GR5065 Homework 3

Ben Goodrich

Due March 7, 2022 at 4PM

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

Read Baicker and Finkelstein (2019)

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

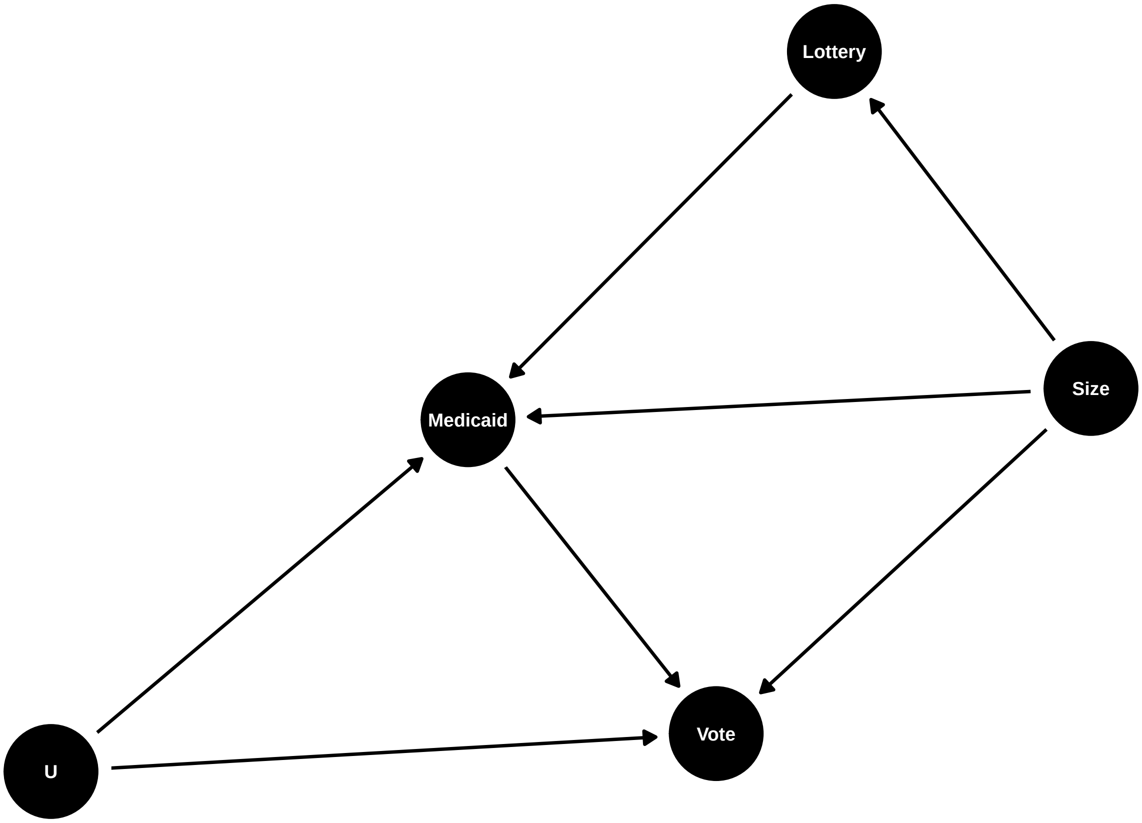
including the appendices. The essence of it is that in 2008, the state of Oregon conducted a lottery among households that entered to decide who would be eligible to apply for government-provided health insurance (Medicaid). It is rare to have a randomized variable in such a large dataset that could 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 simply did not follow up and many who did turned out not to be eligible because their income was not low enough. Also, 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")))
```

A basic DAG for the base model in the paper (although you could add more variables) is

