© 2017 IJSRSET | Volume 3 | Issue 6 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A Comparative Study of Bayesian Methods for the Analysis of Binomial Proportion Data in Agricultural Research

Shagufta Yasmeen, Athar Ali Khan
Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh, India

ABSTRACT

Agricultural research incorporates data which are in the form of discrete counts or proportions based on counts. This kind of data are usually non-normally distributed that can cause issues with parameter estimation and prediction if the usual general linear model framework is applied as they do not satisfy the assumptions of linear models. Here we studied the performance of generalized linear model within Bayesian framework with some asymptotic analytic tools and simulation techniques to the analysis of proportion data with different link functions. The results obtained through the exact as well as asymptotic inference have been compared through open source software such as R and JAGS.

Keywords: Bayesian inference, independence Metropolis, JAGS, Metropolis within Gibbs, R, sampling importance resampling.

I. INTRODUCTION

In many areas of application of agricultural research, one encounters observations made on individual experimental units that take one of two possible forms. For example, a seed may germinate or fail to germinate under certain experimental conditions; an insect in an insecticidal trial may survive or die when exposed to a particular dose of the insecticide. Such data are said to be binary, although an older term is quantal, and the two possible forms for each observation are often described generically by the terms success and failure. In some circumstances, interest centers not just on the response of one particular experimental unit (seed or insect) but on a group of units that have all been treated in a similar manner. Thus a batch of seeds may be exposed to conditions determined by the relative humidity and temperature, for example, and the proportion of seeds germinating in each batch recorded. The resulting data are then referred to as grouped binary data, and represent the number of successes out of the total number of units exposed to a particular set of experimental conditions. Data in the form of proportions are often modeled using the binomial distribution; hence do not satisfy the assumptions of linear models, being both discrete and non-normal. To deal this type of data, two approaches are possible (i) to find some transformation of the data or (ii) to use an alternative form of analysis that takes account of the distributional form of the data. The development of the generalized linear models (Nelder and Wedderburn, 1972) provided a solution to the latter approach. The aim of this paper is to review the binomial proportion data under the framework of Bayesian generalized linear model. A comparison has been made among the different link functions for the binomial regression model.

Bayesian inference is based on Bayes' rule which provides a rational method for updating our beliefs in the light of new information. The Bayes' rule states that posterior distribution is the combination of prior and data information. It does not tell us what our beliefs should be; it tells us how they should change after seeing new information. The prior distribution is important in Bayesian inference since it influences the posterior. When no information is available, we need to specify a prior which will not influence the posterior distribution. Such priors are called weakly-informative

or non-informative or vague priors. This type of priors will be used throughout the paper. The posterior distribution contains all the information needed for Bayesian inference and the objective is to calculate the numeric summaries of it via integration. In cases, where the conjugate family is considered, posterior distribution is available in a closed form and so the required integrals are straightforward to evaluate. However, the posterior is usually of non-standard form and evaluation of integrals is difficult. For evaluating such integrals various methods are available such as Laplace's method (see, for example, Carlin and Louis 2009, Erkanli 1994, Tierney and Kadane 1986, Tierney, Kass, and Kadane 1989) and numerical integration methods (Davis and Rabinowitz 1975, Evans and Swartz 1996). Simulation can also be used as an alternative technique. Simulation based on Markov chain Monte Carlo (MCMC) is used when it is not possible to sample θ directly from posterior $p(\theta|y)$. For a wide class of problems, this is the easiest method to get reliable results (Gelman et al, 2014). Gibbs sampling and Metropolis-Hastings algorithm are the two MCMC techniques which render difficult computational tasks quite feasible. A variant of MCMC techniques are performed such as independence Metropolis and Metropolis within Gibbs sampling. To make computation easier, software such as R and JAGS (Just Another Gibbs Sampler) are used.

II. BINARY AND BINOMIAL REGRESSION

Binary and binomial regression are frequently encountered in modern science, especially in the field of agricultural research for the analysis of proportions that have been calculated from discrete counts, for instance, number of plants out of 20 affected by a disease. Suppose that $y_i \sim \text{binomial}(N_i, p_i)$ with N_i known. Choosing the logit transformation of the probability of success, $g(p_i) = \log(p_i/(1-p_i))$, as the link function leads to the logistic regression model. The likelihood for the data y is

$$p(y|\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{y_i} \left(\frac{1}{1+e^{\eta_i}}\right)^{n_i - y_i}$$

where $\eta_i = X_i\beta$ is the linear predictor for the i^{th} case. Alternatives to logit link are the *probit* $\Phi^{-1}(p)$, and *complementary log-log*, $\log(-\log(1-p))$. The data distribution for the probit and complementary log-log model is

$$p(y|\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} (\Phi(\eta_i))^{y_i} (1-\Phi(\eta_i))^{n_i - y_i}$$

$$p(y|\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} (1-exp[-e^{\eta_i}])^{y_i} (exp[-e^{\eta_i}])^{n_i - y_i}$$

respectively. In practice, the probit and logit models are similar, symmetric about $p = 0.5$, differing mainly in the extremes of the tails, and the relationship $g(p) = -g(1-p)$ holds for both. By contrast, the complementary log-log link allows asymmetry by specifying $\log(-\log(1-p)) = X^T\beta$.

