mth422_assignment-5

Charitha

2024-04-16

Assignment - 5

question - 1

Given a dataset called **gambia**, in this Y_i is the response variable which is the binary indicator that child i tested positive for malaria (pos) and the remaining seven variables as X_{ij} are covariates.

(a)

To fit the logistic regression model

$$logit[Prob(Y_i = 1)] = \sum_{j=1}^{p} X_{ij}\beta_j$$

with uninformative priors for the β_j

```
## Loading required package: coda
```

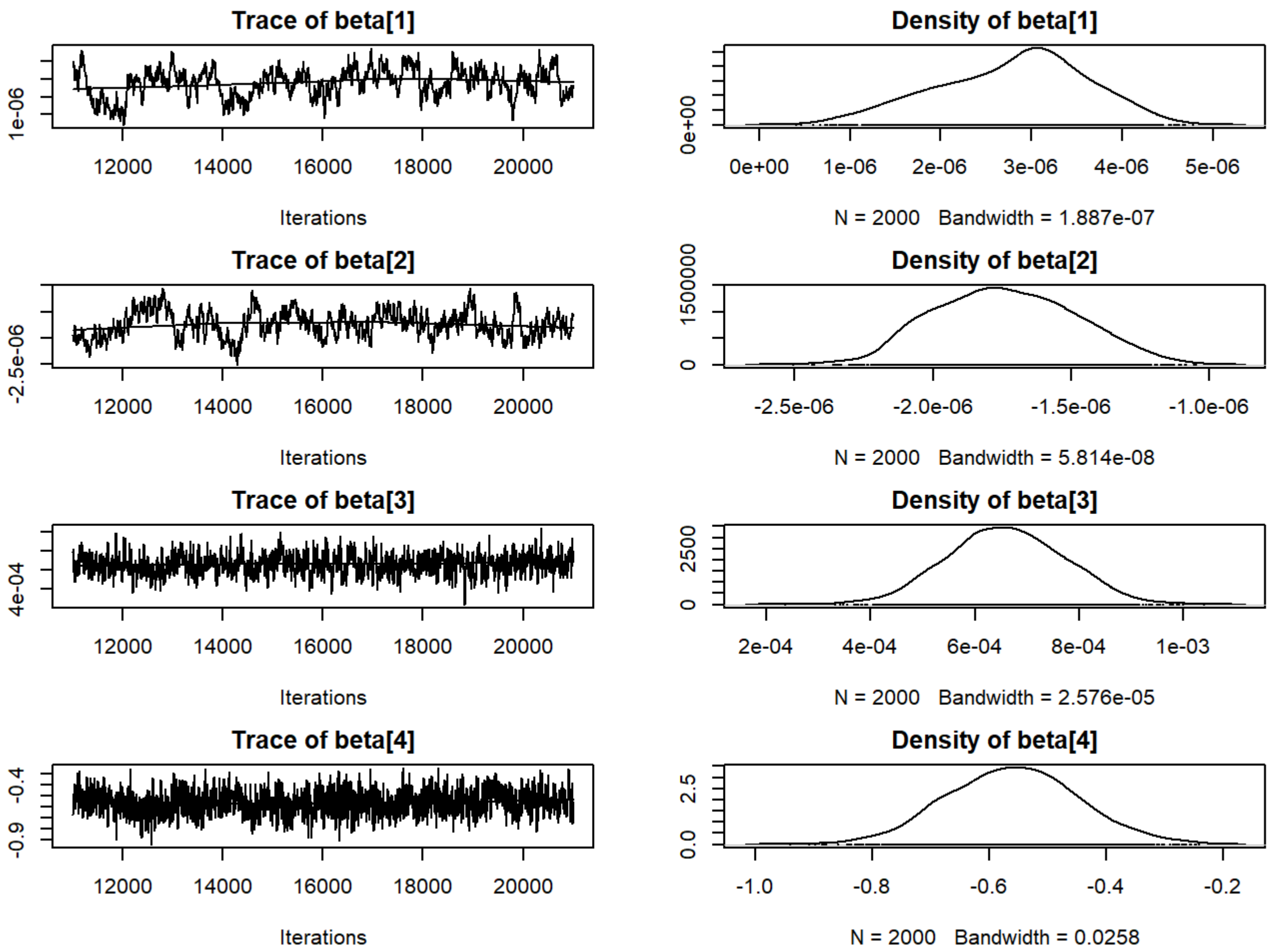
```
## Linked to JAGS 4.3.1
```

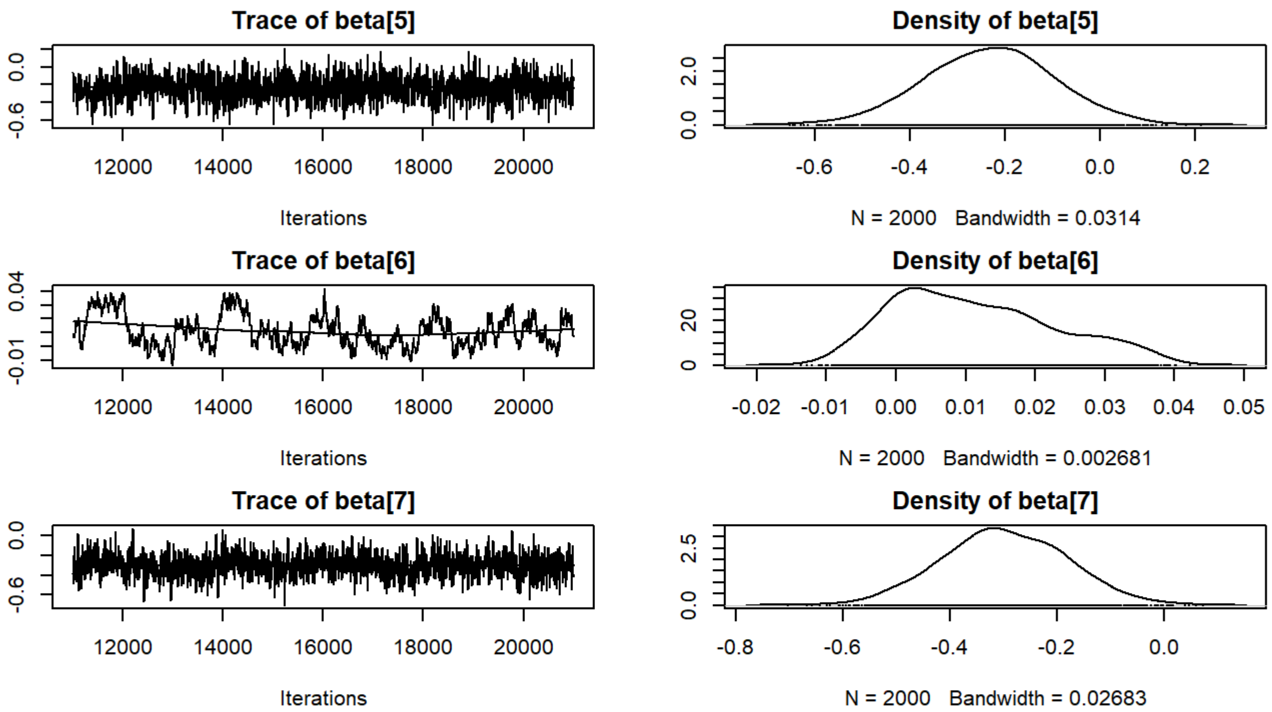
```
## Loaded modules: basemod,bugs
```

```
## --------------------------------------------------------------
##  Analysis of Geostatistical Data
##  For an Introduction to geoR go to http://www.leg.ufpr.br/geoR
##  geoR version 1.9-4 (built on 2024-02-14) is now loaded
## --------------------------------------------------------------
```

```
## Summary of beta's
```

```
