Assignment5

Wen-Shiuan, Liang

11/30/2017

Question1

Consider the data(rugged) data on economic development and terrain ruggedness examined in Chapter 7. One of the African countries in the example, Seychelles, is far outside the cloud of other nations, being a rare country with both relatively high GDP and high ruggedness. Seychelles is also unusual, in that it is a group of islands far from the coast of mainland Africa, and its main economic activity is tourism.One might suspect that this one nation is exerting a strong influence on the conclusions. In this problem, we want you to drop Seychelles from the data and re-evaluate the hypothesis that the relationship of African economies with ruggedness is different from that on other continents.

```
rm(list=ls())                           # clear memory

## R code 7.1
library(rethinking)

## Loading required package: rstan

## Loading required package: ggplot2

## Loading required package: StanHeaders

## rstan (Version 2.16.2, packaged: 2017-07-03 09:24:58 UTC, GitRev: 2e1f913d
3ca3)

## For execution on a local, multicore CPU with excess RAM we recommend calli
ng
## rstan_options(auto_write = TRUE)
## options(mc.cores = parallel::detectCores())

## Loading required package: parallel

## rethinking (Version 1.59)

data(rugged)
d <- rugged

# make log version of outcome
d$log_gdp <- log( d$rgdppc_2000 )
```

```r
#extract countries with GDP data
d1 <- d[ complete.cases(d$rgdppc_2000) , ]

# exclude Seychelles
d2 <- d1[!d1$country=='Seychelles',]

#The interaction model (the model in question was wrong)
m1 <- map(
  alist(
    log_gdp ~ dnorm( mu , sigma ) ,
    mu <- a + bA*cont_africa +gamma*rugged,
    gamma <- bR + bAr*cont_africa,
    a ~ dnorm( 8 , 100 ) ,
    bA ~ dnorm( 0 , 1 ) ,
    bR ~ dnorm(0 ,1 ),
    bAr ~ dnorm(0, 1 ),
    sigma ~ dunif( 0 , 10 )
  ) ,
  data=d1)

m2 <- map(
  alist(
    log_gdp ~ dnorm( mu , sigma ) ,
    mu <- a + bA*cont_africa +gamma*rugged,
    gamma <- bR + bAr*cont_africa,
    a ~ dnorm( 8 , 100 ) ,
    bA ~ dnorm( 0 , 1 ) ,
    bR ~ dnorm(0 ,1 ),
    bAr ~ dnorm(0, 1 ),
    sigma ~ dunif( 0 , 10 )
  ) ,
  data=d2)

precis(m1)
```

```
##        Mean StdDev  5.5% 94.5%
## a      9.18   0.14  8.97  9.40
## bA    -1.85   0.22 -2.20 -1.50
## bR    -0.18   0.08 -0.31 -0.06
## bAr    0.35   0.13  0.14  0.55
## sigma  0.93   0.05  0.85  1.01
```

```r
precis(m2)
```

```
##        Mean StdDev  5.5% 94.5%
## a      9.19   0.14  8.97  9.40
## bA    -1.78   0.22 -2.13 -1.43
## bR    -0.19   0.08 -0.31 -0.07
```

```
## bAr    0.25   0.14  0.04  0.47
## sigma  0.93   0.05  0.85  1.01
```

(a)Compare the inference from this model fit to the data without Seychelles to the same model fit to the full data. Does it still seem like the effect of ruggedness depends upon continents? How much has the expected relationship changed?

