Effect of bottom coverage to larval presence

Week6-ex1, solution

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")

# Vector of area indexes. Let's change the area names to numerical area codes.
area = as.numeric(factor(data$AREANAME))

# 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.

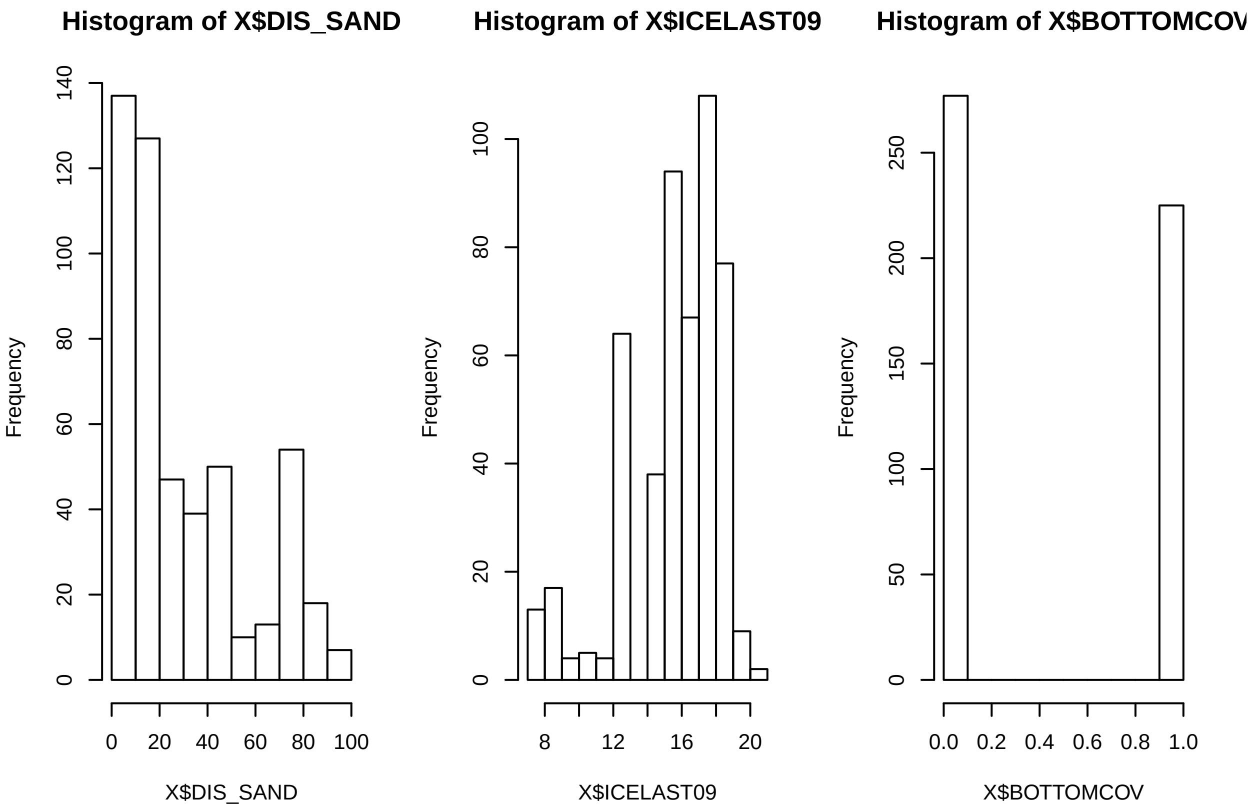
$$
\begin{aligned}
y_i &\sim \text{Bernoulli}(\theta_i) \\
\text{logit}(\theta_i) &= \alpha + X\beta \\
\alpha, \beta_1, \beta_2, \beta_3 &\sim N(0, 10)
\end{aligned}
$$

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.

```r
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))
```

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 correlation between α and β_3. How does this differ from the prior correlation 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.

Solution

1.

