National Taiwan University
Quantitative Business Science

Individual Assignment 4

Kyle (Wen-Shiuan, Liang)
R05724080

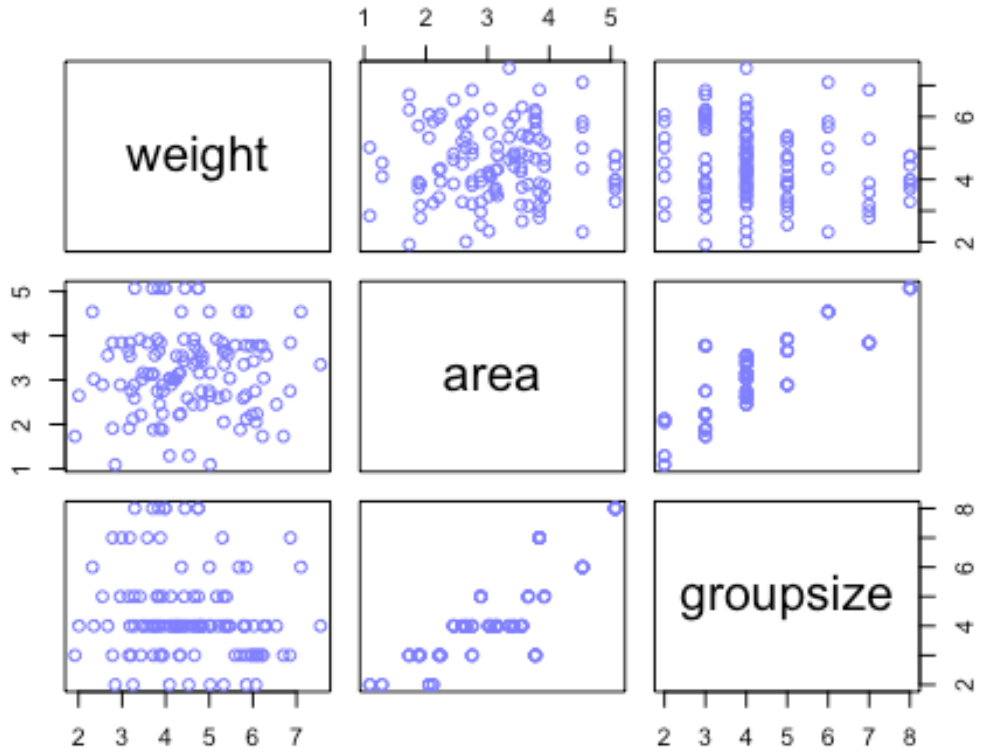
10/30/2017

Question1

Fit two bivariate Gaussian regressions, using map:

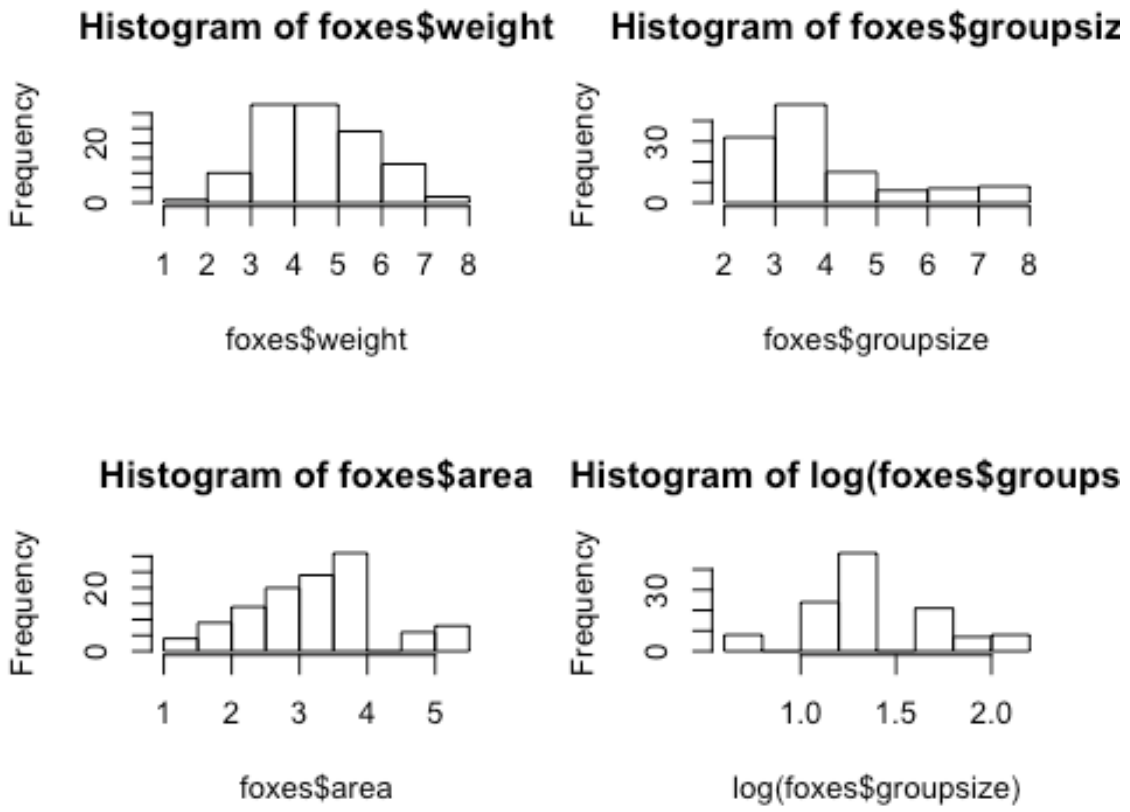
Data overview

```
library(rethinking)
data(foxes)
pairs(~ weight + area + groupsize, data = foxes, col=rangi2 )
```



