Climate Change Analysis – The Bayesian Approach

DS6040 Final Project – Summer 2023 – UVA School of Data Science

Abigail Snyder – syc6vs, Elena Tsvetkova – rrm3nh, Suraj Kunthu – sk9km

Problem Statement

The objective of this analysis is to uncover how demographic factors influence opinions regarding climate change. This investigation aims to provide insights that could potentially contribute to the development of effective educational and policy strategies concerning climate change awareness and environmental sustainability. The central question driving this project is: To what extent does demographic information, including income, age, gender, religion, marital status, and education, contribute to shaping viewpoints on climate change?

To address this question, we conducted a comprehensive examination of the relationship between demographic variables and opinions on climate change. By systematically analyzing a diverse dataset encompassing the previously mentioned predictors, we aimed to find meaningful patterns in the response *happening.* Through statistical model-based approaches, we analyzed the model performance of these demographic influences on climate change opinions. These findings could potentially provide insights into the ways in which different demographic segments perceive climate change issues. These insights could serve as a valuable resource for policymakers and educators seeking to cater their policies to resonate better with specific demographic groups. By understanding the complex relationship between demographics and opinions on climate change, we attempt to contribute to a more informed approach to environmental awareness and, potentially, sustainable practices.

Data Description

The chosen data set is survey data from the Yale Program on Climate Change Communication & George Mason University Center for Climate Change Communication. The survey data catalogs American adults, ages 18 and over, and their beliefs, attitudes, and policy opinions regarding global warming from 2008 through 2022. The data includes demographic data from the U.S. Census regarding gender, race and ethnicity, and level of education. All the data is in categorical form and stored as a *'.SAV'* file. Along with the data set there is a Survey Methods PDF which functions as a codebook for the data. The total weighted sample size of the entire dataset is (n=30,136). There are 30,136 observations and 54 features.

We started by importing the data and performing some exploratory data analysis. Fortunately, not much data cleaning was necessary with this data set. We decided to start by creating a model with *happening* as the response variable, and trimmed down the predictor variables to those we are most interested in *gender*, *age_category, educ_category, income_category, race, ideology, party, region4, religion,* and *marit_status*. We chose these predictors as they best described the individual survey takers for determining demographics while the other predictors were the responses to the survey. The variables were represented in the form of questions in the survey with multiple choice answers for everyone to answer. *Happening* describes whether the individual believes global warming is occurring with *Yes*, *No*, *Don't Know*, and *Refused* as the possible responses. *Gender* describes the individual's gender identity with *Male* and *Female* as the possible responses. *Age_category* is computed based on the individual's age input with *18-34 years, 35-54 years,* and *55+ years* as the possible responses. *Educ_category* is computed based on the individual's highest completed school level input with *Less than high school, High school, Some college, bachelor's degree or higher* as the possible inputs. I*ncome_category* is computed based on the individual's total household income input with *Less than $50,000, $50,000 to $99,999, $100,000 or more* as the possible responses. *race,* describes the individual's race with *White, non-Hispanic, Black, non-Hispanic, Other, non-Hispanic, Hispanic* as the possible inputs. *Ideology* describes the individuals' views with *Refused, Very liberal, Somewhat liberal, Moderate, middle of the road, Somewhat* *conservative,* and *Very conservative* as possible inputs. P*arty* describes the individuals' political views *Refused, Republican, Democrat, Independent, Other; please specify:,* and *No party/not interested in politics* as possible inputs. *Region4* is computed based on the individual's response to what region of the United States they live in with N*ortheast, Midwest, South,* and *West* as possible responses. *Religion* describes an individual's religious preference with *Refused, Baptist – any denomination, Protestant (e.g., Methodist, Lutheran, Presbyterian, Episcopal), Catholic, Mormon, Jewish, Muslim, Hindu, Buddhist, Pentecostall, Eastern Orthodox, Other Christian, Other – non-Christian (Please specify), Agnostic [Apr 2014 on], Atheist [Apr 2014 on],* and *None of the above [Apr 2014 on; "None" Nov 2008 – Dec 2013]* as possible inputs. *Marit_status* describes an individual's marital status with *Married, Widowed, Divorced, Separated, Never married,* and *Living with partner* as possible inputs. Although there were a few other demographic predictor variables in the dataset, we felt the ones we selected cast the widest net to capture the data for modeling.

