Assignment\_4\_Q1

Cheng

5/1/2019

# Question1

## (a)

# load the data  
data('rugged', package = 'rethinking')  
d <- rugged; rm(rugged)  
  
# extract countries with gdp data  
d <- d[complete.cases(d$rgdppc\_2000), ] # comlete.cases() ==> drop na row  
  
# make log version of outcome  
d$loggdp <- log(d$rgdppc\_2000)  
  
# create another dataframe without variable Seychelles  
d2 <- d[d$country != "Seychelles", ]

# Rstan setup  
ma = "  
data {  
int N;  
vector[N] loggdp;  
vector[N] rugged;  
vector<lower=0,upper=1>[N] cont\_africa;  
}  
parameters {  
real a;  
real bR;  
real bA;  
real bAR;  
real<lower=0, upper=10> sigma;  
}  
model {  
//transfer variable  
vector[N] mu;  
for (i in 1:N) {  
mu[i] = a + bR \* rugged[i] + bAR \* rugged[i] \* cont\_africa[i] + bA \* cont\_africa[i];  
}  
//prior  
a ~ normal(8,100);  
bR ~ normal(0,1);  
bA ~ normal(0,1);  
bAR ~ normal(0,1);  
  
//likelihood  
loggdp ~ normal(mu,sigma);   
}  
generated quantities {  
vector[N] log\_lik;  
{  
 vector[N] mu; vector[N] gamma;  
 for(n in 1:N) {  
 gamma[n] = bR + bAR \* cont\_africa[n];  
 mu[n] = a + gamma[n] \* rugged[n] + bA \* cont\_africa[n];  
 log\_lik[n] = normal\_lpdf(loggdp[n] | mu[n], sigma);  
 }  
}  
}  
"  
  
# fitm1: include Seychelles  
dat <- list(  
 N = NROW(d),  
 loggdp = d$loggdp,  
 rugged = d$rugged,  
 cont\_africa = d$cont\_africa  
)  
  
fitm1 = stan(model\_code = ma,   
 data = dat,   
 cores = 4,   
 chains = 4,   
 iter = 3000)  
  
# fitm2: exclude Seychelles  
dat <- list(  
 N = NROW(d2),  
 loggdp = d2$loggdp,  
 rugged = d2$rugged,  
 cont\_africa = d2$cont\_africa  
)  
  
fitm2 = stan(model\_code = ma,   
 data = dat,   
 cores = 4,   
 chains = 4,   
 iter = 3000)

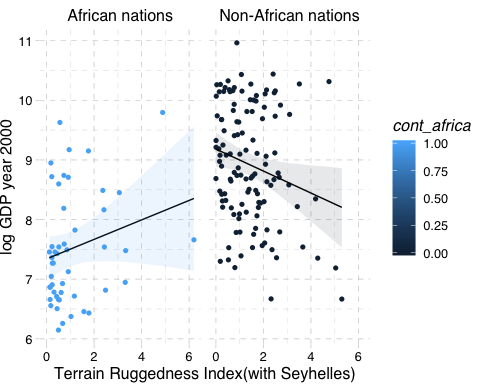
# Comparing estimates  
my\_coef\_tab <-  
 rbind(tidy(fitm1), tidy(fitm2)) %>%  
 mutate(model = c(rep("fitm1", times = nrow(tidy(fitm1))),  
 rep("fitm2", times = nrow(tidy(fitm2))))) %>%  
 filter(term != "lp\_\_") %>%  
 select(model, everything())  
  
coef <-  
 my\_coef\_tab %>%  
 complete(term = distinct(., term), model) %>%  
 select(model, term, estimate) %>%  
 mutate(estimate = round(estimate, digits = 2)) %>%  
 spread(key = model, value = estimate)  
  
head(coef, n=4)

## # A tibble: 4 x 3  
## term fitm1 fitm2  
## <chr> <dbl> <dbl>  
## 1 a 9.18 9.19  
## 2 bA -1.84 -1.79  
## 3 bAR 0.35 0.25  
## 4 bR -0.18 -0.19

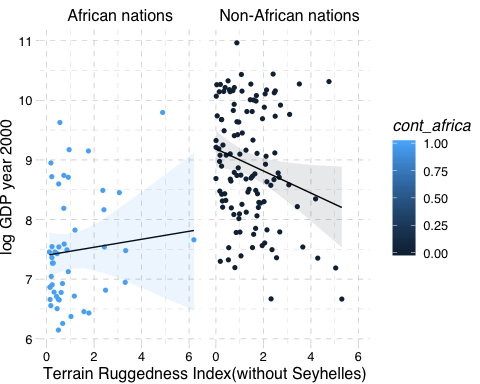
From the table, we can see that the coefficient of interaction variable dropped from 0.34 to 0.25 which seems like the effect of ruggedness upon continent decreased.