III. BAYESIAN LOGISTIC REGRESSION MODEL WITH DEMETHYLATION EXPERIMENT

To implement the regression model within Bayesian framework, let us consider a data set taken from Welham et al. (2015) also discussed in Amoah et al. (2008). In this experiment, a demethylation agent is applied to plants: the agent has the effect of converting methylated nucleotides to non-methylated form, causing epigenetic changes that lead to abnormal phenotypes such as stunting and deformation. The study aimed to investigate the relationship between dose and the resulting proportion of plants with a normal phenotype. Seed is treated with the demethylation agent at six doses (0, 0.01, 0.1, 0.5, 1.0, 1.5), including a zero control dose. Plants are grown in trays, each tray sown with seeds treated with the same dose of agent and each dose is replicated in four trays: two with 60 plants, and two with 100 plants. The trays are arranged as a completely randomized design. The header part of the data, which contains explanatory variate `Dose`, response variate `Normal` and variate `Total` containing the number of plants for each tray, each identified using dummy index variate `DTray`, is

DTray	Dose	Total	Normal
1	0.00	60	59
2	0.00	60	58
3	0.00	100	99
4	0.00	100	98
5	0.01	60	58
6	0.01	60	59

For the demethylation experiment, a Binomial GLM with logit link for the normal plants in the i^{th} tray (y_i) in terms of the dose applied to that tray (X_i) can be written as

$$y_i \sim binomial(N_i, p_i) \quad i = 1, \ldots, 24$$

$$logit(p_i) = \beta_0 + \beta_1 X_i = \eta_i$$

where p_i is the probability that $X_i(Dose_i)$ gives a normal phenotype, and η_i is its logit transformation. The likelihood for the independent binomial observations is

$$p(y|\beta_0, \beta_1) = \prod_{i=1}^{n} \binom{N_i}{y_i} \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{y_i} \left(\frac{1}{1+e^{\eta_i}}\right)^{N_i - y_i}$$

A weakly informative logistic prior with scale 10^3 is assumed for the regression coefficients

$$\beta_j \sim Logistic(0, 10^3), \qquad j = 0,1$$

Thus, by Bayes' rule the joint posterior density can be obtained as

$$p(\beta_0, \beta_1|y) \propto p(y|\beta_0, \beta_1)\, p(\beta_0, \beta_1)$$

$$p(\beta_0, \beta_1|y) \propto \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}\right)^{y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 X_i}}\right)^{N_i - y_i} \frac{e^{(\beta_0+\beta_1)/10^3}}{(1+e^{(\beta_0+\beta_1)/10^3})^2}$$

The marginal posterior densities for β_0 and β_1 are

$$p(\beta_0|y, X) = \int p\,(\beta_0, \beta_1|y, X) d\beta_1$$

$$p(\beta_1|y, X) = \int p\,(\beta_0, \beta_1|y, X) d\beta_0$$

These integrals, like many encountered in Bayesian inference, are almost always analytically intractable. So inference for a single parameter in a multi parameter setting involves two potentially difficult calculations: first, of the joint posterior, second, of the marginal posterior from the joint. Fortunately, these two tasks are easily accomplished by some asymptotic tool such as Laplace approximation or the simulation techniques like Markov chain Monte Carlo (MCMC). We implement the analytic tool to approximate the posterior density and use the results as starting values in the MCMC algorithm. A variant of MCMC algorithm known as Independence Metropolis is implemented to approximate the joint posterior density, through the package `LaplacesDemon` in `R` and Metropolis within Gibbs sampling through `R2jags`.

A. Analysis with LaplacesDemon

For the computation of marginal posterior densities of each βs the Laplace approximation technique is used to approximate the integral. Parallel simulation tools are also implemented to draw the samples from marginal posterior densities to approximate the results with SIR method and one of the MCMC algorithms. Simulation via MCMC algorithm is done through independent Metropolis. These required techniques are implemented in `LaplacesDemon` package. For this purpose, functions `LaplaceApproximation` and `LaplacesDemon` are used and for doing this first data must be created in `R`.

1) Creation of demethylation data

In the demethylation data set, the vector y contains the individual values for normal phenotype and `Dose` is an explanatory variate, both can be extracted from the data object `demethy` which is read through the function `read.table` in `R` in which `path` is a path where data is stored. The model matrix X has two columns which are denoted by J, the vector of 1's has been inserted for intercept and second column is for `Dose`. The number of plants for each tray is entered with `N`.

```
demethy<- read.table
("path/DEMETHYLATION.DAT",
 header=TRUE)
y<-demethy$Normal
Dose<-demethy$Dose
X<-cbind(1,Dose)
N<-demethy$Total
J <- 2
mon.names <- "LP"
parm.names<-as.parm.names(
list(beta=rep(0,J)))
MyData <- list( J=J, N=N,
mon.names=mon.names, parm.names= parm
.names, X=X, y=y)
```

The two parameters beta[1] and beta[2] are organized in the vector `parm.names` with the function `as.parm.names`. The function `mon.names` is used to monitor the log posterior (LP). `MyData` is a listed object having all the defined objects within it.

2) Model specification

The logarithm of the unnormalized joint posterior distribution is defined in order to make the computation numerically more stable

$$log\ p(\beta_0,\beta_1|y,X) \propto \sum_{i=1}^{24} log\left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{y_i} + log\left(\frac{1}{1+e^{\eta_i}}\right)^{N_i-y_i} + \frac{\beta_0+\beta_1}{10^3} - log(1+e^{(\beta_0+\beta_1)/10^3})$$

$$\eta_i = \beta_0 + \beta_1 X_i$$

The R codes for the above equation are

```
Model <- function(parm, Data){
### Parameters
beta <- parm[1:Data$J]
### Log-Prior
beta.prior <sum(dlogis(beta,0,
1000, log=TRUE))
### Log-Likelihood
mu<- tcrossprod(Data$X,t(beta))
p <- invlogit(mu)
LL<-um(dbinom(Data$y,Data$N,p,
log=TRUE))
### Log-Posterior
LP <- LL + beta.prior
Modelout <- list(LP=LP, Dev=-
2*LL, Monitor=LP,
yhat=rbinom(length(p), Data$N, p), pa
rm=parm)
return(Modelout)}
```

