GR5065 Homework 3

Ben Goodrich

Due February 23, 2021 at 8PM New York Time

```
# call the set.seed function once here to make the knitting conditionally deterministic
```

1 Police Stops in North Carolina

Read the paper at

https://projecteuclid.org/journals/annals-of-applied-statistics/volume-11/issue-3/The-problem-of-infra-marginality-in-outcome-tests-for-discrimination/10.1214/17-AOAS1058.full

The number on traffic stops by race of the driver in various jurisdictions of North Carolina (which is a state in the southeast part of the United States) over several years can be obtained — if your working directory is the same as your .Rmd file — via

```
stops <- readRDS("north_carolina.rds")
head(stops)
```

```
##                                        Asian  Black Hispanic  White
## Charlotte-Mecklenburg Police Department 13599 419873    77727 282761
## Raleigh Police Department                8065 199297    40638 173329
## Greensboro Police Department             4812 143572    13711 105563
## Fayetteville Police Department           2535 128176    13486  76564
## Winston-Salem Police Department          1666  98631    22355  89247
## Durham Police Department                 2392  82629    17326  37466
```

1.1 Prior Predictive Distribution

Write an R function that draws (once) from the prior predictive distribution of "searches" (of cars, drivers, etc.) and "hits" (searches that find something illegal) for every police department in North Carolina and each of the four race categories using the model described in section 2.3 of the paper. Your R function could put values into a three-dimensional array and start like:

```
prior_PD <- function(stops) {
  D <- nrow(stops)
  R <- ncol(stops)
  searches_hits <- array(0, dim = c(D, R, 2), dimnames =
                         list(rownames(stops), colnames(stops), c("searches", "hits")))
  # modify all the cells of searches_hits by drawing from appropriate distributions
  return(searches_hits)
}
```

The `stops` data have already been sorted such that the first row corresponds to the largest police department, which is handled a bit differently from the other police department under the authors' model of the data-generating process.

We previously discussed the prior predictive distribution in relatively simple contexts, such as vaccine trials, where there was only one unknown parameter and only one observation on only one outcome variable. This situation is more complicated because there are 100 counties, 4 race categories, two outcome variables (searches and hits) that are not independent, and many parameters that are related to each other in a multilevel fashion. This situation is a bit like our model for bowling where there was one unknown parameter, Υ, that the probability distribution of the first roll was conditioned on and then the second roll of that frame was conditioned on the result of the first roll. In models with multilevel parameters, the "top-level" parameters are drawn from a marginal prior distribution and then the parameters at the next level down are drawn the their conditional prior distributions, given the realizations of the "top-level" parameters. This process can proceed downward through as many levels as are necessary until you reach the bottom level(s) where the data can be drawn conditional on the realizations of some of the parameters at higher levels. Here are some additional hints:

- The DAG in Figure 2 of the paper is helpful, which includes all of the parameters in addition to the observable variables searches, (S_{rd}), and "hits" (H_{rd}). All of the nodes inside the plate marked R are race-specific and all of the nodes inside the plate marked D are specific to the police department. All of the parameters outside of a plate are at the state (of North Carolina) level.
- When the authors use the $\sim$ symbol, it means that the variable to the left is distributed as (and can be drawn from) the distribution on the right.
- The paper parameterizes the Beta distribution in terms of the expectation and a positive "concentration" parameter, whereas the `rbeta` function in R parameterizes the Beta distribution in terms of two positive shape parameters, `shape1` and `shape2`. As implied by footnote 3, you can map from the parameterization in the paper to the parameterization used by R by using $\alpha = \phi\lambda$ and $\beta = (1 - \phi)\,\lambda$, where ϕ is the expectation, λ is the concentration, α is the first shape parameter, and β is the second shape parameter.
- The "inverse logit" function, denoted $\text{logit}^{-1}(x)$ is just $\frac{1}{1+e^{-x}}$, which is already implemented as `plogis` in R because it is also the CDF of the standard logistic distribution.
- By $\mathcal{N}_+(0, 2)$, the authors mean a "half-normal" distribution with location 0 and scale 2 that excludes negative values (and thus 0 and 2 are not the expectation and standard deviation of the truncated random variable). You can draw from a "half-normal" distribution with a location of 0 by drawing from the corresponding untruncated normal distribution and then taking the absolute value.
- On page 1201, the authors go into some detail about the binomial likelihood function, which makes sense for computational reasons if you were drawing from the posterior distribution of the parameters. Since you are merely drawing from the prior predictive distribution of searches and hits, these computational considerations do not apply and it is much easier if you draw repeatedly from Bernoulli distributions and sum them up yourself in order to obtain realizations of searches and hits for each racial group and police department. Since the Bernoulli distribution is a special case of the binomial distribution with $n = 1$, you can (repeatedly) call the `rbinom` function in R with `size = 1`. Thus, you really only need to follow the authors' description of the data-generating process under their model through page 1200.