## (b) plotting the interaction

# with Seyhelles  
post1 <- as.data.frame(fitm1)  
f\_mu\_1 <- function(rugged, cont\_africa) with(post1,   
 a + bR \* rugged + bAR \* rugged \* cont\_africa + bA \* cont\_africa )  
mu\_1 <- mapply(f\_mu\_1, rugged = d$rugged, cont\_africa = d$cont\_africa)  
mu\_1\_mean <- apply(mu\_1, 2, mean)  
mu\_1\_pi <- apply(mu\_1, 2, rethinking::PI, prob = .97)  
d\_1 <- d %>%  
 mutate(mu\_mean = mu\_1\_mean,  
 mu\_pi\_l = mu\_1\_pi[1,],  
 mu\_pi\_h = mu\_1\_pi[2,],  
 inAfrica = ifelse(cont\_africa, 'African nations', 'Non-African nations'))  
  
  
  
d\_1 %>%  
 ggplot(aes(x = rugged)) +  
 geom\_point(aes(rugged, loggdp, color = cont\_africa), shape = 16) +  
 theme(legend.position = '') +  
 geom\_line(aes(rugged, mu\_mean)) +  
 geom\_ribbon(aes(x=rugged,   
 ymin=mu\_pi\_l,   
 ymax=mu\_pi\_h,   
 fill = cont\_africa),   
 alpha = .1) +  
 facet\_wrap(~inAfrica) +  
 theme\_pander() +  
 labs(x = 'Terrain Ruggedness Index(with Seyhelles)', y = 'log GDP year 2000')



# without Seyhelles  
post2 <- as.data.frame(fitm2)  
f\_mu\_2 <- function(rugged, cont\_africa) with(post2,   
 a + bR \* rugged + bAR \* rugged \* cont\_africa + bA \* cont\_africa )  
mu\_2 <- mapply(f\_mu\_2, rugged = d$rugged, cont\_africa = d$cont\_africa)  
mu\_2\_mean <- apply(mu\_2, 2, mean)  
mu\_2\_pi <- apply(mu\_2, 2, rethinking::PI, prob = .97)  
d\_2 <- d %>%  
 mutate(mu\_mean = mu\_2\_mean,  
 mu\_pi\_l = mu\_2\_pi[1,],  
 mu\_pi\_h = mu\_2\_pi[2,],  
 inAfrica = ifelse(cont\_africa, 'African nations', 'Non-African nations'))  
  
  
  
d\_2 %>%  
 ggplot(aes(x = rugged)) +  
 geom\_point(aes(rugged, loggdp, color = cont\_africa), shape = 16) +  
 theme(legend.position = '') +  
 geom\_line(aes(rugged, mu\_mean)) +  
 geom\_ribbon(aes(x=rugged,   
 ymin=mu\_pi\_l,   
 ymax=mu\_pi\_h,   
 fill = cont\_africa),   
 alpha = .1) +  
 facet\_wrap(~inAfrica) +  
 theme\_pander() +  
 labs(x = 'Terrain Ruggedness Index(without Seyhelles)', y = 'log GDP year 2000')



If we drop the data of Seyhelles, the slope became slower in African nations which imply the effect of interaction variable is weaker than expected.

## (c) model comparison

# Rstan setup  
  
mc\_1="  
data {  
int N;  
vector[N] loggdp;  
vector[N] rugged;  
}  
parameters {  
real a;  
real bR;  
real<lower=0, upper=10> sigma;  
}  
model {  
vector[N] mu = a + bR \* rugged;  
//prior  
a ~ normal(8,100);  
bR ~ normal(0,1);  
//likelihood  
loggdp ~ normal(mu,sigma);   
}  
generated quantities {  
vector[N] log\_lik;  
{  
 vector[N] mu;  
 for(n in 1:N) {  
 mu[n] = a + bR \* rugged[n];  
 log\_lik[n] = normal\_lpdf(loggdp[n] | mu[n], sigma);  
 }  
}  
}  
"  
  
dat <- list(  
 N = NROW(d2),  
 loggdp = d2$loggdp,  
 rugged = d2$rugged  
)  
fitmc\_1 = stan(model\_code = mc\_1,   
 data = dat,   
 cores = 4,   
 chains = 4,   
 iter = 3000)