'groupsize' seems to be correlated with 'area'.

```
par(mfrow=c(2,2))
hist(foxes$weight)
hist(foxes$groupsize)
hist(foxes$area)
hist(log(foxes$groupsize))
```



Seems that the groupsize is right skewed. Yet it might not be a good choice to logarithm an integer variable with small range. The plot is no longer countinous.Perhaps we should still use groupsize instead of log(groupsize).

(1)Body weight as a linear function of territory size (area)

```
summary(foxes)
```

```
##      group           avgfood        groupsize          area       
##  Min.   : 1.00   Min.   :0.3700   Min.   :2.000   Min.   :1.090  
##  1st Qu.:11.75   1st Qu.:0.6600   1st Qu.:3.000   1st Qu.:2.590  
##  Median :18.00   Median :0.7350   Median :4.000   Median :3.130  
##  Mean   :17.21   Mean   :0.7517   Mean   :4.345   Mean   :3.169  
##  3rd Qu.:24.00   3rd Qu.:0.8000   3rd Qu.:5.000   3rd Qu.:3.772  
##  Max.   :30.00   Max.   :1.2100   Max.   :8.000   Max.   :5.070  
##      weight     
##  Min.   :1.920  
```

```
##  1st Qu.:3.720
##  Median :4.420
##  Mean   :4.530
##  3rd Qu.:5.375
##  Max.   :7.550
```

```
reg1 = map(
  alist (
    weight ~ dnorm( mu , sigma ) ,
    mu <- a + bR * area ,
    a ~ dnorm( 5 , 5 ) ,
    bR ~ dnorm( 0 , 10 ) ,
    sigma ~ dunif( 0 , 10 )
  ) , data = foxes)
precis(reg1,digits=5)
```

```
##          Mean  StdDev     5.5%   94.5%
## a     4.45428 0.38963  3.83157 5.07699
## bR    0.02387 0.11806 -0.16481 0.21254
## sigma 1.17868 0.07738  1.05501 1.30236
```

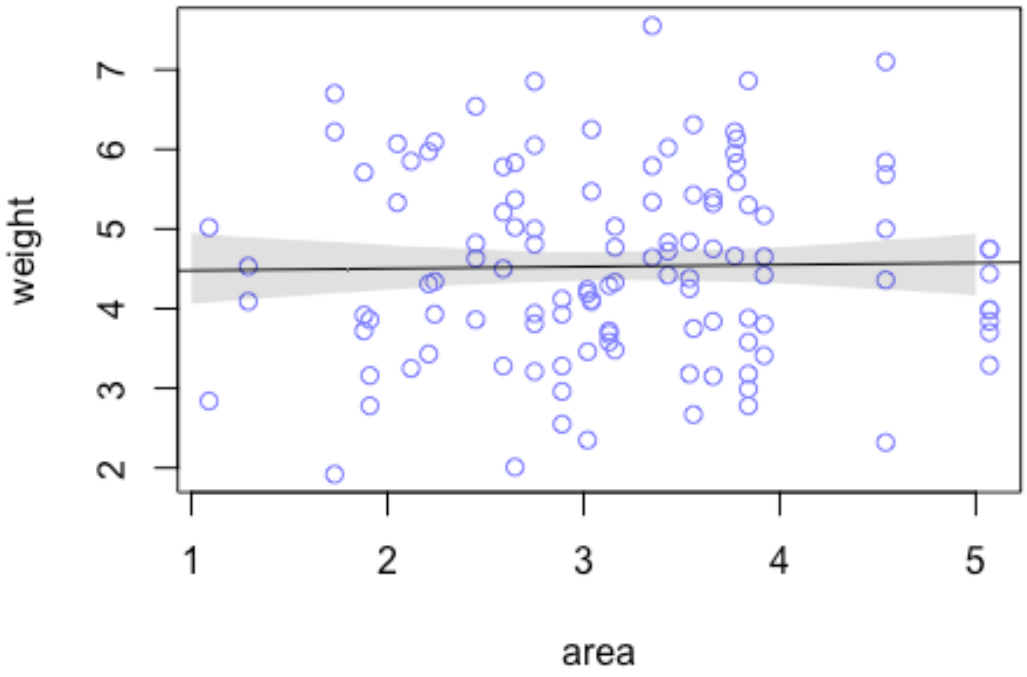
```
area.seq <- seq( from=1 , to=5 , length.out=30 )
mu <- link( reg1 , data=data.frame(area=area.seq) )
```

