Data 102 Final Report

Alan Jian, Shreya Chowdhury, Yannie Li, Matilda Ju

January 25, 2024

0.1 Data Cleaning

Since we were only interested in trends between the years 2000-2018, we started by limiting our dataset to those who responded in these years. When we explored our data, we started by analyzing levels of non-response to see if there were any general trends in them, and if some groups failed to respond more than others. Once we confirmed that individuals were missing almost completely at random, we dropped any non-respondents for the purposes of our visualization.

In preparing our data for Bayesian Hierarchical Modeling, we started by numericalizing self-reported political leaning (see the **Methods** section of the **Question 2** for more details) and binarizing all of our response variables (see Table 1), since it was easier to work with Bernoulli-distributed observations since our problem formulation involves some classification.

#	Original Question	Binarized Question	0	1
1	Are we (the government) spending too much, too little, or about the right amount on the environment?	Do we spend too little protecting the environment?	Too Much, Just Right	Too Little
2	Are we (the government) spending too much, too little, or about the right amount on national defense?	Do we spend too little on national defense?	Too Much, Just Right	Too Little
3	Are we (the government) spending too much, too little, or about the right amount on welfare?	Do we spend too little on welfare?	Too Much, Just Right	Too Little
4	Are we (the government) spending too much, too little, or about the right amount on social security?	Do we spend too little on social security?	Too Much, Just Right	Too Little
5	Would you favor a law requiring a a permit before a person could buy a gun?	N/A	Oppose	Favor
6	Should a pregnant woman be able be able to obtain a legal abortion for any reason?	N/A	No	Yes
7	Do you think sexual relations between two adults of the same sex is wrong at all?	N/A	Not Wrong at All,	Always Wrong, Almost Always Wrong, Sometimes Wrong
8	On average, (Blacks/African-Americans) have worse jobs, income, and housing than white people. Do you think these differences are mainly due to discrimination?	N/A	No	Yes

Table 1: Political Questions of Interest and their Binarized Counterparts

0.2 Data Overview

The General Social Survey is a regular, ongoing interview survey administered randomly to NORC (National Opinion Research Center) national samples independently drawn from English-speaking persons 18 years of age or over living in non-institutional arrangements within the United States, producing a high-quality, representative sample of the adult population of the U.S. Participants respond to a randomly subsampled standard questionnaire via personal interview as conducted by The National Data Program for the Social Sciences. It collects information on a wide range of demographic characteristics of respondents including behavioral items such as group membership and voting, personal psychological evaluations, and attitudinal questions on public issues such as abortion. Given the GSS's high response rate (¿70%) and rigorous design, we do not expect to encounter any selection bias or convenience sampling. The data and documentation can be easily downloaded from the GSS website (https://gss.norc.org/), with each row in our data representing a de-identified individual.

0.3 Research Question 1 (Causal Inference): Is there a causal relationship between race and opinion of police brutality?

0.3.1 Motivations

In the wake of the #BlackLivesMatter movement, there has been a lot of debate over police practices and the use of excessive force. For any given individual, there might be a higher or lower chance that they oppose excessive use of force by officers, something politicians and government officials must navigate. To better understand this relationship, we decided to use causal inference, since we are looking to understand whether there exists a causal relationship between a given respondent's race and their opinions on police brutality, and alternative metrics such as a risk difference are only sufficient to establish correlation.

0.3.2 Methods

To establish a causal whether self-identifying as African American causes a shift in opinion regarding police brutality, we began by compiling a number of questions regarding police brutality into a measure-able score for each individual. In order to create a valid score, we limited our study to individuals who answered all 5 questions, and decided upon a simple score out of 5 based on the number of scenarios in which he/she supported police brutality. The scenarios that respondents were queried about were as follows (in each, respondents answered with a Yes/No response):

1. Would you approve of a policeman striking a citizen who...
   (a) had said vulgar and obscene things to the policeman?
   (b) was being questioned as a suspect in a murder case?
   (c) was attacking the policeman with his fists?
   (d) was attempting to escape from custody?
2. Are there any situations you can imagine in which you would approve of a policeman striking an adult male citizen?

