Assignment4\_B04704016林家毅

# Question 1 (a)

# Load data  
data(rugged)  
data\_rugged <- rugged  
  
# Drop incomplete rows  
data\_rugged <- data\_rugged[ complete.cases( data\_rugged$rgdppc\_2000 ) , ]  
  
# Get log GDP  
data\_rugged$log\_gdp <- log( data\_rugged$rgdppc\_2000 )  
  
# Rescale variables  
data\_rugged$log\_gdp\_std <- data\_rugged$log\_gdp / mean(data\_rugged$log\_gdp)  
data\_rugged$rugged\_std <- data\_rugged$rugged / max(data\_rugged$rugged)  
mean(data\_rugged$rugged\_std)

## [1] 0.2149601

# Drop Seychelles  
data\_rugged\_without\_seychelles <- data\_rugged[ data\_rugged$country!="Seychelles" , ]  
mean(data\_rugged\_without\_seychelles$rugged\_std)

## [1] 0.2115714

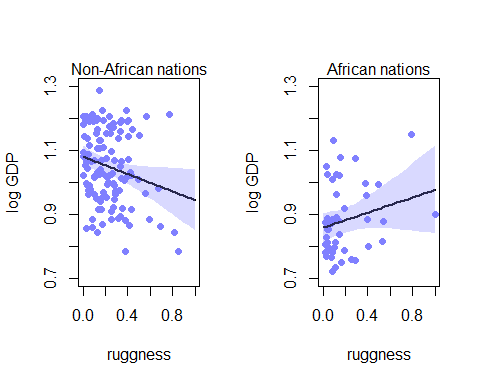
# Set input data and generate models  
rugged\_seq <- seq( from=0 , to=1 , length.out=30 )  
data\_rugged.tomodel <- list(  
 N = NROW(data\_rugged),  
 cont\_africa = as.integer(data\_rugged$cont\_africa),  
 rugged\_std = data\_rugged$rugged\_std,  
 log\_gdp\_std = data\_rugged$log\_gdp\_std  
)  
str(data\_rugged.tomodel)

## List of 4  
## $ N : int 170  
## $ cont\_africa: int [1:170] 1 0 0 0 0 0 0 0 0 1 ...  
## $ rugged\_std : num [1:170] 0.138 0.553 0.124 0.125 0.433 ...  
## $ log\_gdp\_std: num [1:170] 0.88 0.965 1.166 1.104 0.915 ...

model.rugged <- ulam(  
 alist(  
 log\_gdp\_std ~ dnorm( mu , sigma ),  
 mu <- a + bA\*cont\_africa + bR\*(rugged\_std - 0.2149601) + bAR\*cont\_africa\*(rugged\_std - 0.2149601),  
 a ~ dnorm( 1 , 0.1 ),  
 bA ~ dnorm( 0 , 0.3 ),  
 bR ~ dnorm( 0 , 0.3 ),  
 bAR ~ dnorm( 0 , 0.3 ),  
 sigma ~ dexp( 1 )  
 ),  
 data=data\_rugged.tomodel, chains=4 , cores=4 , iter=1000  
)  
precis( model.rugged , 2 )

## mean sd 5.5% 94.5% n\_eff Rhat  
## a 1.0504569 0.010415509 1.0332602 1.06730171 1793.228 1.001950  
## bA -0.1667918 0.019981178 -0.1980719 -0.13460221 1682.453 1.001018  
## bR -0.1352191 0.054973156 -0.2214415 -0.04550675 1394.387 1.003941  
## bAR 0.2537691 0.090236397 0.1056882 0.39938989 1390.423 1.003996  
## sigma 0.1114638 0.006067479 0.1024047 0.12166193 2126.295 1.000410

par(mfrow=c(1,2))  
for ( s in 0:1 ) {  
 idx <- which( data\_rugged$cont\_africa==s )  
 plot( data\_rugged$rugged\_std[idx] , data\_rugged$log\_gdp\_std[idx] , xlim=c(0,1) , ylim=c(0.7,1.3) ,  
 xlab="ruggness" , ylab="log GDP" , pch=16 , col=rangi2 )  
 mu <- link( model.rugged , data=data.frame( cont\_africa=s , rugged\_std=rugged\_seq ))  
 mu\_mean <- apply( mu , 2 , mean )  
 mu\_PI <- apply( mu , 2 , PI , prob=0.97 )  
 lines( rugged\_seq , mu\_mean , lwd=2 )  
 shade( mu\_PI , rugged\_seq , col=col.alpha(rangi2,0.3) )  
 if ( s==0 ) {  
 mtext("Non-African nations")  
 } else {  
 mtext("African nations")  
 }  
}



