Assignment 5

Jasmin Capka

12/08/2021

As basis for this assignment, I loaded the `Trolley` dataset:

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

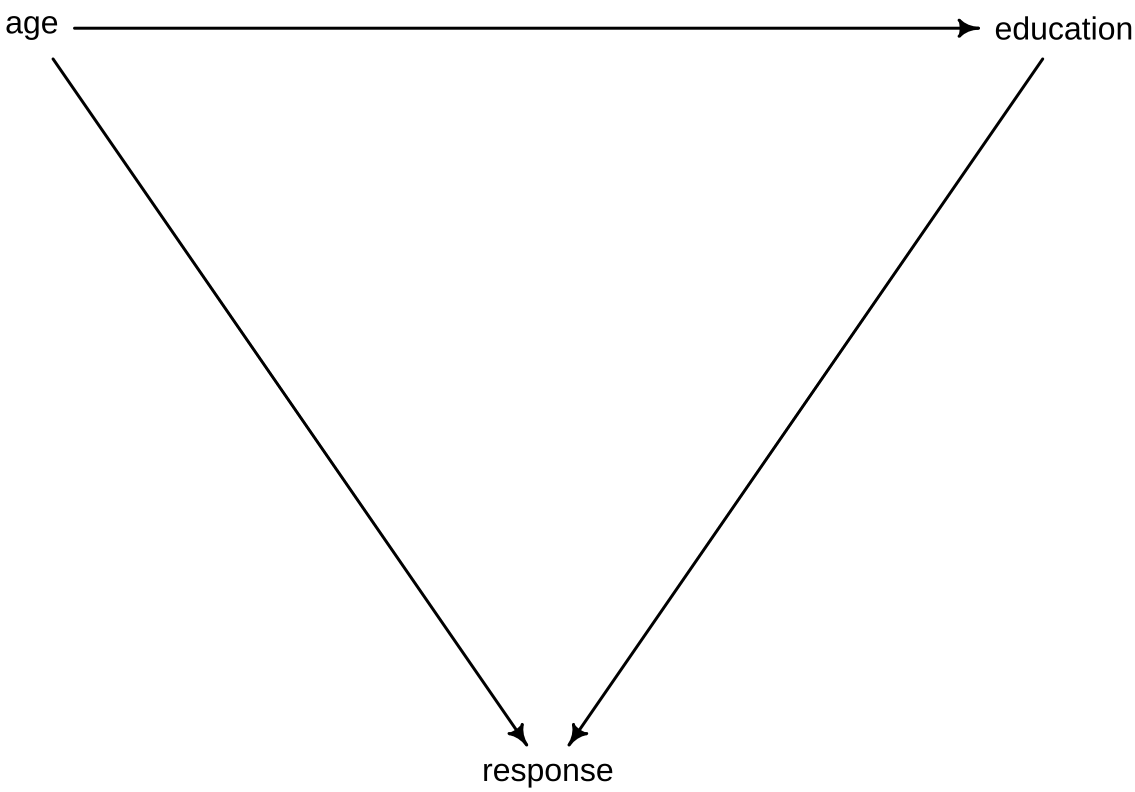
```
##                 mean         sd 5.5% 94.5%     histogram
## case             NaN         NA   NA    NA
## response   4.1992951  1.9050530    1     7
## order     16.5005035  9.2939946    2    31
## id               NaN         NA   NA    NA
## age       37.4894260 14.2336424   18    61
## male       0.5740181  0.4945159    0     1
## edu              NaN         NA   NA    NA
## action     0.4333333  0.4955606    0     1
## intention  0.4666667  0.4989128    0     1
## contact    0.2000000  0.4000201    0     1
## story            NaN         NA   NA    NA
## action2    0.6333333  0.4819187    0     1
```

Exercise 1

Part 1.a)

Education is associated with moral judgment. Now introducing age, to find out whether this association is causal: Age can have an influence on response through education, as well as directly. This is represented in the following DAG:

```
# define DAG
dag_trolley1 <- dagitty("dag{
  age -> response
  age -> education -> response
}")

# set coordinates
coordinates(dag_trolley1) <- list(x = c(age = 0, education = 2, response = 1),
                                  y = c(age = 0, education = 0, response = 1))
# draw
drawdag(dag_trolley1)
```



Age influences education, as the level of education depends on age. Additionally, age may influence response directly, as age might lead to different attitudes and behaviors among people and thus affect their responses.

Part 1.b)

To evaluate the causal influence of education on response, a model blocking the backdoor through age (fork: education <- age -> response) is required. This is confirmed by the following code, as well:

```
adjustmentSets(dag_trolley1, exposure = "education", outcome = "response")
```

```
## { age }
```

Firstly, I put the data in a list:

```
edu_levels <- c( 6 , 1 , 8 , 4 , 7 , 2 , 5 , 3 )
d$edu_new <- edu_levels[ d$edu ]

dat1 <- list(
  R = d$response,
  action = d$action,
  intention = d$intention,
  contact = d$contact,
  E = as.integer(d$edu_new), # edu_new as an index
  age = normalize(d$age),     # normalized age
  alpha = rep(2, 7)           # delta prior
)
```

It is basically the same list as in 12.34 on page 394, except age is added as new variable. I normalized age to be between 0 and 1. This allows using the same prior for age as for the other variables.

