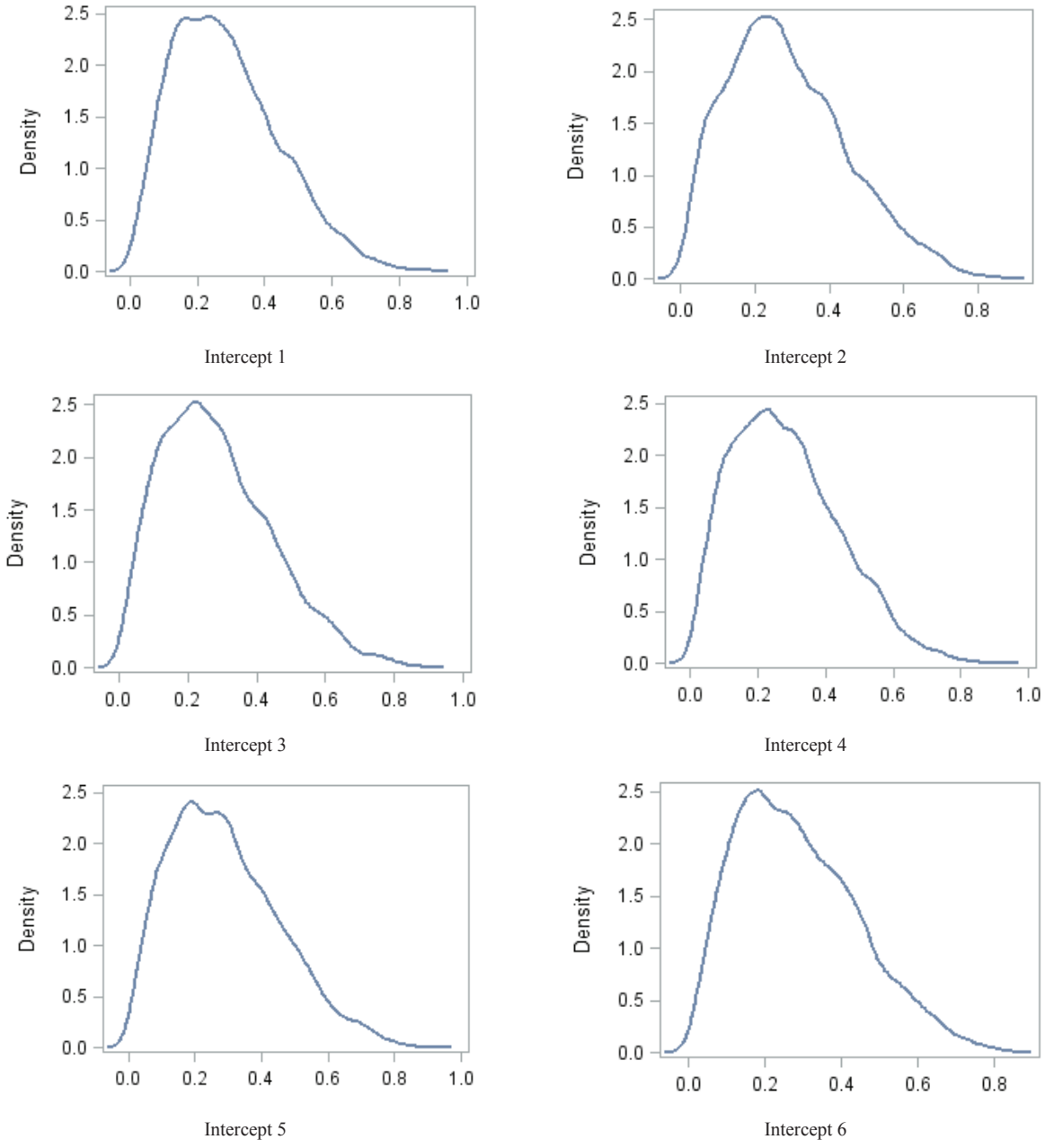

Fig. 2. Random effect densities for Mohanpur under normal distribution

Here Fig. 2 represents the posterior density graph of some random intercepts. Table 2 and 3 provide the posterior summaries of the odds ratios and median odds ratios of different clusters of different observations of the two locations when the intercept terms follow Normal distribution and Beta distribution respectively. Here each cluster denotes the different temperature, relative humidity with different number of aphid populations. For example, the odds ratio for comparing two individuals of 2nd and 1st cluster is 1.46, 3rd and 1st cluster is 1.71, hence we can say that these clusters are heterogeneous in nature.

Here in Table 2 and Table 3 provides posterior summaries of the odds ratio distributions for some particular comparisons of Mohanpur and Bharatpur is described with their corresponding 95% HPD intervals. Table 2 describes the OR when random intercepts follow normal distribution. OR1 compares the cluster 1 and cluster 2 which indicates probable growth of aphid population in cluster 1 is 1.46 times as likely as cluster 2; OR2 compares the cluster 1 and 3 which indicates probable growth of aphid population in cluster 1 is 1.71 times as likely as cluster 3; similarly, OR10 compares the cluster 14 and 7 which indicates probable growth of aphid population in cluster 14 is 5.06 times as likely as cluster 7. Table 3 describes the OR when