Variable selection for the Gambia data

The gambia data in the geoR package includes data for 1332 children in the Gambia. The binary response Y_i is the indictor that child i tested positive for malaria. We use five covariates in X_{ij}.

1. Age: age of the child, in days
2. Netuse: indicator variable denoting whether (1) or not (0) the child regularly sleeps under a bed-net
3. Treated: indicator variable denoting whether (1) or not (0) the bed-net is treated (coded 0 if netuse=0)
4. Green: satellite-derived measure of the greenness of vegetation in the immediate vicinity of the village (arbitrary units)
5. PCH: indicator variable denoting the presence (1) or absence (0) of a health center in the village

We use the logit regression model

$$\text{logit}[\text{Prob}(Y_i = 1)) = \alpha + \sum_{j=1}^{p} X_{ij}\beta_j.$$

The spike-and-slab prior for $\beta_j = \gamma_j \delta_j$ where $\gamma_j \sim \text{Bernoulli}(0.5)$ and $\delta_j \sim \text{Normal}(0, \tau^2)$.

Load the data and necessary libraries

```
library(geoR)
```

```
## ----------------------------------------------------------
##  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
## ----------------------------------------------------------
```

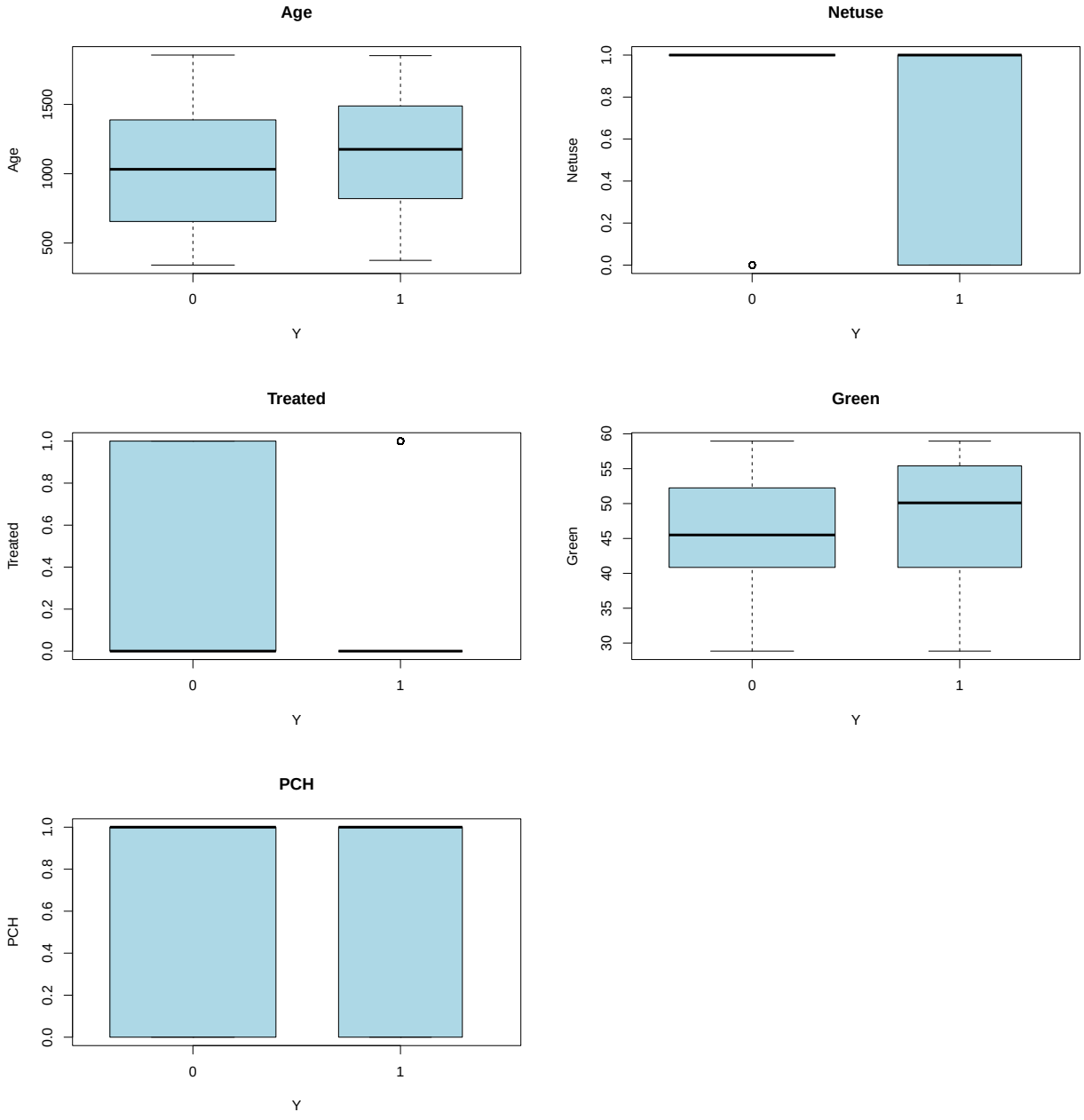
```
data(gambia)
Y <- gambia[,3]
X <- gambia[,4:8]
Y[1:5]
```

```
## [1] 1 0 0 1 0
```

```
X[1:5,]
```

```
##       age netuse treated green phc
## 1850 1783      0       0 40.85   1
## 1851  404      1       0 40.85   1
## 1852  452      1       0 40.85   1
## 1853  566      1       0 40.85   1
## 1854  598      1       0 40.85   1
```

```r
# Define the names of the variables
variable_names <- c("Age", "Netuse", "Treated", "Green", "PCH")
for (i in seq_along(variable_names)) {
  boxplot(X[, i] ~ Y, main = variable_names[i],
          xlab = "Y", ylab = variable_names[i], col = "lightblue",
          varwidth = TRUE)
}
```




```r
# Standardize X
X <- scale(X)
X[1:5,]
```

```
##            age     netuse    treated      green       phc
## 1850  1.650148 -1.5683351 -0.6167739 -0.8462609 0.6802564
## 1851 -1.588791  0.6373055 -0.6167739 -0.8462609 0.6802564
## 1852 -1.476050  0.6373055 -0.6167739 -0.8462609 0.6802564
## 1853 -1.208292  0.6373055 -0.6167739 -0.8462609 0.6802564
## 1854 -1.133132  0.6373055 -0.6167739 -0.8462609 0.6802564
```

```
n <- length(Y)
p <- ncol(X)
```

