Part I: Estrogen Bioassay

Altamash Rafiq, Christy Hu, Ronald Nhondova, Srishti Saha, Tien Yu Liu

Introduction

At present, it is widely known that environmental estrogens may disrupt body functions, including the growth of bones, skin, and other organs and tissues. On this matter, the rat uterotrophic bioassay data set, provided by Dr. Akande from IDS 702: Modeling and Representation of Data, is utilized in order to re-investigate this issue and use a multi-level model to answer the questions of interest shown below:

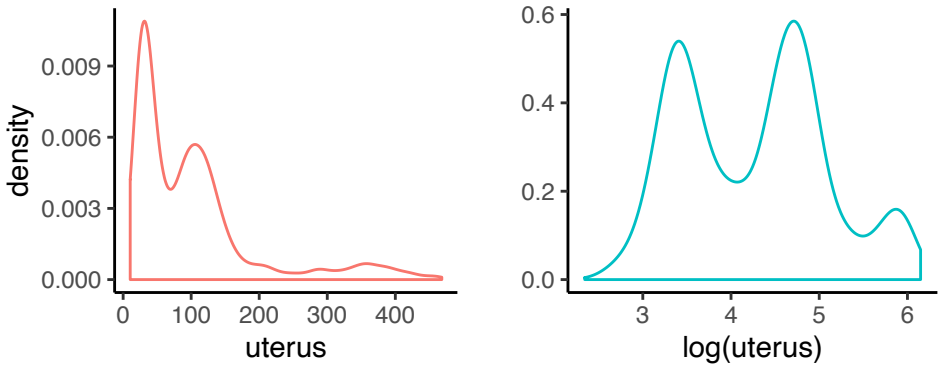
- Is the uterotrophic bioassay successful at identifying estrogenic effects of EE and anti-estrogenic effects of ZM. That is, after controlling for predictors and random effects, does uterus weight exhibit an increasing dose response trend for EE and a decreasing dose response trend for ZM?
- Does the dose response vary across labs? If so, are there certain labs that appear to be outliers?
- Do the protocols differ in their sensitivity to detecting EE and ZM effects? If so, is there one protocol that can be recommended?

Data

The data set, *bioassay*, contains seven variables (*protocol*, *uterus*, *weight*, *EE*, *ZM*, *lab* and *group*) and 2681 observations. Within the data set, 4 missing values are removed since those missing values seem to be randomly distrubuted. In addition, the variable *group* is excluded for it is an indicator for the lab settings (combinations of EE and ZM). Moreover, the variable *weight* is centered (*weight_c*) to improve potential interpretability.

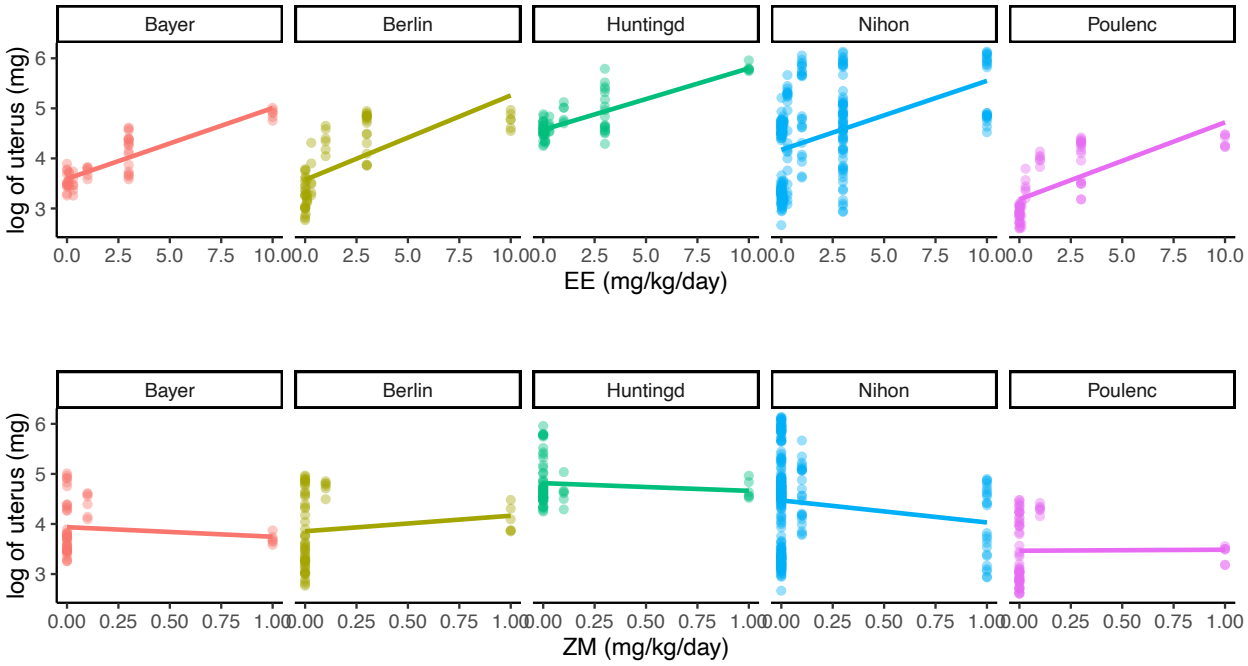
Exploratory Data Analysis

The distributions for *uterus* and the log of *uterus* are illustrated below. Although it seems that the transformation does not alleviate the violation of the normality assumption, it resolves the skewness issue in *uterus*. Therefore, the log of *uterus* is used as the response variable in this analysis.

