ps5-1

Haowei Lee

12/10/2021

CSP Assignment ps5

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.
* The assignment is split into three Rmd files and the pdf was created by putting the results of all three Rmd files. The reason is that the training time for each MCMC model takes a long time and calculation, making my laptop super hot.

Import libraries

```
library(rethinking)
library(dagitty)
library(cmdstanr)
```

Import and inspect the dataset

```
options(digits = 2) #set to two decimal
data(Trolley)
d <- Trolley
precis( d )
```

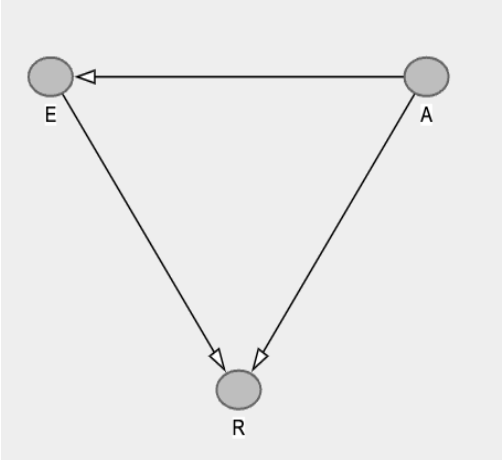
```
##             mean    sd 5.5% 94.5%     histogram
## case         NaN    NA   NA    NA
## response    4.20  1.91    1     7
## order      16.50  9.29    2    31
## id           NaN    NA   NA    NA
## age        37.49 14.23   18    61
## male        0.57  0.49    0     1
## edu          NaN    NA   NA    NA
## action      0.43  0.50    0     1
## intention   0.47  0.50    0     1
## contact     0.20  0.40    0     1
## story        NaN    NA   NA    NA
## action2     0.63  0.48    0     1
```

Question 1

a) Draw a DAG that represents hypothetical causal relationships among response, education, and age.

The DAG was built on dagitty.net with the model code below, and then screenshot as a picture.

DAG1



E Education
A Age
R Response

```
# Model code to create the DAG on on dagitty.net
dag {
  bb="-1.7,-1.95,2.1,2.1"
  A [pos="0.779,-0.118"]
  E [pos="-0.714,-0.118"]
  R [pos="0.033,1.100"]
  A -> E
  A -> R
  E -> R
}
```

b) Identify which statistical model or models are required to evaluate the causal influence of education on responses.

```
dag <- dagitty("dag{
    E -> R <- A
    A -> E
}")
adjustmentSets( dag , exposure="E" , outcome="R" , effect="total" )
```

```
## { A }
```

To study the impact on R from E, we should block all possible back doors from E to R through A because there might be a hidden one. To achieve our goal, we need to condition on A, as the code shown above.

Start building the model

In the Trolley dataset, we have 8 categories of education level. The code below reorder them in the proper order.

```
# reorder and transform the categorical variable: education level
edu_levels <- c( 6 , 1 , 8 , 4 , 7 , 2 , 5 , 3 )
d$edu_new <- edu_levels[ d$edu ]
```

Build data list containing only features / predictors that will be used in our model.

```
data <- list(
  R = d$response,
  action = d$action,
  intention = d$intention,
  contact = d$contact,
  E = as.integer( d$edu_new ),

  edu_norm = normalize( d$edu_new),  # normalize education to ensure values only ranges
from 0 to 1
  age = standardize( d$age),         # standardize age

  alpha = rep(2,7)                   # delta prior
)
```

Build the model (The model was built referring to code block 12.34 in the textbook.)

```
model_Q1 <- ulam(
  alist(
    R ~ ordered_logistic( phi , kappa ),
    phi <- bE*sum( delta_j[1:E] ) + bA*action + bC*contact + BI*intention + bAge*age,
    BI <- bI + bIA*action + bIC*contact ,
    c(bA,bI,bC,bIA,bIC,bE,bAge) ~ normal( 0 , 0.5 ),
    kappa ~ normal( 0 , 1.5 ),
    vector[8]: delta_j <<- append_row( 0 , delta ),
    simplex[7]: delta ~ dirichlet( alpha )
  ), data=data , chains=4 , cores=4, cmdstan=TRUE)
```