A numerical approximation algorithm iteratively maximizes the logarithm of the unnormalized joint posterior density as specified in this `Model` function. During each iteration in which a numerical approximation algorithm is maximizing the logarithm of the unnormalized joint posterior density, two arguments are passed to `Model`: `parm` and `Data`, where `parm` is short for the set of parameters, and `Data` is a list of data. The `Model` function is evaluated and returned `Modelout` in a form of list which contains log posterior (`LP`) along with the deviance (`Dev`), a vector (`Monitor`) of any variables desired to be monitored in addition to the parameters, y^{rep} (`yhat`) or replicates of y, and the parameter vector `parm`.

3) Asymptotic Approximation

In order to approximate the joint posterior density, Laplace's method is used with the function `LaplaceApproximation`. For the purpose of optimization, several algorithms have been implemented in this function. Among all the algorithms, we have found that the BFGS, LBFGS, NM, and TR perform well in most of the cases. However, for this particular case Trust region (TR) algorithm of Nocedal and Wright (1999) is safe due its efficiency of convergence in the fewest number of iterations. To start the optimization, some initial values for the parameters must be defined and in this way zero is set to the regression coefficients.

```
Initial.Values <- rep(0,J)
Fit<LaplaceApproximation(Model=
Model, parm=Initial.Values,
Data=MyData, Method="TR",
Iterations=500, sir=TRUE)
```

Table 1. Marginal posterior densities summaries of the parameters using the function LaplaceApproximation.

Parameters	Mode	SD	LB	UB
beta[1]	2.79	0.13	2.53	3.05
beta[2]	-7.62	0.35	-8.32	-6.92

Table 2. Summaries of the posterior samples drawn with sampling importance resampling (SIR)

Parameters	Mode	SD	LB	Median	UB
beta[1]	2.82	0.13	2.57	2.82	3.09
beta[2]	-7.69	0.35	-8.35	-7.67	-7.06
Deviance	233.58	2.02	231.62	232.89	239.49
LP	-133.38	1.01	-136.33	-133.03	-132.40

From the summary output of analytic Laplace approximation method via `LaplaceApproximation` function reported in Table 1, it may be noted that the posterior mode of parameter β_0 is -2.79 ± 0.13 with 95% credible interval (2.53,3.05), which is statistically significant, whereas the posterior mode of β_1 is -7.62 ± 0.35 with 95% credible interval (−8.32, -6.92), which is also statistically significant. The simulated results due to sampling importance resampling algorithm using the same function reported in Table 2 says that the posterior mode of β_0 is 2.82 ± 0.13 with 95% credible interval (2.57, 3.09), whereas the posterior mode of β_1 is -7.69 ± 0.35 with 95% credible interval (−8.35, -7.06).

4) Posterior simulation with independence Metropolis

MCMC simulation of logistic regression is done through the function `LaplacesDemon` which implements the independent Metropolis algorithm. Before fitting the model with `LaplacesDemon` it is necessary to specify initial values for each of the parameters. Each initial value is a starting point for the estimation of a parameter. If all initial values are set to

zero for MCMC, the LaplacesDemon function optimizes initial values using a spectral projected gradient algorithm in the LaplaceApproximation function. Hence, it is better to use the last fitted object Fit with the function as.initial.values to get a vector of initial values from LaplaceApproximation for fitting with LaplacesDemon. Thus, to get a vector of initial values and model fitting, the R command is

```
Initial.Values<-as.initial.values(Fit)
FitDemon.TR<- LaplacesDemon(
Model=Model,Data=MyData,
Initial.Values,Covar=Fit1$Covar,Itera
tions=15000,Status=100,
Thinning=10, Algorithm="IM",
Specs=list(mu=Fit1$Summary1[1:
length(Initial.Values),1]))
```

Table 3. Marginal posterior summaries based on the MCMC samples using independence-Metropolis algorithm

Parameters	Mean	SD	MCSE	ESS	LB	Median	UB
beta[1]	2.80	0.08	0.00	1500	2.64	2.80	2.94
beta[2]	-7.63	0.21	0.01	1500	-8.03	-7.63	-7.23
Deviance	232.27	0.71	0.02	1500	231.60	232.06	234.14
LP	-132.72	0.35	0.01	1500	-133.66	-132.62	-132.39

The reduced posterior standard deviations for the parameters based on MCMC (IM) samples as reported in Table 3, depict a better posterior approximation.

B. Analysis with JAGS

In order to interact with JAGS, the interface R2jags is used for the Bayesian analysis of the logistic regression for the same demethylation data, which basically throws R data structures at JAGS and sucks the results back into R, as suitable objects for further processing. This enables us to maintain a completely reproducible record of the initial data-manipulation (in R), estimation (in JAGS) and reporting of results (in R).