Extending the code 12.34 from page 394 (the model with education as categorical predictor) with the model from 12.24 from page 387 (the model containing interaction effects between action/contact and intention), I fit the new model containing age:

```
m1a <- ulam(
  alist(
    R ~ ordered_logistic(phi, kappa),
    phi <- bE * sum(delta_j[1:E]) + bAC * action + BI * intention + bC * contact
            + bAG * age,
    BI <- bI + bIA * action + bIC * contact,
    c(bAC, bC, bE, bAG, bI, bIA, bIC) ~ normal(0, 0.5),
    vector[8]:delta_j <<- append_row(0, delta),
    simplex[7]:delta ~ dirichlet(alpha),
    kappa ~ normal(0, 1.5)
  ),
  data = dat1,
  chains = 4,
  cores = 4
)
```

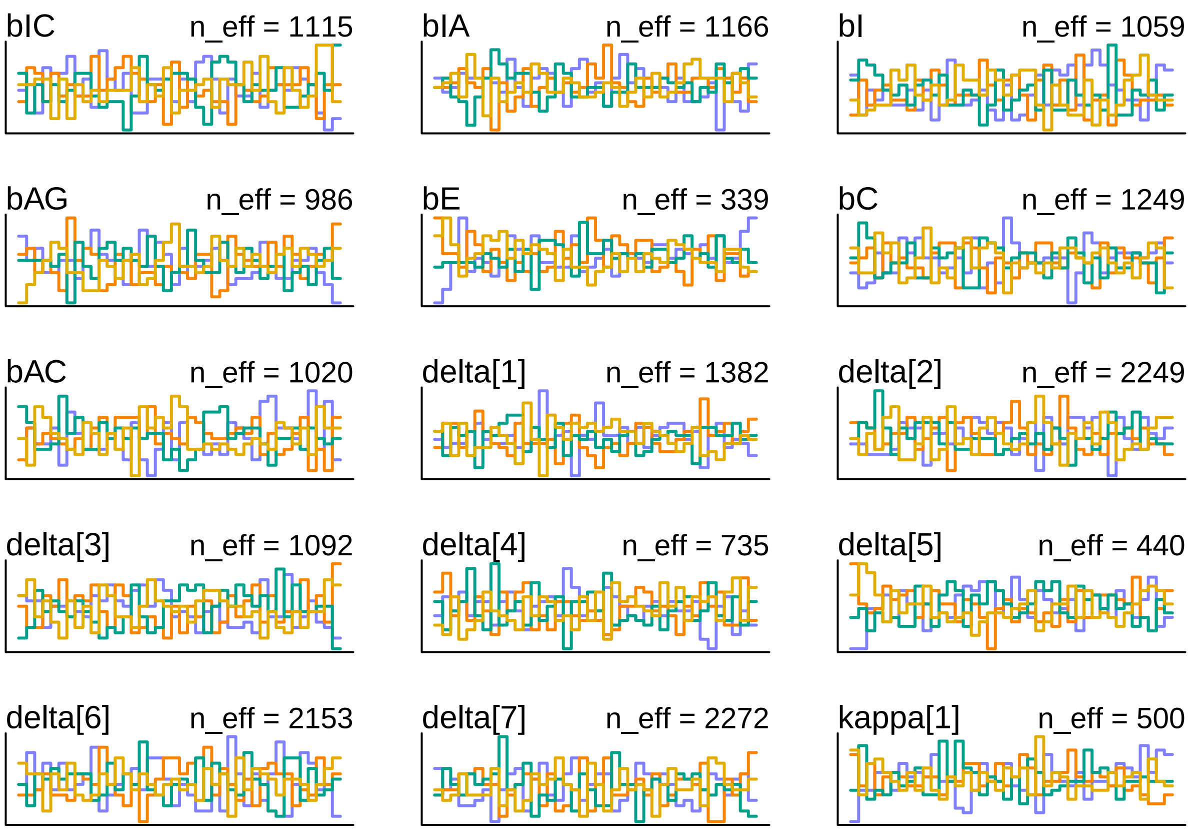
```
## Trying to compile a simple C file
```

```
## Warning: Tail Effective Samples Size (ESS) is too low, indicating posterior variances and tail quant
## Running the chains for more iterations may help. See
## https://mc-stan.org/misc/warnings.html#tail-ess
```

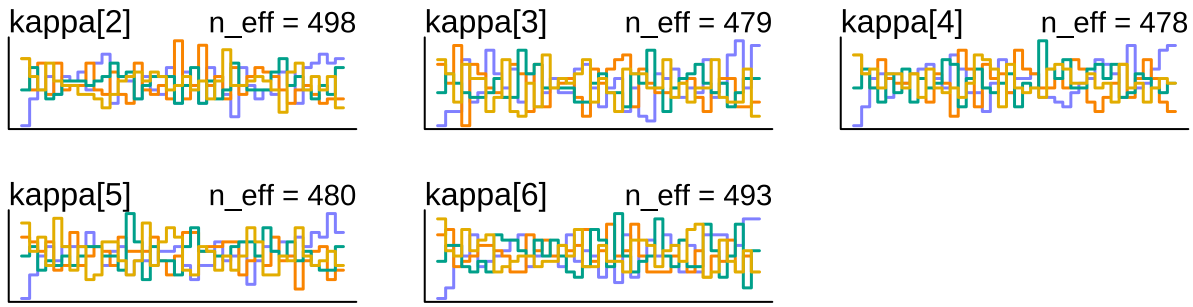
As the model produced a warning about the Tail Effective Samples Size (ESS) being too low, I used the `traceplot` and `trankplot` functions in order to check the chains for convergence:

```
trankplot(m1a)
```