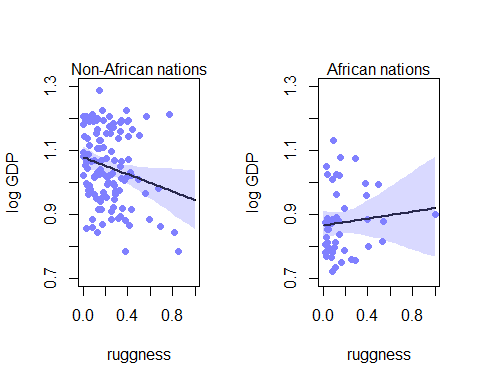
data\_rugged\_without\_seychelles.tomodel <- list(  
 N = NROW(data\_rugged\_without\_seychelles),  
 cont\_africa = as.integer(data\_rugged\_without\_seychelles$cont\_africa),  
 rugged\_std = data\_rugged\_without\_seychelles$rugged\_std,  
 log\_gdp\_std = data\_rugged\_without\_seychelles$log\_gdp\_std  
)  
str(data\_rugged\_without\_seychelles.tomodel)

## List of 4  
## $ N : int 169  
## $ cont\_africa: int [1:169] 1 0 0 0 0 0 0 0 0 1 ...  
## $ rugged\_std : num [1:169] 0.138 0.553 0.124 0.125 0.433 ...  
## $ log\_gdp\_std: num [1:169] 0.88 0.965 1.166 1.104 0.915 ...

model.without.seychelles <- ulam(  
 alist(  
 log\_gdp\_std ~ dnorm( mu , sigma ),  
 mu <- a + bA\*cont\_africa + bR\*(rugged\_std - 0.2115714) + bAR\*cont\_africa\*(rugged\_std - 0.2115714),  
 a ~ dnorm( 1 , 0.1 ),  
 bA ~ dnorm( 0 , 0.3 ),  
 bR ~ dnorm( 0 , 0.3 ),  
 bAR ~ dnorm( 0 , 0.3 ),  
 sigma ~ dexp( 1 )  
 ),  
 data=data\_rugged\_without\_seychelles.tomodel, chains=4 , cores=4 , iter=1000  
)  
precis( model.without.seychelles , 2 )

## mean sd 5.5% 94.5% n\_eff Rhat  
## a 1.0504417 0.010060453 1.03411410 1.06677492 1488.676 1.001858  
## bA -0.1741345 0.019359040 -0.20485861 -0.14317889 1460.549 1.000779  
## bR -0.1349147 0.053832102 -0.21914027 -0.04739414 1509.103 1.000194  
## bAR 0.1891665 0.095534759 0.03933674 0.34643205 1655.030 0.999348  
## sigma 0.1108416 0.005846444 0.10170017 0.12048719 1456.905 1.002333

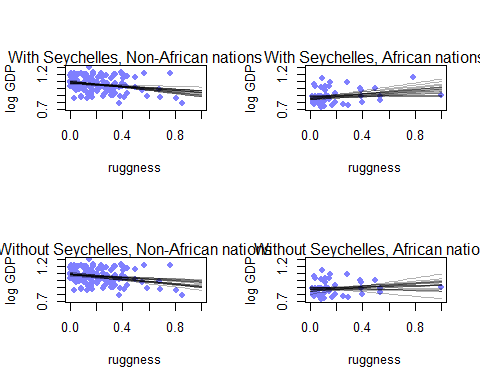
par(mfrow=c(1,2))  
for ( s in 0:1 ) {  
 idx <- which( data\_rugged\_without\_seychelles.tomodel$cont\_africa==s )  
 plot( data\_rugged\_without\_seychelles.tomodel$rugged\_std[idx] , data\_rugged\_without\_seychelles.tomodel$log\_gdp\_std[idx] , xlim=c(0,1) , ylim=c(0.7,1.3) ,  
 xlab="ruggness" , ylab="log GDP" , pch=16 , col=rangi2 )  
 mu <- link( model.without.seychelles , data=data.frame( cont\_africa=s , rugged\_std=rugged\_seq ))  
 mu\_mean <- apply( mu , 2 , mean )  
 mu\_PI <- apply( mu , 2 , PI , prob=0.97 )  
 lines( rugged\_seq , mu\_mean , lwd=2 )  
 shade( mu\_PI , rugged\_seq , col=col.alpha(rangi2,0.3) )  
 if ( s==0 ) {  
 mtext("Non-African nations")  
 } else {  
 mtext("African nations")  
 }  
}