Let's load the needed libraries into R and write the Stan model

```
library(ggplot2)
```

```
## Warning: replacing previous import 'vctrs::data_frame' by 'tibble::data_frame'
## when loading 'dplyr'
```

```
library(StanHeaders)
library(rstan)
```

```
## rstan (Version 2.19.3, 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)
```

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

whitefish.model = "data{
  int<lower=0> n;       // number of sampling sites
  int<lower=0> d;       // number of covariates
  int<lower=0> y[n];    // white fish presence/absence per site
  matrix[n,d] X;        // matrix of covariates
}
parameters{
  vector[d] beta;
  real alpha;
}
model{

  beta ~ normal(0,sqrt(10));
  alpha ~ normal(0,sqrt(10));

  y ~ bernoulli_logit(alpha + X*beta);
  // Note that the above line is the same as follows
  // for( i in 1 : n ){
  //  y[i] ~ bernoulli_logit(alpha + X[i,]*beta);
  //}
}"
```

Next, we create our data lists and do the sampling with both models

```
data.stan <- list ("n"=length(y),"d"=dim(X)[2], "y"=y, "X"=X)    # no vegetation data
post=stan(model_code=whitefish.model,data=data.stan,
                warmup=200,iter=400,chains=3)
```

```
## Trying to compile a simple C file
## Running /usr/lib/R/bin/R CMD SHLIB foo.c
## gcc -std=gnu99 -I"/usr/share/R/include" -DNDEBUG   -I"/home/local/jpvanhat/R/x86_64-pc-linux-gnu-lib
## In file included from /home/local/jpvanhat/R/x86_64-pc-linux-gnu-library/3.6/RcppEigen/include/Eigen/
##                  from /home/local/jpvanhat/R/x86_64-pc-linux-gnu-library/3.6/RcppEigen/include/Eigen/
##                  from /home/local/jpvanhat/R/x86_64-pc-linux-gnu-library/3.6/StanHeaders/include/stan
##                  from <command-line>:0:
## /home/local/jpvanhat/R/x86_64-pc-linux-gnu-library/3.6/RcppEigen/include/Eigen/src/Core/util/Macros.h
##  namespace Eigen {
##  ^~~~~~~~~
## /home/local/jpvanhat/R/x86_64-pc-linux-gnu-library/3.6/RcppEigen/include/Eigen/src/Core/util/Macros.h
##  namespace Eigen {
##                  ^
## In file included from /home/local/jpvanhat/R/x86_64-pc-linux-gnu-library/3.6/RcppEigen/include/Eigen/
##                  from /home/local/jpvanhat/R/x86_64-pc-linux-gnu-library/3.6/StanHeaders/include/stan
##                  from <command-line>:0:
## /home/local/jpvanhat/R/x86_64-pc-linux-gnu-library/3.6/RcppEigen/include/Eigen/Core:96:10: fatal erro
##  #include <complex>
##           ^~~~~~~~~
## compilation terminated.
## /usr/lib/R/etc/Makeconf:168: recipe for target 'foo.o' failed
## make: *** [foo.o] Error 1
```

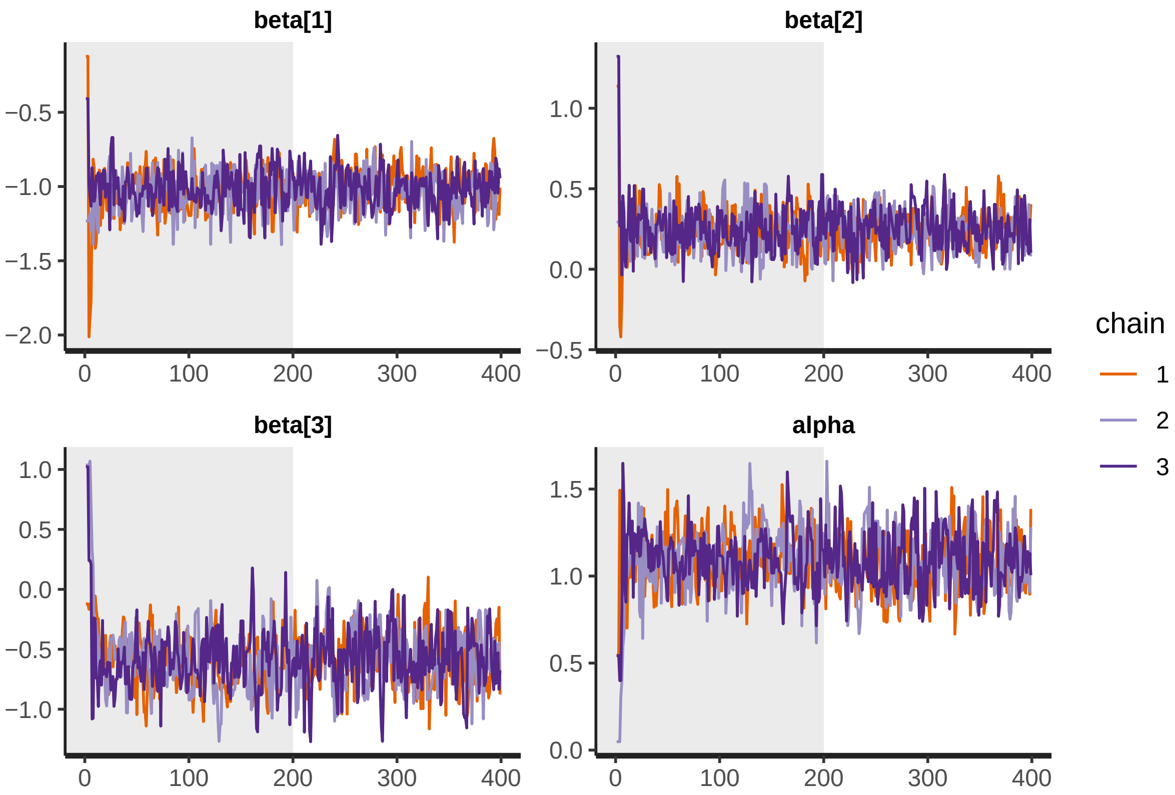
```
## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#bulk-ess
```

Let's examine convergence and autocorrelation:

