Estrogen Bioassay

Akshay Punwatkar, Andrew Patterson, Derek Wales, Xinwen Li,Tzu-Chun Hsieh

November 4, 2019

Summary

Analysis of a bioassay examining a possible estrogen agonist and antagonist. The agonist, Ethinylestradiol or EE, was confirmed in this analysis to have estrogen-like effects on the weight of the rat uterus. Additionally, ZM, a potential estrogen antagonist, was shown to possibly reduce the estrogenic effects of EE. However, without proper controls in the study, ZM may also operate on another mechanism of action than the one proposed here.

Introduction

Estrogen is an important hormone in mammals which controls numerous primary and secondary sex characteristics in the organism. Estrogen agonists and antagonists are classes of compounds which act on the estrogen receptor in place of estrogen to either activate or inhibit the receptor. In this study, several experiments were carried out to determine if the potential estrogen agonist Ethinylestradiol, or EE, had similar effects to estrogen on estrogen-free female rat uteruses. Additionally, the effects of a potential estrogen antagonist, ZM, were also examined in the study. Data from this project was then used to construct a model to determine if EE and ZM had estrogen agonistic or antagonistic effects on the rat uterus.

Data

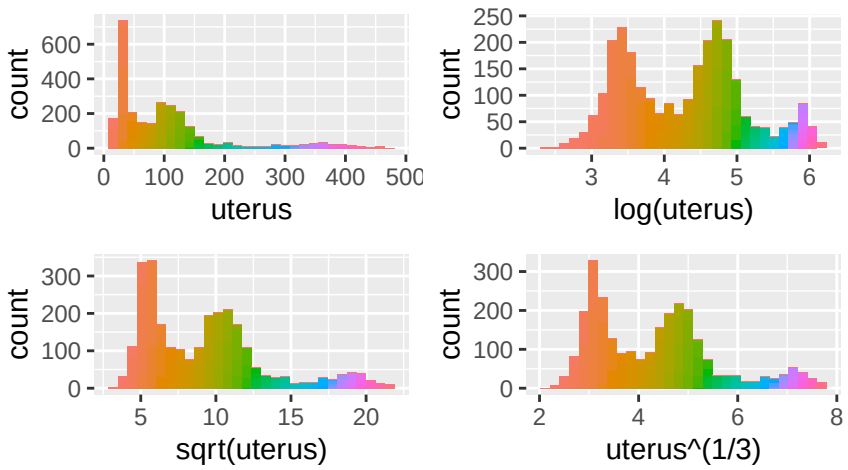
Data used in this analysis contained 2681 observations from different labs conducting research on whether the estrogen level would affect the uterus weights of rats. A cursory investigation of the data revealed there were 4 rows missing uterus weight values and 2 of rows missing weight values. Since these missing rows occurred randomly, which means these rows are in different groups, protocol types, or labs. Therefore, these rows were deleted.

For the variables in the data set, uterus, weight, EE, ZM were treated as numeric variables and protocal, lab, group were treated as categorical variables. In the research, there are only 3 kinds of dosage of ZM and 7 kinds of dosage of EE, however they were still treated as numerical variables because if treated as categorical variables, information would be lost between different dosages. A different dosage isn't just different from another dosage. For example, a 10 mg dose is 10 times a 1 mg dose. This information would be lost in a categorical variable. Another variable was added to the dataset, a new binary variable- mature, based on the value of protocol, to indicate whether the rats were mature. If the rat was categorized as protocol A or B, it would have value 0, and if the rat was categorized as protocol C or D, it would have value 1.