# Conclusion:  
# According to the graphs above, we could find out that no matter Seychelles is included in the model or not, ruggness is positively related to log GDP of African nations. On the other hand, ruggness is negatively related to log GDP of non-African nations in both models. However, the difference between two models is that for the one that does not include Seychelles, the slope of ruggness and log GDP of African nations seems to be flatter, compared to the original model.

# Question 1 (b)

# Plot the predictions  
par(mfrow=c(2,2))  
for ( s in 0:1 ) {  
 idx <- which( data\_rugged$cont\_africa==s )  
 plot( data\_rugged$rugged\_std[idx] , data\_rugged$log\_gdp\_std[idx] , xlim=c(0,1) , ylim=c(0.7,1.3) ,  
 xlab="ruggness" , ylab="log GDP" , pch=16 , col=rangi2 )  
 mu <- link( model.rugged , data=data.frame( cont\_africa=s , rugged\_std=rugged\_seq ))  
 for ( i in 1:20 ) lines( rugged\_seq , mu[i,] , col=col.alpha("black",0.3) )  
 if ( s==0 ) {  
 mtext("With Seychelles, Non-African nations")  
 } else {  
 mtext("With Seychelles, African nations")  
 }  
}  
for ( s in 0:1 ) {  
 idx <- which( data\_rugged\_without\_seychelles$cont\_africa==s )  
 plot( data\_rugged\_without\_seychelles$rugged\_std[idx] , data\_rugged\_without\_seychelles$log\_gdp\_std[idx] , xlim=c(0,1) , ylim=c(0.7,1.3) ,  
 xlab="ruggness" , ylab="log GDP" , pch=16 , col=rangi2 )  
 mu <- link( model.without.seychelles , data=data.frame( cont\_africa=s , rugged\_std=rugged\_seq ))  
 for ( i in 1:20 ) lines( rugged\_seq , mu[i,] , col=col.alpha("black",0.3) )  
 if ( s==0 ) {  
 mtext("Without Seychelles, Non-African nations")  
 } else {  
 mtext("Without Seychelles, African nations")  
 }  
}



# Conclusion:  
# As for the predictions, we could find out that for each model, the predictions on log GDP of non-African nations are quite the same. However, as for predicting log GDP of African nations, the model not including Seychelles generates predictions of more flatter slopes than the other one.

# Question 1 (c)

model.c1 <- map2stan(  
 alist(  
 log\_gdp\_std ~ dnorm( mu , sigma ),  
 mu <- a + bR\*(rugged\_std - 0.2115714),  
 a ~ dnorm( 1 , 0.1 ),  
 bR ~ dnorm( 0 , 0.3 ),  
 sigma ~ dexp( 1 )  
 ),  
 data=data\_rugged\_without\_seychelles.tomodel, chains=4 , cores=4 , iter=1000  
)  
precis( model.c1 , 2 )

## mean sd 5.5% 94.5% n\_eff Rhat  
## a 0.99957486 0.010482074 0.9830805 1.01645306 1819.670 0.9998952  
## bR -0.01223602 0.058704862 -0.1040155 0.08246863 1665.950 0.9999027  
## sigma 0.13829379 0.007854328 0.1262975 0.15093151 1705.367 0.9988094

model.c2 <- map2stan(  
 alist(  
 log\_gdp\_std ~ dnorm( mu , sigma ),  
 mu <- a + bA\*cont\_africa + bR\*(rugged\_std - 0.2115714),  
 a ~ dnorm( 1 , 0.1 ),  
 bA ~ dnorm( 0 , 0.3 ),  
 bR ~ dnorm( 0 , 0.3 ),  
 sigma ~ dexp( 1 )  
 ),  
 data=data\_rugged\_without\_seychelles.tomodel, chains=4 , cores=4 , iter=1000  
)  
precis( model.c2 , 2 )

