HomeWork4

Yung-Tang Chou

2019年5月1日

Here is the rmarkdown file for the Multivariate Analysis Homework3.

Before we start let’s import some libraries and do data transformation.

library(rstan)

## Loading required package: ggplot2

## Loading required package: StanHeaders

## rstan (Version 2.18.2, 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)

## For improved execution time, we recommend calling  
## Sys.setenv(LOCAL\_CPPFLAGS = '-march=native')  
## although this causes Stan to throw an error on a few processors.

library(rethinking)

## Loading required package: parallel

## rethinking (Version 1.88)

data('rugged')  
d <- rugged  
d$log\_gdp <- log(d$rgdppc\_2000)  
d <- d[complete.cases(d$rgdppc\_2000),]  
d$log\_gdp\_std <- d$log\_gdp / mean(d$log\_gdp)  
d$rugged\_std <- d$rugged / max(d$rugged)  
d2 <- d[(d$country != 'Seychelles'), ]

|  |
| --- |
| For Question 1 |
| Problem A |
| Firstly we build two models for full data and full data without Seychelles. |
| ```r model.01 <- quap(alist( log\_gdp\_std ~ dnorm(mu, sigma), mu <- a + br \* rugged\_std + bc \* cont\_africa + brc*rugged\_std*cont\_africa, a ~ dnorm(1, 0.3), br ~ dnorm(0, 0.3), bc ~ dnorm(0.5,0.25), brc ~ dnorm(0.5, 0.25), sigma ~ dexp(1) ), data = d) |
| model.02 <- quap(alist( log\_gdp\_std ~ dnorm(mu, sigma), mu <- a + br \* rugged\_std + bc \* cont\_africa + brc*rugged\_std*cont\_africa, a ~ dnorm(1, 0.3), br ~ dnorm(0, 0.3), bc ~ dnorm(0.5,0.25), brc ~ dnorm(0.5, 0.25), sigma ~ dexp(1) ), data = d2) ``` |
| Before moving on, let’s peak into the parameter estimate information using precis function. For model.01, the parameter estimate is as follows. |
| ## mean sd 5.5% 94.5% ## a 1.0808683 0.015775077 1.0556567 1.10607994 ## br -0.1442445 0.053488965 -0.2297302 -0.05875884 ## bc -0.2251857 0.025302165 -0.2656234 -0.18474791 ## brc 0.2904814 0.087891356 0.1500141 0.43094878 ## sigma 0.1094981 0.005935976 0.1000113 0.11898494 |
| For model.02, the parameter estimate is as follows. |
| ## mean sd 5.5% 94.5% ## a 1.0815454 0.015661787 1.05651486 1.1065760 ## br -0.1470696 0.053124524 -0.23197290 -0.0621664 ## bc -0.2196152 0.025305258 -0.26005791 -0.1791725 ## brc 0.2328666 0.092800680 0.08455315 0.3811800 ## sigma 0.1086483 0.005908113 0.09920599 0.1180906 |
| Given the mean and slope of all the parameters shown above, there are not too much difference between model.01, which uses all data, and model.02, which uses all data except Syechelles. |
| Problem B |
| Before doing the plotting let’s first prepare items needed. |
| r par(mfrow=c(2,2)) d.A1 <- d[d$cont\_africa == 1,] d.A0 <- d[d$cont\_africa == 0,] d2.A1 <- d2[d2$cont\_africa == 1,] d2.A0 <- d2[d2$cont\_africa ==0, ] rugged\_seq <- seq(from=-0.1, to = 1.1, length.out = 30) |
| With all the prepatation, let’s first plot all African countries. |
| Then let’s plot non-african countries. |
| Similarly, let’s work on all African countries without Seychelles and all non-African countries without Seychelles. |
| From the plot we can get some simple facts. Firstly, for African nations, the extent of logarithmic ruggedness is positively related to the magnitude of GDP, which in turn implys that ruggedness is positively related to GDP. FOr non-African nations, the extent of logarithmic ruggedness is negatively related to the magnitude of GDP, which in turn implys that ruggedness is negatively related to GDP. |
| Problem C |
| Firstly build the three models for model comparison analysis. |
| ```r model.03 <- quap(alist( log\_gdp\_std ~ dnorm(mu, sigma), mu <- a + br \* rugged\_std, a ~ dnorm(1, 0.3), br ~ dnorm(0, 0.3), sigma ~ dexp(1) ), data = d2) |
| model.04 <- quap(alist( log\_gdp\_std ~ dnorm(mu, sigma), mu <- a + br \* rugged\_std + bc \* cont\_africa, a ~ dnorm(1, 0.3), br ~ dnorm(0, 0.3), bc ~ dnorm(0.5,0.25), sigma ~ dexp(1) ), data = d2) |
| model.05 <- quap(alist( log\_gdp\_std ~ dnorm(mu, sigma), mu <- a + br \* rugged\_std + bc \* cont\_africa + brc*rugged\_std*cont\_africa, a ~ dnorm(1, 0.3), br ~ dnorm(0, 0.3), bc ~ dnorm(0.5,0.25), brc ~ dnorm(0.5, 0.25), sigma ~ dexp(1) ), data = d2) ``` |
| Using WAIC to conduct model comparison analysis. This can be done with “Compare” under the rethinking package. |
| r compare(model.03, model.04, model.05) |
| ## WAIC pWAIC dWAIC weight SE dSE ## model.05 -260.6623 4.691640 0.000000 8.066182e-01 15.28181 NA ## model.04 -257.8059 4.007366 2.856368 1.933818e-01 14.33494 4.137158 ## model.03 -187.5600 2.824929 73.102242 1.078246e-16 13.43963 15.730079 |
| Given the results shown above, we can tell that model.05, which incorporates two vairables as well as the interaction, is the best prediction model in that it has the lowest WAIC value. |
| We can get the model-averaged predictions of these models using the “Ensemble” function. |
| r rugged.seq <- seq(from = 0, to = 8, length.out = 30) rugged.seq.std <- rugged.seq - mean(rugged.seq) data = data.frame(rugged\_std = rugged.seq.std, cont\_africa = 1) model.ensemble <- ensemble(model.03, model.04, model.05, data = data) |
| To better understand the predictions, let’s simply plot it out. |
| This plot is drawn based on the facts below. 1. The dataset includes all African and non-African nations excep Seychelles. 2. This is the model-averaged prediction of model.03, model.04, model.05. |
| The result shows positive relationship between the two variables, ruggedness and GDP. This is the combination of both African and non-African countries. Comparing with the plots from problem B, we can find out that if we do not consider the difference in countries, we will have wrong inplications for non-African countries. |

