PLSC 40502: Problem Set 2

YOUR NAME

January 22, 2025

This problem set is due at **11:59 pm on Wednesday, February 5th**.

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.

Problem 1

This problem will have you implementing a Bayesian version of the classic probit regression model. The dataset comes from Barilleaux and Rainey (2014) which studies the decisions of U.S. state governors to accept or reject the Affordable Care Act's Medicaid expansion. The full citation is:

> Barrilleaux, Charles, and Carlisle Rainey. "The politics of need: Examining governors' decisions to oppose the "Obamacare" Medicaid expansion." State Politics & Policy Quarterly 14, no. 4 (2014): 437-460.

The code below will load in the dataset

```
aca <- read_csv("data/politics_and_need.csv")
```

Part A

We will need to first construct our outcome and covariates of interest. Create a binary indicator variable, `gov_opposes`, for whether a state's governor **opposes** medicaid expansion by transforming the `gov_position` variable. Your predictors/regressors will be:

- `gop_governor` - An indicator for whether the governor is a Republican (transform from `gov_party`)
- `percent_favorable_aca` - Percent of residents favorable to the ACA
- `gop_legislature` - An indicator for whether both houses of the state legislature are controlled by the Republican party (transform from `sen_party` and `house_party`)
- `percent_uninsured` - Percentage of residents who are uninsured.
- `bal2012` - Fiscal health (state year end reserves as a share of total spending)
- `multiplier` - State medicaid multiplier
- `percent_nonwhite` - Percentage of residents who are non-white (transform from `percent_white`)
- `percent_metropolitan` - Percentage of residents who live in metropolitan areas

Some of the above variables do not exist in the data frame. Generate them by transforming the relevant character variables into binary indicators.

Rescale each of the continuous regressors so that it has mean zero and standard deviation .5 (de-mean each variable and divide it by 2 times its sample standard deviation). Rescale each binary regressor so that it is mean zero.

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 governor rejects ACA Medicaid expansion has the following DGP:

$$Y_i = \mathbf{1}(Y_i^* > 0)$$

where $\mathbf{1}()$ 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 $\mathbf{Y}^* = \{Y_1^*, Y_2^*, Y_3^*, \ldots, Y_N^*\}$:

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

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 rejecting medicaid spending conditional on the covariates from Part A. Include the covariates additively, omitting any interactions. 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 $B_0 = \frac{1}{9}\mathbf{I}$ 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 "GOP Governor". Obtain a posterior mean estimate along with a 95% credible interval. Interpret your results substantively. Compare the results to the estimates from a conventional GLM estimated via MLE (use the `glm` package to fit a probit regression). Why do we need to use a Bayesian approach for this particular regression setting?

Part I

Now implement the model above in Stan without the data augmentation step (define y conventionally using the normal CDF `Phi`). Instead of using a normal prior on the coefficients, use a Cauchy prior with center parameter 0 and scale parameter 2.5. Obtain the posterior mean and 95% credible interval for the coefficient on "GOP Governor" and compare the results to your analysis above. How does changing the prior change the results?