GR5065 Homework 4

Ben Goodrich

Due March 16, 2021 at 8PM New York Time

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

1 Minimum Wage Increases

Read Alan Manning's paper, including the appendices, which addresses the question of why increasing the minimum wage seems to have a noticeable effect on wages but has an "elusive" effect on the level of employment, which is to say that economists have not found the large adverse effects on employment that they theorized decades ago would occur.

The actual dataset Manning uses is not yet available (apparently due to a problem with the OpenICPSR website), but I was able to approximately reconstruct what the dataset should have been using various sources mentioned in the appendix, which can be loaded into R via

```
Manning <- readRDS("Manning.rds") # assuming your working directory is HW4
str(Manning)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    2009256 obs. of  7 variables:
##  $ time     : Ord.factor w/ 164 levels "1979.1"<"1979.2"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ state    : chr  "Maine" "Maine" "Maine" "Maine" ...
##  $ age      : num  27 28 24 24 23 24 24 29 25 21 ...
##  $ rw       : num  30.3 18.12 9.72 13.18 8.73 ...
##  $ teen_perc: num  8.91 8.91 8.91 8.91 8.91 ...
##  $ unem_rate: num  5.48 5.48 5.48 5.48 5.48 ...
##  $ min_mw   : num  2.9 2.9 2.9 2.9 2.9 ...
##   ..- attr(*, "label")= chr "Quarterly State Minimum"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  - attr(*, "label")= chr "CEPR ORG Extract, Version 2.5, 1979"
##  - attr(*, "notes")= chr  "1" "Age 16 and older only"
##  - attr(*, "na.action")= 'omit' Named int  1 2 3 5 6 7 10 14 16 17 ...
##   ..- attr(*, "names")= chr  "1" "2" "3" "5" ...
```

In the `Manning` dataset, the unit of observation is a person-quarter, i.e. a person whose data are observed during some quarter of the year but the same people are not observed in multiple quarters:

- `time` is an ordered factor from quarter 1 of 1979 to quarter 4 of 2019
- `state` is the state (or the District of Columbia, which is techically not a state) that the person lives in
- `age` is the age of the person in years
- `rw` is the person's real hourly wage, which is to say that it is measured in 2019 dollars after adjusting for inflation between 2019 and whenever the person's nominal wage was observed
- `teen_perc` is the percentage of teens in the total population for the `state` that the person lives in during that `time` period
- `unem_rate` is the "prime-age" unemployment rate, i.e. the percentage of people between 25 and 54 who do not have a job but are looking for one relative to the number of people in the work force , for the `state` that the person lives in during that `time` period

- `min_mw` is the maximum value between the federal minimum wage and the minimum wage in the `state` that the person lives in during that `time` period, which is in nominal dollars (i.e. not adjusted for inflation)

1.1 Frequentist Inference

Manning uses Ordinary Least Squares and states with regard to Figures 1 through 4 that "we will summarize results and confidence intervals for the coefficient on the log minimum wage in these regressions". To what extent are the inferences Manning draws Frequentist, in either the sense of Fisher or Neyman? For which of the inferences Manning wants to make should he have used Bayesian methods?

1.2 Bayesian Inference

Use the `stan_lm` function in the rstanarm package to estimate the parameters of the model (where the logarithm of the real wage is the outcome variable) listed in "Specification 2" of Table 2 separately for people whose age is in $[16, 19]$ and whose age is in $[20, 24]$. You might need to read `help(formula)` to understand how R processes the syntax but basically

- Character or factor variables in R will be converted into a sequence of dummy variables that is one less than the number of levels of the categorical variable, with no dummy variable for the reference category
- The "state time trend" can be created via an interaction term, `state : as.integer(time)`
- You can (and should) use the `log` function inside a formula to do transformations

You will need to specify both `prior_intercept`, which can be a call to the `normal` function whose `location` and `scale` arguments should correspond to your beliefs about the distribution of the average log real wage among people in a particular age bracket. Technically, it should be the distribution of the log real wage for a person with average values on all predictors, but that number is not too different from the average log real wage in linear models. You will also need to specify `prior = R2(...)` to express your prior beliefs about the proportion of variance in log real wages among people in that age bracket that is attributable to all the predictors in "Specification 2" of Table 2 under a linear model. Finally, you should pass an integer to the `seed` argument to make Stan deterministic.

1.3 Interpretation

How would you describe your posterior beliefs about the coefficient(s) on the logarithm of the minimum wage (in the two models)? To what extent are your conclusions similar to those of Manning?

1.4 Prediction

You may have seen in the news that Democrats like Bernie Sanders have been pushing to raise the federal minimum wage to $15 per hour, which would be much greater than its current value and larger than in most states or cities. However, such a change in the law is opposed by all Republicans in the Senate, and several of them are needed to at least allow a vote to happen on the minimum wage (as opposed to preventing the vote using a fillibuster).

You can select the most recent observations on teenagers with something like

```
recent <- dplyr::filter(Manning, age <= 19, time == "2019.4")
```

Call the `posterior_predict` function on the object created by `stan_lm` with `newdata = recent` in order to obtain posterior predictions for log wages, conditional on all of the past data. Then, use the `exp` function to convert these predictions to dollars.

Next, create a counterfactual dataset where the federal minimum wage is raised to $15 if the state's minimum wage is less than $15 via

```
recent_ <- dplyr::mutate(recent, min_mw = pmax(min_mw, 15))
```

and call the `posterior_predict` function on the object created by `stan_lm` with `newdata = recent_` to obtain predictions of log wages that can be mapped to dollars via `exp`.

How would you describe your posterior beliefs about the effect on teenage wages of increasing the federal minimum wage to $15 by state (and the District of Columbia) if it were to have gone into effect in 2019?