Since our data comes from an observational study, we needed to find ways to guarantee $Z \perp\!\!\!\perp (Y_i(0), Y_i(1))$, where Z represents self-identification as Black/African American, and $Y_i(0), Y_i(1)$ represent possible number of scenarios the ith individual supports police brutality in, depending on whether or not they self-identify as Black/African American. Given that we had a fairly large group of responses, we decided to apply the conditional independence assumption ($Z \perp\!\!\!\perp (Y_i(0), Y_i(1))|X$) via matching. To guarantee that the unconfoundedness assumption holds, we decided to condition on variables that are highly correlated with both race and opinion of police brutality (Note: While the technical definition of a confounder requires a causal relationship with both treatment and outcome, there are no variables in the study that would cause race. It is for this reason that we decided to use variables that are strongly correlated with race). In our study, we chose as many plausible confounder variables as possible and came up with income (INCOME), age (AGE), sex (SEX), education (EDUC), marital status (MARITAL), and work status (WRKSTAT). We believe the unconfoundedness assumption holds because by controlling for income, age, marital status, work status, education, and sex, a respondent's race is conditionally independent of the number of scenarios in which he/she supports police brutality. There are no colliders in the dataset.

In our efforts to match, we initially considered possible options such as 1:1 matching, and k:1 matching for various values of k. However, each of these failed to yield meaningful results. Since the number of Blacks in our study were outnumbered by non-Black respondents at a ratio of roughly 4:1, 1:1 matching eliminated a lot of meaningful information, particularly in instances where an African American respondent could be matched to several non-Black respondents. In the case of k:1 matching, however, we were seeing poor performance for the exact opposite reason: by enforcing a strict matching ratio, African Americans with rarer combinations of confounders were being forced to match with their nearest neighbors, which were not always very similar. As a result, we decided not to enforce a pairing ratio on our study, and decided to use an entirely new approach to estimate the ATE: the conditional simple difference in mean outcomes (SDO). Our inspiration for using this measure comes from the idea that, in randomized controlled experiments when $Z \perp\!\!\!\perp (Y_i(0), Y_i(1))$ we can trivially compute the ATE using the SDO since

$$\mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1)|Z = 1] - \mathbb{E}[Y_i(0)|Z = 0]$$

Assuming that $Z \perp\!\!\!\perp (Y_i(0), Y_i(1))|X$ holds, we can extend this idea to the calculation of conditional ATE via conditional SDO:

$$\mathbb{E}[Y_i(1) - Y_i(0)|X = x] = \mathbb{E}[Y_i(1)|Z = 1, X = x] - \mathbb{E}[Y_i(0)|Z = 0, X = x]$$

This equivalence can be trivially proven; since $Y_i(1)|Z = 1, X \sim Y_i(1)|X$ and $Y_i(0)|Z = 0, X \sim Y_i(0)|X$ via the unconfoundedness assumption, the above can be written as the true statement

$$\mathbb{E}[Y_i(1) - Y_i(0)|X = x] = \mathbb{E}[Y_i(1)|X = x] - \mathbb{E}[Y_i(0)|X = x]$$

After computing the conditional ATE from the conditional SDO, we then used iterated expectation to find the ATE (true prevalence of each set of confounders was estimated by each set's prevalence in the data).

0.3.3 Results

Assuming unconfoundedness given income, age, sex, education level, marital status, and work status, we found there to be a negative causal relationship between race and opinion of police brutality. Our computed ATE of -0.52 suggested that Black Americans support police brutality in slightly fewer scenarios compared to those of non-Black Americans. The largest sources of uncertainty in our estimate of the ATE were the issues that we had in formulating the study, which are discussed in detail in the **Discussion** section.

0.3.4 Discussion

Our study has one primary flaw: the confounding variable of income. Although the literature says definitively that income is highly correlated with race and has a causal relationship with opinions on police brutality, the confounding variable that we controlled for here is poorly designed for this use case. The income variable is badly binned, with a majority of individuals self-identifying a salary above $25000 for obvious reasons. The question asked in the survey was clearly more interested in those living in poverty, since there were several bins that divided up income levels between $0-25000 in annual salary. As a result, there is a chance that the unconfoundedness assumption might not hold in our study, since the income bins only limited matches in respondents who earned an income far below the poverty line. If we had a better income variable, controlling for income's confounding effects would be much more effective.

Additionally, there are some other confounders that we did not consider. Since interactions with police can be extremely correlated with race and opinions of police brutality (if you're constantly getting arrested or having run-ins with police there might be a differential amount of empathy you feel in this case), it would have been nice to have access to this information for purposes of matching.

The questions we were able to explore about police brutality were very specific. For example, they generally follow the format "Would you approve of a policeman striking a citizen who..." We would have been more interested in exploring DEFUND, which asks if people favor or oppose reducing funding for police departments (which didn't have enough responses for us to use), or more general questions about police brutality.