Initially, we wanted to see what the overall demographics of the survey takers were. We found that most of the survey takers were female, over the age of 55, from the southern US, married, white, non-hispanic, have a bachelor's degree or higher, politically democrat, ideologically moderate, and make less than $50K/year, see **Figure 1**. This gave us an idea of the initial spread of data. Now we can focus on our response variable *happening.* An overwhelming majority of the survey takers believe that climate change **is** happening, **Figure 2.** However, we were interested in the spread of responses in each demographic predictor variable. We observed some interesting insights in a few of the more detailed bar charts of the *happening* variable. In **Figure 3**, we can see the distribution of responses to *happening* by *age_category*. All 3 age groups tend to follow the same pattern as the overall bar chart in **Figure 2**. It does look like younger people tend to agree that climate change is happening. In **Figure 4**, we can see the distribution of responses to *happening* by *ideology*. Here we can start to see the divergences in the data where certain subsets of people tend to respond with climate change not happening. Particularly 'Very conservative' individuals display this behavior the most of out the 6 groups. In **Figure 5**, we can see the distribution of responses to *happening* by *party*. If we consider 'Democrat' and 'Republican' at opposite ends of a spectrum, it appears that the vast majority of democrats do believe that climate change is happening while more republicans are not convinced climate change is happening. We could continue analyzing this dataset for every predictor, but we understand the overall pattern of data. Most people believe that climate change is happening.

Model Description

With this in mind, we started with the model-building process, working our way through several different models and comparing them to see what would be the best fit for the data. First, we created a 60/40 training and testing data split for validation. Then we utilized the `bambi` v0.12.0 package in Python which is the Bayesian Model Building Interface in Python that is built on top of `PyMC3` and `ArviZ`.

We developed four Bernoulli logistic regression models to predict the influence of certain demographic variables on the response *happening* using the training data. Two models with all possible predictors but differing priors and two models with a select few predictors and differing priors. The first model utilized the default Normal priors provided by the `bambi` package. These priors are weakly informed by loosely scaling the priors to the observed data, as per the documentation. The means for each regression coefficient were all 0 and the standard deviations for each coefficient can be seen in **Figure 6** along with the model. For the second model, we used our own set of weakly informed Normal priors with larger standard deviations. The means for each regression coefficient were all 0 and the standard deviations for each coefficient were all 100 and can be seen in **Figure 7** along with the model. For models 3 and 4, we subset the data to the only using two predictors: *educ_category* and *ideology* to test how the simplified model performs compared to the more complex model. These were chosen because the belief of whether climate change is happening is a well noted political topic that is often divided along party lines. However, more potential exposure to science education could offset some of the effects of political ideology. The third model utilized the default normal priors provided by the `bambi` package. The means for each regression coefficient were all 0 and the standard deviations for each coefficient can be seen in **Figure 8** along with the model. The fourth and final model utilized our own set weakly informed Normal priors with larger standard deviations. The means for each regression coefficient were all 0 and the standard deviations for each coefficient were all 100 and can be seen in **Figure 9** along with the model.

Results

From the models developed, we utilized Leave-One-Out Cross-Validation and compared each model's performance with ELPD (expected log predictive density). We can see from the results in Figure **9** and **Figure 10** that models 1 and 2 with all predictors have a higher ELPD and perform better than the simpler model. Model 1 included `bambi`'s default priors and model 2 included weakly informed priors (mean=0, standard deviation=100). Both model 1 & 2 have extremely similar ELPD values, suggesting that the priors don't have much impact on our analysis, when using Normal priors. We will move forward with model 1. We then summarized the model performance of each predictor variable using the trace plots and forest plots in **Figure 11** and **Figure 12.** For Gender, assuming all other variables are the same, there is a 23% reduction in belief that global warming is happening for men compared to women. For Education category, completing college increases the odds that a person believes global warming is happening. The HDI's (highest density intervals) do not overlap with zero, and the coefficients for some college/high school/less than high school is negative, meaning you are more likely to believe global warming is happening if that person has completed college. For Ideology, there is a very clear pattern in ideology and climate change. Generally, when a person identifies as being somewhat/very liberal, they are more likely to believe global warming is happening, relative to moderates. In contrast, conservatives were much less likely to believe global warming is happening. This follows the behavior we observed during the exploratory data analysis phase. For religion, even though it varies for some religions, belief in some religion is generally associated with a decreased chance in believing global warming is happening, relative to agnostics. Although the posterior mean for those claiming to be atheist is positive, the HDI did overlap with zero, suggesting the effect may be small relative to agnostics. This makes sense as both are not acknowledging a higher power. Finally for age, it appears to have little effect on whether a person believes global warming is happening (HDI's overlap with zero).

