assignment_5_blaufuss

Dennis Blaufuss

12/10/2021

For this assignment, you will work with the Trolley dataset:

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

	mean	sd	5.5%	94.5%	histogram
	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
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	▇▁▁▁▁▁▁▁▁▂

1-10 of 12 rows Previous 1 2 Next

The basis of the assignment are the models developed in sections 12.3-12.4 of the textbook. So, you should first implement that code before starting this problem set.

Note: I will use a combination of m12.5 and m12.6 to include Education but the Interaction effect as well.

1. We see that education, modeled as an ordered category, is associated with moral judgments. Is this association causal? One possible confound is that education is also associated with age through a causal process: namely, people are older when they finish a level of education than when they begin it.

Reconsider the Trolley data in light of this issue. Specifically,

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

(b) Identify which statistical model or models are required to evaluate the causal influence of education on responses. Hint: you should use the code in the book to reorder the education level labels:

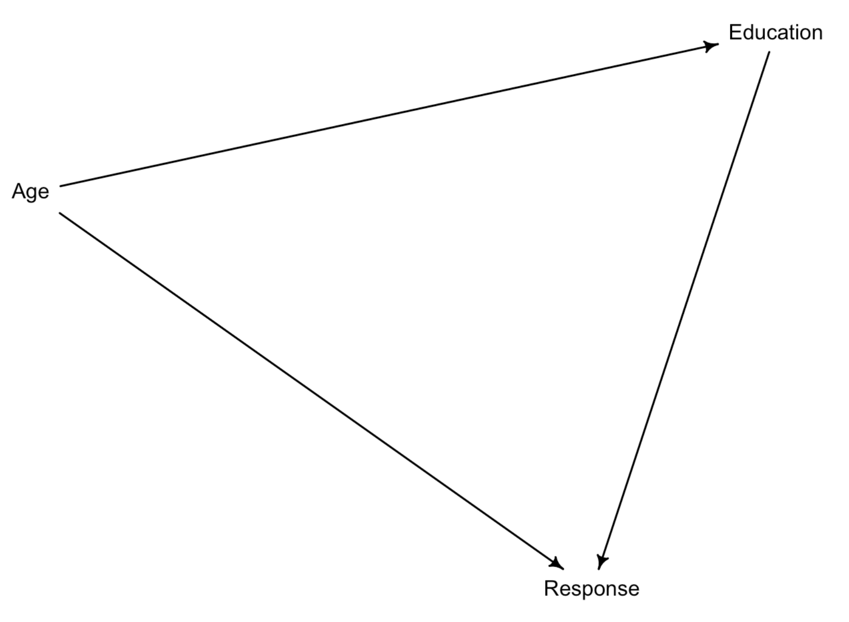
```
# R code 12.31
edu_levels <- c( 6 , 1 , 8 , 4 , 7 , 2 , 5 , 3 )
d$edu_new <- edu_levels[ d$edu ]
```

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

Answer:

a. As we include Age in this relationship, the DAG will look like the following:

```
dag <- dagitty("dag{
                    Education -> Response
                    Age -> Education
                    Age -> Response
               }"
              )
drawdag(dag)
```



b. We should first check for backdoors:

```
adjustmentSets(dag, exposure = "Education", outcome = "Response")
```

```
## { Age }
```

As we see in the DAG from a) and the function above reassures, to investigate on the effect of Education on Response we need to close the backdoor over Age with conditioning on Age (as of including Age in our model). Furthermore, as always it makes sense to standardize age here. Thus the following alteration of the model(s) of the book will be required:

```
dat_list <- list(
    R = d$response ,
    A = d$action,
    I = d$intention,
    C = d$contact,
    E = as.integer( d$edu_new ),
    Y = standardize( d$age ),
    alpha = rep(2,7)
)

m1 <- ulam(
    alist(
        R ~ ordered_logistic( phi , kappa ),
        phi <-  bE*sum(delta_j[1:E]) + bA*A + BI*I + bC*C + bY*Y,
        BI <- bI + bIA*A + bIC*C ,
        c(bA,bI,bC,bIA,bIC,bE,bY) ~ normal( 0 , 0.5 ),
        kappa ~ normal( 0 , 1.5 ),
        vector[8]: delta_j <<- append_row( 0 , delta ),
        simplex[7]: delta ~ dirichlet( alpha )
    ), data=dat_list , chains=4 , cores=4     )
```