##
## Iterations = 11005:21000
## Thinning interval = 5
## Number of chains = 1
## Sample size per chain = 2000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##               Mean        SD  Naive SE Time-series SE
## beta[1]  2.787e-06 8.139e-07 1.820e-08      1.138e-07
## beta[2] -1.738e-06 2.508e-07 5.608e-09      3.881e-08
## beta[3]  6.590e-04 1.112e-04 2.485e-06      5.017e-06
## beta[4] -5.604e-01 1.113e-01 2.489e-03      7.281e-03
## beta[5] -2.324e-01 1.355e-01 3.029e-03      3.773e-03
## beta[6]  1.128e-02 1.156e-02 2.586e-04      2.255e-03
## beta[7] -2.991e-01 1.157e-01 2.588e-03      3.694e-03
##
## 2. Quantiles for each variable:
##
##               2.5%        25%        50%        75%      97.5%
## beta[1]  1.076e-06  2.224e-06  2.888e-06  3.351e-06  4.194e-06
## beta[2] -2.174e-06 -1.925e-06 -1.747e-06 -1.560e-06 -1.246e-06
## beta[3]  4.434e-04  5.853e-04  6.576e-04  7.350e-04  8.740e-04
## beta[4] -7.769e-01 -6.386e-01 -5.597e-01 -4.853e-01 -3.403e-01
## beta[5] -4.982e-01 -3.249e-01 -2.298e-01 -1.422e-01  3.405e-02
## beta[6] -7.016e-03  2.034e-03  9.652e-03  1.894e-02  3.551e-02
## beta[7] -5.226e-01 -3.761e-01 -3.023e-01 -2.181e-01 -6.944e-02
```