Discussion

In **Figure 14** we have a separation plot of the model's performance using test data. The vertical lines are represented as model predictions. Light blue is class 0 (global warming is not happening) and dark blue is class 1 (yes, it is happening). However, the position of the vertical lines represents the truth - further to the left, a person does not believe global warming and further to the right, a person does believe it is happening. Based on the plot, we do see light blue vertical lines towards the left and more dark blue lines to the right. This model does a decent job at predicting whether a person believes global warming is happening or not. However, does this model address the central question driving this project: To what extent does demographic information, including income, age, gender, religion, marital status, and education, contribute to shaping viewpoints on climate change? We can clearly see from the results that some predictors, such as education, ideology, and religion, have an impact to varying degrees on whether a person believes in climate change or not. There are also predictors that have little to no impact on determining if a person believes in climate change or not. Future studies could further explore the dataset by not only examining various modeling options but also investigating the extent to which interactions between predictors impact the overall model. This information could be utilized to inform policymakers and educators to tailor their approach for specific demographic groups by targeting the issues they are most concerned about. Knowing your audience can help persuade those who are not swayed by scientific facts and figures.

Appendix

region4	count
Midwest	7065
Northeast	5493
South	10643
West	6935

gender	count
Female	15235
Male	14901

income_category	count	percent
Less than $50,000	10953	36.345235
$50,000 to $99,999	10107	33.537961
$100,000 or more	9076	30.116804

marit_status	count
Divorced	2889
Living with partner	1628
Married	15786
Never married	4975
Separated	454
Widowed	1343

race	count
Black, Non-Hispanic	2675
Hispanic	3104
Other, Non-Hispanic	1987
White, Non-Hispanic	22370

age_category	count
18-34 years	6259
35-54 years	9912
55+ years	13965

educ_category	count
Bachelor's degree or higher	10748
High school	8305
Less than high school	2223
Some college	8860

ideology	count
Moderate, middle of the road	11994
Refused	591
Somewhat conservative	6635
Somewhat liberal	5375
Very conservative	3317
Very liberal	2224

party	count	percent
Democrat	9903	32.861030
Republican	8336	27.661269
Independent	7737	25.673613
No party/not interested in politics	2992	9.928325
Other; Please specify:	859	2.850411
Refused	309	1.025352

