Multivariate Analysis HW3 \_ b04704017

Yung-Tang Chou

2019年4月10日

Here is the rmarkdown file for the Multivariate Analysis Homework3.

Before we start let’s import some libraries

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(foxes)  
d <- foxes  
  
attach(foxes)

|  |
| --- |
| For Question 1 |
| We will use rstan to build the model for this bivariate Gaussian regressions. |
| For the first model (model.01), let’s build a model for the relationship between body weight and area. |
| ```r model.01 <- " data { int N; vector[N] weight; vector[N] area; } parameters { real a; real bA; real sigma; } model { vector[N] mu = a + bA \* area; |
| weight ~ normal(mu, sigma); a ~ normal(3, 2); bA ~ normal(0, 2); sigma ~ uniform(0,3); }" ``` |
| We make up another data in order to put into the stan model fit.01. We use 2 cores and 2 chains, with 10,000 iteration. |
| ```r data <- list( N = NROW(d), weight = darea ) |
| fit.01 <- stan(model\_code = model.01, data = data, cores = 2, chains = 2, iter = 1e4) ``` |
| After the model is fitted, simply use the extract command to extract samples from the fitted model. Then we use the relationship along with samples to find out the “mu” of the prediceted body weight. |
| r la <- extract(fit.01, permuted = TRUE) mu.link <- function(area) la$a + la$bA \* area |
| When it’s done, let’s start drawing the plot for the relationship. |
| r xseq <- seq(from = min(d$area)-0.3, to=max(d$area)+0.3, length.out = 1e4) mu <- sapply(xseq, mu.link) mu.mean <- apply(mu, 2, mean) mu.PI <- apply(mu, 2, PI, prob=.95) |
|  |
| Let’s use the command precis to look into this model. |
| ## mean sd 5.5% 94.5% n\_eff Rhat ## a 4.39319962 0.39121408 3.7658832 5.0155837 3543.266 1.000113 ## bA 0.04136276 0.11896063 -0.1475478 0.2295697 3616.685 1.000094 ## sigma 1.20132681 0.08043219 1.0808709 1.3357333 5336.053 1.000140 From the output we can tell that the relationship between body weight and merely area is low, with the mean slope being only 0.04. |
| Now let’s move on to the linear model of body weight and groupsize. Similarly, let’s start by building the stan model and the needed data for this relationship. |
| ```r model.02 <- " data { int N; vector[N] weight; vector[N] groupsize; } parameters { real a; real bG; real sigma; } model { vector[N] mu = a + bG \* groupsize; |
| weight ~ normal(mu, sigma); a ~ normal(3, 2); bG ~ normal(0, 2); sigma ~ uniform(0,3); }" |
| data.02 <- list( N = NROW(d), weight = dgroupsize ) ``` |
| Once it’s done, we put the data into the stan model model.02 Likewise, We choose 2 cores and 2 chains, with 10,000 iteration for our fitted model fit.02. |
| r fit.02 <- stan(model\_code = model.02, data = data.02, cores = 2, chains = 2, iter = 1e4) |
| After the model is fitted, simply use the extract command to extract samples from the fitted model. Then we use the relationship along with samples to find out the “mu” of the prediceted body weight. After that comes the plotting. |
| ```r la <- extract(fit.02, permuted = TRUE) mu.link <- function(groupsize) labG \* groupsize |
| xseq <- seq(from = min(dgroupsize), length.out = 1e4) mu <- sapply(xseq, mu.link) mu.mean <- apply(mu, 2, mean) mu.PI <- apply(mu, 2, PI, prob=.95) ``` |
|  |
| Again, let’s use the command precis to look into this model. |
| ## mean sd 5.5% 94.5% n\_eff Rhat ## a 5.006851 0.32687301 4.473880 5.529159428 3663.281 1.000174 ## bG -0.110880 0.07110222 -0.224194 0.004614677 3594.174 1.000182 ## sigma 1.186911 0.08053431 1.064519 1.321036381 4880.295 1.000242 From the output we can tell that the relationship between body weight and groupsize is slightly higher than that of body weight and area. However, the variable is still considered not important for predicting body weight. |
| To sum up, for model.01 (body weight vs. area) we can see the relationship between the two variables is quite low. We can conclude that one cannot use variable “Area” to predict body weight effectively. Likewise, for model.02 (body weight vs. groupsize), though the correlation is somhow slightly higher, it is still low. As a result, both Area and Groupsize are not important for predicting Body Weight. |

For Question 2

For this question I am going to plot two counterfactural plots. Before continuing, let’s standadize variables needed first.

d$A <- scale(d$area)  
d$G <- scale(d$groupsize)  
d$W <- scale(d$weight)  
attach(d)

## The following objects are masked from foxes:  
##   
## area, avgfood, group, groupsize, weight

For later comparison, let’s divide the plotting area into two part.

par(mfrow=c(1,2))

The first one is for the model of “Weight vs. Area”, holding Groupsize at its mean. First, let’s create the data for the counterfactural plot, along with the model. Note that of the pred\_data\_1, we keep groupsize at its mean, using ‘G = mean(d$G)’.