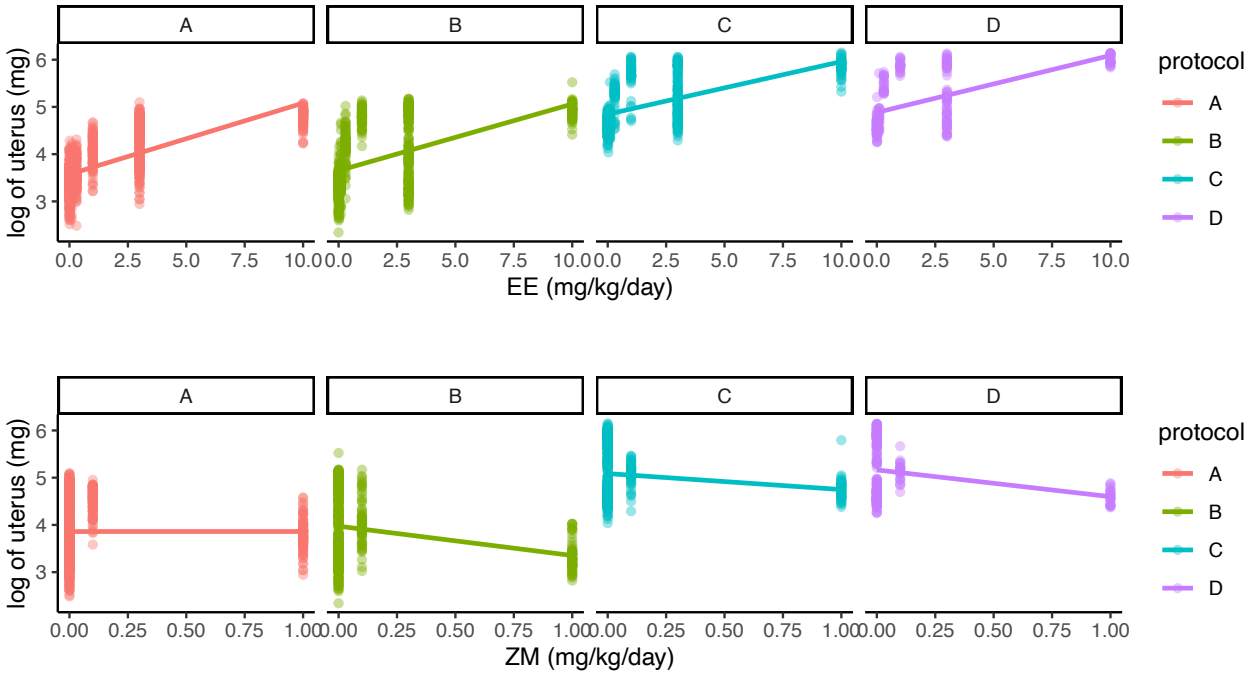


In order to study the relationships between the log of *uterus* of and the predictor *EE* and *ZM*, and whether these relationships vary in different labs, five labs are randomly selected: Poulenc, Nihon, Berlin, Huntingd

and Bayer. Based on the scatter plots illustrated below, in general, the increase of log of *uterus* is associated to the increase of *EE* and decrease of *ZM*. However, lab Berlin and Poulenc seem to deviate from the other labs in ZM. Therefore, further investigation needs to be conducted.



Considering that the relationships between the log of *uterus* and predictor *EE* and *ZM* might vary in *protocol*, their corresponding scatter plots are generated. These plots show that those relationships seem to be consistent under different laboratory settings; nevertheless, the slope for *ZM* under protocal B suggests that protocal B may potentially produce a different result compared to other protocols.



Model

In order to answer the aforementioned questions of interest, a varying-intercept model is constructed: the fixed effect part of this mixed effects model contains the main effect variable *EE*, *ZM*, *protocol*, and the interaction terms of *protocol* and the variable *EE* and *ZM*; Regarding the random effect part, since it is difficult to determine the hierarchical structure of *protocol*, and the issue of small sample size does not present, only the varying intercept for *lab* is included. In addition, *weight_c* is excluded to build the final model based on the insignificant p-value, 0.072, from the likelihood ratio test.

