GR5065 Midterm Exam Answer Key

1 Regression

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

1.1 Generative Model

If the generative model is

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

then we can use dplyr to draw from the joint distribution of parameters and outcomes from the bottom up.

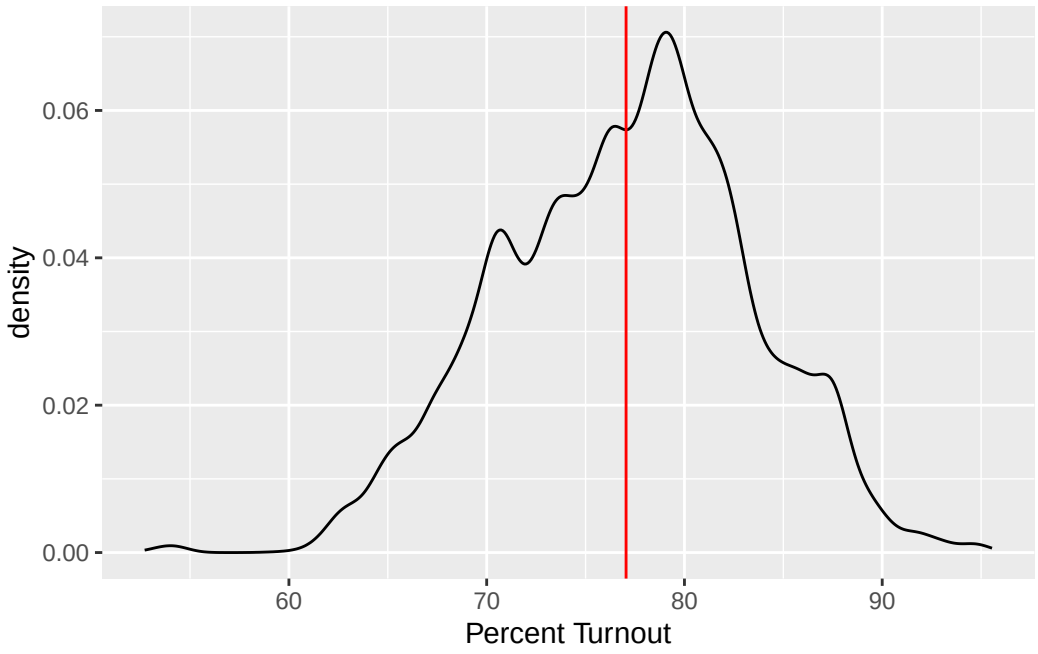
```
library(dplyr)
m_0 <- mean(France$turnout_2001)
s_0 <- 6
m_1 <- 1
s_1 <- 0.75
m_2 <- 0
s_2 <- 1
m_3 <- 0
s_3 <- 0.5
```

```
r <- 1 / sd(France$turnout_2001)

PR_bar <- mean(France$PR)
log_Ratio_bar <- mean(France$log_Ratio)
interaction_bar <- mean(France$PR * France$log_Ratio)
R <- 500
draws <- tibble(beta_1 = rnorm(R, mean = m_1, sd = s_1),
                beta_2 = rnorm(R, mean = m_2, sd = s_2),
                beta_3 = rnorm(R, mean = m_3, sd = s_3),
                mu = rnorm(R, mean = m_0, sd = s_0),
                alpha = mu - beta_1 * PR_bar - beta_2 *
                  log_Ratio_bar - beta_3 * interaction_bar,
                sigma = rexp(R, rate = r)) %>%
  rowwise %>%
  summarize(gamma = beta_1 + beta_3 * France$log_Ratio,
            epsilon = rnorm(nrow(France), mean = 0, sd = sigma),
            y = alpha + gamma * France$PR + beta_2 * France$log_Ratio) %>%
  ungroup
```

1.2 Plotting the Prior Predictive Distribution

```
library(ggplot2)
ggplot(draws) +
  geom_density(aes(x = y), adjust = 3) +
  geom_vline(aes(xintercept = m_0), color = "red") +
  labs(x = "Percent Turnout")
```



