# Final Report

Hannah Gong

10/17/2020

# PART I

## Introduction

The group of hormones known as estrogens are responsible for regulating bones, skin, and organ growth, among other essential bodily functions. Estrogens often serve to disrupt processes within the endocrine system (hormone pathways), which is completely benign when occurring as a natural bodily process. Over the past ten years, however, the increased production of many synthetic and plant products has raised concerns over the potentially adverse impacts these chemicals have on the human body. Chemicals that have estrogenic effects have been labeled environmental estrogens.

One method for testing the estrogenic effects of these environmental estrogens is the rat uterotrophic bioassay, which observes the uterus growth response to varying levels of estrogen agonists and antagonists. The expectation is that uterus weight will increase as the amount of prescribed estrogen agonist increases and decreases as the amount of estrogen antagonist increases.

This analysis examines the data from one such uteroptophic bioassay study to determine if the process was successful in identifying both the estrogenic effects of environmental estrogens, and the anti-estrogenic effects of antagonist chemicals. This analysis also examines if and how these effects vary across the laboratories involved in the study and examines how different dosing protocols may have affected the ability to detect the effects of the two chemicals.

## Data Pre-processing

The dataset was accessed via international multilaboratory study on the effect of estrogen agonists and antagonists on the weight of uterus with 2677 non-empty observations. The cleaned data includes three categorical variables – protocol, group and lab – and four numeric variables – uterus weight in mg, the body weight of rats in grams, dose of estrogen agonist (EE) and dose of estrogen antagonist (ZM). The character and numeric values for protocol, group, and lab were factorized. The nature of the data allows us to organize the data with multiple levels. Potential rat groupings to be explored are by lab, protocol, and group.