EDA

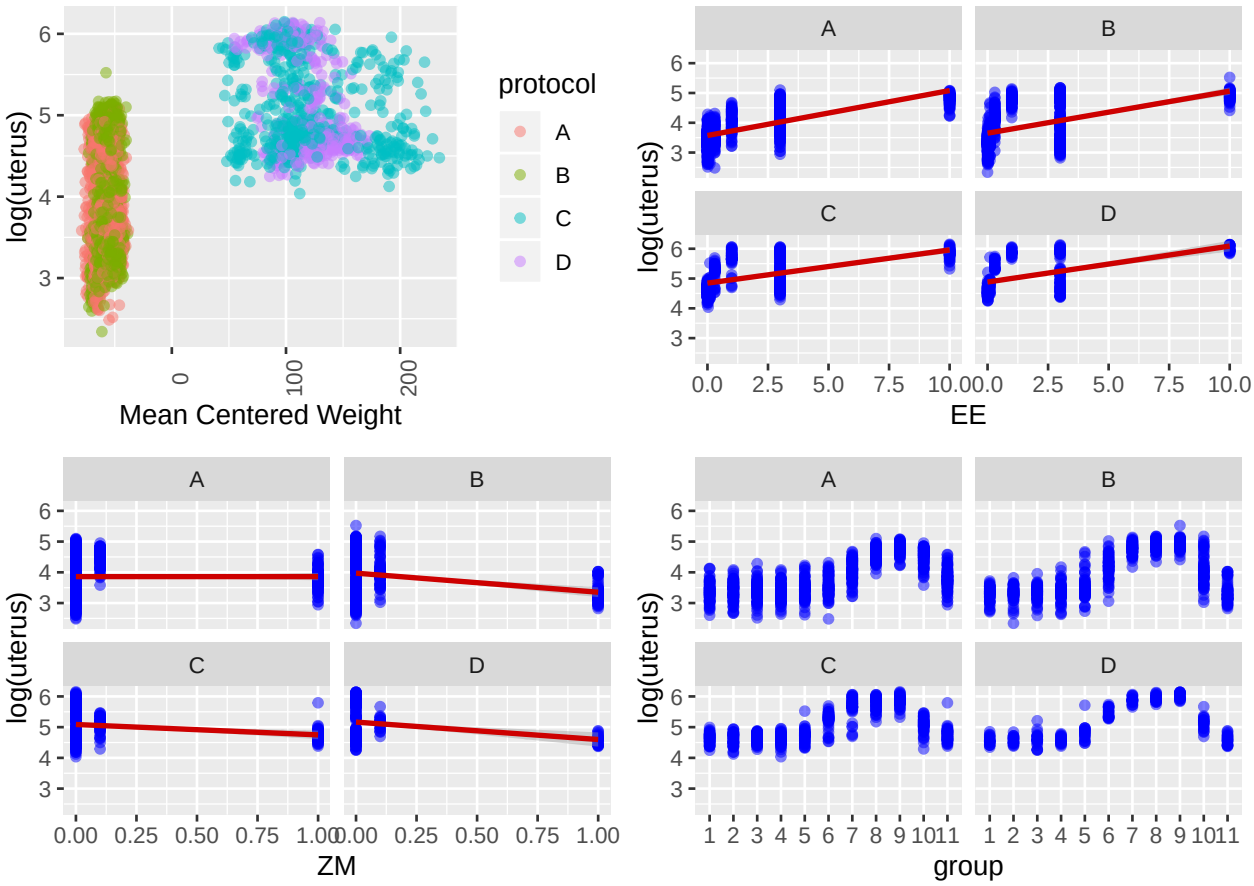
For the data set used, collinearity was investigated as the data was explored. Mature versus Protocol: Since mature is the new variable created based on protocol types, there is high collinearity between them. ZM and EE versus Group: Since groups is seperated depends on the different combination of dosage of ZM and EE, they are highly correlated with each other.

First, a histogram was plotted of the response variable- uterus weight. However, the distribution is skewed and is not a normal distribution. Out of the transformations square root, cube root, and log transformation, it seems that log transformation improves the distribution the most. As a result, even though it is still hard to say that the distribution of log uterus weight is a normal distribution, in the following analysis, log uterus weight was used as the response variable.



Moreover, plots were used to check the relatonship between log uterus weight with other variables. First, the difference of log uterus weight between labs was examined. By the distribution of log uterus weight of each lab, there appeared to be some differences. However, the data is further separated by different protocol types, in each protocol type group, each lab has a similar distribution of log uterus weight. This difference observed is caused by the fact that not all labs conduct experiment for every protocol type. Moreover, an apparent pattern seemed to indicate that the data points were clustered by group when plotting for log uterus weight and mean centered weight. There are four clusters in the plot. For the relation between log uterus rate and protocol, rates categorized as protocols C or D apparently have higher log uterus weight comparing to protocols A or B.