Final model:

$$\text{uterus}_{i\ell} = \beta_{0\ell} + \beta_1 + \beta_2 \times \text{EE} + \beta_3 \times \text{ZM} + \beta_{4p} \times \text{protocol}_p + \beta_{5p} \times \text{EE} \times \text{protocol}_p + \beta_{6p} \times \text{ZM} \times \text{protocol}_p + \epsilon_{i\ell}$$

$i = 1, \ldots, 2677$, $\ell = 1, \ldots, 19$ and $p =$ A, B, C, D

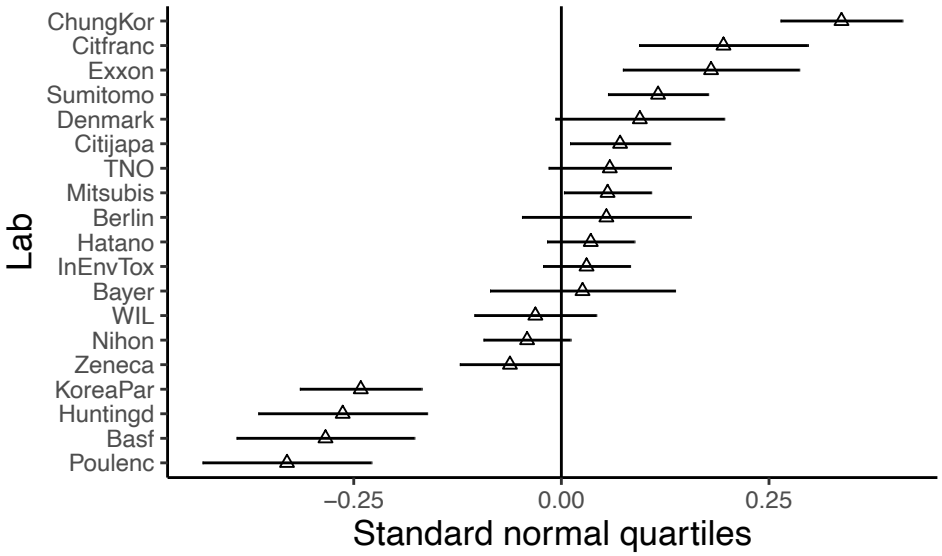
$\epsilon_{i\ell} \sim N(0, \sigma^2)$ and $\beta_{0\ell} \sim N(\beta_0, \tau_0^2)$

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
final_model	14	3159	3242	-1566	3131	NA	NA	NA
final_model_with_weight_c	15	3158	3246	-1564	3128	3.23	1	0.072

Fixed Effects						
	Est.	2.5%	97.5%	t val.	d.f.	p
(Intercept)	3.56111	3.47363	3.64859	79.78457	21.40672	0.00000
EE	0.15388	0.14467	0.16309	32.75431	2647.03349	0.00000
ZM	-0.20442	-0.29617	-0.11266	-4.36637	2646.56216	0.00001
protocolB	0.11565	0.06238	0.16891	4.25508	2623.96868	0.00002
protocolC	1.35544	1.29568	1.41520	44.45691	2591.63280	0.00000
protocolD	1.35747	1.27992	1.43502	34.30777	2662.86101	0.00000
EE:protocolB	-0.00091	-0.01497	0.01316	-0.12636	2646.70575	0.89946
EE:protocolC	-0.03547	-0.05077	-0.02017	-4.54454	2646.66248	0.00001
EE:protocolD	-0.02288	-0.04340	-0.00237	-2.18599	2646.56791	0.02890
ZM:protocolB	-0.63161	-0.77173	-0.49150	-8.83526	2646.49767	0.00000
ZM:protocolC	-0.29868	-0.45152	-0.14585	-3.83034	2646.48985	0.00013
ZM:protocolD	-0.54397	-0.74901	-0.33893	-5.19982	2646.47183	0.00000

p values calculated using Satterthwaite d.f.

Random Effects		
Group	Parameter	Std. Dev.
lab	(Intercept)	0.18000
Residual		0.43000



Conclusions and Limitations

Based on the summary for the final model, the questions of interest may be answered:

- Is the uterotrophic bioassay successful at identifying estrogenic effects of *EE* and anti-estrogenic effects of *ZM*. That is, after controlling for predictors and random effects, does uterus weight exhibit an increasing dose response trend for *EE* and a decreasing dose response trend for *ZM*?
  - Yes, the uterotrophic bioassay successfully identify the effects of *EE* and *ZM*. After controlling the predictors and the random effect, the positive coefficient for *EE* (0.154), negative coefficient for *ZM* (-2.04) and their significant p-values ($<$ 2e-16 and 1.3e-05) show that the uterus weight increases when *EE* increases or *ZM* decreases.
- Does the dose response vary across labs? If so, are there certain labs that appear to be outliers?
  - Yes, the dose response seems to be vary across labs even though the standard deviation of the random intercept for *lab* is only 0.18, which is considered to be small. According to the standard normal quartiles plot, lab Poulenc, Basf, Huntingd, KoreaPar and ChungKor may be considered outliers, and Citfranc, Exxon, Sumitomo might be potential outliers.
- Do the protocols differ in their sensitivity to detecting EE and ZM effects? If so, is there one protocol that can be recommended?
  - Yes, based on the significant p-values for the interaction terms in the final model, protocol B, C and D are statistically different from protocol A in terms of detecting EE and ZM effects. According to the insignificant p-value of the interaction term of *EE* and *protocol B*, 0.899, it seems that both *protocol A* and *B* are good at detecting *EE* compared to other protocols; however, regarding *ZM* detection, protocol B has the best sensitivity with the coefficient estimate -0.632. Therefore, protocol B is recommended.