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

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

# plot it all
plot( weight ~ area , data=foxes , col=rangi2 )
abline( reg1 )
```

```
## Warning in abline(reg1): only using the first two of 3 regression
## coefficients
```

```
shade( mu.PI , area.seq )
```



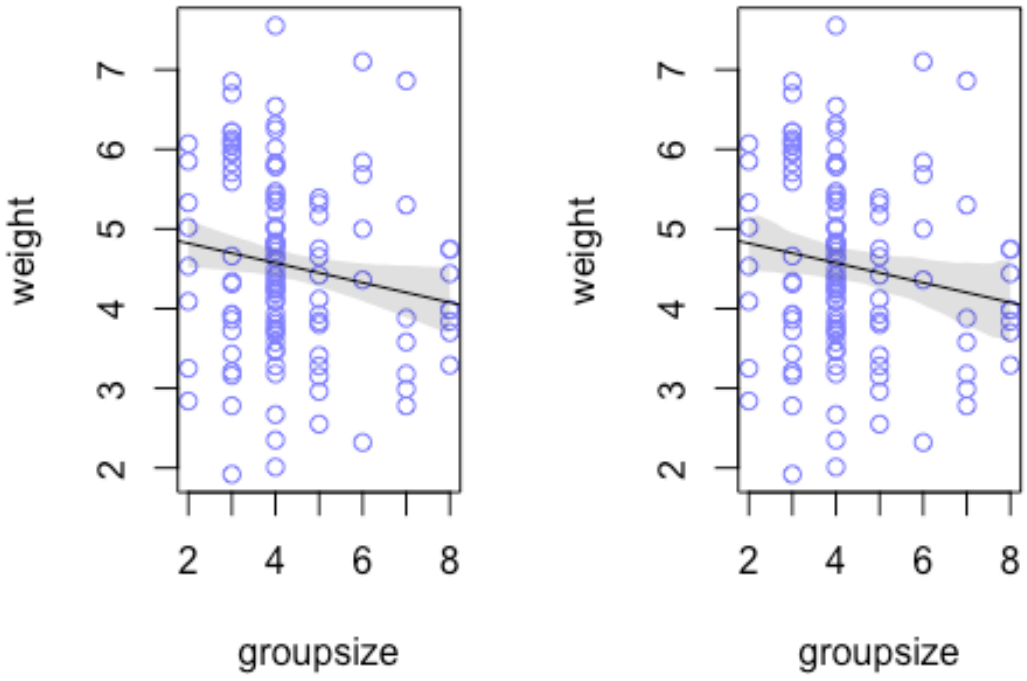
(2) body weight as a linear function of groupsize. Plot the results of these regressions, displaying the MAP regression line and the 95% interval of the mean.

```
reg2 = map(
  alist (
    weight ~ dnorm( mu , sigma ) ,
    mu <- a + bR * groupsize ,
    a ~ dnorm( 5 , 5 ) ,
    bR ~ dnorm( 0 , 10 ) ,
    sigma ~ dunif( 0 , 10 )
  ) , data = foxes)
precis(reg1,digits=5)
```

```
##           Mean  StdDev     5.5%   94.5%
## a     4.45428 0.38963  3.83157 5.07699
## bR    0.02387 0.11806 -0.16481 0.21254
## sigma 1.17868 0.07738  1.05501 1.30236
```

```
groupsize.seq <- seq( from=2 , to=8 , length.out=30 )
mu <- link( reg2 , data=data.frame(groupsize=groupsize.seq) )
```

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

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

par(mfrow=c(1,2))
# plot it all (HPDI)
plot( weight ~ groupsize , data=foxes , col=rangi2 )
abline(reg2)

## Warning in abline(reg2): only using the first two of 3 regression
## coefficients

shade(mu.PI , groupsize.seq)
mu.HPDI <- apply( mu , 2 , HPDI,prob=.95 )
# plot it all (PI)
plot( weight ~ groupsize , data = foxes , col=rangi2 )
abline(reg2)

## Warning in abline(reg2): only using the first two of 3 regression
## coefficients

shade(mu.HPDI , groupsize.seq)
```



It seems that HPDI has a bit thinner shade area than PI. (Not sure the reason)

Since the 95% of beta coeficient has covered 0, neither area nor groupsize is significant solely.

