IDS 702 Team Project 2: StreetRx and Voting in NC

10/25/2021

Part I: StreetRx

Aarushi Verma as Programmer/Writer/Coordinator and Mohammad Anas as Programmer/Writer/Checker

Introduction and Summary

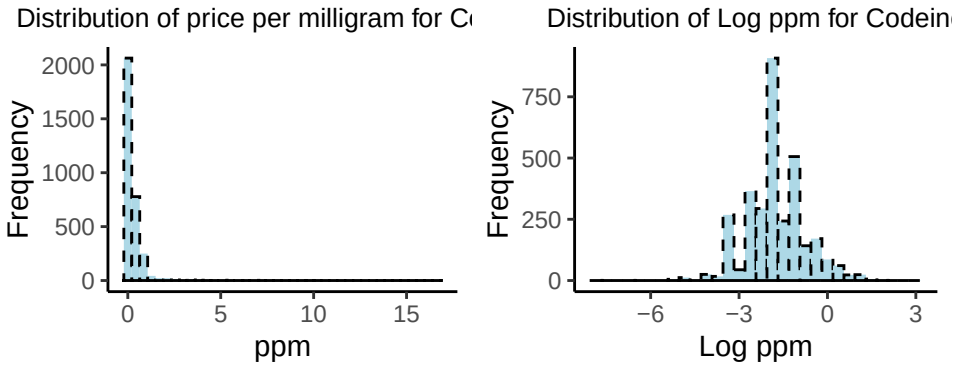
StreetRx (streetrx.com) is a web-based citizen reporting tool enabling real-time collection of street price data on diverted pharmaceutical substances. Users anonymously report prices they paid for prescription drugs on the web. As part of our analysis, we built hierarchical models to investigate how location among other factors may influence the pricing per milligram of a certain drug. According to our analysis, the drug Codeine is found to cheaper in the Michigan, Missouri, Illinois and Texas states. We also found out that purchasing the drug in higher dosages and in huge quantities cad reduce the price paid for Codiene per milligram.

Data

The data used in this analysis pertains to the drug, *Codeine*. We examined our data and observed missing values in columns ppm (price per milligram), mgstr (dosage strength in mg) and source (source of the reported price). We dropped the rows with missing values in ppm and mgstr column. Our original data had 4134 observations for Codeine which reduced to 3125 after missing value removal. Since the source column had high number of missing values and multiple unique values, we categorized the column into *Personal*, *Heard it*, *Internet*, *Not Indicated*. The column mgstr is a discrete variable however, it only contained 3 unique values. We created a new column called dosage to indicate the potency of the drug based on the mgstr values as low, medium or high dosage. In the state column we noted that some states had been incorrectly updated as USA hence, we replaced this value with "Others". We did not consider the form variable in our analysis since it only contained one value - pill/tablet. for the drug Codiene.

EDA

The first step in our EDA was to plot the response variable *ppm* to check whether it follows a normal distribution in order to build a linear regression model. We observed the distribution of *ppm* is highly skewed to the right. We used a log transformation on ppm to address the skewness. The distribution of log ppm was relatively normal and we decided to move ahead with our EDA with log ppm as our response variable.



To explore the data further, we plotted our variables to establish any interesting associations with the response variable. Since all our variables are factor variables, we plotted box plots to identify relationships between our variables.

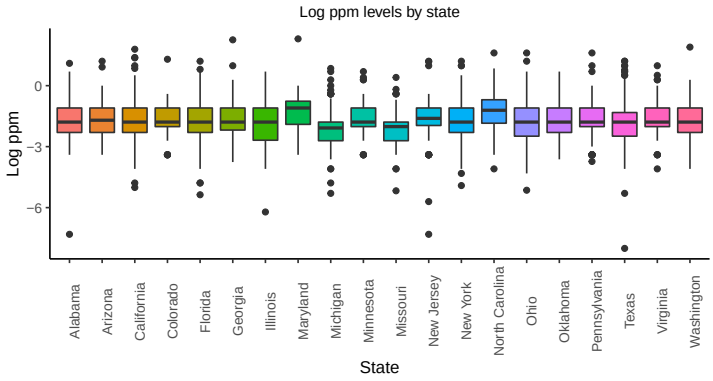
To establish main effects we plotted log ppm against the variables *Source*, *Bulkpurchase* and *Dosage*. We did not observe any change in trend in the plots for log ppm vs. Source and Bulk Purchase. We did note a variation in the trend plot for log ppm and dosage indicating an association between the two. However we must also bear in mind that both variables are derived from the price entered by the user for the drug.

After assessing the relationships between the response and explanatory variables, we went on to further explore the interactions between the explanatory variables . The box plots for interaction between Dosage vs Bulk Purchase and Dosage vs. Source showed no change in trend. In particular, we observed some variation in trend of log(ppm) and bulk_purchase when looked at separately for each source. We concluded to explore this interactions further in our model.

We also plotted the response variable against our location variables - State and USA region to establish whether we should include the random intercept for these variables.