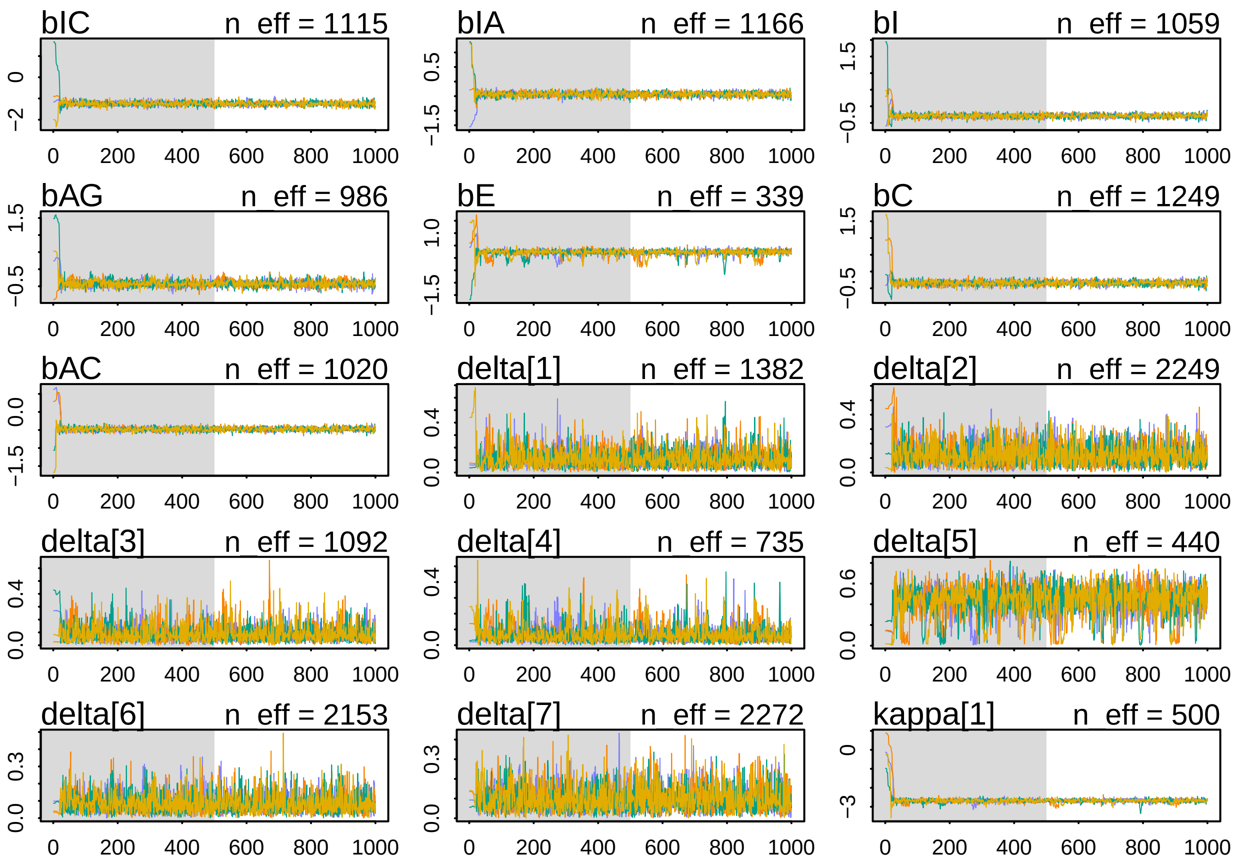
```
## Waiting to draw page 2 of 2
```

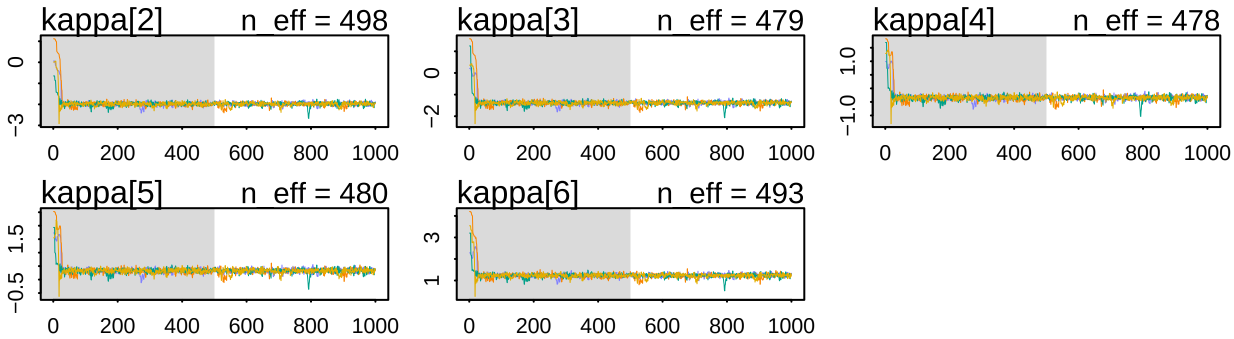


```
traceplot(m1a)
```




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

## Waiting to draw page 2 of 2
```





The plots show that the chains are converging for model m1a and, thus, that the model produces a reliable posterior distribution.

To be able to compare and see the influence of age on the relationship between education and response, I also created a model without age as reference. This means the following model is not blocking the backdoor through age:

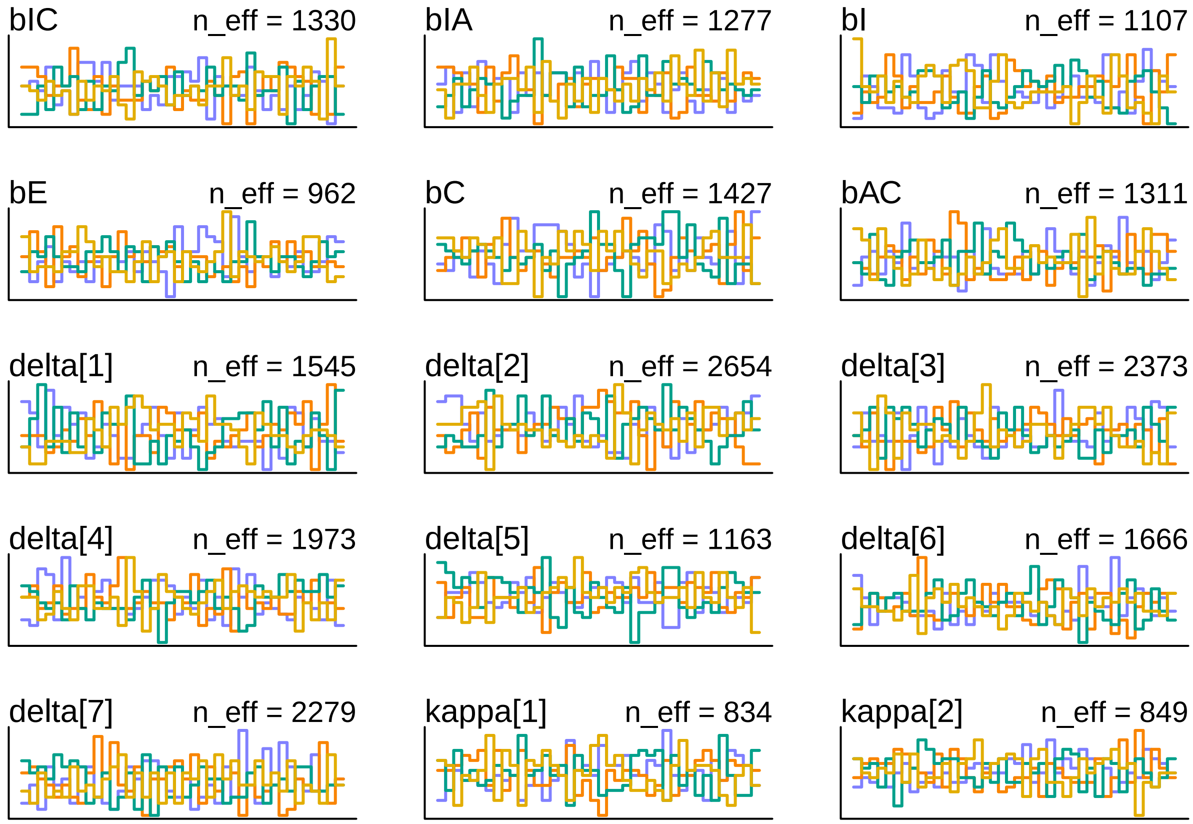
```
dat2 <- list(
  R = d$response,
  action = d$action,