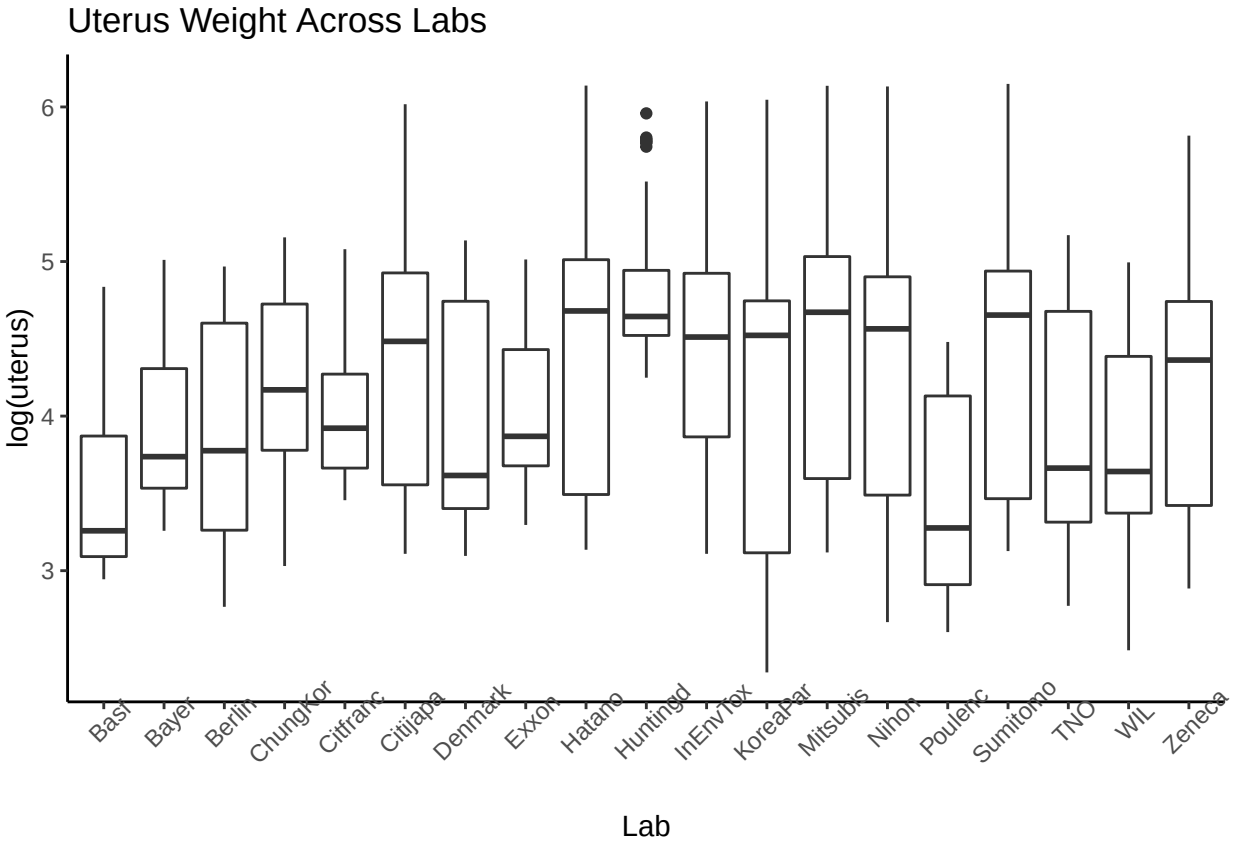
## EDA

### Response variable Uterus

The distribution of the response variable, uterus weight, is heavily skewed right. Applying a log transformation improves the normality of distribution and will be used moving forward.

**Control by Lab**

The distribution of Uterus Weight is not identical across all labs but there do not appear to be any significant outliers. All labs have sufficient rat observations. Including variable intercepts by Lab will be explored in the final model to account for this variance, but we do not see any signs that variable slopes are needed.