Nevertheless, there are limitations that may hinder the predictive and inferential abilities:

1. Extrapolation (outside the ranges of EE and ZM) might raise risks of implausible prediction for uterus weight.
2. There is a better modeling approach that can be used to handle the bimodal issue.

Part II: Modeling Voter Turnout in North Carolina

Introduction

This report covers an extensive analysis of voters across several demographic characteristics who registered and voted in North Carolina in 2016. The data is sourced from the North Carolina State Board of Elections (NCSBE) which is the agency responsible for the administration of the elections process, campaign finance, disclosure and compliance. The main objective questions being answered in this report are around the trends observed in the registration and voting processes of the different demographic sections. Of primary interest are the following 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?

For our analysis, we have developed a multilevel logistic regression model (with counties as the main hierarchy group) that represents a binomial relationship between voters who registered versus those who actually voted as a function of demographic predictors like party affiliations, age, sex, race, etc.

Data

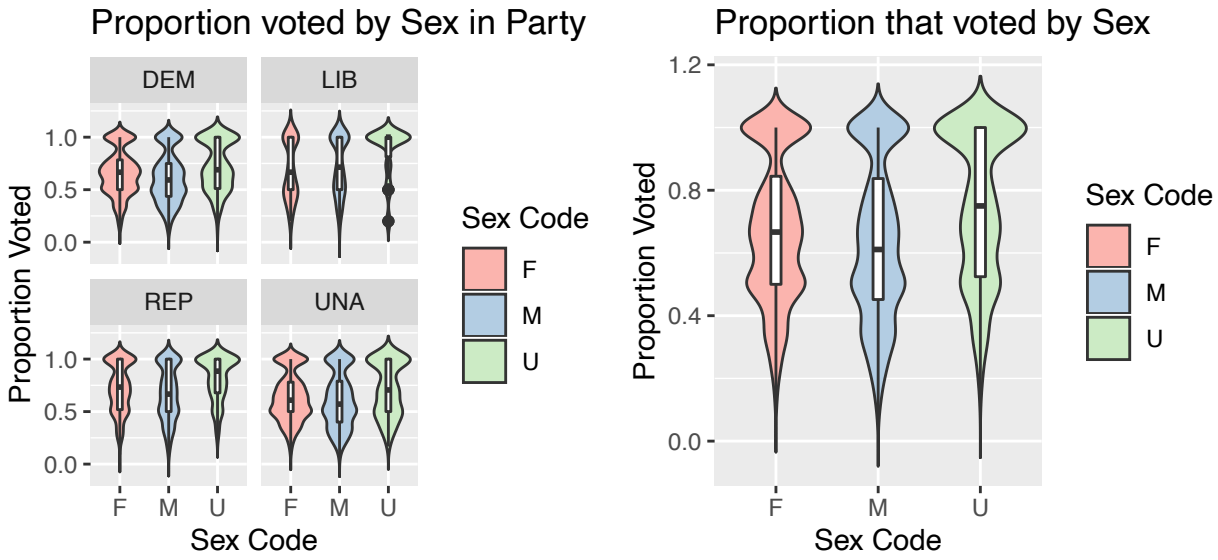
Data Transformation

The data for this analysis was extracted from two files, one containing data on the people who registered to vote and the other containing data on the voters who eventually cast a vote. In the first step of our analysis, we randomly sampled 20 counties from the data on which to focus our analysis. Our 20 counties were: Watauga, Anson, Cumberland, Greene, Perquimans, Pamlico, Richmond, Columbus, Jones, Lincoln, Hertford, Yancey, Alleghany, Craven, Yadkin, Davie, Gates, Alexander, Davidson, and Wilson. These datasets were then joined and aggregated by county, party, race, ethnicity, sex, and age to get the total number of voters in demographic groups represented by these characteristics.