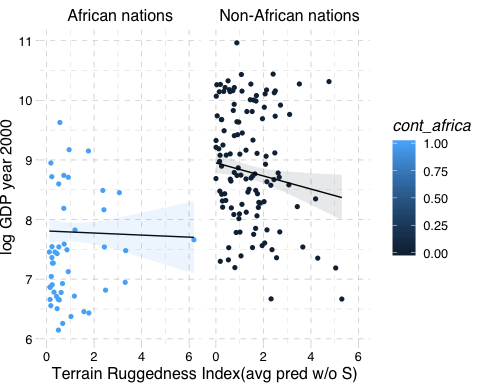
mc\_2 ="  
data {  
int N;  
vector[N] loggdp;  
vector[N] rugged;  
vector<lower=0,upper=1>[N] cont\_africa;  
}  
  
parameters {  
real a;  
real bR;  
real bA;  
real<lower=0, upper=10> sigma;  
}  
  
model {  
vector[N] mu = a + bR \* rugged + bA \* cont\_africa;  
//prior  
a ~ normal(8,100);  
bR ~ normal(8,100);  
bA ~ normal(8,100);  
  
//likelihood  
loggdp ~ normal(mu,sigma);   
}  
  
generated quantities {  
vector[N] log\_lik;  
{  
 vector[N] mu;  
 for(n in 1:N) {  
 mu[n] = a + bR \* rugged[n] + bA \* cont\_africa[n];  
 log\_lik[n] = normal\_lpdf(loggdp[n] | mu[n], sigma);  
 }  
}  
}  
"  
  
dat <- list(  
 N = NROW(d2),  
 loggdp = d2$loggdp,  
 rugged = d2$rugged,  
 cont\_africa = d2$cont\_africa  
)  
fitmc\_2 = stan(model\_code = mc\_2,   
 data = dat,  
 cores = 4,   
 chains = 4,   
 iter = 3000)

mc\_3 = "  
data {  
int N;  
vector[N] loggdp;  
vector[N] rugged;  
vector<lower=0,upper=1>[N] cont\_africa;  
}  
parameters {  
real a;  
real bR;  
real bA;  
real bAR;  
real<lower=0, upper=10> sigma;  
}  
model {  
//transfer variable  
vector[N] mu;  
for (i in 1:N) {  
mu[i] = a + bR \* rugged[i] + bAR \* rugged[i] \* cont\_africa[i] + bA \* cont\_africa[i];  
}  
//prior  
a ~ normal(8,100);  
bR ~ normal(0,1);  
bA ~ normal(0,1);  
bAR ~ normal(0,1);  
  
//likelihood  
loggdp ~ normal(mu,sigma);   
}  
generated quantities {  
vector[N] log\_lik;  
{  
 vector[N] mu; vector[N] gamma;  
 for(n in 1:N) {  
 gamma[n] = bR + bAR \* cont\_africa[n];  
 mu[n] = a + gamma[n] \* rugged[n] + bA \* cont\_africa[n];  
 log\_lik[n] = normal\_lpdf(loggdp[n] | mu[n], sigma);  
 }  
}  
}  
"  
  
dat <- list(  
 N = NROW(d2),  
 loggdp = d2$loggdp,  
 rugged = d2$rugged,  
 cont\_africa = d2$cont\_africa  
)  
  
fitmc\_3 = stan(model\_code = mc\_3,   
 data = dat,   
 cores = 4,   
 chains = 4,   
 iter = 3000)

# WAIC comparison  
log\_lik\_mc1 <- extract\_log\_lik(fitmc\_1, merge\_chains = FALSE)  
log\_lik\_mc2 <- extract\_log\_lik(fitmc\_2, merge\_chains = FALSE)  
log\_lik\_mc3 <- extract\_log\_lik(fitmc\_3, merge\_chains = FALSE)  
  
waic1 <- waic(log\_lik\_mc1)  
waic2 <- waic(log\_lik\_mc2)  
waic3 <- waic(log\_lik\_mc3)  
  