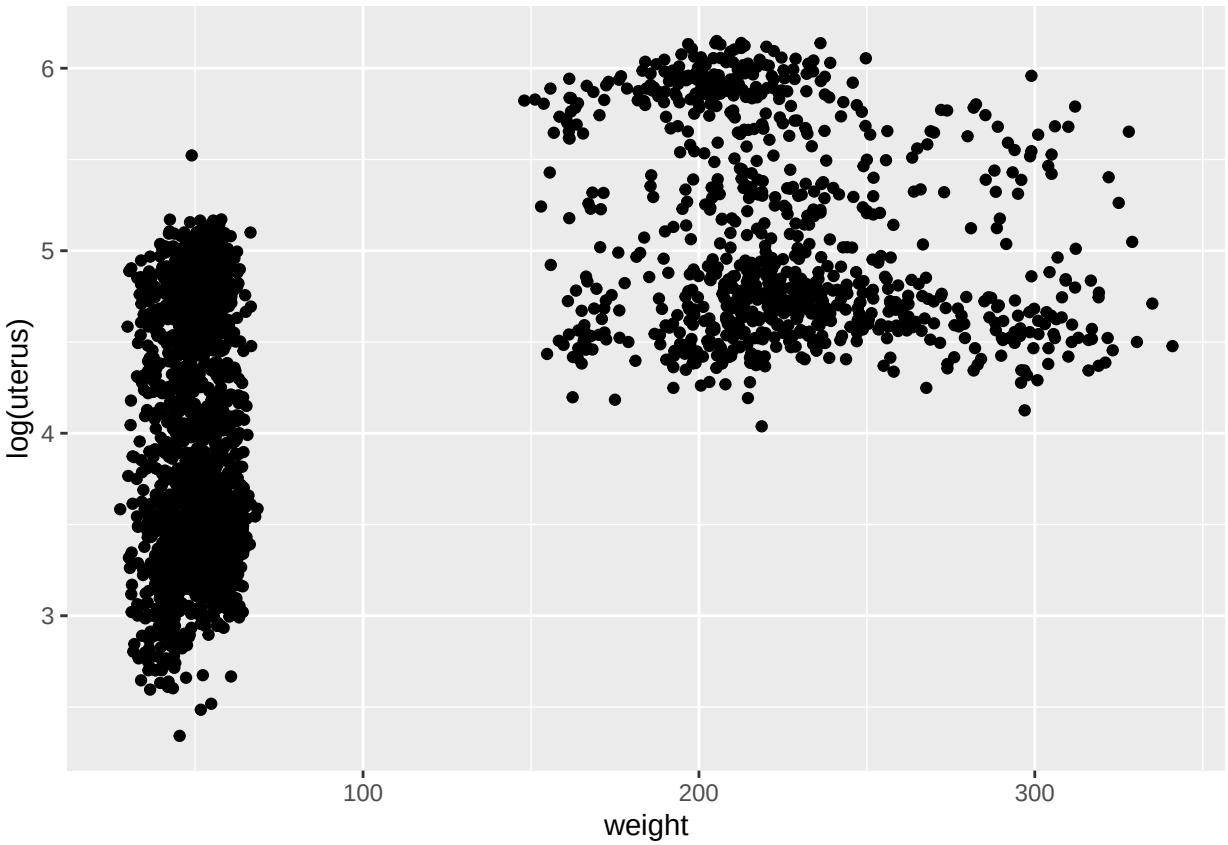



**The impact of weight on Uterus**

The distribution of Uterus Weight across protocols appears to fall into a high weight and low weight cluster - with protocols A and B in the lower cluster and protocols C and D in the upper cluster. While not every lab carried out every protocol, the trend of A/B lower than C/D holds across all labs. All protocols have sufficient rat observations. Including variable intercepts by Protocol will be explored in the final model to try and account for this variance. The two clusters could be due to the two types of rats (mature vs immature) used in protocols A/B vs C/D.

Uterus Weight is not consistent across all groups, with groups 6, 7, 8 and 9 having higher weights than the others. This trend is consistent across all the labs in the study. Further analysis of the data shows that the differences between groups are differences of Protocol and EE/ZM dosage, which are variables whose relationships we are more interested in exploring. As such we will not be including the Group variable moving forward.

A scatterplot of Weight vs Uterus Weight shows 2 distinct clusters of rats - similar to that seen when grouping by protocol. This is consistent across all labs. When coloring points by EE dosage the heavier rats in both clusters appear to be correlated with higher EE doses. When coloring by ZM dosage, rats with higher ZM dosages tend to fall in the bottom to mid-range of uterus weight. Since the distribution of Uterus Weight by Weight has two distinct clusters, we created a low_weight variable to indicate rats with weights above and below 100g.




Further exploration of Uterus Weight by EE/ZM dosage shows a positive correlation between increased EE levels and Uterus Weight, and a negative correlation between increase ZM levels and Uterus Weight. For EE this trend appears hold constant across all protocols, however for ZM there does appear to be variance in the slopes by protocol. Including variable EE and ZM slopes by Protocol will be explored in the final model to try and account for this variance.

It doesn't look like we need to do varying-slopes for each lab, only varying intercepts. We may want to include varying intercepts by protocol to account for the weight differences due do differently ages mice and varying.

Moving forwards our variables of interest are the low_weight, EE, and ZM, with varying intercepts by Lab and Protocol, as well as varying slopes by protocol for ZM and EE.

## Model Selection

The model selection is based on two methods: AIC forward selection and ANOVA F-test. For a baseline, we start with a model fit with only EE and ZM and varying intercepts by Lab, and get a AIC of 5487. Fitting a model with all the variables EE, ZM, varying intercepts by Lab and Protocol, and varying slopes for EE and ZM by protocol gives a deviance of 3210. Fitting a final model with the variables EE, ZM, the low_weight, varying intercepts by Lab and Protocol, and varying slopes for EE and ZM by protocol gives a resulting deviance of 3201.

For each additional element added, we incrementally tested each element against the previous model with an ANOVA F-test. Using a 0.05 threshold, we confirm that varying intercepts by Protovol, varing slopes for EE and ZM by protocols and the low_weight variable significant in predicting the uterus weight.

## Final Model

$$log(Y_{uterus}) = (4.87+\gamma_{0\ lab}+\gamma_{0\ protocol})+(0.14+\gamma_{1\ protocol})*EE_{ij}+(-0.56+\gamma_{2\ protocol})*ZM_{ij}+(-1.211*LowWeight)+\epsilon_{ij}$$