1) Creation of data in R

The creation of data is same as defined for LaplacesDemon package. Here, n is the total number of observations of the data demethy. The object j.dat is the data which is provided in the form of a named list, one element per data-structure (usually vector or matrix).

```
y<-demethy$Normal
Dose<-demethy$Dose
X<-cbind(1,Dose)
N<-demethy$Total
n<-nrow(demethy)
j.dat<-list(y=y, X=X, n=n,N=N)
```

2) Model specification

The full Bayesian structure of the binomial GLM with logit link is defined as

$$y_i \sim binomial(N_i, p_i)$$

$$logit(p_i) = \beta_0 + \beta_1 X_i = \eta_i$$

where p_i is the probability that $X_i(Dose_i)$ gives a normal phenotype, and η_i is its logit transformation. The regression coefficients are logistically distributed with zero mean and 0.001 precision.

$$\beta_j \sim Logistic(0, 0.001) \quad j = 0,1$$

The program specifying the above model must be put in a separate file which is then read by JAGS. When working in R this is most conveniently done using the R-function cat() which behaves pretty much like paste() with the exception that the result is not a character object but directly written to a specified file. Here is the BUGS code specifying the above model, using cat to put it in the file "demeth.txt":

```
cat("model{
 for(i in 1:n){
  y[i] ~ dbin(p[i],N[i])
  logit(p[i])<-inprod(X[i,],beta[])
 }
 for(j in 1:2){
  beta[j]~dlogis(0,0.001)}
}",file="demeth.txt")
```

3) Initial values and parameters

To start MCMC simulation, some starting values must be supplied. In order to be able to monitor convergence atleast three chains must be supplied as a starting values for each chain. j.ini is used for this purpose. params is defined for the monitored parameters.

```
j.ini<-function(){list(beta=
 rnorm(2))}
params<-c("beta")
```

4) Calling JAGS from R

With these specifications we can now use JAGS to first compile and initialize the model and then run the model for some 20000 iterations.

```
jagsfit<- jags(jdat,j.ini,
params,model.file="demeth.txt",
n.iter=20000,progress.bar=NULL)
```

The summary of the posterior estimates of the parameters can be obtained with the function `print`.

Table 4. Summary of the marginal posterior distribution of the parameters obtained by Metropolis within Gibbs algorithm implemented in JAGS.

Parameters	Mean	SD	2.5%	50%	97.5%	Rhat	n.eff
beta[1]	2.80	0.13	2.56	2.80	3.07	1.003	1000
beta[2]	-7.65	0.34	-8.33	-7.65	-6.99	1.004	670
Deviance	233.55	1.93	231.62	232.95	238.65	1.001	3000

From this JAGS output, it is noticed that the values of the posterior mean of both beta parameters are very close to the values obtained from sampling importance resampling (SIR).

IV. BAYESIAN BINOMIAL REGRESSION WITH PROBIT LINK

Logit models are the most popular stochastic formulations for binary responses and are cited as logistic regression models. Another popular link is the probit link (Bliss, 1935), which gives results similar to those for the logit link. For the Bayesian implementation of binomial regression with probit link, same demethylation data is discussed here. A binomial GLM with probit link for the normal plants in the $\mathbf{i^{th}}$ tray ($\mathbf{y_i}$) in terms of the dose applied to that tray is defined as

$$y_i \sim binomial(N_i, p_i) \quad i = 1, \ldots, 24$$

$$\Phi^{-1}(p_i) = \beta_0 + \beta_1 X_i = \eta_i$$

where Φ is the cumulative probability function of the standardized normal distribution and Φ^{-1} is its corresponding inverse function. The likelihood for the independent binomial observation is

$$p(y|\beta_0, \beta_1) = \prod_{i=1}^{n} \binom{N_i}{y_i} (\Phi(\eta_i))^{y_i} (1 - \Phi(\eta_i))^{N_i - y_i}$$

A weakly informative logistic prior is assumed for the regression coefficients

$$\beta_j \sim Logistic(0, 10^3), \qquad j = 0,1$$

According to the Baye's rule, the joint posterior density is

$$p(\beta_0, \beta_1|y) \propto \prod_{i=1}^{n} \binom{N_i}{y_i} (\Phi(\eta_i))^{y_i} (1 - \Phi(\eta_i))^{N_i - y_i} \frac{e^{(\beta_0+\beta_1)/10^3}}{(1 + e^{(\beta_0+\beta_1)/10^3})^2}$$

In Bayesian inference the target is to obtain the marginal posterior densities. Under the proportionality, the marginal posterior density for $\boldsymbol{\beta_0}$ can be obtained by integrating out the joint posterior density over $\boldsymbol{\beta_1}$. Similarly, for the marginal posterior density for $\boldsymbol{\beta_1}$ can be obtained by integrating out the joint posterior density over $\boldsymbol{\beta_0}$. Since, these marginal posterior densities for parameters do not have a closed form expression, so that their posterior summaries cannot be evaluated analytically. The most convenient approach seems to be the numerical integration technique and the Markov chain Monte Carlo (MCMC) methods.

A. Analysis with LaplacesDemon

To avoid computational overflows and underflows, the logarithms of the posterior densities are performed.

$$p(\beta_0, \beta_1|y) \propto \sum_{i=1}^{24} log(\Phi(\eta_i))^{y_i} (1 - \Phi(\eta_i))^{n_i - y_i} + \frac{\beta_0 + \beta_1}{10^3} - 2 * log(1 + e^{(\beta_0+\beta_1)/10^3})$$

The R codes for the above model would be same as in logistic regression except for the link function. Here, we use probit link instead of logit link. The code for the probit link is

```
p <- pnorm(mu)
```

After executing all the codes with probit link, the Model function is evaluated and returned Modelout in a form of list containing log posterior (LP) along with the

deviance (Dev), a vector (Monitor) of any variables desired to be monitored in addition to the parameters, $\boldsymbol{y^{rep}}$ (yhat) or replicates of y, and the parameter vector parm.

1) Asymptotic approximation

The function LaplaceApproximation is used to approximate the joint posterior density with the algorithm Trust region. Some initial values for the parameters are assigned to start the optimization. In the LaplacesDemon package the order of the elements of the vector of initial values must match the order of the parameters associated with each element of parm passed to the Model function.

```
Initial.Values <- rep(0 ,J)
Fit <- LaplaceApproximation
(Model=Model, parm=Initial.Values,Dat
a=MyData, Method="TR",
Iterations=1000, sir=TRUE)
```

Table 5. Asymptotic posterior summaries of the parameters with probit link based on Laplace approximation.