```
library(dagitty)
library(ggplot2)
library(ggdag)
G <- dag("{ Size -> Lottery -> Medicaid -> Vote; Size -> Vote;
           Size -> Medicaid;
           Medicaid <- U -> Vote;
           U [unobserved] }")
ggdag(G, text_size = 2.5) + theme_void()
```



where `Size` is the number of adults in the household, `Lottery` is a binary variable for whether any adult in the household wins the Medicaid lottery and thus can apply for `Medicaid`, which is a binary variable indicating whether the household has Medicaid, and `Vote` is whether the adult in question votes in the November 2008 presidential election between Barack Obama and John McCain.

1 Frequentist

1.1 Confounding

What is at least one unobserved variable (U) that is a parent of both whether the person's household has Medicaid and whether the person votes, which leads to confounding, i.e. `Medicaid` $\leftarrow U \rightarrow$ `Vote`

1.2 Intent-to-Treat

As can be seen from

```r
adjustmentSets(G, exposure = "Lottery", outcome = "Vote")
```

```
## { Size }
```

the Average Causal Effect of winning the Medicaid lottery (irrespective of whether the household actually got Medicaid) can be identified simply by conditioning on household size (because `Medicaid` is a collider that naturally blocks the path from `Size` to `U` to `Vote`).

According to the paper, it is absolutely critical to condition on the number of adults in the household because if any adult wins the Medicaid lottery, then the entire household can apply for Medicaid (and get it if it satisfies the eligibility requirements). Thus, larger households have a greater chance of winning the Medicaid lottery, which may also be a parent of the outcome variable (in this case `Vote`).

If you execute,

```r
with(oregon, table(Lottery = treatment, Vote = vote_presidential_2008_1, Size = numhh_list))
```

```
## , , Size = signed self up
##
##        Vote
## Lottery     0     1
##       0 24912 12787
##       1 13027  6802
##
## , , Size = signed self up + 1 additional person
##
##        Vote
## Lottery     0     1
##       0  4913  2458
##       1  6424  3441
##
## , , Size = signed self up + 2 additional people
##
##        Vote
## Lottery     0     1
##       0    17     1
##       1    94    46
```

you can see the number of people with each combination of these three variables (note that `oregon$treatment` indicates whether someone won the Medicaid lottery even though we ordinarily would say that having Medicaid is the treatment variable).

For each of the three levels of `Size` — which are basically whether the person signed up 0, 1, or ≥ 2 additional adults from their household for the Medicaid lottery — calculate the difference in the proportion of voters between Medicaid lottery winners and Medicaid lottery losers. Then, created a weighted sum of those three differences, where the weight is the proportion of observations that have that value of `Size`. This weighed sum is a point estimate of the so-called Intent-to-Treat (ITT) effect, because the researchers "intended" for the Medicaid lottery to completely determine whether the household gets Medicaid (but they knew it would not work out that way perfectly). This weighed sum is also referred to as the stratified estimator in the Elwert reading.

1.3 Fisher Transformation

One of Fisher's most striking results

https://en.wikipedia.org/wiki/Fisher_transformation

is that, under the null hypothesis that a correlation, ρ, is zero in the population, the inverse hyperbolic tangent (`atanh` in R) of the estimated correlation, $\widehat{\rho}$, is asymptotically standard normal (across datasets of size N) with an expectation zero and standard deviation $\sqrt{\frac{1}{N-3}}$.

Set

```r
N <- nrow(oregon) # 74922
S <- 1000         # number of simulations
```

then, repeat the following `S` times:

1. Draw `N` realizations of V from a standard normal distribution
2. Draw `N` realizations of W from a standard normal distribution (independent of V)
3. Estimate the sample correlation between V and W using the `cor` function
4. Transform the estimated sample correlation with the `atanh` function

Plot the `S` realizations of (4) (using `geom_density` or `density`) and compare it to the PDF of a normal distribution (using `geom_function` or `curve`) with expectation zero and standard deviation $\sqrt{\frac{1}{N-3}}$. How close are these two densities?

1.4 Sensitivity to the Bivariate Normal Assumption

Fisher derived this transformation under the assumption that the two variables are bivariate normal with some correlation ρ (that is not necessarily zero). However, in the Oregon case, most of the relevant variables are binary. Repeat the previous subproblem, except draw V and W independently from Bernoulli distributions with expectation $\frac{1}{2}$. Now how close is the density of the estimated transformed correlation and the normal PDF with expectation zero and standard deviation $\sqrt{\frac{1}{N-3}}$?

1.5 First Stage of Instrumental Variables

As can be seen from