Since our finding is supported in sociology, we are highly confident that there is a causal relationship between race and opinions on police brutality. If there would be an uncertainty that we have about our result, it would be the fact that the ATE is calculated only on the sample data that we have; and given its closeness to 0, it wouldn't be a surprise to see that the 95% confidence interval on the ATE would include 0, indicating a chance of no causal effect.

0.3.5 Conclusions

In our study, we found evidence to suggest that self-identifying as African American makes you support police brutality in fewer scenarios. Since our finding is supported in sociology, we are highly confident that there is a causal relationship between race and opinions on police brutality. Given that our ATE is close to 0, future studies should seek to determine whether the 95% confidence interval includes 0, which would lend this finding more credibility. This information should be used to inform political parties and politicians' tune and modify their messaging with respect to different audiences.

0.4 Research Question 2 (Bayesian Hierarchical Model): Is there a discrepancy between a person's self-reported political leaning and their true political leaning?

0.4.1 Motivations

As more and more issues become increasingly politicized, an individual's opinion may shift from the political ideologies they once held. We wanted to see if there exists a perception gap between perceived and actual political leaning, and its distribution. To answer this question, we utilized a Bayesian Hierarchical Model, which is uniquely suited to capturing information about latent variables.

0.4.2 Methods

j	Question	$y_{ij} = 0$	$y_{ij} = 1$
1	Do we spend too little protecting the environment?	No	Yes
2	Do we spend too little on national defense?	No	Yes
3	Do we spend too little on welfare?	No	Yes
4	Do we spend too little on social security?	No	Yes
5	Would you favor a law requiring a permit before a person could buy a gun?	Oppose	Favor
6	Should a pregnant woman be able to obtain a legal abortion for any reason?	No	Yes
7	Do you think sexual relations between two adults of the same sex is wrong at all?	No	Yes
8	On the average (Blacks/African-Americans) have worse jobs, income, and housing than white people. Do you think these differences are mainly due to discrimination?	No	Yes

Table 2: Binarized Political Questions

To find a solution to this research question, we started with the assumption that an individual's self-reported political leaning was a noisy estimate of his/her true political leaning. Given $\{1, ..., m\}$ individuals with self-reported leanings $\alpha_i \in \{-3, ..., 3\}$ corresponding to the ith individual, we modeled individuals' true leanings with the expression $\vec{\alpha} + \epsilon 1$, where $\vec{\alpha}, 1 \in \mathbb{R}^m$, and ϵ is some random variable that represents a perception gap between actual and perceived leaning. Since individuals' true leanings seem to be reasonable predictors of how they might respond to $\{1, ..., n\}$ political questions (see Table 2 for mapping), we used individuals' observed binary responses to give us information about this hidden variable. As a result, our problem required a blend of inference and regression in which each ith individual's observed response to each question j could be modeled by $y_{ij} \sim Bernoulli(\sigma(\beta_j(\alpha_i + \epsilon)))$, where each jth element of n-vector $\vec{\beta}$ represents the weight $\alpha_i + \epsilon$ has on predicting the value y_{ij} (as shown in Figure 0.4.2). Given a matrix of observed answers $Y \in \mathbb{R}^{m*n}$ from m individuals across n questions, we couched the problem as the following *maximum a posteriori* estimate:

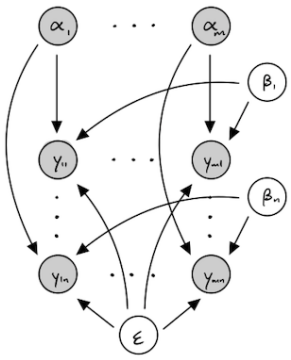

Figure 1: Graphical Model

$$\max_{\epsilon, \vec{\beta}} \mathbb{P}(Y|\vec{\alpha}, \vec{\beta}, \epsilon)\mathbb{P}(\vec{\beta}, \epsilon)$$

where

$$\mathbb{P}(\epsilon) \sim Normal(0, 0.0001)$$

and $\vec{\beta}$ starts with a flat prior. The prior on ϵ was chosen to encourage model convergence, and to reflect the fact that we expect an individual's self-reported leaning to be an unbiased estimator of true leaning. After updating priors on ϵ and $\vec{\beta}$, we analyzed the distribution of ϵ to better understand the perception gap between an individual's response and their actual political leaning, and visualize the elements of $\vec{\beta}$ as a sanity check to make sure its values match our expectations.

