Effect of bottom coverage to larval presence

Week6-ex1, problem statement

In this exercise, we continue the analysis of the white fish larval areas (week 2, exercise 3 and week4, exercise 2). This time we extend the model to include regression along two continues covariates in addition to the vegetation cover status. The additional covariates that we are interested in are the distance to sandy shore and the length of ice cover during winter. Many fishermen have observed that white fish are caught more easily from sandy shores than elsewhere during their spawning season. Moreover, white fish spawn their eggs during fall but the larvae hatch only in the spring. Hence, it has been suggested that longer ice cover period works as a shelter for the eggs. Hence, let's take a look whether there is statistical signal to these covariates.

Let's load the data and construct covariate matrix X (a matrix where the i'th row contains the covariates for the i'th sampling site), vector of area indexes a (areas in the code) and vector of white fish presence-absence observations y.

```r
# Read the data
data = read.csv("white_fishes_data.csv")
head(data)
```

```
##     AREANAME DIS_SAND ICELAST09 WHIBIN BOTTOMCOV BOTTOM
## 1 Bjuroklubb       17        16      1         0      3
## 2 Bjuroklubb       17        16      1         0      2
## 3 Bjuroklubb       17        16      1         0      3
## 4 Bjuroklubb       15        16      1         0      4
## 5 Bjuroklubb       15        16      1         0      4
## 6 Bjuroklubb       14        16      1         0      3
```

```r
# Let's then take the covariates to matrix X and standardize them
X = data[,c("DIS_SAND","ICELAST09","BOTTOMCOV")]

# And for last let's take the presence-absence observations of white fish larvae into Y
y = data$WHIBIN
```

Unlike in our previous analyses of this data we treat each sampling site as one observation and consider the triplets $\{y_i, a_i, X_i\}$ (X_i is the i'th row of X) exchangeable.

We will first build the following model to analyze the data.

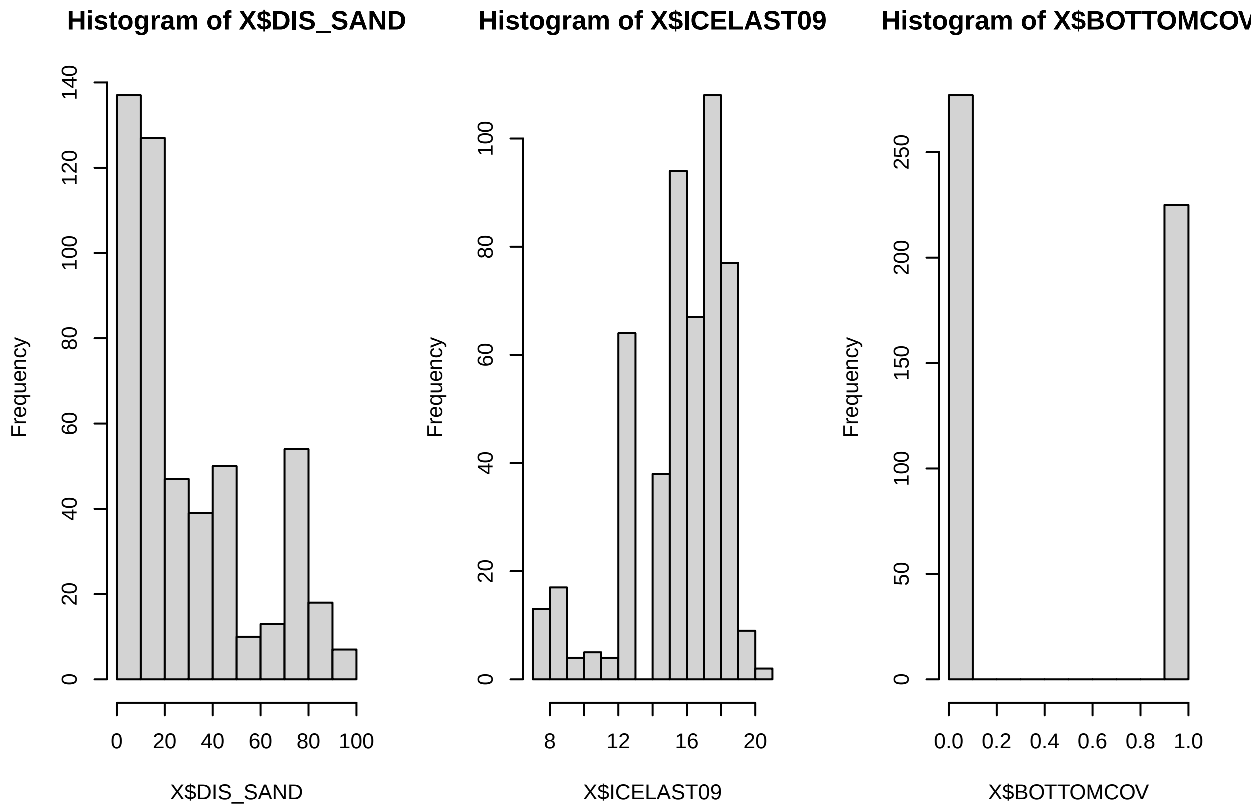
$$
\begin{aligned}
y_i &\sim \text{Bernoulli}(\theta_i) \\
\text{logit}(\theta_i) &= \alpha + X\beta \\
\alpha, \beta_1, \beta_2, \beta_3 &\sim N(0, 10)
\end{aligned}
$$

Hence, we assume that the prior expectation of the probability to observe white fish larvae ($E[y_i] = \theta_i$) follows logit linear model where α is the intercept and β is a 3×1 vector of (fixed) effects of covariates. Note that the matrix notation $X\beta$ is the same as writing

$$X\beta = \beta_1 \times \text{DISSAND} + \beta_2 \times \text{ICELAST09} + \beta_3 \times \text{BOTTOMCOV}$$

Note also that the DIS_SAND and ICELAST09 are continuous covariates whereas BOTTOMCOV is a categorical covariate getting value 1 if the bottom is covered by vegetation and 0 if the bottom is not covered by vegetation.

```
par(mfrow=c(1,3))
hist(X$DIS_SAND)
hist(X$ICELAST09)
hist(X$BOTTOMCOV)
```



Hence, the parameter β_3 corresponds to the effect of vegetation to the observation probability of white fish larvae.