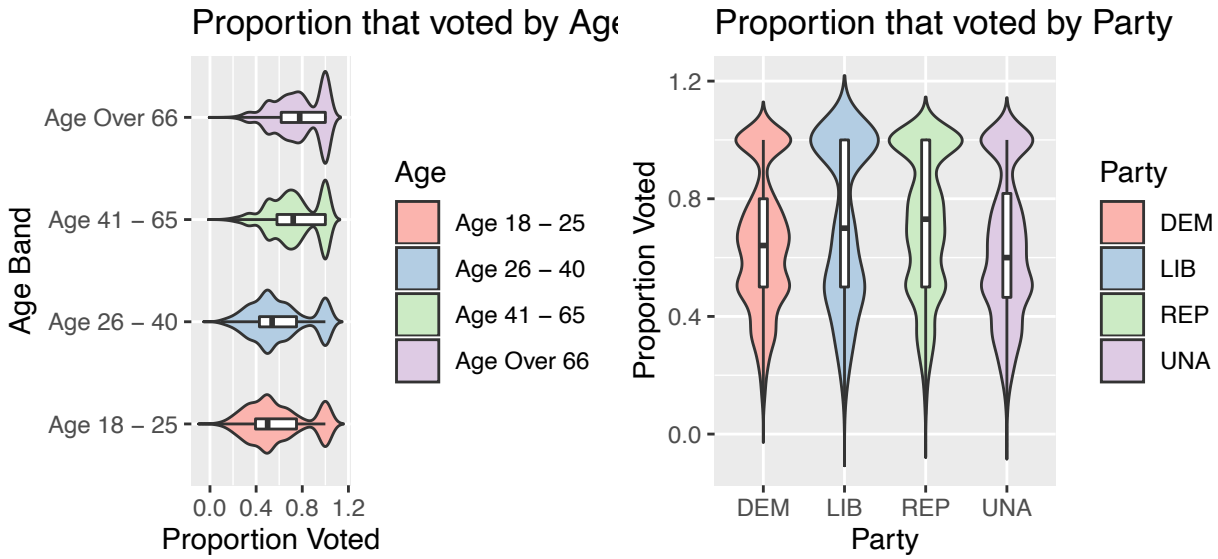
It is important to note here that 241 rows of our merged dataset had more people who voted than those who were registered to vote. One potential explanation for this is that people voted in different counties and precincts than those where they had registered to vote, leading to a turnout that was higher than the total expected turnout for that precinct. As the true reason for this discrepency is unknown, we removed these rows and did not consider them in our analysis. Additionally, for some rows, voter turnout data was unavailable so these rows were not included in the analysis.

Exploratory Data Analysis

In our exploratory data analysis, a new variable indicating the proportion of people that voted (i.e. voters/registered voters) was calculated. This was done to standardize the comparisons as using the absolute voter values would make comparisons difficult specially for significantly differently sized counties.



Sex Code highlighted slightly higher median turn-out for Females than Males but Undesignated sex-code was notably higher than both. Similar patterns were observed when voter turn-out over the different sex codes was observed per party except for Libertarians; they observed slightly higher turn-out for males than females.



Voter turn-out displayed evidence of increasing by age for the counties of interest. In terms of voter turn-out by party, Republicans had the highest median turn-out, with Democrats being the party with the least turnout. Unaffiliated voters had approximaltely a 50% median turn-out which was lower than the turnout for the party affiliated voters. In plotting other variables, little to no discernable relationships were observed.

Model

Model Selection

Model selection was undertaken by creating various models to incorporate different interactions and effects, considering both random and fixed effects. Forward selection was used to incrementally add predictors and interaction terms to our model and model fit was assessed at each step by comparing models using ANOVA chi-squared tests. Additionally, the AIC of the models as well as the p-values of the predictors were assessed at each step to assess model fit and prevent overfitting. In our final, and largest, model, the majority of our

predictors are significant and the AIC is lower than our next best model by 2000 points. This reinforced the decision of selecting the largest model for our final analysis and inference.

In terms of random effects, as we were producing a multi level model, counties were used as the only hierarchy. While nested hierarchies, like by precinct, were considered, our model did not converge for these nested hierarchies so they were abandoned. Additionally, random slopes, such as for race were considered however, again, the model failed to converge for any random slopes. as such, only random intercepts by county were used in the model building process.

Final Model

The final model produced after the data analysis is :-

$$\log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \beta_0 + \gamma_{0j}^{County} + \beta_1 Age_{ij} + \beta_2 Sex_{ij} + \beta_3 Ethnicity_{ij} + \beta_4 Race_{ij} + \beta_5 Party_{ij} +$$
$$\beta_6 Sex_{ij} : Party_{ij} + \beta_7 Race_{ij} : Party_{ij} + \beta_8 Age_{ij} : Party_{ij}$$
$$\text{where} \quad y_{ij}|x_{ij} \sim Bernoulli(\pi_{ij}) \quad \text{and} \quad \gamma_{0j}^{County} \sim N(0, \sigma_0^2)$$

In this model, race = white, sex = female, ethnicity = not latino, and party = Democratic were used as the baseline factors that are absorbed into the intercept for greater ease in interpretation of these predictors' impact on the voter turnout. As all the predictors and their interaction terms, are categorical, there are a total of 49 distinct factors in the model, the majority of which are significant as per our model. The coefficients of note among these factors are displayed below :-