For Question 2

Before entering each problem, first let’s import data and libraries needed.

library(rstan)  
library(rethinking)  
data('nettle')  
d <- nettle  
str(d)  
  
d$lang.per.cap <- d$num.lang / d$k.pop  
d$log.lang.per.cap <- log(d$lang.per.cap)  
d$log.area <- log(d$area)  
d$mean.growing.season

Problem A

The first hypothesis is that language diversity is positively associated wit hthe average length of the growing season. The model is shown below.

model.06 <- quap(alist(  
 log.lang.per.cap ~ dnorm(mu, sigma),  
 mu <- a + bm \* d$mean.growing.season + ba \* d$log.area,  
 a ~ dnorm(0,30),  
 bm ~ dnorm(0,30),  
 ba ~ dnorm(0,30),  
 sigma ~ dexp(10)  
), data = d)

precis(model.06)

## mean sd 5.5% 94.5%  
## a -3.8454986 1.80980313 -6.73791355 -0.9530836668  
## bm 0.1435976 0.05135259 0.06152628 0.2256689950  
## ba -0.2027194 0.12670883 -0.40522460 -0.0002142509  
## sigma 1.2825938 0.09392344 1.13248598 1.4327015766

The result above shows that mean growing season is positively associated with language diversity, with slope = 0.14.

Problem B

model.07 <- quap(alist(  
 log.lang.per.cap ~ dnorm(mu, sigma),  
 mu <- a + bs \* sd.growing.season + ba \* log.area,  
 a ~ dnorm(0,30),  
 bs ~ dnorm(0,30),  
 ba ~ dnorm(0,30),  
 sigma ~ dexp(10)  
), data = d)

precis(model.07)

## mean sd 5.5% 94.5%  
## a -1.9943281 1.7277623 -4.7556260 0.76696973  
## bs -0.2089830 0.1717148 -0.4834164 0.06545045  
## ba -0.2402122 0.1436598 -0.4698084 -0.01061606  
## sigma 1.3247524 0.0966836 1.1702334 1.47927150

The result above shows that the standard deviation growing season is negatively associated with language diversity, with slope = -0.21.

Problem C

model.08 <- quap(alist(  
 log.lang.per.cap ~ dnorm(mu, sigma),  
 mu <- a + bm \* mean.growing.season + bs \* sd.growing.season + bms \* mean.growing.season \* sd.growing.season,  
 a ~ dnorm(0,30),  
 bm ~ dnorm(0,30),  
 bs ~ dnorm(0,30),  
 bms ~ dnorm(0,30),  
 sigma ~ dexp(10)  
), data = d)

precis(model.08)

## mean sd 5.5% 94.5%  
## a -6.9715370 0.54824353 -7.8477361 -6.0953380  
## bm 0.2994080 0.06680743 0.1926368 0.4061792  
## bs 0.4210809 0.34516333 -0.1305568 0.9727185  
## bms -0.1087363 0.04362694 -0.1784605 -0.0390120  
## sigma 1.2112743 0.08921400 1.0686931 1.3538555

The result above shows that with interaction, given mean growing season, standard deviation of the growing season will be positively associated with language diversity. Likewise, given standard deviation of the growing season, mean growing season will be positively associated with language diversity. This is the counterfactual result.

When considering the interaction effect of mean growing season as well as the standard deviation, the relationship is negatively associated. Intuitively, the higher the combination of mean growing seanson and standard deviation of growing season, social integration is more important than trading with other nations. Thus, language diversity should decrease.