Put the models in JAGS

```
library(rjags)
```

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

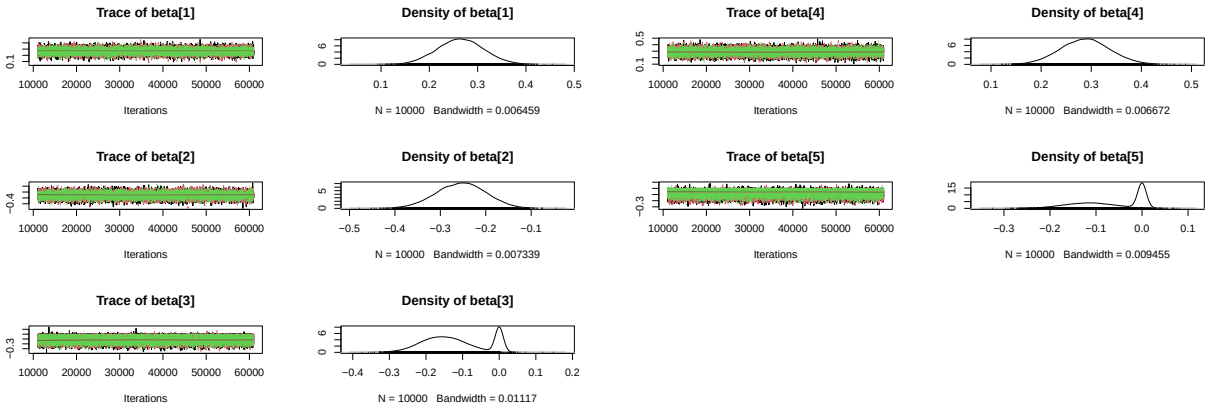
```
## Linked to JAGS 4.3.1
```

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

```
m <- textConnection("model{
 for(i in 1:n){
   Y[i]          ~ dbern(pi[i])
   logit(pi[i]) <- alpha          + X[i,1]*beta[1] + X[i,2]*beta[2] +
                   X[i,3]*beta[3] + X[i,4]*beta[4] + X[i,5]*beta[5]
 }
 for(j in 1:5){
    beta[j] <- gamma[j]*delta[j]
    gamma[j] ~ dbern(0.5)
    delta[j] ~ dnorm(0,tau)
 }
 alpha ~ dnorm(0,0.01)
 tau   ~ dgamma(0.1,0.1)
}")
```