All these observations indicate that different protocol types would have different log uterus distributions. Therefore, a plot was made to examine the relationship between log uterus weight with each variable by different protocol types. This indicated that when comparing log uterus weight and mean centered weight, the mature rats have a negative pattern, while the immature rats do not appear to have a pattern. For the different dosages of ZM and EE, there appears to be a positive relationship between log uterus weight and ZM and a negative relationship between log uterus weight and EE. Lastly, it appeared that different groups of rats had different distribution of log uterus weights. However, since grouping is based on the dosage of ZM and EE and the experiment is concerned with studying the effect of ZM and EE, only ZM and EE will be included in the final model and the group variable will be excluded.



Model Selection Process

Based on the EDA, including lab and protocols as level predictors, EE, ZM, mean centered weight as numeric predictors, the first model constructed:

$$Log(Uterus) = (B_0 + \gamma_{lab} + \eta_{protocal}) + B_1 EE + B_2 ZM + B_3 MWeight + \varepsilon_{ijk}$$

The AIC for this model was 3314, and BIC was 3356. It was expected that both scores would be smaller for the final model. Mean centered weight was not very significant with a -0.72 t-value but the protocal variable does contain information of the weight of the rat and weight cannot help distinguish the difference between mature and inmature rats here. Instead of including the weight, another binary variable was created indicating either mature or immature rats. Because protocols A and B were done on immature rats, and protocls C and D were done on mature rats, the mature variable could help identify this difference between the two sets of protocols. Replacing weight with mature, The mature variable had a t-value of 40.33, and therefore seems to be a very significant predictor of uterus weight. However the normality issue was still not solved. There does not appear to be a multicollinearity issue in the model by checking the VIF scores, but the normality assumption seems to be violated from the qq-plot. Ignoring normality at this time, random slopes based on the protocols were created to test the sensitivity of different protocols detecting the effects ZM or EE. However, the model had a convergence issue. Due to this issue, protocol was instead included as a normal predictor instead of a level predictor. Interaction terms between protocol and EE and protocol and ZM were added to see the sensivity of the protocols in detecting ZM or EE as compared to the base protocol A. Weight was then added back into the model in hopes of distinguishing between the mature vs immature rats. However, mean centered weight was not very significant in the model and also caused a multicollinearity issue between weight and protocol. Thus, Weight was not included in the final model. The group variable was also added into the model as an extra predictor variable in the model once. This appeared to fix the normality and equal variance issues. However, by adding the group variable, the model loses information on

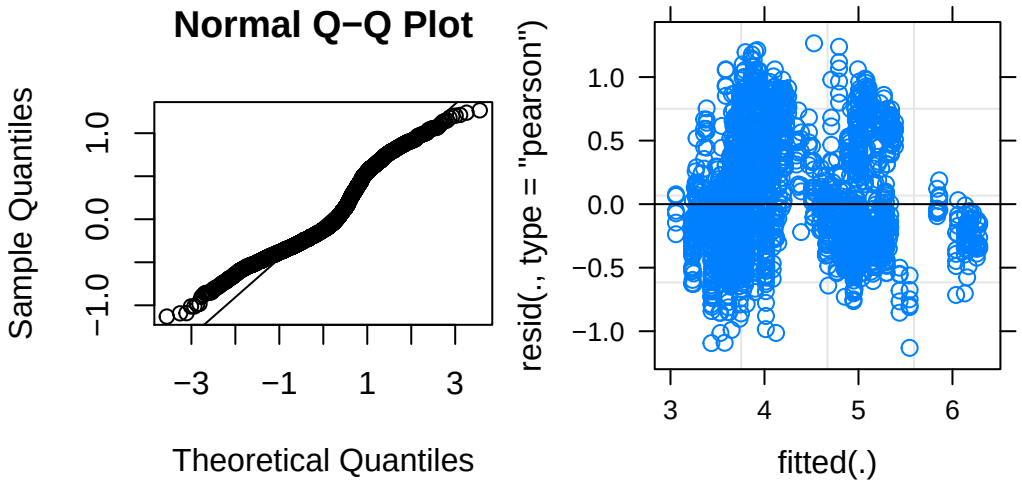
EE and ZM and the interpretation would not make sense to say different groups of rats can have different impacts. Adding group isn't a great idea here.

