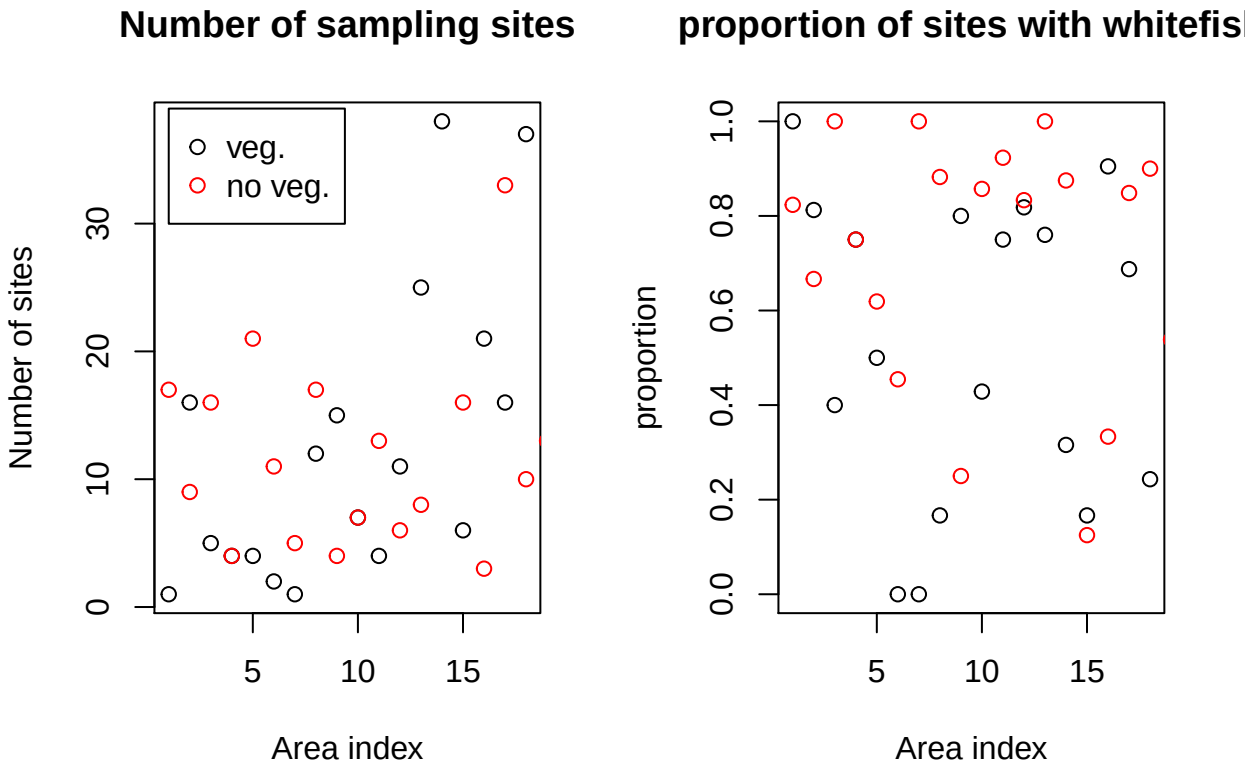
Effect of bottom coverage to larval presence

Week4-ex3, solution

In this exercise, we continue the analysis of the white fish larval areas (week 2, exercise 3). We are again interested in analysing whether or not bottom vegetation affects white fish larvae occurrence probability. However, instead of having a common probability of presence parameter across the Gulf of Bothnia, we expand the model so that it allows the probability of presence to vary between sampling areas. This modification to the model encodes an assumption that some areas may be more preferable to white fish than others.

Let's first explore the data a bit more.

```
# Read the data
data = read.csv("white_fishes_data.csv")
# Form a data table for sites without bottom vegetation
y.noveg = table(data$AREANAME[data$BOTTOMCOV==0], data$WHIBIN[data$BOTTOMCOV==0])
colnames(y.noveg) <- c("y=0","y=1")
N.noveg = rowSums(y.noveg)
# Form a data table for sites with bottom vegetation
y.veg = table(data$AREANAME[data$BOTTOMCOV==1], data$WHIBIN[data$BOTTOMCOV==1])
colnames(y.veg) <- c("y=0","y=1")
N.veg = rowSums(y.veg)

par(mfrow=c(1,2))
plot(N.veg, main="Number of sampling sites", xlab="Area index", ylab="Number of sites")
points(N.noveg, col="red")
legend(1, 39, c("veg.","no veg."),col=c("black","red"), pch=1, cex=1, box.lty=1)
plot(y.veg[,2]/N.veg, main="proportion of sites with whitefish", xlab="Area index", ylab="proportion")
points(y.noveg[,2]/N.noveg, col="red")
```




```
print(y.veg)
```

```
##
##                 y=0 y=1
##   Bjuroklubb      0   1
##   Bygdea          3  13
##   Haaparanta      3   2
##   Hailuoto        1   3
##   Harnosand       2   2
##   Hornslandet     2   0
##   Lohtaja         1   0
##   Luvia          10   2
##   Mikkelinsaaret  3  12
##   Mjolefjarden    4   3
##   Nordingra       1   3
##   Pietarsaari     2   9
##   Pitea           6  19
##   Pori           26  12
##   Siipyy          5   1
##   Storsand        2  19
##   Tore            5  11
##   Vaasa          28   9
```

The first figure above shows the number of sampling sites for each of the 19 study areas and both bottom vegetation types (with and without). The second figure shows the proportion of the sites with white fish

larvae within each area and bottom vegetation type. It is rather evident that there is considerable variation in the sample proportions of the second figure. However, we would want to know how much of this is actually due to varying probability of presence vs. pure chance. Note also, that there are no sampling sites in Kalajoki (sampling area number 7 below) with vegetation cover. Hence, we have missing data there.

```
N.veg
```

```
##     Bjuroklubb         Bygdea     Haaparanta       Hailuoto      Harnosand
##              1             16              5              4              4
##    Hornslandet        Lohtaja         Luvia Mikkelinsaaret   Mjolefjarden
##              2              1             12             15              7
##      Nordingra    Pietarsaari         Pitea           Pori         Siipyy
##              4             11             25             38              6
##       Storsand          Tore         Vaasa
##             21             16             37
```

We will denote by $\theta_{i,c}$ the probability that white fish larvae are present in area i at sites with $(c = 1)$ or without $(c = 0)$ bottom vegetation. The data will be denoted by $y_{i,c}$ and $N_{i,c}$ where the former denotes the number of sites with white fish larvae and the latter the total number of sites inside an area i with $(c = 1)$ or without $(c = 0)$ bottom vegetation. We will now implement the following model

$$
\begin{aligned}
y_{i,c} &\sim \text{Binom}(\theta_{i,c}, N_{i,c}) \\
\theta_{i,c} &\sim \text{Beta}(\mu_c s_c, s_c - \mu_c s_c) \\
\mu_c &\sim \text{Unif}(0, 1) \\
s_c &\sim \text{log-}N(4, 4).
\end{aligned}
$$

where μ_c is the prior mean of $\theta_{i,c}$ and s_c governs the uncertainty about it. The parametrization of log-Gaussian distribution $s_c \sim \text{log-}N(m, \sigma^2)$ is such that $E[\log(s_c)] = m$ and $Var[\log(s_c)] = \sigma^2$

1. Implement the model in Stan and sample from the posterior for the parameters. Check for convergence for all parameters, and examine what is the autocorrelation for s_c, μ_c and few $\theta_{i,c}$. Visualize the posterior for μ_c, s_c and $\theta_{i,c}, i = 1, \ldots, 19$.

2. Visualize also the posterior distributions of $\Delta\mu = \mu_0 - \mu_1$ and $\phi_i = \theta_{i,0} - \theta_{i,1}$ for each area $i = 1, \ldots, 19$.

3. Sample from the posterior predictive distribution of outcome $\tilde{y}_{19,c}$ of a new sampling with $\tilde{N}_{19} = 10$ in the sampling area $i = 19$ for both vegetated and non-vegetated sites. Visualize the resulting posterior samples as well as the posterior distribution for $\tilde{y}_{19,0} - \tilde{y}_{19,1}$.

4. Sample from the posterior predictive distribution of outcome $\tilde{y}_{20,c}$ of a new sampling with $\tilde{N}_{20} = 10$ in a new sampling area $i = 20$ (an area from where we don't have data yet) within the Gulf of Bothnia. Do this for both vegetated and non-vegetated sites. Visualize the resulting posterior samples as well as the posterior distribution for $\tilde{y}_{20,0} - \tilde{y}_{20,1}$.

5. The posterior distributions calculated in exercise 3 of week 2 correspond to the so called pooled estimate of θ_c. Discuss how does the posterior of the pooled θ_c differ from the population mean, μ_c, and from the individual $\theta_{i,c}$ in the hierarchical model? Which model seems more justified in your opinion and why?