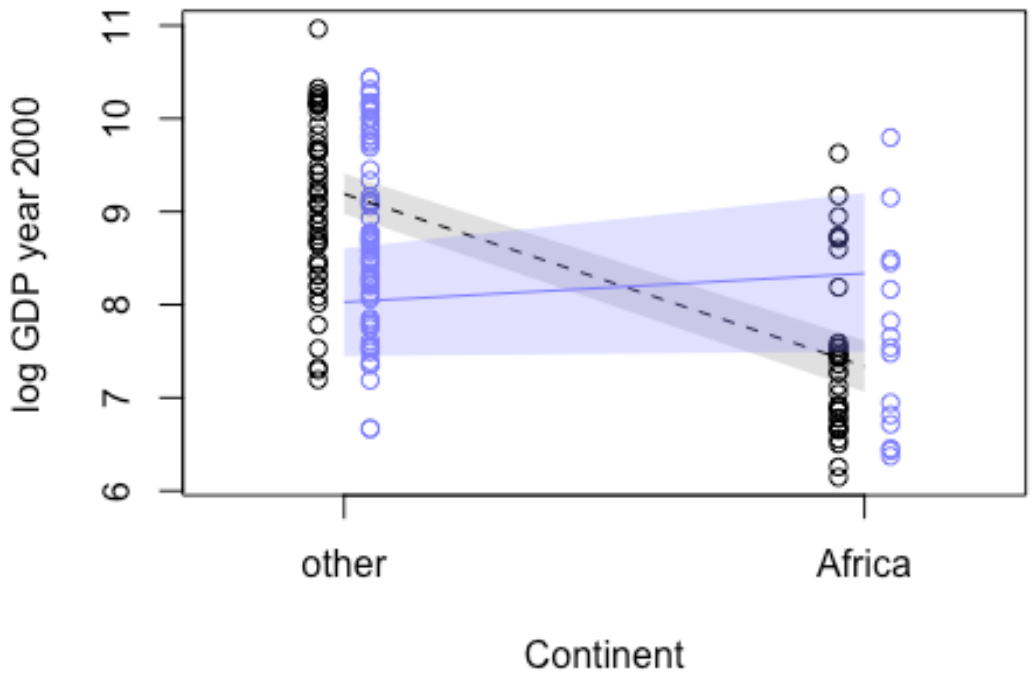
Yes, it still seem like the effect of ruggedness depends upon continents, because the interaction bAr is still between 0 and 1. However, the effect of the interaction does have reduce by removing the sample,Seychelles. The relationship has changed about 0.1.

(b) Now plot the predictions of the interaction model, with and without Seychelles. Does it still seem like the effect of ruggedness depends upon continents? How much has the expected relationship changed?

The original interaction plot (m1)

```
q.rugged <- range(d1$rugged)
mu.ruggedlo <- link( m1 ,
                data=data.frame(rugged=q.rugged[1],cont_africa=0:1) )

## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]

mu.ruggedlo.mean <- apply( mu.ruggedlo , 2 , mean )
mu.ruggedlo.PI <- apply( mu.ruggedlo , 2 , PI )
mu.ruggedhi <- link( m1 ,
                data=data.frame(rugged=q.rugged[2],cont_africa=0:1) )

## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]
```

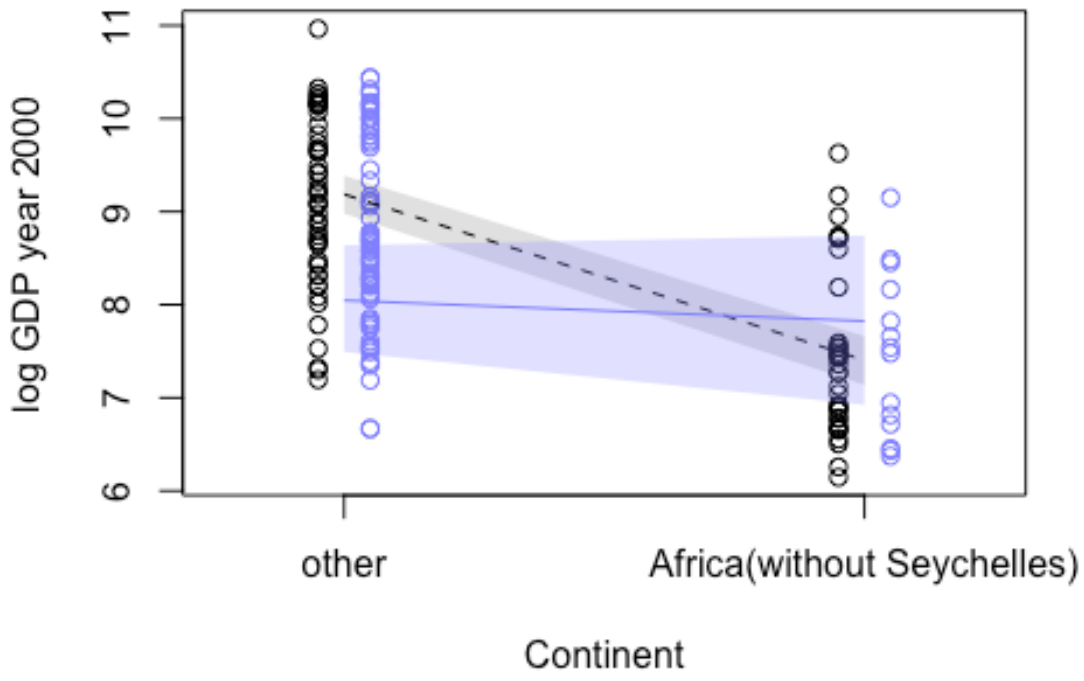
```
mu.ruggedhi.mean <- apply( mu.ruggedhi , 2 , mean )
mu.ruggedhi.PI <- apply( mu.ruggedhi , 2 , PI )

# plot it all, splitting points at median
med.r <- median(d1$rugged)
ox <- ifelse( d1$rugged > med.r , 0.05 , -0.05 )
plot( d1$cont_africa + ox , log(d1$rgdppc_2000) ,
      col=ifelse(d1$rugged>med.r,rangi2,"black") ,
      xlim=c(-0.25,1.25) , xaxt="n" , ylab="log GDP year 2000" , xlab="Contin
ent" )
axis( 1 , at=c(0,1) , labels=c("other","Africa") )
lines( 0:1 , mu.ruggedlo.mean , lty=2 )
shade( mu.ruggedlo.PI , 0:1 )
lines( 0:1 , mu.ruggedhi.mean , col=rangi2 )
shade( mu.ruggedhi.PI , 0:1 , col=col.alpha(rangi2,0.25) )
```