## Model Interpretation

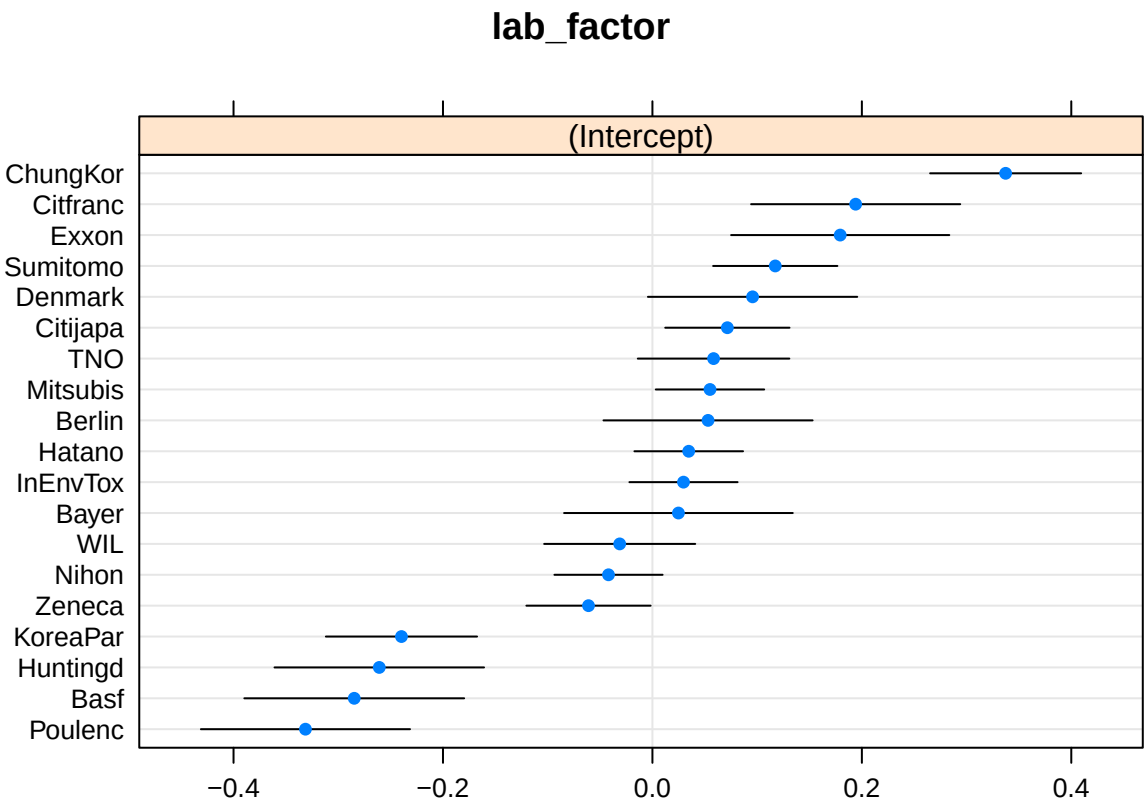
The cross-Lab variance in the intercepts for Uterus Weight is .0322 (standard deviation = .179), the cross-Protocol variance in the intercepts for Uterus Weight is .00469 (standard deviation = .0685). The cross-Protocol variance in slopes for Uterus Weight vs EE is .0784 (standard deviation = .28) and for Uterus Weight vs ZM is .000273 (standard deviation = .0165) for ZM. The residual inter-group variance is .185 (standard deviation = .43)

For any rats in high weight group, across all labs and protocol conditions, at the baseline of 0 EE and ZM, the average Uterus Weight is e4.873 = 130.32mg (p-value <.001). Holding all other variables constant, a one unit increase in EE is correlated with an 14% (e.1396 = 1.15) increase in Uterus Weight (p-value <.001).

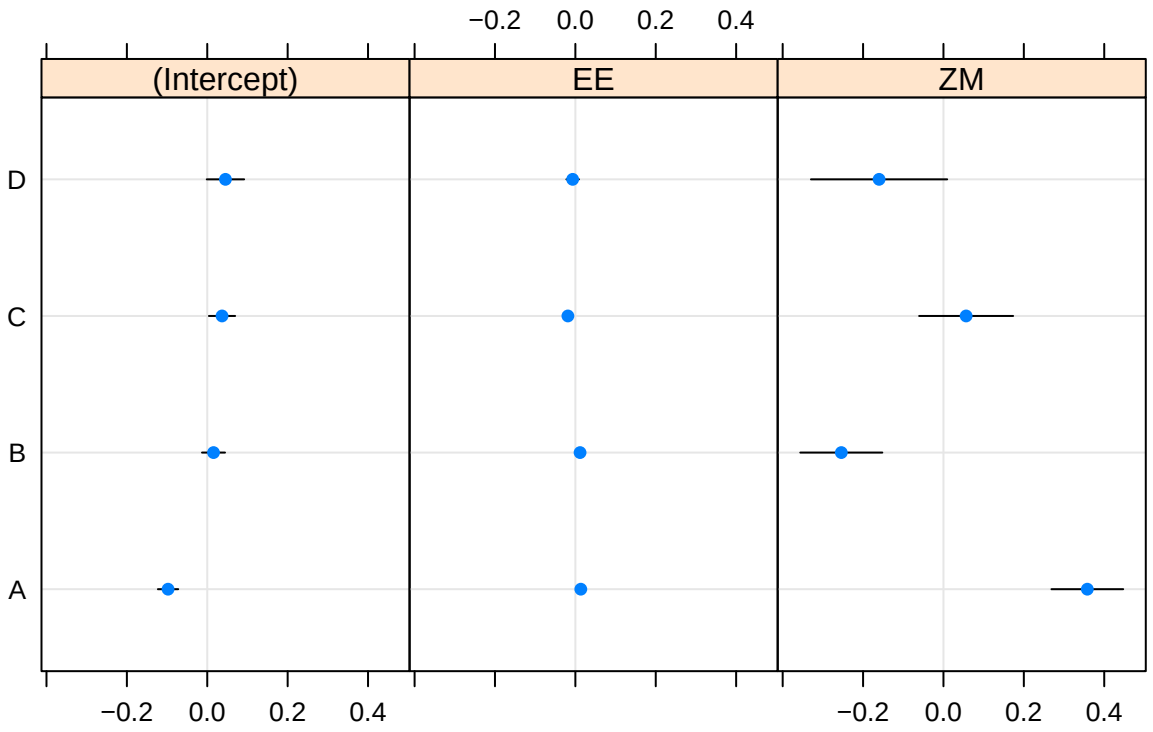
In the same situation, a one unit increase in ZM is correlated with a 43% (e.570 = 1.77) decrease in Uterus Weight (p-value <.001). Holding all other variables constant, rats in the lower weight class are correlated with a 70% (e-1.211 = 0.298) lower Uterus Weight than rats in the higher weight class (p-value <.001).