Question2

Now fit a multiple linear regression with weight as the outcome and both area and groupsize as predictor variables. Plot the predictions of the model for each predictor, holding the other predictor constant at its mean.

```
mreg1 = map(
  alist (
    weight ~ dnorm( mu , sigma ) ,
    mu <- a + b1 * groupsize+ b2 * area ,
    a ~ dnorm( 5 , 5 ) ,
    b1 ~ dnorm( 0 , 10 ) ,
    b2 ~ dnorm( 3 , 5 ) ,
    sigma ~ dunif( 0 , 10 )
  ) , data = foxes)
precis(mreg1,digits=5)
```

```
##           Mean  StdDev     5.5%    94.5%
## a      4.44939 0.36977  3.85842  5.04035
## b1    -0.43443 0.12069 -0.62732 -0.24154
## b2     0.62100 0.19981  0.30166  0.94034
## sigma  1.11845 0.07343  1.00110  1.23581
```

```
#Prepare counterfactual data
G.avg = mean(foxes$groupsize)
A.seq = seq( from = 1, to = 5.5, length.out = 30)
pred.data = data.frame(
  groupsize = G.avg,
  area = A.seq)

#Compute counterfactual mean weight (mu)
mu = link(mreg1, data = pred.data)
```

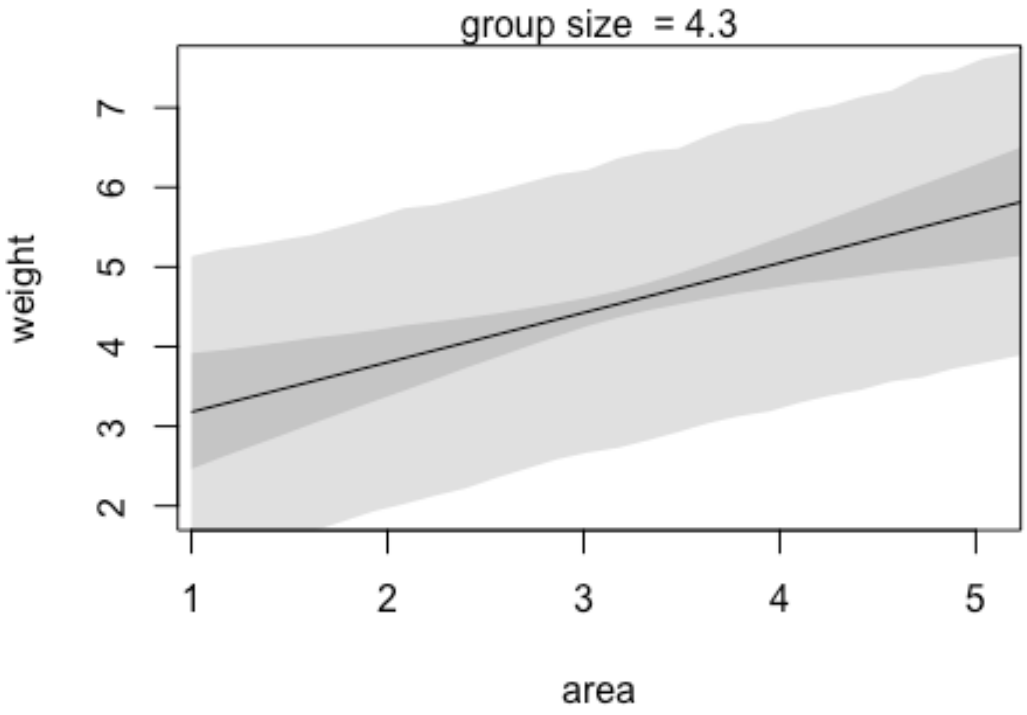
```
## [ 100 / 1000 ]
[ 200 / 1000 ]
[ 300 / 1000 ]
[ 400 / 1000 ]
[ 500 / 1000 ]
[ 600 / 1000 ]
[ 700 / 1000 ]
[ 800 / 1000 ]
[ 900 / 1000 ]
[ 1000 / 1000 ]
```

```
mu.mean =  apply(mu, 2, mean)
mu.PI = apply( mu, 2, PI)

#Simulate counterfactual weight outcomes
A.sim = sim(mreg1, data = pred.data, n=1e4)
```

```
## [ 1000 / 10000 ]
[ 2000 / 10000 ]
[ 3000 / 10000 ]
[ 4000 / 10000 ]
[ 5000 / 10000 ]
[ 6000 / 10000 ]
[ 7000 / 10000 ]
[ 8000 / 10000 ]
[ 9000 / 10000 ]
[ 10000 / 10000 ]
```