The new interaction plot

```
q.rugged <- range(d2$rugged)
mu.ruggedlo <- link( m2 ,
                data=data.frame(rugged=q.rugged[1],cont_africa=0:1) )

## [ 100 / 1000 ]
[ 200 / 1000 ]
```

```
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]

mu.ruggedlo.mean <- apply( mu.ruggedlo , 2 , mean )
mu.ruggedlo.PI <- apply( mu.ruggedlo , 2 , PI )
mu.ruggedhi <- link( m2 ,
                data=data.frame(rugged=q.rugged[2],cont_africa=0:1) )

## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]

mu.ruggedhi.mean <- apply( mu.ruggedhi , 2 , mean )
mu.ruggedhi.PI <- apply( mu.ruggedhi , 2 , PI )

# plot it all, splitting points at median
med.r <- median(d2$rugged)
ox <- ifelse( d2$rugged > med.r , 0.05 , -0.05 )
plot( d2$cont_africa + ox , log(d2$rgdppc_2000) ,
      col=ifelse(d2$rugged>med.r,rangi2,"black") ,
      xlim=c(-0.25,1.25) , xaxt="n" , ylab="log GDP year 2000" , xlab="Contin
ent" )
axis( 1 , at=c(0,1) , labels=c("other","Africa(without Seychelles)") )
lines( 0:1 , mu.ruggedlo.mean , lty=2 )
shade( mu.ruggedlo.PI , 0:1 )
lines( 0:1 , mu.ruggedhi.mean , col=rangi2 )
shade( mu.ruggedhi.PI , 0:1 , col=col.alpha(rangi2,0.25) )
```



