GR5065 Homework 6

Ben Goodrich

Due April 13, 2021 at 8PM New York Time

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

1 Household Pulse Survey

Download (once) the Week 25 Public Use File of the Household Pulse Survey from

https://www2.census.gov/programs-surveys/demo/datasets/hhp/2021/wk25/HPS_Week25_PUF_CSV.zip

to your working directory. From there, it can be unzipped and loaded into R via:

```
unzip("HPS_Week25_PUF_CSV.zip")
library(readr)
pulse <- read_csv("pulse2021_puf_25.csv")
```

The wording of the survey questions and answers is given at

https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/Phase%203%20Questionnaire_02.25.21_English.pdf

1.1 Posterior Distribution

The binary outcome variable of interest is Q9: "Have you, or has anyone in your household experienced a loss of employment income **since March 13, 2020**?" This variable is called `WRKLOSS` in `pulse`. The primary predictor of interest is Q5: "What is the highest degree or level of school you have completed?" This variable is called `EEDUC` in `pulse`. We are going to assume that the log-odds of a loss of employment income is monotonically decreasing in the level of schooling completed. You should include some other predictors of loss of employment income that may or may not be ordinal and if any of them are ordinal, may or may not be monotonically related to the log-odds of employment income loss. It may be helpful to look at the spreadsheet that was unzipped in your working directory called pulse2021_data.dictionary_CSV_25.xlsx to understand how the variables in `pulse` correspond to the questions in the survey.

Use the `brm` function in the brms package to draw from the posterior distribution of the parameters. You should use proper priors on all the parameters in your model.

1.2 Interpretation

How would you describe the posterior relationship between the log-odds of employment income loss and the level of schooling completed?

1.3 Frequentism

As is explained in the technical documentation https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/Phase3_Source_and_Accuracy_Week_26.pdf

the Household Pulse Survey was conducted by dividing the United States into 66 subregions and within each subregion, households were randomly sampled from the Census Bureau's database of (approximately 140 million) addresses. A complicated (but necessary) procedure was applied to ultimately create "household weights" that arguably are appropriate in this case because Q9 refers to "anyone in your household".

In the section entitled "Standard Errors and Their Use" (starting on page 10), the Census Bureau correctly defines Frequentist concepts, such as standard errors, confidence intervals, etc. They recommend estimating standard errors — which are nontrivial due to the stratified random sampling design as opposed to simple random sampling assumed by most software — by "successive difference replication". Basically, this entails obtaining an estimate once using (something proportional to) the `HWEIGHT` variable in `pulse` as weights and then 80 more times using as weights each of the 80 columns of

```r
pulse_weights <- read_csv("pulse2021_repwgt_puf_25.csv")[ , 3:82]
```

The estimated (by successive difference replication) standard error of the estimate of θ is

$$\sqrt{\frac{4}{80}\sum_{i=1}^{80}\left(\theta_i - \theta\right)^2}$$

Use the `glm` function that comes with R to obtain point estimates of the same parameters as in your model above specifying `weights = HWEIGHT / sum(HWEIGHT)`. Then, do likewise 80 times using the renormalized columns of `pulse_weights` as the `weights` argument in order to estimate the standard error of the coefficient on having a graduate degree using the above formula. Is this number approximately equal to the posterior standard deviation for this coefficient that you obtained earlier? Is this number conceptually similar to this posterior standard deviation?

2 General Social Survey

Install (once, outside of your .Rmd file) the gssr package via

```r
remotes::install_github("kjhealy/gssr")
```

at which point you can load the General Social Survey (GSS) cumulative (through 2018) dataset via

```r
data(gss_all, package = "gssr")
```

See the documentation of the gssr package at

https://kjhealy.github.io/gssr/

to obtain more information about the variables (from inside R) or go to

http://gss.norc.org/get-documentation

particularly Appendix V.

There are thousands of ordinal variables in the GSS. Choose one of them that has been asked in at least two different years to be your outcome variable. You can model it however you want using `brms::brm` but you should use proper priors on all of the parameters in your model, and your model should somehow allow for the possibility that the probability of someone responding in each outcome category can change over time (the `gss_all` data.frame has observations taken in different years, but the respondents are different each time).

2.1 Prior Predictive Distribution

Call `brm` and specify `sample_prior = "only"` to draw from the prior distribution of the parameters. You can then use `posterior_predict`, `posterior_epred`, etc. to explore your beliefs about the outcome under your model. Tweak your priors until all the outcome categories have roughly the same prior probability, aggregating across all of the observations you are conditioning on.

2.2 Posterior Distribution

Call `brm` omitting the `sample_prior` argument to draw from the posterior distribution of the parameters. How would you describe your posterior beliefs about some quantity that is entailed by your model?