PLSC 40502: Problem Set 2

YOUR NAME

January 19, 2024

This problem set is due at **11:59 pm on Wednesday, January 31st**.

Please upload your solutions as a .pdf file saved as "Yourlastname_Yourfirstinitial_pset2.pdf"). In addition, an electronic copy of your .Rmd file (saved as "Yourlastname_Yourfirstinitial_pset2.Rmd") must be submitted to the course website at the same time. We should be able to run your code without error messages. In addition to your solutions, please submit an annotated version of this `.rmd` file saved as "Yourlastname_Yourfirstinitial_pset2_feedback.rmd", noting the problems where you needed to consult the solutions and why along with any remaining questions or concerns about the material. In order to receive credit, homework submissions must be substantially started and all work must be shown.

Problem 1

The **Congressional Election Study** is an annual, large, nationally representative survey of the American population designed to understand how Americans view their representatives. We will be using this dataset to predict the level of support for U.S. trade policy, particularly the Section 232 Steel and Aluminum tariffs implemented by the Trump administration and currently in place under the Biden administration.

You will find the `CCES Guide 2020.pdf` very useful for understanding the definition and structure of different variables (see after page 27, the "common content") and this problem will assume that you are able to use this reference to identify relevant outcomes and covariates. Note that the numeric values in the CSV dataset below correspond in order to the categories listed in the PDF. For example, the Guide lists two responses for `gender`: Male and Female. Therefore, "Male" is coded as a 1 and Female is coded as a "2"

Be very careful in reading the variable names and definitions!

The code below will load in the Common Content from the 2020 CES. Please download the file directly from `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/E9N6PH` and place it into your data folder:

```r
ces <- read_csv("data/CES20_Common_OUTPUT_vv.csv")
```

Our main outcome of interest is `CC20_338b`, whether the respondent supports a policy of "25% tariffs on imported steel and 10% on imported aluminum, EXCEPT from Canada and Mexico." (1 = Yes, 2 = No)

We'll be predicting this using data on age, sex, race, education and income. Start by cleaning the data and setting up the preliminaries

Part A

We'll need to coarsen our covariates into a higher level of aggregation since some bins are very, very small. Specifically, we want to match the bins for sex, race, age, education and income used to generate the national vote breakdown in the CCES Guide (pages 25/26). Using the definitions given by those bins, and the following variables, generate a factor variable for gender, race, age and education (you should have four factor variables in total):

- `birthyr` Year of birth (also remember that this survey was conducted in 2020)
- `gender` - Gender of respondent (1 = Male, 2 = Female)
- `race` - "What racial or ethnic group best describes you?"
- `hispanic` - "Are you of Spanish, Latino, or Hispanic origin or descent?" (Note for the 'hispanic' category in the 5-category race variable, you should code respondents who answer yes to this question *or* "hispanic" to `race`)
- `educ` - "What is the highest level of education you have completed?"

You can validate that your coarsenings are correct using the "% of electorate" column in the national vote breakdown from pages 25/26 of the CCES Guide. Subset the data to respondents without missing `vvweight_post` (matched to voter file + filled out both waves) and without missing `CL_2020gvm` and take weighted averages using the sampling weights.

Generate a dataset that contains an indicator variable for whether the respondent supports the tariff `CC20_338b == 1` along with your four predictors and remove all missing data. Set the baselines for gender to male, age to "18-29", race to White and education to "High school or less".

To check your work, your final dataset should contain `60398` observations.

Part B

Next, let's write down a model. We'll assume that the data-generating process for the indicator $Y_i \in \{0, 1\}$ of whether the respondent supports the tariffs has the following DGP:

$$Y_i = \nVdash(Y_i^* > 0)$$

where $\nVdash()$ denotes the indicator function. That is, when Y_i^* is greater than 0, $Y_i = 1$ and when Y_i^* is less than or equal to 0, $Y_i = 0$.

$$Y_i^* \sim \text{Normal}(X_i'\beta, 1)$$

and

$$\beta \sim \text{Normal}(b_0, B_0^{-1})$$

where b_0 is a vector of hyperparameters denoting the prior mean of β and B_0 is a matrix of hyperparameters denoting the inverse of the prior variance-covariance matrix of β.