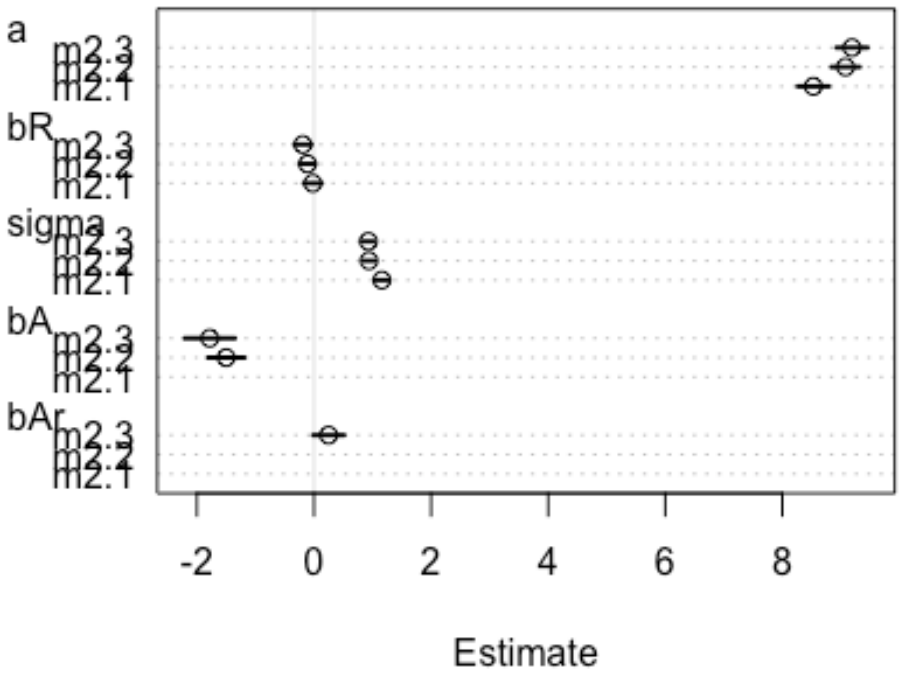
After removing the outlier, Seychelles, the reverse intepretation for high ruggness has eliminated. It means that no matter the country has high ruggness or not, the the GDP would be higher for those which is not Africa country. Though the effect would be 'much higher' for those with low ruggness.

Therefore, it still seem like the effect of ruggedness depends upon continents. Based on the coeficient, originally African countries gain premium on ruggness of 0.35, but now they only get 0.25 premium, meaning that Seychelles has misled that "High ruggedness is good for African countries".

(c)Finally, conduct a model comparison analysis, using WAIC. Fit three models to the data without Seychelles.

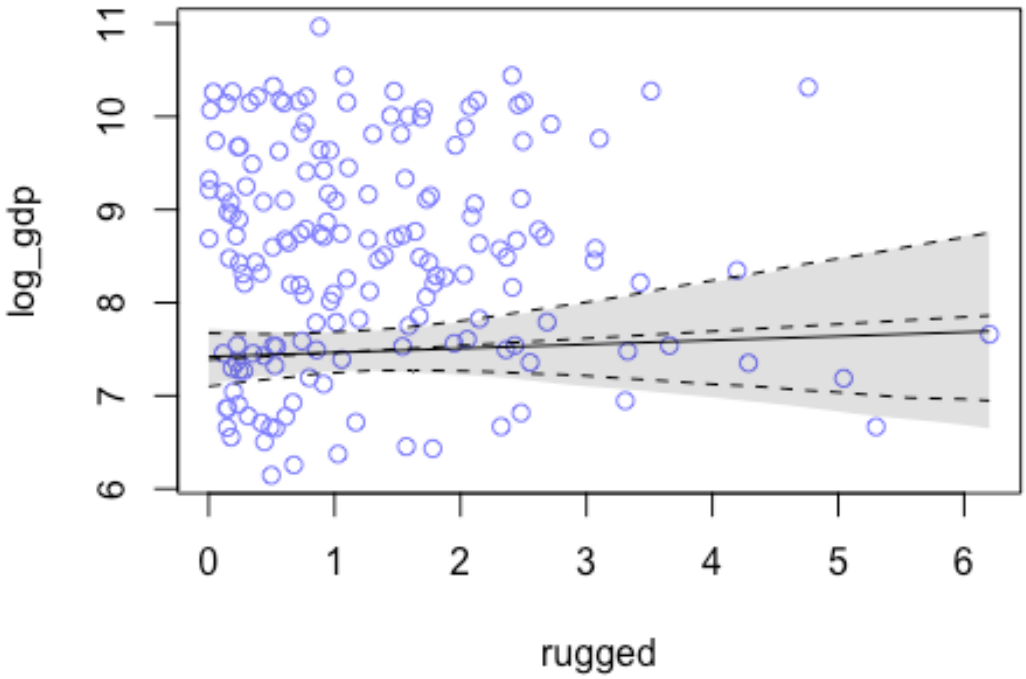
```
m2.1 <- map(
  alist(
    log_gdp ~ dnorm( mu , sigma ) ,
    mu <- a + bR*rugged,
    a ~ dnorm( 8 , 100 ),
    bR ~ dnorm(0 ,1 ),
    sigma ~ dunif( 0 , 10 )
  ) ,
  data=d2)
m2.2 <- map(
```

```
  alist(
    log_gdp ~ dnorm( mu , sigma ) ,
    mu <- a + bA*cont_africa +bR*rugged,
    a ~ dnorm( 8 , 100 ) ,
    bA ~ dnorm( 0 , 1 ) ,
    bR ~ dnorm(0 ,1 ),
    sigma ~ dunif( 0 , 10 )
  ) ,
  data=d2)
m2.3 <- map(
  alist(
    log_gdp ~ dnorm( mu , sigma ) ,
    mu <- a + bA*cont_africa +gamma*rugged,
    gamma <- bR + bAr*cont_africa,
    a ~ dnorm( 8 , 100 ) ,
    bA ~ dnorm( 0 , 1 ) ,
    bR ~ dnorm(0 ,1 ),
    bAr ~ dnorm(0, 1 ),
    sigma ~ dunif( 0 , 10 )
  ) ,
  data=d2)
plot(coeftab(m2.1 , m2.2 , m2.3))
```



```
rugged_model <- compare( m2.1 , m2.2 , m2.3 , func=WAIC )

#For Africa
int.seq <- seq(from=0,to=6.2,length.out=30)
d.predict <- list(
  loggdp = rep(0,30), # empty outcome
  rugged = int.seq,     # sequence of rugged
  cont_africa = rep(1,30) # For africa
)
pred.m2.3 <- link( m2.3 , data=d.predict )

## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]

mu <- apply( pred.m2.3 , 2 , mean )
mu.PI <- apply( pred.m2.3 , 2 , PI )

# plot it all
plot( log_gdp ~ rugged , d2 , col=rangi2 )
lines( int.seq , mu , lty=2 )
lines( int.seq , mu.PI[1,] , lty=2 )
lines( int.seq , mu.PI[2,] , lty=2 )

## R code 6.30
rug.ensemble <- ensemble( m2.1, m2.2 , m2.3, data=d.predict )

## Constructing posterior predictions

## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]

## Constructing posterior predictions

## [ 100 / 1000 ]
[ 200 / 1000 ]
```

```
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]

## Constructing posterior predictions

## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]

mu <- apply( rug.ensemble$link , 2 , mean )
mu.PI <- apply( rug.ensemble$link , 2 , PI )
lines( int.seq , mu )
shade( mu.PI , int.seq )
```