Before starting the analysis we standardize the continues covariates but not the categorical BOTTOMCOV covariate. If we standardized the categorical variable the interpretation of β_3 parameter would change.

```
mx = colMeans(X[,1:2])
stdx = apply(X[,1:2],2,sd)
X[,1:2] = (X[,1:2]-t(replicate(dim(X)[1],mx)))/t(replicate(dim(X)[1],stdx))

mx
```

```
##  DIS_SAND ICELAST09
##  30.78088  16.06175
```

```
stdx
```

```
##  DIS_SAND ICELAST09
## 26.151787  2.889545
```

Your tasks are now the following:

1. Implement the model in Stan and sample from the posterior for the parameters α and β. Check for

convergence of the MCMC chain and examine the autocorrelation of the samples. Visualize the posterior for α and β and discuss the results.

2. Calculate the posterior covariance between α and β_3. How does this differ from the prior covariance and why?
3. Visualize the posterior of θ as a function of ICELAST09 when DISSAND is set to its mean value and in both cases when BOTTOMCOV=0 and BOTTOMCOV=1. That is, draw the median and 95% credible interval of the prediction function within the range from minimum to maximum value of ICELAST09 in the data.
4. Visualize the posterior distribution of θ at location where DIS_SAND is 60 and ICELAST is 18 for both vegetated and non-vegetated bottom types as well as their difference.
5. How does the difference in θ for vegetated and non-vegetated bottom differ from $\phi = \Delta\theta = \theta_0 - \theta_1$ in exercise 3 of week 2 and $\delta\mu$ in exercise 2 of week 4? Would you say that the result concerning the effect of vegetation is consistent in all these different analyses? Which analysis would you prefer?
6. Visualize the posterior distribution of $\tilde{y}$ corresponding to the number sampling occasions where white fish is present out of a total 10 repeated sampling occasions at location where DIS_SAND is 60 and ICELAST is 18 for both vegetated and non-vegetated bottom types.

Task 1

```r
library(ggplot2)
library(StanHeaders)
library(rstan)
```

```
## rstan (Version 2.21.2, GitRev: 2e1f913d3ca3)

## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)

## Do not specify '-march=native' in 'LOCAL_CPPFLAGS' or a Makevars file
```

```r
library(gridExtra)
library(see)
library(boot)
library(Rlab)
```

```
## Rlab 2.15.1 attached.

##
## Attaching package: 'Rlab'

## The following objects are masked from 'package:stats':
##
##     dexp, dgamma, dweibull, pexp, pgamma, pweibull, qexp, qgamma,
##     qweibull, rexp, rgamma, rweibull

## The following object is masked from 'package:datasets':
##
##     precip
```

```r
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
set.seed(123)
```

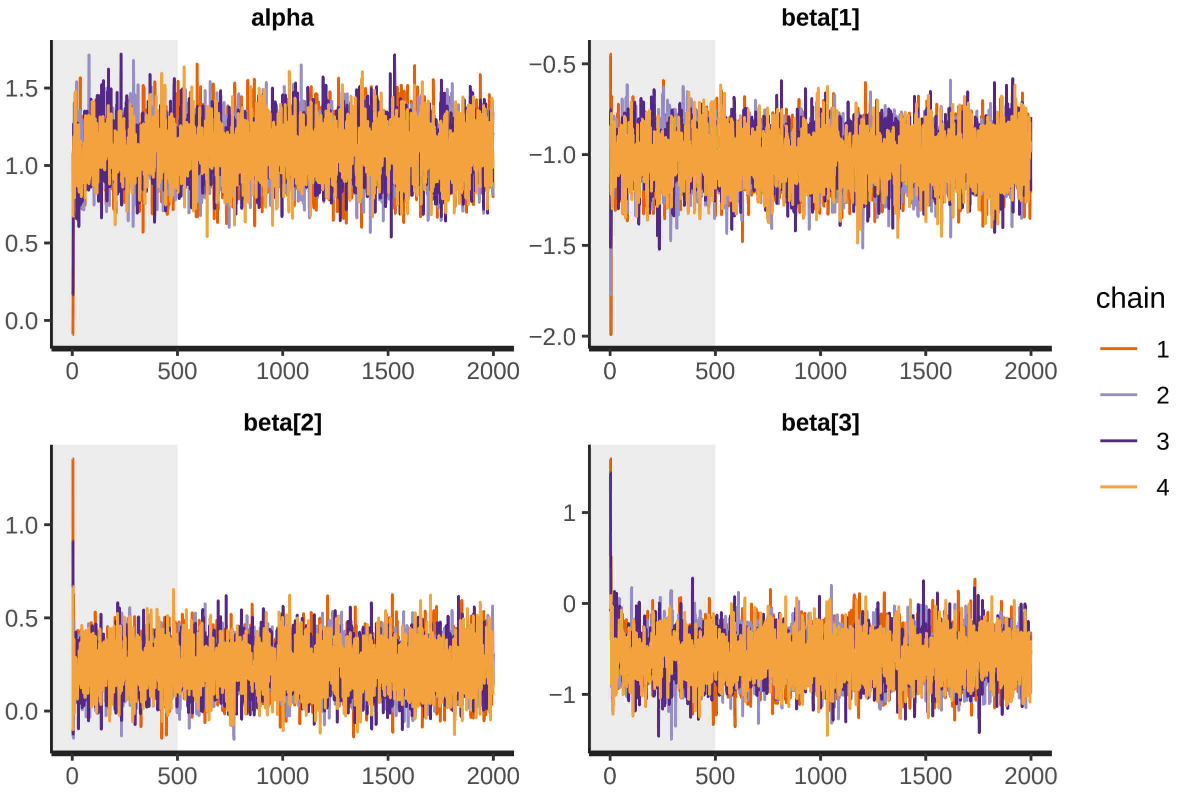
```r
whitefish_model = "data{
  int<lower=0> n;     // number of sampling sites
  matrix[n, 3] X;     // the covariate matrix
  int<lower=0> y[n];  // vector of white fish presence-absence observations
}
parameters{
  real alpha;
  vector[3] beta;
}
model{
  y ~ bernoulli_logit_glm(X, alpha, beta);
  alpha ~ normal(0, sqrt(10));
  beta ~ normal(0, sqrt(10));
}
"
```

```
data <- list (n=length(y), X=X, y=y)

set.seed(123)
post = stan(model_code=whitefish_model, data=data, warmup=500, iter=2000, chains=4, thin=1, control = li
```