We observed a slight trend change between log ppm and USA region. For state, we saw variation across multiple states. Since states are nested within regions, we decided to assess the impact of location on the price of the drug at a more granular level i.e. at the state level. Hence we included state as our hierarchical variable.

The state variable has more than 50 levels, therefore to look at the trend we plotted a subset of these states. We filtered states with more than 50 observations in our data and plotted them against log ppm to see if there was any variation. We did see a variation and thus we decided to control our intercept of the model by state.



We also investigated whether we need to include any random slopes in our model based on the interaction between the main effects and our grouping variable State. We plotted box plots to investigate the associations and concluded that we may need to include random slopes by State for the Source variables and Bulk Purchase variable based on change in trends across states for ppm for these variables.

Model

To build our model, we first built our baseline linear regression which included all our main effects without controlling any of the intercepts or slopes by state. Next we used step wise selection using AIC to generate our final linear model with main effects. We then included some interaction effects which we thought to be significant, or they answered questions with respect to the study. With the help of anova tests we assessed if these interaction were significant to our model. We then incorporated our random intercepts and random slopes and tested them again using anova to arrive on our final model.

Our first model included the main effect of every variable. Since all effects are factor variables we did not need to center them to improve our interpretation. Here our response variable is **log ppm** and the predictors are **source, dosage and bulk purchase**. Next we used step wise selection using AIC and BIC to assess which variables should we retain in our model. Based on this, we removed the variable source from our model. Given that source was an interesting variable we also tested it using anova and the p value was insignificant at 95% confidence level.

To ensure our final model is the best fit for our data, we also included the one interaction effect we found interesting during our EDA to the model and used the anova test to conclude whether the interaction between source and bulk purchase had a significant impact on our model or not. However, based on anova the interaction also came out to be insignificant which meant there seemed to be no additional impact of those interactions.

Once we had our linear model, we proceeded to include the random intercepts and slopes that were of interest to us. Based on our EDA we included a varying intercept for the state variable. Further, we also included varying slopes by state for bulk

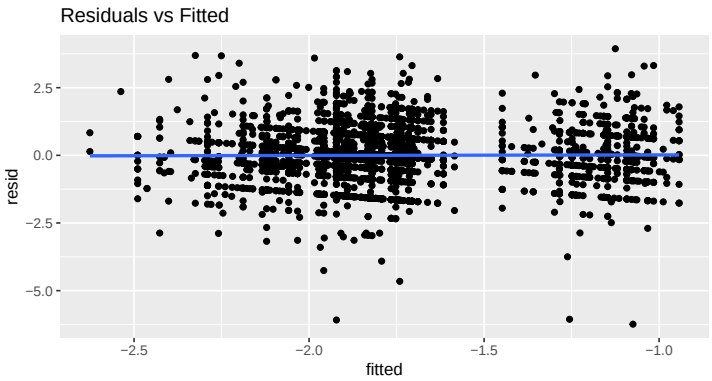
purchase and dosage. In order to analyze whether including varying slopes improved the fit of the model, we used anova to compare each of the varying slope models to our model, which only included the varying intercept. However based on anova, the p value for both the models with varying slopes was insignificant at 95% confidence level indicating that controlling for slopes by state did not improve the fit of our model.

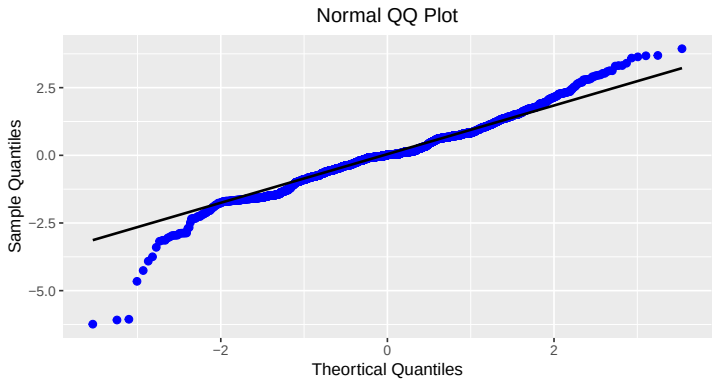
$$y_{i\ state} = (\beta_0 + \gamma_{0\ state}) + (\beta_1)dosage_{1i\ state} + (\beta_1)bulk\_purchase_{1i\ state} + \epsilon_{i\ state};\ i = 1, ..., n_{state};\ state = 1, ..., 59$$

Model Assessment

To assess our final model we checked if the assumptions of Linearity, Normality, Equal variance and Independence were violated. Since all our variables are factor variables, we were unable to verify the linearity assumption.

To check the independence and equal variance assumptions we plotted the residuals against the fitted values. The points seemed randomly distributed with no discernible pattern and the spread of variables seemed constant above and below the line. There did seem to be some points on the x axis that may have violated the equal variance of errors assumptions however, they were only few points and it is safe to say that neither of the above mentioned assumptions were violated. However, this does indicate that there are outliers are present in our data. To check for normality, we plotted the Q-Q plot. For our model we observed that majority of the points lie on the 45 degree line. Both the *Q-Q plot* and the _Residuals vs fitted plot are shown below.