1.2 Description

Call your `prior_PD` function 1000 times, pulling out the results for the police department of the University of North Carolina at Chapel Hill, which is row 85 (this is the largest university in North Carolina and is famous for its men's basketball team). How would you describe the authors' prior beliefs about the distribution of searches and hits by this police department? Since the absolute figures are not too relevant and heavily influenced by the size of the surrounding population, it is presumably better to transform the counts of the prior predictive distribution into proportions or ratios.

1.3 Criticism

What is one criticism of the authors' model of the data-generating process and why do you think it is the most substantial criticism?

2 Medicaid Expansion in Oregon

Read

https://www-nowpublishers-com.ezproxy.cul.columbia.edu/article/Details/QJPS-19026

including the appendices. The essence of it is that in 2008, the state of Oregon conducted a lottery among households with sufficiently low income to decide who would be eligible for government-provided health insurance (Medicaid). It is rare to have a randomized variable in such a large dataset that could make such a substantial difference to the people in the study. Economists have considered the effect of (eligibility for) Medicaid on a variety of outcomes, and in this study they consider voting turnout and registration. However, not everyone who won the Medicaid lottery actually obtained Medicaid (some did not follow up and many who did turned out not to be eligible because their income was not low enough) and a small number of people who did not win the Medicaid lottery subsequently obtained Medicaid (usually by getting married and / or having children, which changes the eligibility criteria). Thus, the analysis is more complicated because those who obtain Medicaid may be more (or less) likely to vote for many spurious reasons.

Under Supplementary Information, click on the link that says "Replication Data" to download a file called 100.00019026_supp.zip to your working directory. Then, the following R syntax will get the dataset into R:

```
library(haven)
unzip("100.00019026_supp.zip")
oregon <- as_factor(read_dta(file.path("19026_supp", "Data", "individual_voting_data.dta")))
```

These data can be rewritten in a form suitable for the CausalQueries package by calling

```
library(dplyr)
oregon <- transmute(oregon,
                    V = vote_presidential_2008_1,           # voted in Nov 2008?
                    M = ohp_all_ever_nov2008 == "Enrolled", # had Medicaid in Nov 2008?
                    L = treatment,                          # won lottery in spring 2008?
                    N = numhh_list != "signed self up")     # registered additional adults?
```

The variable called `N` in `oregon` is a slight simplification of the `numhh_list` factor in the original dataset. In Oregon, if any adult in a household won the Medicaid lottery, then all adults in the household could obtain Medicaid (if they were eligible). Thus, if the household included two or more adults, they had a much higher probability of being able to obtain Medicaid than a household with only one adult. Including all levels of `numhh_list` is theoretically important because of how it affects the probability of winning the Medicaid lottery but in practice it was not that important because the number of adults in the household has a negligible effect on the probability that any of them vote. In addition, there were only 158 observations from households with three or more adults, so `N` collapses this to a logical variable where `FALSE` indicates there was only one adult in the household and `TRUE` indicates there was more than one adult in the household.

2.1 *p*-value

On the top line of Table 1, the authors report a p-value of 0.073. What is the null hypothesis being tested and what is this p-value the probability of?

2.2 Directed Acyclic Graph

Call the `make_model` function (plus `set_confound` and `set_restrictions`) in the CausalQueries package to specify a Directed Acyclic Graph (DAG) for the authors' model of the data-generating process for the purpose of estimating the causal effect of Medicaid *coverage*. Use the symbols in the `compacted` data.frame and you can ignore gender, age, language, zipcode, and previous vote history in your DAG as the authors did at the top of Table 1. Call `plot` on the resulting object to see the DAG.

2.3 Posterior Distribution

Call the `update_model` function in the CausalQueries package to obtain the posterior distribution of the unknown parameters conditional on the data and the DAG (you can use the default priors). You should also specify `data = "oregon", chains = 1` when calling `update_model`.

This will take a long time to run, so it is a good idea to knit it overnight or when you are taking a break. You should put the following into the header of your R chunk (the line that starts with three backticks: `posterior, cache = TRUE, results = "hide"` This way the result of the `update_model` call will be cached and it will not be run again each time you knit (unless you change your code). There is an example of this feature of RMarkdown in Week06/Slides06.Rmd .

2.4 Interpretation

How would you describe the posterior distribution of the ...

- Average Treatment Effect of Medicaid
- Average Intent to Treat Effect of winning the Medicaid lottery
- Average Treatment Effect of Medicaid among those with Medicaid
- Average Treatment Effect of Medicaid among those without Medicaid
- Average Treatment Effect of living in a household with more than one adult

2.5 Citizenship

Legal immigrants with sufficiently low income are eligible for Medicaid in Oregon but are not eligible to vote (unless they become citizens). However, the researchers did not collect data on which people are citizens, which is difficult because collecting such data tends to make legal immigrants less likely to participate in a study. Discuss to what extent not including a citizenship variable in the DAG is a problem for estimating the Average Treatment Effect of Medicaid with this dataset.