random intercepts follow Beta distribution. OR1 of Mohanpur compares the cluster 2 and cluster 1 which indicates probable growth of aphid population in cluster 2 is 4.21 times as likely as cluster 1. The other comparisons follow similarly.

We know that odds ratio depends on the random intercepts. So, when odds ratio is computed for different clusters of two different models for different data sets it is seen that for different models, odds ratio varies to a great extent. Similarly we calculated median odds ratio and it is seen that median odds ratio is more robust that the odds ratio at least for the present example.

Table 2. Posterior summaries for some OR and MOR for both locations under Normal distribution

Location: Mohanpur				Location: Bharatpur			
Odds Ratio	**Mean**	**Median**	**HPD**	**Odds Ratio**	**Mean**	**Median**	**HPD**
OR1	1.46	1.17	[0.08, 7.74]	OR1	0.39	0.47	[0.09, 1.31]
OR2	1.71	1.27	[0.03, 6.40]	OR2	0.69	0.69	[0.10, 1.14]
OR3	1.90	1.05	[0.06, 6.64]	OR3	0.32	0.53	[0.34, 1.62]
OR4	2.61	1.07	[0.02, 6.42]	OR4	1.15	1.09	[0.08, 1.16]
OR5	2.46	1.08	[0.01, 7.56]	OR5	0.57	0.93	[0.13, 1.29]
OR6	3.36	1.10	[0.05, 7.94]	OR6	0.60	0.94	[0.08, 1.17]
OR7	0.40	1.11	[0.01, 7.79]	OR7	0.88	0.84	[0.12, 1.06]
OR8	3.71	0.95	[0.02, 7.67]	OR8	2.54	2.14	[0.40, 2.94]
OR9	2.99	1.07	[0.05, 7.32]	OR9	0.65	0.58	[0.25, 0.99]
OR10	5.06	0.97	[0.05, 7.16]	OR10	0.76	0.78	[0.30, 1.08]

Table 3: Posterior summaries for some OR and MOR for both locations under Beta distribution

