GR5065 Midterm Exam

- You can use all of the course materials, readings, etc. to complete this exam. The only thing that is prohibited is communicating with other people (besides Ben) about the exam before everyone has finished (not all students are taking the exam at the same time).
- You have 1 hour and 50 minutes to upload your .qmd file to the Assignments section of Canvas. You can subsequently upload the .pdf file after it finishes rendering, but you might need to clear out of the classroom so that the next class can get in.
- It is not necessary to copy the text of the questions into your .qmd file; just use appropriate section headings and type your answers.
- You are not required to utilize LaTeX for any part of the exam; just type clearly.
- Each of the three main problems is weighted equally and each of the subproblems within each problem is weighted equally. Thus, you should try all of the subproblems so that you can at least receive partial credit.

1 Regression

Many political scientists have argued, largely empirically but somewhat theoretically, that democratic countries with proportional representation (PR) electoral rules have a higher percentage of citizens that actually turn out to vote, compared with similar democratic countries with plurality rule elections like we have in the United States where candidates usually first have to win a primary in order to represent their political party in an election against an individual from the other party.

In a PR system, voters vote for one of several (i.e. more than two) political parties, and if the k-th political party receives 40% of the votes, then the k-th party receives approximately 40% of the seats in the parliament or other political institution. The people who fill those 40% of seats are chosen by the party, rather than the voters. Often no party receives more than 50% of the votes (and hence seats), in which case the party with the largest percentage of votes has to form a coalition with another party in order to make any new laws.

Andy Eggers published a (Frequentist) paper in 2015 that analyzed the hypothesis that PR leads to higher voter turnout by looking primarily at French elections at the municipal level. French municipal elections are somewhat unusual in that France has a law that requires municipalities with a population greater than 3500 to have PR elections for local government, while municipalities with a population less than 3500 must have plurality rule elections. The idea of a "regression discontinuity design" is to estimate the effect on voter turnout of the population going from slightly below 3500 to slightly above 3500 and thereby switching from a plurality rule system to a PR system for municipal elections.

Eggers has a pre-print of the paper on his website at https://andy.egge.rs/papers/Eggers_ProportionalityTurnoutFrance.pdf but you do not need to read it closely, in order to answer the following questions. A subset of the data used for this paper can be downloaded from the Files/ folder on Canvas into your working directory and then loaded into R via

```
France <- readRDS("France.rds")
```

1.1 Generative Model

The essence of the simplest model for the percentage of voter turnout in the n-th municipality, y_n, can be written as

$$\begin{aligned}
\forall n : y_n &\equiv \alpha + \gamma_n \text{PR}_n + \beta_2 \ln\left(\text{Ratio}_n\right) + \epsilon_n \\
\forall n : \epsilon_n &\sim \mathcal{N}\left(0, \sigma\right) \\
\sigma &\sim \mathcal{E}\left(r\right) \\
\forall n : \gamma_n &\equiv \beta_1 + \beta_3 \ln\left(\text{Ratio}_n\right) \\
\alpha &\equiv \mu - \beta_1 \overline{\text{PR}} - \beta_2 \overline{\ln\left(\text{Ratio}\right)} - \beta_3 \overline{\text{PR} \times \ln\left(\text{Ratio}\right)} \\
\mu &\sim \mathcal{N}\left(m_\mu, s_\mu\right) \\
\forall k : \beta_k &\sim \mathcal{N}\left(m_k, s_k\right)
\end{aligned}$$

where $\text{Ratio}_n = \frac{\text{Population}_n}{3500}$ is the ratio of the population in the n-th municipality to the threshold of 3500 and hence the natural logarithm of this ratio is zero if and only if $\text{Population}_n = 3500$. Thus,

- α can be interpreted as the expected percentage of voter turnout for a municipal election when Population_n is infinitesimally *below* 3500, in which case $\text{PR}_n = 0$ and it has a plurality rule system.
- β_1 can be interpreted as the expected increase in voter turnout when Population_n is infinitesimally *above* 3500, in which case $\text{PR}_n = 1$ and it has a PR system.
- β_2 can be interpreted as the expected increase decrease in voter turnout — relative to a municipality whose population is infinitesimally below 3500 — when $\text{Population}_n \approx 1288$, in which case $\text{Ratio}_n \approx e^{-1}$ and $\ln\left(\text{Ratio}_n\right) \approx -1$
- β_3 can be interpreted as the slope in γ_n as $\ln\left(\text{Ratio}_n\right)$ changes. If $\beta_3 \approx 0$, then all municipalities with PR systems have about the same treatment effect (i.e. β_1) regardless of how much their population exceeds the threshold of 3500.