## mean sd 5.5% 94.5% n\_eff Rhat  
## a 1.04968285 0.010041797 1.0336449 1.065586749 1603.458 1.0018456  
## bA -0.17924738 0.018898692 -0.2106270 -0.149487664 1549.038 1.0023849  
## bR -0.07945738 0.048180854 -0.1544851 -0.002452869 1973.963 0.9991725  
## sigma 0.11197277 0.005988461 0.1029189 0.122177988 2046.410 1.0000745

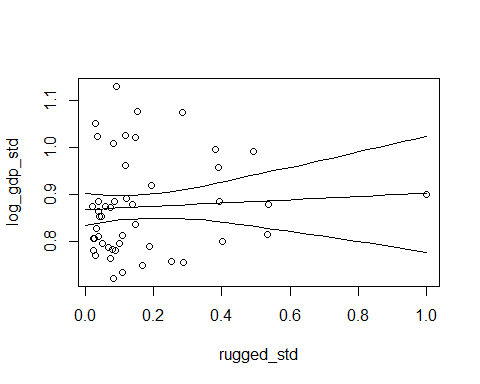
model.c3 <- map2stan(  
 alist(  
 log\_gdp\_std ~ dnorm( mu , sigma ),  
 mu <- a + bA\*cont\_africa + bR\*(rugged\_std - 0.2115714) + bAR\*cont\_africa\*(rugged\_std - 0.2115714),  
 a ~ dnorm( 1 , 0.1 ),  
 bA ~ dnorm( 0 , 0.3 ),  
 bR ~ dnorm( 0 , 0.3 ),  
 bAR ~ dnorm( 0 , 0.3 ),  
 sigma ~ dexp( 1 )  
 ),  
 data=data\_rugged\_without\_seychelles.tomodel, chains=4 , cores=4 , iter=1000  
)  
precis( model.c3 , 2 )

## mean sd 5.5% 94.5% n\_eff Rhat  
## a 1.0507617 0.010202909 1.03463208 1.06720005 1866.025 0.9983744  
## bA -0.1742625 0.019171862 -0.20458797 -0.14361878 1633.730 0.9986479  
## bR -0.1384647 0.053596927 -0.22438894 -0.05391544 1481.409 1.0000886  
## bAR 0.1947107 0.093929588 0.04265394 0.34689113 1501.219 1.0001914  
## sigma 0.1108486 0.006134432 0.10149318 0.12132089 1862.304 0.9993884

compare( model.c1 , model.c2 , model.c3 )

## WAIC pWAIC dWAIC weight SE dSE  
## model.c3 -261.3866 4.170152 0.000000 8.310429e-01 14.55945 NA  
## model.c2 -258.2005 3.732954 3.186073 1.689571e-01 13.85545 3.370127  
## model.c1 -187.8565 2.609183 73.530119 8.969351e-17 13.01597 15.118374

par(mfrow=c(1,1))  
prediction.data <- data.frame( rugged\_std = rugged\_seq, cont\_africa = 1 )  
mu.ensemble <- ensemble( model.c1 , model.c2 , model.c3 , data = prediction.data )  
mu.mean <- apply( mu.ensemble$link , 2 , mean )  
mu.PI <- apply( mu.ensemble$link , 2 , PI , prob=0.89 )  
data.plot <- data\_rugged\_without\_seychelles[(data\_rugged\_without\_seychelles$cont\_africa == 1) , ]  
plot( log\_gdp\_std ~ rugged\_std , data=data.plot )  
lines( rugged\_seq , mu.mean )  
lines( rugged\_seq , mu.PI[1,] )  
lines( rugged\_seq , mu.PI[2,] )

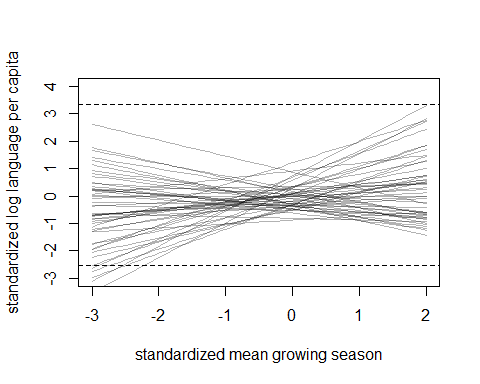


# Conclusion:  
# The inferences are quite similar to the results of graph in lower-right in part(b), because both predictions drop Seychelles out of the models. Also, both models consider the interactions between variables.