Figure 1: Initial survey taker frequency tables

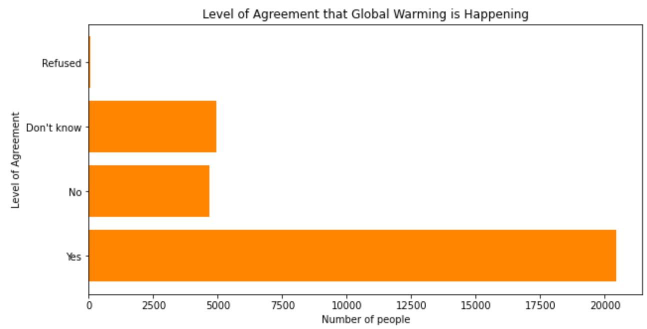

Figure 2: Overall *happening* bar chart

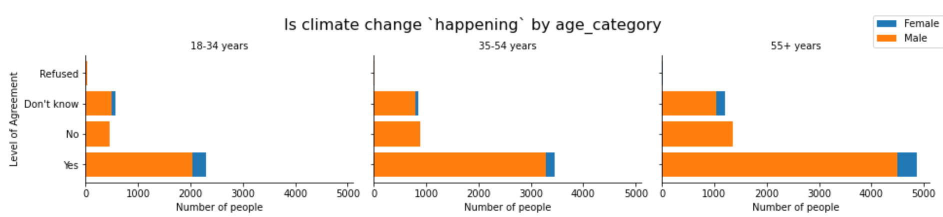

Figure 3: *happening* bar chart by *age_category*

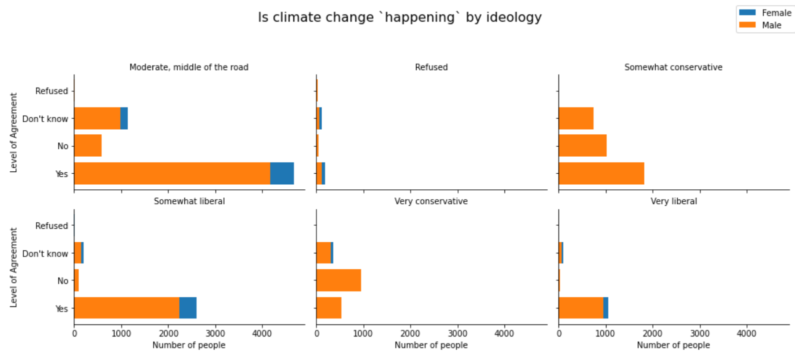

Figure 4: *happening* bar chart by *ideology*

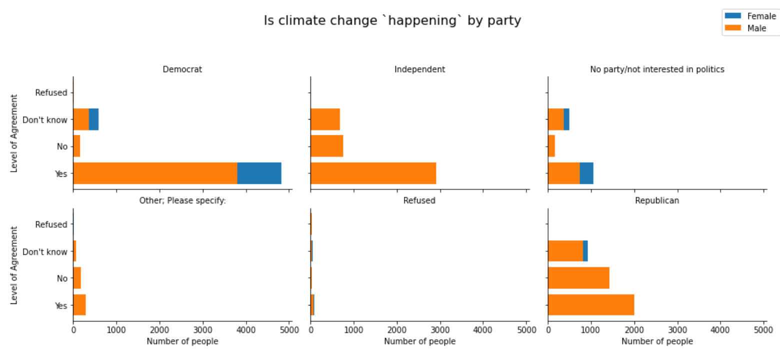

Figure 5: *happening* bar chart by *party*

```
      Formula: yes['1'] ~ gender + age_category + educ_category + income_category + race + ideology + party + region4 + religion + marit_status
       Family: bernoulli
         Link: p = logit
 Observations: 13488
       Priors:
  target = p
       Common-level effects
           Intercept ~ Normal(mu: 0.0, sigma: 9.3088)
           gender ~ Normal(mu: 0.0, sigma: 5.0)
           age_category ~ Normal(mu: [0. 0.], sigma: [5.3083 5.018 ])
           educ_category ~ Normal(mu: [0. 0. 0.], sigma: [5.6679 9.9517 5.474 ])
           income_category ~ Normal(mu: [0. 0.], sigma: [5.2643 5.2212])
           race ~ Normal(mu: [0. 0. 0.], sigma: [8.1456 9.8719 5.7021])
           ideology ~ Normal(mu: [0. 0. 0. 0. 0.], sigma: [21.1589  6.1762  6.229   8.1917  9.2036])
           party ~ Normal(mu: [0. 0. 0. 0. 0.], sigma: [ 5.7448  8.9414 14.5967 29.1427  5.6895])
           region4 ~ Normal(mu: [0. 0. 0.], sigma: [6.4396 5.2648 5.8706])
           religion ~ Normal(mu: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.], sigma: [13.1565  7.6317
               33.399   5.8725 41.9824 35.8267 15.5319 19.0278 40.7335
             7.3533 19.9156  8.0154 15.5954  6.1366 25.9876])
           marit_status ~ Normal(mu: [0. 0. 0. 0. 0.], sigma: [10.6782  5.0785  6.4347 19.8255 11.768 ])
```

Figure 6: Model 1 with `Bambi` default Normal priors. ALL possible predictors

```
      Formula: yes['1'] ~ gender + age_category + educ_category + income_category + race + ideology + party + region4 + religion + marit_status
       Family: bernoulli
         Link: p = logit
 Observations: 13488
       Priors:
  target = p
       Common-level effects
           Intercept ~ Normal(mu: 0.0, sigma: 100.0)
           gender ~ Normal(mu: 0.0, sigma: 100.0)
           age_category ~ Normal(mu: 0.0, sigma: 100.0)
           educ_category ~ Normal(mu: 0.0, sigma: 100.0)
           income_category ~ Normal(mu: 0.0, sigma: 100.0)
           race ~ Normal(mu: 0.0, sigma: 100.0)
           ideology ~ Normal(mu: 0.0, sigma: 100.0)
           party ~ Normal(mu: 0.0, sigma: 100.0)
           region4 ~ Normal(mu: 0.0, sigma: 100.0)
           religion ~ Normal(mu: 0.0, sigma: 100.0)
           marit_status ~ Normal(mu: 0.0, sigma: 100.0)
```