Model

The final model:

$$Log(Uterus) = (B_0 + \gamma_{lab}) + B_1 EE + B_2 ZM + B_3 Protocal + B_4 Protocol * EE + B_5 Protocol * ZM + \varepsilon_{ijk}$$

Model assumptions were examined by plotting residual vs fitted and qq-plot. As mentioned in the model selection process, the model does not satisify the normality assumption. There appears to be clustering in the residual plot, because the model cannot distinguish mature vs inmature rats, while Equal variance also appear to be violated.



In addition, multicollinearity was examined, and the VIF scores for each predictor in our final model were all between 2.6 to 5.6, indicating some multicollinearity but not enough to cause an issue for our interpretation. This is the best model considering all the questions asked of this data. The AIC score for this model is 3228 and BIC is 3310, which is lower than the first model above.

Table 1: Data Dictionary

Variable	Estimates (exp)	t-value
Intercept	35.202	79.785
EE	1.166	32.754
ZM	0.815	-4.366
Protocol B	1.123	4.255
Protocol C	3.878	44.457
Protocol D	3.887	34.308
EE: Protocol B	0.999	-0.126
EE: Protocol C	1.036	-4.545
EE: Protocol D	0.977	-2.186
ZM: Protocol B	0.531	-8.835
ZM: Protocol C	0.741	-3.830
ZM: Protocol D	0.580	-5.200

Results

There are three independent variables, a single hierarchical variable, and two interaction terms included in the model which are significant at predicting the weight of the rat uterus.

EE is a significant predictor of the weight of the uterus; as one unit of EE increases, the uterus weight will increase by a multiplicative effect of 1.17.The absolute t-value for this effect is 32.74, indicating this effect is significant at predicting uterus weight. This indicates that treating an estrogen free mouse with EE results in an increase in uterus weight.

For the independent variable ZM, as one unit of ZM increases, the uterus weight will decrease by a multiplicative effect of 0.81. The absolute t-value of this effect is 4.38, which is lower than EE but still significant at predicting the weight of the uterus. This indicates that the estrogen antagonist does have a negative effect on uterus weight. However, without proper controls as discussed in the conclusion, this study cannot conclusively say that this result is due to an estrogen antagonist effect.

For the random effects, each lab contained a different random intercept. The highest and lowest outliers were Chungkor and Poulenc labs, respectively at 0.337 and -0.330 intercepts.

The four protocols each had a significant effect on the weight of the uterus, with protocols C and D having an increased effect of as compared to A and B. Using protocol A as the baseline, protocol B had an multiplicative effect of on uterus weight of the rats by 1.12 with a t-value at 4.26. Protocols C and D had much higher effects, at 4.6 and 4.6 at t-values 44.46 and 34.31, respectively. This indicates protocols C and D had very significant effects on log uterus weight as compared to protocol A.

For the interaction terms, the interactions between the protocols and ZM and EE were significant. Of particular interest is the difference in the interaction term between ZM and protocol B versus the interaction term between EE and protocol B. The difference between these interaction terms, at 0.63, is larger than the other protocol interaction terms, indicating protocol B is the most sensitive protocol for determining the difference in effects between ZM and EE.

Conclusion

From these results, this study can conclude that EE has a positive effect on uterus weight while ZM has a negative effect on uterus weight. The different laboratories all have an effect on uterus weight, however this does not interfere with the overall conclusions of the study. Protocol B is the most sensitive protocol for detecting EE and ZM effects.

For further research, the point of this study was to see the effects of an estrogen agonist, EE, on uterus weight and to see the effects of a uterus antagonist, ZM, on uterus weight. From this study alone, ZM can be said to have an estrogen antagonistic effect. However, there were no controls for ZM as compared to EE. Without including a control ZM without EE, the actual mechanics of ZM on uterus weight cannot be determined. A further control group with ZM and without EE could determine if ZM is actually an estrogen antagonist, as expected, or is instead working by another mechanism on the uterus, such as a testosterone analog. The expectation of the estrogen antagonist is that, in the absence of all estrogen or estrogen analogs, there would be no decrease in uterus weight- only in the presence of EE would ZM show a reduction in uterus weight. If ZM were acting by another mechanism, then it would have a negative effect on uterus weight even in the absence of EE. Additional followups could investigate why there were differences in the laboratories' intercepts. A possible way to improve is to fit a mixture model to this dataset, since there is still bimodal pattern in the histogram of log(uterus).