Table 1: Binomial Regression Summary

Coefficients	Log-Odds Estimate	Odds Estimate	Log-Odds Std. Error	z-value	Pr(>|z|)	Signif.
Intercept	0.2087	1.2321	0.039217	5.321	1.03e-07	***
Age 26 - 40	0.0653	1.0675	0.013752	4.747	2.06e-06	***
Age 41 - 65	0.9437	2.5695	0.012988	72.660	< 2e-16	***
Age Over 66	0.9761	2.6541	0.014254	68.478	< 2e-16	***
Sex = Male	-0.3205	0.7258	0.007828	-40.937	< 2e-16	***
Ethnicity = Hispanic or Latino	-0.0428	0.9581	0.018285	-2.339	0.019331	*
Race = Black	0.1297	1.1384	0.008565	15.138	< 2e-16	***
Party = Republican	0.2494	1.2832	0.019184	13.000	< 2e-16	***
Republican : Age 26 - 40	0.1291	1.1379	0.020942	6.167	6.96e-10	***
Republican : Age 41 - 65	0.0383	1.0390	0.019646	1.948	0.051379	.
Republican : Age Over 66	0.1112	1.1176	0.022148	5.020	5.15e-07	***
Republican : Male	0.2660	1.3047	0.012299	21.627	< 2e-16	***
Republican : Black	-1.1863	0.3054	0.029941	-39.621	< 2e-16	***

Note:
See appendix for full summary table.

As per the model summary statistics, the maximum likelihood estimates for the coefficients of sex = male, race = black, ethnicity = latino, and party = republican are significant. In addition, the interaction between sex and party, age and party, and race and party were significant for a majority of the factor pairings. This implies that all these factors impact the log-odds of whether or not a person will vote (and collectively the voter turnout) as the null hypotheses that the slopes for these predictors are not different from 0 are rejected. The largest z-values (absolute value) among the predictors are for Age (specifically Age 41-65 and Age Over 66), sex = male, and the interaction between race = black and party = republican, suggesting that these factors are the strongest predictors of whether or not a person is expected to vote in the election. Moreover, the random effect on County has a grand mean at log odds of 0.2087 with a standard deviation of 0.1645 across the counties. Additionally, the AIC of the model is 34166.1.

Model Assessment

Model Assessment was primarily undertaken by considering the in-sample accuracy of the model. The total voter turnout by each sub category was unpacked and transformed to a binary response. The training data was then fed into the model to assess the model's in-sample predictive power. With a threshold of 0.5, the in-sample accuracy of the model was 0.954 the sensitivity of the model was 0.9654 while the specificity was 0.9307. The confusion matrix for this model is displayed below :-

Table 2: Confusion Matrix

	Reference	
	Not Voted	Voted
Pred: Not Voted	247178	18285
Pred: Voted	18402	510471

Out-of-sample accuracy was not used as a measure of model assessment as that would require removing a large portion of the grouped data leading to a significant detrimental impact on the model's predictive power, rendering any model assessment inaccurate. Similarly, residual and binned residual plots were not used in the model assessment. The residual plots were not used as the response has a bernoulli distribution making the residual plots uninterpretable. Binned residual plots could not be constructed as none of the predictors were continuous.

Additionally, the Variance Inflation Factor (VIF) metric was used to check for multicollinearity. As all the available predictors are categorical and there are interaction terms for several of them, there was a high VIF for all the predictors except Sex. As these high multicolinearity scores are to be expected in the case of only categorical predictors, the analysis was continued without modifying the model.

Results

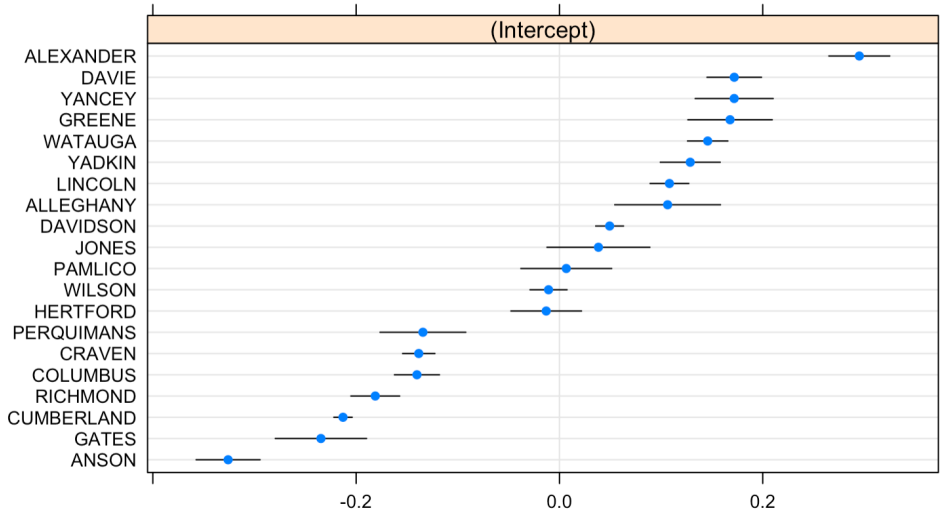
Of primary interest from the summary statistics of the model is that the Maximum Likelihood Estimate of the coefficients for the demographic indicators for age, ethnicity, sex, and race are significant. This suggests that voter turnouts differs significantly between demographic groups. For example, our model suggests that, everything else being constant, compared to females, male voters have lesser odds of voting by a multiplicative effect of 0.726 ($e^{-0.3205}$) i.e. a 27.4% decrease in the odds of voting compared to females.