Location: Mohanpur				Location: Bharatpur			
Odds Ratio	**Mean**	**Median**	**HPD**	**Odds Ratio**	**Mean**	**Median**	**HPD**
OR1	4.21	5.74	[3.52, 6.46]	OR1	0.65	0.64	[0.31, 1.08]
OR2	5.42	4.69	[3.52, 6.44]	OR2	0.49	0.46	[0.31, 1.10]
OR3	0.07	0.05	[3.49, 6.45]	OR3	0.85	0.87	[0.29, 1.12]
OR4	0.28	0.31	[3.52, 6.42]	OR4	1.06	0.95	[0.31, 1.09]
OR5	1.67	1.62	[3.49, 6.51]	OR5	0.85	0.90	[0.29, 1.10]
OR6	5.51	5.20	[3.54, 6.47]	OR6	0.89	0.82	[0.32, 1.08]
OR7	1.90	1.86	[3.50, 6.48]	OR7	0.81	0.31	[0.30, 1.08]
OR8	1.52	1.63	[3.46, 6.39]	OR8	0.33	0.64	[0.31, 1.12]
OR9	2.90	2.80	[3.53, 6.50]	OR9	0.63	0.75	[0.32, 1.08]
OR10	5.42	5.19	[3.48, 6.47]	OR10	0.86	0.87	[0.29, 1.09]

5. CONCLUSION

In this paper, an attempt has been made to show that one can assume a parametric distribution of a random effects model. As an illustration, aphid population dynamics in two locations namely Bharatpur, Rajasthan and Mohanpur, West Bengal have been considered. The parameter interpretation of logistic regression with random intercepts is carried out in Bayesian perspective. In general when we compute odds ratio we ignore the randomness of the intercept terms. But in the present investigation it is shown that odds ratio very much depends on random intercepts. Two probability distributions *i.e.* Normal and beta distribution are assumed for prior of random intercepts. For different clusters and intercepts different odds ratio and median odds ratio is calculated. The results show that median OR are robust in comparison to mean OR, at least for the present example.

REFERENCES

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in general linear mixed models. *J. Amer. Statist. Assoc.,* **88**, 9-25.

Dey, D.K., Chen, M.H. (2000). Bayesian model diagnostics for correlated binary data. In: D.K. Dey, S.K. Ghosh and B.K. Mallick (Eds.) *Generalized linear models: A Bayesian perspective* (pp. 139-159), Marcel Dekker, New York.

Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. 2nd edition. Oxford University Press, New York.

Hedeker, D. and Gibbons, R.D. (2006). *Longitudinal Data Analysis.* Wiley, New York.

Larsen, K., Petersen, J.H., Budtz-Jorgensen, E. and Endahl, L. (2000). Interpreting parameters in the logistic regression model with random effects. *Biometrics*, **56**, 909-914.

Litiere, S., Alonso, A. and Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statist. Medi.*, **27**, 3125-3144.

Longford, T.N. (1994). Logistic regression with random coefficients. *Comput. Statist. Data Anal.*, **17**, 1-15.

Liu, J. and Dey, D.K. (2008). Skew random effects in multilevel binomial models: an alternative to nonparametric approach. *Statist. Model.*, **8**, 221-241.

Ord, K. (1975). Estimation methods for models of spatial interaction. *J. Amer. Statist. Assoc.,***70**, 120-126.

Paulino, C.D., Silva, G., Achcar, J.A. (2005). Bayesian analysis of correlated misclassified binary data. *Comput. Statist. Data Anal.*, **49**, 1120-131.

Santos, C.C., Loschi, R.H. and Arellano-Valle, R.B. (2013). Parameter interpretation in skewed logistic regression with random intercept. *Baye. Anal.,* **8(2),** 381–410**.**

Saunakiya1, A.K., Tiwari, N. (2014). Economic injury and threshold level of lipaphis erysimi (Kalt.) *Inter. J. Life Sci. Res.*, **2**, 178-184.

Souza, A. and Migon, H. (2010). Bayesian outlier analysis in binary regression. *J. Appl. Statist.*, **37**, 1355-1368.

Sun, D., Speckman, P. and Tsutakawa, R.K. (1999). *Random effects in generalized linear mixed models*. Technical Report, **97**.

Ten Have, T.R., Landis, J.R. and Weaver, S.L. (1995). Association models for periodontal disease progression: a comparison of methods for clustered binary data. *Statist. Medi.,***14**, 413-429.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44,** 1049-1060.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Amer. Statist. Assoc.*, **86**, 79-86.