Analysis of Voting in NC (2016 General Elections)

Akshay Punwatkar, Melody(Xinwen) Li, Derek Wales, Andrew Patterson, Tzu-Chun (Angela)

11/04/2019

SUMMARY

Analysis and modeling of the voter registration and participation data from the 2016 US General election for the state of North Carolina were performed. And the effects of the demographics on the voter turnout were analyzed and quantified. The analysis was preceded by several data processing steps and was performed on a subset of the data for 20 Counties within North Carolina. Voter turnout based on different genders, ethnicity, race, age groups, and party affiliations were analyzed. A Multilevel/Hierarchical logistic regression model used to quantify the effects of demographics on voter turnout.

INTRODUCTION

United States General Elections in 2016 was one of the most anticipated election. The election saw an average turnout of about 63%, which was reported as the lowest turnout in the past 20 years. Voter turnout is assumed to be affected by several demographic factors such as gender, age, race, ethnicity, county, and party affiliation. Other factors, such as the election campaign, accessibility to voting booths, and many more, also affects the turnout. The following analysis is aimed at analyzing and quantifying the effects of such demographic factors in determining the voter turnout. The scope of the analysis is limited to the state of North Carolina.

The North Carolina State Board of Elections (NCSBE) is the agency charged with the administration and the election process and campaign finance disclosure and compliance for the state. They are also required to keep extensive records to ensure electoral compliance, as part of their duties, they also keep information on likely voters and registered voters. Using the data obtained from NCSBE, the analysis was done primarily to gain insights about regarding a few questions :

- How did demographic subgroups vote in 2016?
- Did the overall probability or odds of voting differ by county in 2016?
- How did the turnout rates differ between females and males for the different party affiliations?

DATA

Dataset obtained from NCSBE contained demographic related information about the registered voters and voters who voted from 102 counties in North Carolina. Since the data was provided in two parts, extensive pre-processing of the data was performed. Overall, voter turnout as per the provided data was ~ 72.0 %, and it was made sure throughout the data processing process that the turnout occurs in a similar range.

Data Processing and Transformation :

The registered voter dataset, initially containing duplicates and null observations were processed to obtain the unique observations without any null values (0.2 % observations were null). Similarly, the voted voters dataset was processed to remove any null values (4.0 % observations were null). Subsequently, both the datasets were merged based on the demographics.

However, due to the variation in the methods used for voting or change in party affiliation, the voted dataset had multiple observations for the same demographics as in the registered voter's dataset. In order to eliminate redundancy of the total registered voter post merging, data were aggregated based on the demographics and total voters. In the process, few of the features, such as the voting method and voting method description, were dropped. Few more

- Since only a fraction of the population changed their party affiliations during voting, **party_cd** (original party affiliation) was used in the final dataset instead of **voted_party_cd**. Also, **voted_party_cd** was leading to duplication of total_registered voters.
- Few observations (0.84%) had **more voted voters than the registered voters** for a demographic. It could be explained by assuming that the voters might have changed their county/precinct and hence having the same demographic voted under different precincts. For such cases, the count of total registered voters was increased to match the voted voters, because eliminating such observations would have led to information loss for the entire demographic.
- **Precinct** and **voter district** was dropped from the final dataset. And the data was again aggregated based on the demographics.

Data Description :

The data dictionary used as part of the final analysis is as follows:

county_desc - *Name of the County of the voters belonging to a demographic group*
age - *Age group of the voters belonging to a demographic group*
race - *Race of the voters belonging to a demographic group*
ethnicity - *Enthnicity of the voters belonging to a demographic group*
sex_code - *Gender of the voters belonging to a demographic group*
party_cd - *Party affiliation of the voters belonging to a demographic group*
total_voters - *Number of registered voters belonging to a demographic group*
voted_voters - *Number of voters who voted belonging to a demographic group*

Features not used in the final dataset are not mentioned in the dictionary

Data Selection :

