Assessing Results of 2020 American Election by Region

Daniel Grant

02/11/2020

Assessing Results of 2020 American Election by Region

Daniel Grant

Nov 2nd 2020

Intro

Our objective of this report is to use the supplied data from IPUMS USA and Nationscape by Democracy Fund to predict the election results of the 2020 American elections. This will be preformed using multilevel regression and poststratification on our supplied data. These methods will help in our analysis of our data. When we are looking at our data, we can see that it is divided up by different regions that can be used to decide how each region is looking to vote in the upcoming election. We will be looking particularly if someone in the region is going to vote for Trump or not, a binary outcome. Because our data is binary, we are going to approach this problem with a Bayesian multilevel regression model. This model was chosen because we want to know what the chances that, based upon the age, gender, and region that the voter lives in. This paper would not be possible without the generosity of IPUMS USA and Nationscape as well as the help from both Rohan and Sam. In the following paper we will use and describe the models that have been discussed.

Model

Here we are interested in predicting the popular vote outcome of the 2020 American federal election (include citation). To do this we are employing a post-stratification technique. In the following sub-sections I will describe the model specifics and the post-stratification calculation.

Model Specifics

In this paper we will be using a multilevel regression method to approach our data. Multilevel models (MLMs) are useful for the application that we are looking at because it can offer a great amount of flexibility. Through using a MLM we can model data on different levels at the same time. Our focus is on how each region of the United States will vote in the coming 2020 election. Our model is multilevel and is as follows,

$$y = \beta_{0j} + \beta_1 x_{age} + \beta_2 x_{gender} + \epsilon$$
$$\beta_{0j} = \mu_{00} + \mu_{01} W_{region} + \rho_{0j}$$

Through this model we are able to preform our multilevel regression. Our y represents the proportion of our voters that will be voting for Donald Trump in the 2020 election. Our β_{0j} is our intercept which we will have multiple of as it is part of our multilevel regression model, which is why we have a second equation present above. As seen in the second equation that is present, we are going to have multiple intercepts that will correspond to our region variable.In our second equation we have β_{0j} which will be the result of our multiple intercepts. In our β_{0j} equation we have μ_{00} which will be th intercept of our intercepts. So, for everyone one

unit increase in age, we expect a β_1 increase in the probability of voting for Donald Trump and if one is Male then we expect a β_2 increase in the probability of voting for Donald Trump. The model that we are using is Bayesian and was chosen because our outcome in binary and the brms package can be used with great effect towards this data. With this step of our analysis we are trying to estimate our response (voting for Donald Trump) within each cell (Region). When working with an MLM we are dealing with multiple variables, which we can divide into individual level and group level variables, our independent variables would be age and gender while our region data will be our group level variable.

Post-Stratification

In order for us to be able to estimate the portion of voters that will be voting for Donald Trump in the 2020 election we will be preforming a post-stratification analysis. To do this we will be creating different cells that are based on the region of the United States that the person lives in. As mentioned previously there are 9 regions of the united states that one could reside in. We will first estimate the proportion of voters in each region, then we will weight each proportion estimate within each region and divide by our population total. We will be doing so using the following equation:

$$y^{\hat{P}S} = \sum N_j \hat{y} / \sum N_j$$

By finding Y hat we are able to estimate our values for each cell. By finding these cells we are now able to find the proportion of each region of the United States that will be voting for Donald Trump in the upcoming election.

Results

When looking at the results of our post-stratification of our MLM we are able to see what estimated proportion of the population of each region that may vote for Donald Trump. This can be visualized through both an informative table of our results and a density graph.

