APSTA-GE 2123 Assignment 2

Due by 8:45 AM EST on May 4, 2020

1 The Impact of Medicaid Expansion on Voter Participation

For this problem, we are going to reanalyze a recently-published paper entitled "The Impact of Medicaid Expansion on Voter Participation: Evidence from the Oregon Health Insurance Experiment" by Katherine Baicker and Amy Finkelstein, which is available from

https://www-nowpublishers-com.proxy.library.nyu.edu/article/Details/QJPS-19026

Read the paper and the appendices, but the essence of it was 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 where the (intermediate and final) outcomes could make a tangible 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 behavior.

To make things somewhat simpler, we will consider the "intent-to-treat" (ITT) estimates, where some outcome variable is modeled as a function of whether someone in the household won the lottery, the size of the household, and perhaps other control variables. However, not all households that won the lottery actually signed up for Medicaid, which would make it more difficult to estimate the perhaps more relevant causal effect of having health insurance. Thus, the ITT estimates the effect of winning the lottery, which is a lower bound for the effect of having health insurance.

Check with Ben on CampusWire for which part of which table you should reanalyze. Not that all of the estimates in those tables are multiplied by 100 for some reason. Also, it is too much effort to get the standard errors (and p-values) from R to match those in the paper, which are clustered by household. Finally, your R formula should refer to the factor called `numhh_list`, which indicates the household size, rather than `nnnnumhh_li_2` and `nnnnumhh_li_3` which are just dummy variables created to indicate the second and third categories respectively, compared to the first.

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(brms)
library(haven)
unzip("100.00019026_supp.zip")
oregon <- as_factor(read_dta(file.path("19026_supp", "Data", "individual_voting_data.dta")))
table(oregon$treatment) # this indicates who won the Medicaid lottery
```

```
##
##     0     1
## 45088 29834
```

If you understand basic Stata syntax, it may be helpful to refer to the text file called `oregon_voting_replication.log` in the `19026_supp` directory.

1.1 Priors and Prior Predictive Distribution with brms

For the renalysis of the part of the table assigned to you, choose reasonable priors for the parameters in a Bernoulli model (even though the original paper does not use Bernoulli models). You cannot use the default

priors, although `get_prior(...)` with `family = bernoulli` will tell you what everything is called.

Then call `brm` and specify the argument `sample_prior = "only"` to draw from the prior distribution without conditioning on the observed data. The resulting object can be passed to the `pp_expect` to calculate the prior distribution of μ_n, which is the probability that $y_n = 1$. Demonstrate somehow that these prior success probabilities are fairly reasonable.

1.2 Posterior Distribution

Proceed as in the previous problem but do not specify `sample_prior = "only"`, in which case `brm` will condition on the observed data to produce draws from the posterior distribution of the parameters. What is the posterior probability that winning the Medicaid lottery has a positive effect?

1.3 Alternative Model

Consider some other variable(s) in the `oregon` dataset that you could use as predictors. When added to your model (with appropriate priors on their coefficients), does the estimated Expected Log Predictive Density (ELPD) improve or worsen? You may need to call the `loo_subsample` function in the brms package to estimate the ELPD on a subset of observations if your computer does not have enough RAM to estimat the ELPD over all observations. The `loo_subsample` function works just like the `loo` function.

2 Coronavirus in NYC

Download data on the number of confirmed coronavirus cases by day in New York City with

```
ROOT <- "https://raw.githubusercontent.com/nychealth"
NYC <- readr::read_csv(paste0(ROOT, "/coronavirus-data/master/case-hosp-death.csv"))
```

The outcome of interest is called `NEW_COVID_CASE_COUNT` and the only predictor is the number of days that have passed, which you can create via

```
NYC$day <- 1:nrow(NYC)
```

2.1 Negative Binomial Model

You can include time and its square in a model by using the formula `NEW_COVID_CASE_COUNT ~ poly(day, degree = 2, raw = FALSE)`. However, this utilizes orthogonal polynomials, which are preferable to raw polynomials for numerical stability reasons. Choose priors (again, you cannot use the defaults) and draw from the posterior distribution of a negative binomial model using `brm`.

2.2 Poisson Model

Draw from the posterior distribution of the same model but specify `family = poisson` to change the likelihood to the Poisson distribution, which is a limiting case of the negative binomial.

2.3 Model Comparison

What graphs and numbers can you produce to substantiate a claim that one of these two models is preferable to the other in this case?

2.4 Posterior Prediction

Create a new dataset that extends the `day` variable an additional seven days. Call the `posterior_predict` function on your preferred model from the previous subproblem, specifying the `newdata` argument to be your newly created dataset. How would you describe your implied beliefs about the number of new confirmed coronavirus cases over the coming week?