A dotplot of the random effects shows that the varying slopes for ZM by protocol are significant as all the confidence intervals in each group do not contain zero. The random slopes for EE by protocol do not appear to be significant as all the values lay at zero.

```
## $lab_factor
```




```
##
## $protocol_factor
```




## Model assessment

## Conclusion

The uterotrophic bioassay is successful in identifying the estrogenic effects of EE and anti-estrogenic effects of ZM. Increased levels of EE administered to rats is correlated with an increase in rat Uterus Weight while increased levels of ZM administered to rats is correlated with a decrease in rat Uterus Weight. All three of these effects were found to be significant at the .001 significance level.

The analysis also finds that while the effects (slope) of EE and ZM on Uterus Weight did not vary across labs, the baseline (intercept) Uterus Weight variance between labs accounts for roughly 3% of the variance in the all the response variable. The labs whose intercepts are significantly different from the average (fixed effect) are Poulenc, BASF, Huntingd, and Korea Par (low) and Sumitomo, Exxon, Citfranc and ChungKor (high).

The different protocols were found to differ in their sensitivity to detecting ZM effects but not EE effects. When including the random effects for ZM by protocol, protocols B and D have larger coefficients in the negative direction than protocols A and C. Since ZM is negatively correlated with our response, we can conclude that protocols B and D (by injection) are more sensitive in detecting the ZM effects, the most sensitive of which being protocol B – immature female rats dosed by injection (3 days).