mod\_comp <- loo::compare(waic1, waic2, waic3)  
mod\_comp

## elpd\_diff se\_diff elpd\_waic p\_waic waic   
## waic3 0.0 0.0 -231.6 4.4 463.1  
## waic2 -1.5 1.6 -233.1 3.9 466.2  
## waic1 -36.4 7.3 -268.0 2.5 536.0

w <-   
 mod\_comp %>%   
 as\_tibble()  
  
w <-  
 w %>% mutate(ratio = w$waic / sum(w$waic) )  
  
postmc1 <- as.data.frame(fitmc\_1)  
postmc2 <- as.data.frame(fitmc\_2)  
postmc3 <- as.data.frame(fitmc\_3)  
  
mu\_c\_1 <- function(rugged) with(postmc1,   
 a + bR \* rugged)  
mu\_c\_2 <- function(rugged, cont\_africa) with(postmc2,   
 a + bR \* rugged + bA \* cont\_africa )  
mu\_c\_3 <- function(rugged, cont\_africa) with(postmc3,   
 a + bR \* rugged + bAR \* rugged \* cont\_africa + bA \* cont\_africa )  
  
mu\_c\_avg <-   
 w$ratio[1] \* mapply(mu\_c\_1, rugged = d2$rugged) +   
 w$ratio[2] \* mapply(mu\_c\_2, rugged = d2$rugged, cont\_africa = d2$cont\_africa) +  
 w$ratio[3] \* mapply(mu\_c\_3, rugged = d2$rugged, cont\_africa = d2$cont\_africa)  
  
mu\_c\_mean <- apply(mu\_c\_avg, 2, mean) # 2 compute by row  
mu\_c\_pi <- apply(mu\_c\_avg, 2, rethinking::PI, prob = .97)  
d\_c <- d2 %>%  
 mutate(mu\_mean = mu\_c\_mean,  
 mu\_pi\_l = mu\_c\_pi[1,],  
 mu\_pi\_h = mu\_c\_pi[2,],  
 inAfrica = ifelse(cont\_africa, 'African nations', 'Non-African nations'))  
  
d\_c %>%  
 ggplot(aes(x = rugged)) +  
 geom\_point(aes(rugged, loggdp, color = cont\_africa), shape = 16) +  
 theme(legend.position = '') +  
 geom\_line(aes(rugged, mu\_mean)) +  
 geom\_ribbon(aes(x=rugged,   
 ymin=mu\_pi\_l,   
 ymax=mu\_pi\_h,   
 fill = cont\_africa),   
 alpha = .1) +  
 facet\_wrap(~inAfrica) +  
 theme\_pander() +  
 labs(x = 'Terrain Ruggedness Index(avg pred w/o S)', y = 'log GDP year 2000')



The result of the model-averaged predictions are even more easier to see the weak effect of interaction. The slope looks nearly flat.

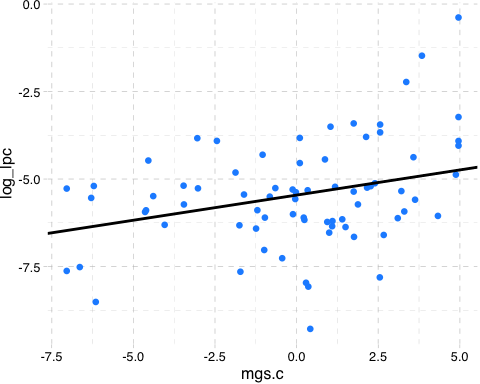
# Question2

## (a)

data('nettle', package = 'rethinking')  
d3 <- nettle; rm(rugged)  
  
# data-preprocessing  
d3$lang.per.cap <- d3$num.lang / d3$k.pop  
d3$log\_lpc <- log(d3$lang.per.cap)  
d3$log\_area <- log(d3$area)  
d3$log\_area.c <- d3$log\_area - mean(d3$log\_area)  
d3$mgs.c <- d3$mean.growing.season - mean(d3$mean.growing.season)  
d3$sgs.c <- d3$sd.growing.season - mean(d3$sd.growing.season)

m2a\_0="  
data {  
int N;  
vector[N] log\_lpc;  
vector[N] mgsc;  
}  
  
parameters {  
real a;  
real bM;  
real<lower=0, upper=10> sigma;  
}  
  
model {  
vector[N] mu = a + bM \* mgsc;  
//prior  
a ~ normal(-5,5);  
bM ~ normal(0,5);  
//likelihood  
log\_lpc ~ normal(mu,sigma);   
}  
"  
  