Review the training process

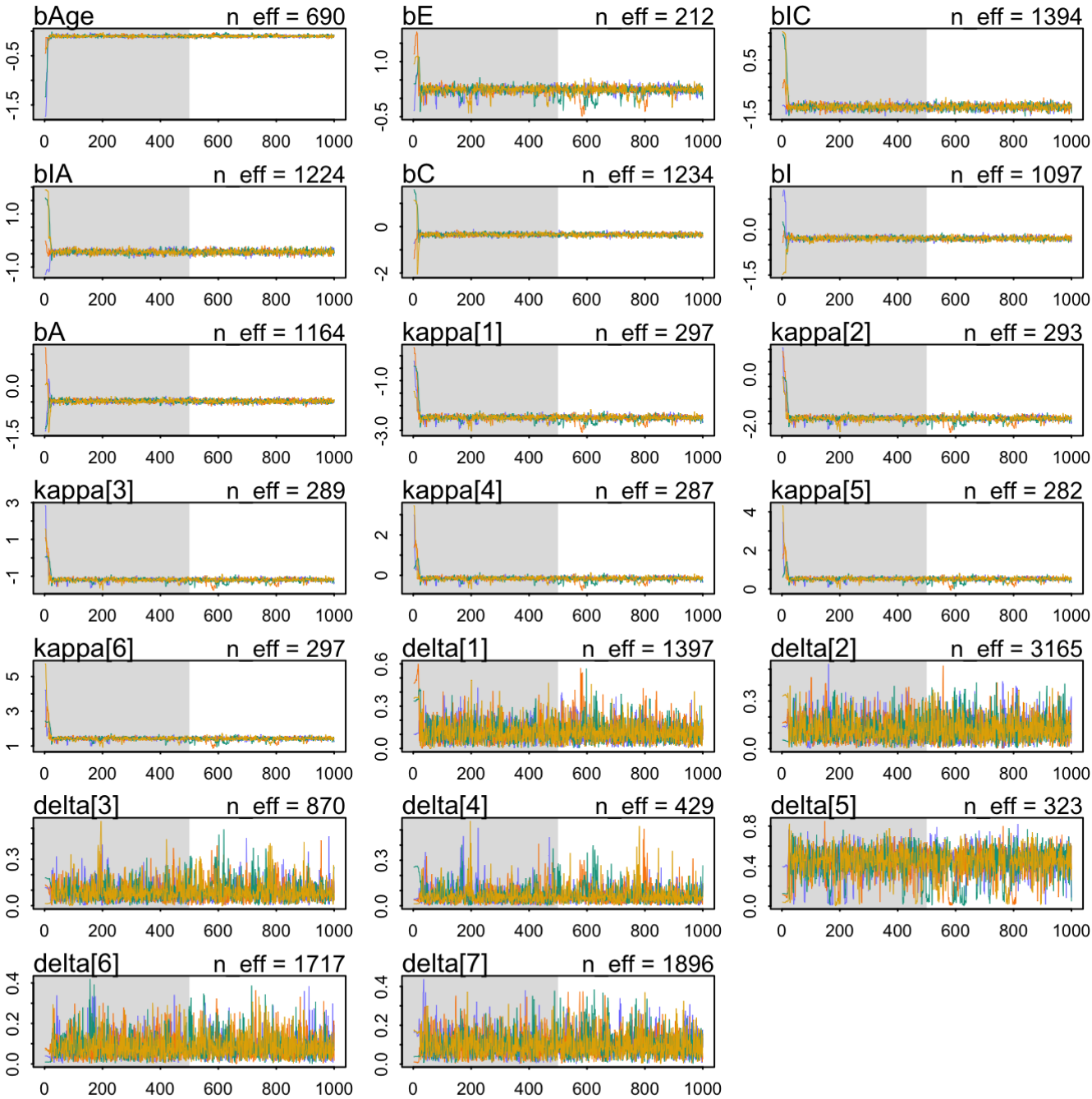
```
show(model_Q1)
```

```
## Hamiltonian Monte Carlo approximation
## 2000 samples from 4 chains
##
## Sampling durations (minutes):
##         warmup sample total
## chain:1    6.4    1.9   8.2
## chain:2    5.6    1.9   7.4
## chain:3    5.8    2.1   7.8
## chain:4    6.4    1.9   8.3
##
## Formula:
## R ~ ordered_logistic(phi, kappa)
## phi <- bE * sum(delta_j[1:E]) + bA * action + bC * contact +
##     BI * intention + bAge * age
## BI <- bI + bIA * action + bIC * contact
## c(bA, bI, bC, bIA, bIC, bE, bAge) ~ normal(0, 0.5)
## kappa ~ normal(0, 1.5)
## vector[8]:delta_j <<- append_row(0, delta)
## simplex[7]:delta ~ dirichlet(alpha)
```