### Limitations

PART II

# PART II

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
## embedded nul(s) found in input
```

## Introduction

The North Carolina State Board of Elections (NCSBE) is responsible for administering the election process and campaign finance disclosure and compliance. The NCSBE also provides voter registration and turnout data.

The purpose of this analysis is to examine the NCSBE voter data to come to a conclusion on the following questions: 1. How did demographic subgroups vote in 2016? For example, how did the turnout for males compare to the turnout for females after controlling for other potential predictors?
2. Did the overall probability or odds of voting differ by county in 2016? 3. How did the turnout rates differ between females and males for the different party affiliations?

### Key Findings

This analysis finds that voter turnout increases substantially as the age group of those measured increases. We also uncovered variation in voter turnout by race within political parties, with no party having equal turnout rates for all or even most racial demographics. A similar trend to what was uncovered about race was also uncovered in regard to gender, however, voter turnout for people not affiliated with a particular party was fairly consistent across both males and females. Also, it was found that voting odds did in fact differ from county to county. While these finding are significant to our analysis, there is also a significant amount of unexplained variance that prevents our model from tightly fitting out data.

## Data Pre-processing

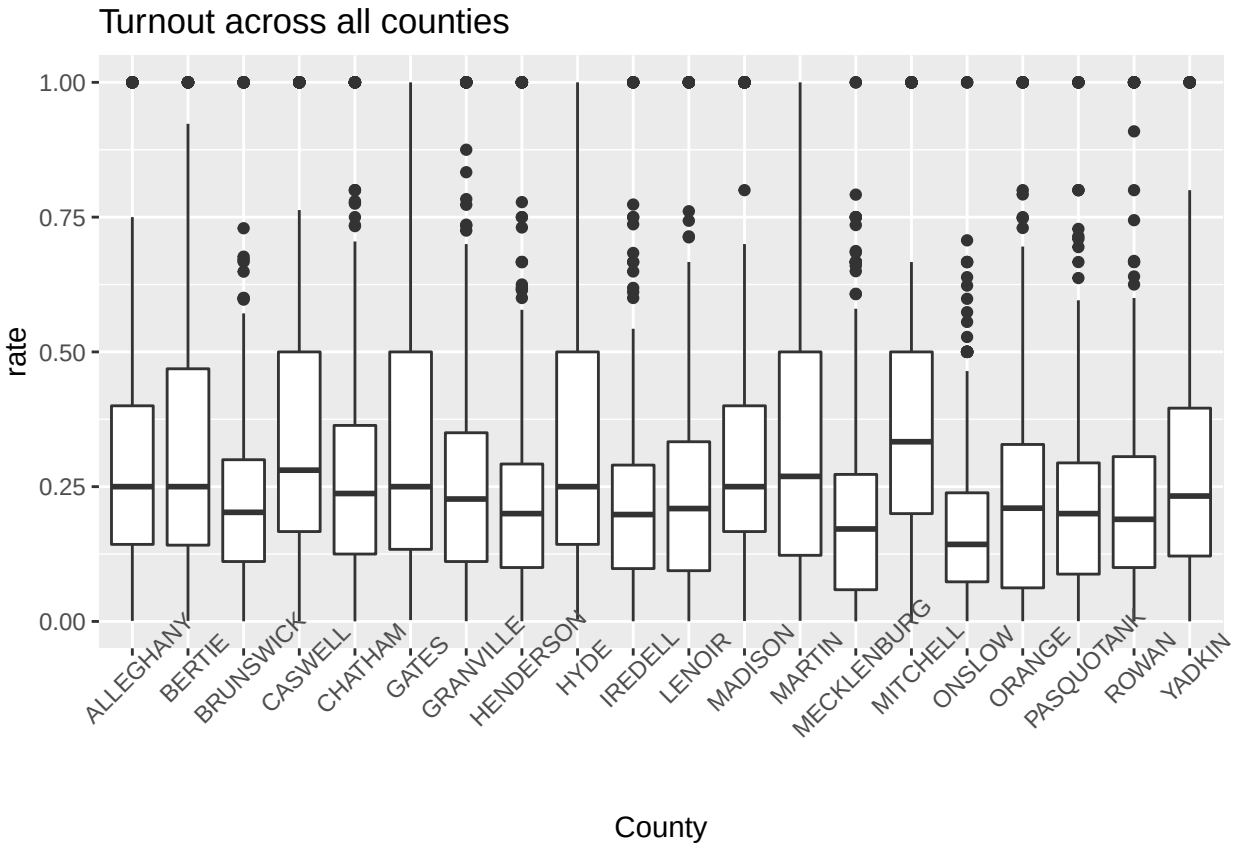
The two datasets used in this analysis were the 2016 voter history data and 2016 voter registration data. To prepare for the analysis of voting habits by demographic group the two datasets were grouped by county, age group, ethnicity, race, party, and gender (each row having a unique combination of the variables) and aggregated the number of actual voters or registered voters for the respective dataset.
To effectively join the two datasets along all the variable columns we first needed to ensure that the levels for the factor variables for each dataset were consistent. Invalid voter groups levels that did not match, like the age group below 18 years from the voting records dataset were removed. Missing values were also removed from the dataset as any methods to impute the missing values did not seem feasible. Also, the two datasets have a relatively small number of predictor variables compared to a very large number of observations that we were not concerned about the degree of information loss resulting from omitting observations with missing values.

The two cleaned datasets were then merged along all common predictor variable columns. Voter turnout for each row was calculated by dividing the number of actual voters by the number of registered voters. This voter turnout variable was then used as the response variable as we explored relationships in the data.

## EDA

Only boxplots were used to explore relationships as there are no continuous or discrete predictors in the dataset.

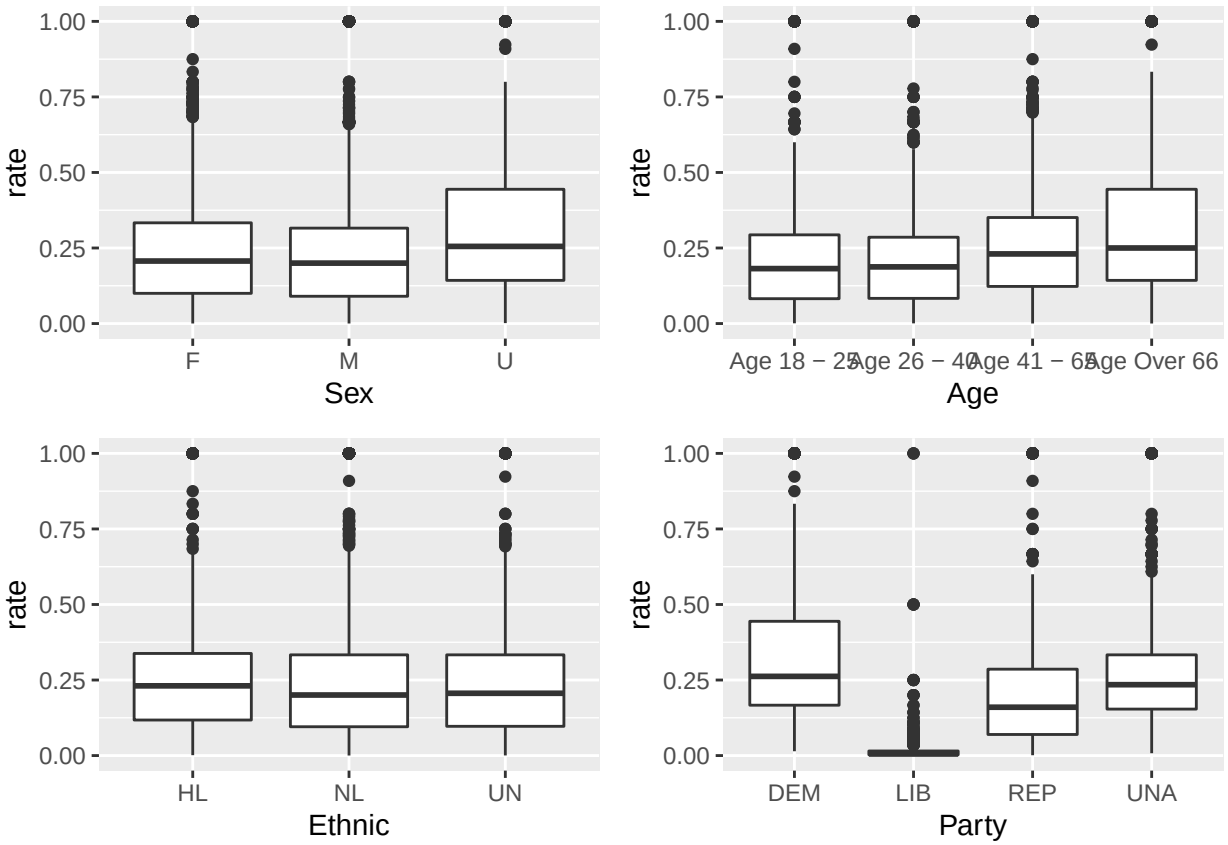
Voter turnout rate varies from one county to another. However, consistent across the entire dataset, and within parties and counties is the increasing trend in voter turnout as age increases.