To check if outliers were affecting our model we removed them from our data and ran the model again. However, the standard estimates and their p-values did not change and hence we can conclude that the outliers were not affecting our model.

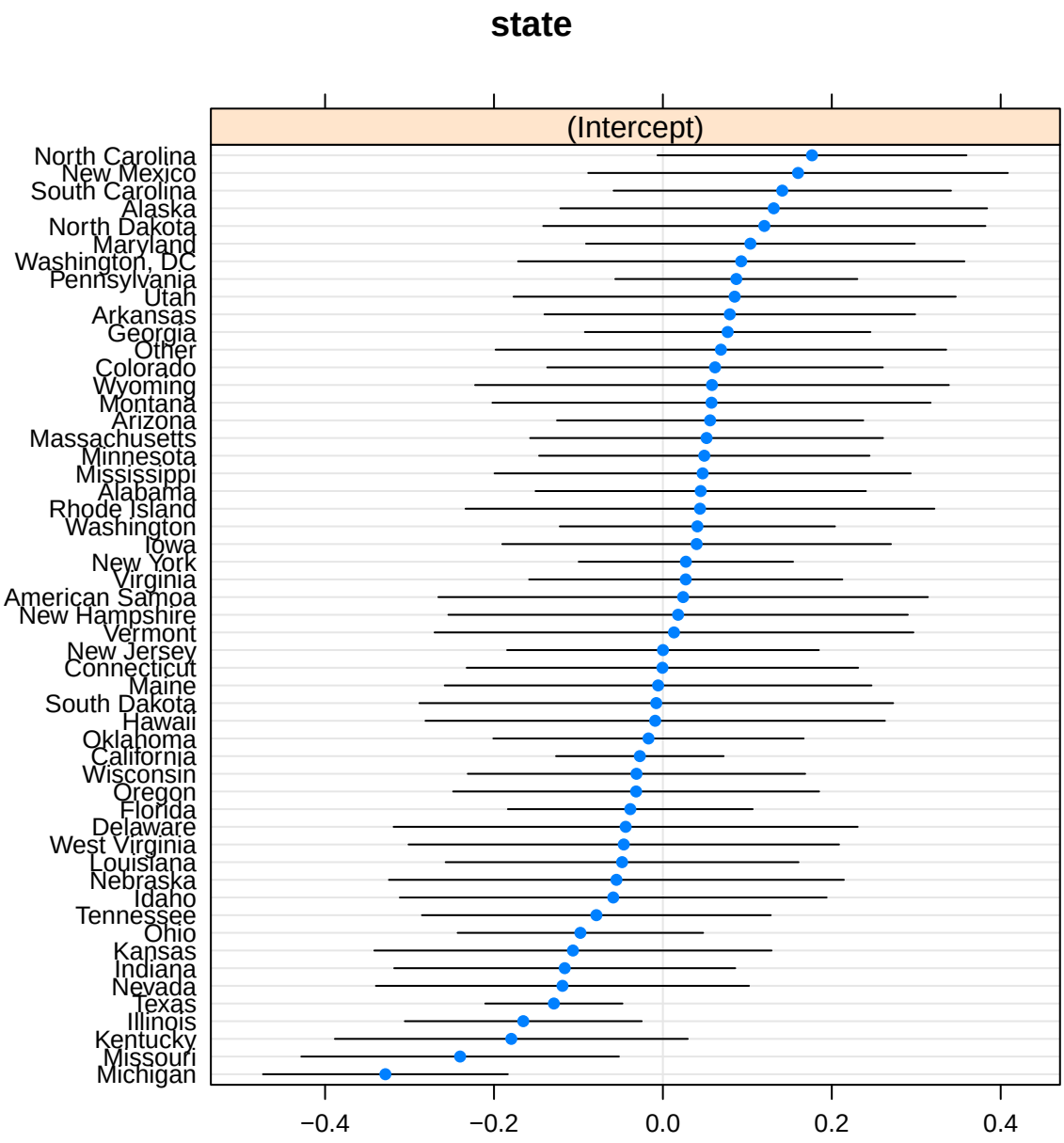
Model Interpretations

We found out that all our explanatory variables had a significant effect on log_ppm. To make the interpretation of the standard estimates simpler we exponentiate them and measure the effect on price. We notice that for low dosage and drug not being purchase in bulk the price was 0.33 USD. If the drug was purchased in high dosage the price per milligram decreased by 65% compared to if it was purchase in low dosage. When purchase in medium dosage the price/mg decreases by 51%. If the drug was purchase in bulk, the price/mg dropped by 87%. We also note that controlling our intercept for state only explains 13.2% of the variation in log(ppm). The summary of our model is shown below.

Looking at the dot plot of random effects we notice that the the drug Codeine is found to be cheaper in the states Michigan, Missouri, Illinois and Texas. The confidence interval of the random effect for the other states contains zero, hence, we can say that the price of the drug does not vary significantly for them. The dotplot is shown below.