# Question 2 (a)

# Load data  
data(nettle)  
data\_nettle <- nettle  
data\_nettle <- data\_nettle[ complete.cases( data\_nettle$num.lang , data\_nettle$k.pop , data\_nettle$area ) , ]  
  
# Generate variables  
data\_nettle$lang.per.cap <- data\_nettle$num.lang/data\_nettle$k.pop  
data\_nettle$log.lang.per.cap <- log(data\_nettle$lang.per.cap)  
data\_nettle$log.area <- log(data\_nettle$area)  
data\_nettle$log.k.pop <- log(data\_nettle$k.pop)  
data\_nettle$pop.dens <- data\_nettle$log.k.pop/data\_nettle$log.area  
  
# Standardize variables  
data\_nettle$log\_lang\_per\_cap\_std <- (data\_nettle$log.lang.per.cap - mean(data\_nettle$log.lang.per.cap))/sd(data\_nettle$log.lang.per.cap)  
min\_lang <- min(data\_nettle$log\_lang\_per\_cap\_std)  
max\_lang <- max(data\_nettle$log\_lang\_per\_cap\_std)  
data\_nettle$mean\_growing\_season\_std <- (data\_nettle$mean.growing.season - mean(data\_nettle$mean.growing.season))/sd(data\_nettle$mean.growing.season)  
min\_mean\_grow <- min(data\_nettle$mean\_growing\_season\_std)  
max\_mean\_grow <- max(data\_nettle$mean\_growing\_season\_std)  
data\_nettle$sd\_growing\_season\_std <- (data\_nettle$sd.growing.season - mean(data\_nettle$sd.growing.season))/sd(data\_nettle$sd.growing.season)  
min\_sd\_grow <- min(data\_nettle$sd\_growing\_season\_std)  
max\_sd\_grow <- max(data\_nettle$sd\_growing\_season\_std)  
data\_nettle$area\_std <- (data\_nettle$area - mean(data\_nettle$area))/sd(data\_nettle$area)  
  
# Modeling  
model.2a <- map2stan(  
 alist(  
 log\_lang\_per\_cap\_std ~ dnorm( mu , sigma ),  
 mu <- a + bM\*mean\_growing\_season\_std + bA\*area\_std,  
 a ~ dnorm( 0 , 0.5 ),  
 bM ~ dnorm( 0 , 0.5 ),  
 bA ~ dnorm( 0 , 0.5 ),  
 sigma ~ dexp( 1 )  
 ),  
 data=data\_nettle, chains=4 , cores=4 , iter=1000  
)  
  
# Inspect prior predictions  
par(mfrow=c(1,1))  
prior <- extract.prior( model.2a )  
plot( NULL , xlim=c(floor(min\_mean\_grow),ceiling(max\_mean\_grow)) , ylim=c(floor(min\_lang),ceiling(max\_lang)) ,  
 xlab="standardized mean growing season" , ylab="standardized log language per capita" )  
abline( h=min\_lang , lty=2 )  
abline( h=max\_lang , lty=2 )  
growing\_season\_seq <- seq( from=floor(min\_mean\_grow) , to=ceiling(max\_mean\_grow) , length.out=50 )  
mu <- link( model.2a , post=prior , data=data.frame( mean\_growing\_season\_std=growing\_season\_seq, area\_std=0 ))

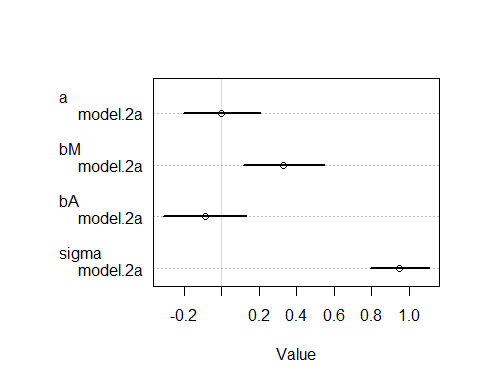
for ( i in 1:50 ) lines( growing\_season\_seq , mu[i,] , col=col.alpha("black",0.3) )



precis( model.2a )

## mean sd 5.5% 94.5% n\_eff Rhat  
## a -0.0001297975 0.10517281 -0.1730193 0.16337724 2616.292 1.0005140  
## bM 0.3314653225 0.11031814 0.1506461 0.50887353 2322.638 0.9996611  
## bA -0.0916580668 0.11193167 -0.2710691 0.08107019 2101.952 1.0006676  
## sigma 0.9465225738 0.07956058 0.8276919 1.08118946 2091.664 0.9987633

plot( coeftab( model.2a ) )