0.4.3 Results

After running our model, our results suggest that there is a real, consistent perception gap between an individual's real and perceived political leaning. When viewing the distribution of samples from the posterior distribution of ϵ (see Figure 0.4.3), we can see that over 97% of the distribution is greater than or equal to zero, and that an individual's perceived political leaning is off by an average of -0.07 from their true political leaning. Assuming that our data are representative of the U.S population, and that each individual's value of epsilon is independent and identically distributed, this data indicates that an overwhelming majority of Americans misrepresent (either intentionally or unintentionally) their political leaning as being more liberal than they truly are. This has major implications for government officials, and how they choose to best serve their constituents.

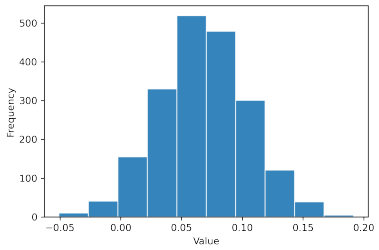

Figure 2: Posterior Distribution of ϵ

Furthermore, upon investigating the elements of $\vec{\beta}$ as shown in Figure 0.4.3, we see that the ethical and political questions that we used to build our model's intuition about each individual's true political leaning were informative in the right ways. Being liberal

was positively associated with beliefs that the national government was not spending enough on protecting the environment or expanding welfare, while being conservative was associated with the belief that felt the government was not spending enough on national defense. Additionally, we can see that being conservative also made you more likely to oppose homosexuality and oppose abortion rights, beliefs that are consistent with politics today. The only variable that was (correctly) identified to be evenly supported by both liberals and conservatives was national spending on social security. In this case, both liberals and conservatives felt that national spending on social security is not enough, with the value of $\beta_{socialsecurity}$ hovering very close to zero.

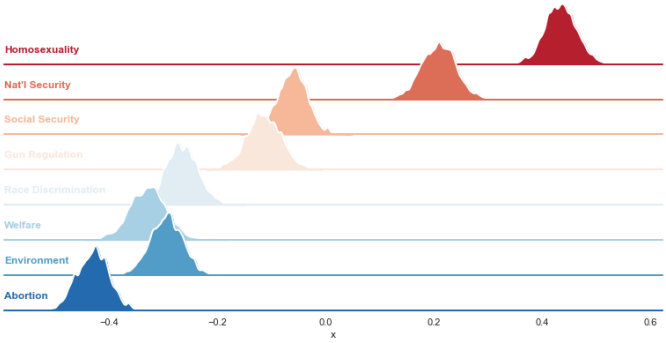

Figure 3: Posterior Distribution of the elements of β

0.4.4 Discussion

For all this models benefits, it also came with some major limitations. Firstly, due to the setup of our model, it easily diverged when initializing ϵ with a Gaussian who's $\sigma > 0.05$. When this occurs, the posterior of ϵ becomes split, literally diverging towards two different ends of the number line. Based on our tuning, it seems like this problem arises due to lack of informative features. Even though this collection of survey questions helped the model get a better idea of the relationship between true political leaning and opinion on these issues, the model clearly needed more questions to improve accuracy, or more respondents to improve confidence (both of these are equivalent to expanding Y in our graphical model), something that we tried to solve via imputation. Rather than drop all data entries that were not asked every single question of interest, we tried using PyMC3's automatic value imputation via posterior predictive distribution to avoid throwing away a treasure trove of information. Given that our data was missing completely at random (MCAR) via study design, we knew that this imputation would not bias our results. Though this idea sounded great in theory, it proved intractable in practice due to the sheer scale of imputation required. Internally, when PyMC3 encounters non-response, it treated it like another latent variable with a prior distribution to be updated along with everything else. For a dataset with a total size of over 28000, in which a majority of individuals were not asked every question of interest, this meant adding well over 30000 random variables to the calculation, a computation cost that was simply too much.

0.4.5 Conclusions

In our study, we found concrete evidence that U.S citizens are very likely to perceive themselves as more liberal than they truly are, as inferred based on their opinions on political issues. Furthermore, we also have evidence to suggest that an individual's perceived political leaning is not an unbiased estimator of their true political leaning. This finding has major implications both for individuals who hold political positions, as well citizens. For politicians, this evidence may be useful in shifting campaign strategy. Given that individuals see themselves as more liberal-leaning than they actually are, politicians may increasingly turn to performative activism as a way to pull in more voters. For citizens, it calls into question whether our sources of entertainment and information are distracting or pulling us away from the issues and matter most to us. Still, more needs to be done to understand why this perception gap exists, and what effects it has on our already fragile democracy.