From the plot it seems that there exists no effect for rugged for African countries, which is different from the results of Question B.

Here we try to see what makes the difference.

```
precis(m2.3)
```

```
##         Mean StdDev  5.5% 94.5%
## a       9.19   0.14  8.97  9.40
## bA     -1.78   0.22 -2.13 -1.43
## bR     -0.19   0.08 -0.31 -0.07
## bAr     0.25   0.14  0.04  0.47
## sigma   0.93   0.05  0.85  1.01
```

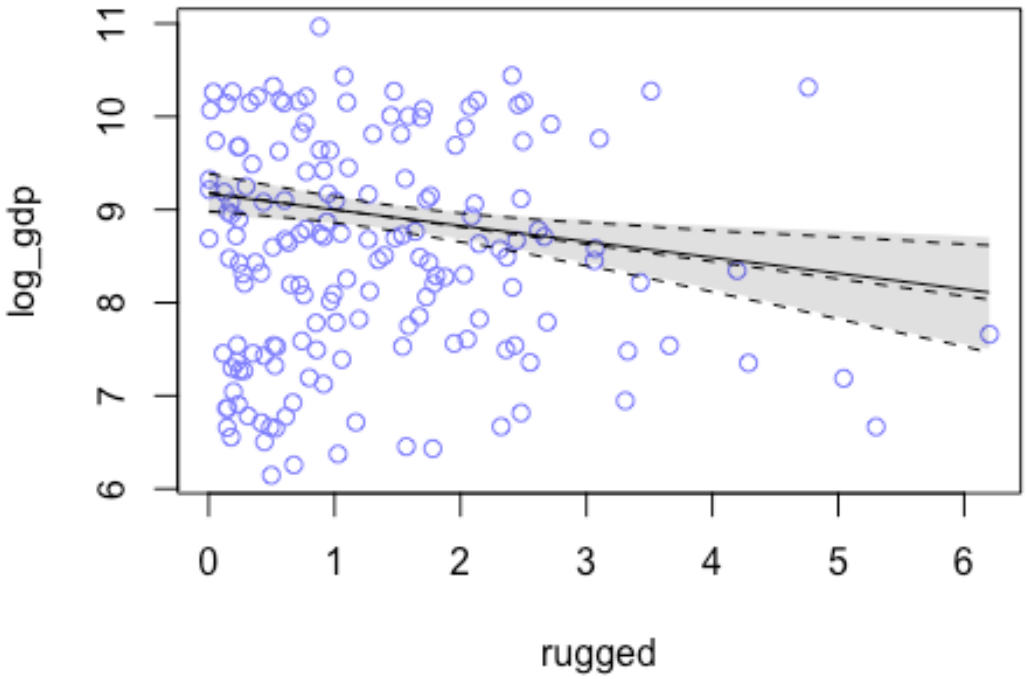
```
precis(m2.2)
```

```
##         Mean StdDev  5.5% 94.5%
## a       9.08   0.12  8.88  9.28
## bA     -1.50   0.16 -1.76 -1.25
## bR     -0.11   0.06 -0.21 -0.01
## sigma   0.94   0.05  0.86  1.02
```