Parameters	Mode	SD	LB	UB
beta[1]	1.45	0.06	1.33	1.56
beta[2]	-3.47	0.13	-3.73	-3.22

Table 6. Summaries of the posterior samples drawn with sampling importance resampling (SIR) when sir=TRUE, given the point-estimated posterior modes and the covariance matrix with the bounds that constitute a 95% probability interval.

Parameters	Mode	SD	LB	Median	UB
beta[1]	1.45	0.06	1.32	1.45	1.57
beta[2]	-3.48	0.13	-3.75	-3.48	-3.23
Deviance	324.51	2.16	322.45	323.72	330.51
LP	-171.01	1.08	-174.01	-170.61	-169.98

2) Posterior simulation with independence Metropolis

The function `LaplacesDemon` is used for the MCMC simuation with the independent Metropolis algorithm in order to get the posterior estimates of the parameters.

```
Initial.Values<-as.initial.values(Fit)
FitDemon<- LaplacesDemon(
Model=Model, Data=MyData, Initial.Val
ues, Covar=Fit$Covar, Iterations=5000
, Status=1000, Algorithm="IM",
Specs=list(mu=Fit$Summary1[1:length(I
nitial.Values),1]))
```

Table 7. Marginal posterior summaries based on the MCMC samples using independence-Metropolis algorithm

Parameters	Mean	SD	MCSE	ESS	LB	Median	UB
beta[1]	1.45	0.04	0.00	344.46	1.38	1.45	1.51
beta[2]	-3.48	0.07	0.00	436.53	-3.62	-3.48	-3.34
Deviance	323.11	0.72	0.03	500.00	322.42	322.90	324.82
LP	-170.31	0.36	0.02	500.00	-171.17	-170.20	-169.96

B. Analysis with JAGS

The interface of JAGS with R, i.e.; `R2jags`, is used for the Bayesian analysis of demethylation experiment with probit link. The creation of data is same as for the logistic regression. The full Bayesian model for demethylation data with probit link is defined as

$$y_i \sim binomial(N_i, p_i)$$

$$\Phi^{-1}(p_i) = \beta_0 + \beta_1 X_i = \eta_i$$

The regression coefficients are logistically distributed with mean zero and scale 10^3 (or precision 10^{-3}).

$$\beta_j \sim Logistic(0,10^{-3}) \quad j = 0,1$$

The `BUGS` codes specifying the above model, using cat to put it in the file `demeth.txt` are

```
cat("model{
 for(i in 1:n){
  y[i]~dbin(p[i],N[i])
  p[i]<-phi(inprod(X[i,],beta[]))
for(j in 1:2){
 beta[j]~dlogis(0,0.001)}
}",file="demeth.txt")
```

We use the same initial values and parameters which already defined for logistic regression in JAGS. With these specifications we can now use JAGS to first compile and initialize the model and then run the model for some 2000 iterations.

```
jagsfit<-jags(jdat, j.ini,
params, model.file="demeth.txt", n.it
er=2000, progress.bar=
```

```
NULL)
```

Table 8. Summary of the marginal posterior distribution of the parameters obtained by Metropolis within Gibbs algorithm implemented in JAGS.

Parameters	Mean	SD	2.5%	50%	97.5%	Rhat	n.eff
beta[1]	1.45	0.06	1.34	1.45	1.56	1.001	3000
beta[2]	-3.48	0.13	-3.73	-3.48	-3.23	1.001	3000
Deviance	324.39	2.02	322.45	323.76	329.65	1.001	3000

V. BAYESIAN BINOMIAL REGRESSION WITH COMPLEMENTARY LOG-LOG LINK

A transformation of the binary response probability that is not as widely used as the logistic or probit transformations is the complementary log-log transformation (Fisher, 1922), a less popular link, but it models more efficiently the tails of the distribution, especially when asymmetry between low and high probability values is observed. Collett (2002) explains the complementary log log model in the context of a bioassay which can be derived by supposing that the tolerances of individuals have an extreme value distribution known as the Gumbel distribution. For the Bayesian implementation of Binomial regression with complementary log-log link, demethylation data is revisited. A Binomial GLM with complementary log-log link for the number of normal plants in the $\mathbf{i^{th}}$ tray ($\mathbf{y_i}$) in terms of the dose applied to that tray is defined as

$$y_i \sim binomial(N_i, p_i) \quad i = 1, \dots, 24$$

$$log(-log(1 - p_i)) = \beta_0 + \beta_1 X_i = \mu_i$$

where $\mathbf{p_i}$ is the probability that $\mathbf{X_i}(\mathbf{Dose_i})$ gives a normal phenotype, and $\boldsymbol{\mu_i}$ is its complementary log-log transformation. The likelihood for the independent binomial observations is

$$p(y|\beta_0, \beta_1) = \prod_{i=1}^{n} \binom{N_i}{y_i} (1 - exp[-e^{\mu_i}])^{y_i} (exp[-e^{\mu_i}])^{N_i - y_i}$$

The regression coefficients are assigned to a weakly informative logistic prior with scale 10^3

$$\beta_j \sim Logistic(0, 10^3), \qquad j = 0,1$$

According to the Baye's rule, the joint posterior density is

$$p(\beta_0, \beta_1|y) \propto \prod_{i=1}^{n} \binom{N_i}{y_i} (1 - exp[-e^{\mu_i}])^{y_i} (exp[-e^{\mu_i}])^{N_i - y_i} \frac{e^{(\beta_0+\beta_1)/10^3}}{(1 + e^{(\beta_0+\beta_1)/10^3})^2}$$

The marginal densities for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are not in closed form, so that some asymptotic or simulation techniques are required. In order to this, tools such as Laplace approximation, independence Metropolis, and Metropolis within Gibbs are used.

