CSP Assignment ps3

Module: Computational Statistics & Probability - MADS
Professor: Dr. Gregory Wheeler
Student: Haowei Lee
Textbook: Statistical Rethinking 2nd edition
* The pdf file is produced by knitting rmd into html and export the result as pdf.

Import libraries

```
library(rethinking)
library(ggplot2)
```

Question 1

Import and inspect foxes data

```
data(foxes)
d <- foxes
precis(d)
```

```
##                 mean        sd    5.5%   94.5%     histogram
## group     17.2068966 8.0027357 4.00000 28.6750
## avgfood    0.7517241 0.1983158 0.42000  1.2100
## groupsize  4.3448276 1.5385111 2.00000  8.0000
## area       3.1691379 0.9283539 1.77875  5.0700
## weight     4.5296552 1.1840226 2.78000  6.2905
```

a)

Does territory size have a causal influence the weight of foxes? Construct a quap model to infer the total causal influence of area on weight. Does increasing the area available to each fox make it healthier (i.e., heavier)? I recommend that you standardize your variables and use prior predictive simulation to show that your models predictions stay within the possible outcome range.

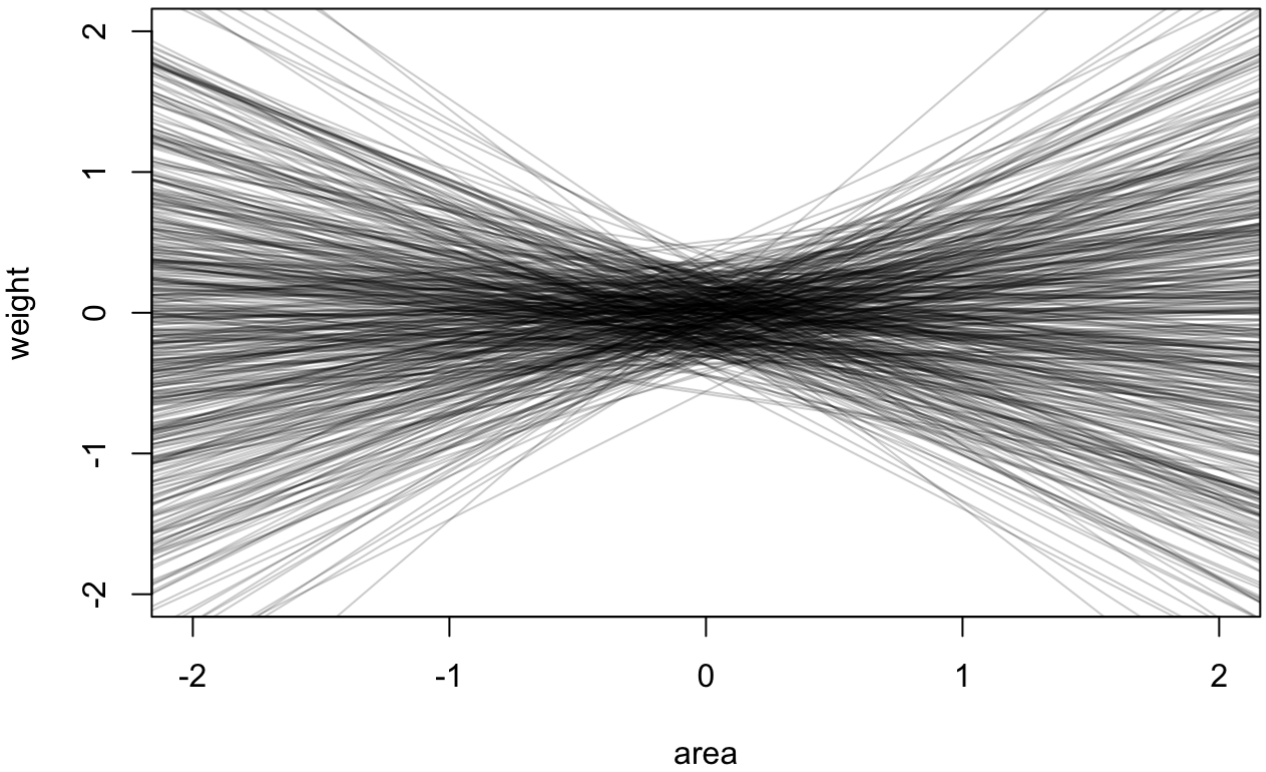
```
# standardize the variables
d$w_s <- standardize(d$weight)
d$a_s <- standardize(d$area)
```

```
# build a linear model
modelA <- quap(
  alist(
    w_s ~ dnorm( mu , sigma ),  # weight
    mu <- a + b_area * a_s,
    a ~ dnorm(0,0.2),
    b_area ~ dnorm(0,0.5),
    sigma ~ dexp(1)
  ), data=d )

precis(modelA)   # inspect the result
```

```
##                 mean         sd       5.5%     94.5%
## a      -6.922871e-09 0.08360867 -0.1336228 0.1336228
## b_area  1.883359e-02 0.09089582 -0.1264355 0.1641027
## sigma   9.912661e-01 0.06466649  0.8879166 1.0946157
```