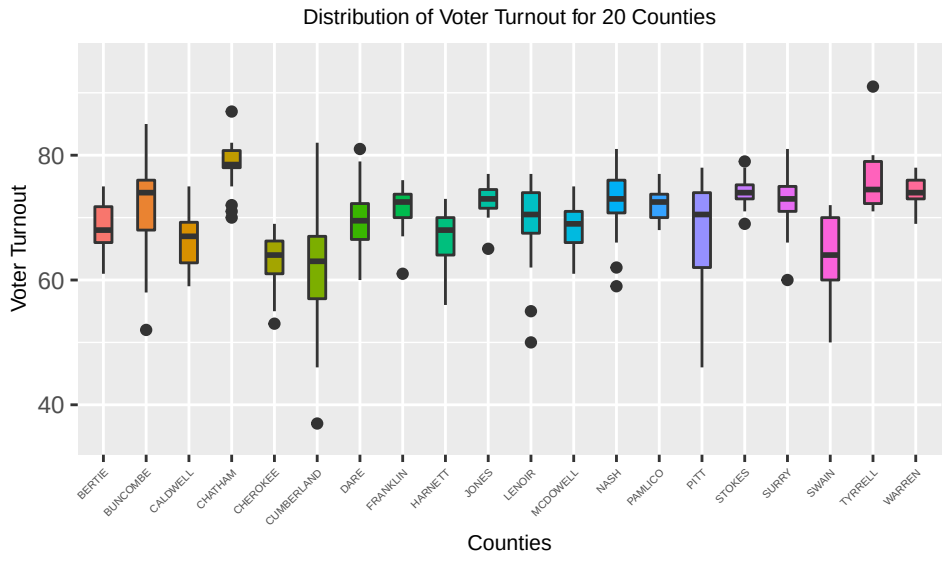
Voting data about 20 Counties was randomly selected from the primary dataset for the final analysis.

County (1-5)	County(6-10)	County (10-15)	County (15-20)
BERTIE	CALDWELL	FRANKLIN	DARE
HARNETT	JONES	CHATHAM	SWAIN
PAMLICO	BUNCOMBE	MCDOWELL	MCDOWELL
STOKES	PITT	NASH	SURRY
WARREN	TYRRELL	LENOIR	CUMBERLAND

Data Analysis :

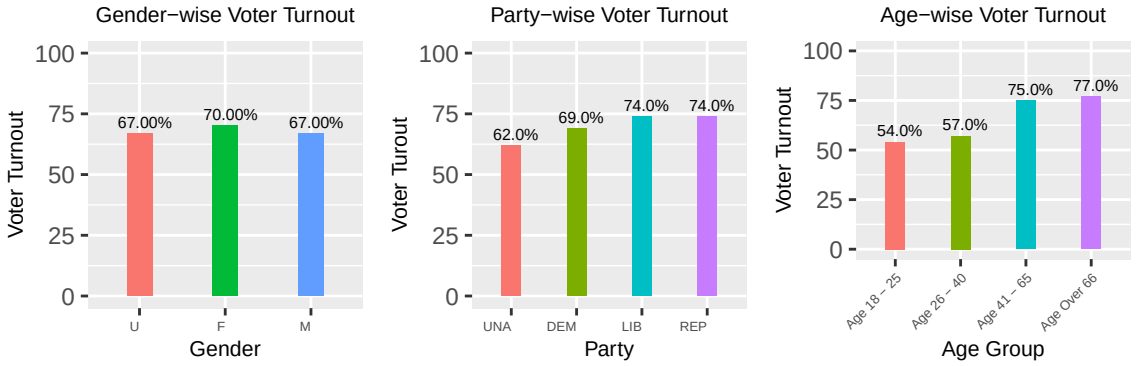
Initial analysis of data using visualization provides several keys insights about the variation of voter turnout among and within the counties. It also highlighted several relationships among gender, race, party, and counties. Following are the key observations from the analysis :

- Although **average voter turnout** for every **county** was in the **similar** range of ~70%, **distribution** of voter turnout within the counties didn't appear to be that similar.