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

```
## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be unreliabl
e.
## Running the chains for more iterations may help. See
## https://mc-stan.org/misc/warnings.html#bulk-ess
```

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

Note: I use the BI instead of the bI approach (meaning we include the interaction) as of m12.5 of the book as we have as stated in the a book and observable in those charts a large interaction between contact and intention. Thus also the tighter priors.

```
precis(m1, 2, omit="kappa")
```

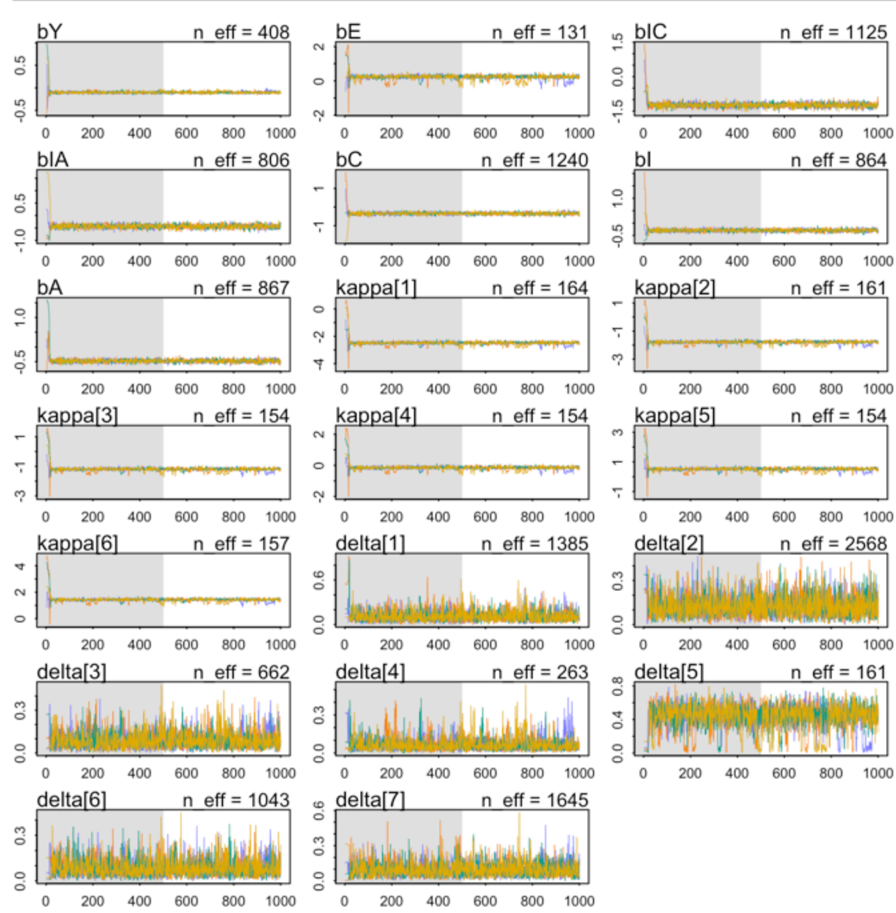
	mean	sd	5.5%	94.5%	n_eff	Rhat4
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
bY	-0.10087074	0.02212597	-0.13361413	-0.06382878	408.1232	1.0084215
bE	0.21523357	0.13196548	-0.06466159	0.36761423	131.1520	1.0318072
bIC	-1.24065262	0.09826204	-1.39250965	-1.07786489	1124.7972	0.9985121
bIA	-0.43615387	0.07983286	-0.56307386	-0.30624127	806.0485	0.9989759
bC	-0.34476257	0.06929354	-0.45874362	-0.23967816	1240.3310	0.9984875
bI	-0.28973310	0.05763773	-0.38535933	-0.20094591	863.5703	0.9986487
bA	-0.47458363	0.05487493	-0.56218573	-0.38699726	867.2203	0.9994198
delta[1]	0.11341469	0.07555643	0.02502715	0.25240190	1385.0705	1.0017015
delta[2]	0.12145191	0.07797504	0.02461366	0.26522202	2567.7973	0.9986851
delta[3]	0.08925497	0.06241932	0.01844296	0.20141897	661.5132	1.0052054