Plot out the MCMC chains

```
traceplot(model_Q1)
```

```
## [1] 1000
## [1] 1
## [1] 1000
```



c) What do you conclude about the causal relationships among these three variables?

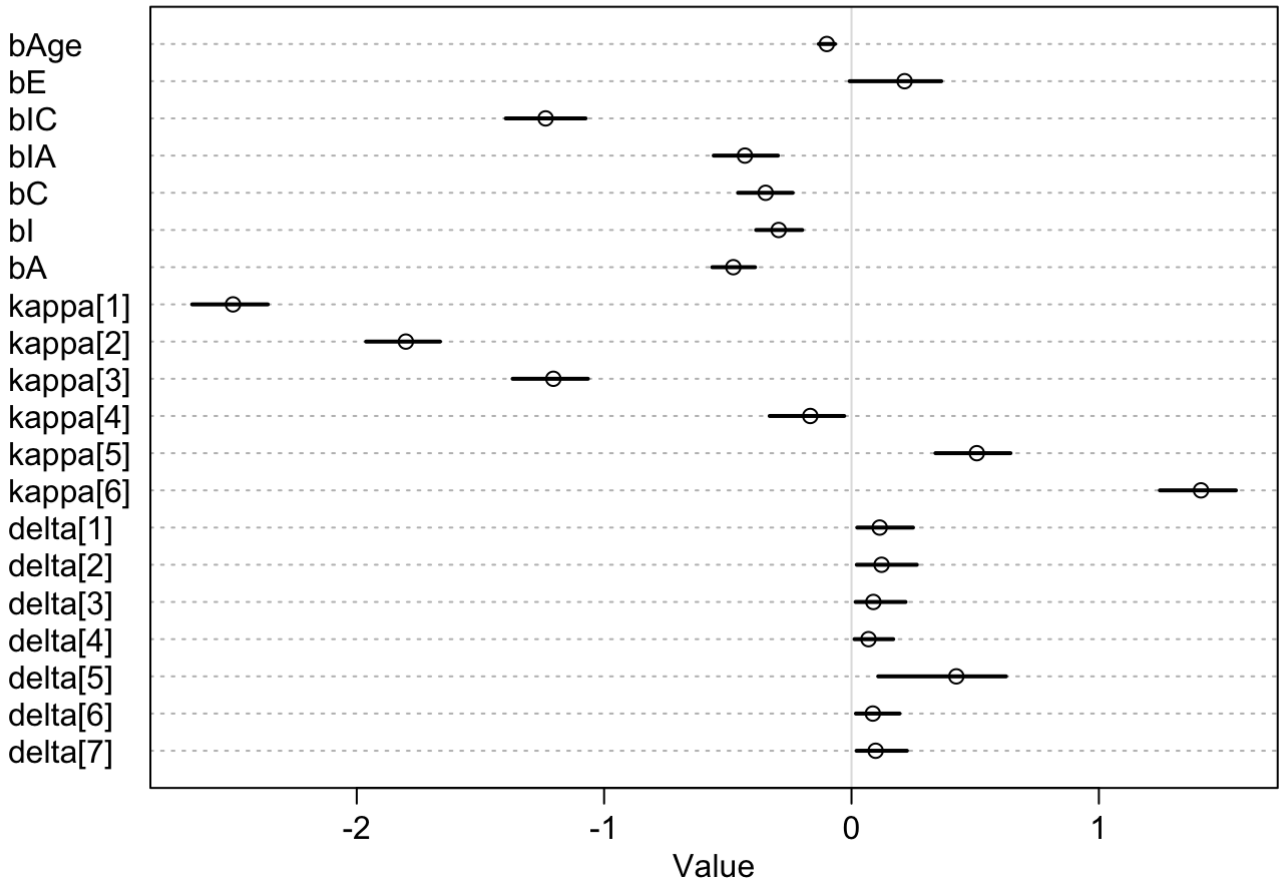
Inspect the result

```
options(digits = 2) #set to two decimal
precis(model_Q1)
```

```
##       mean    sd    5.5%  94.5% n_eff Rhat4
## bAge -0.10 0.021 -0.1322 -0.067   690     1
## bE    0.21 0.123 -0.0074  0.363   212     1
## bIC  -1.24 0.100 -1.3975 -1.075  1394     1
## bIA  -0.43 0.082 -0.5571 -0.297  1224     1
## bC   -0.35 0.069 -0.4587 -0.238  1234     1
## bI   -0.29 0.059 -0.3856 -0.199  1097     1
## bA   -0.48 0.054 -0.5628 -0.390  1164     1
```