Table 1: Hierarchical Model Summary

	Dependent variable:
	log(ppm)
dosagehigh	-1.04^{***} $(-1.18, -0.90)$
dosagemedium	-0.67^{***} $(-0.76, -0.58)$
bulk_purchase1 Bulk purchase	-0.14^{***} $(-0.23, -0.05)$
Constant	-1.12^{***} $(-1.21, -1.02)$
Observations	3,215
Log Likelihood	−4,537.65
Akaike Inf. Crit.	9,087.31
Bayesian Inf. Crit.	9,123.76
Note:	*p<0.1; **p<0.05; ***p<0.01



Limitations and Conclusions

There are a few potential limitations in our model. Firstly, we removed 919 data observations when dealing with missing values. This can be avoided using missing value imputation methods. The second major limitation is that the data used is not reliable as it is crowd sourced and any one can put in any value for the price of the drug.

To conclude, we note that the drug Codeine can be bought for a cheaper price in the Michigan, Missouri, Illinois and Texas states. Purchasing the drug in bulk quantity and at high dosages can also reduce the price of the drug. It was surprising to find out that the source variable did not have a significant affect on the price/mg as we initially hypothesized that prices on the internet to be more expensive than the prices they observed through personal experiences.

Part II: Voting in NC (2020 General Elections)

Deekshita Saikia as Programmer/Writer/Checker, Tego Chang as Programmer/Writer/Coordina
Sydney Donati-Leach as Programmer/Writer/Checker and Aarushi Verma as Presenter

Summary

This analysis is to investigate the potential factors, especially the demographic ones, that affect the turnout rates of the US 2020 General elections in North Carolina. During the process, data cleaning, aggregation, and merging have been performed to combine the registered voters information with the actual voter turnouts. Exploratory data analysis was then performed and the effects of demographic factors on the actual turnouts were examined. The finalized hierarchical model helps us dive into how demographics like age impact the turnout rates in the election, and serves to provide a possible explanation as to why the Republicans outperform the Democrats in certain counties.

Introduction

It is widely acknowledged that the voting behaviors in the election are related to several factors, e.g., demographic factors, location, or the party you are affiliated to. This analysis constructs a hierarchical model to answer the following questions:

- How did different demographic subgroups vote in the 2020 general elections?
- Did the odds of voting differ by county? Which counties differ the most?
- How did the turnout rate differ between males and females for different parties?
- How did the turnout rate differ among age groups for different parties?

Data

Data Pre-Processing

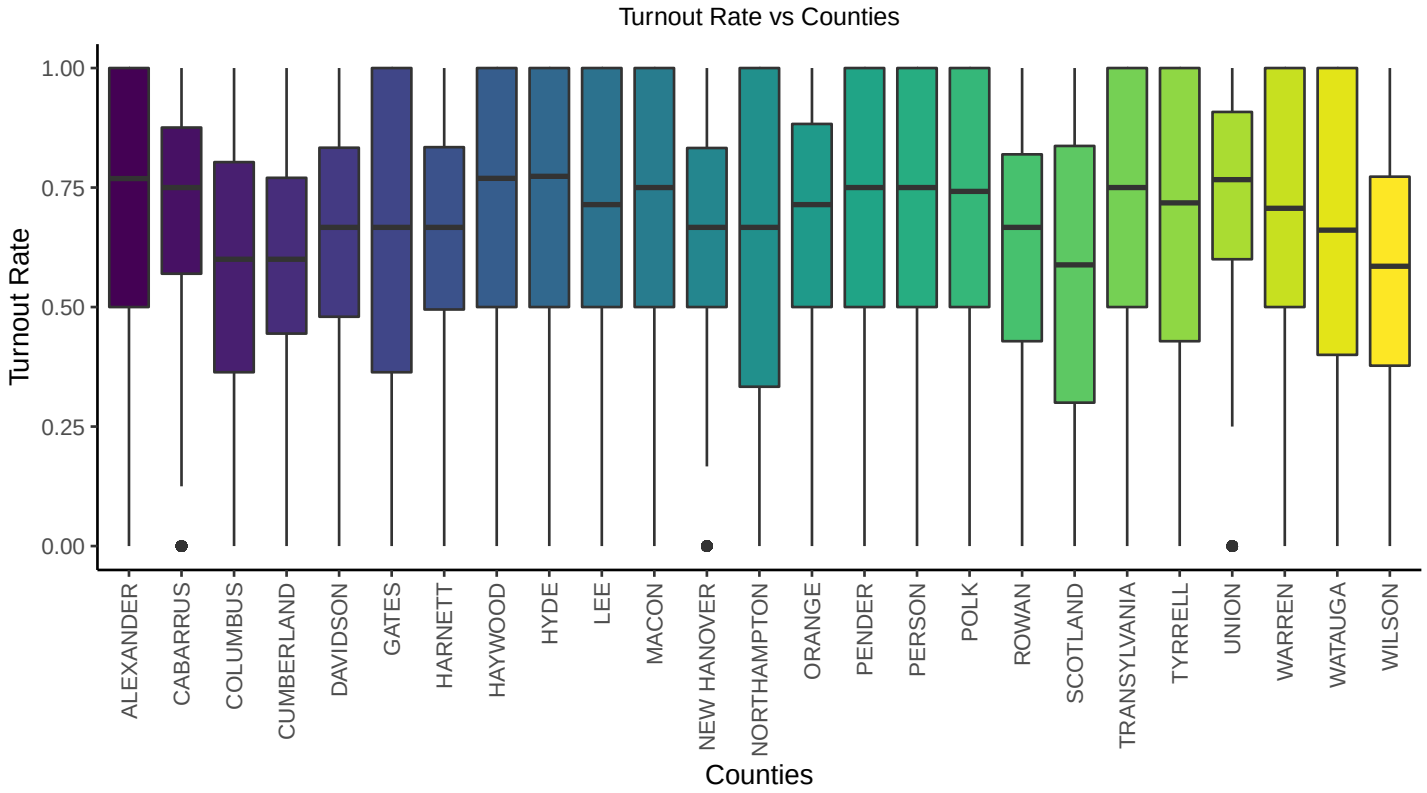
The data used for this analysis was extracted from two files available with The North Carolina State Board of Elections (NCSBE), which is the agency charged with the administration of the elections process and campaign finance disclosure and compliance. One file contains the voter registration records, while the other contains data on the actual turnout (https://www.ncsbe.gov/index.html, https://www.ncsbe.gov/results-data). The code book can be found in the appendix.