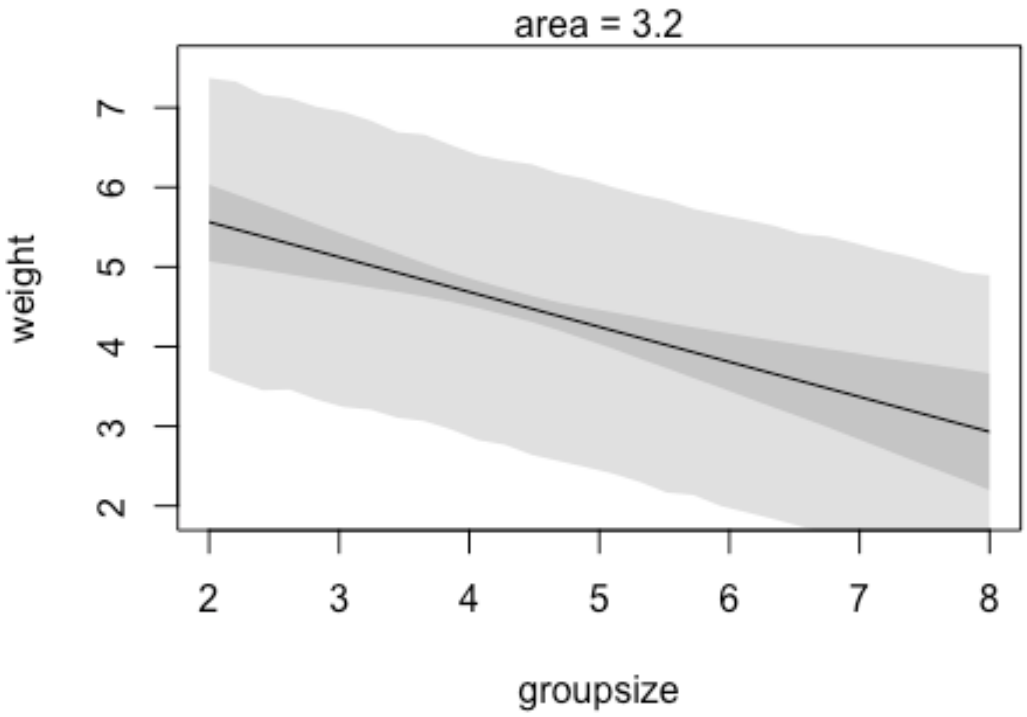
```
A.PI = apply(A.sim,2,PI)

#display predictions, hiding ras data with type = 'n'
plot( weight ~ area, data = foxes, type = 'n')
mtext('group size  = 4.3')
lines ( A.seq,mu.mean)
```

```
shade(mu.PI, A.seq)
shade(A.PI, A.seq)
```




```
#Prepare counterfactual data again for groupsize
A.avg = mean(foxes$area)
G.seq = seq( from = 2, to = 8, length.out = 30)
pred.data2 = data.frame(
  groupsize = G.seq,
  area = A.avg)

#Compute counterfactual mean weight (mu)
mu = link(mreg1, data = pred.data2)

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

```
mu.mean =  apply(mu, 2, mean)
mu.PI = apply( mu, 2, PI)

#Simulate counterfactual weight outcomes
G.sim = sim(mreg1, data = pred.data2, n=1e4)

## [ 1000 / 10000 ]
[ 2000 / 10000 ]
[ 3000 / 10000 ]
[ 4000 / 10000 ]
[ 5000 / 10000 ]
[ 6000 / 10000 ]
[ 7000 / 10000 ]
[ 8000 / 10000 ]
[ 9000 / 10000 ]
[ 10000 / 10000 ]

G.PI = apply(G.sim,2,PI)