Answer

```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
set.seed(123)

options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
model="
data{
    int<lower = 0> i;
    int y[i];
    int N[i];
}

parameters{
  real<lower = 0, upper = 1> theta[i];
  real<lower = 0, upper = 1> theta_20;
  real<lower = 0> mu;
  real<lower = 0> s;
}

model {
  s ~ lognormal(4,sqrt(4));
  mu ~ uniform(0,1);
  for (a in 1:i){
      theta[a] ~ beta(mu*s, s-mu*s);
      y[a] ~ binomial(N[a], theta[a]);
      }
  theta_20~ beta(mu*s, s-mu*s);
}
generated quantities {
  real ytilde_19;
  real ytilde_20;
  ytilde_19 = binomial_rng(10,theta[19]);
  ytilde_20 = binomial_rng(10,theta_20);
}
"
```

```r
# set data into named list
y_initial <- y.noveg[,2]
# y_initial <- append(y_initial, round(mean(y_initial)), 6)
N_initial <- N.noveg
```

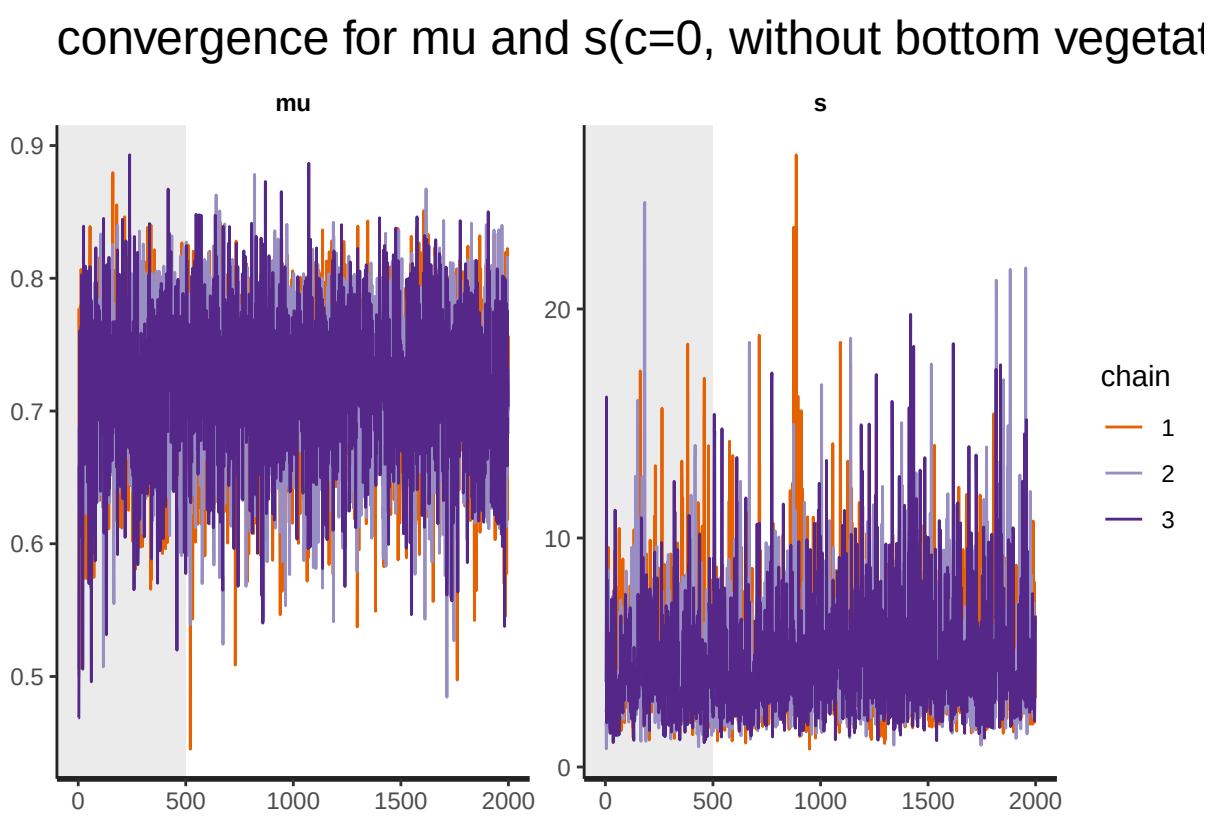
```
# N_initial <- append(N_initial, round(mean(N.noveg)), 6)
data = list('i'=19,'y'=y_initial, 'N'=N_initial)

# initialize parameters
init1 = list(theta = rep(0.5,19), mu=0.1, s=1,theta_20=0.5)
init2 = list(theta = rep(0.5,19), mu=0.5, s=2,theta_20=0.5)
init3 = list(theta = rep(0.5,19), mu=0.7, s=3,theta_20=0.5)
inits <- list(init1, init2, init3)

# Fit a model defined in the Stan modeling language and return the fitted result as an instance of stan

post_noveg=stan(model_code=model,data=data,warmup=500,iter=2000,chains=3,thin=1,
          init=inits,control = list(adapt_delta = 0.8,max_treedepth = 15))
```

```
# set data into named list
y_initial <- y.veg[,2]
y_initial <- append(y_initial, round(mean(y_initial)), 6)
N_initial <- N.veg
N_initial <- append(N_initial, round(mean(N_initial)), 6)

data = list('i'=19,'y'=y_initial, 'N'=N_initial)

# initialize parameters
init1 = list(theta = rep(0.5,19), mu=0.1, s=1,theta_20=0.5)
init2 = list(theta = rep(0.5,19), mu=0.5, s=2,theta_20=0.5)
init3 = list(theta = rep(0.5,19), mu=0.7, s=3,theta_20=0.5)
inits <- list(init1, init2, init3)

post_veg=stan(model_code=model,data=data,warmup=500,iter=2000,chains=3,thin=1,
          init=inits,control = list(adapt_delta = 0.8,max_treedepth = 15))
```

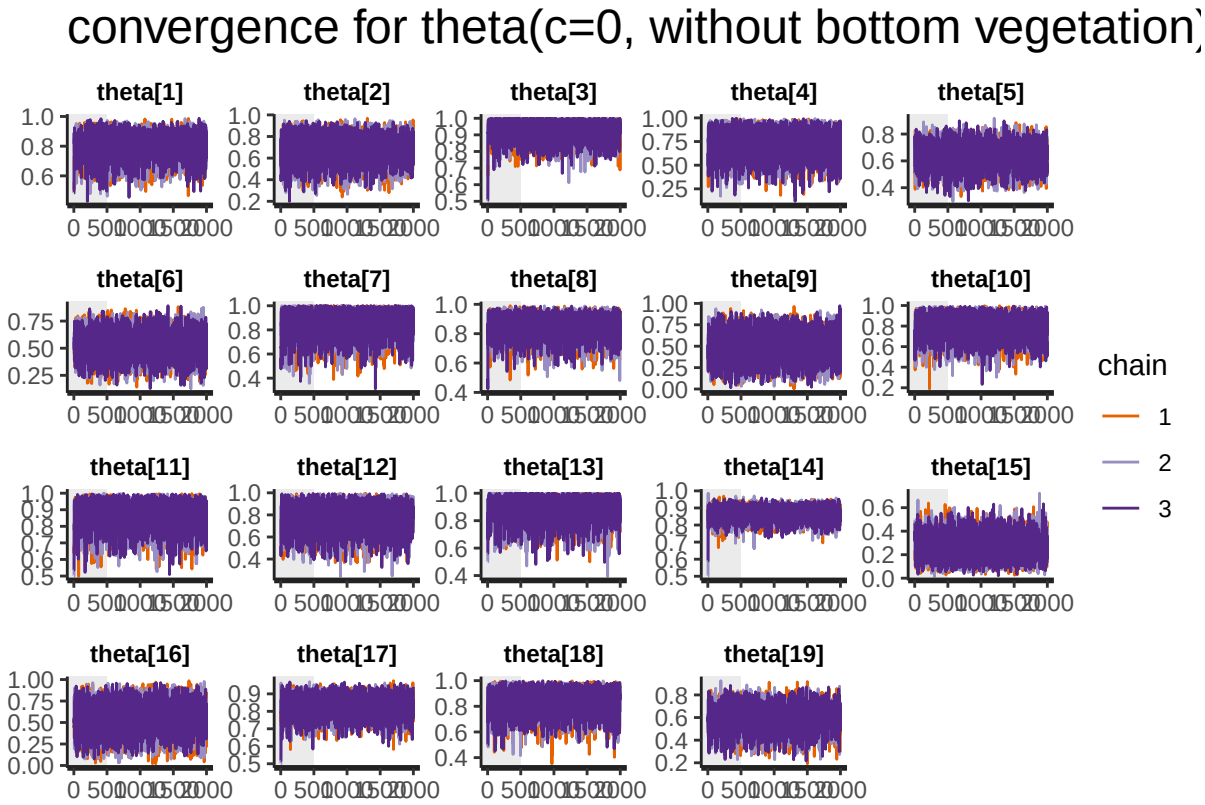
1-answer

1-1 Check for convergence for all parameters (c=0, without bottom vegetation)

```
#check the convergence visually, plot the sample chains
plot(post_noveg, pars =c("mu","s"),plotfun= "trace", inc_warmup = TRUE)+ggtitle("convergence for mu and
```

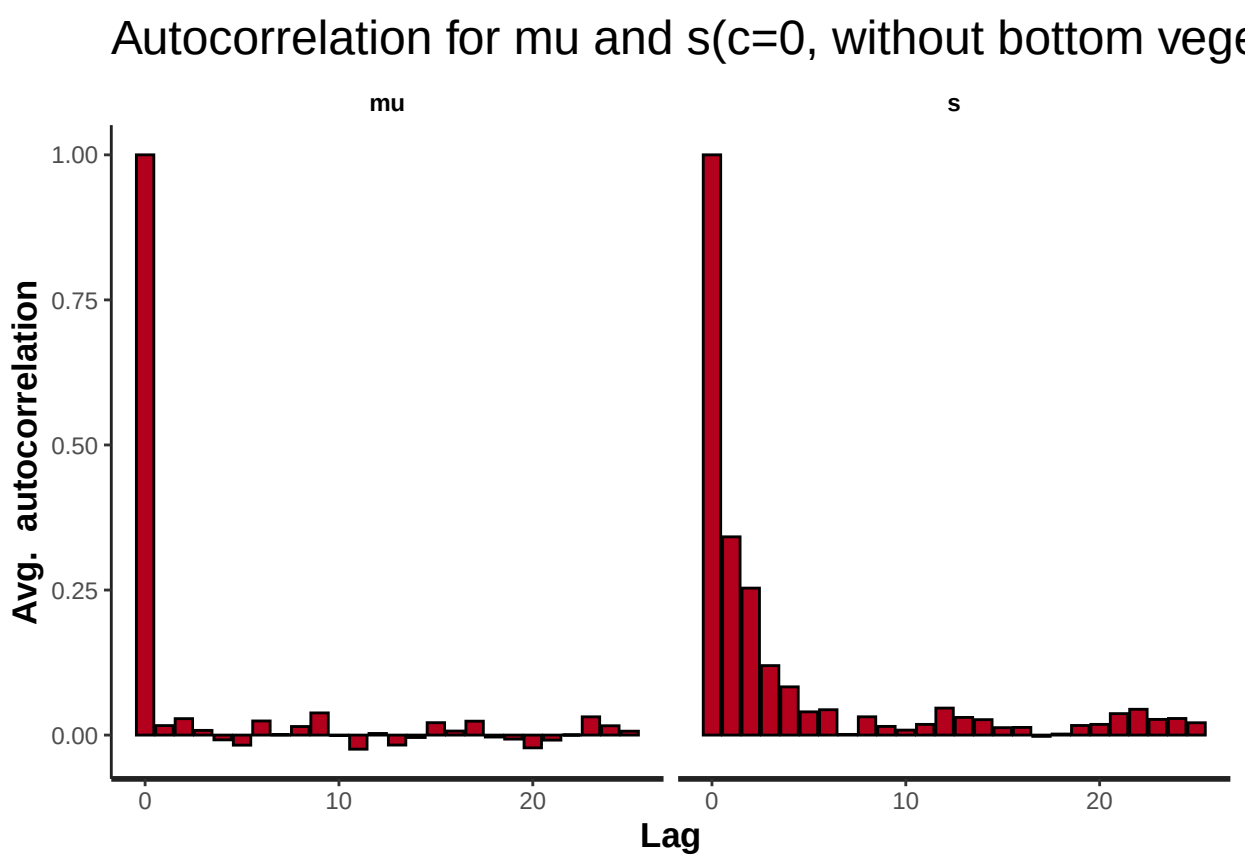



```
plot(post_noveg, pars ="theta",plotfun= "trace", inc_warmup = TRUE)+ggtitle("convergence for theta(c=0,
```