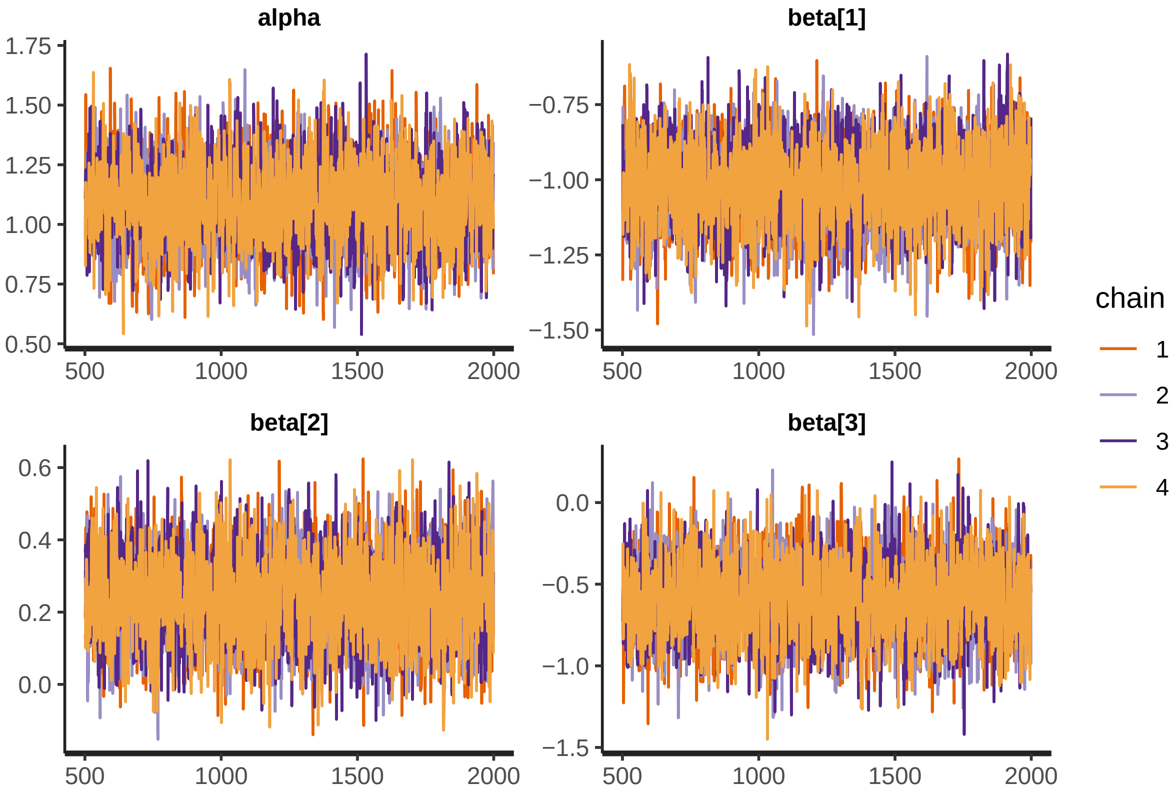
```
print(post, pars=c("alpha","beta"))
```

```
## Inference for Stan model: 2729fef54968f012f10591193140aee1.
## 4 chains, each with iter=2000; warmup=500; thin=1;
## post-warmup draws per chain=1500, total post-warmup draws=6000.
##
##          mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## alpha    1.09       0 0.17  0.77  0.98  1.09  1.20  1.42  3438    1
## beta[1] -1.02       0 0.13 -1.29 -1.10 -1.01 -0.92 -0.77  3901    1
## beta[2]  