Trace of beta[1]
1e-06
12000
14000
16000
18000
20000
Iterations
Density of beta[1]
0e+00
0e+00
1e-06
2e-06
3e-06
4e-06
5e-06
N = 2000 Bandwidth = 1.887e-07
Trace of beta[2]
-2.5e-06
12000
14000
16000
18000
20000
Iterations
Density of beta[2]
0
1500000
-2.5e-06
-2.0e-06
-1.5e-06
-1.0e-06
N = 2000 Bandwidth = 5.814e-08
Trace of beta[3]
4e-04
12000
14000
16000
18000
20000
Iterations
Density of beta[3]
0
2500
2e-04
4e-04
6e-04
8e-04
1e-03
N = 2000 Bandwidth = 2.576e-05
Trace of beta[4]
-0.9
-0.4
12000
14000
16000
18000
20000
Iterations
Density of beta[4]
0.0
2.5
-1.0
-0.8
-0.6
-0.4
-0.2
N = 2000 Bandwidth = 0.0258


Trace of beta[5]
Iterations
Density of beta[5]
N = 2000 Bandwidth = 0.0314
Trace of beta[6]
Iterations
Density of beta[6]
N = 2000 Bandwidth = 0.002681
Trace of beta[7]
Iterations
Density of beta[7]
N = 2000 Bandwidth = 0.02683

(b)

Now we will be using random effect term that are labels of the e location for observation i. To fit the random effects logistic regression model

~~~~
\begin{align*}
  logit[Prob(Y_i = 1)] & = \sum_{j=1}^p X_{ij} \beta_j +
\alpha_{s_i} \\
  \alpha_l & \sim Normal(0,\tau^2)
end{align*}
with  uninformative priors $\beta_j$ and $\tau$.