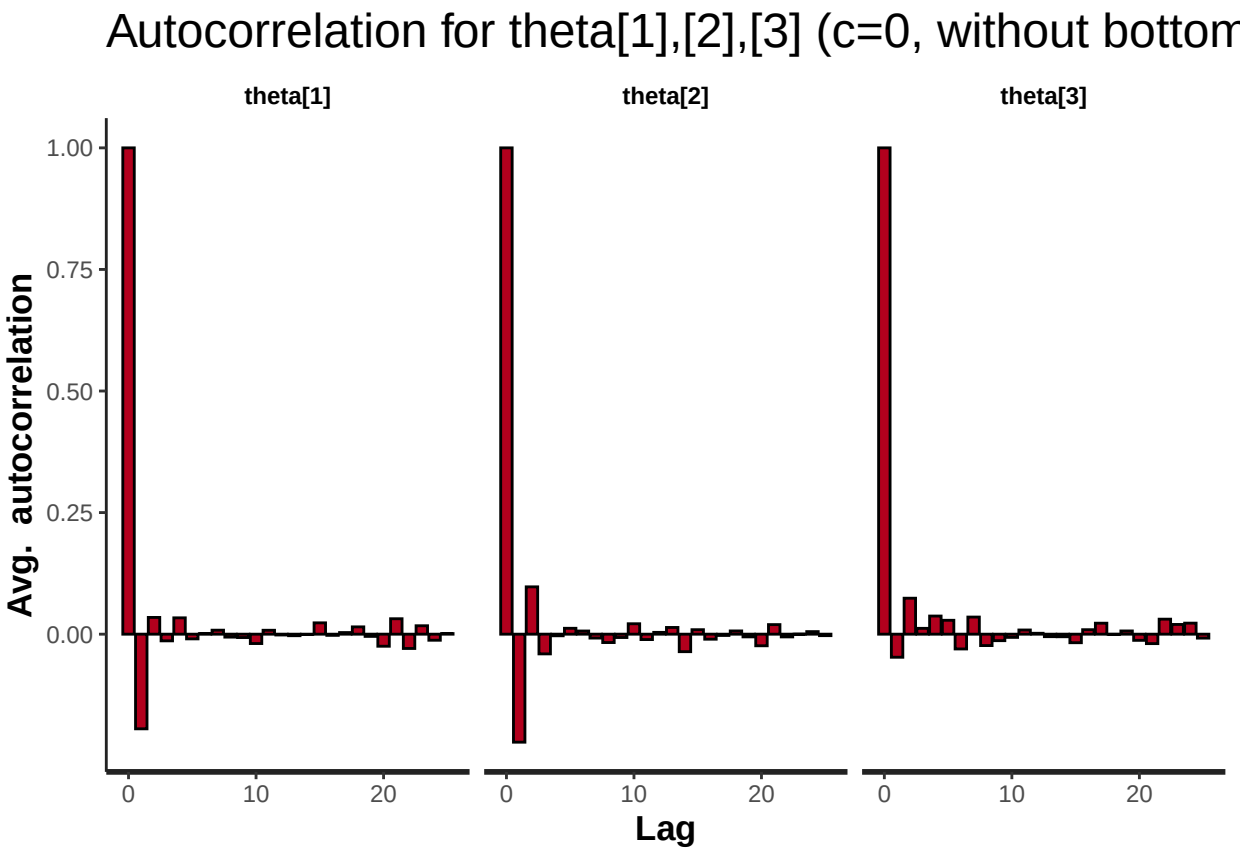


1-2 examine what is the autocorrelation for s_c, μ_c and few $\theta_{i,c}$.(c=0, without bottom vegetation)

```
# Check for convergence for all parameters
par(mfrow=c(2,2))
stan_ac(post_noveg,c("mu","s"),inc_warmup = FALSE, lags = 25)+ggtitle("Autocorrelation for mu and s(c=0
```



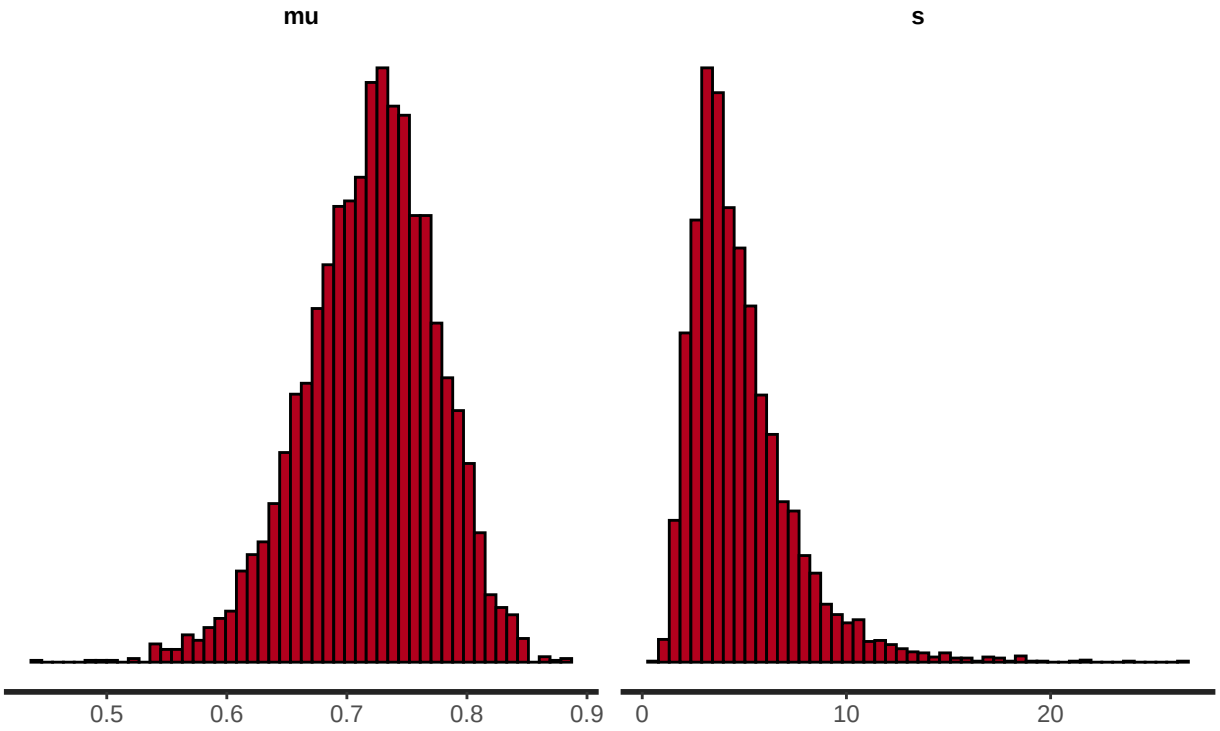

```
# Few $\theta_{i,c}$. Few=2 here.:)
stan_ac(post_noveg,c("theta[1]","theta[2]","theta[3]"),inc_warmup = FALSE, lags = 25)+ggtitle("Autocorr
```



1-3 Visualize the posterior for μ_c, s_c and $\theta_{i,c}, i = 1, \ldots, 19$.(c=0, without bottom vegetation)

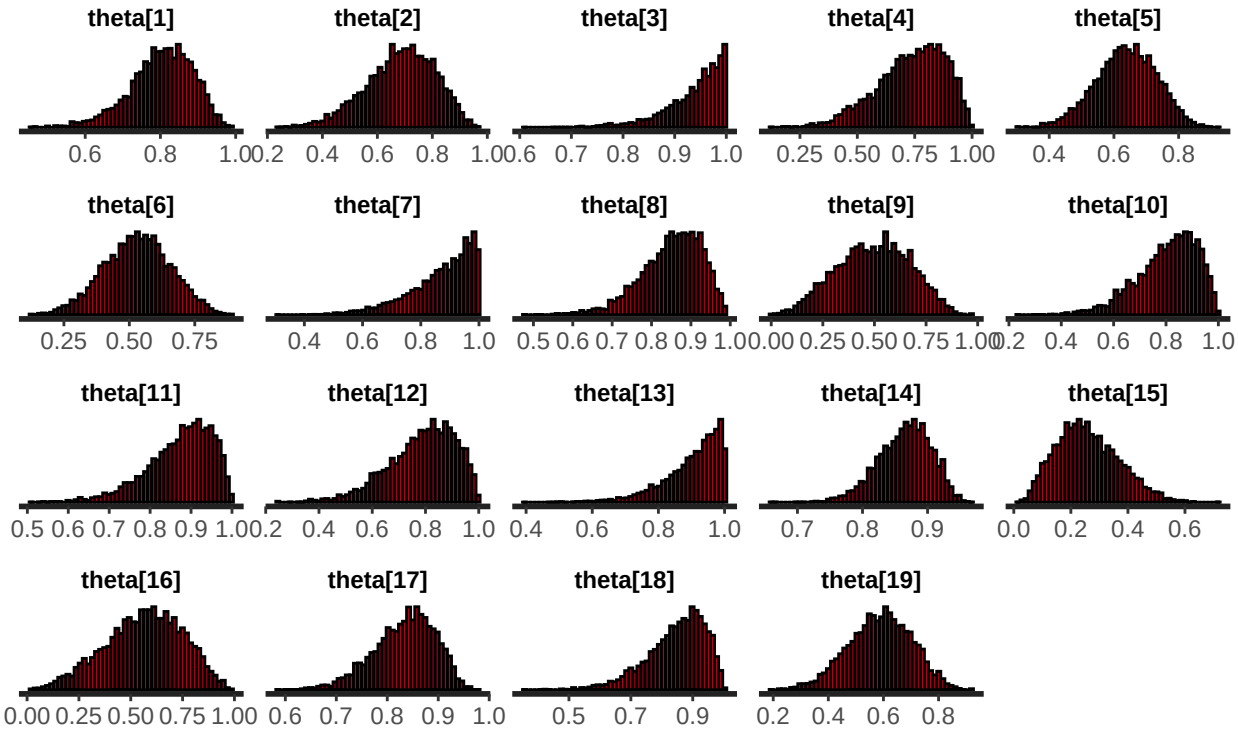
```
plot(post_noveg, plotfun = "hist", pars = c("mu","s"),bins=50)+ggtitle("Visualize the mu and s(c=0, wit
```