0.24       0 0.12  0.01  0.16  0.24  0.32  0.47  3644    1
## beta[3] -0.59       0 0.23 -1.04 -0.74 -0.59 -0.43 -0.13  3344    1
##
## Samples were drawn using NUTS(diag_e) at Mon Dec 07 20:10:51 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

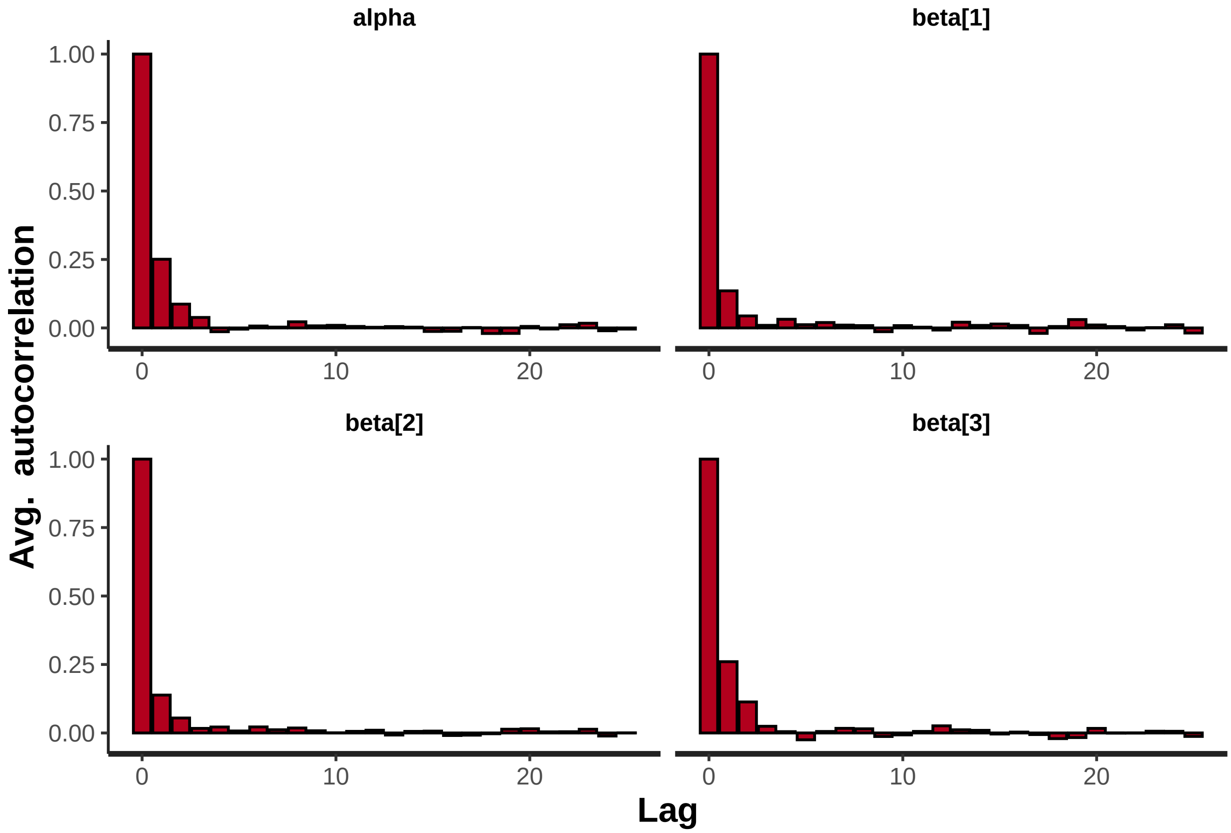
```
plot(post, pars=c("alpha","beta"), plotfun= "trace", inc_warmup = TRUE)
```



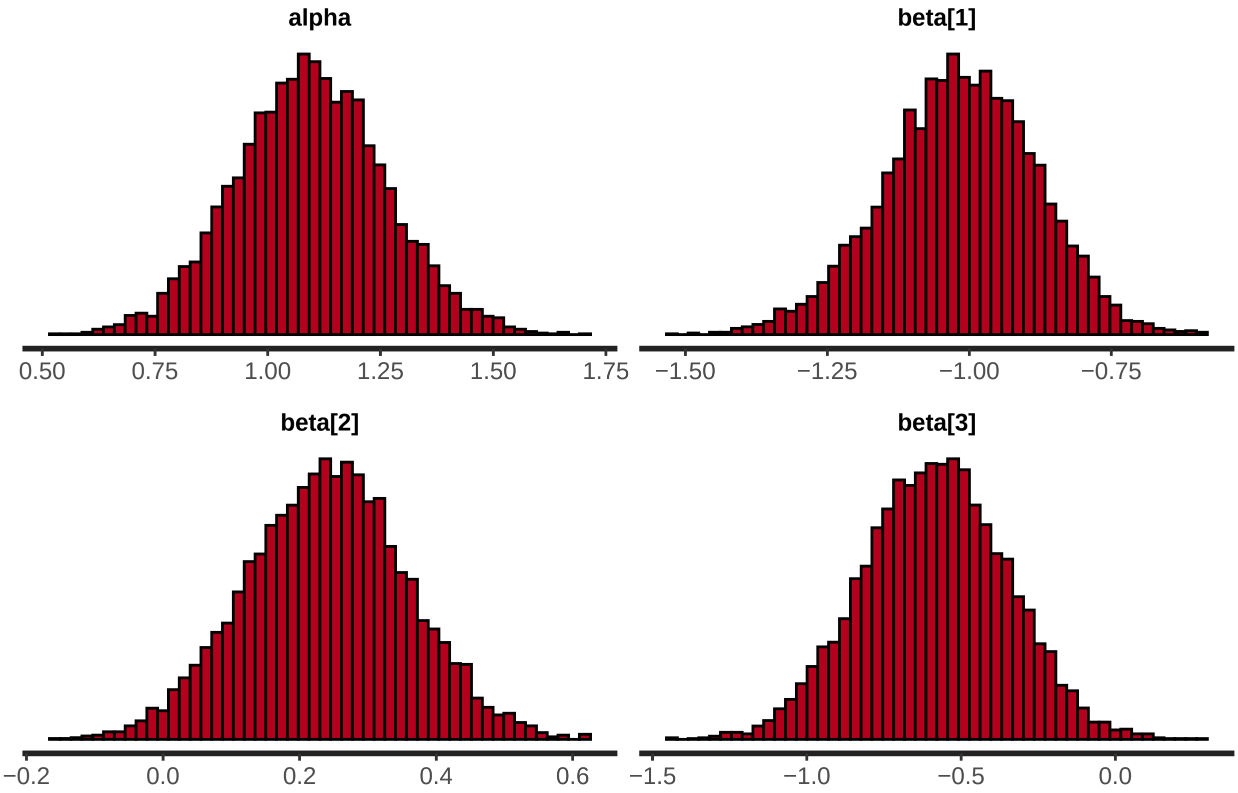
```r
plot(post, pars=c("alpha","beta"), plotfun= "trace", inc_warmup = FALSE)
```