b_area, the coefficient of area, is around 0.018, meaning that territory size barely has a casual influence on weight.

```
set.seed(2971)
N <- 500
priors <- extract.prior( modelA )
a <- priors$a           # a ~ dnorm( 178 , 20 )
b_area <- priors$b_area         # b1 ~ dlnorm( 0 , 1 )

plot( NULL , xlim=c(-2,2) , ylim=c(-2,2) ,
      xlab="area" , ylab="weight" )
# mtext( "b ~ dnorm(0,10)" )
xbar <- mean(d$w_s)
for ( i in 1:N ) curve( a[i] + b_area[i]*(x - xbar) ,
                        from=min(d$w_s) , to=max(d$w_s) , add=TRUE ,
                        col=col.alpha("black",0.2) )
```



b)

Now infer the causal impact of adding food (avgfood) to a territory. Would this make foxes heavier? Which covariates do you need to adjust to estimate the total causal influence of food?

```
d$f_s <- standardize(d$avgfood)
modelB <- quap(
  alist(
    w_s ~ dnorm( mu , sigma ),
    mu <- a + b_food * f_s,
    a ~ dnorm(0,0.2),
    b_food ~ dnorm(0,0.5),
    sigma ~ dexp(1)
  ), data=d )

precis(modelB)   # inspect the result
```

```
##                 mean         sd       5.5%     94.5%
## a       4.044263e-08 0.08360009 -0.1336091 0.1336091
## b_food -2.421164e-02 0.09088492 -0.1694633 0.1210400
## sigma   9.911429e-01 0.06465840  0.8878062 1.0944795
```

b_food, the coefficient of food, is around -0.024, meaning that average food barely has a casual influence on weight.

c)

Now infer the causal impact of group size (groupsize). Which covariates do you need to adjust to make this estimate? Inspect the posterior distribution of theresulting model. What do you think explains these data? Specifically, explain the estimates of the effects of area, avgfood, and groupsize on weight. How do they make sense together? (Hint: we covered an example in class which exhibited a similar relationship between predictors and outcome variable.)

```
d$g_s <- standardize(d$groupsize)
modelC <- quap(
  alist(
    w_s ~ dnorm( mu , sigma ),
    mu <- a + b_food * f_s + b_group * g_s,
    a ~ dnorm(0,0.2),
    b_food ~ dnorm(0,0.5),
    b_group ~ dnorm(0,0.5),
    sigma ~ dexp(1)
  ), data=d )

precis(modelC)   # inspect the result
```

```
##                  mean         sd       5.5%      94.5%
## a       -3.102318e-06 0.08014636 -0.1280925  0.1280863
## b_food   4.778414e-01 0.17913510  0.1915490  0.7641339
## b_group -5.741556e-01 0.17915304 -0.8604768 -0.2878345
## sigma    9.421601e-01 0.06177153  0.8434372  1.0408829
```

With the mean of coefficient being -0.57 and 0.48 for group and food, respectively, we can know that group have a negative causal impact on weight, controlling food. In contrast, food has a positive causal effect on weight. To explain, when the size increase in a group of fox, each fox will lose weight. And if average food increases, and only if the group size keeps the same, foxes will gain weight. However, through the analysis on part a, we know that the total casual impact of average food on fox is hardly anything. This implies that there is a masking effect.

This effect could exist between area and average food. When average food increase in an area, foxes tend to move to that area and will soon wash out the advantages on food supply. This kind of active change is like the natural system constantly changing something to reach equilibrium and ensure no place is better than other places in the long run.

Question 2

Explain the difference between model selection and model comparison. What information is lost under model selection?

The main difference between model selection and comparison is that model selection picks the model with the best information criteria value without considering the relative performance of other models, while model comparison does.

To explain, model selection applied different strategies, either cross-validation or information criteria, and determined the accuracy of models by choosing different information criterion such as AIC, DIC, or WAIC. The final model picked would be the one with the best score by using these methods. Doing this will lose information about the relative accuracy of other models.

However, model comparison cares about the relative accuracy. And by observing the standard error of the difference in WAIC between models, we can know how different a model is from the other. Thus, using model comparison, we keep information on relative accuracy and could conduct a more comprehensive analysis of the models and data.