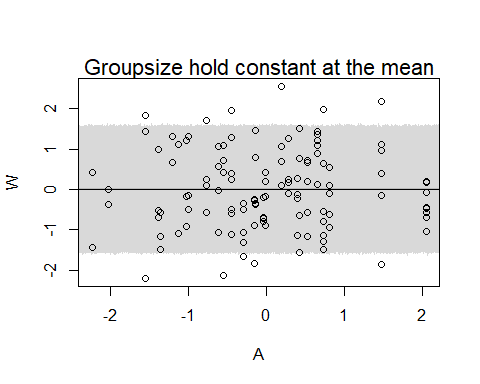
CFP\_seq <- seq(from=-3, to=3, length.out = 1e4)  
pred\_data\_1 <- data.frame(A = CFP\_seq, G = mean(d$G))  
model.03 <- quap(  
 alist(  
 W ~ dnorm(mu,sigma),  
 mu <- a + bA \* A + bG \* G,  
 a ~ dnorm(0,0.2),  
 bA ~ dnorm(0,0.5),  
 bG ~ dnorm(0,0.5),  
 sigma ~ dexp(1)  
 ), data = pred\_data\_1  
)

With the “link” command, we can fit the model with data prepared for this counterfactural plot.

mu\_model.03 <- link(model.03, data = pred\_data\_1)  
mu\_mean.03 <- apply(mu\_model.03, 2, mean)  
mu\_PI.03 <- apply(mu\_model.03, 2, PI, prob=.89)

We can simulate counterfactural outcome using the ‘sim’ command. Meanwhile, we draw the Percentage Interval of the simulated outcome.

D\_sim.03 <- sim(model.03, data= pred\_data\_1, n=1e4)  
D\_PI.03 <- apply(D\_sim.03, 2, PI)

Let’s draw the plot for the model “Weight vs. Area”, holding Groupsize at its mean.  From the plot we can tell that there’s no relationship between weight and area, holding groupsize at its mean.

The second model is for the model “Weight vs. Groupsize”, holding Area at its mean. First, let’s create the data for the counterfactural plot, along with the model. Note that of the pred\_data\_2, we keep area at its mean, using ‘A = mean(d$A)’.

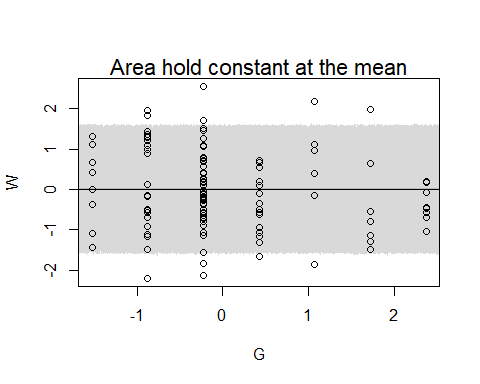
pred\_data\_2 <- data.frame(G = CFP\_seq, A = mean(d$area))  
model.04 <- quap(  
 alist(  
 W ~ dnorm(mu,sigma),  
 mu <- a + bA \* A + bG \* G,  
 a ~ dnorm(0,0.2),  
 bA ~ dnorm(0,0.5),  
 bG ~ dnorm(0,0.5),  
 sigma ~ dexp(1)  
 ), data = pred\_data\_2  
)

Similarly, with the “link” command, we can fit the model with data prepared for this counterfactural plot.

mu\_model.04 <- link(model.04, data = pred\_data\_2)  
mu\_mean.04 <- apply(mu\_model.04,2, mean)  
mu\_PI.04 <- apply(mu\_model.04, 2, PI, prob=.89)

Then we keep on simulating counterfactural outcome and display the predictions.

#Simulate counterfactural outcome  
D\_sim.04 <- sim(model.04, data=pred\_data\_2, n=1e4)  
D\_PI.04 <- apply(D\_sim.04, 2, PI)



From the plot, again we can tell that there’s no relationship between weight and groupsize, holding area at its mean.

To sum up, with the help of counterfactural plot, we know that both groupsize and area are not important variables that can be used to predict body weight.

Though slightly correlation exists in Question 1 for both variables, from the counterfactural plot we can tell that both variables cannot be an important factor when predicting body weight only by itself.