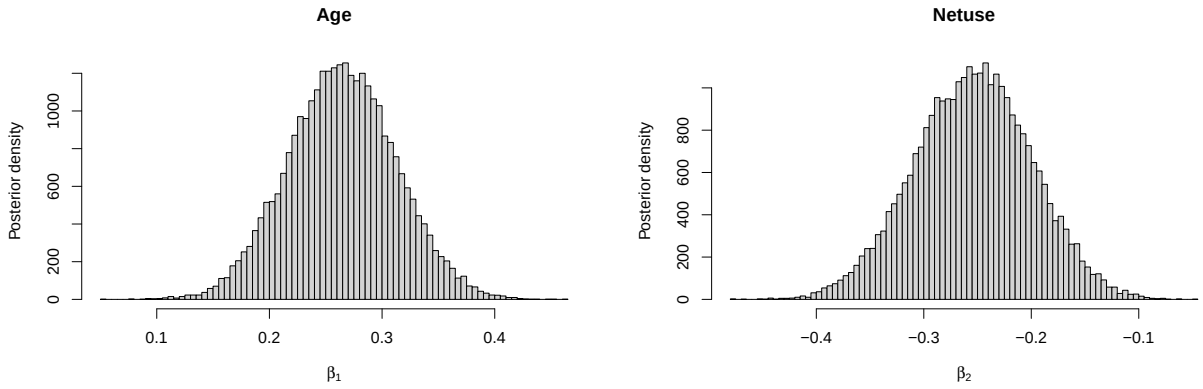
Fit the model

```
data   <- list(Y=Y,X=X,n=n)
burn   <- 10000
iters  <- 50000
chains <- 3
model  <- jags.model(m,data = data, n.chains=chains,quiet=TRUE)
update(model, burn, progress.bar="none")
samps  <- coda.samples(model, variable.names=c("beta"),
                       thin=5, n.iter=iters, progress.bar="none")
plot(samps)
```