```
## Loading required package: viridisLite
```

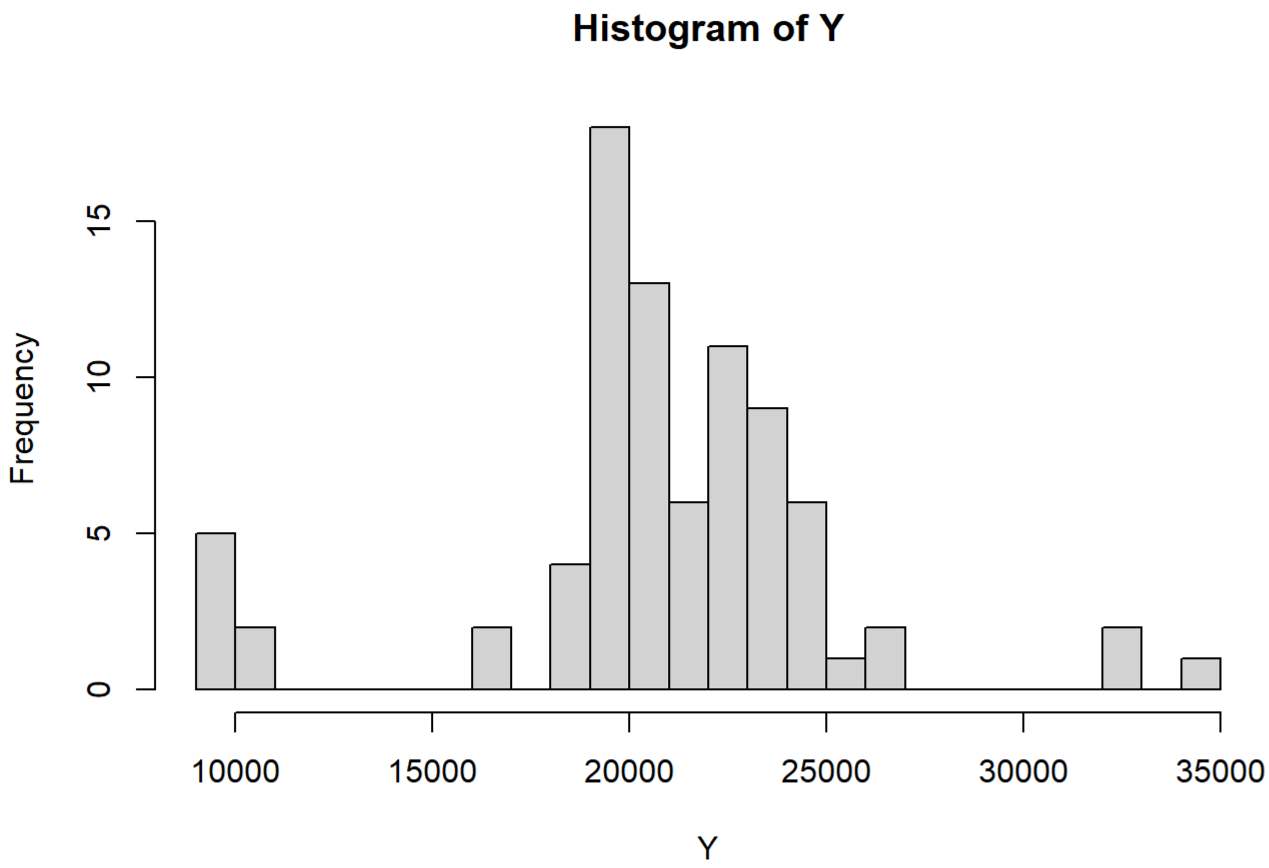

<img
src="MTH422_assignment-5_files/figure-html/unnamed-chunk-2-1.png"
width="672" />

- The random effects logistic regression model is useful when there is
clustering or grouping within the data, as it allows for variability
between these groups. In this case, the children in the dataset are
located in $65$ unique locations, each with its own characteristics that
may influence the outcome variable $(Y_i = 1 \text{ or } Y_i = 0)$. By
incorporating random effects ($\alpha_{s_i}$), we account for the
potential differences between these locations that may affect the
probability of the outcome.

- Adding random effects to the model can lead to differences in the
posteriors of the regression coefficients compared to a standard
logistic regression model. This is because the random effects capture
the variation between locations, which can affect the estimates of the
fixed effects ($\beta_j$). In particular, the coefficients may shrink
towards zero or show different patterns of association with the outcome
variable when accounting for location-specific effects.

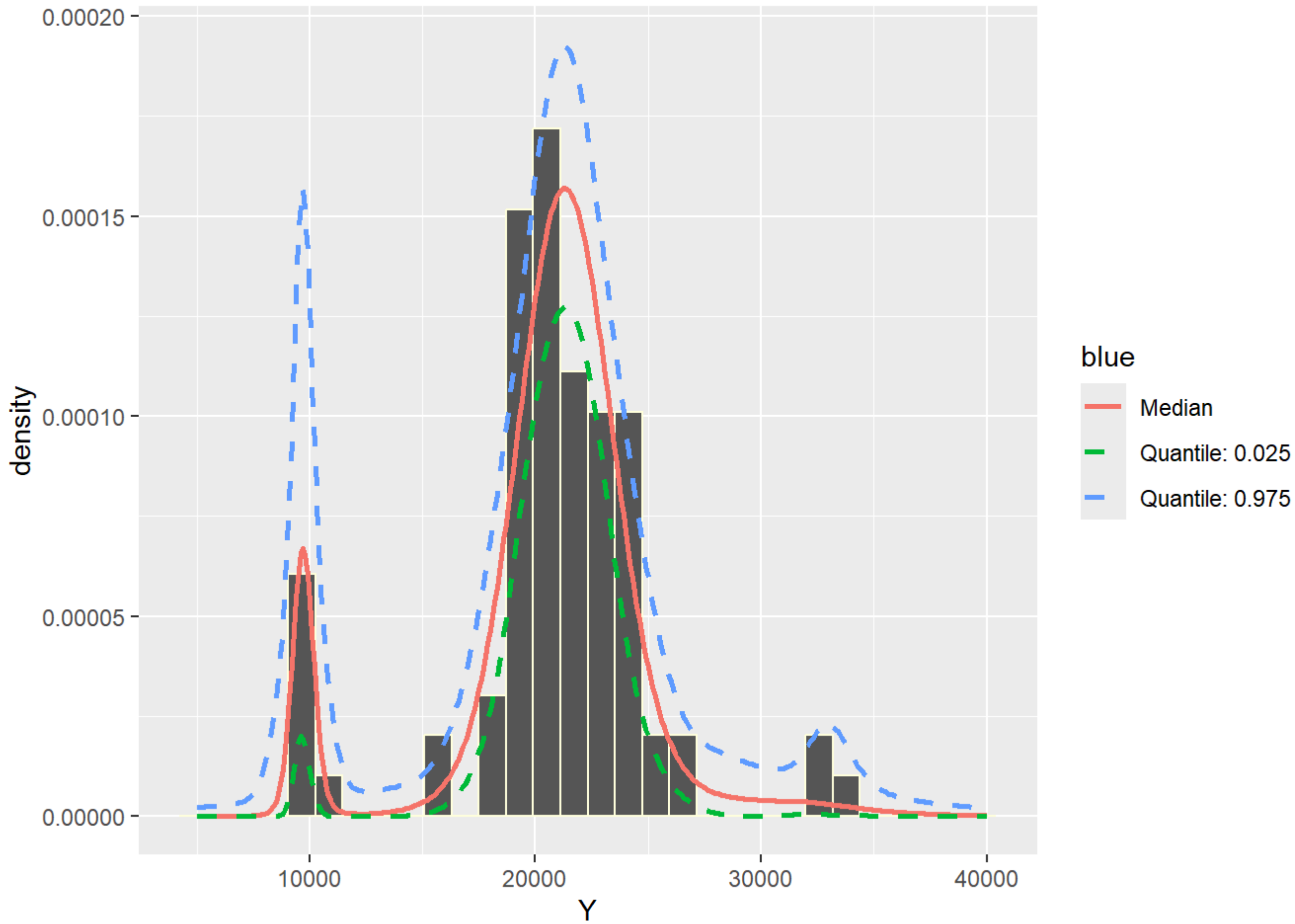

### question - 2
Given the **galaxies** dataset, we have to model the observations using
mixture of $K = 3$ normal distributions.

\begin{align*}
  Y & = \theta_1 Normal(\mu_1,\sigma_1^2) + \theta_2
Normal(\mu_2,\sigma_2^2) + \theta_3 Normal(\mu_3,\sigma_3^2)
\end{align*}
~~~~