For Question 3

Finally we include variable avgfood into consideration. For Question 3-1, let’s find out whether avgfood by itself can be a significant predictor for body weight.

d$AVG <- scale(d$avgfood)  
attach(d)

## The following objects are masked from d (pos = 3):  
##   
## A, area, avgfood, G, group, groupsize, W, weight

## The following objects are masked from foxes:  
##   
## area, avgfood, group, groupsize, weight

model.05 <- quap(  
 alist(  
 W ~ dnorm(mu,sigma),  
 mu <- a + bAVG \* AVG,  
 a ~ dnorm(0,0.2),  
 bAVG ~ dnorm(0,0.5),  
 sigma ~ dexp(1)  
 ), data = d  
)

## mean sd 5.5% 94.5%  
## a 2.016421e-07 0.08360013 -0.1336090 0.1336094  
## bAVG -2.421145e-02 0.09088497 -0.1694632 0.1210403  
## sigma 9.911434e-01 0.06465849 0.8878066 1.0944801

For the table above we can tell that simply by avgfood itself cannot be a good predictor.

Let’s compare the two model, ‘Weight as a function of avgfood and groupsize’ and ‘Weight as a function of area and groupsize’.

First, let’s build a model for Weight, avgfood, and groupsize.

model.06 <- quap(  
 alist(  
 W ~ dnorm(mu,sigma),  
 mu <- a + bAVG \* AVG + bG \* G,  
 a ~ dnorm(0,0.2),  
 bAVG ~ dnorm(0,0.5),  
 bG ~ dnorm(0,0.5),  
 sigma ~ dexp(1)  
 ), data = d  
)

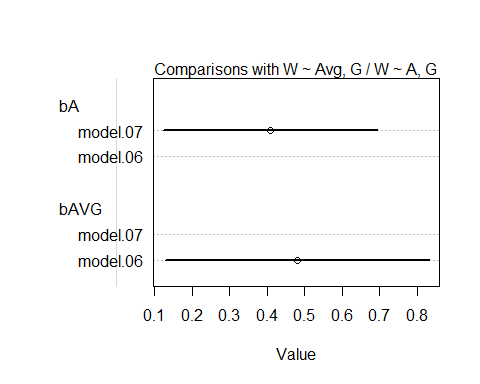
## mean sd 5.5% 94.5%  
## a -1.824237e-07 0.08013809 -0.1280763 0.1280760  
## bAVG 4.772542e-01 0.17912327 0.1909807 0.7635278  
## bG -5.735265e-01 0.17914177 -0.8598296 -0.2872233  
## sigma 9.420443e-01 0.06175262 0.8433517 1.0407369

From the result we can tell that within this model there are significant correlations.

Let’s continue building model for Weight, area, and groupsize.

model.07 <- quap(  
 alist(  
 W ~ dnorm(mu,sigma),  
 mu <- a + bA \* A + bG \* G,  
 a ~ dnorm(0,0.2),  
 bA ~ dnorm(0,0.5),  
 bG ~ dnorm(0,0.5),  
 sigma ~ dexp(1)  
 ), data = d  
)

## mean sd 5.5% 94.5%  
## a 8.466215e-05 0.08012670 -0.1279733 0.1281426  
## bA 4.059433e-01 0.14535390 0.1736397 0.6382469  
## bG -4.821029e-01 0.14536381 -0.7144223 -0.2497834  
## sigma 9.418848e-01 0.06158413 0.8434614 1.0403081

 From the result, again we find out that within this model there are significant correlations.

In the models above, both of them show significant relationships for the body weight. Avgfood, along with groupsize, can be an important set of predictors. Likewise, area, along with groupsize, can be an important set of predictors.

For Question 3-2, let’s create a model consisting all three variables: avgfood, area, and groupsize.

model.08 <- quap(  
 alist(  
 W ~ dnorm(mu,sigma),  
 mu <- a + bA \* A + bG \* G + bAVG \* AVG,  
 a ~ dnorm(0,0.2),  
 bA ~ dnorm(0,0.5),  
 bG ~ dnorm(0,0.5),  
 bAVG ~ dnorm(0,0.5),  
 sigma ~ dexp(1)  
 ), data = d  
)

## mean sd 5.5% 94.5%  
## a -4.138628e-07 0.07936201 -0.126836226 0.1268354  
## bA 2.782400e-01 0.17011226 0.006367741 0.5501122  
## bG -6.396192e-01 0.18161482 -0.929874806 -0.3493637  
## bAVG 2.968948e-01 0.20960025 -0.038086877 0.6318765  
## sigma 9.312063e-01 0.06100008 0.833716381 1.0286962

From the table above we can tell that with all three variables, there seems to be relationships between body weight and each variable. We can simply conclude that there are masked associations at first.

Before we end this question, we can draw a plot to compare the three models. 