  intention = d$intention,
  contact = d$contact,
  E = as.integer(d$edu_new),
  alpha = rep(2, 7)
)

m1b <- ulam(
  alist(
    R ~ ordered_logistic(phi, kappa),
    phi <- bE * sum(delta_j[1:E]) + bAC * action + BI * intention + bC * contact,
    BI <- bI + bIA * action + bIC * contact,
    c(bAC, bC, bE, bI, bIA, bIC) ~ normal(0, 0.5),
    vector[8]:delta_j <<- append_row(0, delta),
    simplex[7]:delta ~ dirichlet(alpha),
    kappa ~ normal(0, 1.5)
  ),
  data = dat2,
  chains = 4,
  cores = 4
)
```

```
## Trying to compile a simple C file
```

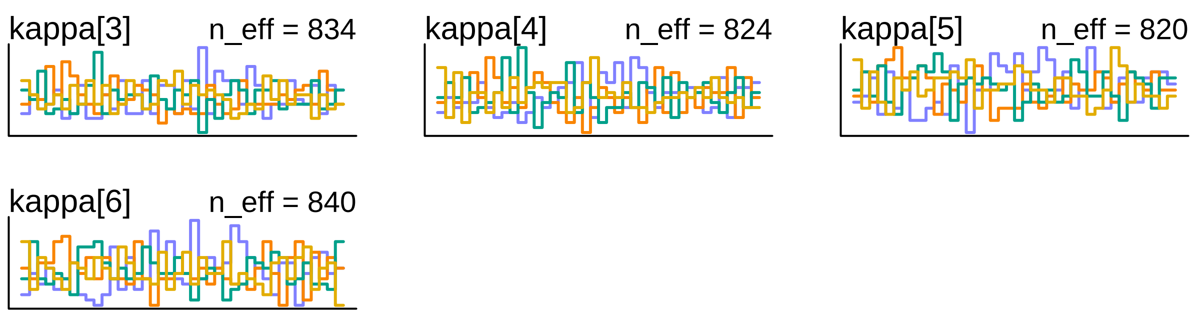
In order to check the chains for convergence, I use the `traceplot` and `trankplot` functions again:

```
trankplot(m1b)
```

```
## Waiting to draw page 2 of 2
```

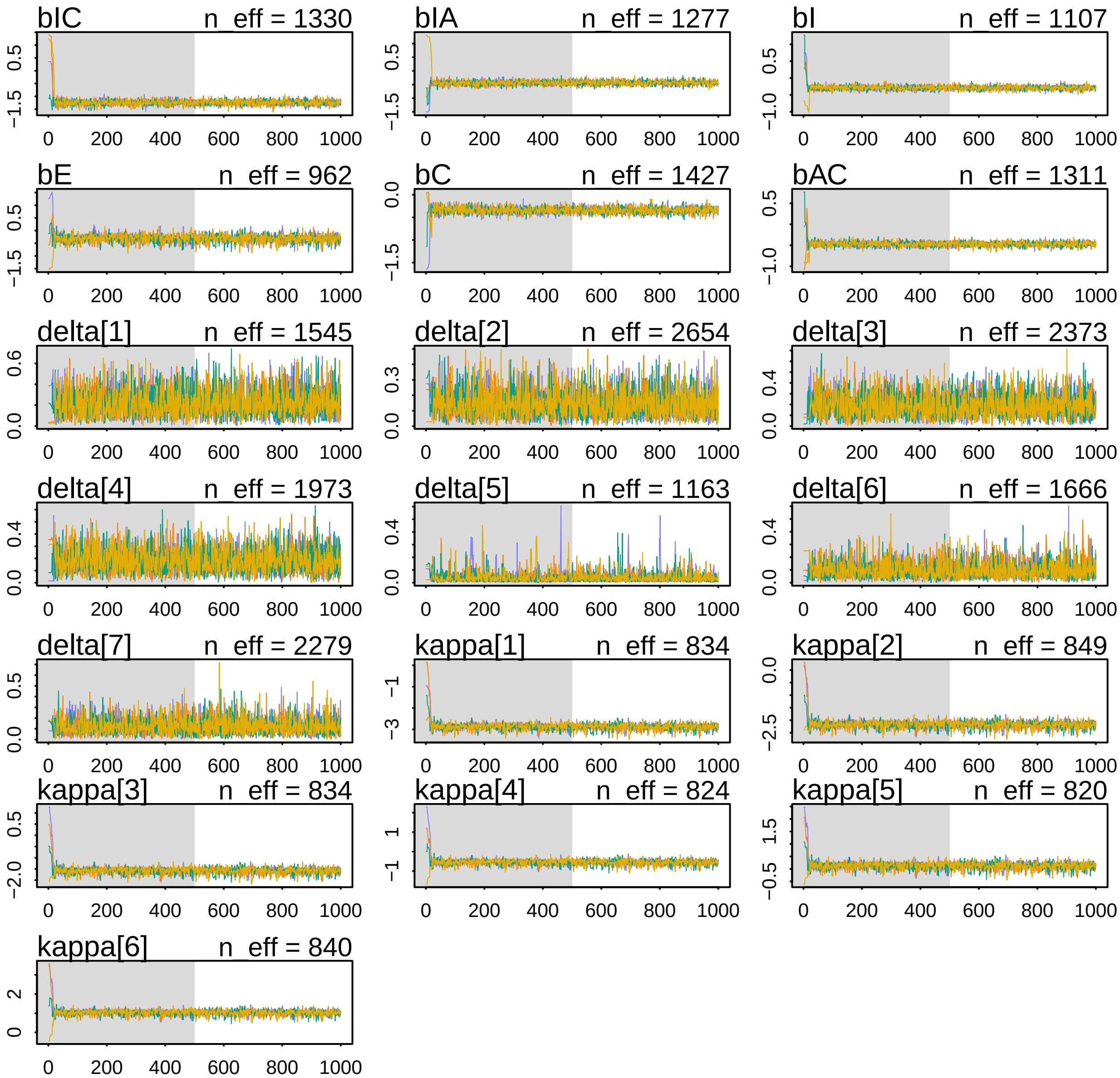



```
traceplot(m1b)
```




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

```
## Waiting to draw page 2 of 2
```



The plots show that the chains are converging for m1b and, thus, that the model produces a reliable posterior distribution.

Part 1.c)

To draw conclusions concerning the models and the causal effect of education on response, I took a look at the posterior distributions:

```
precis(m1a)
```

```
## 13 vector or matrix parameters hidden. Use depth=2 to show them.

##            mean         sd        5.5%      94.5%     n_eff    Rhat4
## bIC -1.2455736 0.09575077 -1.39885139 -1.0939330 1115.1474 1.002912
## bIA -0.4380788 0.07796473 -0.56388030 -0.3134728 1166.2261 1.000009
## bI  -0.2883147 0.05611758 -0.37571807 -0.1976156 1059.1974 1.002306
## bAG -0.4220824 0.09488474 -0.56486469 -0.2632953  985.8039 1.003772
## bE   0.2149456 0.12678265 -0.02336044  0.3591873  339.3764 1.009975
```

```
## bC  -0.3422422 0.06774991 -0.45072464 -0.2336532 1249.4902 1.001818
## bAC -0.4727846 0.05217685 -0.55489568 -0.3893505 1019.7421 1.001120
```