This prior predictive distribution looks reasonable because it is centered at the average turnout in 2001 (red line), with no predictions that are negative or greater than 100, which would be inappropriate for a percentage. Moreover, the range in 2001

```
range(France$turnout_2001)
```

```
[1]   0.5922166 100.0000000
```

is similar to that in the plot. Note that the *marginal* distribution of the outcome under the model but irrespective of the parameters looks non-normal in the plot, even though the *conditional* distribution of each y_n is normal by construction because ϵ_n is normal. Thus, the fact that the observed outcome looks not normal is not, by itself, evidence against the model.

```
rm(draws)
```

1.3 Posterior Distribution

```
library(rstanarm)
options(mc.cores = parallel::detectCores())
post <- stan_glm(turnout_2008 ~ PR * log_Ratio, data = France,
                 prior_intercept = normal(m_0, s_0), # on mu
                 prior = normal(c(m_1, m_2, m_3), c(s_1, s_2, s_3)),
                 prior_aux = exponential(r))
```

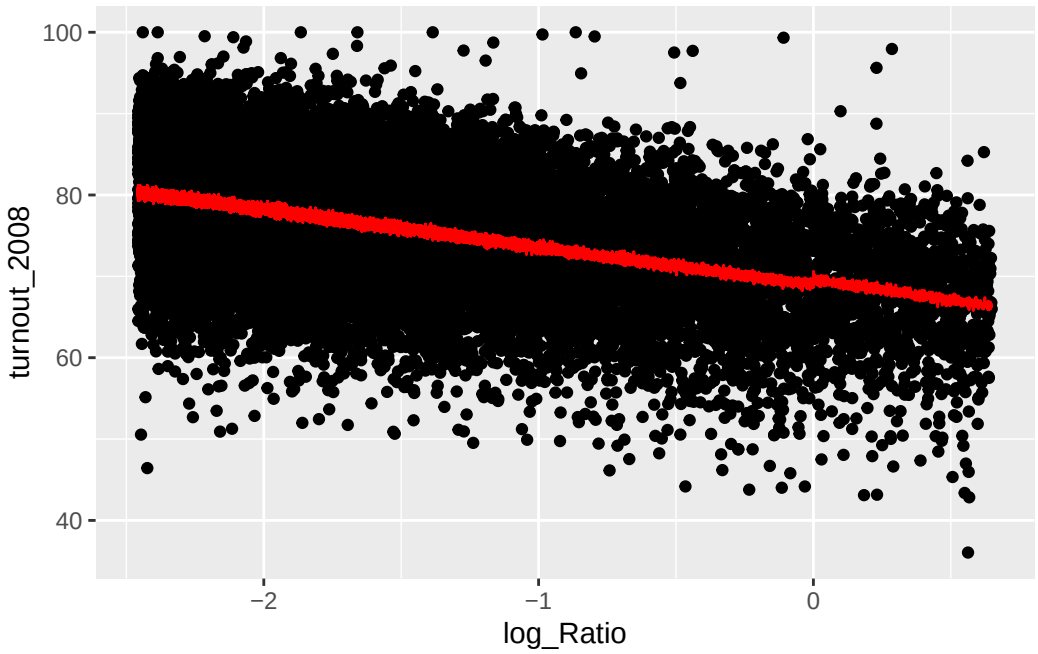
The posterior probability that crossing the 3500 threshold — and thereby switching to a PR system — has a positive effect on the percentage of voter turnout is

```
summarize(as_tibble(post), prob = mean(PR > 0))
```

```
# A tibble: 1 x 1
   prob
  <dbl>
1 0.992
```

1.4 Posterior Predictive Checking

```
predictive_means <- colMeans(posterior_predict(post, draws = R))
ggplot(France, aes(x = log_Ratio)) +
  geom_point(aes(y = turnout_2008)) +
  geom_line(aes(y = predictive_means), color = "red")
```



The red points are linear by construction because they are drawn from a generative model that is linear in `log_Ratio`, but the red points seem to be fairly consistent with the black points that are observed — with considerable noise — in 2008. Thus, there is no obvious need to use a smaller window around the 3500 threshold and perhaps one could use an even larger window.

2 MH 370

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

Much more information about how Australia used a Bayesian process to search for MH 370 can be found in this book.