We are going to substitute for α so that μ can be interpreted as expected turnout in a municipality with "average" predictors, which is similar to the expected turnout in a municipality irrespective of its predictors. If we substitute for γ_n in the first line, we equivalently get

$$\begin{aligned}
\forall n : y_n &= \alpha + \left(\beta_1 + \beta_3 \ln\left(\text{Ratio}_n\right)\right) \text{PR}_n + \beta_2 \ln\left(\text{Ratio}_n\right) + \epsilon_n \\
&= \alpha + \beta_1 \text{PR}_n + \beta_2 \ln\left(\text{Ratio}_n\right) + \beta_3 \text{PR}_n \times \ln\left(\text{Ratio}_n\right) + \epsilon_n
\end{aligned}$$

which is the form of the model with an interaction term that estimation software would use.

Use dplyr to create a tibble of $R = 500$ draws from the prior distribution of the parameters (using `rexp` for σ and `rnorm` for every other unknown), each of which can be used to obtain a draw from the prior predictive distribution of the outcome in 2008 for each of the $N = 18698$

municipalities. You will need to chose values for r, m_μ, s_μ, m_1, s_1, m_2, s_2, m_3, and s_3. You can look at the `turnout_2001` variable inside `France`, which we are not otherwise using, to help chose the prior hyperparameters.

1.2 Plotting the Prior Predictive Distribution

Using the ggplot2 package, plot the density of the prior predictive distribution of the outcome in 2008. In what senses is your prior predictive distribution reasonable in this context for subsequently estimating the effect of PR systems at the municipal level?

You may need to call

```
rm(draws)
```

to clear the object you have been using from RAM so that you can create other large objects.

1.3 Posterior Distribution

Use the `stan_glm` function in the rstanarm package along with the choice of priors that you used before (recall that in order to specify a normal prior on the three coefficients, you need to pass a vector of three expectations as the first argument to `normal` and pass a vector of three standard deviations as the second argument to `normal`) to draw from the posterior distribution of the parameters given the data in `France` (where `turnout_2008` is the outcome variable).

What is the posterior probability that crossing the 3500 threshold — and thereby switching to a PR system — has a positive effect on the percentage of voter turnout?

1.4 Posterior Predictive Checking

The subset of data in `France` only includes municipalities whose population is within ∓ 3200 of the 3500 threshold. It is commonplace in regression discontinuity designs to choose some window on either side of the threshold where a linear model is thought to hold reasonably well and avoid including observations with more extreme populations that would call the linearity assumptions into doubt.

Use the `posterior_predict` — without the `newdata` argument but with `draws = 500` option — function to draw 500 times from the posterior predictive distribution for each of these $N = 18698$ municipalities. The resulting matrix will have 500 rows and N columns, one for each municipality. The `colMeans` of this matrix will be a vector of size N that can be plotted on the vertical axis against `France$log_Ratio` on the horizontal axis. If you overlay the actual values of `France$turnout_2008` in a different color, does the resulting plot support the assumption a linear relationship over the entire 3500 ∓ 3200 range?

2 MH 370

In March 2014, a passenger airplane — Malaysia Airlines 370 — took off from Malaysia headed north to Beijing, but somehow got pointed south where there was a lack of extensive radar and satellite coverage. Eventually the plane ran out of fuel and crashed somewhere in the south Pacific ocean without conveying any communication from the pilots.

Someone named Andrew Heneen licensed the following map CC BY 4.0, which summarizes what is publicly known about flight 370.