The unit of observation in the registered voters file is *county_desc*, *precinct_abbrv*, *vtd_abbrv*, *party_cd*, *race_code*, *ethnic_code*, *sex_code* and *age*. The turnouts file had data at a more granular level (*voting_method*), and it was aggregated to match the unit of observation in the registered voter file. The registered voters file had 592,265 observations, while the turnouts file had 928,532 observations. Post aggregation of the turnouts file, there were 492,567 observations remaining. The actual turnout numbers were then merged with the total voter file to create a model ready dataset. The dataset was further reduced to a sample of 25 counties, post which we were left with 13,162 observations, grouped at a *county_desc*, *party_cd*, *race_code*, *ethnic_code*, *sex_code* and *age* level, with the total number of registered voters and turnouts in demographic groups represented by these characteristics.

EDA

Our EDA is split into 4 different sections. First, we will look at a plot of our hierarchy, county, to determine if it has varying intercept or varying slope. Second, we will look at our main effects, which are all factor variables in our data, so we will only look at boxplots. Third, we will look at all of the interactions with our main effects, again by analyzing boxplots. Finally, we will look at the interactions between our main effects and our set hierarchy to see if we need to control for varying slope.

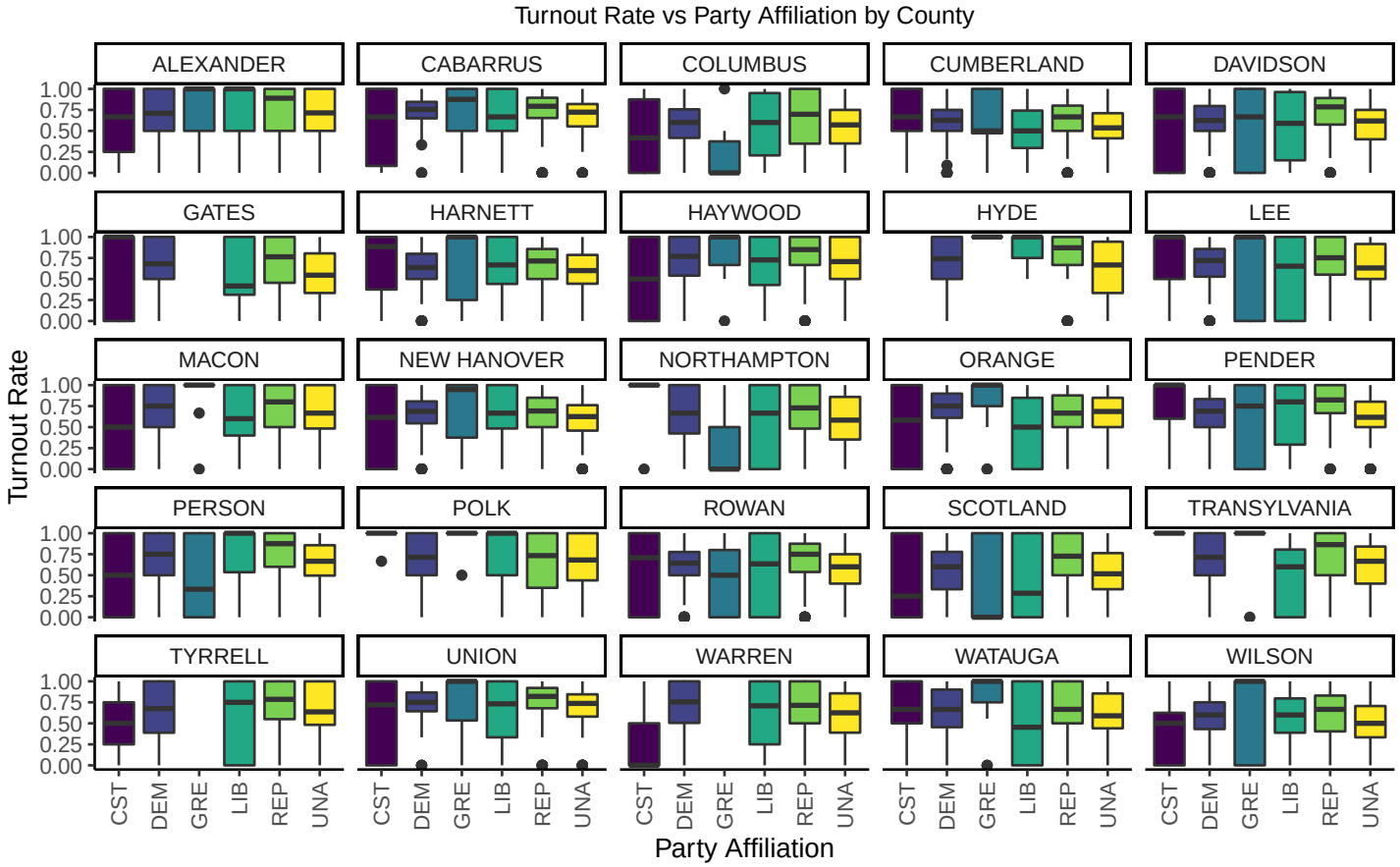
When we look at county, the hierarchy in our data, it clearly varies from county to county. Alexander, Haywood and Hyde have high voter turnout compared to Columbus, Cumberland, Scotland and Wilson. Therefore, we will move forward knowing that we need to include varying intercept in our model building. Next, when we look at our main effects, we can pull some basic information from the boxplots. For example, older age groups coming out to vote more than younger age groups, or that most people choose not disclose their sex when registering to vote.