Figure 7: Model 2 with weakly informed Normal priors. ALL possible predictors

```
     Formula: yes['1'] ~ educ_category + ideology
      Family: bernoulli
        Link: p = logit
Observations: 13488
      Priors:
  target = p
      Common-level effects
          Intercept ~ Normal(mu: 0.0, sigma: 4.019)
          educ_category ~ Normal(mu: [0. 0. 0.], sigma: [5.6679 9.9517 5.474 ])
          ideology ~ Normal(mu: [0. 0. 0. 0. 0.], sigma: [21.1589  6.1762  6.229   8.1917  9.2036])
```

Figure 8: Model 3 with `Bambi` default Normal priors. With select predictors.

```
     Formula: yes['1'] ~ educ_category + ideology
      Family: bernoulli
        Link: p = logit
Observations: 13488
      Priors:
  target = p
      Common-level effects
          Intercept ~ Normal(mu: 0.0, sigma: 100.0)
          educ_category ~ Normal(mu: 0.0, sigma: 100.0)
          ideology ~ Normal(mu: 0.0, sigma: 100.0)
```

Figure 9: Model 4 with weakly informed Normal priors. With select predictors.

	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
FullModel_BambiDefaultPriors	0	-4997.378975	45.974433	0.000000	0.931106	67.946761	0.000000	False	log
FullModel_WeakPriors	1	-4997.953187	46.564411	0.574212	0.000000	67.941599	0.184091	False	log
SimpleModel_BambiDefaultPriors	2	-5224.408353	9.126264	227.029377	0.068894	67.751997	22.568917	False	log
SimpleModel_WeakPriors	3	-5224.543081	9.265582	227.164106	0.000000	67.770853	22.572547	False	log