Visualize the mu and s(c=0, without bottom vegetation)



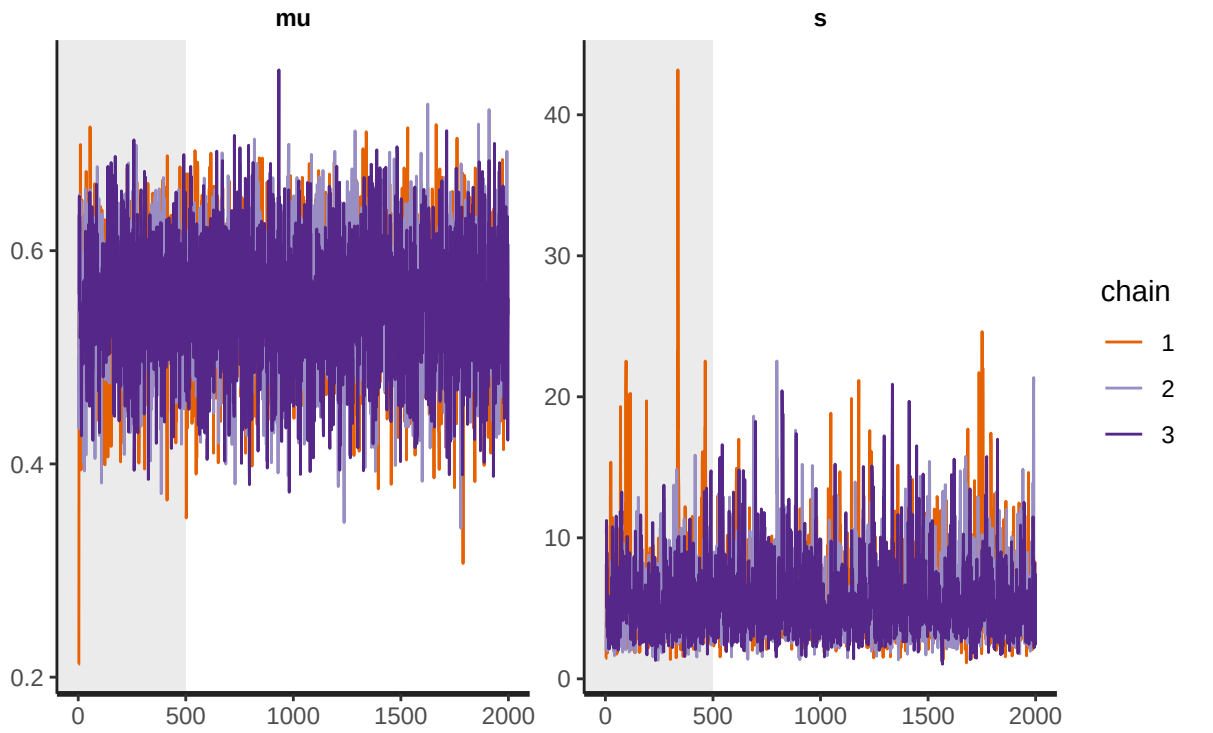

```
plot(post_noveg, plotfun = "hist", pars = "theta",bins=50)+ggtitle("Visualize the theta (c=0, without b
```

Visualize the theta (c=0, without bottom vegetation)

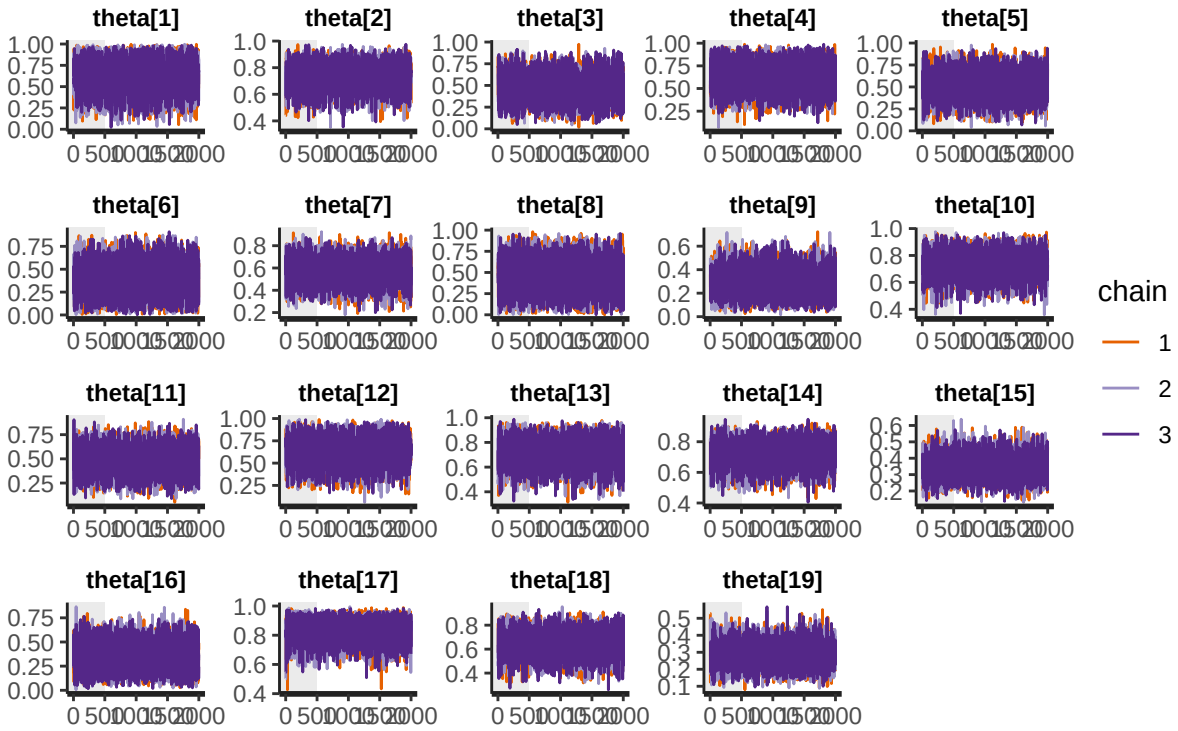


1-1 Check for convergence for all parameters (c=1, with bottom vegetation)

```
plot(post_veg, pars =c("mu","s"),plotfun= "trace", inc_warmup = TRUE)+ggtitle("convergence for mu and s
```

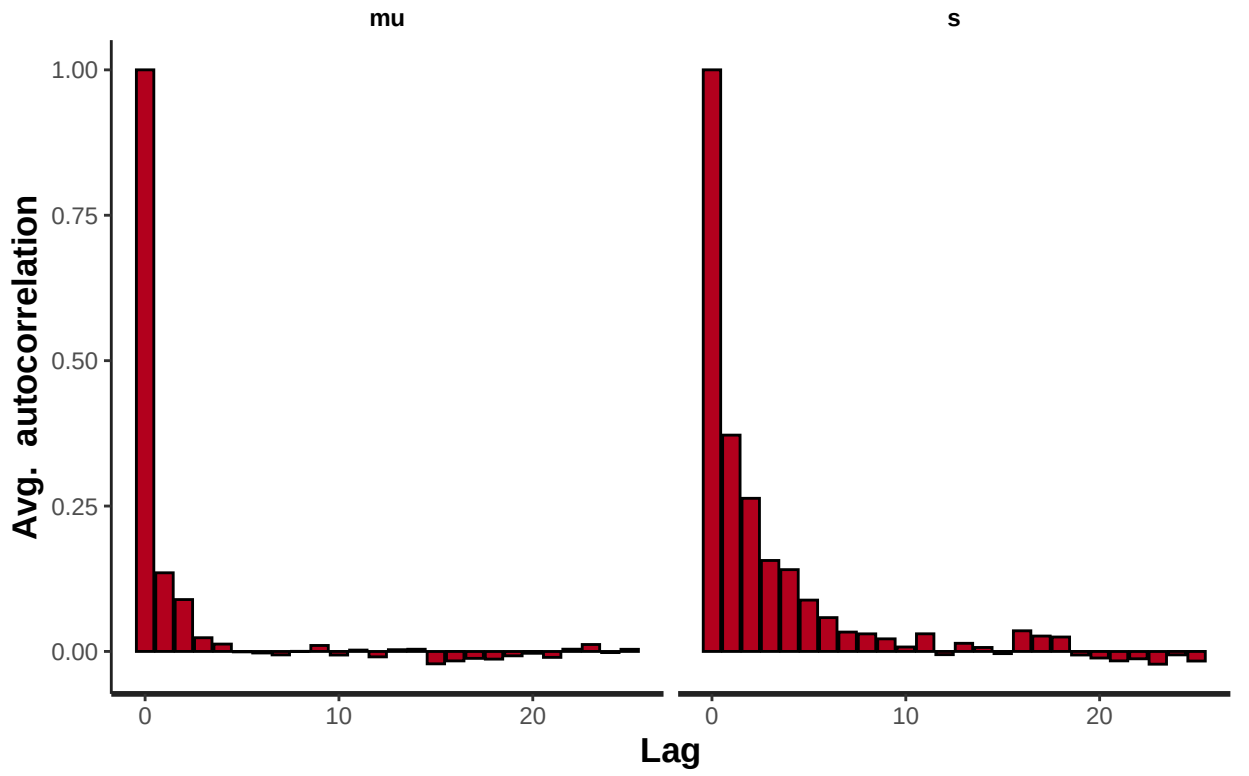



```
plot(post_veg, pars ="theta",plotfun= "trace", inc_warmup = TRUE)+ggtitle("convergence for theta(c=1, w
```