```
adjustmentSets(G, exposure = "Medicaid", outcome = "Vote")
```

There is no set of observed variables that blocks the non-causal path from `Medicaid` to `Vote`, which would be necessary to satisfy the Adjustment Criterion to identify the Average Causal Effect of Medicaid on voting. However,

```
instrumentalVariables(G, exposure = "Medicaid", outcome = "Vote")
```

```
##  Lottery |  Size
```

indicates that `Lottery` is an instrumental variable that — if you also condition on `Size` — can be used to identify the Local Average Treatment Effect (LATE). The LATE is also called the Compiler Average Treatment Effect because it is the proportion of people who are "compilers" in the Principal Stratification scheme:

- Compliers: People who would obtain Medicaid if and only if they won the Medicaid Lottery
- Defiers: People who would obtain Medicaid if and only if they did not win the Medicaid Lottery
- Always takers: People who would obtain Medicaid regardless of whether they won the Medicaid Lottery
- Never takers: People who would not obtain Medicaid regardless of whether they won the Medicaid Lottery

A person can be a "Never taker" if their household is not actually eligible for Medicaid (or other reasons). A few "Always takers" got Medicaid despite not winning the Medicaid lottery, and the paper is assuming they would have gotten Medicaid if they had won the Medicaid lottery. Thus, neither "Never takers" nor "Always takers" contribute to the estimation of the effect of Medicaid on voting. The paper is also assuming there are no "Defiers"; i.e., everyone who got Medicaid despite not winning the Medicaid lottery is an "Always taker". Thus, the LATE is just the Average Causal Effect of `Medicaid` on `Vote` among "Compliers" only and can be estimated using instrumental variable techniques, which have a long history in economics.

In theory, if `Size` were also a parent of `Vote`, then you would need to stratify on `Size` to isolate the LATE, which is what you should do in actual scientific research. However, in practice, the effect of the person's household size on the probability that they vote turns out to be very small, so for this homework, it is simpler to disregard it so that all the (relevant) observed variables are binary.

A point estimate of the LATE can be obtained in several equivalent ways:

1. The ratio of the covariance between `Lottery` and `Vote`, divided by the covariance between `Medicaid` and `Vote`
2. The ratio of the standard deviation of `Vote` to the standard deviation of `Medicaid`, times the ratio of the correlation between `Lottery` and `Vote` divided by the correlation between `Lottery` and `Medicaid`

3. The ratio of the difference between the proportion of voters among lottery winners and lottery losers, divided by the difference between the proportion of people with Medicaid among lottery winners and lottery losers.
4. Using `AER::ivreg` (which also handles other situations with non-binary variables and / or more conditioning)

The first three forms clearly (and thus the fourth equivalent form, albeit less clearly) have a problem if the denominator in the population is zero, in which case the LATE is not a well-defined thing to estimate because there are no "Compliers".

Thus, the first order of business in a Frequentist analysis is to test (and hopefully reject) the null hypothesis that the denominators are zero. This is perhaps easiest to do with form (2) because the standard deviations are necessarily positive, so we can just test the null hypothesis that the correlation between winning the Medicaid lottery and actually obtaining Medicaid insurance is zero. Using the theory from the previous two subproblems, calculate the p-value for this null hypothesis (noting that `oregon$treatment` indicates whether the person's household won the Medicaid lottery and `oregon$ohp_all_ever_nov2008 == "Enrolled"` indicates whether the person has Medicaid insurance) against the one-sided alternative hypothesis that the correlation is positive.

1.6 Second Stage of Instrumental Variables

If, and only if, we reject the null hypothesis that the denominator is zero, we can proceed to test the null hypothesis that the LATE is zero by testing the null hypothesis that the correlation between winning the Medicaid lottery and voting in 2008 is zero. Using the theory from the previous three subproblems, calculate the p-value for this null hypothesis (noting that `oregon$vote_presidential_2008_1` indicates whether the person voted) against the one-sided alternative hypothesis that the correlation is positive.

If, and only if, we reject the null hypothesis that the numerator is zero, we can estimate the LATE by multiplying the ratio of these two correlations by the ratio of the standard deviation of voting to the standard deviation of having Medicaid. What is this estimate? Note that it should be close to but not exactly identical to the corresponding LATE estimate in the paper because we are ignoring household size for simplicity.

1.7 Joint Test

In the previous two subproblems, we have tested the null hypothesis that a correlation is zero, irrespective of the other correlation. In order to test the null hypothesis that both correlations are zero simultaneously — against the alternative hypothesis that at least one is non-zero — we can take two (independent) test statistics that are distributed standard normal under the null hypothesis, square them, and add them together to obtain a single test statistic that is distributed χ^2 with two degrees of freedom under the null hypothesis that both correlations are zero. The CDF for the χ^2 distribution is called `pchisq` in R. What is the p-value for the test of the joint null hypothesis?

1.8 Distribution of the Instrumental Variable Estimator

The estimated standard error for the estimated LATE in this case is about 0.013395. Under the null hypothesis that the true LATE is zero, the estimated LATE is asymptotically distributed standard normal with expectation zero and some standard error that can be consistely estimated. Based on this information, can you reject a null hypothesis that the LATE is zero in favor of an alternative hypothesis that the LATE is positive? Is the conclusion compatible with the null hypothesis tests in the previous subproblems?

1.9 Confidence Interval Estimator for the LATE

An asymptotic 95% confidence interval estimator for the LATE is given by evaluating