We can suppose $p = 0.9$ because the rectangle is constructed to contain 90% of the posterior predictive distribution and it looks like $r = 0.08$ is to the northeast of the rectangle. But in order to evaluate $r^* = r\frac{1}{1-pq}$, we need a value of q, which is the probability of finding the plane given that it is within the searched square. Of course, no one *knows* q and it would be more Bayesian to put a (perhaps beta) prior on it. Although the ocean is deep and very dark at the bottom, a commercial plane is a rather large object. If, for example, $q = 0.85$, then $r^* \approx 0.34$, which is much higher than r but still not that high in absolute terms. Thus, for this q, in the absence of additional information, MH 370 is still likely to be in the rectangle but was overlooked by the first search. Only if $q \approx 1$, does $r^* \approx 0.8$.

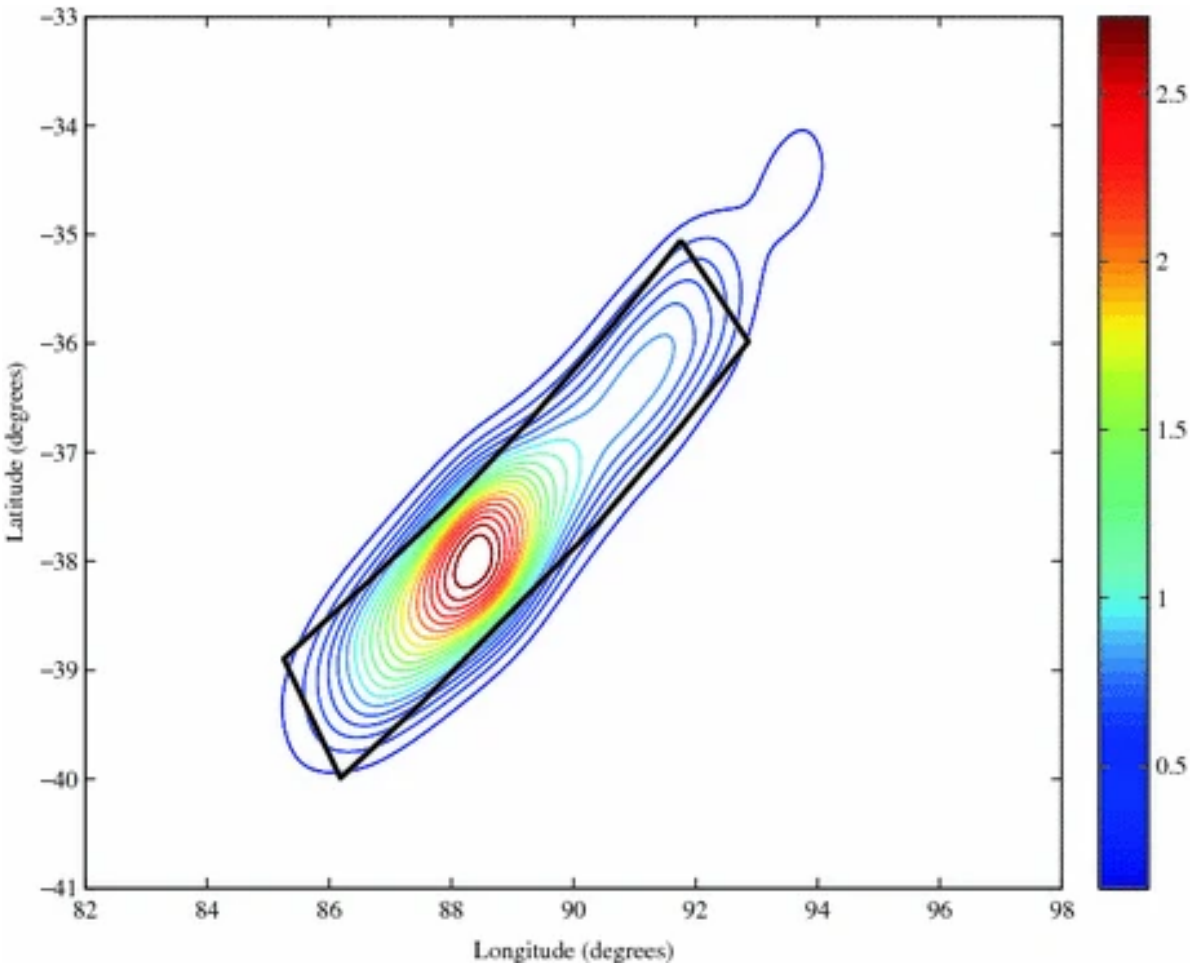


Figure 1: Davey, Sam, et al. Bayesian Methods in the Search for MH370. Springer Nature, 2016. p.97