From the traceplot in section b, we can see that the chains are converging. However, the n_eff (effective samples) are far away from the 2000 samples we have. On the other hand, the Rhats are all one, and that is a good sign.

```
plot(precis(model_Q1, 2))
```



By observing the result of precis and its plot, we can conclude the following:

- Age has a slightly negative impact on the response, for the mean at -0.10. Because the effect is so small, this should be some effect other than a casual effect.
- Education has a positive impact on the response, for the mean at -0.23. This implies that education could affect people to consider the action, intention and contact more acceptable.

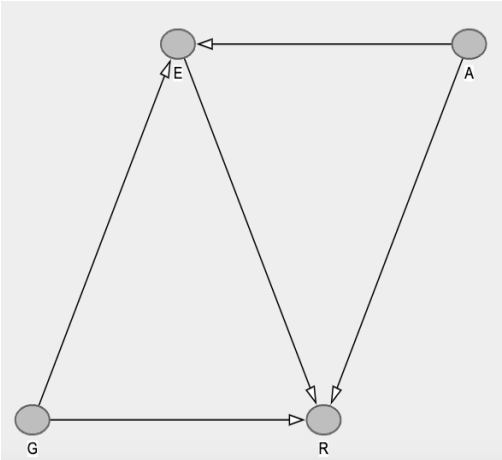
ps5-2

Haowei Lee

12/10/2021

Question 2

a) Draw a DAG DAG2




```
# Model code to create the DAG on on dagitty.net
dag {
  bb="-1.7,-1.95,2.1,2.1"
  A [pos="0.779,-0.118"]
  E [pos="-0.714,-0.118"]
  G [pos="-1.460,1.791"]
  R [pos="0.033,1.791"]
  A -> E
  A -> R
  E -> R
  G -> E
  G -> R
}
```

To find out which variable should we condition on to block back doors casual influence.

```
dag2 <- dagitty("dag{
    E -> R <- A
    A -> E
    G -> E
    G -> R
}")
adjustmentSets( dag2 , exposure="E" , outcome="R" , effect="total" )
```

```
## { A, G }
```

Just as problem 1, we should also condition on A(age). And so should we do the same on G(gender). We never know if there is any hidden back door for casual influence among the variables. So, the condition on all variables in the DAG would be a save strategy.

b) Define any additional models you need to infer the causal influence of education on response.