#display predictions, hiding ras data with type = 'n'
plot( weight ~ groupsize, data = foxes, type = 'n')
mtext('area = 3.2')
lines ( G.seq,mu.mean)
shade(mu.PI, G.seq)
shade(G.PI, G.seq)
```


area = 3.2
weight
2
3
4
5
6
7
2
3
4
5
6
7
8
groupsize

What does this model say about the importance of each variable?

Controling each variable could still produce correlation with dependent variable, telling that both variables, groupsize and area are necessary and significant.

Why do you get different results than you got in the questions just above?

The contrast result of Q1 and Q2 might imply that there exists "Masked association" between predictors. That is, since 'groupsize' and 'area' has 'opposite influence' to weight in simple regressions, thus the sole effect may be eliminated by the other unobserved variables. However, adding both variables in multiregression could control the effect and produce significant outcomes for both variables.

Question3

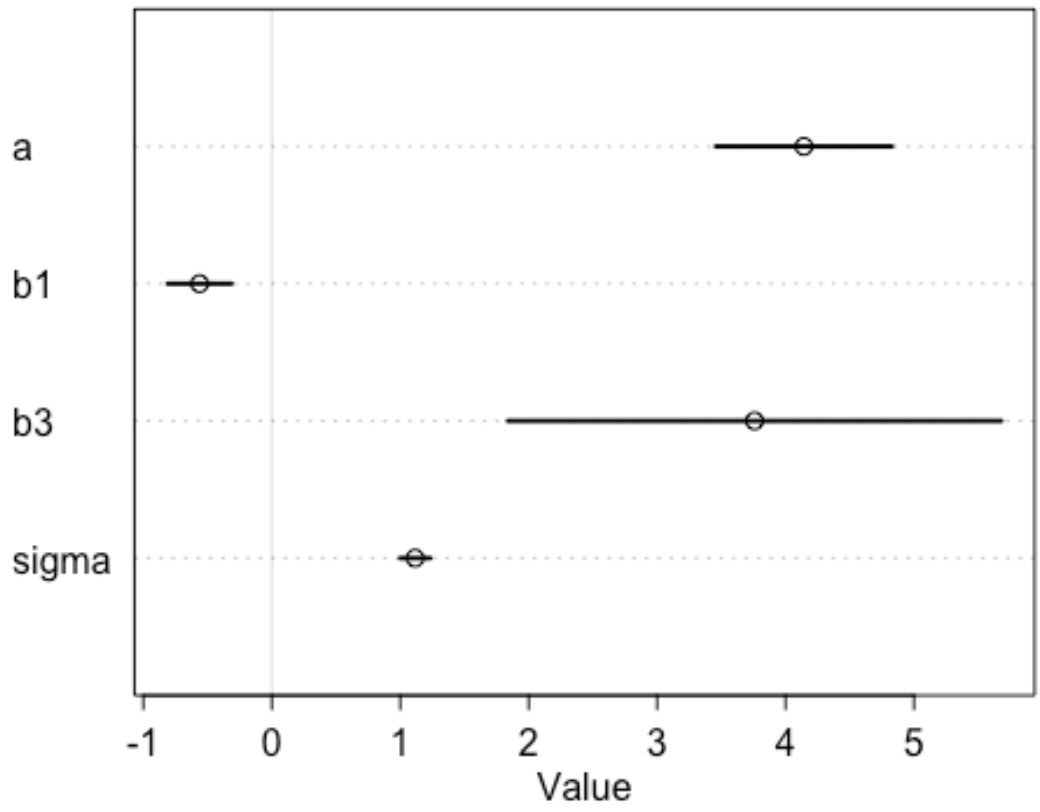
Finally consider the avgfood variable. Fit two more multiple regressions:

(1) body weight as an additive function of avgfood and groupsize

```
mreg2 = map(
  alist (
    weight ~ dnorm( mu , sigma ) ,
    mu <- a + b1 * groupsize+b3*avgfood ,
    a ~ dnorm( 5 , 5 ) ,
    b1 ~ dnorm( 0 , 10 ) ,
    b3 ~ dnorm( 0 , 10 ),
    sigma ~ dunif( 0 , 10 )
  ) , data = foxes)
precis(mreg2,digits=5)
```

```
##           Mean  StdDev     5.5%    94.5%
## a      4.14329 0.42921  3.45733  4.82924
## b1    -0.56138 0.15537 -0.80969 -0.31307
## b3     3.75915 1.20207  1.83802  5.68029
## sigma  1.11661 0.07331  0.99945  1.23378
```