2.2 Clues

However, additional information came to light when 18 to 20 small pieces of washed ashore in the islands off the coast of Africa. If MH 370 crashed in the southern part of the search rectangle, the ocean currents would have carried the debris from the plane east to Australia. Only if MH 370 crashed farther north would the ocean currents carry the debris toward Africa, and the timing of the discovered plane pieces strongly suggested to one group of scientists that MH 370 is most likely near -35.6 degrees latitude and 92.8 degrees longitude, even though it was not found there by a private marine robotics company called Ocean Infinity.

We can thus set

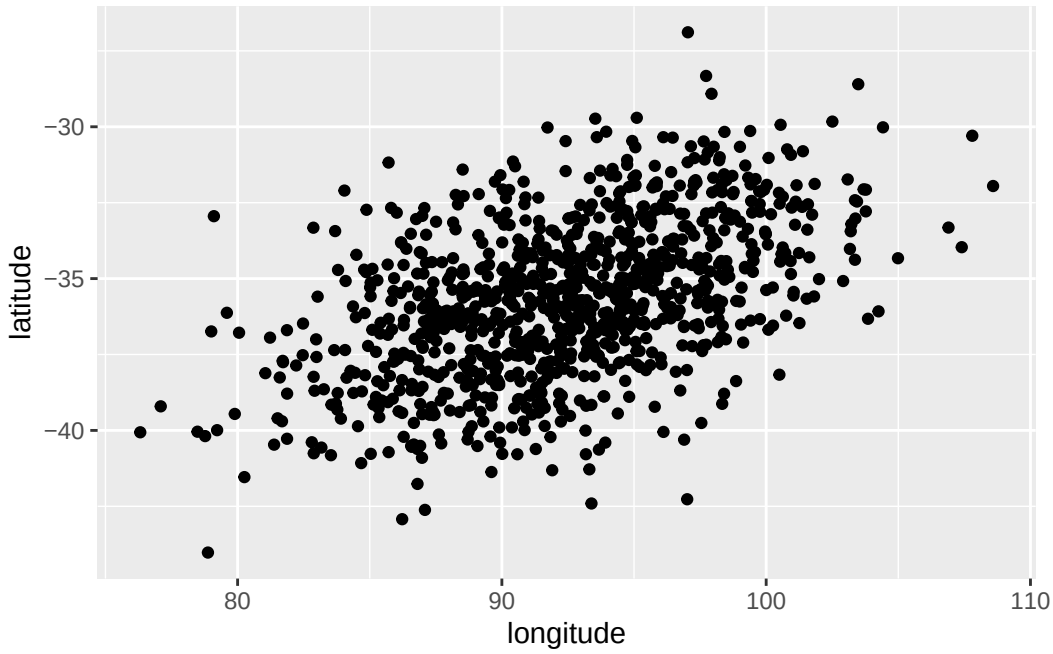
```
mu_X <- 92.8
mu_Y <- -35.6
rho <- 0.5
```

where ρ is rather positive in order to obtain draws along the upward sloping arc where a satellite implied MH 370 was when it made its last ping. Unfortunately, σ_X and σ_Y still need to be considerably large because we still have a lot of uncertainty about where MH 370 is. So, perhaps

```
sigma_X <- 5
sigma_Y <- 2.5
```

We can then draw from a bivariate normal distribution of longitude (X) and latitude (Y) of MH 370 with

```
R <- 1000
draws <- tibble(longitude = rnorm(R, mean = mu_X, sd = sigma_X),
                latitude = rnorm(R, mean = mu_Y +
                                   rho * sigma_Y / sigma_X *
                                   (longitude - mu_X),
                                 sd = sigma_Y * sqrt(1 - rho^2)))
ggplot(draws) +
  geom_point(aes(x = longitude, y = latitude))
```



Ocean Infinity is in negotiations with the Malaysian government to search for MH 370 again, but they have been unable to reach an agreement similar to their agreement in 2018 where Ocean Infinity would cover all the costs and Malaysia would pay $75 million only if MH 370 is found.