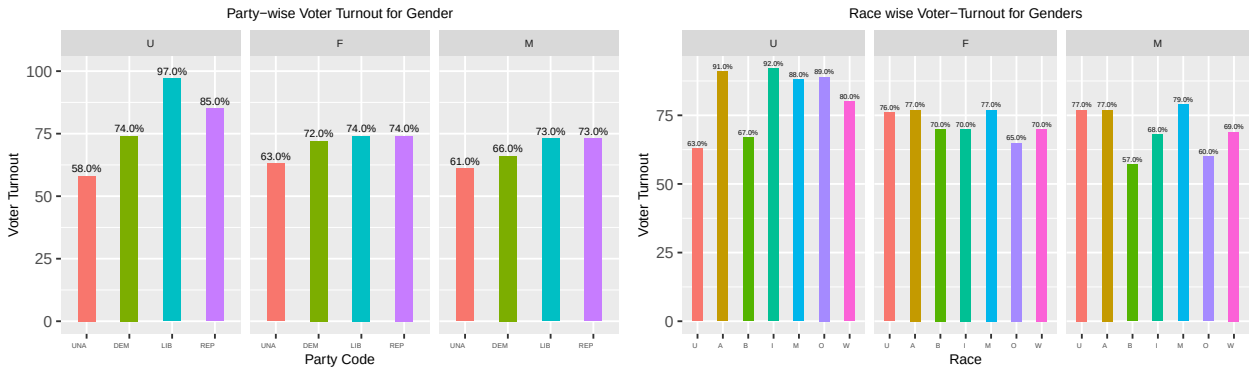


- **Female** voters had **3%** more turnout as compared to male counterparts.
- **Republican** and **Libertarian** voters had the highest turnout, 5% more as compared to Democrats.
- Only **half** of the registered voters **aged between 18-40** showed up for voting.

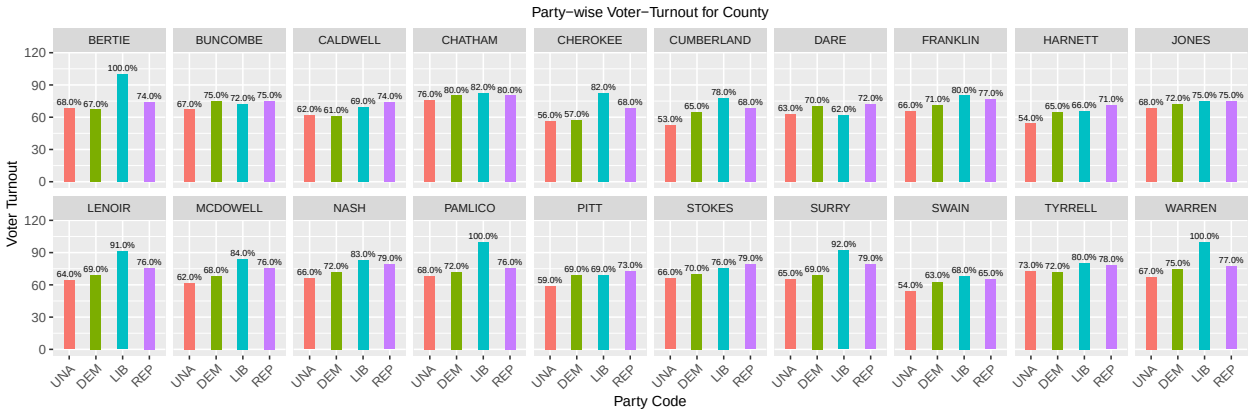
.



- **Female Democrats** had **8%** more turnout than the male democrats
- **Black-Female** had **13%** more turnout than their male counterparts.



- Interestingly several counties had **100%** voter turnout for **Libertarian** party which is much higher than the overall average turnout of ~ 70%.



MODEL

Post the initial analysis, to quantify the effects of the demographic variables, a logistic regression model was used. In addition, as discussed earlier, although all the counties had similar average voter turnout in the range of ~70 %, inter-county distributions were different. To capture this in-county variance in voter turnout, along with the information of overall voter turnout, a **multi-level** hierarchical model (random intercept for counties in this case) was used to quantify the demographic effects.

A series of modeling attempts, using *county* as a random intercept and rest of demographics as fixed effects along with few interactions between gender, age, party, and the race was made and tested using ANOVA. However, since voter turnout is primarily a function of the demographics, each model highlighted high significance towards all the fixed demographic variables. Moreover, with the increment of a number of interactions, effects were getting distributed and hence diminished among interactions.