When we go through all of the interactions of our main effects, we have to consider both the trends in the boxplots as well as the number of observations. This is because a changing trend from one category to the next could be due to a lack of observations rather than an interesting interaction taking place. One example of this is the race of P (Pacific Islander). We sometimes have zero observations of this race when we split it up into different categories, so we have to look at changing trends without consideration for it. This becomes complicated when we also want to look at the interaction between race and party because there are two parties with far fewer observations: CST and GRE. Now we have three different categories that we should *not* take into consideration when looking at trends. Due to these limitations, we decided not to include the interaction between race and party in our model building. On the other hand, there are two interactions of high importance to us: age versus party and sex versus party. These are questions of interest, so regardless of what kind of trend we see, we must include them when model building. We did see that there was a higher median for republicans of ages 26-40 as compared to democrats of the same age group, so we will keep that in mind moving forward.

	CST	DEM	GRE	LIB	REP	UNA
A	10	355	6	55	324	365
B	76	677	59	181	438	580
I	7	347	4	55	303	362
M	7	481	5	89	362	470
O	35	577	23	158	500	585
P	0	29	0	0	31	33
U	106	710	89	239	663	715
W	177	747	162	448	754	763

Finally, we looked at interactions of our main effects and county. We did not find the interaction between county and age or the interaction between county and ethnicity to be as interesting because the trend only changed in 8 or 9 counties out of all 25. When we look at interactions with race and party, we must still keep in mind the low count of observations for some categories. If we do not consider race P when analyzing our race interaction, the trend changes from one county to the next. We also saw the trend continually changing in the interaction between party and county, even when we did not take the CST or GRE parties into consideration. So now we have two interactions with county that we found interesting: race and county and party and county. Even though both of these interactions are interesting, we had to choose only one we wanted to focus on so that we could interpret our final model. We made the decision that we should control the random slope for county versus party, and we will further discuss this in our model building section.



Model

Model selection was performed by accounting for different interactions and effects, which included both random and fixed effects. Our main interactions of interest are analyzing how the turnouts differed by the sexes, and if party affiliations played a role. We are also interested in exploring by the turnouts differed for the age groups for different party affiliations.

We fit a hierarchical model to explore the random effects that different counties may contribute to the model. Counties are used as the only hierarchy. In addition, random slopes for party affiliations were also considered. The model with random intercepts by county was then compared to the model with random slopes for party affiliations with an ANOVA Chi-squared test, and we observe that incorporating the random slope significantly improves model fit.

The final model equation is as follows:

$$y_i|x_i \sim Bernoulli(\pi_i);\ i = 1, 2, \ldots, 13162;\ j = 1, 2, \ldots, 25$$

$$(\beta_0 + \gamma_{0j|i|}) + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \beta_3 * x_{i3} + \beta_4 * x_{i4} + \beta_5 * x_{i5} + \beta_6 * x_{i6} + (\beta_7 + \gamma_{1j|i|})x_{i7};$$

where, x_{i1} is *age*, x_{i2} is *race_code*, x_{i3} is *ethnic_code*, x_{i4} is *sex_code*, x_{i5} is the interaction effect of *age* and *party_code*, x_{i6} is the interaction between *sex_code* and *party_cd*, and x_{i7} is *party_cd*.