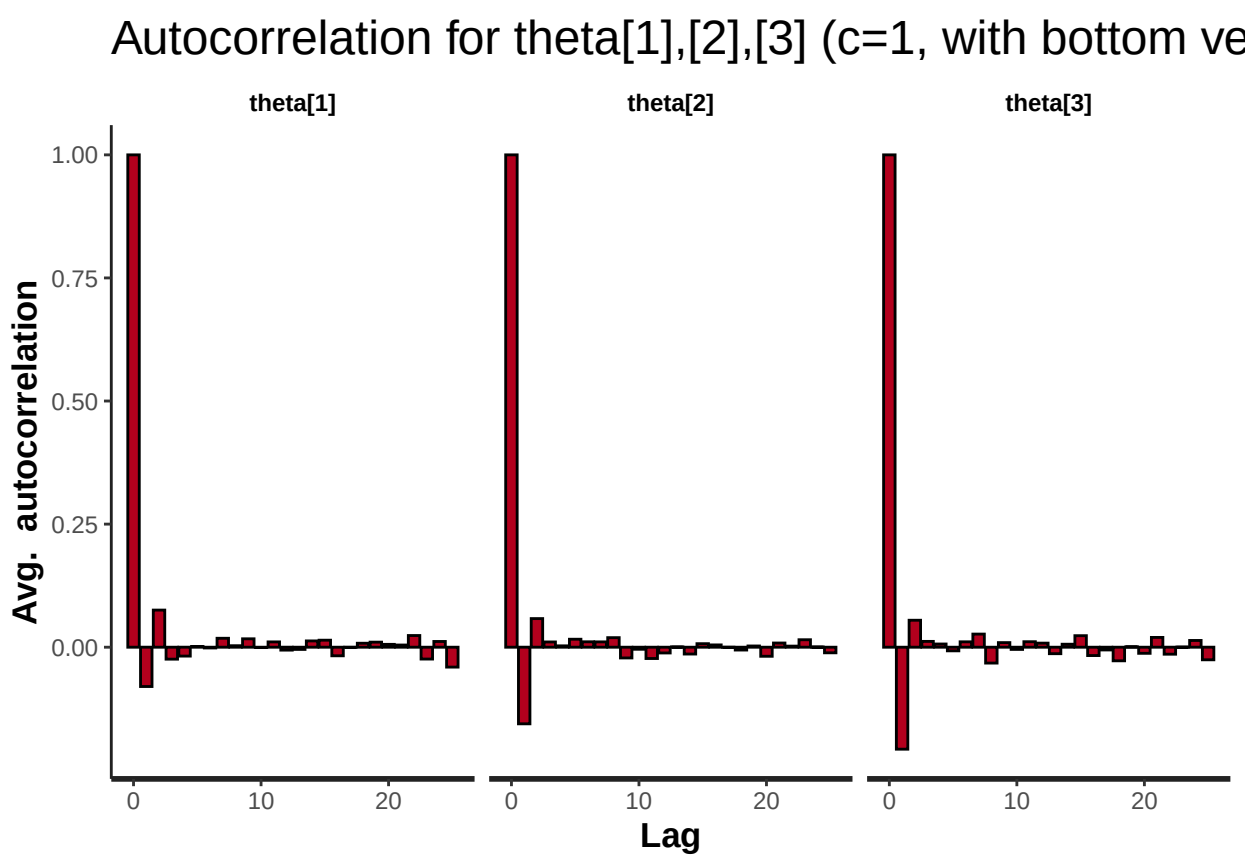


1-2 examine what is the autocorrelation for s_c, μ_c and few $\theta_{i,c}$.(c=1, with bottom vegetation)

```
stan_ac(post_veg,c("mu","s"),inc_warmup = FALSE, lags = 25)+ggtitle("Autocorrelation for mu and s(c=1, v
```



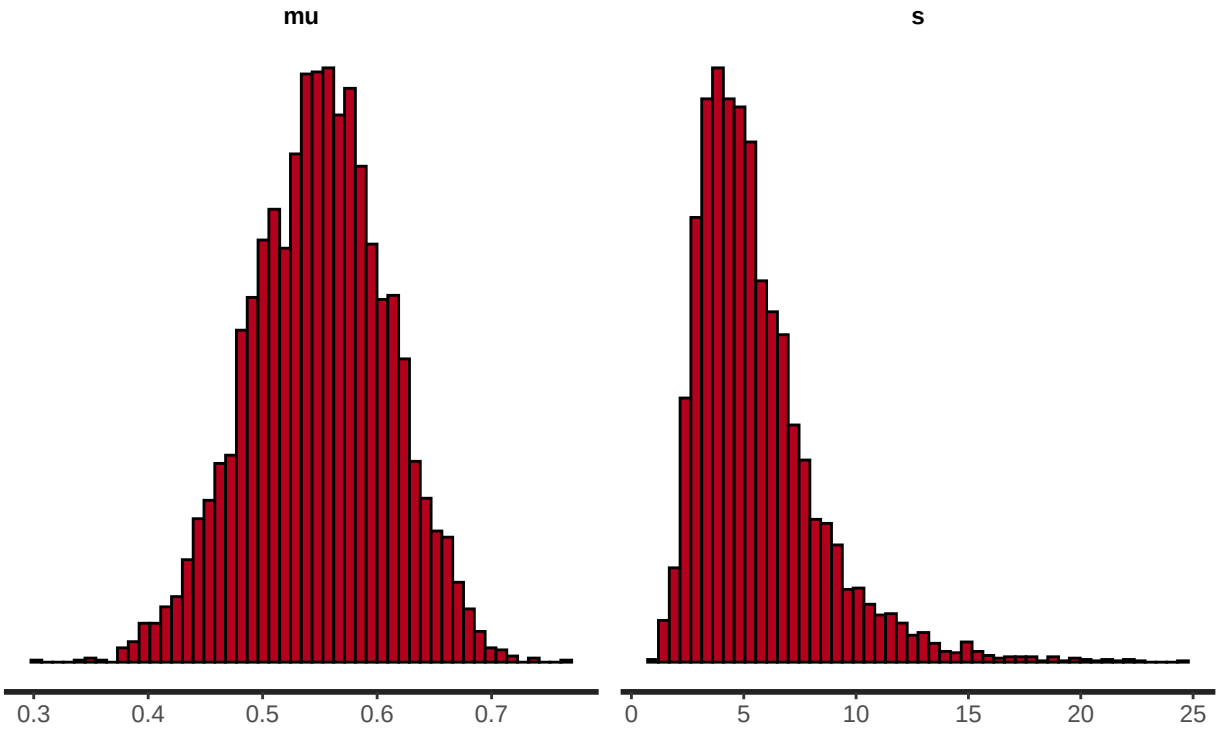

```
# Few $\theta_{i,c}$. Few=3 here.:)
stan_ac(post_veg,c("theta[1]","theta[2]","theta[3]"),inc_warmup = FALSE, lags = 25)+ggtitle("Autocorrela
```



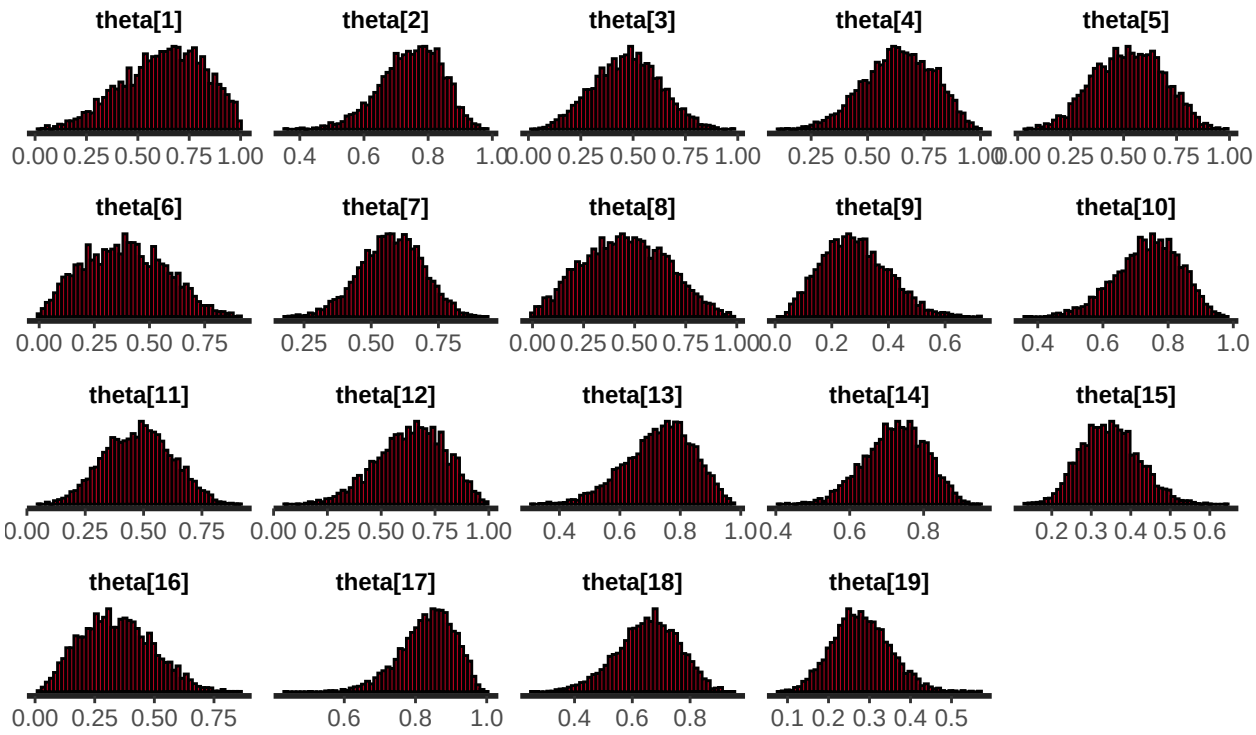
1-3 Visualize the posterior for μ_c, s_c and $\theta_{i,c}, i = 1, \ldots, 19$.(c=1, with bottom vegetation)

```
par(mfrow=c(2,2))
plot(post_veg, plotfun = "hist", pars = c("mu","s"),bins=50)+ggtitle("Visualize the mu and s(c=1, with
```