Finally, a model (with lowest AIC and variation) using all the demographic features (age, gender, race, ethnicity, and party affiliations) along with an interaction between gender and party affiliations were selected.

$$log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = \hat{\beta}_0 + \gamma_{0j[i]}^{County} + \hat{\beta}_g G_i + \hat{\beta}_r R_i + \hat{\beta}_E E_i + \hat{\beta}_P P_i + \hat{\beta}_A A_i + \sum_{k=2}^{K} \hat{\beta}_6 G_{ik} : P_i$$

Where: **G** = *Gender*, **R** = *Race Code*, **E** = *Ethnicity code*, **P** = *Party Code*, **A** = *Age group*

RESULT

Quantifying the demographics effects provided several key insights :

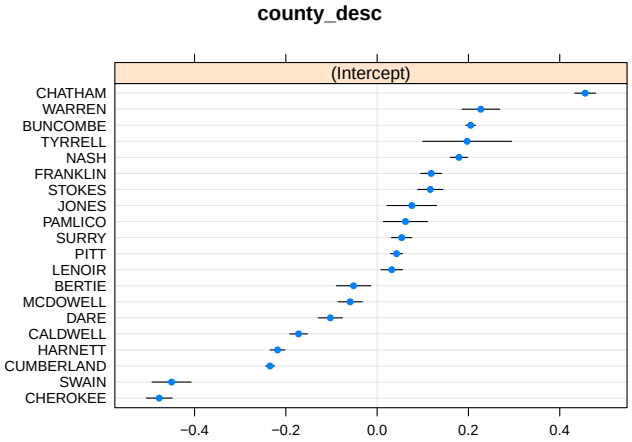
- The base odds of turnout is **1.09**. **Chatham** county had the highest baseline odds of turnout at **1.54**, while **Swain** and **Cherokee** had the lowest baseline odds of turnout at **0.5**

Given control of other potential predictors, **INCREASE** in Odds of Turnout being a :

- **Female** voters (1.14) is **13%** more the **Male** voters (1.01).
- **Libertarians** was surprisingly high, which could be explained by the very high voter turnout for the Libertarian party in several counties, as discussed during the analysis.
- **Republican** voters (3.82) is **117%** more than the **Democrats** voters (2.05).
- Voter in **age groups above 40 years** is nearly **150%** more than the lower age groups.
- **Hispanic or Latino** voter is nearly **30%** more than **Non-Hispanic or Latino** voters.

Controlling all other potential variables and varying only the gender and party, the prediction was made to quantify the effects of gender w.r.t to party. Below were the observations:

- Odds of Voter turnout being a **Female Democrats** exceeded **Male Democrats** by a factor of **0.65**.
- However, Odds of Voter turnout being a **Female Republic** exceeded **Male Republic** by only a factor of **0.05**.
- Moreover, Odds of Voter turnout being a **Male Republic** exceeded **Male Democrats** by a factor **0.93**.
- Moreover, Odds of Voter turnout being a **Female Republic** exceeded **Female Democrats** by a factor **0.35**.



Beta	Exp Val	Beta.	Exp Val.	Beta..	Exp Val..	Beta...	Exp Val...
Intercept	1.09	Age 26-40	1.16	>2 Races	1.91	F Dem	0.64
Female	1.14	Age 41-65	2.55	Other Race	0.69	M Dem	0.54
Male	1.01	Age over 66	2.85	White	0.76	F Lib	0.08
Democrats	2.05	Asian	1.44	Hispanic or Latino	1.36	M Lib	0.08
Libertarians	25.59	Black or AA	0.73	Non Hispanic or Latino	1.07	F Rep	0.39
Republican	3.82	AI or AN	0.99			M Rep	0.43

The above table contains the exponentiated estimates for the variables.

- The Model generated an **in-sample of accuracy** of **95%**.
- Based on the dot-plot highlighting the 95% confidence interval for the random effects of counties, a multi-level (random intercept) model seemed a good fit.

CONCLUSION

The analysis provided several important highlights about the demographics features such as County, Gender, age, race, and ethnicity, which seemed to affect voter turnout in the state of NC, among other variables. Quantifying the effect of these features using a multi-level logistic regression model provided an idea of the extent of effects of the demographics on the voter turnout. However, given more information about the polling stations and election campaigns along with the demographics, it could have assisted in better analysis. Moreover, since all the features were categorical, the model was highly susceptible to overfitting, and due to the absence of test data, testing model performance could not be done. Also, the interpretation of the model to quantify the effects of interactions could not be carried out directly. Prediction using controlled data was performed to quantify these effects. To summarize, an extensive analysis needs to done using more demographic features in the data, as mentioned earlier, and a better modeling algorithm to analyze and model the effects leading to voter turnout.