A. Analysis with LaplacesDemon

For the computation of marginal posterior densities of each βs the Laplace approximation technique is used to approximate the integral through the function `LaplaceApproximation`. Simulation via MCMC algorithm is done through independent Metropolis through the function `LaplacesDemon`. To make the computation comparatively easier, the logarithm of the unnormalized joint posterior density is used

$$p(\beta_0, \beta_1|y) \propto \sum_{i=1}^{24} log(1 - exp[-e^{\mu_i}])^{y_i} (exp[-e^{\mu_i}])^{N_i - y_i} + \frac{\beta_0 + \beta_1}{10^3} - 2 * log(1 + e^{(\beta_0+\beta_1)/10^3})$$

The first step is to create data in R and the creation of demethylation data is same for all link functions. All the R codes for the above model would be same as in logistic regression except for the link function. Here, we use complementary log-log link instead of logit or probit link. The code for the complementary log-log link is

```
Model <- function(parm, Data){
### Parameters
beta <- parm[1:Data$J]
### Log-Prior
beta.prior<-sum(dlogis(beta,
0, 1000, log=TRUE))
### Log-Likelihood
mu<-tcrossprod(Data$X, t(beta))
p <- invcloglog(mu)
LL<-sum(dbinom(Data$y, Data$N,
```

```
p, log=TRUE))
### Log-Posterior
LP <- LL + beta.prior
Modelout <- list(LP=LP, Dev=-
2*LL, Monitor=LP,
yhat=rbinom(length(p), Data$N, p), pa
rm=parm)
return(Modelout)}
```

Here, μ is the inner product of the parameters and regressor and the function `invcloglog` is the inverse of the complementary log-log which returns the probability `p` in the interval [0, 1].

1) Asymptotic approximation

To approximate the joint posterior density, the function `LaplaceApproximation` with the algorithm Trust region is used. To start the optimization, some initial values are given to the parameters through the object `Initial.Values`.

```
Initial.Values <- rep(0,J)
Fit <- LaplaceApproximation(
Model=Model, Initial.Values,
Data=MyData, Method="TR", Iterations=
1000, sir=TRUE)
```

Table 9. Asymptotic posterior summaries of the parameters with complementary log-log link.

Parameters	Mode	SD	LB	UB
beta[1]	1.29	0.05	1.19	1.40
beta[2]	-5.29	0.23	-5.74	-4.83

Table 10. Summaries of the posterior samples drawn with sampling importance resampling (SIR).

Parameters	Mode	SD	LB	Median	UB
beta[1]	1.29	0.05	1.19	1.29	1.40
beta[2]	-5.29	0.23	-5.79	-5.28	-4.83
Deviance	150.34	2.00	148.35	149.74	156.01
LP	-91.76	1.00	-94.59	-91.46	-90.76

2) Posterior simulation with independence Metropolis

MCMC simulation is done with the independence Metropolis algorithm implemented in the function `LaplacesDemon`. To fit the function `LaplacesDemon`, a vector of initial values from `LaplaceApproximation` is used.

```
Initial.Values<-as.initial.values(Fit)
FitDemon<-LaplacesDemon(Model=
Model, Data=MyData, Initial.Values,Co
var=Fit$Covar, Iterations=7000, Statu
s=100, Algorithm=
"IM",Specs=list(mu=Fit$Summary1[1:len
gth(Initial.Values),1]))
```

Table 11. Marginal posterior summaries based on the MCMC samples using independent Metropolis algorithm.

Parameters	Mean	SD	MCSE	ESS	LB	Median	UB
beta[1]	1.30	0.03	0.00	700	1.24	1.30	1.36
beta[2]	-5.29	0.14	0.01	700	-5.56	-5.29	-5.01
Deviance	149.01	0.72	0.03	700	148.33	148.79	151.07
LP	-91.09	0.36	0.01	700	-92.12	-90.98	-90.75

B. Analysis with JAGS

The package `R2jags` is used to conduct the same Bayesian analysis of demethylation data with the complementary log-log link. The simulation technique is performed with Metropolis within Gibbs algorithm which is implemented in JAGS. The data creation, initial values, and monitored parameters are same, only the difference is in the model specification. The full Bayesian Binomial regression model with complementary log-log link is

$$y_i \sim binomial(N_i, p_i) \quad i = 1, \ldots, 24$$

$$log(-log(1 - p_i)) = \beta_0 + \beta_1 * X_i = \mu_i$$

The regression coefficients are assigned to a weakly informative logistic prior with scale 10^3.

$$\beta_j \sim Logistic(0, 10^{-3}) \quad j = 0,1$$

Here is the `BUGS` code specifying the above model, using `cat` to put it in the file `"demeth.txt"`:

```
cat("model{
 for(i in 1:n){
  y[i]~dbin(p[i],N[i])
  cloglog(p[i])<-inprod(X[i,],beta[])
for(j in 1:2){
 beta[j]~dlogis(0,0.001)}
```

```
}",file="demeth.txt")
```

By specifying the initial values and parameters to be monitored, we can use the function `jags` to first compile and initialize the model and then run the model for some 1000 iterations.

```
jagsfit<- jags(jdat, j.ini,
params, model.file="demeth.txt", n.it
er=1000, progress.bar=
NULL)
```

Table 12. Summary of the marginal posterior distribution of the parameters obtained by Metropolis within Gibbs algorithm implemented in JAGS.