```
plot(precis(mreg2,digits=5))
```

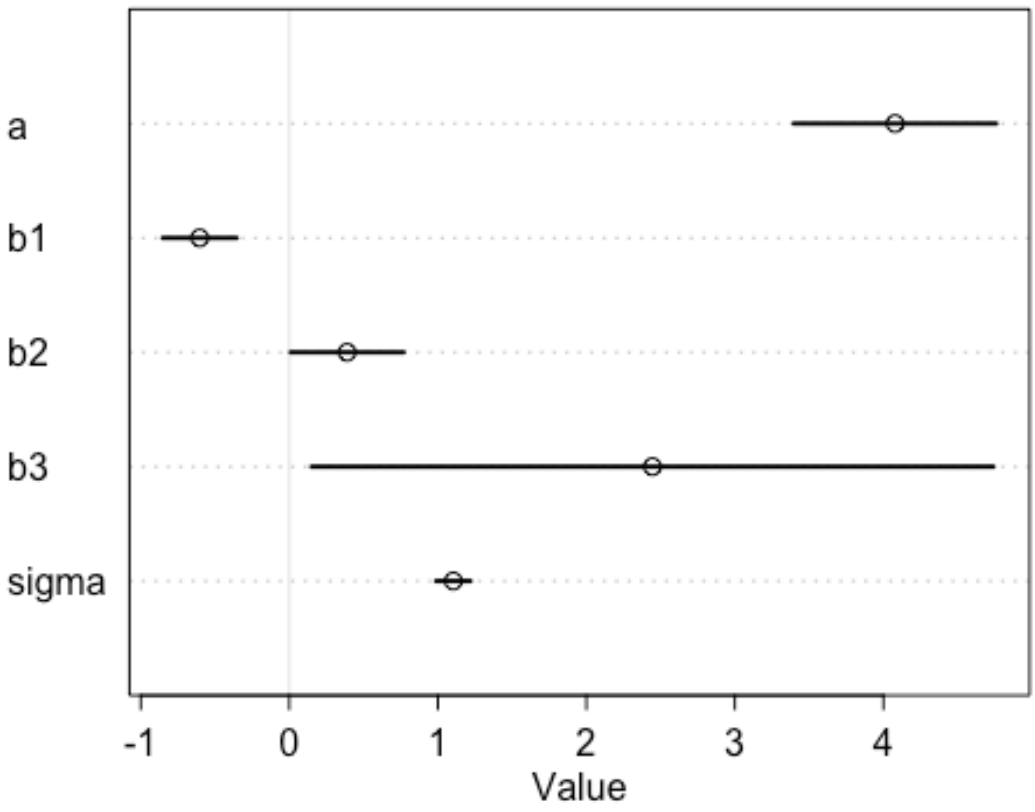


(2) body weight as an additive function of all three variables, avgfood and groupsize and area.

```
mreg3 = map(
  alist (
    weight ~ dnorm( mu , sigma ) ,
    mu <- a + b1 * groupsize+ b2*area+b3*avgfood ,
    a ~ dnorm( 5 , 5 ) ,
    b1 ~ dnorm( 0 , 10 ) ,
    b2 ~ dnorm( 0 , 10 ) ,
    b3 ~ dnorm( 0 , 10 ),
    sigma ~ dunif( 0 , 10 )
  ) , data = foxes)
precis(mreg3,digits=5)
```

```
##           Mean  StdDev     5.5%    94.5%
## a      4.07874 0.42638  3.39731  4.76017
## b1    -0.60312 0.15578 -0.85209 -0.35414
## b2     0.38926 0.23847  0.00814  0.77037
## b3     2.44522 1.43632  0.14970  4.74074
## sigma  1.10436 0.07251  0.98848  1.22024
```

```
plot(precis(mreg3,digits=5))
```



Compare the results of these models to the previous models you've fit, in the first two questions.

(a) Is avgfood or area a better predictor of body weight? If you had to choose one or the other to include in a model, which would it be? Support your assessment with any tables or plots you choose.

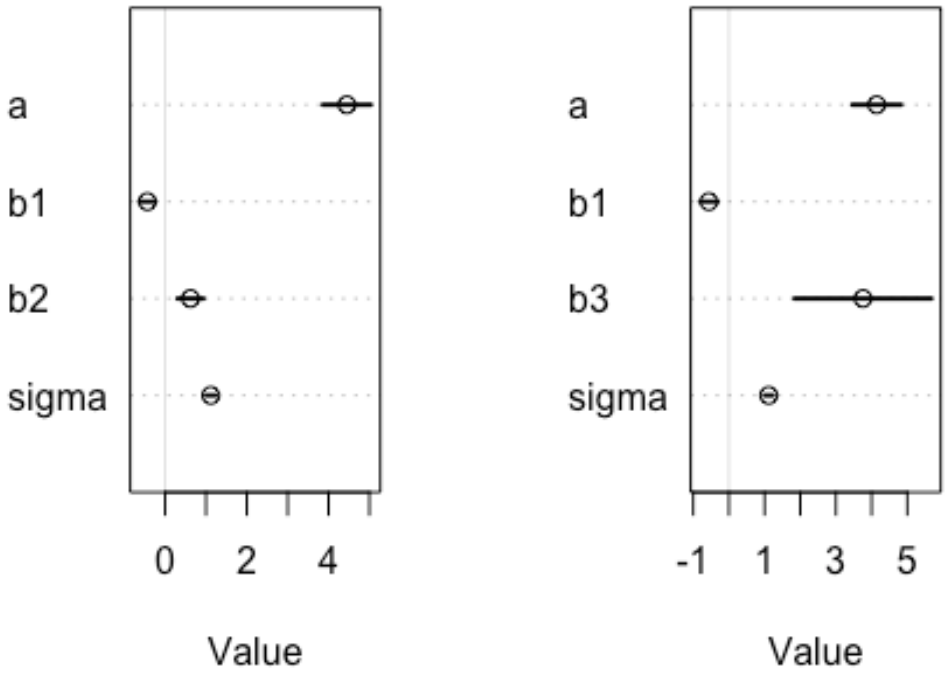
```
precis(mreg1,digits=5)
```