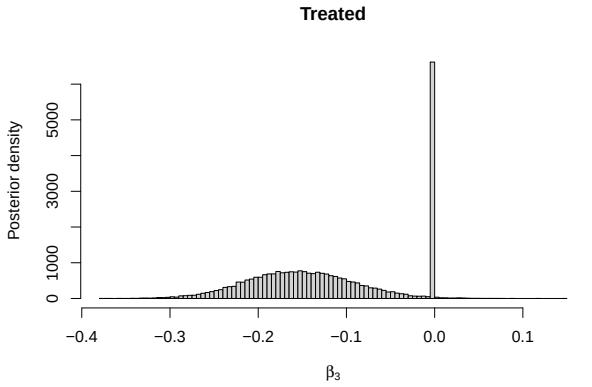


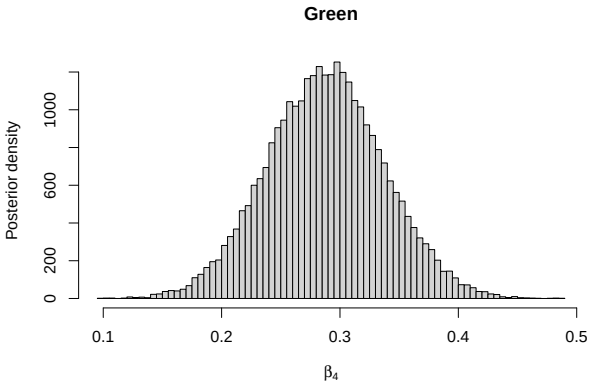
Marginal distributions of the β_j

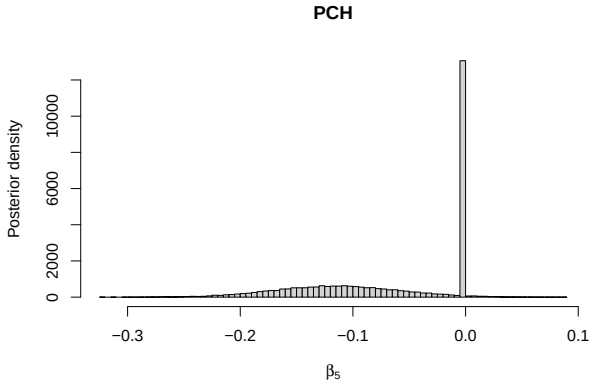
```r
# Define the variable names
names <- c("Age", "Netuse", "Treated", "Green", "PCH")

# Combine samples from all chains into a single matrix
beta_combined <- do.call(rbind, samps)

# Ensure beta_combined has correct dimensions
if (ncol(beta_combined) != length(names)) {
  stop("Number of columns in beta_combined does not match the length of names")
}

# Assign column names to beta_combined
colnames(beta_combined) <- names

# Plot histograms for each variable
for (i in 1:5) {
  hist(beta_combined[, i], xlab = bquote(beta[.(i)]), ylab = "Posterior density",
       breaks = 100, main = names[i])
}
```



	Inc_Prob	50%	5%	95%
Age	1.00	0.26	0.19	0.34
Netuse	1.00	-0.25	-0.35	-0.17
Treated	0.78	-0.13	-0.24	0.00
Green	1.00	0.29	0.21	0.37
PCH	0.57	-0.05	-0.19	0.00








```r
# Load required library
library(kableExtra)

# Calculate inclusion probabilities
Inc_Prob <- colMeans(beta_combined != 0)

# Calculate quantiles
Q <- t(apply(beta_combined, 2, quantile, probs = c(0.5, 0.05, 0.95)))

# Combine the results into a single matrix
out <- cbind(Inc_Prob, Q)

# Create a formatted table
kbl(round(out, 2)) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

```r
# Construct model strings based on beta values
models <- rep("Intercept", nrow(beta_combined))
for (j in 1:5) {
  models <- paste0(models, ifelse(beta_combined[, j] == 0, "", "+"), ifelse(beta_combined[, j] == 0, ""
}

# Print first 5 models and corresponding beta values
print(models[1:5])
```

```
## [1] "Intercept+Age+Netuse+Treated+Green+PCH"
## [2] "Intercept+Age+Netuse+Treated+Green+PCH"
## [3] "Intercept+Age+Netuse+Treated+Green+PCH"
## [4] "Intercept+Age+Netuse+Treated+Green+PCH"
## [5] "Intercept+Age+Netuse+Treated+Green"
```

```r
print(beta_combined[1:5, ])
```

```
##            Age     Netuse     Treated     Green         PCH
## [1,] 0.2580549 -0.2327265 -0.06832056 0.3270162 -0.16317827
## [2,] 0.2540875 -0.1766476 -0.26656317 0.3176130 -0.06695072
## [3,] 0.3610022 -0.3351425 -0.10067628 0.3356066 -0.12249523
## [4,] 0.2885323 -0.2377438 -0.14408006 0.2557963 -0.12620481
## [5,] 0.2873900 -0.1706000 -0.22408183 0.3176870  0.00000000
```

```r
# Calculate and sort model probabilities
model_probs <- table(models) / length(models)
model_probs <- sort(model_probs, decreasing = TRUE)
round(model_probs, 3)
```

```
## models
##     Intercept+Age+Netuse+Treated+Green Intercept+Age+Netuse+Treated+Green+PCH
##                                  0.416                                  0.364
##         Intercept+Age+Netuse+Green+PCH             Intercept+Age+Netuse+Green
##                                  0.203                                  0.016
```

Summary: Age, bed-net use and greenness are included with posterior probability one and are thus clearly important predictors of malaria. Treatment of the bed net and proximity to a health center are included with posterior probability more than a half and so there is moderate evidence that these variables are important predictors of malaria prevalence. The posterior distribution of these parameters has an unusual shape: they are a combination of a Gaussian curve for samples that include the variable and a spike at zero for samples that exclude the variable.

Summary: Three models dominate the posterior probability:

Intercept + Age + Netuse + Green + Treated

Intercept + Age + Netuse + Green + Treated + PCH

Intercept + Age + Netuse + Green + PCH

Therefore it is clear that age, bed net use and greeness should be included in the model, but uncertainty about whether one or both of the remaining two variables should be included.