Visualize the mu and s(c=1, with bottom vegetation)



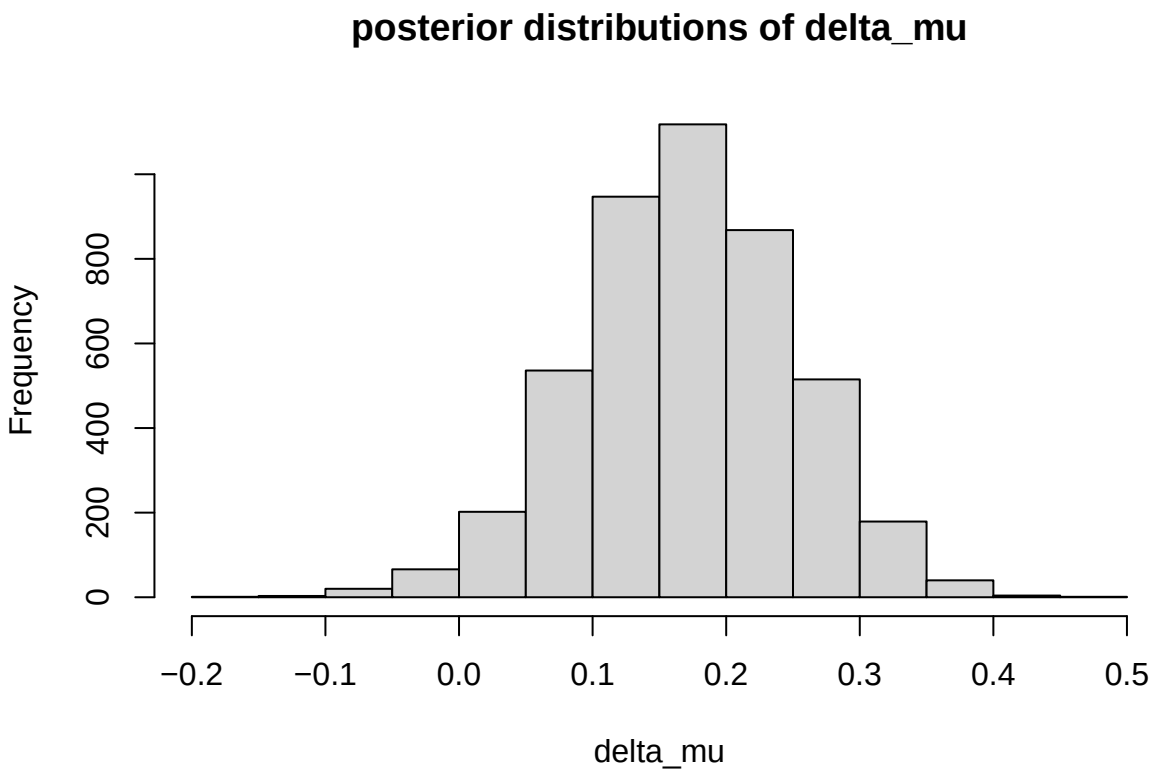

```
plot(post_veg, plotfun = "hist", pars = "theta",bins=50)+ggtitle("Visualize the theta (c=1, with bottom
```



2-answer

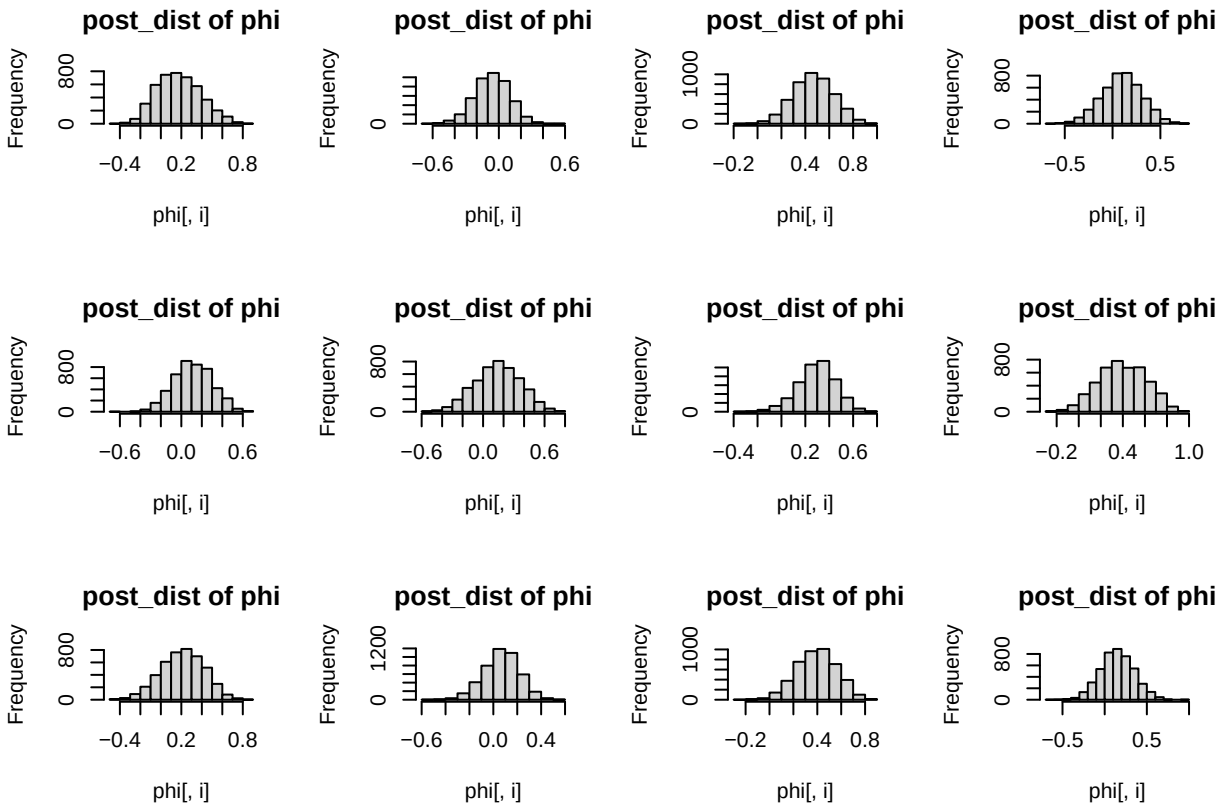
2-1 Visualize the posterior distributions of $\Delta\mu = \mu_0 - \mu_1$

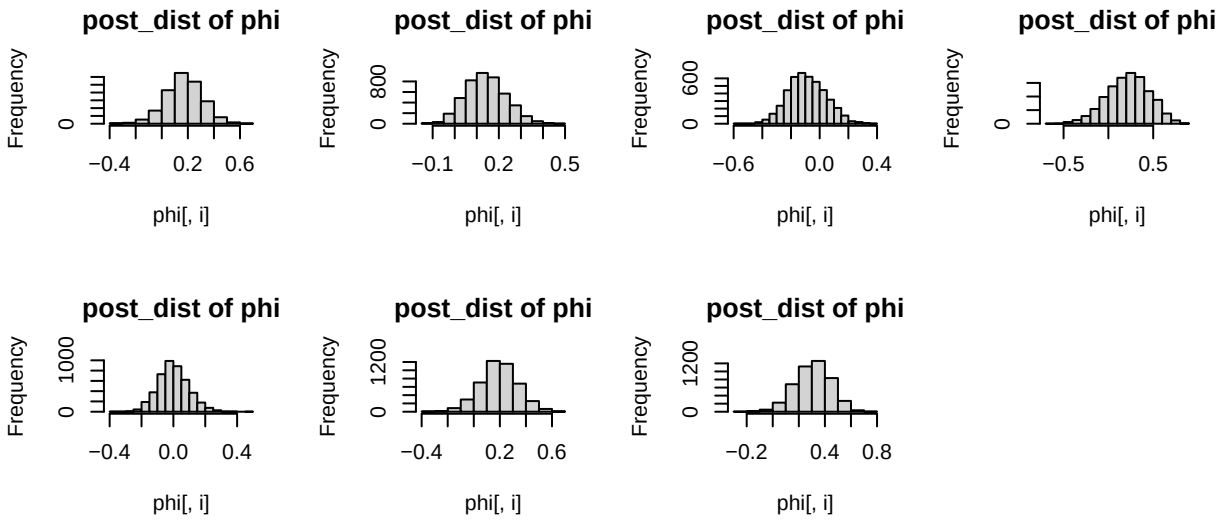
```
#2-1 Visualize the posterior distributions of delta_mu
samples_noveg <- as.matrix(post_noveg, pars =c("mu","s","theta","ytilde_19","ytilde_20"))
samples_veg <- as.matrix(post_veg, pars =c("mu","s","theta","ytilde_19","ytilde_20"))
delta_mu <- samples_noveg[,"mu"]-samples_veg[,"mu"]
hist(delta_mu,main = "posterior distributions of delta_mu")
```



2-2 Visualize the posterior distributions of $\phi_i = \theta_{i,0} - \theta_{i,1}$ for each area $i = 1, \ldots, 19$.

```
#2-2 Visualize the posterior distributions of $\phi_i = \theta_{i,0}-\theta_{i,1}$ for each area $i=1,\
phi <- matrix(nrow=4500,ncol=19)
# To save space I just plot 12 of Phi each graph, par(mfrow=c(4,5)) dose't work when I knit to PDF.
par(mfrow=c(3,4))
for (i in 1:19){
  phi[,i]<- samples_noveg[,i+2]-samples_veg[,i+2]
}
for (i in 1:19){
  hist(phi[,i],main = "post_dist of phi")
  }
```