Applied One-Hot-Encoding to transform categorical variable

```
data$female <- ifelse( d$male==1 , 0L , 1L )
```

Build the model (The model was built referring to code block 12.34 in the textbook.)

```
model_Q2 <- ulam(
    alist(
        R ~ ordered_logistic( phi , kappa ),
        phi <- bE*sum( delta_j[1:E] ) + bA*action + bC*contact + BI*intention +
               bAge*age + bF*female,
        BI <- bI + bIA*action + bIC*contact ,
        c(bA,bI,bC,bIA,bIC,bE,bAge,bF) ~ normal( 0 , 0.5 ),
        kappa ~ normal( 0 , 1.5 ),
        vector[8]: delta_j <<- append_row( 0 , delta ),
        simplex[7]: delta ~ dirichlet( alpha )
    ), data=data , chains=4 , cores=4, cmdstan=TRUE)
```

Review the training process

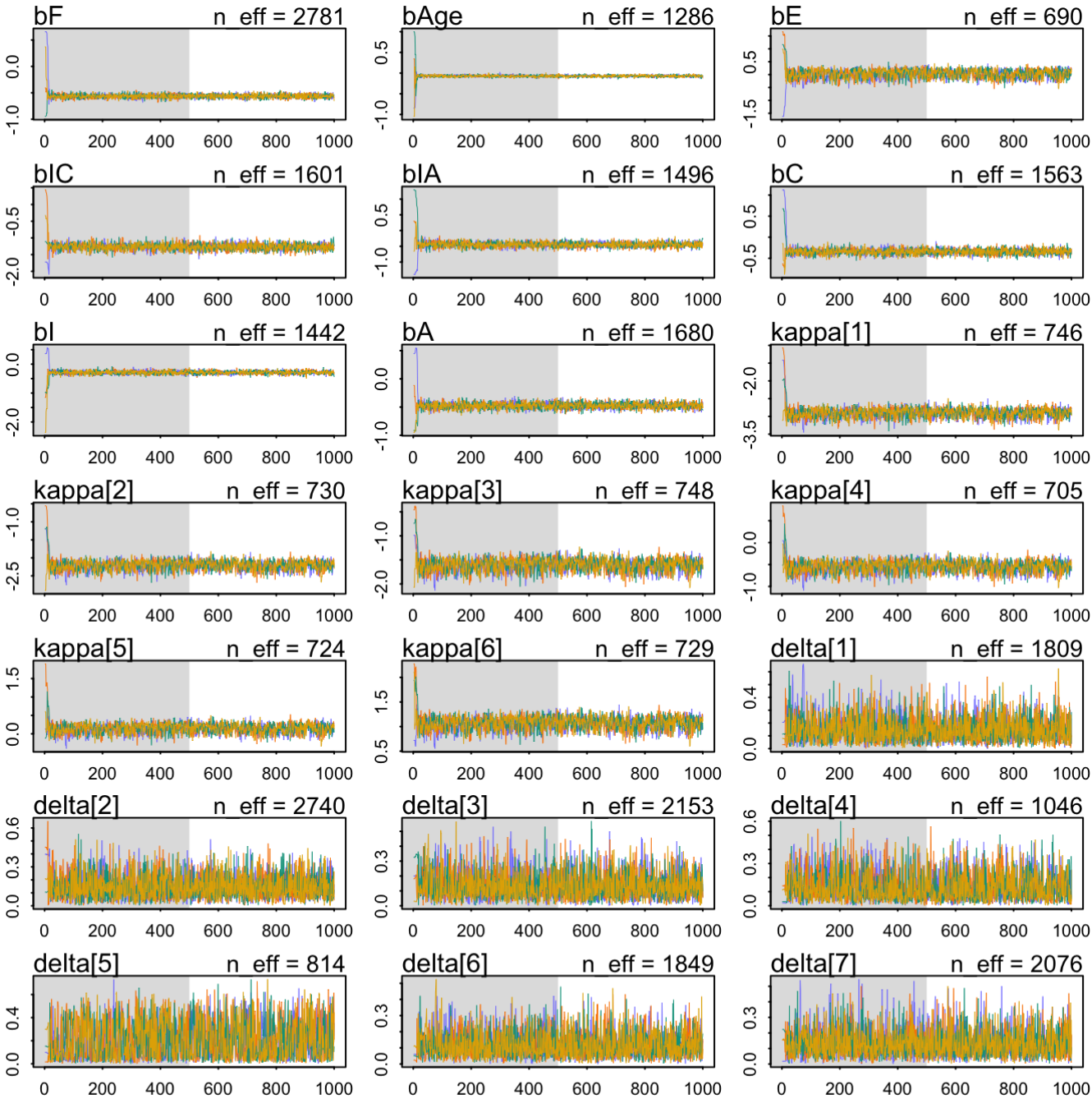
```
show(model_Q2)
```

```
## Hamiltonian Monte Carlo approximation
## 2000 samples from 4 chains
##
## Sampling durations (minutes):
##         warmup sample total
## chain:1   7.73   4.03 11.76
## chain:2   8.05   4.06 12.11
## chain:3   7.55   3.87 11.42
## chain:4   8.14   3.66 11.79
##
## Formula:
## R ~ ordered_logistic(phi, kappa)
## phi <- bE * sum(delta_j[1:E]) + bA * action + bC * contact +
##     BI * intention + bAge * age + bF * female
## BI <- bI + bIA * action + bIC * contact
## c(bA, bI, bC, bIA, bIC, bE, bAge, bF) ~ normal(0, 0.5)
## kappa ~ normal(0, 1.5)
## vector[8]:delta_j <<- append_row(0, delta)
## simplex[7]:delta ~ dirichlet(alpha)
```