Overall, turnout rate for males and females appears to be equal, and people who specified undesignated for sex had a higher turnout rate than both. However, when exploring how turnout rates for men and women differ by party, female democrats appear to have a significantly higher turnout compared to male democrats, while the turnout rates are identical among republicans. Further investigation is needed to determine any significance of a relationship between sex and political affiliation and voter turnout rate.

Exploring the relationship between race and turnout rate we found that indigenous Americans and mixed have the highest average turnout rate, followed by White and Asian, and Black have the lowest average turnout. Further exploration showed that this relationship was not consistent across all counties. Further investigation in needed to determine the significance of any relationship between race and county on voter turnout rate.

There also appears to be an apparent relationship between ethnicity (Hispanic vs non-Hispanic) and party on voter turnout rate that will be explored in the latter parts of this analysis.




There's a possible interaction effect between race and party since the turnout rate for each racial group changes according to party; for example, Black democrats have a significant higher turnout compared to Black republicans. Overall, registered democrats appear to have the highest turnout rate followed by unaffiliated voters, then by registered republicans, then lastly registered libertarians.

## Model Selection

The model selection is based on two methods: AIC forward selection and ANOVA chi-square test To determine a baseline for which to compare later models to we fit the following null model which yielded a deviance of 1,200,702:

$$Null\ model : age + sex + race + varying\ intercepts\ by\ county$$

To try and account for the variation in voter turnout by race across a counties, and the apparent interaction between race and political affiliation we fit the following model that yielded a deviance score of 180,080.1:

$$Intermediate\ model : age + sex + party + party : race + (race \mid county)$$

which was a significant improvement over out null model with a decrease in deviance of over 1,000,000. In an attempt to account for the variance in voter turnout seen between ethnic groups and between gender classifications within political parties we fit the following model which yielded a deviance of 164,559.6:

$$Final\ model : age + sex + ethnic + party + party : race + sex : party + (race \mid county)$$

which we referred to for the remainder of our analysis.

## Final Model

$$log(\frac{\pi}{1-\pi}) = (+\gamma_{0\ county})+(+\gamma_{1\ county})*Race_{ij}+()*Age_{ij}+()*Sex_{ij}+()*Ethic+()*Party+Party : Race+Sex : Party+\epsilon_{ij}$$

## Model Interpretation

## Model assessment

Assessment for this model consisted of an analysis of the binned residual plot. Typically, the criteria is that 95% of the observations are contained within the plot boundaries, and that the trends residuals do not appear to be following a trend depended on some other factor. Our plot does not meet these criteria, however, since we are analyzing voter turnout rates by controlling only for demographic group, we know that there will be other dependencies not accounted for in the analysis that will affect our results. As such we proceeded with the analysis. These limitations will be discussed in further details in a section of this document.