Start by classifying each of the variables in the DGP. What is the **observed data**? What is a **latent variable** and what is a **known constant**?

Part C

Our goal is to do inference on the posterior distribution

$$f(\beta|\mathbf{Y}, \mathbf{X}) \propto f(\mathbf{Y}|\beta, \mathbf{X})f(\beta|\mathbf{X}) \equiv f(\mathbf{Y}|\beta, \mathbf{X})f(\beta)$$

However, this is difficult. Instead, we simulate from the joint posterior that includes the $\mathbf{Y}^* = \{Y_1^*, Y_2^*, Y_3^*, \dots, Y_N^*\}$:

$$f(\beta, \mathbf{Y}^*|\mathbf{Y}, \mathbf{X})$$

We will then marginalize over this joint distribution by keeping the draws of β.

Use Bayes' rule to write this posterior distribution as proportional to a **likelihood** multiplied by a **prior**. Factor the density as much as possible and use the independencies and conditional independencies implied by the model to eliminate terms from the conditioning set of some of the densities. You should have three main component densities: one related to the prior on β and two likelihood terms.

Part D

What is the conditional distribution of $Y_i^*|Y_i, \beta, X_i$?

Hint: What do we already know is the conditional distribution of $Y_i^*|\beta, X_i$? What information does adding Y_i provide?

Part E

What is the form of the conditional distribution of $\beta|Y_i^*, Y_i, X_i$ (just the form/type of the distribution, you don't need to derive its mean/variance though this is a standard result)?

Hint 1: What can we get rid of on the right-hand side of the conditioning bar (using conditional independence)? Hint 2: Once we've gotten rid of that variable, what familiar posterior distribution from a different model does this remind you of?

Part F

Let's estimate the probit model to predict the probability of supporting the tariff conditional on age, sex, race and education level. For now, include the covariates additively, omitting any interactions (e.g. education-race). Use `model.matrix()` to generate the regression matrix $\mathbf{X}$.

Implement a Gibbs sampler to obtain draws from the joint posterior distribution. Use a fairly diffuse prior for β that sets $b_0 = 0$ and $B_0 = \frac{1}{10}\mathbf{I}$ where $\mathbf{I}$ is the identity matrix.

Pick some reasonable starting values for β and start by sampling from the conditional distribution of Y_i^* to generate "starting" values of Y_i^*. Iterate between sampling from the conditional distribution of β and the conditional distribution of the Y_i^*. You only need to store all of the samples from β (this should save on memory as you do not need to store all 2000 iterations of the $\approx 60,000$ draws from Y_i^*).

Set the starting seed to `60637` and run the sampler for `2000` iterations after discarding the first `500` as a "burn-in" (so you should run it for 2,500 iterations in total). My implementation took about 1.5 minutes to run on my desktop, so be aware that you may want to test your code with fewer iterations before launching the full chain.

Hint: You will find the`truncnorm` R package useful.

Hint 2: Using well-known results, the distribution from Part E has variance-covariance matrix

$$\mathbf{V} = (B_0 + \mathbf{X}'\mathbf{X})^{-1}$$

and mean

$$\mathbf{M} = \mathbf{V}(B_0 b_0 + \mathbf{X}'\mathbf{Y}^*)$$

Part F

Make some traceplots of your coefficients - does it look like the Gibbs Sampling chain has sufficiently "mixed"?

Part G

Plot the posterior distribution of the coefficient on "College Graduate". Obtain a posterior mean estimate along with a 95% credible interval. Interpret your results substantively.

Part H

Carry out a posterior predictive check. For each observation in the data, predict $\tilde{Y}_i$ given β across all of the sampled values of β (derive the probability that $Y_i = 1$ given the coefficients and the functional form and simulate from the bernoulli distribution). How well does the model predict the outcomes on average?

Part I

Implement the model above in Stan without the data augmentation step (define y conventionally using the normal CDF `Phi`). Compare your results to your gibbs sampler. Do they line up?