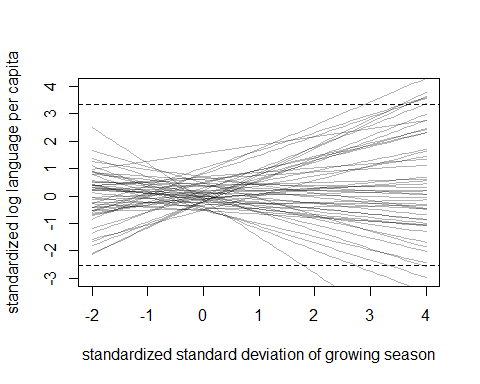


# Conclusion:  
# The results seem to support the hypothesis. As the bM coefficient is positive, we could consider that language diversity is positively associated with the average length of the growing season.

# Question 2 (b)

model.2b <- map2stan(  
 alist(  
 log\_lang\_per\_cap\_std ~ dnorm( mu , sigma ),  
 mu <- a + bS\*sd\_growing\_season\_std + bA\*area\_std,  
 a ~ dnorm( 0 , 0.5 ),  
 bS ~ dnorm( 0 , 0.5 ),  
 bA ~ dnorm( 0 , 0.5 ),  
 sigma ~ dexp( 1 )  
 ),  
 data=data\_nettle, chains=4 , cores=4 , iter=1000  
)  
  
# Inspect prior predictions  
par(mfrow=c(1,1))  
prior <- extract.prior( model.2b )  
plot( NULL , xlim=c(floor(min\_sd\_grow),ceiling(max\_sd\_grow)) , ylim=c(floor(min\_lang),ceiling(max\_lang)) ,  
 xlab="standardized standard deviation of growing season" , ylab="standardized log language per capita" )  
abline( h=min\_lang , lty=2 )  
abline( h=max\_lang , lty=2 )  
growing\_season\_seq <- seq( from=floor(min\_sd\_grow) , to=ceiling(max\_sd\_grow) , length.out=50 )  
mu <- link( model.2b , post=prior , data=data.frame( sd\_growing\_season\_std=growing\_season\_seq, area\_std=0 ))

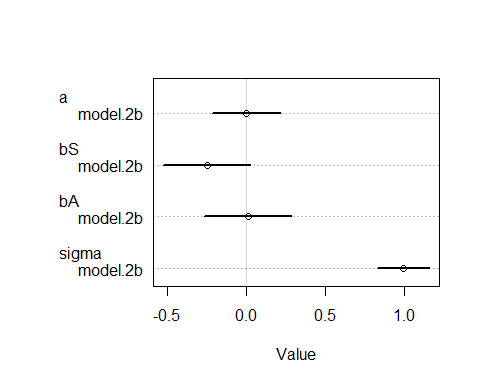
for ( i in 1:50 ) lines( growing\_season\_seq , mu[i,] , col=col.alpha("black",0.3) )



precis( model.2b )

## mean sd 5.5% 94.5% n\_eff Rhat  
## a -0.002510227 0.10948799 -0.1764304 0.17331041 2131.605 1.0003010  
## bS -0.251042951 0.13988992 -0.4802409 -0.01906016 1598.843 0.9991786  
## bA 0.011425351 0.14024356 -0.2093896 0.23701782 1525.646 0.9988575  
## sigma 0.988360918 0.08363497 0.8619342 1.13431593 1915.333 0.9997966

plot( coeftab( model.2b ) )

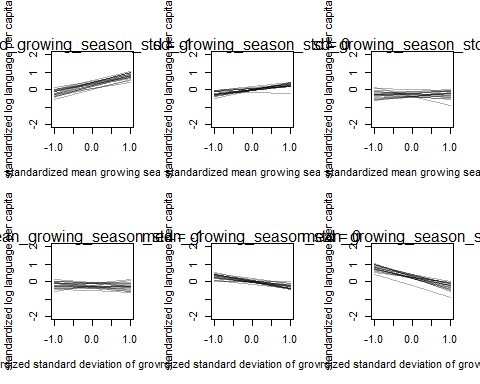


# Conclusion:  
# The results seem support the hypothesis. As the bS coefficient is negative, we could consider that language diversity is negatively associated with the standard deviation of length of the growing season.