2.1 Decision Theory for Searches

When something important goes missing, government agencies call on Bayesians to find it. The general idea is discussed on

https://en.wikipedia.org/wiki/Bayesian_search_theory

where the Mathematics section claims that if a grid square on the map has a probability p of having the missing object in it and there is a probability q of finding the missing object given that it is in the grid square, then the posterior probability of it being in the grid square given that the search did not find it simplifies to $p^* = p\frac{1-q}{1-pq}$, and the posterior probability of it being in an another, unsearched, grid square with prior probability r is $r^* = r\frac{1}{1-pq}$.

In 2015, the Australian government sent a crew to look for flight 370 in a 100,000+ square meter rectangle of the Pacific ocean shown in Figure 1 below, which had a posterior predictive probability of about 0.9 for containing the missing plane.

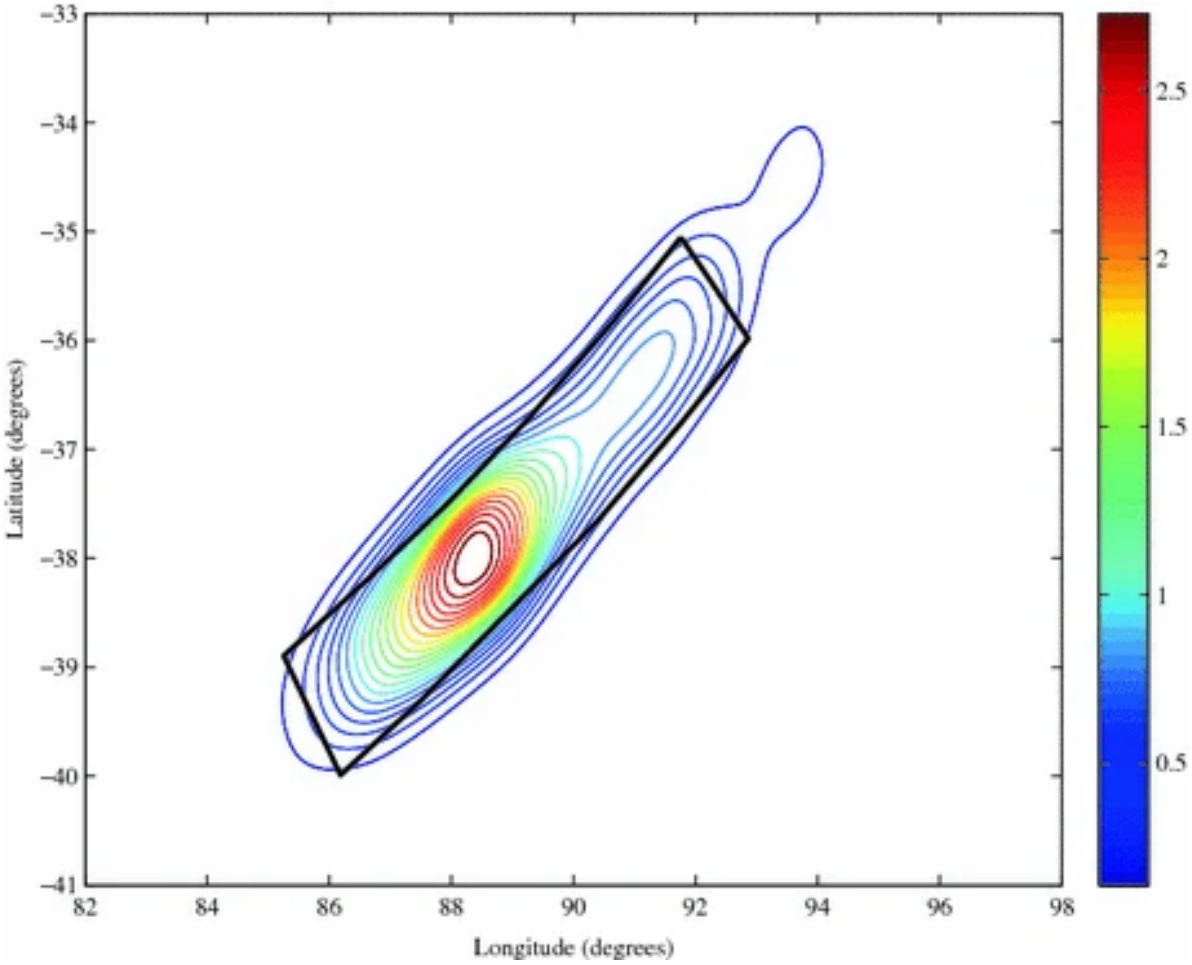

Figure 1: Davey, Sam, et al. Bayesian Methods in the Search for MH370. Springer Nature, 2016. p.97

The posterior distribution was formed by analyzing the initial flight path of MH370, other flights that day in southeast Asia, computer simulations, etc.