1-10 of 14 rows Previous 1 2 Next

```
traceplot(m1)
```

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

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



Before concluding anything about the casual relationship we first should check our chains and precis function output for any bad signs:

- Our Chains are looking fine: We still see the convergence although it isn't as strong as in previous assignments.
- Our n_effs are looking a little bit low and we get a warning message while running the model as well: the Tail Effective Sample Size (ESS) is too low. So running chains for more iterations may help.
- Our Rhats aren't precisely 1 as in previous assignments. But still close enough to not consider it as a bad sign.

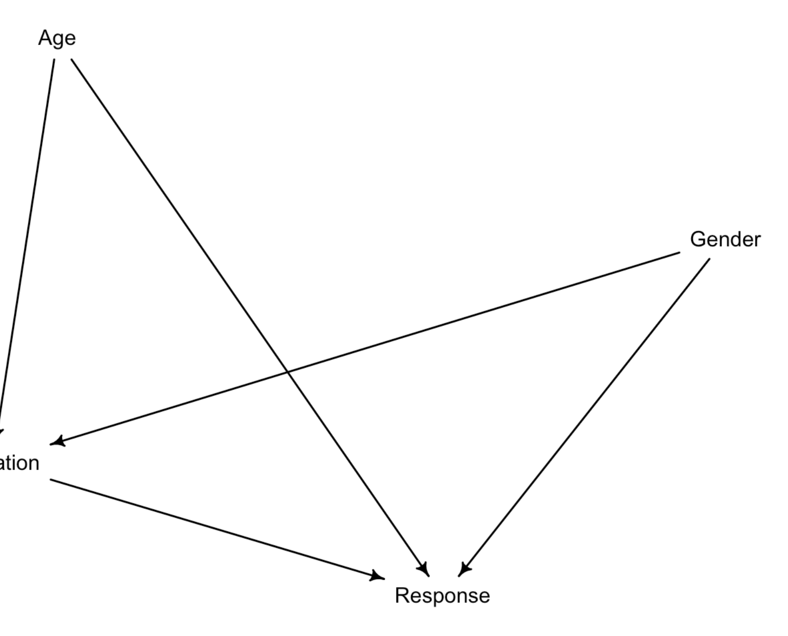
c. What we now observe is that education has a small positive effect on response. Recall here that in the "old" model in the book that's not including age this effect was actually negative. So with this information we can state for sure that age somewhat interferes here and the backdoor may be real. Obviously, there could be a "third" variable that we haven't yet taken into account that is the "real driver" for this relationship. And maybe even in this model with the mentioned variables there could be an interaction effect that is not yet considered.

Still to sum this up I would conclude that we are on the right track in understanding the casual relationship between these three considered variables.

2. Now consider one more variable in the Trolley data: Gender. Suppose that gender might influence education as well as response directly. Draw the DAG now that includes response, education, age, and gender. Is it possible that any of the inferences from Question 1 are confounded by gender? If so, define any additional models you need to infer the causal influence of education on response. What do you conclude?