Figure 10: Model comparison table with LOO Cross Validation

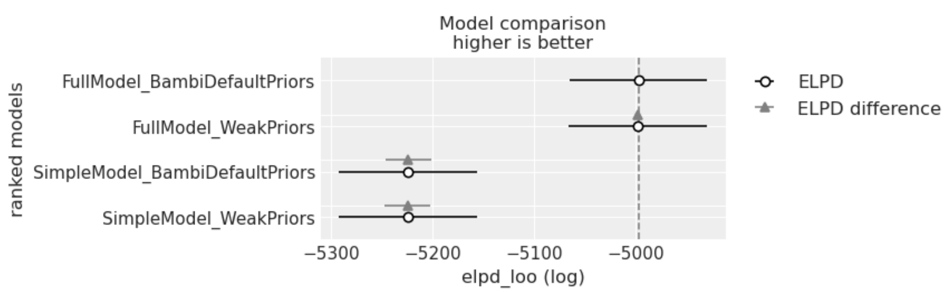

Figure 11: Model comparison ELPD plot with LOO Cross Validation

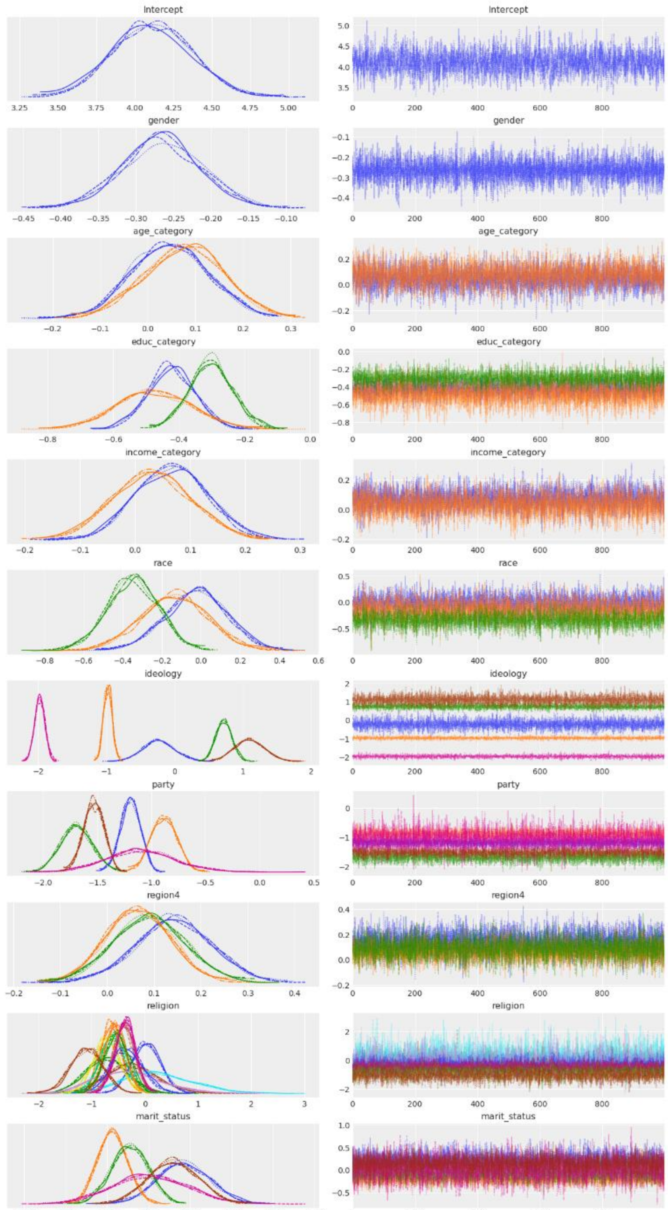

Figure 12: Trace plots summary of model 1

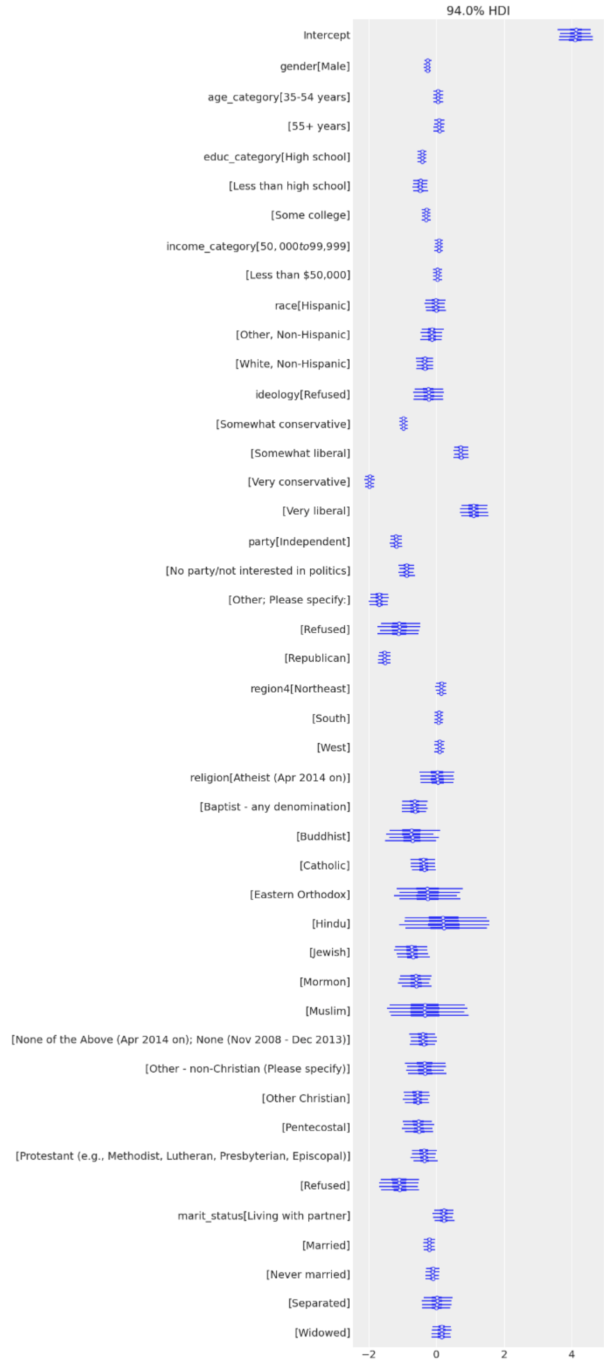

Figure 13: Forest plots summary of model 1

Figure 14: Separation plot using test data from model 1

Citations

Yale Program on Climate Change Communication (YPCCC) & George Mason University Center for Climate Change Communication (Mason 4C). (2022). Climate Change in the American Mind: National survey data on public opinion (2008-2022) [Data file and codebook]. doi: 10.17605/OSF.IO/JW79P

Ballew, M. T., Leiserowitz, A., Roser-Renouf, C., Rosenthal, S. A., Kotcher, J. E., Marlon, J. R., Lyon, E., Goldberg, M. H., & Maibach, E. W. (2019). Climate Change in the American Mind: Data, tools, and trends. Environment: Science and Policy for Sustainable Development, 61(3), 4-18. doi: 10.1080/00139157.2019.1589300