The reason could be resulted from weighted m2.2(0.22), which provides overall less effect for parameters on loggdp. Thus the ruggeness has less effect.

```
#For non-Africa
int.seq <- seq(from=0,to=6.2,length.out=30)
d.predict <- list(
  loggdp = rep(0,30), # empty outcome
  rugged = int.seq,     # sequence of rugged
  cont_africa = rep(0,30) # non africa
)
pred.m2.3 <- link( m2.3 , data=d.predict )

## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]

mu <- apply( pred.m2.3 , 2 , mean )
mu.PI <- apply( pred.m2.3 , 2 , PI )

# plot it all
plot( log_gdp ~ rugged , d2 , col=rangi2 )
lines( int.seq , mu , lty=2 )
lines( int.seq , mu.PI[1,] , lty=2 )
lines( int.seq , mu.PI[2,] , lty=2 )

## R code 6.30
rug.ensemble <- ensemble( m2.1, m2.2 , m2.3, data=d.predict )

## Constructing posterior predictions

## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]

## Constructing posterior predictions

## [ 100 / 1000 ]
[ 200 / 1000 ]
```

```
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]

## Constructing posterior predictions

## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]

mu <- apply( rug.ensemble$link , 2 , mean )
mu.PI <- apply( rug.ensemble$link , 2 , PI )
lines( int.seq , mu )
shade( mu.PI , int.seq )
```



From the plot it seems that there exists negative effect for rugged for non-African countries, which consistent with question B.

Question2

Use these data to evaluate the hypothesis that language diversity is partly a product of food security. The notion is that, in productive ecologies, people don't need large social networks to buffer them against risk of food shortfalls. This means ethnic groups can be smaller and more self-sufficient, leading to more languages per capita. In contrast, in a poor ecology, there is more subsistence risk, and so human societies have adapted by building larger networks of mutual obligation to provide food insurance. This in turn creates social forces that help prevent languages from diversifying.

```
rm(list=ls())                          # clear memory

## R code 7.1
library(rethinking)
data(nettle)
d <- nettle
```

```
d$lang.per.cap <- d$num.lang/d$k.pop
summary(d)
```

```
##        country      num.lang           area             k.pop
##  Algeria   : 1   Min.   :  1.00   Min.   :  12189   Min.   :   102
##  Angola    : 1   1st Qu.: 17.25   1st Qu.: 167708   1st Qu.:  3829
##  Australia : 1   Median : 40.00   Median : 434796   Median :  9487
##  Bangladesh: 1   Mean   : 89.73   Mean   : 880698   Mean   : 33574
##  Benin     : 1   3rd Qu.: 93.75   3rd Qu.:1080316   3rd Qu.: 24744
##  Bolivia   : 1   Max.   :862.00   Max.   :8511965   Max.   :849638
##  (Other)   :68
##   num.stations    mean.growing.season sd.growing.season
##  Min.   :  1.00   Min.   : 0.000      Min.   :0.0000
##  1st Qu.: 10.00   1st Qu.: 5.348      1st Qu.:0.9375
##  Median : 20.50   Median : 7.355      Median :1.6900
##  Mean   : 37.91   Mean   : 7.041      Mean   :1.6992
##  3rd Qu.: 44.75   3rd Qu.: 9.283      3rd Qu.:2.1075
##  Max.   :272.00   Max.   :12.000      Max.   :5.8700
##
##   lang.per.cap
##  Min.   :0.0000931
##  1st Qu.:0.0019901
##  Median :0.0041066
##  Mean   :0.0206464
##  3rd Qu.:0.0100059
##  Max.   :0.6809816
##
```