```
stan_ac(post, c("alpha","beta"),inc_warmup = FALSE, lags = 25)
```



```r
plot(post, plotfun = "hist", pars = c("alpha", "beta"), bins=50)
```



The distribution for beta[1] indicates that as the distance to sandy shores increases the probability of presence of whitefish decreases because the mean is at around -1. The distribution of beta[3] als indicates a negative relationship between vegetation cover status and probability of presence of whitefish. For beta[2] the mean is at around 0.25, indicating a positive relationship. As the length of the ice cover during the winter increases the probability of presence of whitefish also increases. The distribution for alpha shows that the y-intercept for theta is positive, normally distributed around 1.1.

Task 2

```
alpha = as.matrix(post, pars="alpha")
beta1 = as.matrix(post, pars="beta")[1:length(alpha)]
beta2 = as.matrix(post, pars="beta")[(length(alpha)+1):(2*length(alpha))]
beta3 = as.matrix(post, pars="beta")[(2*length(alpha)+1):(3*length(alpha))]

cov(alpha, beta3)
```

```
##              [,1]
## alpha -0.02751622
```

```
cov(as.matrix(rnorm(2000, 0, sqrt(10))))
```

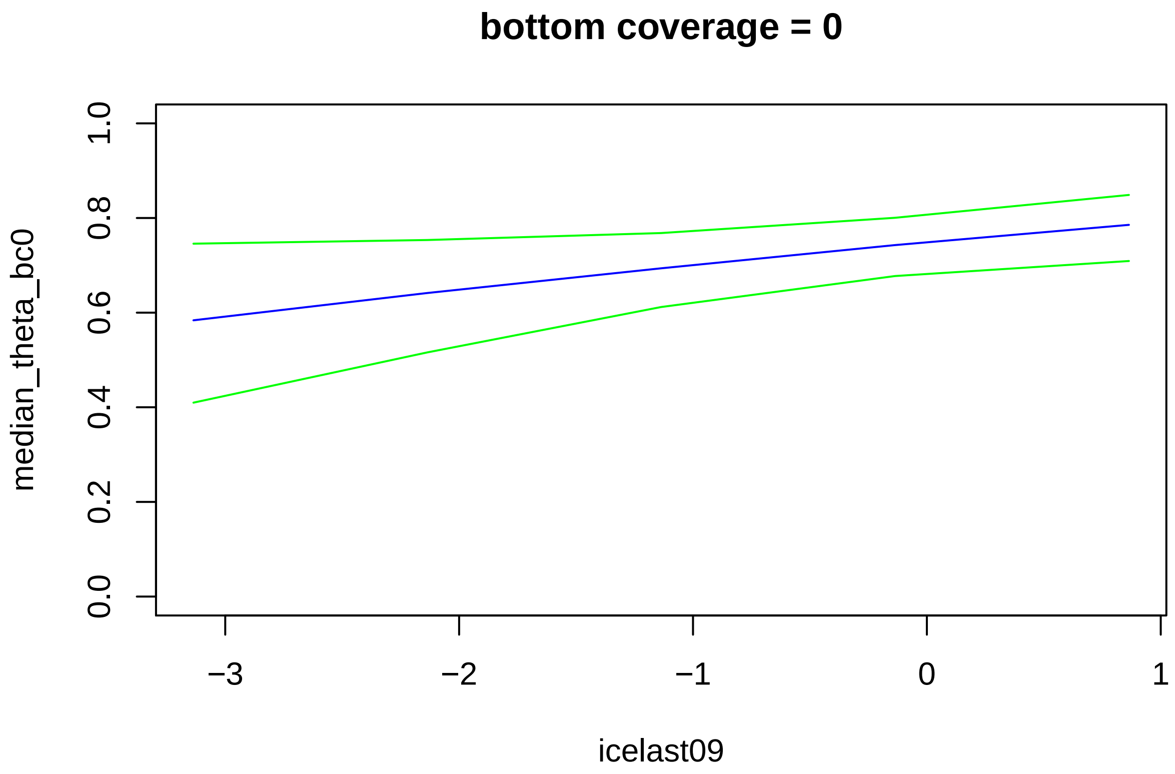
```
##          [,1]
## [1,] 10.01095
```

The prior covariance of N(0,10) is positive and much bigger than the posterior covariance of alpha and beta3, which is negative. The posterior covariance of alpha and beta3 indicates that as alpha increases beta3 decreases. This is consisent with what was seen in the histograms of task 1, where alpha had a positive mean and beta3 a negative one.

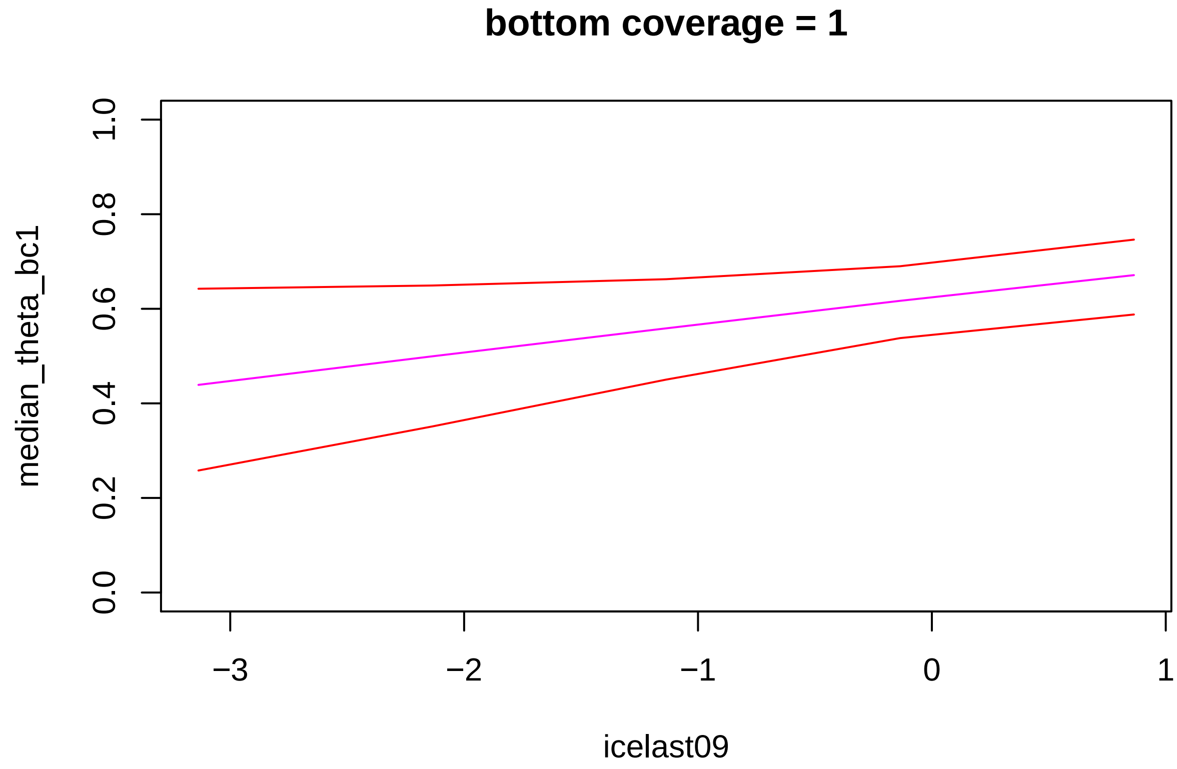
Task 3

```
set.seed(123)
icelast09 = seq(min(X$ICELAST09), max(X$ICELAST09))

theta_bc0 = matrix(NA, length(icelast09), length(alpha))  # matrix of posterior samples of theta when b
median_theta_bc0 = rep(NA, length(icelast09))        # posterior median of theta when bottom coverage = 
int_theta_bc0 = matrix(NA, length(icelast09), 2)  # posterior 95% interval of theta when bottom coverag

theta_bc1 = matrix(NA, length(icelast09), length(alpha))  # matrix of posterior samples of theta when b
median_theta_bc1 = rep(NA, length(icelast09))        # posterior median of theta when bottom coverage = 
int_theta_bc1 = matrix(NA, length(icelast09), 2)  # posterior 95% interval of theta when bottom coverag

for (i in 1:length(icelast09)) {
  theta_bc0[i,] = inv.logit(alpha + beta1 * mean(X$DIS_SAND) + beta2 * icelast09[i] + beta3 * 0)
  median_theta_bc0[i] = median(theta_bc0[i,])
  int_theta_bc0[i,] = quantile(theta_bc0[i,], probs=c(0.025,0.975))

  theta_bc1[i,] = inv.logit(alpha + beta1 * mean(X$DIS_SAND) + beta2 * icelast09[i] + beta3 * 1)
  median_theta_bc1[i] = median(theta_bc1[i,])
  int_theta_bc1[i,] = quantile(theta_bc1[i,], probs=c(0.025,0.975))
}
```