```
qnorm(c(lower = 0.025, upper = 0.975), ...)
```

with mean equal to the estimated LATE and sd equal to its estimated standard error. Is the lower bound of the 95% confidence interval negative or positive? Is that compatible or incompatible with your conclusion in the previous subproblem? Why?

2 Bayesian

2.1 Directed Acyclic Graph

Call the make_model function in the CausalQueries package (and set_restrictions) to specify a DAG like that above, except ignore household size and instead take into account whether the person voted in the previous election (oregon$vote_preperiod). Plot this DAG.

2.2 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 but you will have to rename the relevant variables in oregon so that they correspond to the node names you used for the DAG in the previous subproblem.

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: {r} 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). You should not have any other code in that R chunk. There is an example of this feature of RMarkdown in Week06/Slides06.Rmd toward the end.

2.3 Interpretation

For each of the following, plot the posterior distribution and briefly describe what you see:

- Average Intent to Treat Effect of winning the Medicaid lottery
- Average Treatment Effect of Medicaid among compliers
- Average Treatment Effect of Medicaid among those with Medicaid
- Average Treatment Effect of Medicaid among those without Medicaid
- Average Treatment Effect of Medicaid among those who voted in the previous election
- Average Treatment Effect of Medicaid among those who did not vote in the previous election

Keep in mind that your results may not exactly align with those in the paper because we are incorporating oregon$vote_preperiod but not household size. You may want to refer again to section 9.2 of the CausalQueries reading.

2.4 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 (perhaps because they might be living in the same household as someone who is undocumented). 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.

3 Frequentist vs. Bayesian

Read this opinion piece in today's New York Times by Aubrey Clayton on the FDA's decision to pause Biontech / Pfizer's application to vaccinate children under the age of five. How would you apply Clayton's line of reasoning to the previous subproblems about Frequentist and Bayesian estimation of the effect of Medicaid on outcomes (perhaps things other than voting, such as measures of health or financial security)?