Parameters	Mean	SD	2.5%	50%	97.5%	Rhat	n.eff
beta[1]	1.30	0.05	1.19	1.30	1.40	1.000	1500
beta[2]	-5.30	0.22	-5.73	-5.31	-4.86	1.003	1300
Deviance	150.25	1.82	148.36	149.72	155.21	1.006	580

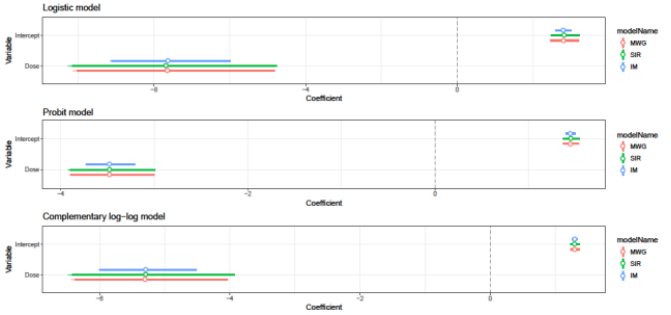


Figure 1. Coefficient estimates for Binomial regression model with all the three link functions. The plot contains the quantiles of the posterior draws. The line corresponds to 95% credible intervals. It is evident that the independent Metropolis (IM) provides shortest interval as compared to sampling importance resampling (SIR) and Metropolis within Gibbs (MWG) sampling. Both Intercept and Dose are significant as they do not contain zero in their credible regions.

VI. MODEL COMPARISON AMONG LINK FUNCTIONS

In Bayesian paradigm, the most common method of assessing the goodness of fit criterion of an estimated statistical model are deviance and deviance information criterion (DIC, Spiegelhalter et al. 2002). Deviance is defined differently in Bayesian inference than frequentist inference. In frequentist inference, deviance is -2 times the log-likelihood ratio of a reduced model compared to a full model, whereas in Bayesian inference, deviance is simply -2 times the log-likelihood. In Bayesian inference, the lowest expected deviance has the highest posterior probability (Gelman et al. 2014).

$$D(y,\theta) = -2log(y|\theta)$$

The DIC is more flexible than deviance since it accounts the model complexity and is the sum of both the mean model-level deviance and the model complexity (pD or pV).

$$DIC = \overline{D} + pV$$

where $pV = var(D)/2$. A smaller DIC and deviance indicates a better fit to the data set. Table 13 provides the deviance and DIC for the three link functions with independent Metropolis implemented in `LaplacesDemon` function and Metropolis within Gibbs sampling implemented in `R2jags`. The deviance as well as the value of DIC is less for complementary log-log model than logit and probit model which shows that the complementary log-log model fits data well. Figure 2 is the graphical representation of deviances for all the three MCMC tools with the three link functions.

Links	IM		MWG	
	Deviance	DIC	Deviance	DIC
Logit	232.27	232.52	233.55	235.40
Probit	323.11	323.37	324.39	326.40
C log-log	149.01	149.27	150.25	151.90

Table 13. Model comparison among logit, probit and complementary log-log model for demethylation data. Both deviance and DIC criterion support Binomial regression model with complementary log-log is a better choice as compared to other two link functions.

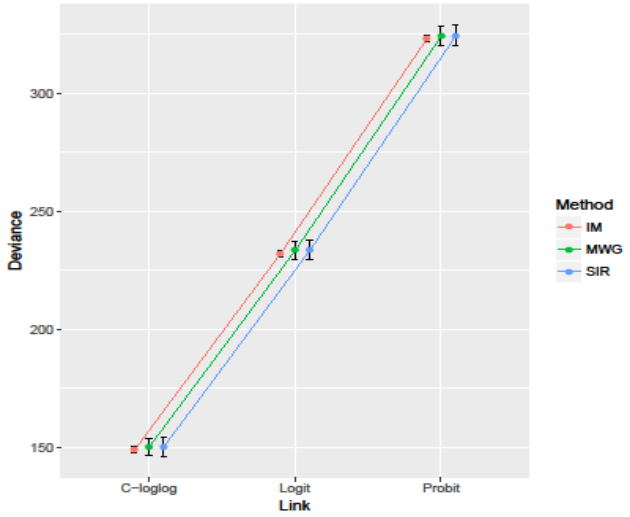
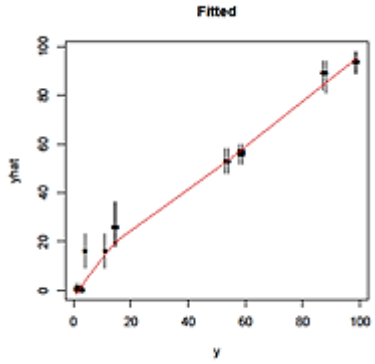
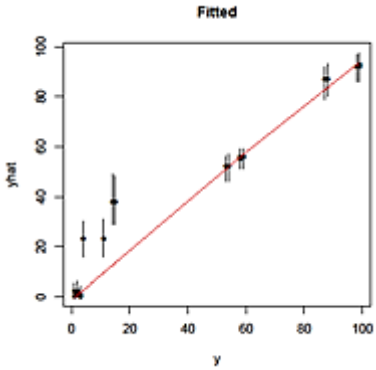


Figure 2. Graphical representation of model comparison among logit, probit and complementary log-log model for demethylation data. The complementary log-log fits data better than other two link functions.