Question 3

Use WAIC or LOO based model comparison on five different models, each using weight as the outcome, and containing the follow sets of predictor variables:

1. avgfood + groupsize + area
2. avgfood + groupsize
3. avgfood + area
4. avgfood
5. area

Can you explain the relative differences in WAIC scores, using the fox DAG from above? Be sure to pay attention to the standard error of the score differences (dSE).

Load the foxes data and store into a DataFrame

```
# load the data
data(foxes)
d <- foxes
d$area <- scale(d$area)
d$avgfood <- scale(d$avgfood)
d$weight <- scale(d$weight)
d$groupsize <- scale(d$groupsize)
```

Construct models

```
# Construct the models
# (1) avgfood + groupsize + area
m1 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + bFood * avgfood + bGroup * groupsize + bArea * area,
    a ~ dnorm(0, 0.2),
    c(bFood, bGroup, bArea) ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = d
)

# (2) avgfood + groupsize
m2 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + bFood * avgfood + bGroup * groupsize,
    a ~ dnorm(0, 0.2),
    c(bFood, bGroup) ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = d
)

# (3) avgfood + area
m3 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + bFood * avgfood + bArea * area,
    a ~ dnorm(0, 0.2),
    c(bFood, bArea) ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = d
)

# (4) avgfood
m4 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + bFood * avgfood,
    a ~ dnorm(0, 0.2),
    bFood ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data = d
)

# (5) area
m5 <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + bArea * area,
    a ~ dnorm(0, 0.2),
    bArea ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
```

```
  ),
  data = d
)
```

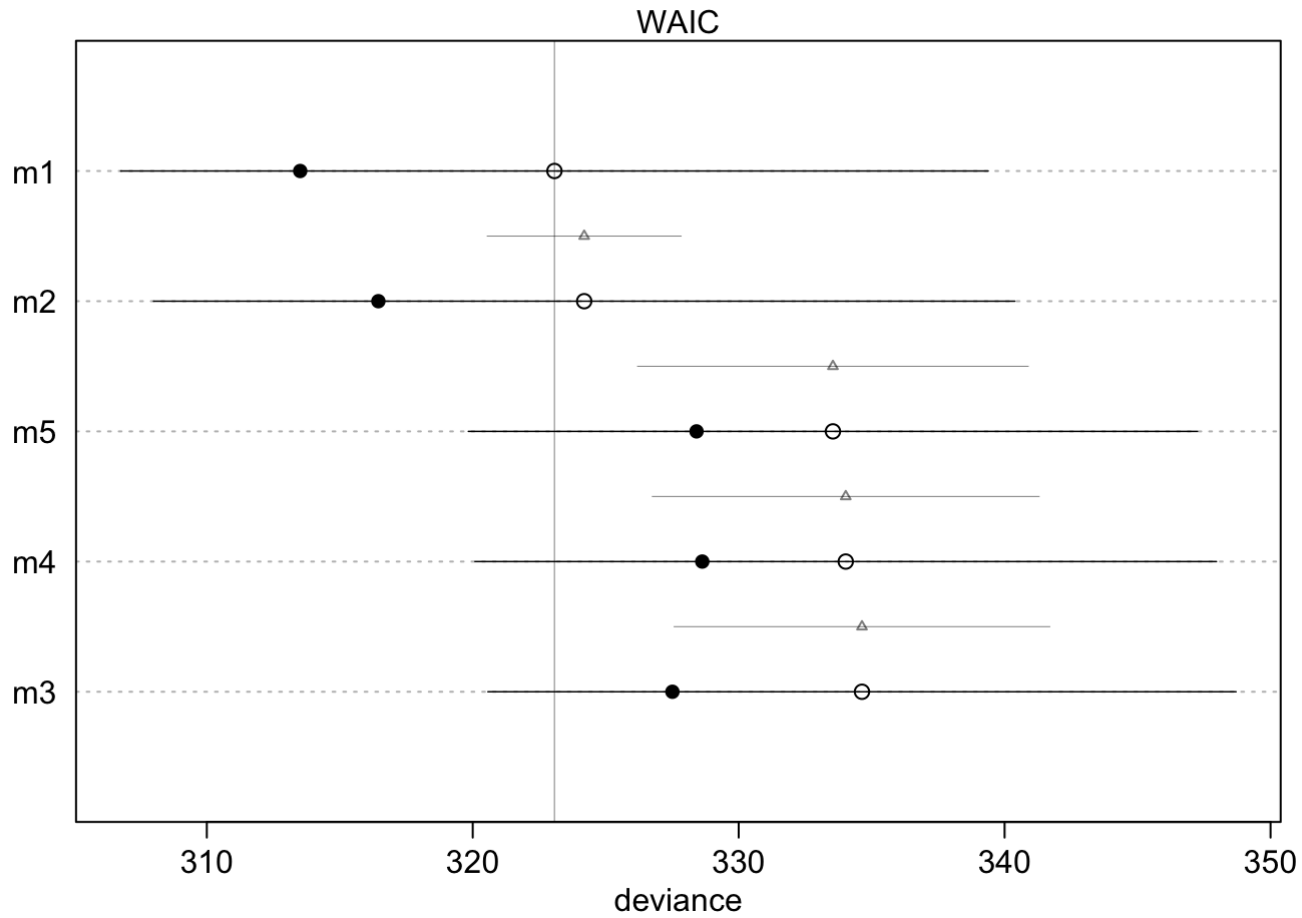
Comparison of models

```
compare(m1, m2, m3, m4, m5)
```

```
##        WAIC       SE      dWAIC      dSE    pWAIC      weight
## m1 323.2172 16.52024  0.0000000       NA 4.871115 0.598414958
## m2 324.0568 15.99376  0.8396345 3.586498 3.779314 0.393258499
## m4 333.6361 13.77807 10.4189151 7.411391 2.513797 0.003270121
## m5 333.6807 13.74738 10.4635410 7.476700 2.627366 0.003197963
## m3 334.7663 14.08982 11.5490734 7.214626 3.648052 0.001858459
```