```
plot(icelast09, median_theta_bc0, type="l", col="blue", main="bottom coverage = 0", ylim=c(0,1)) # post
lines(icelast09, int_theta_bc0[,1], col="green")
lines(icelast09, int_theta_bc0[,2], col="green") # 95% interval of theta when bottom coverage = 0
```



```
plot(icelast09, median_theta_bc1, type="l", col="magenta", main="bottom coverage = 1", ylim=c(0,1)) # p
lines(icelast09, int_theta_bc1[,1], col="red")
lines(icelast09, int_theta_bc1[,2], col="red") # 95% interval of theta when bottom coverage = 1
```

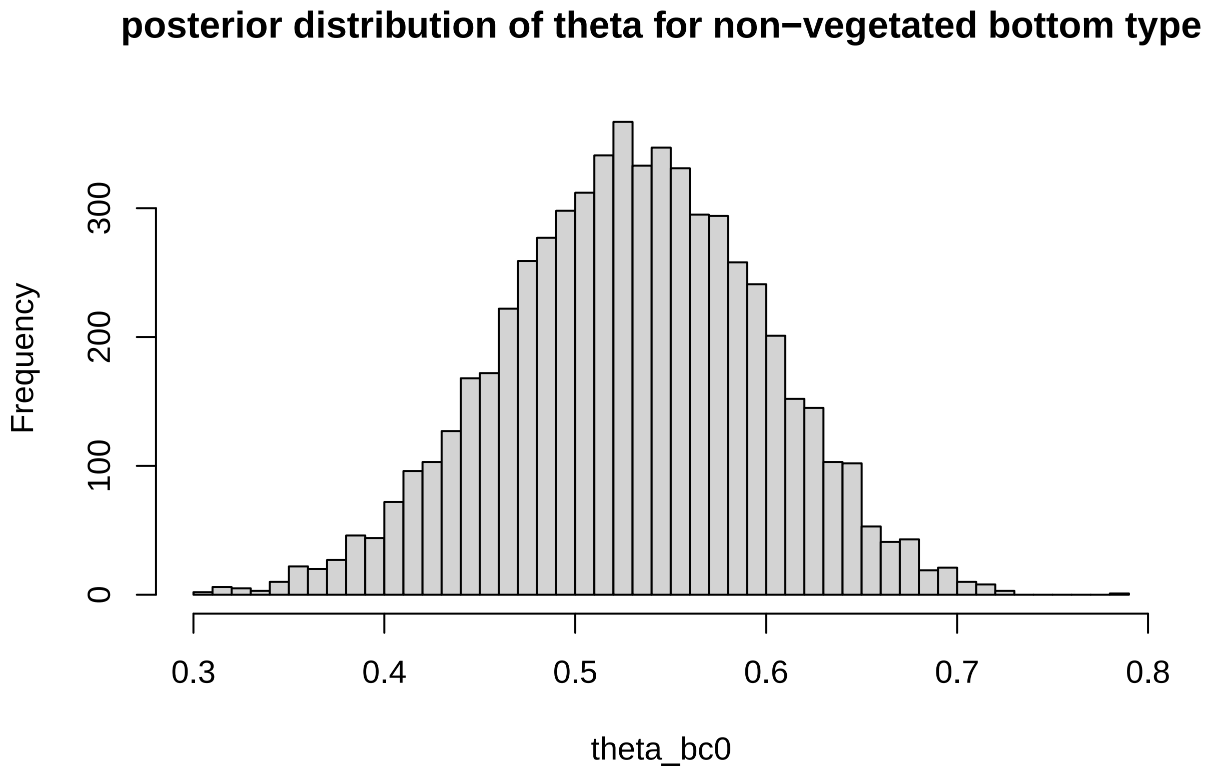


Task 4

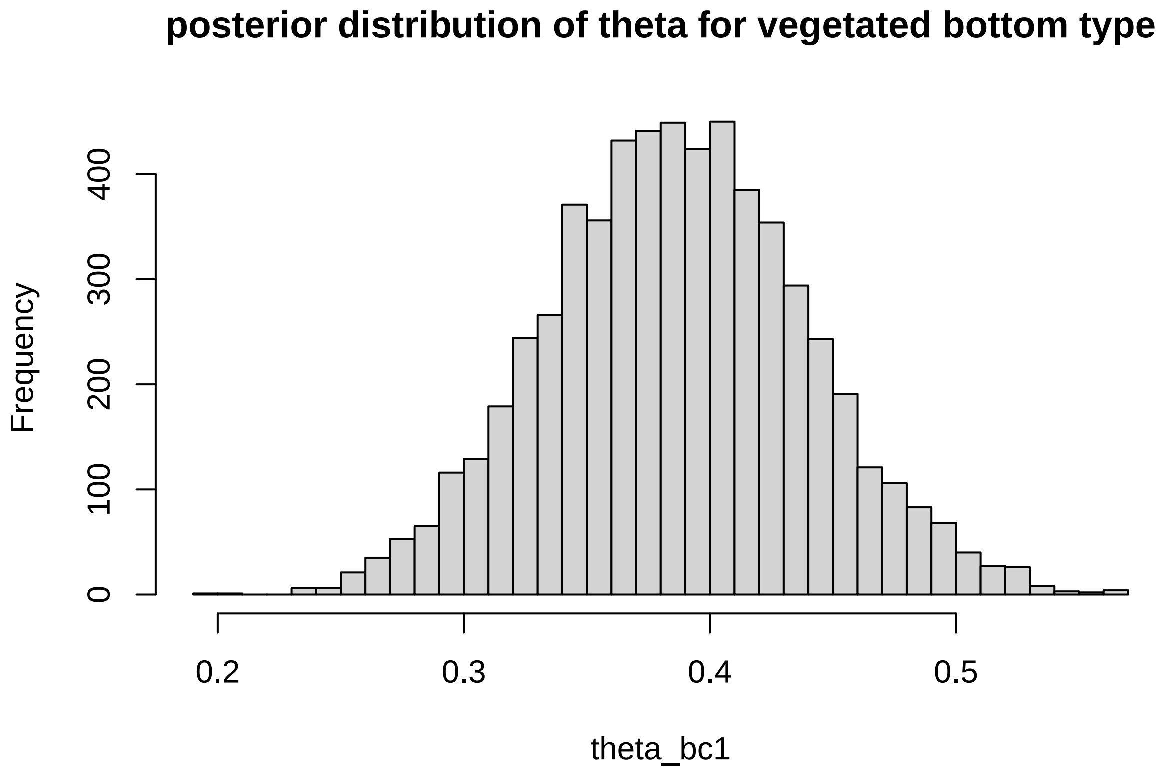
```
(c(60, 18) - t(mx)) / t(stdx)
```