2.3 Addendum

This question illustrates that Bayesian inference is just a generalization of the well-known process of elimination. If you are searching for your keys in your apartment, you might keep a mental list of all the places you have looked, in which the probability that your keeps are at one of the places that you have not looked yet rapidly increases. Bayes Rule does not require that the probability be fully eliminated; anything that diminishes the probability that the sought after object is in one region increases the probability that it is in the complimentary region.

In addition, searching for a commercial plane in the ocean is not unlike searching for θ in a parameter space, except that plane has a discrete mass (although it is tiny relative to the ocean) and that the parameter space can be K-dimensional. Although θ cannot be "found" because it has no mass, your posterior beliefs can become rather concentrated in a small region of the parameter space, which is good enough for almost all social scientific purposes.

3 Theory and Methodology

The main challenge with introducing the debate over whether to utilize Frequentist methods, Bayesian methods, machine learning, etc. is figuring out what can be accomplished by a two-hour discussion on a week's worth of readings. That is far too short to conclusively answer such a wide-ranging question, but the purpose of the Theory and Methodology class is more to raise questions than to settle them.

Moreover, if such material were included, then something else on (one of) the current or past syllabae has to be cut. Different people may (de-)prioritize some topics differently, but I would cut the week on text analysis because there is a whole semester class (GR5067) on it, it is still a niche research topic, and it has been largely conducted with machine learning techniques, which gets back to the broader question of when different quantitative methods should be used in the social sciences.

With such limited time, not much can be spent on the role and scope of probability, which is the main dimension on which Frequentist methods, Bayesian methods, and machine learning differ conceptually. However, it is important to introduce the distinction between objective probability (that can be applied to functions of repeated samples, like cards dealt from a deck or a collection of respondents to the General Social Survey) and subjective probability, which applies to researchers' beliefs about any unknowns, including parameters in models or future data. The prevalence of Frequentist methods in the social sciences is partially due to the fact that Fisher felt that subjective probability was inappropriate for scientific research (and, although he did not do so, a largely identical argument could be made against using Bayesian inference for policymaking or running publicly-traded companies). Many of the supposed evils of Bayesian analysis could be mitigated by requiring "preregistration" of priors or even the entire generative model before collecting the data, and as it turned out, preregistration is also necessary to mitigate the evils of Frequentist analysis. Supervised learning focuses almost exclusively on predicting future data, but to the extent that it references probability at all, it generates the *conditional* probability of success in the testing data, given parameter estimates that were obtained from the testing data, rather than propagating the uncertainty about the parameters through to the predictions.

In section 1.3.1 of Statistical Rethinking, McElreath compares and contrasts Bayesian data analysis with Frequentist analysis. Although McElreath strongly prefers Bayesian data analysis in his scientific research (in anthropology, but by extension the social sciences, if not science generally), he refrains from making any harsh statements about the inappropriateness of Frequentist analysis in science. In his course videos that go along with the book, McElreath emphasizes that Bayesian data analysis is well-accepted in statistics departments without dwelling on the fact that it is not well-accepted, or at least not widely used, in any social science. Indeed, most social scientists would be hard-pressed to define Bayesian data analysis or explain how it differs from Frequentist analysis or machine / supervised learning because Bayesian analysis is not covered. But that merely pushes back the question of why Bayesian analysis is not in the social sciences (and a lot of other fields)?

Moreover, the ultimate goal for Bayesians is not merely to be accepted, but for Bayesian data analysis to be the rule, rather than the exception, in scientific research. In contrast, Fisher's goal was for Bayesian data analysis to be eliminated from scientific research and supervised learning's goal is for Frequentist analysis to be eliminated from business analytics. So, there is a lot at stake in this philosophical debate.

McElreath and the authors of Bayes Rules (in section 8.5) note some of the advantages of the Bayesian approach, namely that applied researchers routinely misinterpret Frequentist results as if they were Bayesian. This would suggest that genuinely Bayesian results are easier to interpret but also that scientists are disposed to asking Bayesian questions despite Fisher's objections. Scientists are often interested in *particular* units of observation (or groups of observations), such as where is MH 370 or will the Republican party win the majority of seats in the 2022 congressional elections. And Fisher insists that nothing probabilistic can be said about such particulars because they are not random variables that can take on different values when the researcher takes a random sample from a population that is characterized by parameters. Fisher's view would preclude scientists from saying anything useful about, e.g., climate change because the average temperature in the distant future is not a random variable that we can randomly sample in the present. In contrast, climate change is almost a perfect example of where Bayesian probability can be combined with decision theory to guide policy.

