Exercises week 5

prof. Richard Torkar

March 29, 2021

> La vie n'est bonne qu'à deux choses: à faire des mathématiques et à les professer.
>
> Siméon Denis Poisson

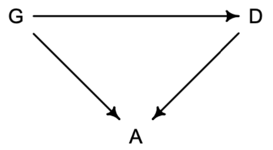
The weekly exercises are mainly from McElreath's notes which can be found on GitHub: `https://github.com/rmcelreath/stat_rethinking_2020/tree/main/homework`

Exercise 1. The data in `data(NWOGrants)` are outcomes for scientific funding applications for the Netherlands Organization for Scientific Research (NWO) from 2010–2012 (see van der Lee and Ellemers doi:10.1073/pnas.1510159112). These data have a very similar structure to the UCBAdmit data discussed in Chapter 11.

I want you to consider a similar question: What are the total and indirect causal effects of gender on grant awards? Consider a mediation path (a pipe) through discipline. Draw the corresponding DAG and then use one or more binomial GLMs to answer the question. What is your causal interpretation? If NWO's goal is to equalize rates of funding between the genders, what type of intervention would be most effective?

That's it for this week. For next week we'll start introducing software engineering data.

Solution 1. The implied DAG can be seen to the right.



Where G is gender, D is discipline, and A is award. The direct causal effect of gender is the path $G \rightarrow A$. The total effect includes that path **and** the indirect path $G \rightarrow D \rightarrow A$. We can estimate the total causal influence (assuming this DAG is correct) with a model that conditions only on gender. I'll use a Normal$(-1, 1)$ prior for the intercepts, because we know from domain knowledge that less than half of applicants get awards.

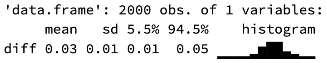
```
library(rethinking)
data(NWOGrants)
d <- NWOGrants
dat_list <- list(
    awards = as.integer(d$awards),
    apps = as.integer(d$applications),
    gid = ifelse( d$gender=="m" , 1L , 2L ) )
m1_total <- ulam(
    alist(
        awards ~ binomial( apps , p ),
        logit(p) <- a[gid],
        a[gid] ~ normal(-1,1)
    ), data=dat_list , chains=4 , cmdstan=TRUE )
precis(m1_total,2)
```

with the following output,

```
      mean   sd  5.5% 94.5% n_eff Rhat4
a[1] -1.53 0.07 -1.64 -1.43  1463     1
a[2] -1.74 0.08 -1.88 -1.61  1190     1
```

Gender 1 here is male and 2 is female. So males have higher rates of award, on average. How big is the difference? Let's look at the contrast on absolute scale:

```
post <- extract.samples(m1_total)
diff <- inv_logit( post$a[,1] ) - inv_logit( post$a[,2] )
precis( list( diff=diff ) )
```



```
'data.frame': 2000 obs. of 1 variables:
     mean   sd 5.5% 94.5%    histogram
diff 0.03 0.01 0.01  0.05
```

which gives the output on the right.

So a small 3% difference on average. Still, with such low funding rates (in some disciplines), 3% is a big advantage.

Now for the direct influence of gender, we condition on discipline as well:

```
dat_list$disc <- as.integer(d$discipline)
m1_direct <- ulam(
    alist(
        awards ~ binomial( apps , p ),
        logit(p) <- a[gid] + d[disc],
        a[gid] ~ normal(-1,1),
        d[disc] ~ normal(0,1) ),
    data=dat_list , chains=4 , cores=4 , cmdstan=TRUE )
precis(m1_direct,2)
```

which gives the following output,

```
      mean   sd  5.5% 94.5% n_eff Rhat4
a[1] -1.31 0.30 -1.79 -0.86   205  1.02
a[2] -1.45 0.31 -1.95 -0.99   203  1.02
d[1]  0.30 0.36 -0.27  0.86   294  1.01
d[2] -0.03 0.33 -0.53  0.50   232  1.02
```

```
d[3] -0.26 0.32 -0.75  0.25   223  1.01
d[4] -0.30 0.35 -0.84  0.27   260  1.01
d[5] -0.37 0.32 -0.85  0.15   219  1.01
d[6] -0.05 0.35 -0.59  0.50   258  1.01
d[7]  0.26 0.38 -0.34  0.88   305  1.01
d[8] -0.48 0.31 -0.97  0.03   214  1.02
d[9] -0.24 0.33 -0.74  0.30   254  1.01
```

Aouch... those chains didn't sample very efficiently. This is because the model is over-parameterized—it has more parameters than absolutely necessary. This doesn't break it. It just makes the sampling less efficient and we should be very careful about interpreting the results. Anyway, now we can compute the gender difference again. On the relative scale:

```
post <- extract.samples(m1_direct)
diff_a <- post$a[,1] - post$a[,2]
precis( list( diff_a=diff_a ) )
```

```
'data.frame': 6000 obs. of 1 variables:
       mean   sd  5.5% 94.5% histogram
diff_a 0.14 0.11 -0.03  0.31 ▁▁▂▅▇▅▁▁▁
```

and we get the output on the right.

Still an advantage for the males, but reduced and overlapping zero a bit. To see this difference on the absolute scale, we need to account for the base rates in each discipline as well. If you look at the `postcheck(m1_direct)` display, you'll see the predictive difference is very small. There are also several disciplines that reverse the advantage. If there is a direct influence of gender here, it is small, much smaller than before we accounted for discipline. Why? Because again the disciplines have different funding rates and women apply more to the disciplines with lower funding rates. But it would be hasty, I think, to conclude there are no other influences. There are after all lots of unmeasured confounds...