2 Voter Turnout in France

Read Andy Eggers' paper on voter turnout in proportional representation (PR) systems compared to plurality-rule systems. In a PR system, voters vote for a political party from a list and the resulting proportion of legislators from each political party is (roughly, perhaps due to things like winner bonuses) equal to the proportion of votes that the party receives. In a plurality-rule system, voters vote for a named candidate (who typically is a member of a political party) and the candidate with the most votes wins the election. The resulting proportion of legislators each political party is equal to the proportion of winning candidates.

It is widely believed that a somewhat larger percentage of eligible voters turn out to vote in PR systems than in plurality-rule systems, but it is difficult to take into account all of the other factors that could drive voter turnout that may be associated (causally or not causally) with the political system. Eggers' estimates this difference in voter turnout among small municipalities in France where there is a national rule that municipalities with at least $3,500$ people must have a PR system and municipalities with fewer than $3,500$ people must have a plurality-rule system. The idea of "regression discontinuity designs" is that the causal effect of a municipality going from a population of $3,499$ to $3,500$ is attributable to the change from a plurality-rule to a PR system, as opposed to merely having one additional person to represent or any other variable that is correlated with the population. Of course, there are not enough municipalities whose population is exactly $3,499$ in one year and exactly $3,500$ the next, but there are many municipalities whose population is close to the threshold of $3,500$. Eggers then chooses a "bandwidth window" for the population that is centered at $3,500$ to estimate a linear regression model on that subset of the data. In this homework, you do not have to worry about how the "bandwidth window" was chosen.

You can load the dataset used in Eggers' paper with

```
Eggers <- readRDS("Eggers.rds") # assuming your working directory is HW4
summary(Eggers$rrv) # log population where 0 corresponds to a municipality with 3500 people
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -8.161  -2.990  -2.218  -2.077  -1.324   4.714
```

```
Eggers$PR <- as.integer(Eggers$rrv >= 0) # has a PR system
```

2.1 Drawing from the Prior Predictive Distribution

As explained on page 144, Eggers estimates a basic model (before adding additional covariates, which you do not need to worry about on this homework) where the percentage of voter turnout is conditionally normal, given the population of the municipality. There are five unknown parameters to estimate:

1. β_0, which is the intercept
2. τ, which is the coefficient on the dummy variable for a PR system (due to population being at least $3,500$)
3. β_1, which is the coefficient on the logarithm of the ratio of population to $3,500$ so that `rrv` would be zero in the data if the population were exactly $3,500$
4. β_2, which is the coefficient on the interaction between `PR` and `rrv`, which can be interpreted as how much more sensitive voter turnout is to log population in PR systems than in plurality-rule systesm

5. σ, which is the standard deviation of the errors when predicting the percentage of voter turnout with only the previous four variables only

Draw 1000 realizations from the prior predictive distribution of voter turnout for each of the $N = 35891$ municipalities in the dataset using the Generalized Lambda Distribution to draw from your priors about each of these five unknown parameters. You can load the necessary functions to do so (if your working directory is `HW4`) by calling

```
source(file.path("..", "..", "Week05", "GLD_helpers.R")) # for GLD-related functions
```

In this case, you should *not* center the predictors because `Eggers$rrv` is already constructed so that it is zero when the population is exactly $3,500$. Thus, the intercept, β_0, can essentially be interpreted as the expected percentage of voter turnout for a municipality that has $3,499$ people and thus a plurality-rule system.

2.2 Checking the Prior Predictive Distribution

Show that (the realizations from) your prior predictive distribution seem reasonable for all municipalities with populations between $1,750$ and $5,250$, which you can achieve by filtering on the `Eggers$PSDC99` variable which contains the raw population of each municipality.

Then, show that (the realizations from) your prior predictive distribution seem unreasonable for Toulouse, which is the municipality with the largest population $(390,350)$.

Finally, explain how the differences in the prior predictive distribution between Toulouse and municipalities with populations between $1,750$ and $5,250$ would provide motivation to only condition on the municipalities with populations between $1,750$ and $5,250$ when estimating the causal effect of switching from a plurality-rule system to a PR system at $3,500$.

2.3 Posterior Distribution

Use the `stan_glm` function in the rstanarm package to draw from the posterior distribution of these five unknown parameters conditional on the data from municipalities with populations between $1,750$ and $5,250$ only, where the outcome variable (`to.2008`) is the percentage of voter turnout in the 2008 municipal elections. Since the `stan_glm` function does not yet support priors based on the Generalized Lambda distributions you can use normal priors on the intercept and coefficients with `location` equal to your prior median and `scale` equal to your prior interquartile-range. Note that the `prior_intercept` argument to `stan_glm` is separate from the `prior` argument, which pertains to the coefficients and can take a vector of size 3 for `location` and `scale` whose elements correspond to your beliefs about τ, β_1, and β_2. The `prior_aux` argument specifies the prior for σ, which you can take to be the exponential distribution with `rate` parameter equal to the reciprocal of your prior median. Finally, you should pass an integer to `seed` to make Stan deterministic.

2.4 Interpretation

How would you describe your posterior beliefs about τ?

2.5 Prediction

There are some municipalities with populations between $1,750$ and $5,250$ where `to.2008` is missing for some reason (and thus were dropped when you called `stan_glm`). You can subset to those observations with something like

```
Eggers_missing <- dplyr::filter(Eggers, PSDC99 >= 1750, PSDC99 <= 5250, is.na(to.2008))
Eggers_missing$to.2008 <- NULL
```

Call the `posterior_predict` function on the object created by `stan_glm` with `newdata = Eggers_missing` to draw from the posterior predictive distribution of voter turnout in these municipalities, conditional on the municipalities with populations between $1,750$ and $5,250$ where `to.2008` was observed. Do these distributions seem reasonable? Why or why not?