```
print(post)  #pars=c("s","mu"),
```

```
## Inference for Stan model: ac639f1bb7d2d4219cd6721b5a8004ca.
## 3 chains, each with iter=400; warmup=200; thin=1; 
## post-warmup draws per chain=200, total post-warmup draws=600.
## 
##            mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff Rhat
## beta[1]   -1.02    0.01 0.13   -1.28   -1.11   -1.01   -0.93   -0.78   386 1.00
## beta[2]    0.24    0.01 0.12    0.02    0.16    0.24    0.32    0.48   400 1.01
## beta[3]   -0.57    0.01 0.24   -1.06   -0.75   -0.57   -0.41   -0.12   317 1.00
## alpha      1.08    0.01 0.17    0.77    0.96    1.07    1.21    1.44   331 1.01
## lp__    -250.82    0.09 1.39 -254.09 -251.53 -250.59 -249.82 -248.97   261 1.01
## 
## Samples were drawn using NUTS(diag_e) at Wed Dec  9 14:26:15 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).
```

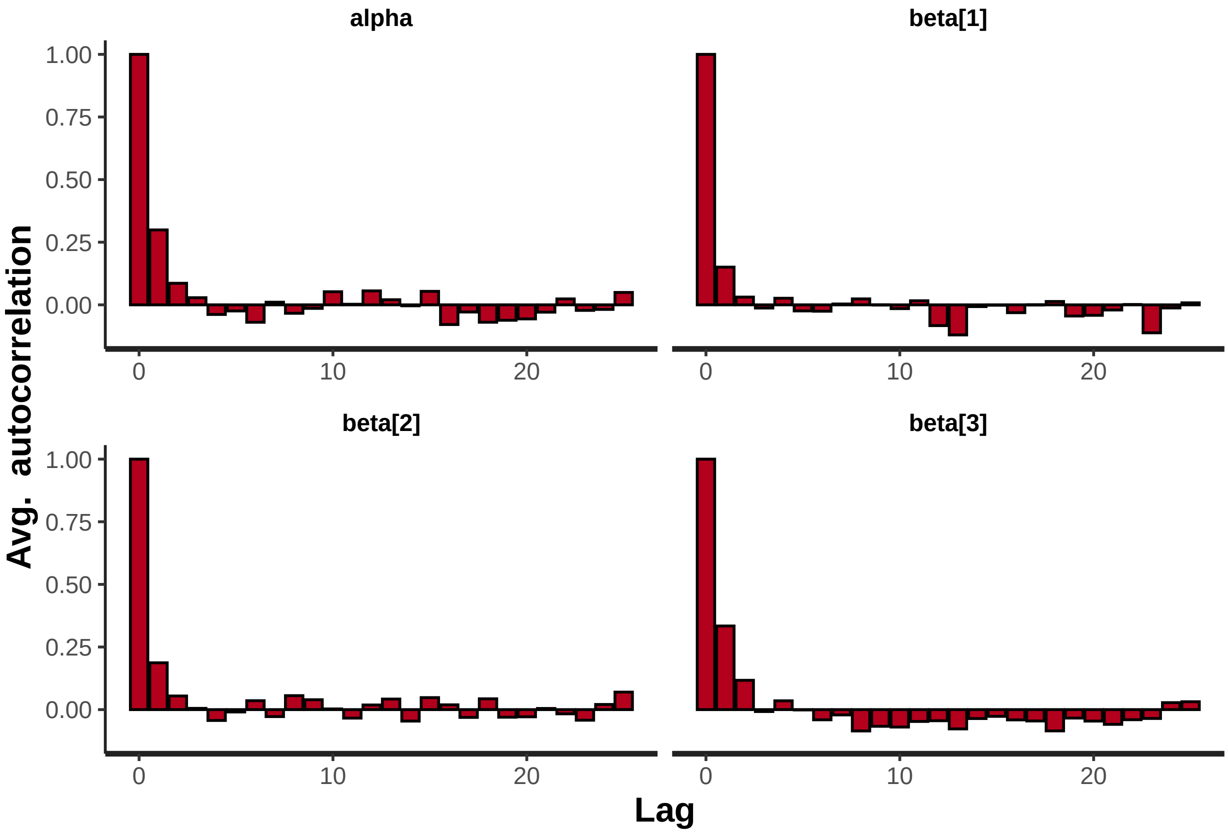
```
plot(post, plotfun="trace", inc_warmup = TRUE)
```



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