Histogram of Y
Frequency
0
5
10
15
10000
15000
20000
25000
30000
35000
Y

```
##
## Iterations = 20001:30000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##               Mean        SD  Naive SE Time-series SE
## mu[1]    9.616e+03 1.335e+03 1.335e+01      1.031e+02
## mu[2]    2.130e+04 2.849e+02 2.849e+00      4.520e+00
## mu[3]    2.563e+04 5.827e+03 5.827e+01      2.274e+02
## tau[1]   8.942e-03 7.408e-01 7.408e-03      7.408e-03
## tau[2]   2.384e-07 5.642e-08 5.642e-10      1.457e-09
## tau[3]   1.714e-07 4.324e-07 4.324e-09      1.554e-08
## theta[1] 9.094e-02 3.257e-02 3.257e-04      9.207e-04
## theta[2] 8.027e-01 7.040e-02 7.040e-04      2.306e-03
## theta[3] 1.064e-01 6.761e-02 6.761e-04      2.659e-03
##
## 2. Quantiles for each variable:
##
##               2.5%       25%       50%       75%     97.5%
## mu[1]    9.219e+03 9.567e+03 9.695e+03 9.821e+03 1.014e+04
## mu[2]    2.074e+04 2.110e+04 2.130e+04 2.149e+04 2.185e+04
## mu[3]    1.223e+04 2.289e+04 2.560e+04 2.977e+04 3.377e+04
## tau[1]   8.120e-07 2.793e-06 4.373e-06 6.540e-06 1.434e-05
## tau[2]   1.536e-07 1.996e-07 2.298e-07 2.681e-07 3.737e-07
## tau[3]   1.969e-09 1.556e-08 3.131e-08 6.834e-08 1.531e-06
## theta[1] 3.555e-02 6.811e-02 8.860e-02 1.105e-01 1.631e-01
## theta[2] 6.344e-01 7.638e-01 8.139e-01 8.538e-01 9.069e-01
## theta[3] 2.077e-02 5.549e-02 9.079e-02 1.421e-01 2.755e-01
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



By observing the graph, we can say that the mixture of $K = 3$ model fit the data well.

question - 3

Given the data of **Mr.October**, $Y_1 = 563, N_1 = 2820, Y_2 = 10, N_2 = 27$

$$M1 : Y_1|\lambda_1 \sim poisson(N_1\lambda_1) \text{ and } Y_2|\lambda_2 \sim poisson(N_2\lambda_2)$$
$$M2 : Y_1|\lambda_0 \sim poisson(N_1\lambda_0) \text{ and } Y_2|\lambda_0 \sim poisson(N_2\lambda_0)$$

To find the bayes factors, DIC and WAIC with priors assumption $\lambda_j \sim Uniform(0, c)$ for c = 1 and 10