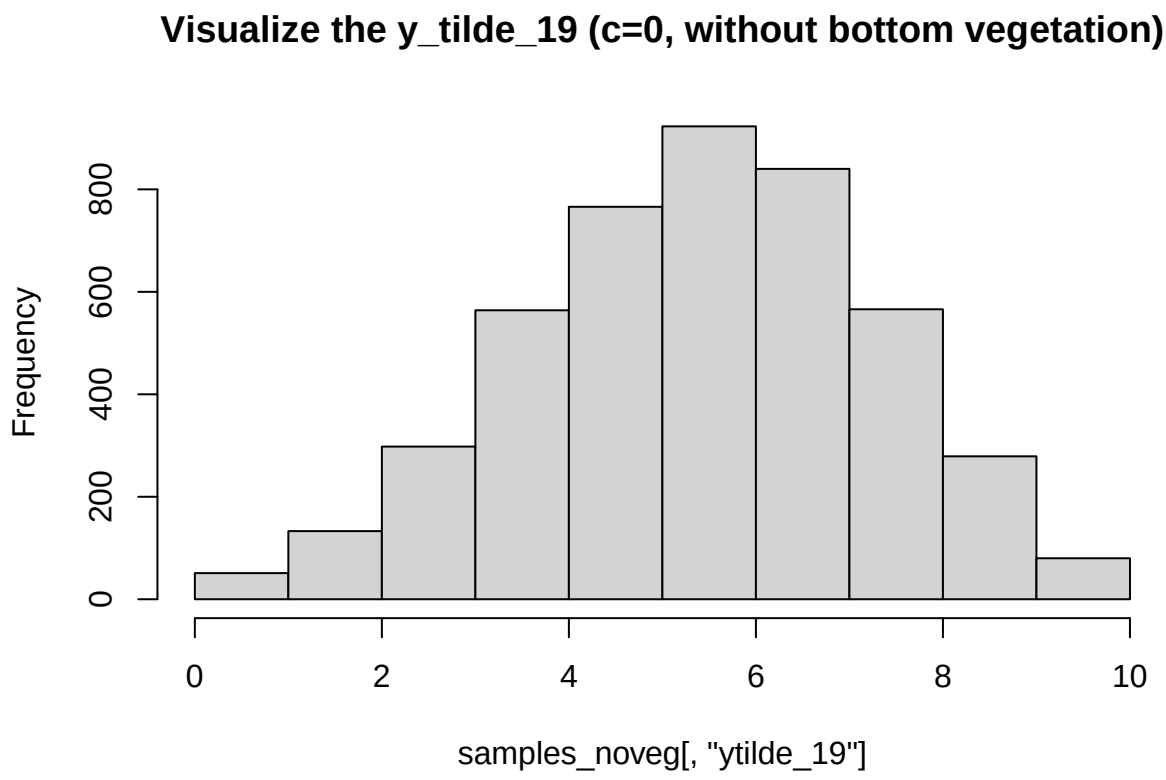

post_dist of phi
Frequency
0
800
−0.4
0.2
0.8
phi[, i]
post_dist of phi
Frequency
0
−0.6
0.0
0.6
phi[, i]
post_dist of phi
Frequency
0
1000
−0.2
0.4
0.8
phi[, i]
post_dist of phi
Frequency
0
800
−0.5
0.5
phi[, i]
post_dist of phi
Frequency
0
800
−0.6
0.0
0.6
phi[, i]
post_dist of phi
Frequency
0
800
−0.6
0.0
0.6
phi[, i]
post_dist of phi
Frequency
0
−0.4
0.2
0.6
phi[, i]
post_dist of phi
Frequency
0
800
−0.2
0.4
1.0
phi[, i]
post_dist of phi
Frequency
0
800
−0.4
0.2
0.8
phi[, i]
post_dist of phi
Frequency
0
1200
−0.6
0.0
0.4
phi[, i]
post_dist of phi
Frequency
0
1000
−0.2
0.4
0.8
phi[, i]
post_dist of phi
Frequency
0
800
−0.5
0.5
phi[, i]



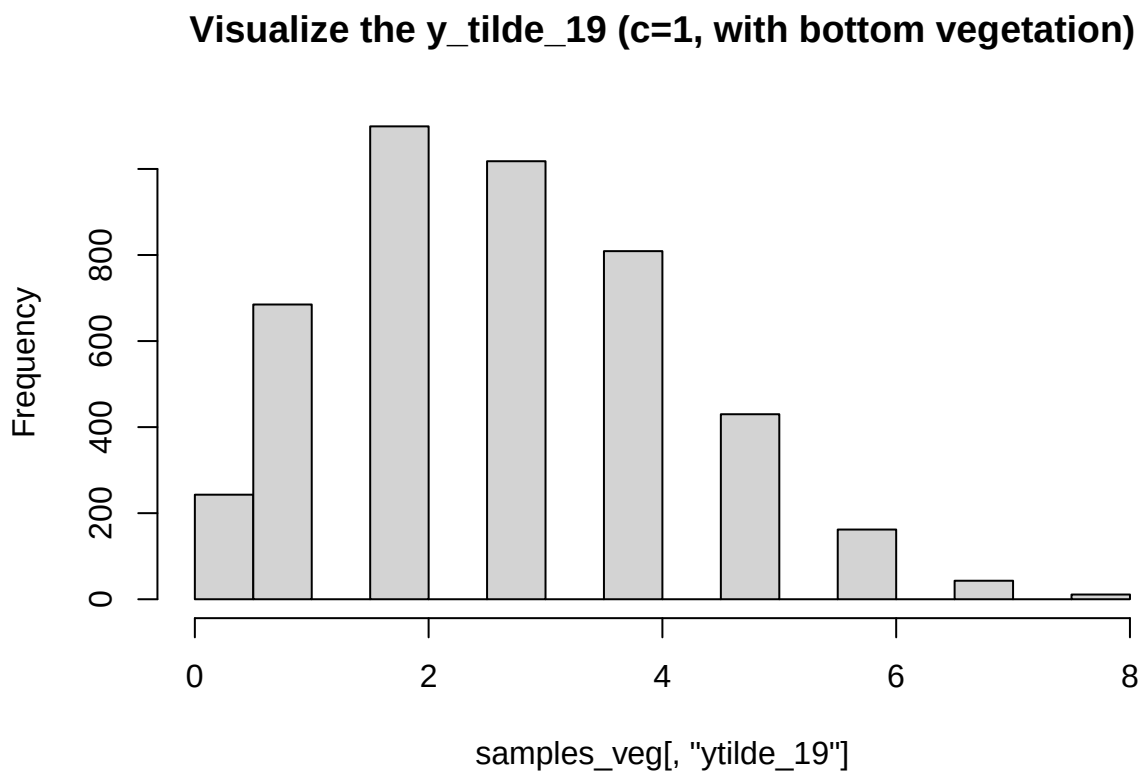
3-answer

Visualize the resulting posterior samples as well as the posterior distribution

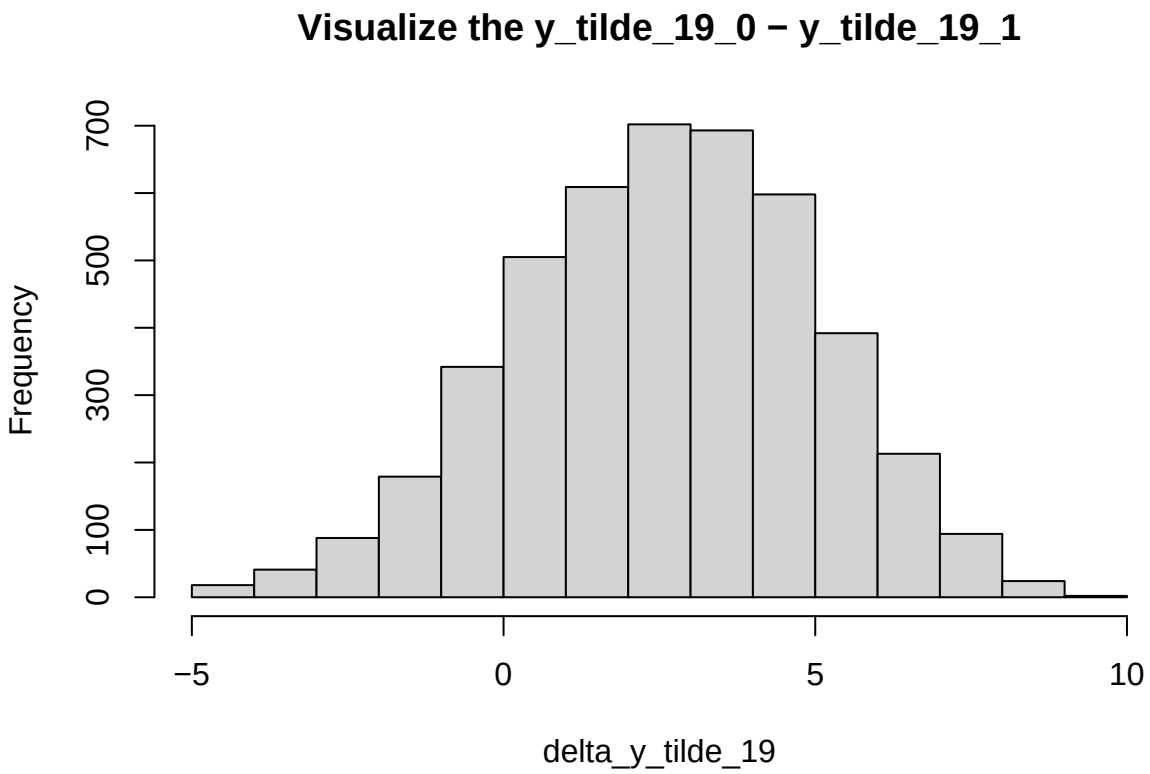
```
hist(samples_noveg[,"ytilde_19"],main = "Visualize the y_tilde_19 (c=0, without bottom vegetation)")
```




```
hist(samples_veg[,"ytilde_19"],main = "Visualize the y_tilde_19 (c=1, with bottom vegetation)")
```