$$\gamma_{0j}, \gamma_{1j} \sim N_2(0, \Sigma)$$

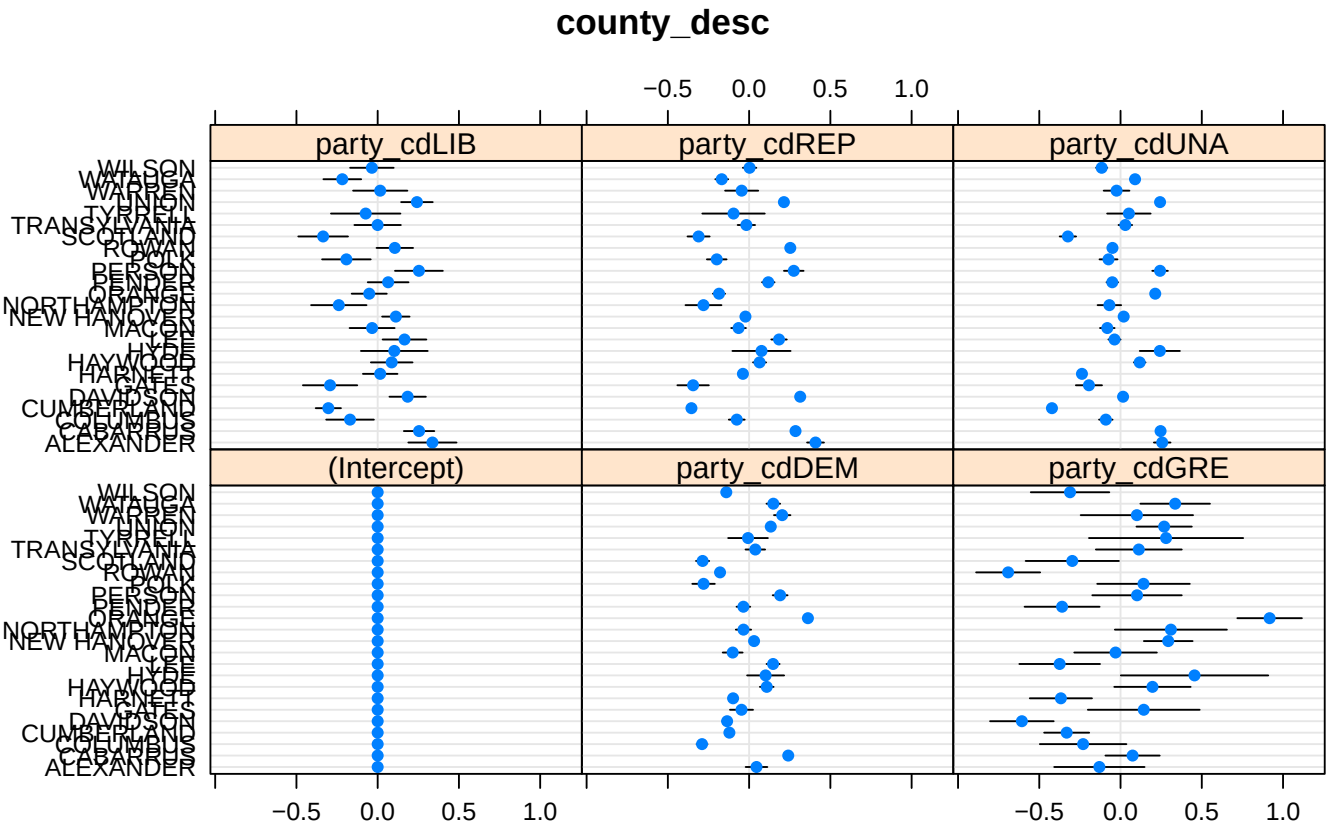
The model results obtained are as shown below:

	Dependent variable:
	cbind(turnout, total_voters - turnout)
ageAge 26 - 40	0.559^{***} (0.175)
ageAge 41 - 65	1.000^{***} (0.183)
ageAge Over 66	0.804^{**} (0.346)
party_cdDEM	0.677^{***} (0.184)
party_cdGRE	0.790^{***} (0.265)
party_cdLIB	0.114 (0.192)
party_cdREP	0.710^{***} (0.186)
party_cdUNA	0.288 (0.184)
race_codeB	-0.155^{***} (0.018)
race_codeI	-0.161^{***} (0.029)
race_codeM	-0.078^{***} (0.027)
race_codeO	-0.227^{***} (0.020)
race_codeP	1.934^{***} (0.337)
race_codeU	0.309^{***} (0.020)
race_codeW	0.197^{***} (0.017)
ethnic_codeNL	0.420^{***} (0.011)
ethnic_codeUN	0.334^{***} (0.012)
sex_codeM	0.058 (0.163)
sex_codeU	0.356^{*} (0.197)
ageAge 26 - 40:party_cdDEM	-0.515^{***} (0.176)
ageAge 41 - 65:party_cdDEM	0.080 (0.183)
ageAge Over 66:party_cdDEM	0.334 (0.346)
ageAge 26 - 40:party_cdGRE	-0.490^{*} (0.256)
ageAge 41 - 65:party_cdGRE	−0.217 (0.306)
ageAge Over 66:party_cdGRE	−0.347 (0.579)
ageAge 26 - 40:party_cdLIB	-0.429^{**} (0.183)
ageAge 41 - 65:party_cdLIB	−0.307 (0.192)
ageAge Over 66:party_cdLIB	0.136 (0.368)
ageAge 26 - 40:party_cdREP	-0.389^{**} (0.176)
ageAge 41 - 65:party_cdREP	0.050 (0.183)
ageAge Over 66:party_cdREP	0.316 (0.346)
ageAge 26 - 40:party_cdUNA	-0.362^{**} (0.176)
ageAge 41 - 65:party_cdUNA	0.111 (0.183)
ageAge Over 66:party_cdUNA	0.624^{*} (0.346)
party_cdDEM:sex_codeM	-0.303^{*} (0.163)
party_cdGRE:sex_codeM	−0.047 (0.256)
party_cdLIB:sex_codeM	−0.052 (0.169)
party_cdREP:sex_codeM	−0.096 (0.163)
party_cdUNA:sex_codeM	−0.176 (0.163)
party_cdDEM:sex_codeU	−0.177 (0.197)
party_cdGRE:sex_codeU	−0.430 (0.280)
party_cdLIB:sex_codeU	0.391^{*} (0.207)
party_cdREP:sex_codeU	−0.004 (0.197)
party_cdUNA:sex_codeU	-0.351^{*} (0.197)
Constant	-0.631^{***} (0.181)
Observations	13,162
Log Likelihood	−32,112.000
Akaike Inf. Crit.	64,356.000
Bayesian Inf. Crit.	64,850.010
Note:	*p<0.1; **p<0.05; ***p<0.01