```r
precis(m1b)
```

```
## 13 vector or matrix parameters hidden. Use depth=2 to show them.
```

```
##           mean         sd       5.5%       94.5%     n_eff     Rhat4
## bIC -1.2377318 0.09464719 -1.3867819 -1.08923948 1330.3524 1.0019175
## bIA -0.4334620 0.07973614 -0.5628066 -0.30205388 1277.1816 0.9999391
## bI  -0.2906800 0.05895307 -0.3850580 -0.19806220 1107.2281 1.0029544
## bE  -0.3175479 0.15424950 -0.5774706 -0.08706233  961.5771 1.0009645
## bC  -0.3416565 0.06921230 -0.4514219 -0.23315047 1426.6179 0.9997005
## bAC -0.4716518 0.05466993 -0.5592551 -0.38574708 1311.1669 0.9986320
```

In model m1b (without age), education has a negative effect on response. Its parameter (`bE`) is distributed below zero. In model m1a (age included), education now has a positive effect on response. The mean of parameter `bE` is 0.21 and, considering the percentile intervals, it has only a slight overlap with zero.

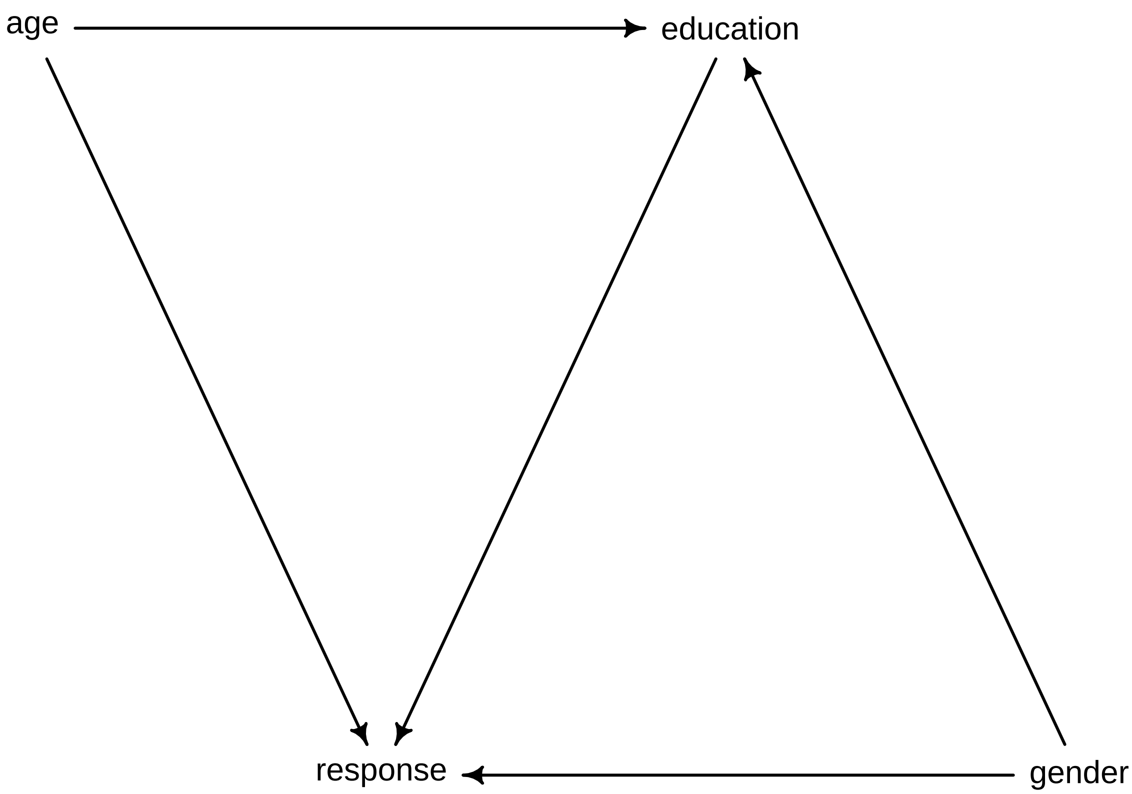
This means that education was probably confounded by age and some of the effect measured for education in model m1b can be assigned to the variable age. Additionally, there might be a causal relationship between age and response, as well as between education and response as both variables have an influence on response.

According to model m1a, age causes people to give lower response values and education causes higher response values.

Exercise 2

In this exercise, the variable gender is added, having a possible direct influence on education as well as on response. Therefore, this DAG now includes response, education, age, and gender:

```r
# define DAG
dag_trolley2 <- dagitty("dag{
  age -> response
  age -> education -> response
  response <- gender -> education
}")

# set coordinates
coordinates(dag_trolley2) <- list(x = c(age = 0, education = 2, gender=3 , response = 1),
                               y = c(age = 0, education = 0, gender=1, response = 1))
# draw
drawdag(dag_trolley2)
```



It is possible that any of the inferences from exercise 1 are confounded by gender. This is because the path `response <- gender -> education` is a fork. Therefore, it is possible that the effect of education on response is still confounded. To check for this, the variable gender needs to be added to the model, as well, to close the backdoor. The code below also confirms that, as well:

```
adjustmentSets(dag_trolley2, exposure = "education", outcome = "response")
```

```
## { age, gender }
```

Therefore, I constructed another model containing gender that can be compared with model m1a. I added an indicator variable `male` to the data list from model m1a. The variable is 1 when a person is male and 0 when a person is female.

```
dat1$male <- ifelse(d$male==1, 1L, 0L)
```

Then, I fitted the model:

```
m2 <- ulam(
  alist(
    R ~ ordered_logistic(phi, kappa),
    phi <- bE * sum(delta_j[1:E]) + bAC * action + BI * intention + bC * contact
            + bAG * age + bM * male,
    BI <- bI + bIA * action + bIC * contact,
    c(bAC, bC, bE, bAG, bI, bIA, bIC, bM) ~ normal(0, 0.5),
    vector[8]:delta_j <<- append_row(0, delta),
    simplex[7]:delta ~ dirichlet(alpha),
    kappa ~ normal(0, 1.5)
  ),
  data = dat1,
  chains = 4,
  cores = 4
```

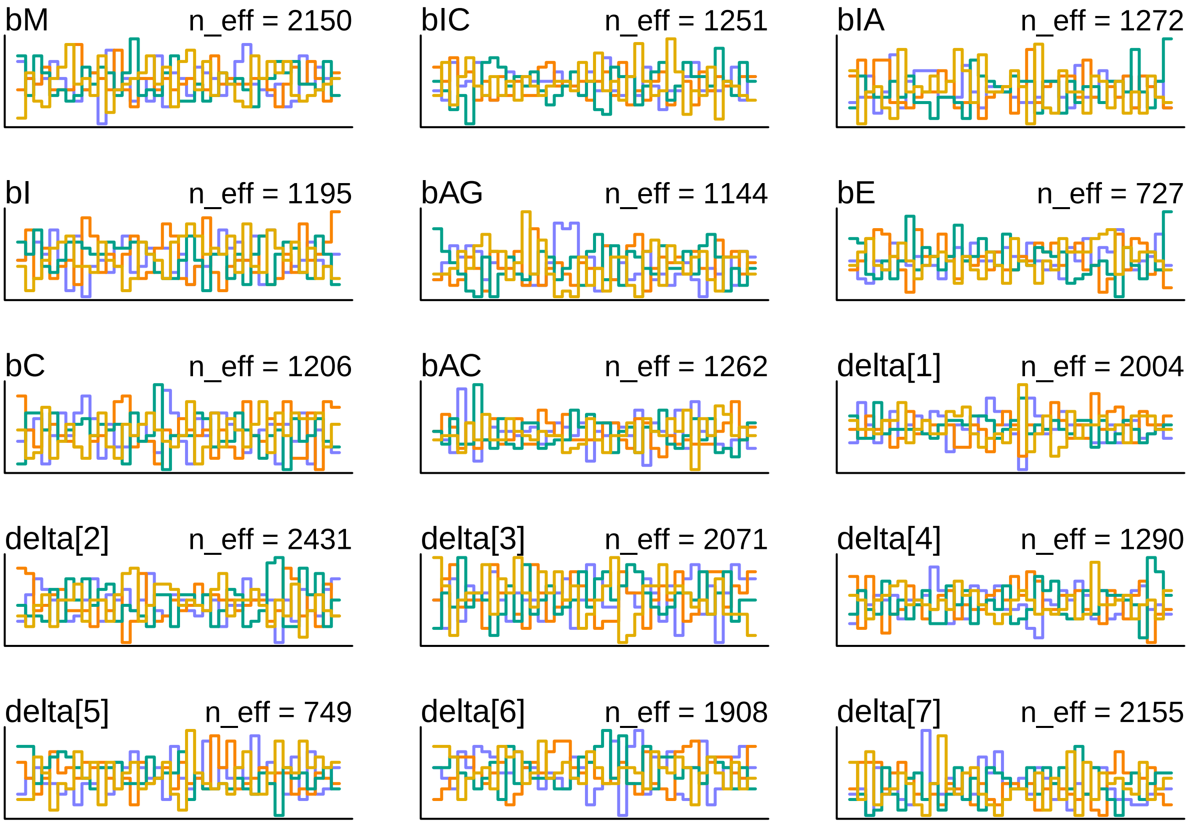
```
)
```

```
## Trying to compile a simple C file
```

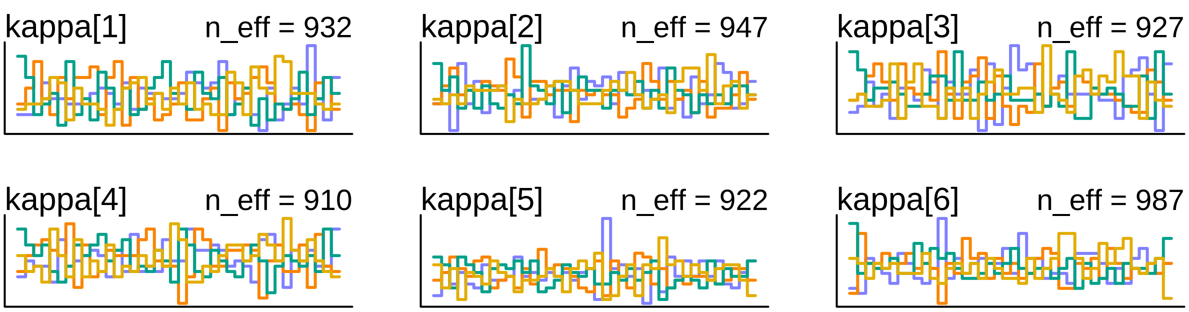
In order to check the chains for convergence, I use the `traceplot` and `trankplot` functions again:

```
trankplot(m2)
```

```
## Waiting to draw page 2 of 2
```

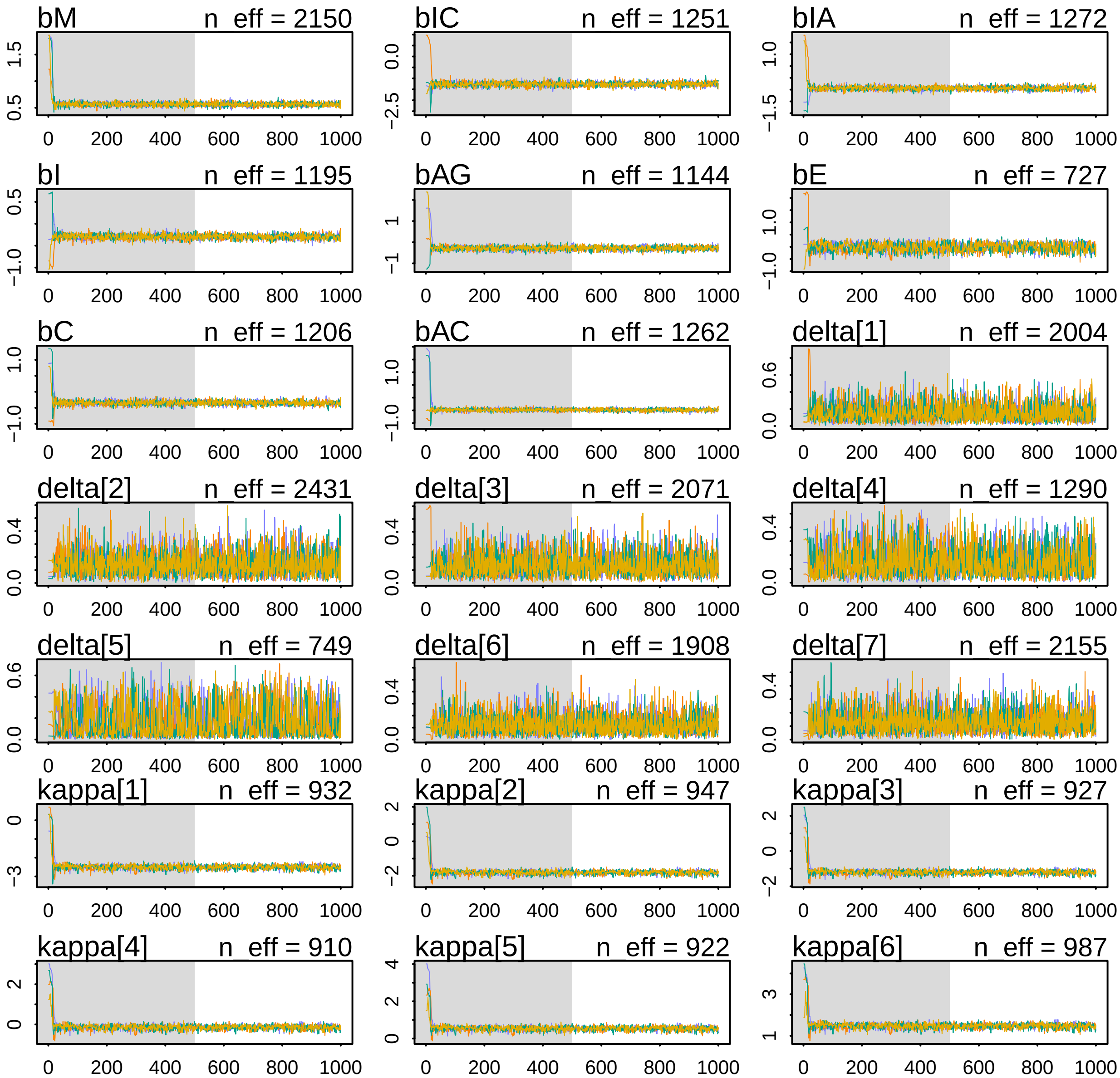



```
traceplot(m2)
```




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

```
## Waiting to draw page 2 of 2
```



The plots show that the chains are converging and, thus, that the model produces a reliable posterior distribution.

In order to draw conclusions about the causal relationships, I took a look at the posterior of model m2 to be able to compare it to the posterior of model m1a.

```
precis(m2)
```

```
## 13 vector or matrix parameters hidden. Use depth=2 to show them.

##             mean         sd       5.5%      94.5%    n_eff     Rhat4
## bM    0.5665047 0.03595644  0.5111113  0.6261524 2149.712 0.9991697
## bIC  -1.2626516 0.09523844 -1.4104110 -1.1111593 1251.382 1.0006144
## bIA  -0.4387705 0.07827418 -0.5667044 -0.3149655 1271.761 1.0006026
## bI   -0.2934150 0.05551542 -0.3838270 -0.2070256 1194.937 1.0000303
## bAG  -0.2798029 0.09485832 -0.4324285 -0.1320679 1144.183 1.0011658
## bE   -0.0218418 0.16375622 -0.2811592  0.2282922  727.392 1.0013048
## bC   -0.3457515 0.06681672 -0.4501223 -0.2377825 1205.624 1.0012902
```

```
## bAC -0.4794083 0.05141869 -0.5642298 -0.3991760 1262.255 0.9991547
```

Comparing model m1a and m2 with each other, the parameter for education (`bE`) has a mean close to zero and a high standard deviation, spreading from values below to values above zero in model m2 (with gender). The parameter for gender (`bM`) is positively associated with response in model m2.

This means that education in model m1a was probably still confounded by gender and some of the effect measured for education in model m1a belongs actually to the variable gender.

Therefore, there might be a causal relationship between gender and response, but it is possible that education is, in fact, not causally influencing response like previously assumed. Education seems to have no impact on response when gender is included.

In general, in model m2, education has no significant impact, age has a negative impact, and gender has a positive impact on response.

Exercise 3

The following model was given:

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

In order to rewrite the following model as a multilevel model, I used a varying intercepts model. Instead of assigning α_{group} to a certain prior distribution like above, the distribution of α_{group} now contains two hyperparameters that are distributed with hyperpriors. This results in the following model:

$$y_i \sim Binomial(1, p_i)$$
$$logit(p_i) = \alpha_{group[i]} + \beta x_i$$
$$\alpha_{group} \sim Normal(\overline{\alpha}, \sigma)$$
$$\overline{\alpha} \sim Normal(0, 1.5)$$
$$\sigma \sim Exponential(1)$$
$$\beta \sim Normal(0, 0.5)$$