Plot out the comparsion

```
## Comparison plot
plot(compare(m1, m2, m3, m4, m5))
```

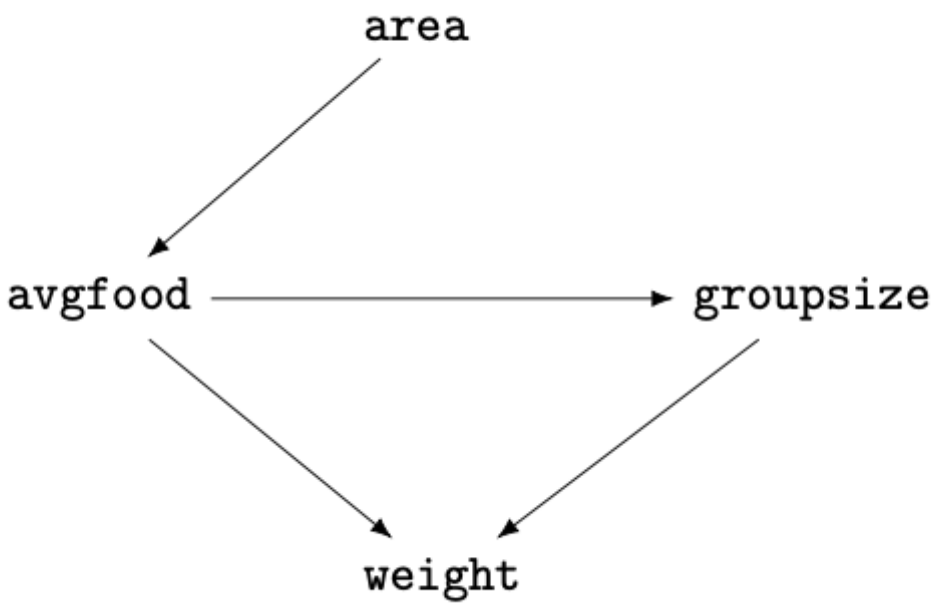


Overall, the WAIC values are very similar. As dSE implies the standard error of the score differences, we can compare the numbers among all models. While dSE of m2 is closer to m1, dSE of m3, m4, and m5 are closer. In addition, plotting out the comparison also helps us put those models into two different groups of model:

Group A: m1 and m2

Group B: m3 and m4 and m5

To understand the reason behind this, we use the DAG provided in question 1 to explore the relationship among the variables.



In Group A, the first two models (m1 and m2) contain groupsize and average food. Since the casual effect of the area goes directly through average food, it is very similar to just considering average food. Therefore, an extra variable area of m1 does not make a lot of difference from m2. The reason of the similarity of these two models is that there is no back-door path from area or avgfood to weight. In other words, the effect of area affecting groupsize is the same as the effect of avgfood affecting groupsize, because the effect of area goes directly and entirely through avgfood.

In Group B, the last three models (m3, m4 and m5) are also almost identical due to the relationship between area to avgfood. Because the effect of area is routed entirely through avgfood, including only avgfood or area or both should result in the same casual inferences.