As per our interaction term between sex and party, this trend/association varries by party. While the direction of this difference in association as well as its significance is directly obtainable via the model summary statistics, an accurate measure of the magnitude of the differences is not directly obtainable. As such, to gain an estimate of this magnitude, we produced new standardized datasets by only varrying our demographics of interest and predicted outcomes on these datasets using our model. These outcomes were then used to calculate the log odds of voting of each subgroup and as well the difference in log odds. In particular, we focused on interpretting the difference in log odds between male democrats and female democrats as well as male republicans and female republicans as these are the two major parties in the US. We found that a female democrat has higher log odds of voting than a male democrat by a multiplicative factor of 1.15 (0.32 in log scale), given that all other variables are constant or same for both records. For Republicans, females higher odds of voting than males by a multiplicative factor of 0.16 (0.05 in log-scale). Comparing male republicans to male democrats, given all other factors are kept constant, a male republican has higher odds of voting by a multiplicative factor of 0.06. For females, this difference is of 0.90 with female Democrats having higher odds than female republicans.

Similarly to these trends in sex, if a voter is of race = black, their odds of voting, compared to a white person (baseline) are higher by a multiplicative effect of 1.14 ($e^{0.13}$). However, this trend/association also varies by party; as per in the interaction term between party and race, compared to the demoncratic party, in the republican party for example, the odds of black people voting are lesser. In terms of age, the odds of older ages voting are higher compared to persons aged 18-25. For a person aged 41-65 years old, for example, the odds of voting are higher than the odds of a person aged 18-25 voting by a multiplicative effect of 2.57 ($e^{0.944}$) i.e. a 157%. This

trend/association tends to vary by party, for example, in the libertarian camp, people aged 41-65 have lower odds of voting compared to people aged 18-25.

As our model is a multilevel model, we were able to compare and contrast how the log odds of a person voting varies across the counties. The overall grand mean of the log odds of a person voting in the counties we selected was 0.209 (odds of 1.23). However, as per our random effect on the intercept of the model, we found that this grand mean has a standard deviation of 0.1645 (odds of 1.179). This means that across these counties, there is significant variance in the odds of a person voting. This is illustrated in the dotplot below :-



As can be seen in the dotplot, the baseline odds of voting in the counties we selected differ widely with the biggest difference being between Alexander (baseline log odds of 0.295 and odds of 1.343) and Anson (baseline log odds of -0.326 and odds of 0.722). This result reprents how a higher proportion of people turnout to vote in some counties compared to others.

Conclusion

To conclude, our final model supports the hypothesis that there is indeed a difference between the turnout for males compared to females and that this turnout differs considerably by party. Additionally, it identifies relations between turnout and age, race, and ethnicity. It also supports the hypothesis that the association between turnout and age, race, and sex varries by party. Being a multilevel model, it also quantifies how the log odds of voting vary across counties.

It is important to note at this stage some limitations of this analysis and modeling framework. Firstly, as residual or binned residual plots for the model could not be made and out of sample accuracy could not be computed without sacrificing important information, the model assessment is limited and it is difficult to determine if the model does any overfitting. Secondly, many data points were deleted for which data for the number of people who voted was greater than the number of people who registered to vote. Perhaps this discrepency is of value and could have been better understood by consulting a domain expert. Dropping these data points has allowed the model to be trained on lesser data than possible. Thirdly, as all the predictors are categorical, interaction terms in our model are not easily interpretable and manual calculation of the odds using standardized predictions is necessary to quantify the magnitude of the change in odds represented by the interaction terms. Finally, while this model does indicate several associations between demographic characteristics and voter turnout, this does not mean that the model is confirming a causal relationship between these variables. In short, correlation does not imply causation in this model. As such, a more detailed analysis, perhaps one that uses domain knowledge to make use of the data that has not been utilized in this analysis and which employs some continuous predictors for the log-odds of voting could lead to a better model and understanding of the relationships in the data.