The Australian government eventually gave up the search after not finding it in the above search area, and plane has not been found since. Based on the above information only, what do you roughly think is the probability that the plane is within the outer contour line to the northeast of the search area? What is your justification for that?

2.2 Clues

Over the next few years, 18 to 20 small pieces of a large airplane like that of MH 370 washed ashore in Réunion and other islands off the coast of Africa as indicated by the dot in the map above. Since all similarly-sized commercial airplanes are accounted for, these pieces were presumably from MH 370. The ocean currents sufficiently off western Australia run west-northwest, which lead a group of Australian scientists to conclude (without formally using Bayes Rule) that MH 370 is most likely near 35.6 degrees latitude and 92.8 degrees longitude. However, a search in 2018 by a private marine robotics company called Ocean Infinity did not find it there.

Suppose the current posterior predictive distribution of the longitude (X) and latitude (Y) of MH 370 is approximately bivariate normal. The bivariate normal distribution has five parameters, μ_X, μ_Y, $\sigma_X > 0$, $\sigma_Y > 0$, and correlation $\rho \in (-1, 1)$. The bivariate normal PDF can be expressed as a *marginal* univariate normal PDF for X with expectation μ_X and standard deviation σ_X times a *conditional* univariate normal PDF for $Y \mid X$ with conditional expectation $\mu = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ and conditional standard deviation $\sigma = \sigma_Y\sqrt{1-\rho^2}$. Thus, you can draw from a bivariate normal distribution by drawing realizations of X from its marginal normal distribution and then using those realizations of X to draw realizations of Y from its conditional normal distribution.

Form a tibble by drawing $R = 1000$ times from a bivariate normal distribution of longitude and latitude that represents your beliefs as to where MH 370 is. Then, create a scatterplot of these R realizations of longitude and latitude. Briefly describe how you chose the values of μ_X, μ_Y, $\sigma_X > 0$, $\sigma_Y > 0$, and $\rho \in (-1, 1)$.

3 Theory and Methodology

GR5065 can be seen as a theory as to what methodology should be applied to research topics in the social sciences, albeit one that has almost no overlap with the course entitled "Theory and Methodology" that QMSS students are required to take or similar courses that were taken by non-QMSS students in GR5065.

A collection of past syllabae for Theory and Methodology can be found at

https://tinyurl.com/TheoryAndMethodology

Suppose you were going to spend just one week in that two-hour course covering what social science researchers should consider when deciding whether to utilize Frequentist methods, Bayesian methods, machine learning, etc. in a research project. Write an essay on what you would cover and what the main points would be. You do not need to provide citations to required or recommended readings, but you should mention specific topics. In particular, your essay should address at least:

1. What week of what syllabus would you eliminate in order to spend one week on whether to Frequentist methods, Bayesian methods, machine learning, etc.? Why?
2. How you would handle the role of probability in each of these estimation methods, given that the students presumably have not been exposed to probability rigorously in their previous classes and have not recently taken calculus?
3. What situations in social science research would be most amenable and least amenable to Frequentist methods, Bayesian methods, machine learning, etc.?
4. To what extent do you agree or disagree with the following brief statements on this topic from section 1.3.1 of McElreath's Statistical Rethinking and section 8.5 of Johnson, Ott, and Dogucu's Bayes Rules?
5. How would the considerations differ for a QMSS student who is thinking of going on to a Ph.D. program in the social sciences where they would be expected to conduct novel research versus a student who is planning to find a job somewhere as a data analyst after receiving their master's degree?
6. If your proposed outline for a week on whether to Frequentist methods, Bayesian methods, machine learning, etc. is so important, why is it not already included in classes on theory and methodology in the social sciences? What are the obstacles to including it?