In this model, age group 18 - 25 year olds, the Constitution Party, the Asian race, the Hispanic/Latino ethnicity and the female sex are used as baseline factors that are absorbed into the intercept, for ease of interpretation of the impact of these predictors in voter turnouts. All the predictors and the interaction terms are categorical, and a total of 44 distinct factors can be seen in the final model. We observe that the fixed effects of the model are significant at the 5% level. The largest

z-values can be observed for *race_code*, *ethnic_code* and *age*, signifying that these factors are the strongest predictors of whether a person if likely to turn up to vote in the elections.

$county_desc



From the dotplot of the random effects above, we observe that introducing the random slope effect of party shows that the log odds of a person voting across different counties differs greatly by the party they are affiliated to. Counties like Alexander, Cabarrus, Orange and Union have odds of voting which are significantly different than zero. People voting for the Green Party have highly varying odds, but this can also be attributed to the lower number of observations recorded against this party. Although the plot confirms that parties account for a lot of the variance in the voting odds we see across counties, it still does not explain all the variation in the model.

Conclusion

The answers to our questions of interest are:

- The odds of turnout for males is 1.06 times compared to females, which is six percent higher. However, it is not statistically significant. The odds of turnout for age group 41 to 65 is found to be the highest, which is 2.7 times compared to the age group 18 to 25. With respect to race, when the Asian race is set as the baseline, the odds of turnout for Pacific Islanders, Undesignated, and White have higher odds of turnout. However, there are very few observations for Pacific Islanders. With respect to ethnic groups, when the Hispanic/Latino category is set as the baseline, the non-hispanic/non-latino category have higher odds of turnout.
- The odds of voting differ by county in 2020 from our EDA, so we added varying intercepts by county in our model. Further, we also observed changes in trends when checking if there is an interaction between party and county. Thus, varying slopes by county were also added to the model. In this case, the varying intercept in our model is 0, but if we select a party, say Democrat, Wake county has the highest odds of turnout, while Randolph and Mcdowell have the lowest ones.

- Observing the interaction between age and party, there are many factors that are statistically significant. One of the interpretations for these combinations is that the odds of turnout for age group 26 to 40 who vote for Democrats is 2.03 times, increasing by 103%, compared to the age group 18 to 25 who vote for CST.

Besides the above questions of interest, we further dive into the voters' age groups and investigate which age group has a higher impact on the actual turnouts between Democrats and Republicans. Observing the interaction between age and party, the 26 to 40 age group is the only statistically significant factor compared with the two other age groups, with age group 18 to 25 as the baseline. Then, we further calculate the odds of turnout between the Democrats and Republicans for this age group, and we observe that the turnouts for Republicans is higher than for Democrats, with odds of turnout observed at 2.41 for the former and 2.05 for the latter when the baseline is the age group of 18 to 25 voting for CST. Another possible explanation behind this might also be the fact that Democrats and Republicans are the two biggest parties, and hence, many observations are available against these parties, compared to the Constitution and Green parties.

Limitations

One limitation of this analysis is that, the NCSBE does not provide the exact difference between the variables *party_cd* and *voted_party_cd* in the actual voter turnouts file. If voters changed their party affiliation at the time of casting their votes, this information is not captured in the dataset, and would not tally with the registered voter numbers. In addition, a random effect model with a varying slope and a varying intercept leads to very complicated interpretations of the model parameters. Since the data is grouped at a combination of different demographics, which makes it granular, we observe that not all groupings have enough observations recorded against them to fit a model on. Certain categories could be aggregated out so that there are sufficient observations against each grouping.