Plot out the MCMC chains

```
traceplot(model_Q2)
```

```
## [1] 1000
## [1] 1
## [1] 1000
```



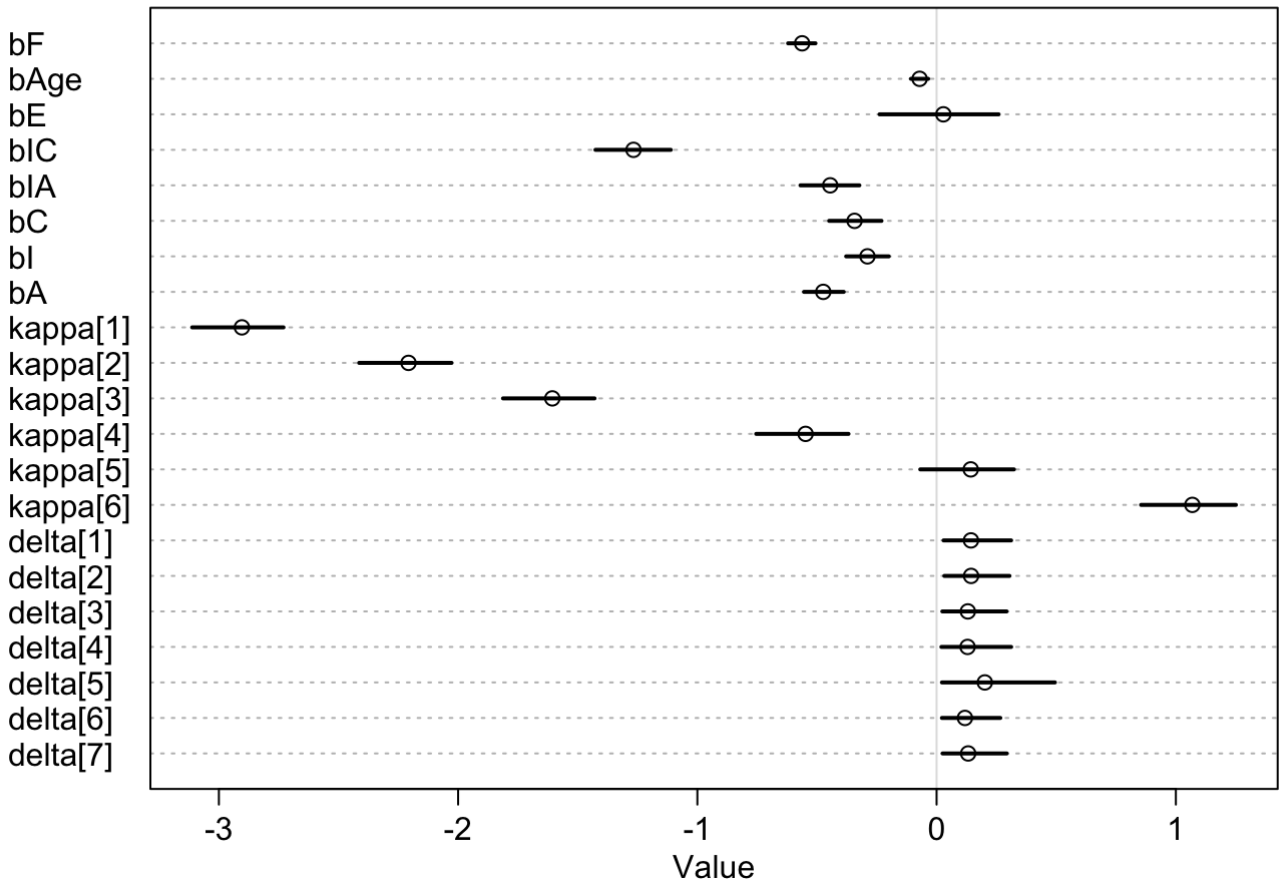
c) What do you conclude?