```
##      DIS_SAND ICELAST09
## [1,]  1.11729 0.6707792
```

```
theta_bc0 = inv.logit(alpha + beta1 * 1.11729  + beta2 * 0.6707792 + beta3 * 0)
theta_bc1 = inv.logit(alpha + beta1 * 1.11729  + beta2 * 0.6707792 + beta3 * 1)

hist(theta_bc0, breaks=50, main='posterior distribution of theta for non-vegetated bottom type')
```

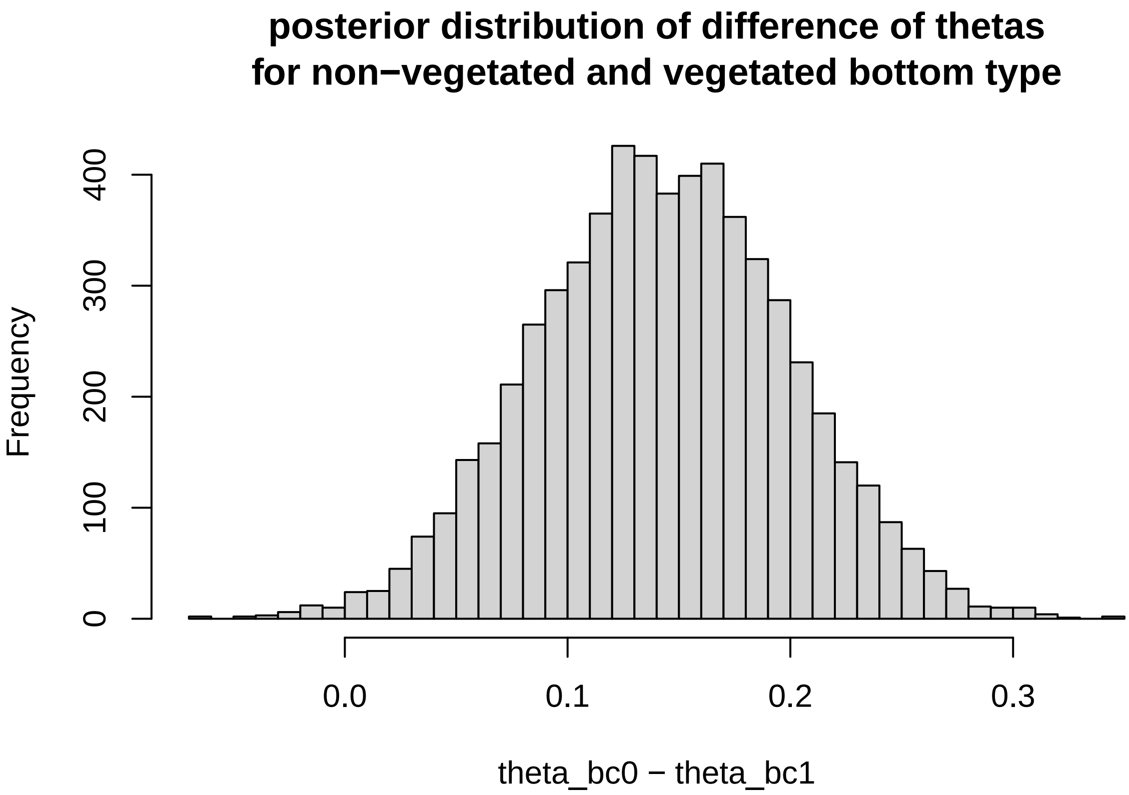


```r
hist(theta_bc1, breaks=50, main='posterior distribution of theta for vegetated bottom type')
```



posterior distribution of theta for vegetated bottom type