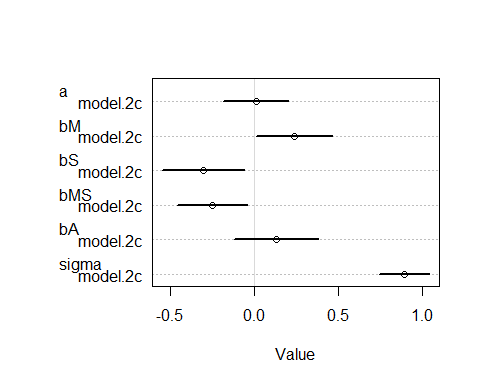
# Question 2 (c)

model.2c <- map2stan(  
 alist(  
 log\_lang\_per\_cap\_std ~ dnorm( mu , sigma ),  
 mu <- a + bM\*mean\_growing\_season\_std + bS\*sd\_growing\_season\_std + bMS\*mean\_growing\_season\_std\*sd\_growing\_season\_std + bA\*area\_std,  
 a ~ dnorm( 0 , 0.5 ),  
 bM ~ dnorm( 0 , 0.5 ),  
 bS ~ dnorm( 0 , 0.5 ),  
 bMS ~ dnorm( 0 , 0.5 ),  
 bA ~ dnorm( 0 , 0.5 ),  
 sigma ~ dexp( 1 )  
 ),  
 data=data\_nettle, chains=4 , cores=4 , iter=1000  
)  
  
# Inspect prior predictions  
par(mfrow=c(2,3))  
for ( s in -1:1 ) {  
 idx <- which( data\_nettle$sd\_growing\_season\_std==s )  
 plot( data\_nettle$mean\_growing\_season\_std[idx] , data\_nettle$log\_lang\_per\_cap\_std[idx] , xlim=c(-1,1) , ylim=c(-2,2) ,  
 xlab="standardized mean growing season" , ylab="standardized log language per capita" , pch=16 , col=rangi2 )  
 mu <- link( model.2c , data=data.frame( sd\_growing\_season\_std=s , mean\_growing\_season\_std=-1:1 , area\_std=0 ) )  
 for ( i in 1:20 ) lines( -1:1 , mu[i,] , col=col.alpha("black",0.3) )  
 if ( s == -1 ) {  
 mtext("sd\_growing\_season\_std = -1")  
 } else if ( s == 0 ) {  
 mtext("sd\_growing\_season\_std = 0")  
 } else if ( s == 1 ) {  
 mtext("sd\_growing\_season\_std = 1")  
 }  
}

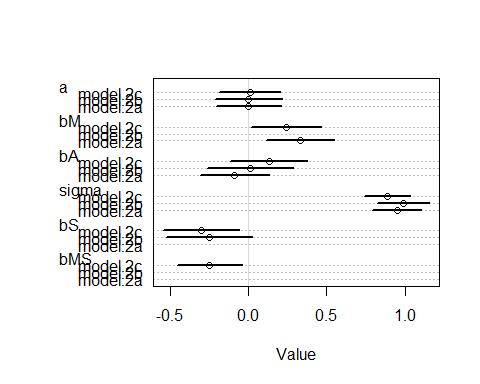
for ( s in -1:1 ) {  
 idx <- which( data\_nettle$mean\_growing\_season\_std==s )  
 plot( data\_nettle$sd\_growing\_season\_std[idx] , data\_nettle$log\_lang\_per\_cap\_std[idx] , xlim=c(-1,1) , ylim=c(-2,2) ,  
 xlab="standardized standard deviation of growing season" , ylab="standardized log language per capita" , pch=16 , col=rangi2 )  
 mu <- link( model.2c , data=data.frame( mean\_growing\_season\_std=s , sd\_growing\_season\_std=-1:1 , area\_std=0 ) )  
 for ( i in 1:20 ) lines( -1:1 , mu[i,] , col=col.alpha("black",0.3) )  
 if ( s == -1 ) {  
 mtext("mean\_growing\_season\_std = -1")  
 } else if ( s == 0 ) {  
 mtext("mean\_growing\_season\_std = 0")  
 } else if ( s == 1 ) {  
 mtext("mean\_growing\_season\_std = 1")  
 }  
}



par(mfrow=c(1,1))  
plot( coeftab( model.2c ) )



plot( coeftab( model.2a , model.2b , model.2c ) )



# Conclusion:  
# The results seem to support the hypothesis that nations with longer average and higher variance of growing seasons might lead to greater social integration and fewer languages per capita.