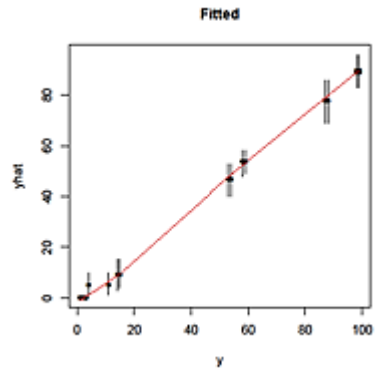
Figure 3 plots the discrepancies between the model and the data which is based on posterior predictive density. A posterior predictive check is a method to assess discrepancies between the model and the data (Gelman, Meng, and Stern 1996). The basic technique for checking the fit of a model to data is to draw simulated values from the joint posterior predictive distribution of replicated data (y^{rep} or $yhat$) and compare these samples to the observed data (y). Any systematic differences between the simulations and the data indicate potential failings of the model (Gelman et al., 2014).



(a) Logit model



(b) Probit model



(c) Complementary log-log model

Figure 3. Posterior predictive fit plots for Bayesian binomial regression with demethylation data under logit, probit, and complementary log-log transformations. The probit model plot shows a worst fit between the dependent variable and its expectation. However, the complementary log-log model fits data well.

VII. CONCLUSION

Analysis of Binomial regression with different link functions under the framework of Bayesian inference have been studied. For the Bayesian implementation, asymptotic technique such as Laplace approximation and simulation with sampling importance resampling, independent Metropolis and Metropolis within Gibbs sampling have been performed. The choice of the link function is quite important for the demethylation experiment. We compared the models using deviance information criterion (DIC) and obtained that complementary log-log model has lowest DIC and it fits data well. After seeing the results from all the mentioned MCMC techniques, it is evident that the independent Metropolis implemented in `LaplacesDemon` function for the complementary log-log model provides the shortest credible regions for the regression coefficients. Moreover, it is also noticed that the independence Metropolis gives shortest standard deviations as compared to sampling importance resampling (SIR) and Metropolis within Gibbs (MWG) sampling. However, the results obtained through SIR and MWG are quite close to each other.

VIII. REFERENCES

[1]. Amoah, S., Wilkinson, M., Dunwell, J. and King, G.J. (2008). Understanding the relationship between DNA methylation and phenotypic plasticity in crop plants. Comparative Biochemistry and Physiology, Part A 150 (Supplement 1), S145.

[2]. Bliss, C. I. (1935). The calculation of the dosage-mortality curve. Annals of Applied Biology, 22, 134-167.

[3]. Carlin, B., P., and Louis, T., A. (2009) Bayesian Methods for Data Analysis (3rd ed.), New York: Chapman and Hall.

[4]. Davis, P. and Rabinowitz, P. (1975). Methods of numerical integration. Academic Press, Waltham, MA.

[5]. Erkanli, A. (1994). Laplace approximations for posterior expectation when the model occurs at

the boundary of the parameter space. Journal of the American Statistical Association, 89, 205-258

[6]. Evans, M., and Swartz, T. (1996). Discussion of methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. Statistical Science, 11, 54-64

[7]. Fisher, R. A. (1922). On the interpretation of chi-square from contingency tables, and the calculation of p. Journal of the Royal Statistical Society B, 85, 87-94.

[8]. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). Bayesian Data Analysis (3rd ed.), Chapman and Hall/CRC, New York.

[9]. Gelman, A., and Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models, New York: Cambridge University Press.

[10]. Gelman, A., Meng, X., Stern, H. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. Statistica Sinica, 6, 773-807.

[11]. Hoff, P. D. (2010). A First Course in Bayesian Statistical Methods. University of Washington, USA 98195–4322

[12]. Korner-Nievergelt, F., Roth, T., Felten, S., Guelat, J., Almasi, B., Korner-Nievergelt, P. (2015). Bayesian Data Analysis in Ecology using R, BUGS and Stan. Elsevier, New York.

[13]. Lunn, D., Jackson, C., Best, N., Thomas, A., Spiegelhalter, D. (2012). The BUGS Book: A Practical Introduction to Bayesian Analysis. Boca Raton, FL: Chapman and Hall/CRC Press.

[14]. McCullagh, P., and Nelder, J. A. (1989). Generalized linear models, second edition. New York: Chapman and Hall.

[15]. Nelder, J. A., and Wedderburn, R.W.M. (1972). Generalized linear models. Journal of the Royal Statistical Society A, 135, 370-384.

[16]. Nocedal, J. and Wright, S. (1999). Numerical Optimization. Springer-Verlag, New York.

[17]. Ntzoufras, I. (2009). Bayesian Modeling Using WinBUGS. John Wiley & Sons, New York.

[18]. Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.

[19]. R Development Core Team,R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austri (2011), ISBN 3-900051-07-0, http://www.R-project.org.

[20]. Robert, C. (2007), The Bayesian Choice, (2nd ed.). Springer-Verlag, New York.

[21]. Statisticat LLC (2013). LaplacesDemon: Complete environment for Bayesian inference. R package version 13.09.01, URL http://www.bayesian-inference.com/software

[22]. Su, Y.-S., and Yajima, M. (2015). Using R to Run 'JAGS', Version 0.5-7

[23]. Tierney, L., and Kadane, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. Journal of the American Statistical Association, 81, 82-86

[24]. Tierney, L., Kass, R., and Kadane, J. (1989), Fully exponential Laplace approximations to expectations and variances of nonpositive functions. Journal of the American Statistical Association, 84, 710-716

[25]. Welham, S. J., Gezan, S.A., Clark, S. J., and Mead, A. (2015). Statistical methods in biology: Design and Analysis of Experiments and Regression. Boca Raton, FL: Chapman and Hall/CRC Press.