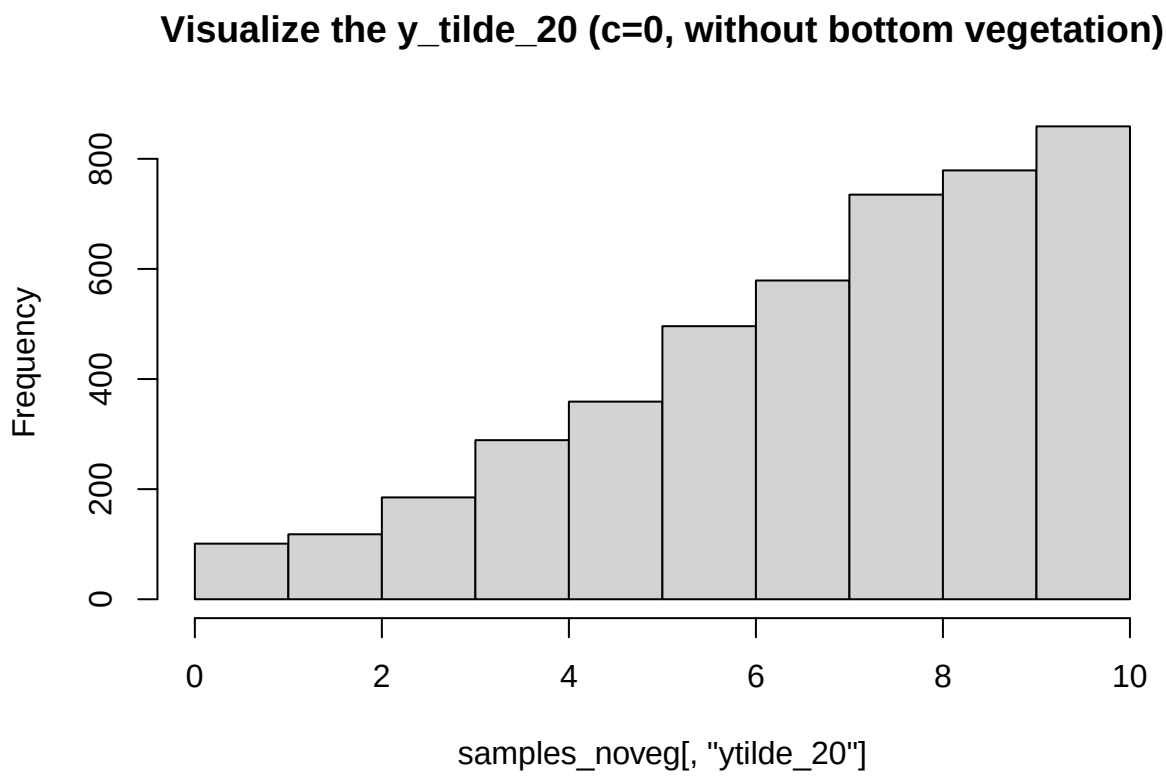



```r
delta_y_tilde_19 <- samples_noveg[,"ytilde_19"]-samples_veg[,"ytilde_19"]
hist(delta_y_tilde_19,main = "Visualize the y_tilde_19_0 - y_tilde_19_1")
```

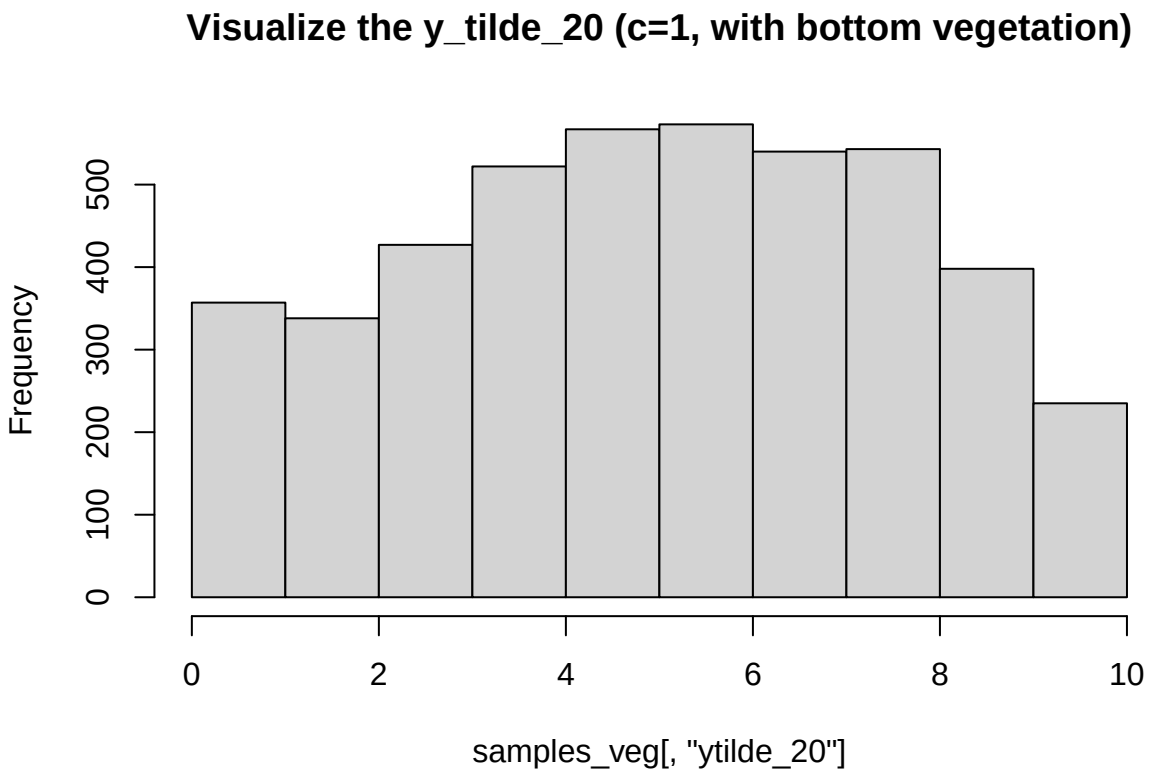


4-answer Visualize the resulting posterior samples as well as the posterior distribution

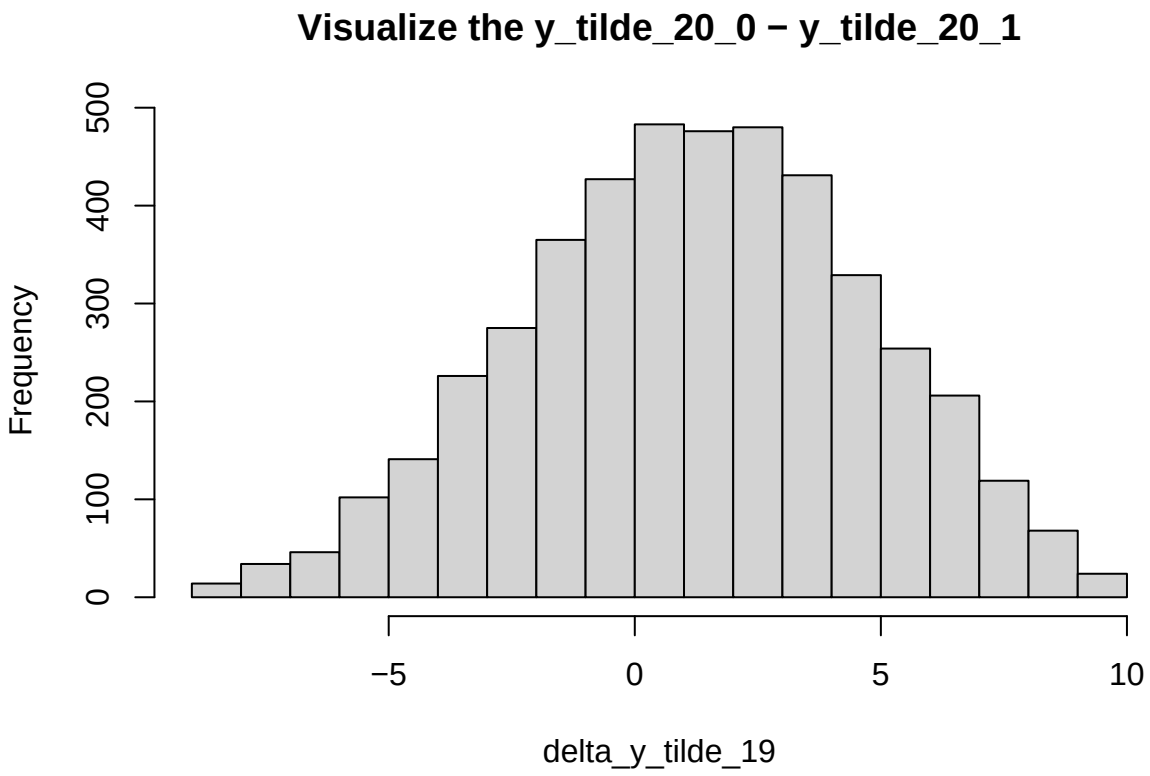
```
#Visualize the resulting posterior samples as well as the posterior distribution

hist(samples_noveg[,"ytilde_20"],main = "Visualize the y_tilde_20 (c=0, without bottom vegetation)")
```




```
hist(samples_veg[,"ytilde_20"],main = "Visualize the y_tilde_20 (c=1, with bottom vegetation)")
```




```
delta_y_tilde_19 <- samples_noveg[,"ytilde_20"]-samples_veg[,"ytilde_20"]
hist(delta_y_tilde_19,main = "Visualize the y_tilde_20_0 - y_tilde_20_1")
```



##5-answer

Pooled estimate of theta θ_c is more accurate and reliable than population meanμ_c because it is based on certain specific values. But both are not as good as individual theta $\theta_{i,c}$ in the hierarchical model. Because this subdivision of regions is more accurate, it is undoubtedly the most justified statistical method. If we can classify the background (variable) we want to observe more carefully, we can reduce random errors, so it is a better and more justified model.

Grading

Total 20 points Each of the steps provides 4 points from correct answer and 2 points from an answer that is towards the right direction but includes minor mistake (e.g. a bug or typo)

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/