```
##            Mean  StdDev     5.5%    94.5%
## a       4.44939 0.36977  3.85842  5.04035
## b1     -0.43443 0.12069 -0.62732 -0.24154
## b2      0.62100 0.19981  0.30166  0.94034
## sigma   1.11845 0.07343  1.00110  1.23581
```

```
precis(mreg2,digits=5)
```

```
##            Mean  StdDev     5.5%    94.5%
## a       4.14329 0.42921  3.45733  4.82924
## b1     -0.56138 0.15537 -0.80969 -0.31307
## b3      3.75915 1.20207  1.83802  5.68029
## sigma   1.11661 0.07331  0.99945  1.23378
```

```
par(mfrow=c(1,2))
plot(precis(mreg1,digits=5))
plot(precis(mreg2,digits=5))
```

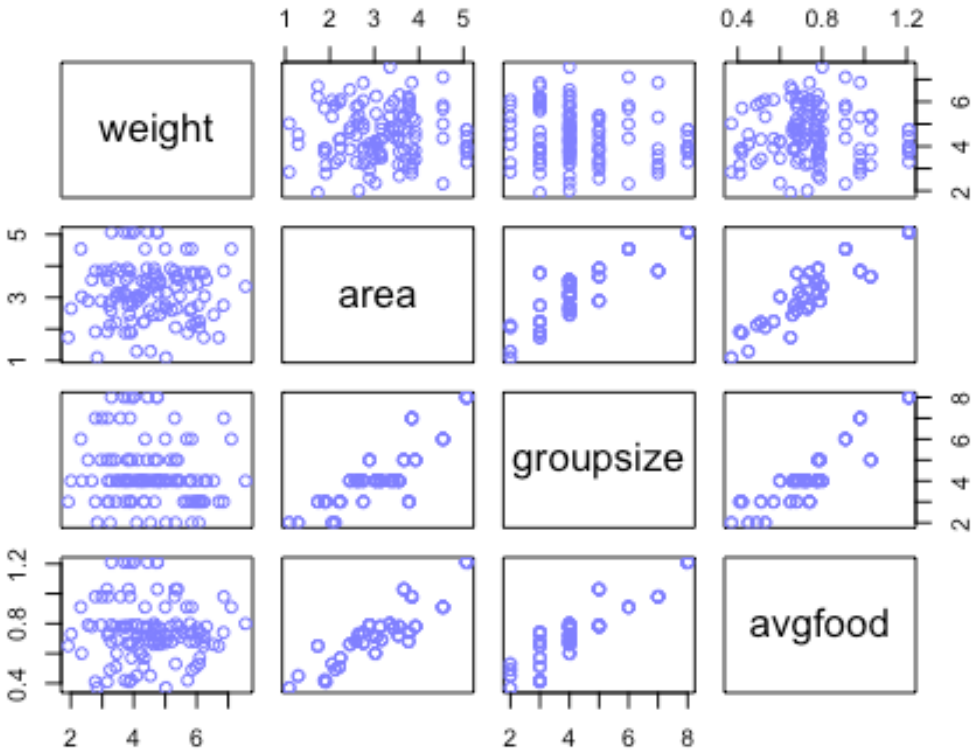


######Simply from the above graphes we could conclude that 'avgfood' is a better predictor than 'area', since the beta coeficient is larger(3.75 compare to 0.62), though with higher standard errors (1.2 compare to 0.17). Moreover, we can also find out the influence of 'groupsize' has improve by controling 'avgfood'. As a result, I would still choose 'avgfood' as predictor.

(b) When both avgfood or area are in the same model, their effects are reduced (closer to zero) and their standard errors are larger than when they are included in separate models. Can you explain this results?

One possible reason of increasing standard errors could be resulted from multicollinearity. That is, avgfood could be correlated with area.

```
pairs(~ weight + area + groupsize + avgfood, data = foxes, col=rangi2 )
```




```
cor(foxes$area,foxes$avgfood)
```

```
## [1] 0.8831038
```

From the graph and stats can find that area and avgfood are highly correlated(0.88), which is not suprising. Since the larger the terriorty is, the richer the food would be.