```
## Bayes Factor for c = 1 is  0.7155077
```

```
## Bayes Factor for c = 10 is 71.55077
```

```
## DIC values when c = 1
```

```
## [[1]]
## [1] 28701.47
##
## [[2]]
## [1] 32865.9
##
## [[3]]
## [1] "Model 1 is preferred"
```

```
## DIC values when c = 10
```

```
## [[1]]
## [1] 28751.59
##
## [[2]]
## [1] 32925.81
##
## [[3]]
## [1] "Model 1 is preferred"
```

```
## WAIC values when c = 1
```

```
## [[1]]
## [1] 16.14651
##
## [[2]]
## [1] 17.3669
##
## [[3]]
## [1] "Model 1 is preferred"
```

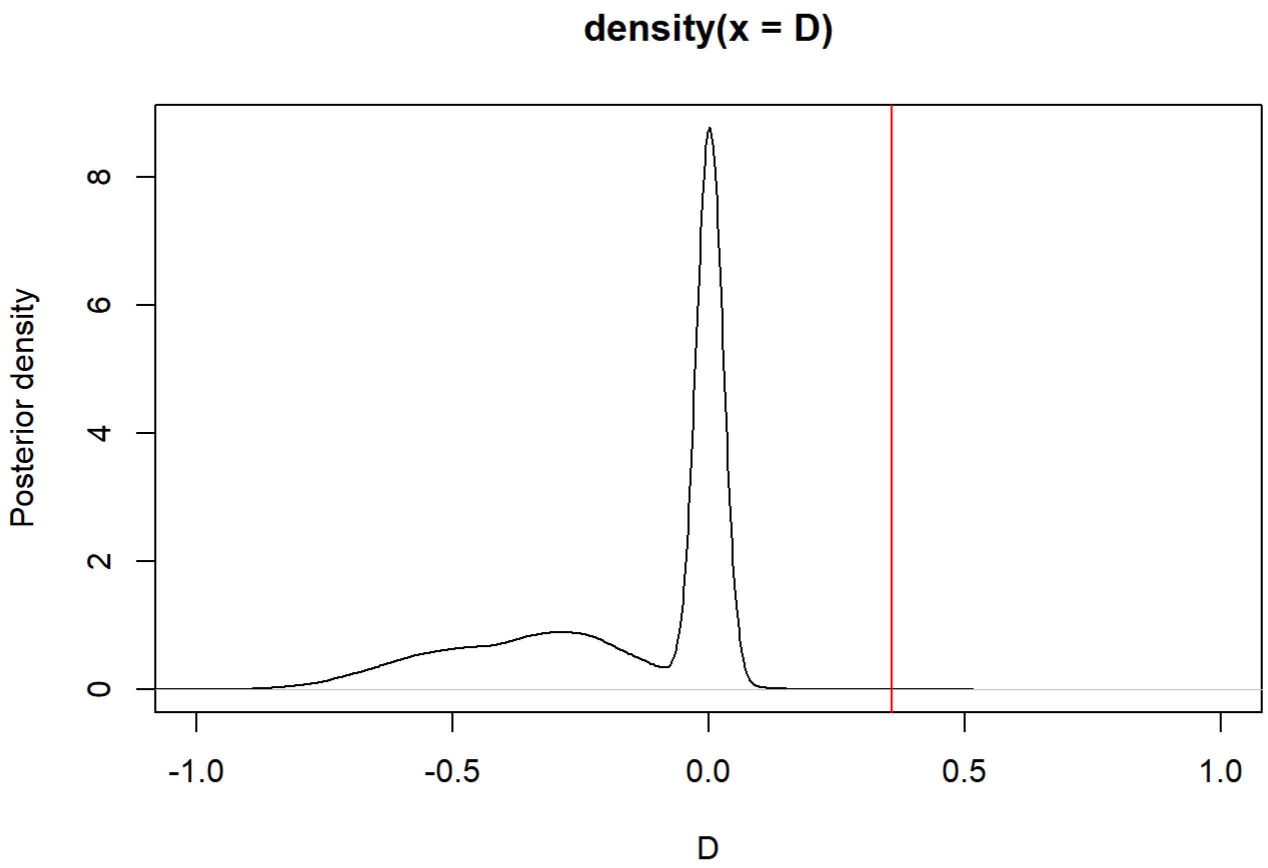
```
## WAIC values when c = 10
```