```
## Warning: Ignoring unknown parameters: fun.y
```

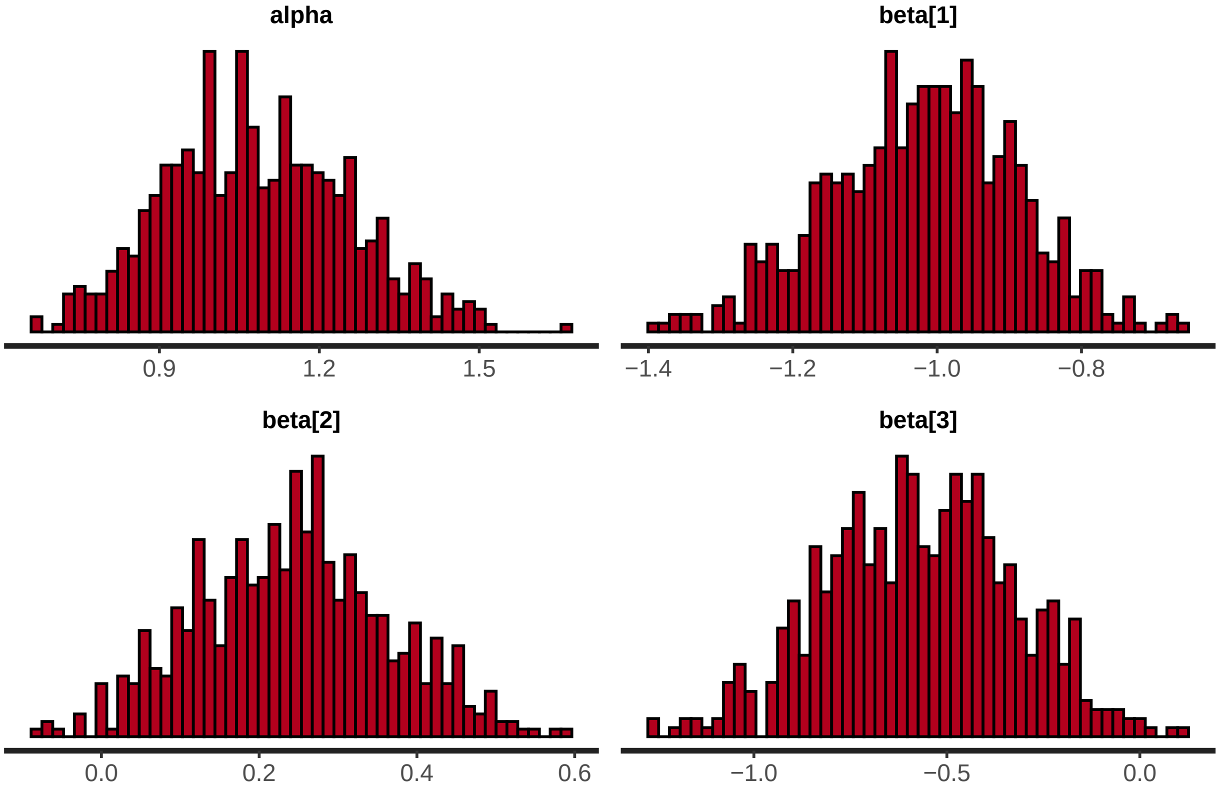
```
## No summary function supplied, defaulting to `mean_se()`
## No summary function supplied, defaulting to `mean_se()`
## No summary function supplied, defaulting to `mean_se()`
## No summary function supplied, defaulting to `mean_se()`
```



According to the Rhat summary and the plotted trace plots of sample chains the chains seem to have converged. The autocorrelation of the Markov chain samples is very small and, hence, not a problem.

Let's then visualize the posterior for α, and β

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



From the above figures and from the earlier summary print-out we can see that all the covariates have significant effect on the white fish larvae occurrence probability. The effect of DIS_SAND is negative (occurrence probability decreases with increasing distance to sandy shore), the effect of ICELAST09 is positive (the probability increases when the length of ice cover season increases) and the effect of BOTTOMCOV (bottom vegetation) is negative.

2

The posterior correlation between α and β is

