APSTA-GE 2123 Assignment 4

Due by 8:45 AM EST on April 20, 2020

1 Oregon Medicaid Experiment

In 2008, Oregon conducted an experiment where all people in some randomly-selected but poor households were granted an opportunity to enroll in Medicaid, which is a public health insurance program. Researchers then analyzed whether this intervention had a causal effect on a variety of outcome measures.

Pre-analysis plans and priors go hand-in-hand, even though these particular researchers only used Frequentist methods. Read

https://www.nber.org/oregon/documents/analysis-plan/analysis-plan-labor-2013-01-22.pdf

especially the parts about how they planned to estimate the effect that the opportunity to enroll in Medicaid has on future income of individuals. To keep things simpler, we are going to focus on the so-called "intent-to-treat" estimation, rather than the "local-average-treatment-effect". The difference arises because although some households were randomly given the *opporunity* to enroll in Medicaid, a non-random subset of those households subsequently filled out the paperwork to *actually* obtain health insurance. Although estimating the effect of having health insurance is more interesting, doing so is more complicated and is often done with so-called "instrumental variables". So, we will only consider the effect of having the *opportunity* to obtain Medicaid on individual income, which should be a lower bound on the effect of actually having Medicaid.

The authors describe the applicable theory as being "ambiguous". On one hand, if people have health insurance, they should go to the doctor more and become or remain sick less, in which case they could earn more money by missing fewer days of work. On the other hand, economists take seriously the idea that if the government subsidizes health insurance, then that could reduce the incentive of some individuals to work in order to pay for medical care out-of-pocket.

You can simulate the predictors with code like

```
J <- 50000 # number of households
dataset <- data.frame(household_ID = as.factor(unlist(lapply(1:J, FUN = function(j) {
  rep(j, each = sample(1:3, size = 1, prob = c(0.5, 0.3, 0.2)))
})))
selection <- rbinom(nrow(dataset), size = 1, prob = 0.2)
dataset$lottery <- ave(selection, dataset$household_ID, FUN = any)
dataset$numhh <- as.factor(ave(dataset$lottery, dataset$household_ID, FUN = length))
head(dataset)
```

```
##   household_ID lottery numhh
## 1            1       0     1
## 2            2       0     1
## 3            3       0     1
## 4            4       1     1
## 5            5       0     2
## 6            5       0     2
```

Here

- `household_ID` is a factor indicating which household an individual belongs to
- `lottery` is a binary variable indicating whether *anyone* in the household won the Medicaid lottery, in which case *all* individuals in the household have the opportunity to obtain Medicaid
- `numhh` is the number of adults in the household, which can be between 1 and 3

As can be seen from

```
table(numhh = dataset$numhh, lottery = dataset$lottery)
```

```
##       lottery
## numhh     0     1
##     1 19921  5002
##     2 19162 10864
##     3 15462 14730
```

there is a positive association between household size and opportunity to enroll in Medicaid, and the lottery is only random among households of a given size. Thus, the authors plan to condition on `numhh` in all of their analyses.

1.1 Actual Prior Predictive Distribution

Draw once from the prior predictive distribution of individual income for each of the N people in the simulated dataset using Generalized λ Distributions for the priors on each of the unknowns, namely the effect of being given the opportunity to enroll in Medicaid and the effect of being in a household with 1, 2, or 3+ adults. You can either use index variables (in which case there is no global intercept) or you can make two dummy variables that indicate whether the household has 2 or 3+ adults relative to the reference category of a household that has only one adult in it (in which case the intercept refers to a household with one adult that did not win the Medicaid lottery). In addition, there is a σ parameter for the standard deviation of the errors in predicting individual income with only `lottery` and `numhh`.

Somehow show that your prior predictive distribution is plausible in the aggregate. Some things that are not plausible include predicting that some people earn negative incomes or predicting that some people will become rich (or even lower middle class) as a result of winning the Medicaid lottery. The average individual income of someone who is poor enough for Medicaid was about $\$14,700$ in 2007.

1.2 Prior Predictive Distribution for a Journal

If you were planning to publish the results of this experiment in a journal, you would undoubtedly be required to assume that the treatment effect has a prior median of zero. Redo the previous subproblem under this assumption but increase the amount of prior uncertainty for the treatment effect so that there is about a 30% chance that the treatment effect is greater than the prior median you used for the treatment effect in the previous subproblem.

2 2018 American Community Survey

Each year, the United States governments surveys a large number of people to ask them a variety of questions that mostly pertain to the financial situation. We are going to use the 2018 version of this survey, which is documented at

https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html

Direct-message Ben on CampusWire to get the link to the dataset for the state you are supposed to analyze, which will be in the form of a zip file. Download that zip file once to your working directory and unzip it to create the CSV file with the dataset, which you can then load into R and eliminate unnecessary variables with

```
dataset <- readr::read_csv(dir(pattern = "csv$"))
dataset <- dataset[ , !startsWith(colnames(dataset), prefix = "PWG")]
dataset <- dataset[ , !startsWith(colnames(dataset), prefix = "F")]
```

The outcome variable we are interested in is called `WAGP`, which is the "wages or salary income past 12 months" earned by the person. Thus, we exclude people with no (or missing) individual income

```
dataset <- dataset[!is.na(dataset$WAGP) & dataset$WAGP > 0, ]
```

2.1 Posterior Distribution

Create a plausible model for `log(WAGP)` that is linear in its parameters using some subset of the other variables as predictors. Note that you can create interaction terms, polynomials, etc. Use `stan_lm` in the rstnanarm package to obtain the posterior distribution of the parameters conditional on the data, which first requires you to choose the modal value of the prior R^2 like we did with the diamonds example in class Monday.

Which predictors in your model have coefficients that are positive with a very high posterior probability?

2.2 Influential Observations

Is there any evidence from the Pareto k shape estimates that any of the observations have an outsized influence on the posterior distribution? If so, which observations are those?

2.3 Posterior Predictions

Call the `posterior_predict` function, specifying the `draws = 100` and `FUN = exp` arguments to draw 100 times from the posterior predictive distribution of income (rather than log-income). The resulting matrix will be $100 \times N$, whose row-wise means are a posterior distribution for average wages among wage-earners. Create a histogram of these row-wise means. How would you describe your uncertainty about the average wage among wage-earners?

2.4 Topcoding

The actual `WAGP` are top-coded at

```
max(dataset$WAGP)
```

which means that even if the individual earned more than this value, it gets recorded as such in the public-use dataset in order to help preserve the anonymity of individuals. For each of the people in the dataset whose incomes are topcoded, what is your posterior expectation for their actual income?