```
hist(theta_bc0 - theta_bc1, breaks=50, main='posterior distribution of difference of thetas\nfor non-veg
```



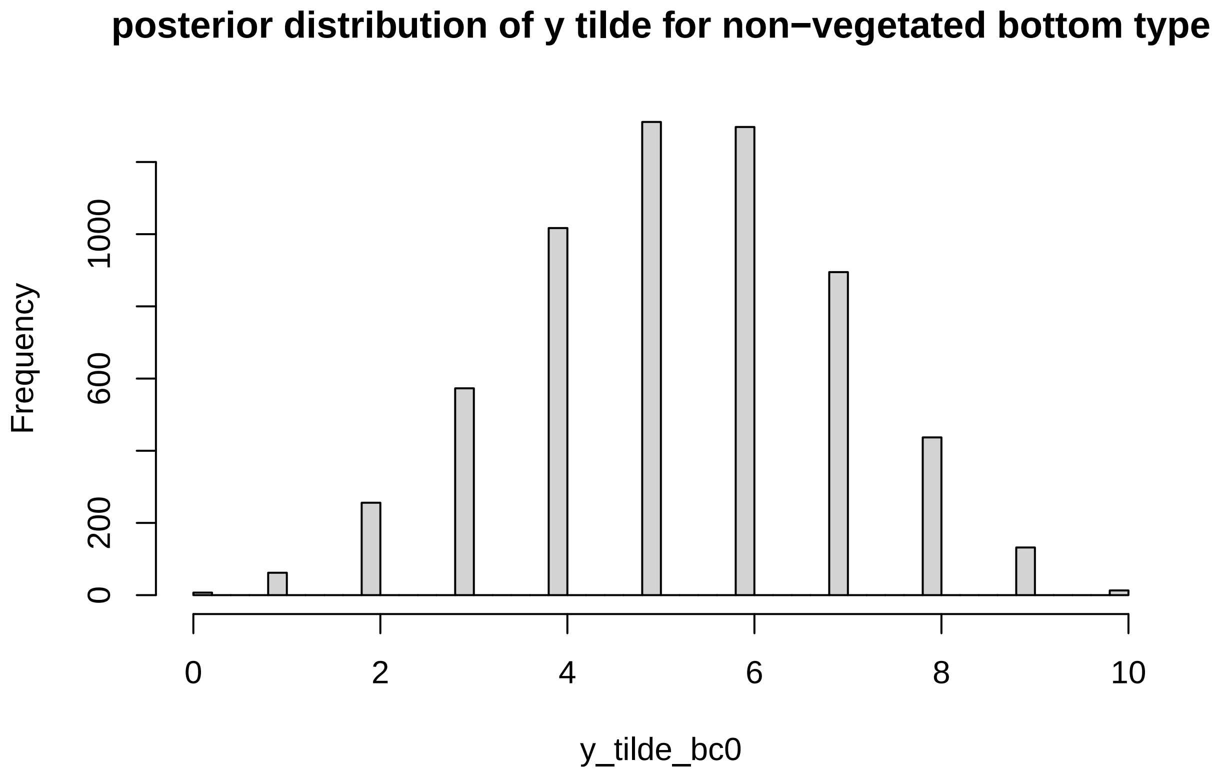
Task 5

The above distribution for the difference in θ for vegetated and non-vegetated bottom has a mean of about 0.14. The distribution of ϕ in week 2 exercise 3 has a mean of approximately 0.22. In week 4 exercise 2 the mean of $\delta\mu$was around 0.16 and thereby only slightly higher than in this week. This means that the probability that $\theta_0 > \theta_1$ was highest in the approach for week 2 and a bit lower and very similar for weeks 4 and 6 (this week). Overall the differences are not very big so I think the results are consistent with each other, especially for weeks 4 and 6. I think the best results might be achieved in week 6. It seemed to make sense to make different models for each area due to the different environmental conditions (week4) but I think the approach this week indirectly accounts for the regional differences by including the teo additional variables. This approach was simpler and made it possible to use all samples to estimate the parameters, rather than relying on a small set of observations for some of the areas as in week4.

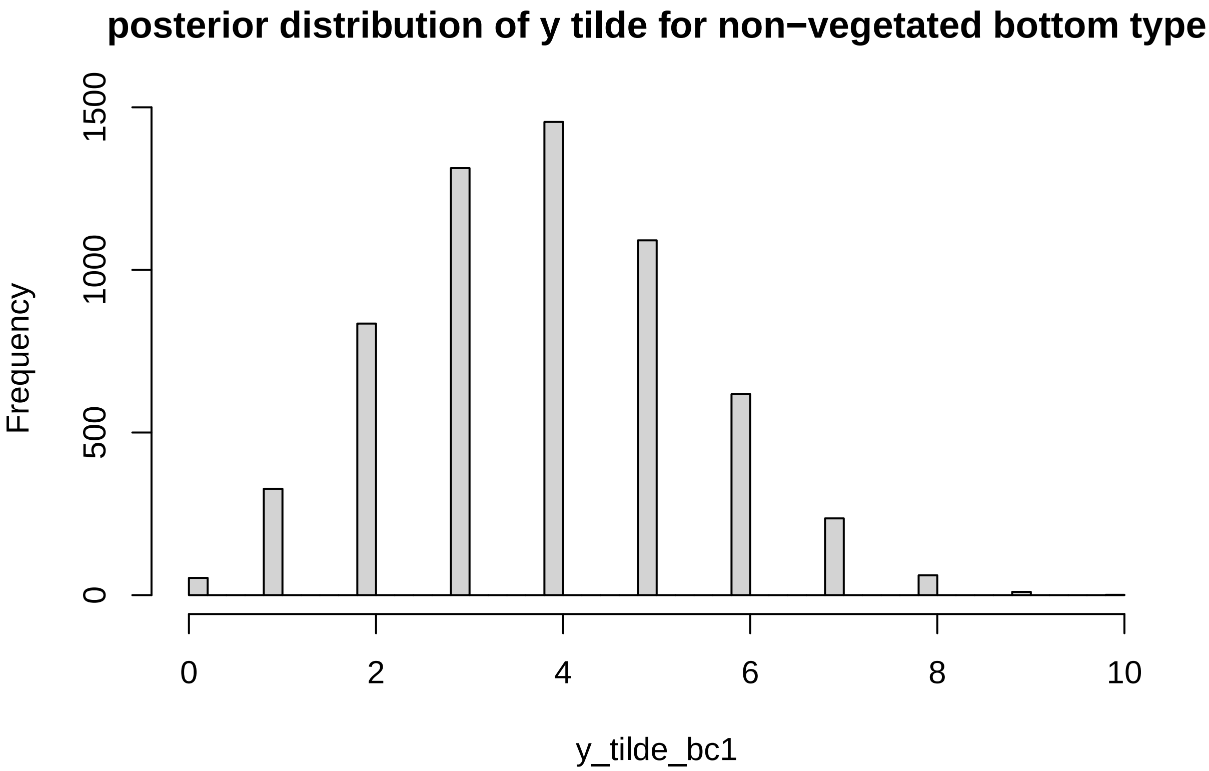
Task 6

```
y_tilde_bc0 = rep(NA, length(alpha))   # posterior of y tilde when bottom coverage = 0
y_tilde_bc1 = rep(NA, length(alpha))   # posterior of y tilde when bottom coverage = 1

for (i in 1:length(alpha)) {
  y_tilde_bc0[i] = sum(rbern(10, inv.logit(alpha[i] + beta1[i] * 1.11729  + beta2[i] * 0.6707792 + beta3

  y_tilde_bc1[i] = sum(rbern(10, inv.logit(alpha[i] + beta1[i] * 1.11729  + beta2[i] * 0.6707792 + beta3
}
```

```
hist(y_tilde_bc0, breaks=50, main='posterior distribution of y tilde for non-vegetated bottom type')
```



```
hist(y_tilde_bc1, breaks=50, main='posterior distribution of y tilde for non-vegetated bottom type')
```



Grading

Total 20 points Steps 1, 3 and 4 give 4 points each, steps 5 and 6 give 2 points each and step 2 gives 1 point if correctly solved. In other steps except 2 you may give half of the points if the step is solved half correctly. This could mean that some of the tasks have not been done (e.g. discussion is missing), there is only small typo that makes the final answer wrong or discussion is clearly not relevant or appropriate.

References

Lari Veneranta, Richard Hudd and Jarno Vanhatalo (2013). Reproduction areas of sea-spawning Coregonids reflect the environment in shallow coastal waters. Marine Ecology Progress Series, 477:231-250. http://www.int-res.com/abstracts/meps/v477/p231-250/