```
samples <- as.matrix(post)    # combine all chains into one matrix in R workspace
cor(samples[,"beta[3]"],samples[,"alpha"])
```

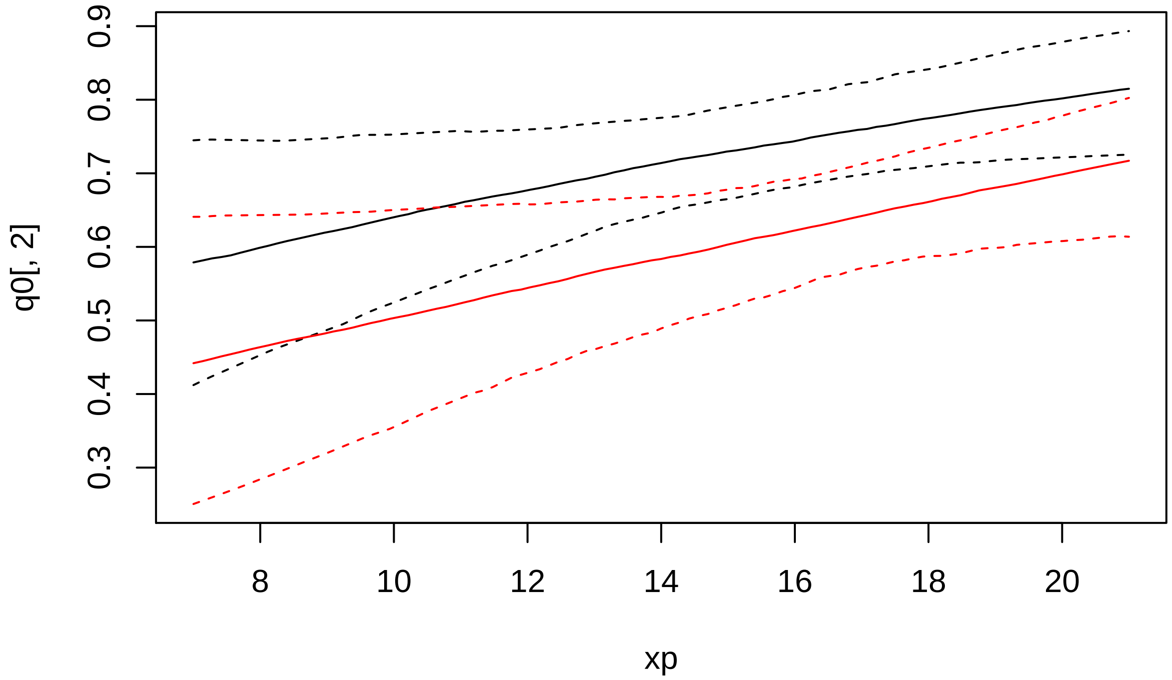
```
## [1] -0.7383351
```

The posterior correlation is significantly less than zero whereas the prior correlation is exactly zero. The reason is that the likelihood function induces correlation between these parameters. ## 3

We will now 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. Note that we can make the prediction with standardized covariates and then plot the response curve with original covariate values. In the standardized covariates the mean of DISSAND is zero so we can ignore it from the prediction

```
xp = seq(min(X[,2]), max(X[,2]), length=101)  # the evaluation points
q0 = matrix(nrow = 101,ncol=3)
q1 = matrix(nrow = 101,ncol=3)