dat <- list(  
 N = NROW(d3),  
 log\_lpc = d3$log\_lpc,  
 mgsc = d3$mgs.c  
)  
  
fitm2a\_0 = stan(model\_code = m2a\_0,   
 data = dat,   
 cores = 4,   
 chains = 4,   
 iter = 3000)

m2a\_1="  
data {  
int N;  
vector[N] log\_lpc;  
vector[N] mgsc;  
vector[N] log\_areac;  
}  
  
parameters {  
real a;  
real bM;  
real bA;  
real<lower=0, upper=10> sigma;  
}  
  
model {  
vector[N] mu = a + bM \* mgsc + bA \* log\_areac;  
//prior  
a ~ normal(-5,5);  
bM ~ normal(0,5);  
bA ~ normal(0,5);  
//likelihood  
log\_lpc ~ normal(mu,sigma);   
}  
  
generated quantities {  
vector[N] log\_lik;  
{  
 vector[N] mu;  
 for(n in 1:N) {  
 mu[n] = a + bM \* mgsc[n] + bA \* log\_areac[n];  
 log\_lik[n] = normal\_lpdf(log\_lpc[n] | mu[n], sigma);  
 }  
}  
}  
"  
  
dat <- list(  
 N = NROW(d3),  
 log\_lpc = d3$log\_lpc,  
 mgsc = d3$mgs.c,  
 log\_areac = d3$log\_area.c  
)  
  
fitm2a\_1 = stan(model\_code = m2a\_1,   
 data = dat,   
 cores = 4,   
 chains = 4,   
 iter = 3000)  
  
post <- as.data.frame(fitm2a\_1)  
  
d3 %>%  
 ggplot(aes(x = mgs.c)) +  
 geom\_point(aes(mgs.c, log\_lpc),   
 shape = 16,   
 color = 'dodgerblue',size = 2) +   
 geom\_abline(slope = mean(post$bM),  
 intercept = mean(post$a),  
 size = 1 ) +  
 theme(text = element\_text(family = "Times"),  
 legend.position = "none") +  
 theme\_pander() +   
 xlab("mgs.c") +  
 ylab("log\_lpc")



#coeff  
my\_coef\_tab <-  
 rbind(tidy(fitm2a\_0), tidy(fitm2a\_1)) %>%  
 mutate(model = c(rep("fitm2a\_0", times = nrow(tidy(fitm2a\_0))),  
 rep("fitm2a\_1", times = nrow(tidy(fitm2a\_1))))) %>%  
 filter(term != "lp\_\_") %>%  
 select(model, everything())  
  
coef <-  
 my\_coef\_tab %>%  
 complete(term = distinct(., term), model) %>%  
 select(model, term, estimate) %>%  
 mutate(estimate = round(estimate, digits = 2)) %>%  
 spread(key = model, value = estimate)  
  
# compare the coefficients of 2 models  
head(coef, n=3)

## # A tibble: 3 x 3  
## term fitm2a\_0 fitm2a\_1  
## <chr> <dbl> <dbl>  
## 1 a -5.45 -5.46  
## 2 bA NA -0.2   
## 3 bM 0.17 0.14

If we includes area, the mean length of growing season was slightly reduced.

## (b)

m2b\_0="  
data {  
int N;  
vector[N] log\_lpc;  
vector[N] sgsc;  
}  
  
parameters {  
real a;  
real bS;  
real<lower=0, upper=10> sigma;  
}  
  
model {  
vector[N] mu = a + bS \* sgsc;  
//prior  
a ~ normal(-5,5);  
bS ~ normal(0,5);  
//likelihood  
log\_lpc ~ normal(mu,sigma);   
}  
  
generated quantities {  
vector[N] log\_lik;  
{  
 vector[N] mu;  
 for(n in 1:N) {  
 mu[n] = a + bS \* sgsc[n] ;  
 log\_lik[n] = normal\_lpdf(log\_lpc[n] | mu[n], sigma);  
 }  
}  
}  
"  
  
dat <- list(  
 N = NROW(d3),  
 log\_lpc = d3$log\_lpc,  
 sgsc = d3$sgs.c  
)  
  
fitm2b\_0 = stan(model\_code = m2b\_0,   
 data = dat,   
 cores = 4,   
 chains = 4,   
 iter = 3000)  
  
  
m2b\_1="  
data {  
int N;  
vector[N] log\_lpc;  
vector[N] sgsc;  
vector[N] log\_areac;  
}  
  