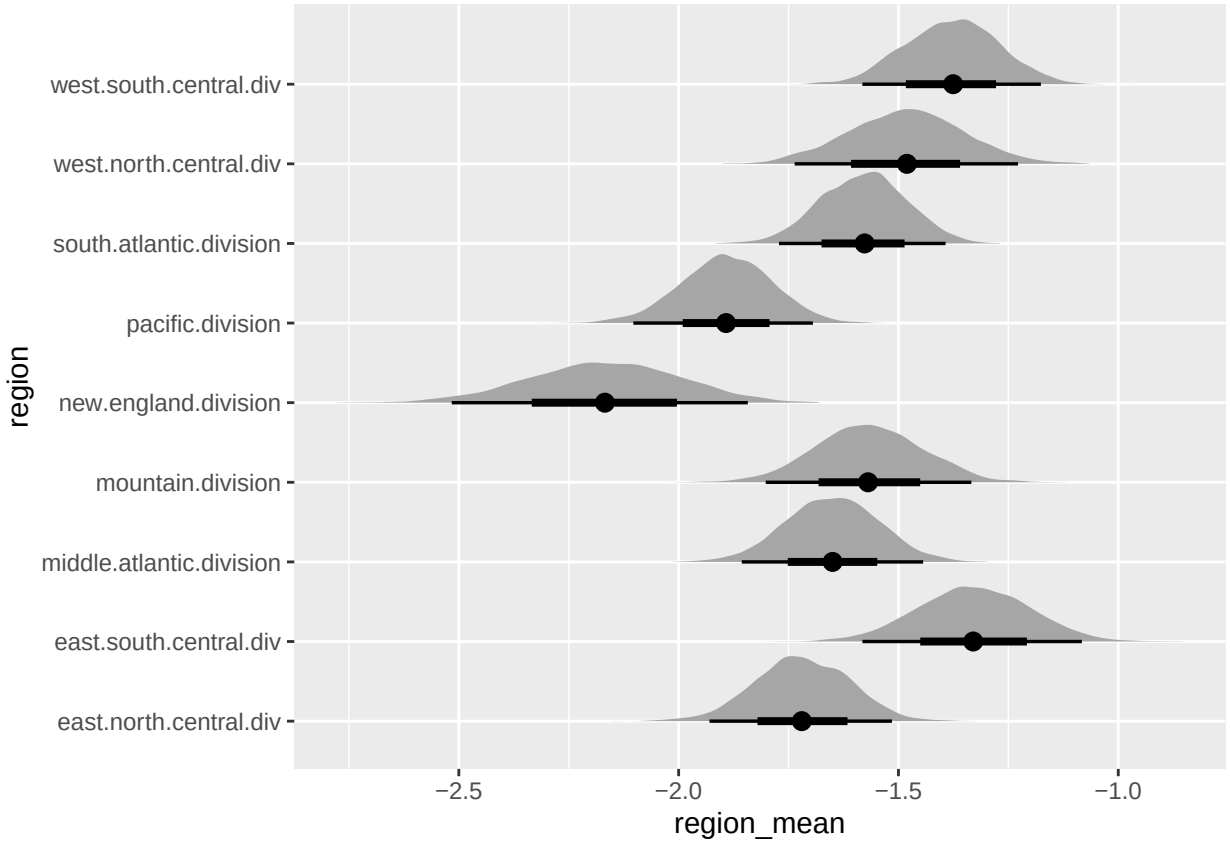
```
post_strat_region
```

```
## # A tibble: 9 x 4
##   region                     mean lower upper
##   <chr>                     <dbl> <dbl> <dbl>
## 1 east north central div   0.358 0.285 0.433
## 2 east south central div   0.444 0.360 0.531
## 3 middle atlantic division 0.374 0.298 0.453
## 4 mountain division        0.387 0.309 0.470
## 5 new england division     0.271 0.190 0.358
## 6 pacific division         0.317 0.244 0.391
## 7 south atlantic division  0.393 0.320 0.469
## 8 west north central div   0.411 0.329 0.494
## 9 west south central div   0.425 0.345 0.503
```

(Table.1)

When we are looking at Table.1 we are able to see that we have been supplied with a mean, lower and upper for the estimates of voting for Donald Trump. Our most pressing question in this paper is, will Donald Trump win a re-election in the 2020 election. By analyzing our Table.1 we are able to see that no matter which region you reside in, the estimated proportion for residents of these regions voting for Donald Trump isn't above .500. These results make it safe to presume that the winner of the 2020 election will not be Donald Trump, as we have not analyzed any other candidates, we cannot presume a winner. To visualize our data, we are able to present the following density graph:

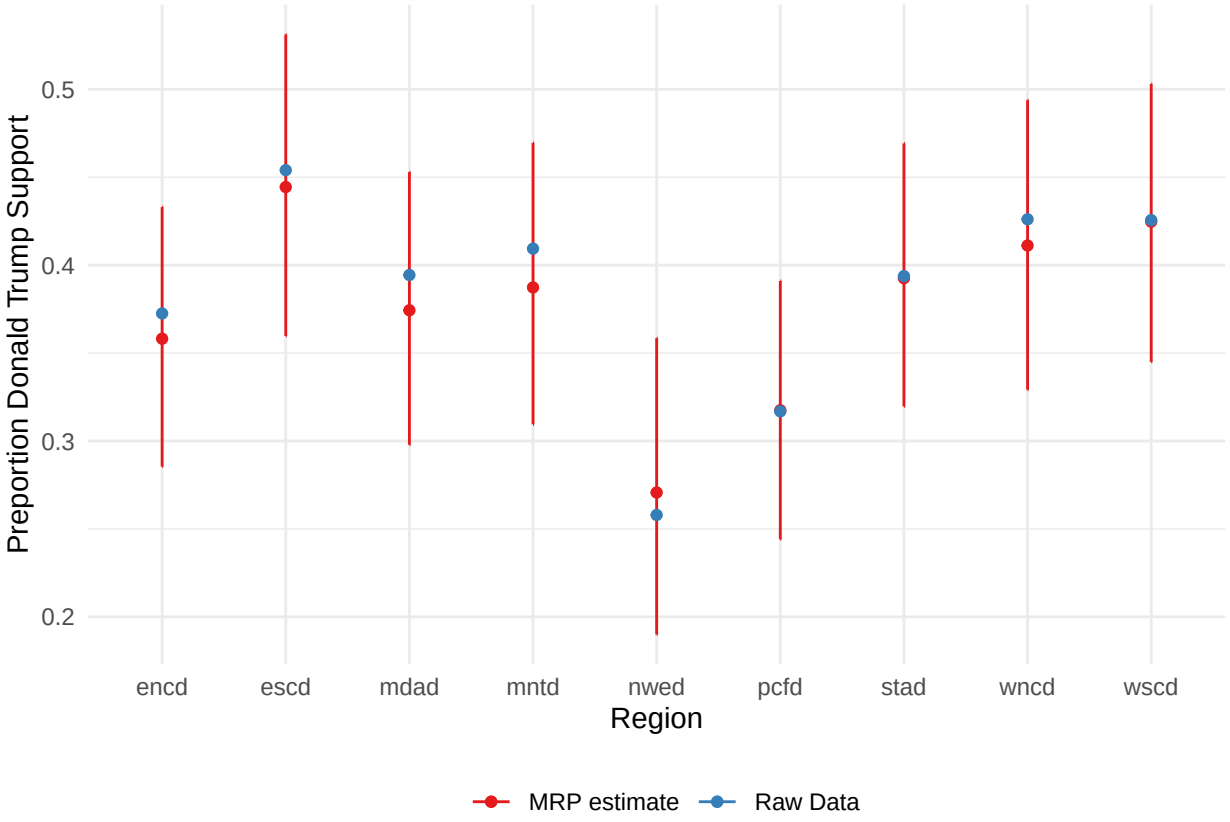
```
density_plot
```



(Figure.1)

When looking at Figure.1, which presents the intercepts of each of the regions as well as the distribution of the data based upon the age and gender of the proportions of the population. We are able to look at another graph that will plot our post-stratification data:

```
MRPplot + scale_x_discrete(labels = abbreviate)
```



(Figure.2)

We are presented with Figure.2, who's code was provided to us by Rohan. We can observe from Plot.2 that there is a significant difference between our raw data and our MRP estimate. We are also able to see, visualized now, that there is a great deal of difference between our divisions of the United States, with the New England Division there is a very low proportion of people that will be voting for Donald Trump at .271 which we can compare to the East South Central Division which sits at a proportion of .446.

Discussion

The goal of this study was to estimate if Donald Trump will win the 2020 United States Election. From the previous sections that we have been discussing we know that the estimated proportion that may be voting Donald Trump is not significant enough to win the election. We can refer back to Table.1 to see our results from our post-stratification. In these results we can see that, by preforming a simple function we can see the national mean which is ~.375. This means that if our data is correct then the proportion of the United States that will vote for Donald Trump will be far below the needed amount for him to win the election.

Weaknesses

In this section of our paper we will be discussing how we can improve in the future and how one could improve the surveying methods or the analysis. The weakness of our model resides in the fact that we have divided the information into only 9 regions. It is hard to find information on these 9 regions and as such it is hard to interpret our data meaningfully when discussing chances of Donald Trump winning the 2020 US election.

Next Steps

In this section of or paper we will be discussing what work can be done in this subject to further our study. The first step that we should approach next is to see what the election results will be on November 3rd 2020. Once the election results are out, we should be able to see if our prediction lined up with the actual results. After that simple step we should be looking if we can get more out of the data perhaps by preforming a deeper analysis on the data. In this report we used the brms package to preform a Bayesian regression on our data because it was what fit the data the best at the time. We should approach our data from a variety of different angles and models to see which one would result in a better fit which may take a much longer amount of time. Because our raw data from both IPUMS and Nationscape are so vast we can perform a deep analysis. As far as improvements towards our surveying and our raw data it would be hard to ask for more than both organizations have provided for us. If we were able to have the same resources as these organizations perhaps we would be able to preform another survey that would focus on the regions of the United States and find out different factors that may affect these regions, for example, are these regions historically Republican or Democrat? Which region contains the most swing states? Knowing information like that would make the region output data more interpretable and have more meaning.

References

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200131). Retrieved from [URL].