McElreath and the authors of Bayes Rules say less about machine / supervised learning, except that the Bayesian approach can also be used for prediction of future observables. Supervised learning approaches essentially operate under the presumption that uncertainty about the parameters is negligible given enough training data, and that the main challenge for the data analyst is to decide which model is most appropriate based solely on which model best predicts the outcomes in the testing data (at the optimal-in-the-training-data parameter values for that model). If the posterior distribution is heavily concentrated around a point, then it is a waste of time, effort, and perhaps electricity to conduct a genuinely Bayesian analysis that will inevitably find that the priors were irrelevant and that the model's parameters have essentially become known. However, it is rather rare for social scientists to have what supervised learning proponents consider to be "Big Data", and whether Data is sufficiently Big to render parameter uncertainty negligible depends on what model is being estimated. The best-case scenario for machine learning in the social sciences might be the analysis of extremely large collections of text that is collected in an automated fashion.

Despite the prevalence of Frequentist estimation in the social sciences, it is rather difficult to come up with an example that is amendable to Frequentist analysis in the sense that it adheres to the Frequentist assumptions that the data are one of many possible random samples from a static population. Outside of the social sciences, it is a bit easier to come up with examples, such as when a manufacturer more-or-less knows its past defect rate and wants to test the null hypothesis that the current defect rate is equal to the past defect rate using a small random sample of products that are pulled off its assembly line. Political polling in the run up to competitive elections has some of the characteristics that are necessary for a Frequentist analysis, in the sense that there are many polls where the respondents are

chosen randomly ostensibly from the same population. You can derive the distribution of the proportion of respondents in a poll who support a candidate or party conditional on the population proportion being $\mu = 0.5$ or any other particular value (i.e. not a probability distribution).

However, while pollsters can randomize who is contacted from lists of registered voters, the people who agree to take the poll are a non-random subset of those who are contacted, which is one of several reasons pollsters have to make non-Frequentist adjustments to their raw data. In addition, no one knows exactly who will constitute the population of voters until (after) Election Day. As a result, it is not clear at all whether the people who respond to the polls are actually a random sample from the population of future voters.

You cannot understand or conduct a Bayesian analysis without understanding probability and many QMSS students would not survive a semester-long course that was based on probability, even if they were motivated to learn it as a result of its importance being introduced in the Theory and Methodology class. Supervised learning emphasizes its differences with Frequentist analysis, but it is differs from Bayesian analysis even more starkly because proponents of supervised learning feel that probability should not be a prerequisite to making predictions. As a result, Bayesian analysis is not highly-valued in many companies, although they love to hire people who are *capable* of Bayesian analysis and can presumably handle whatever non-Bayesian data analyses the company does value. Such companies do not want to become reliant on Bayesian steps in a pipeline because if the Bayesian who wrote the code leaves for another job, it would be difficult to find and evaluate another Bayesian to fill that role.

The case for learning how to conduct a Bayesian analysis is stronger for Ph.D. students or people who are thinking of applying to Ph.D. programs. There will be plenty of times over their research careers where it is appropriate or necessary to do a Bayesian analysis. However, Bayesian analyses are more time consuming, both to conduct and to explain to journal reviewers and editors who are much more familiar with, and perhaps sympathetic to, the Frequentist analyses that have filled previous issues of their journals. Thus, the path of least resistance is to use Frequentist analysis even when it is inappropriate, but that is often acceptable to assistant professors who need to publish several articles and perhaps a book or two during their tenure clocks.

In conclusion, it is certainly understandable why the Theory and Methodology class that is required of QMSS students does not even raise theoretical questions about which estimation methodology to use in social science research. It might be better to have a separate, semester-long class where a month is spent on each of the three major approaches. However, that again raises the question of what can be accomplished in a month studying Bayesian analysis? In GR5065, it took us an entire month just to (rapidly) cover the probability theory that we leverage for the remaining weeks to do Bayesian analyses, and only a small minority of QMSS students could manage that. Nevertheless, it is disturbing that a form of analysis could be too quantitative, too methodological, too scientific, and / or too applicable to social research for it to be taught to graduate students at the top program for Quantitative Methods in the Social Sciences.