(a) Evaluate the hypothesis that language diversity, as measured by log(lang.per.cap), is positively associated with the average length of the growing season, mean.growing.season. Consider log(area) in your regression(s) as a covariate (not an interaction). Interpret your results.

```
m1 <- map(
  alist(
    log(lang.per.cap) ~ dnorm( mu , sigma ) ,
    mu <- a + bG*mean.growing.season + bA*log(area),
    a ~ dnorm( -10 , 5) ,
    bG ~ dnorm( 1 , 2 ) ,
    bA ~ dnorm(1 ,2 ),
    sigma ~ dunif( 0 , 10 )
  ) ,
  data=d)
precis(m1)
```

```
##         Mean StdDev  5.5% 94.5%
## a     -4.76   1.83 -7.68 -1.83
## bG     0.16   0.05  0.07  0.24
## bA    -0.14   0.13 -0.35  0.07
## sigma  1.39   0.11  1.21  1.57
```

The variable mean.growing.season is positive and significant and log(area) is not, meaning that the hypothesis of might be correct. In addition, for bigger the area, people are living less intensive, thus should produce more languages. However, the coeficient is negative, which opposed to our intuition.

(b) Now evaluate the hypothesis that language diversity is negatively associated with the standard deviation of length of growing season, sd.growing.season. This hypothesis follows from uncertainty in harvest favoring social insurance through larger social networks and therefore fewer languages. Again, consider log(area) as a covariate (not an interaction). Interpret your results.

```
m2 <- map(
  alist(
    log(lang.per.cap) ~ dnorm( mu , sigma ) ,
    mu <- a + bG*sd.growing.season + bA*log(area),
    a ~ dnorm( -10 , 5) ,
    bG ~ dnorm( 10 , 5 ) ,
    bA ~ dnorm(1 ,2 ),
    sigma ~ dunif( 0 , 10 )
  ) ,
  data=d)
precis(m2)
```

```
##        Mean StdDev  5.5% 94.5%
## a     -3.02   1.76 -5.84 -0.20
## bG    -0.24   0.18 -0.53  0.06
## bA    -0.16   0.15 -0.39  0.08
## sigma  1.44   0.12  1.25  1.63
```

The variable sd.growing.season and log(area) are both insignificant.

(c) Finally, evaluate the hypothesis that mean.growing.season and sd.growing.season interact to synergistically reduce language diversity. The idea is that, in nations with longer average growing seasons, high variance makes storage and redistribution even more important than it would be otherwise. That way, people can cooperate to preserve and protect windfalls to be used during the droughts. These forces in turn may lead to greater social integration and fewer languages.

```
m3 <- map(
  alist(
    log(lang.per.cap) ~ dnorm( mu , sigma ) ,
    mu <- a + bA*mean.growing.season + bR*gamma*sd.growing.season,
    gamma <- bR + bAr*mean.growing.season,
    a ~ dnorm( -10 , 5 ) ,
    bA ~ dnorm( 1 , 2 ) ,
    bR ~ dnorm(5 ,5 ),
    bAr ~ dnorm(5, 5 ),
```

```
    sigma ~ dunif( 0 , 10 )
    ),
  data = d)
precis(m3)
```

```
##        Mean StdDev  5.5% 94.5%
## a     -7.04   0.58 -7.98 -6.11
## bA     0.31   0.07  0.19  0.42
## bR     0.68   0.27  0.25  1.11
## bAr   -0.17   0.03 -0.21 -0.12
## sigma  1.31   0.11  1.14  1.48
```

All parameters are significant. For the mean.growing.season, the direction of coeficient is consistent with our intuition. However, the coeficient of mean.growing.season is positive, meaning that the larger the sd of growing season, the more diversified the languages would be. It could be explained by the changing season cause migration, and thus leading more diversified group of people. Or there might exist unobserved association. The interaction of these two parameters is negative significance, which is consistent with our intuition.