Inspect the result

```
options(digits = 2) #set to two decimal
precis(model_Q2)
```

```
##        mean    sd  5.5%  94.5% n_eff Rhat4
## bF   -0.562 0.036 -0.62 -0.506  2781     1
## bAge -0.071 0.023 -0.11 -0.035  1286     1
## bE    0.028 0.160 -0.24  0.259   690     1
## bIC  -1.267 0.098 -1.43 -1.111  1601     1
## bIA  -0.445 0.077 -0.57 -0.323  1496     1
## bC   -0.343 0.068 -0.45 -0.230  1563     1
## bI   -0.289 0.055 -0.38 -0.200  1442     1
## bA   -0.474 0.052 -0.55 -0.388  1680     1
```

```
plot(precis(model_Q2, 2))
```



By observing the result of precis and its plot, we can conclude the following:

- Age has a slightly negative impact on the response, for the mean at around -0.07.
- Gender negatively influences response, with a mean around -0.56. This implies that females tend to give lower responses. However, there is one issue that we should take into consideration. That is, whether we have unbalanced data in terms of age and gender. For example, even though the total number of male and female respondents could be the same, they could be distributed differently, making that in a specific age range, there are more women or men, affecting the result of causal inference.
- Education: Comparing with the first model, the coefficient of education in this model drops to 0.03 that is almost zero, showing trivial effect on the response

ps5-3

Haowei Lee

12/10/2021

Question 3

Rewrite the following model as a multilevel model.

$$
\begin{aligned}
y_i &\sim \text{Binomial}(1, p_i) \\
\text{logit}(p_i) &= \alpha_{\text{group}[i]} + \beta x_i \\
\alpha_{\text{group}} &\sim \text{Normal}(0, 1.5) \\
\beta &\sim \text{Normal}(0, 0.5)
\end{aligned}
$$

```
y_i ~ Binomial(1,p_i)
logit(p_i) = a[i] + b*x_i
a[i] ~ Normal(a_bar, sigma_a)
b ~ Normal(0,0.5)
a_bar ~ Normal(0, 1.5)    # extra parameters, hyper-priors, to construct a multi-level
model
sigma_a ~ Exponential(1)
```