for (i in 1:length(xp)) {
  f0 = samples[,"alpha"] + samples[,"beta[2]"]*xp[i]  # the response when there is no vegetation cover
  f1 = samples[,"alpha"] + samples[,"beta[2]"]*xp[i] + samples[,"beta[3]"] # the response when there is
  th0 = 1/(1 + exp(- f0))
  th1 = 1/(1 + exp(- f1))
  q0[i,] = quantile(th0,probs = c(0.025,0.5,0.975))
  q1[i,] = quantile(th1,probs = c(0.025,0.5,0.975))
}
#plot(x,y/n)                           # observed
xp = xp*stdx[2] + mx[2]
plot(xp,q0[,2], type="l", col="black", ylim=c(min(q1[,1]),max(q0[,3]))) # mean
lines(xp,q0[,1], lty=2, col="black")    # 95% interval of f
lines(xp,q0[,3], lty=2, col="black")    # 95% interval of f
lines(xp,q1[,2], type="l", col="red") # mean
lines(xp,q1[,1], lty=2, col="red")    # 95% interval of f
lines(xp,q1[,3], lty=2, col="red")    # 95% interval of f
```

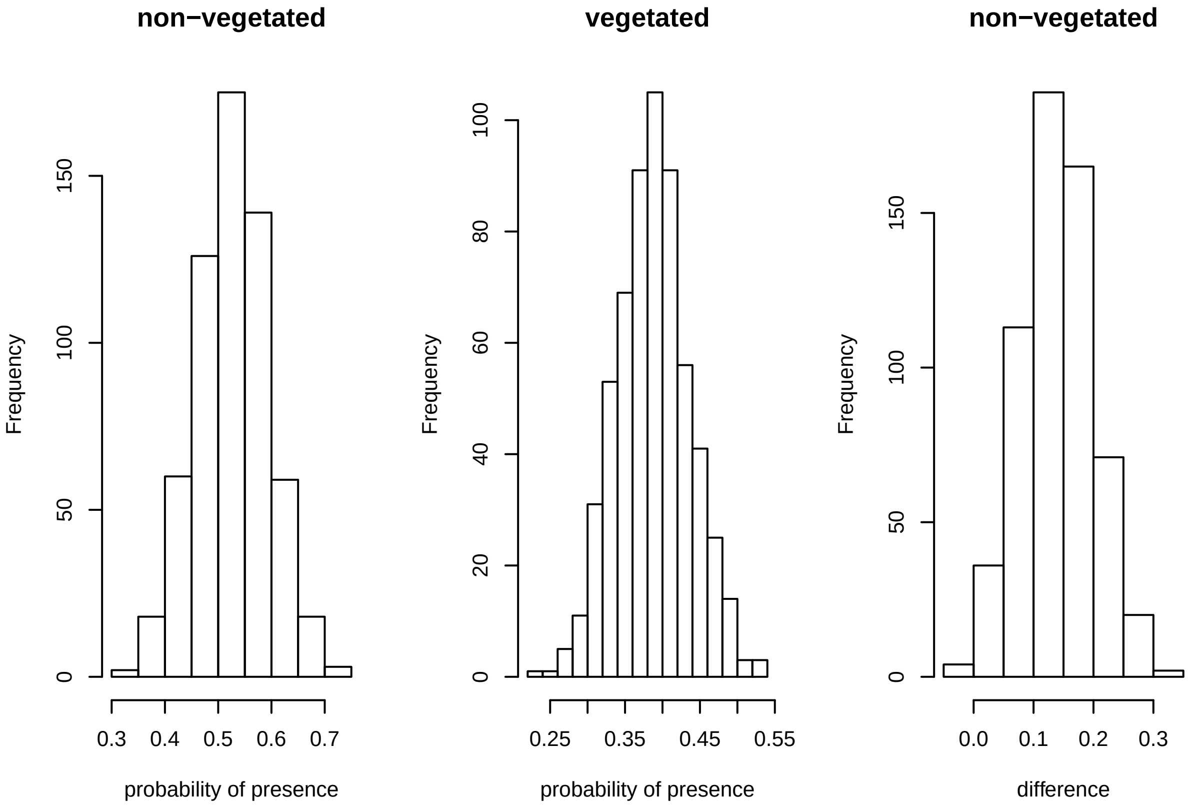


4

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.

```
# the covariates in the prediction location for both vegetated and non-vegetated area
xp =rbind(c(60,18,0),c(60,18,1))
# standardize the covariates
xp[,1:2] = (xp[,1:2]-t(replicate(2,mx)))/t(replicate(2,stdx))
# Make predictions, note that the first column is the
# prediction for non-vegetated bottom and the second
# column for the vegetated bottom
f = replicate(2,samples[,"alpha"]) + samples[,1:3]%*%t(xp)
th = 1/(1 + exp(-f))
par(mfrow=c(1,3))
hist(th[,1], main="non-vegetated", xlab="probability of presence")
hist(th[,2], main="vegetated", xlab="probability of presence")
```