Answer:

```
dag <- dagitty("dag{
                    Education -> Response
                    Age -> Education
                    Age -> Response
                    Gender -> Education
                    Gender -> Response
               }"
              )
drawdag(dag)
```




```
adjustmentSets(dag, exposure = "Education", outcome = "Response")
```

```
## { Age, Gender }
```

As we see with de DAG (again: or the output of the function) we now need to close two backdoors: Age & Gender. For the Gender I chose an indicator variable: Male as 1 and female as 0.

```
dat_list$male <- ifelse( d$male==1 , 1L , 0L )
m2 <- ulam(
    alist(
        R ~ ordered_logistic( phi , kappa ),
        phi <- bE*sum( delta_j[1:E] ) + bA*A + bC*C + BI*I +
               bY*Y + bM*male,
        BI <- bI + bIA*A + bIC*C ,
        c(bA,bI,bC,bIA,bIC,bE,bY,bM) ~ normal( 0 , 0.5 ),
        kappa ~ normal( 0 , 1.5 ),
        vector[8]: delta_j <<- append_row( 0 , delta ),
        simplex[7]: delta ~ dirichlet( alpha )
    ), data=dat_list , chains=4 , cores=4 )
```

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

```
precis(m2, 2, omit="kappa")
```

	mean	sd	5.5%	94.5%	n_eff	Rhat4
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
bM	0.564927665	0.03772576	0.50133245	0.62497757	2487.0980	0.9995180
bY	-0.068082252	0.02188365	-0.10456344	-0.03387347	985.6228	1.0034566
bE	-0.002833077	0.17474328	-0.28619128	0.25105247	530.0495	1.0057019
bIC	-1.264474970	0.09508869	-1.41570136	-1.11150417	1355.8245	1.0014788
bIA	-0.438227516	0.07873868	-0.56240412	-0.30751447	1133.7613	1.0023626
bC	-0.344370274	0.06808029	-0.45095912	-0.23675553	1311.2048	1.0012117
bI	-0.294390979	0.05691804	-0.38505054	-0.20410722	1116.3579	1.0019925
bA	-0.479906123	0.05488137	-0.57070400	-0.39427238	1247.8703	1.0004447
delta[1]	0.153268902	0.10208833	0.03189573	0.33952042	1520.8718	1.0018333
delta[2]	0.144229076	0.09057950	0.03297076	0.30893515	2314.9769	0.9993503

1-10 of 15 rows Previous 1 2 Next

```
traceplot(m2)
```

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

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

Again checking the chains and precis output:

- Our chains still look fine.
- Our n_effs are a little bit but not significantly higher than in model 1 and we don't get a warning message this time.
- Our Rhats look pretty much the same (maybe a little bit better across the board) compared to model 1.

Pretty interesting to see is that the casual influence of Education now seems to be near to none. So at this point it seems like the addition of Gender into the DAG was the correct choice. Weirdly enough our casual influence of Age changes as well. This seems a little odd at first glance since there shouldn't be any mayor influences of Age that are explained better by Gender. If you take a brief look into the data set you will see that we don't have an even distribution across all ages in both genders (meaning in this case the previous idea of Gender not interfering with Age does not apply!). So for the case of this data set I would state that Gender accounts to a lot of the influence on response that was previously allocated to education (and even Age). We may as well state here that males typically show more approval as females.

To Conclude I want to state that to further understand the whole role of Gender in that relationship we require a sample that is better representing the whole population as of with a better / more even distribution across Gender & Age (and maybe even education as well). In the sample as it is right now there is too much risk that we base our result on a bias, just to name two examples: bad representation of older ladies could lead to more impact of Gender instead of Age, underrepresented educational groups may infer with our casual influence of education.

3. Rewrite the following model as a multilevel model.

$y_i \sim Binomial(1, p_i)$

$logit(p_i) = \alpha_{group[i]} + \beta x_i$

$\alpha_{group} \sim Normal(0, 1.5)$

$\beta \sim Normal(0, 0.5)$

Answer:

$y_i \sim Binomial(1, p_i)$

$logit(p_i) = \alpha_{group[i]} + \beta x_i$

$\alpha_{group} \sim Normal(\overline{\alpha}, \sigma)$

$\overline{\alpha} \sim Normal(0, 1.5)$

$\beta \sim Normal(0, 0.5)$

$\sigma \sim Exponential(1)$

Or in the case that the second level is not the highest level you should name α and σ accordingly (e. g. with the subscript "lvl2").