## Conclusion

To summarize our analysis, we will propose our answers to the questions posed at the start of this document.

The analysis finds that the younger demographic is less likely to vote people are less likely to vote than older people. When compared to the baseline, voting odds ratio increases by 13% when the age category changes from 18-25 to 26-40. The voting odds ratio compared to the baseline then increases by 50% when moving from 26-40 to 41-65, and 56% when moving from 26-40 to voters over 66. We also found that when compared to the baseline, the voting odds ratio decreases by 35% when gender is changed to male, and that people with an undesignated gender tend to have, on average when compared to the baseline, a voting odds ratio 15% lower than men.

Also, compared to the baseline, voting odds ratio decreases by 70% when the party changes from Democrat to Republican, however, when assuming the race is white, compared to the baseline, voting odds increases by 136% for when party changes from Democrat to Republican. This stands in stark contrast to the 93% decrease in voting odds for Black republicans when compared to the baseline. A similar (but reversed) trend is also seen within the Democratic Party, highlighting the incredible amount of variance in voting habits between races seen within political parties.

The voting odds ratio differed by county. Our analysis shows that the deviation in voting odds within counties does not fully account for the deviation in voting odds seen between counties. Roughly half of the counties we sampled had voting odds below the average voting odds for all counties, and half of the remaining counties had voting odds that fell above the average voting odds for all counties.

Similar to what was seen with race, voter turnout rates differ between males and females for the different party affiliations. Male republicans have a voting odds ratio nearly 145% higher than that of female republicans when comparing to the baseline, and while male libertarians have a voting odds ratio that is 250% higher than females of the same party when comparing to the baseline. However, these wide variances in voting odds do not hold for unaffiliated voters, where males only have a voting odds ratio that is 35% higher than females with the same affiliation when compared to the baseline.

## Limitations

This analysis focuses entirely on how very specific aspects of a voter's demographic effects the likelihood of them turning out to vote. As such, there are is a significant amount variance that we would expect to be left unaccounted for due to things like income, religion, and education. This variance left unaccounted for is seen in the dependent nature of the binned residual plot for out model.

Another limitation of this analysis is with regards to its usefulness in making decisions for future elections. This election season (2020) has very immensely different to 2016 and seeing as demographics only told part of the story in 2016 that fact could be even more skewed this year. The results of this analysis alone should not be taken as complete fact for what to expect in upcoming elections. One way to strengthen them are to look at how voting trends have historically changed between presidential and non-presidential elections,

and use that information along with information about the 2018 election to better understand what might happen in 2020 (and beyond).