```r
hist(th[,1]-th[,2], main="non-vegetated", xlab="difference")
```



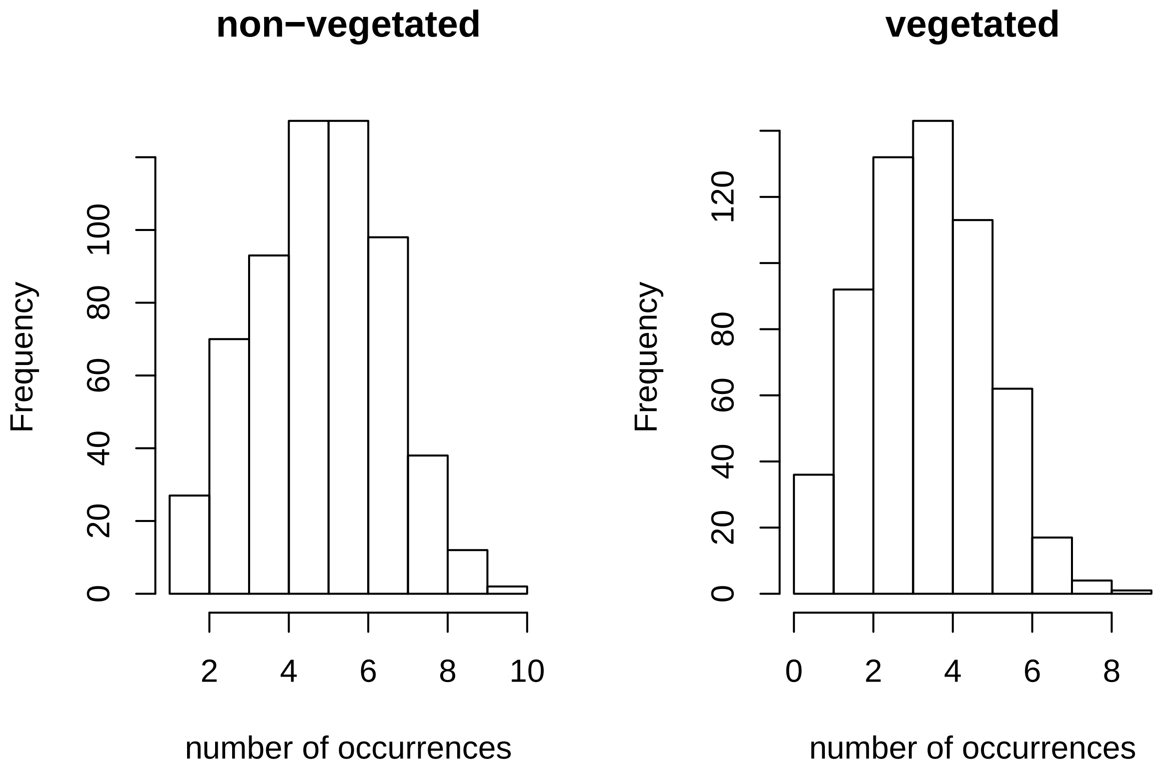
5

When comparing the posterior distribution of the difference in probability of presence in vegetated and non-vegetated bottoms to the respective distributions of ϕ and $\Delta\mu$ in week 2 and week 4 exercises we can conclude that all these three distributions are concentrated around values of same order of magnitude; that is between 0-35%. However, the mean and spread of the distributions differ a bit. For example, the mean of the difference in the the above analyses is the smallest among these three. However, we can conclude that the message from these three analyses is rather consistent. The vegetation cover has negative effect to the white fish larval presence. I would prefer this last analysis since since accounting for other covariates, for example, makes the analysis more detailed.

6

Let's then visualize the posterior distribution of $\tilde{y}$, which corresponds 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.

```r
y.tilde0 = rbinom(th[,1],size = 10,prob = th[,1])
y.tilde1 = rbinom(th[,2],size = 10,prob = th[,2])
par(mfrow=c(1,2))
hist(y.tilde0, main="non-vegetated", xlab="number of occurrences")
hist(y.tilde1, main="vegetated", xlab="number of occurrences")
```



Grading

Total 20 points Steps 1, 3 and 4 give 5 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/