parameters {  
real a;  
real bS;  
real bA;  
real<lower=0, upper=10> sigma;  
}  
  
model {  
vector[N] mu = a + bS \* sgsc + bA \* log\_areac;  
//prior  
a ~ normal(-5,5);  
bS ~ normal(0,5);  
bA ~ normal(0,5);  
//likelihood  
log\_lpc ~ normal(mu,sigma);   
}  
  
generated quantities {  
vector[N] log\_lik;  
{  
 vector[N] mu;  
 for(n in 1:N) {  
 mu[n] = a + bS \* sgsc[n] + bA \* log\_areac[n];  
 log\_lik[n] = normal\_lpdf(log\_lpc[n] | mu[n], sigma);  
 }  
}  
}  
"  
  
dat <- list(  
 N = NROW(d3),  
 log\_lpc = d3$log\_lpc,  
 sgsc = d3$sgs.c,  
 log\_areac = d3$log\_area.c  
)  
  
fitm2b\_1 = stan(model\_code = m2b\_1,   
 data = dat,   
 cores = 4,   
 chains = 4,   
 iter = 3000)  
  
#coeff  
my\_coef\_tab <-  
 rbind(tidy(fitm2b\_0), tidy(fitm2b\_1)) %>%  
 mutate(model = c(rep("fitm2b\_0", times = nrow(tidy(fitm2b\_0))),  
 rep("fitm2b\_1", times = nrow(tidy(fitm2b\_1))))) %>%  
 filter(term != "lp\_\_") %>%  
 select(model, everything())  
  
coef <-  
 my\_coef\_tab %>%  
 complete(term = distinct(., term), model) %>%  
 select(model, term, estimate) %>%  
 mutate(estimate = round(estimate, digits = 2)) %>%  
 spread(key = model, value = estimate)  
  
# compare the coefficients of 2 models  
head(coef, n=3)

## # A tibble: 3 x 3  
## term fitm2b\_0 fitm2b\_1  
## <chr> <dbl> <dbl>  
## 1 a -5.46 -5.46  
## 2 bA NA -0.24  
## 3 bS -0.36 -0.21

when controlling for area, the SD slope is reduced a lot in magnitude (-0.36 to 0.21)

## (c)

m2c="  
data {  
int N;  
vector[N] log\_lpc;  
vector[N] sgsc;  
vector[N] mgsc;  
}  
  
parameters {  
real a;  
real bS;  
real bM;  
real bMS;  
real<lower=0, upper=10> sigma;  
}  
  
model {  
vector[N] mu;  
for (i in 1:N) {  
mu[i] = a + bS \* sgsc[i] + bM \* mgsc[i] + bMS\*mgsc[i]\*sgsc[i];}  
  
//prior  
a ~ normal(-5,5);  
bS ~ normal(0,5);  
bM ~ normal(0,5);  
bMS ~ normal(0,5);  
  
//likelihood  
log\_lpc ~ normal(mu,sigma);   
}  
"  
  
dat <- list(  
 N = NROW(d3),  
 log\_lpc = d3$log\_lpc,  
 sgsc = d3$sgs.c,  
 mgsc = d3$mgs.c  
)  
  
fitm2c = stan(model\_code = m2c,   
 data = dat,   
 cores = 4,   
 chains = 4,   
 iter = 3000)  
  
print(fitm2c)

## Inference for Stan model: 31c8ca49bdd792fea093ccabb019364f.  
## 4 chains, each with iter=3000; warmup=1500; 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  
## a -5.45 0.00 0.16 -5.75 -5.55 -5.45 -5.35 -5.13 6687 1  
## bS -0.34 0.00 0.15 -0.63 -0.45 -0.35 -0.25 -0.04 6390 1  
## bM 0.12 0.00 0.06 0.00 0.08 0.12 0.15 0.23 5694 1  
## bMS -0.11 0.00 0.05 -0.21 -0.14 -0.11 -0.08 -0.01 6108 1  
## sigma 1.37 0.00 0.12 1.16 1.28 1.36 1.44 1.62 6206 1  
## lp\_\_ -59.26 0.03 1.64 -63.39 -60.06 -58.92 -58.06 -57.10 2871 1  
##   
## Samples were drawn using NUTS(diag\_e) at Thu May 2 02:25:43 2019.  
## 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).

The estimates table quite make sense for an interaction.