GR5065 Assignment 4

Due by 4PM EST on April 2, 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.ezproxy.cul.columbia.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(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 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 Monotonic Predictor

Use `brm` with `family = gaussian` (even though the outcome variable is binary) to estimate a Bayesian version of the ITT model in question but specify `mo(numhh_list)` to constrain the effect of household size

to be monotonic. You cannot use the default priors. Also, you should specify the non-default argument `save_all_pars` to `TRUE` when you call `brm` in order to use bridge sampling below.

1.2 Bernoulli Likelihood

Estimate a model with the same formula as in the previous subproblem but use `family = bernoulli` to utilize a Bernoulli likelihood. You still cannot use the default priors. Also, you should specify the non-default argument `save_all_pars` to `TRUE` when you call `brm` in order to use bridge sampling below.

1.3 PSISLOOCV

Use the `loo` function in the brms package to compare the previous two models based on which is expected to predict future data better. Note that you may have to specify `pointwise = TRUE` when you call `loo` in order to reduce the amount of memory consumed for this to be feasible.

1.4 Stacking Weights

Use the `loo_model_weights` function in the loo package to find the vector of weights such that the weighted sum of the two model's predictions is expected to best predict future data. How similar is the conclusion you reach to the conclusion you reached in the previous subproblem?

1.5 Posterior Probability Over Models

Use the `bridge_sampler` function in the brms package to compute `bridge` objects for each of your two models. Then, call the `error_measures` function in the bridgesampling package to verify that the estimated errors are small. Finally, call the `post_prob` function in the bridgesampling package to compute the posterior probability that each of these two models is true, conditional on one of them being true. How similar is the conclusion you reach to the conclusion you reached in the previous subproblem?

1.6 Projection Pursuit

Take the preferred model from the previous subproblem but refit it without constraining the effect of household size to be monotonic. Then, use the functions in the projpred package to try to find a submodel of this model that is expected to predict future data almost as well. Does that submodel include the `treatment` variable as a predictor? How does this conclusion relate to the conclusion in the paper as to whether the effect of the `treatment` variable was or was not statistically significant?

1.7 Unbiasedness

When all of the predictors are categorical and one of the predictors is randomized, it is typical to see papers (such as this one) use least squares to estimate a model even when the variable(s) that they are trying to predict are binary. In this situation, all of the $\widehat{y}$ values will be between 0 and 1 and thus could be interpreted as estimated probabilities. The errors will not be normally distributed, but it is possible to calculate "robust" standard errors of the estimates that are consistent (even though the errors will not be normally distributed) in order to perform a test of the null hypothesis of no treatment effect. But perhaps the most relevant consideration is that the least squares estimator is unbiased across datasets where the predictors are *fixed*, whereas some form of a logit or probit model could be biased if the chosen inverse link function is not correct.

How would a Bayesian respond to the considerations in the previous paragraph? Why might one prefer the Bayesian results in this homework to the Frequentist results in the paper?

2 General Social Survey

The 2018 wave of the General Social Survey (GSS) can be downloaded with a webbrowser from

http://gss.norc.org/Documents/stata/2018_stata.zip

to your working directory once. Then unzip it to create a large Stata-formatted dataset and a PDF codebook.

You can then load it into R with something like

```
GSS <- as_factor(read_dta("GSS2018.dta"))
```

The GSS contains over a thousand variables, many of them ordinal, which you can read about at

http://gss.norc.org/Get-Documentation

Pick an ordinal one to use as your outcome variable and choose a reasonable set of predictor variables (including interactions and polynomials) for that ordinal outcome. Note that you may need to recode IAP (inapplicable) and DK (don't know) into `NA` for some variables

2.1 Prior Predictive Distribution

Call the `brm` function in the brms package, specifying the `prior` argument as well as `sample_prior = "only"` in order to draw from the prior distribution, i.e. without conditioning on the observed outcome. Use `posterior_predict` and / or other functions to establish whether your model implies a reasonable distribution for what your outcome variable could look like.

2.2 Posterior Distribution

Call the `brm` function again but with `sample_prior = "no"` (which is the default) to draw from the posterior distribution, i.e. conditioning on the observed outcome. What substantive conclusion would you draw from the results?

2.3 Posterior Predictive Checks

Use the `pp_check` function to produce evidence that your model does or does not fit the data well.