```
## [[1]]
## [1] 16.49748
##
## [[2]]
## [1] 17.49228
##
## [[3]]
## [1] "Model 1 is preferred"
```

question - 4

To fit logistic regression model to the gambia data and use posterior predictive checks to verify the model fits well

```
##
## Iterations = 11005:31000
## Thinning interval = 5
## Number of chains = 1
## Sample size per chain = 4000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##               Mean        SD  Naive SE Time-series SE
## beta[1]  3.124e-06 8.514e-07 1.346e-08      8.728e-08
## beta[2] -1.724e-06 2.428e-07 3.840e-09      2.096e-08
## beta[3]  6.633e-04 1.147e-04 1.813e-06      3.752e-06
## beta[4] -5.588e-01 1.172e-01 1.854e-03      3.540e-03
## beta[5] -2.264e-01 1.377e-01 2.178e-03      3.926e-03
## beta[6]  7.205e-03 1.176e-02 1.859e-04      1.619e-03
## beta[7] -3.097e-01 1.141e-01 1.804e-03      3.184e-03
##
## 2. Quantiles for each variable:
##
##               2.5%        25%        50%        75%      97.5%
## beta[1]  1.479e-06  2.542e-06  3.116e-06  3.683e-06  4.827e-06
## beta[2] -2.197e-06 -1.890e-06 -1.723e-06 -1.561e-06 -1.238e-06
## beta[3]  4.356e-04  5.876e-04  6.627e-04  7.404e-04  8.889e-04
## beta[4] -7.935e-01 -6.361e-01 -5.562e-01 -4.806e-01 -3.310e-01
## beta[5] -4.932e-01 -3.194e-01 -2.274e-01 -1.370e-01  4.985e-02
## beta[6] -1.627e-02 -8.343e-04  7.738e-03  1.548e-02  2.888e-02
## beta[7] -5.317e-01 -3.856e-01 -3.086e-01 -2.337e-01 -9.265e-02
```



question - 5

Given **WWWusage** dataset, we need to fit the auto regressive model

$$Y_t | Y_{t-1}, \ldots, Y_1 \sim Normal(\beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_L Y_{t-L}, \sigma^2)$$
$$L = \{1, 2, 3, 4\}$$

To select the best time lag L, I have used **WAIC**

```
##   L        WAIC
## 1 1    612.5617
## 2 2    512.2738
## 3 3  21420.2562
## 4 4 261039.6606
```

For time lag $L = 2$, the WAIC value is very less as compared to other time lag model.

So, the best fit model of time lag $L = 2$ is preferred.