Appendix - Full Model Summary

Coefficient	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.2086932	0.0392171	5.3214888	0.0000001
Age 26 - 40	0.0652868	0.0137523	4.7473494	0.0000021
Age 41 - 65	0.9437122	0.0129881	72.6600069	0.0000000
Age Over 66	0.9760901	0.0142540	68.4783389	0.0000000
Sex = Male	-0.3204615	0.0078281	-40.9372551	0.0000000
Sex = Undesignated	0.0384832	0.0308890	1.2458543	0.2128179
Ethnicity = Hispanic or Latino	-0.0427702	0.0182850	-2.3390914	0.0193307
Ethnicity = Undesignated	-0.0977974	0.0066123	-14.7902195	0.0000000
Race = Asian	-0.0043166	0.0557433	-0.0774365	0.9382763
Race = Black	0.1296670	0.0085654	15.1384025	0.0000000
Race = American Indian or Alaskan Native	-0.3006342	0.0446360	-6.7352475	0.0000000
Race = Two or More Races	0.0099247	0.0456694	0.2173157	0.8279623
Race = Other	-0.3556188	0.0266926	-13.3227429	0.0000000
Race = Undesignated	-0.0057028	0.0297160	-0.1919109	0.8478120
Party = Libertarian	-0.3928386	0.0794356	-4.9453715	0.0000008
Party = Republican	0.2493908	0.0191839	13.0000370	0.0000000
Party = Unaffiliated	-0.2683718	0.0173145	-15.4998059	0.0000000
Male : Libertarian	0.4486265	0.0773299	5.8014600	0.0000000
Undesignated : Libertarian	0.7514761	0.2872371	2.6162227	0.0088909
Male : Republican	0.2659912	0.0122989	21.6272028	0.0000000
Undesignated : Republican	0.2939184	0.0563219	5.2185463	0.0000002
Male : Unaffiliated	0.1913049	0.0120845	15.8306504	0.0000000
Undesignated : Unaffiliated	-0.0560309	0.0421861	-1.3281835	0.1841175
Libertarian : Age 26 - 40	0.3487556	0.0888275	3.9262124	0.0000863
Republican : Age 26 - 40	0.1291495	0.0209419	6.1670345	0.0000000
Unaffiliated : Age 26 - 40	0.1382330	0.0185521	7.4510808	0.0000000
Libertarian : Age 41 - 65	-0.2839734	0.1011838	-2.8065112	0.0050081
Republican : Age 41 - 65	0.0382753	0.0196455	1.9482970	0.0513794
Unaffiliated : Age 41 - 65	0.0643484	0.0178808	3.5987348	0.0003198
Libertarian : Age Over 66	0.1575664	0.2643844	0.5959748	0.5511921
Republican : Age Over 66	0.1111942	0.0221482	5.0204566	0.0000005
Unaffiliated : Age Over 66	0.4570485	0.0222023	20.5856190	0.0000000
Libertarian : Asian	1.5432121	0.6385435	2.4167691	0.0156589
Republican : Asian	-0.5319627	0.0820963	-6.4797410	0.0000000
Unaffiliated : Asian	-0.1913439	0.0706786	-2.7072381	0.0067846
Libertarian : Black	-0.1345311	0.1410523	-0.9537673	0.3402015
Republican : Black	-1.1862704	0.0299405	-39.6209283	0.0000000
Unaffiliated : Black	-0.3569671	0.0153919	-23.1918554	0.0000000
Libertarian : American Indian or Alaskan Native	0.7358966	0.6285588	1.1707682	0.2416920
Republican : American Indian or Alaskan Native	0.0012453	0.0821379	0.0151607	0.9879039
Unaffiliated : American Indian or Alaskan Native	-0.0551468	0.0709750	-0.7769891	0.4371652
Libertarian : Two or More Races	0.3668052	0.3334777	1.0999391	0.2713586
Republican : Two or More Races	-0.1251561	0.0931164	-1.3440816	0.1789220
Unaffiliated : Two or More Races	0.0610390	0.0659312	0.9257979	0.3545510
Libertarian : Other	0.5130552	0.1955993	2.6229904	0.0087162
Republican : Other	-0.3686858	0.0418795	-8.8034995	0.0000000
Unaffiliated : Other	-0.0702873	0.0342970	-2.0493732	0.0404256
Libertarian : Undesignated	0.6646543	0.2556598	2.5997603	0.0093289
Republican : Undesignated	-0.0297872	0.0521720	-0.5709415	0.5680393
Unaffiliated : Undesignated	-0.1510652	0.0391177	-3.8618124	0.0001125

Appendix

```
knitr::opts_chunk$set(echo=FALSE,
                      warning=FALSE,
                      message=FALSE,
                      fig.align="center",
                      fig.pos='H')
list_of_packages = c("knitr",
                     "tidyverse",
                     "dplyr",
                     "caret",
                     "broom",
                     "MASS",
                     "xtable",
                     "lme4",
                     "lmerTest",
                     "gridExtra",
                     "jtools",
                     "psycho")
packages = list_of_packages[!(list_of_packages %in% installed.packages()[,"Package"])]
if(length(packages)){install.packages(packages)}
pkg_lib = lapply(list_of_packages, require, character.only = TRUE)
options(digits=3)
bioassay = read_table2("../Data/bioassay.txt") %>%
    mutate_at(vars(protocol, lab, group), as.factor) %>%
    mutate_at(vars(uterus, weight, EE, ZM), as.numeric)
bioassay = bioassay %>%
    na.omit() %>% # remove NAs
    mutate(weight_c=weight-mean(weight)) # centering the data
p1 = bioassay %>%
    ggplot() +
    geom_density(aes(x=uterus), col="#F8766D") +
    theme_classic()

p2 = bioassay %>%
    ggplot() +
    geom_density(aes(x=log(uterus)), col="#00BFC4") +
    theme_classic() +
    labs(y="")

grid.arrange(p1, p2, ncol=2)
set.seed(123)

lab.names = bioassay$lab %>%
    unique()

sample.lab.names = lab.names %>% sample(5)

bio.sample = bioassay %>%
    filter(lab %in% sample.lab.names)

bio.sample %>%
    ggplot(aes(x=EE, y=log(uterus), colour=lab)) +
    geom_point(alpha=0.4, shape=19) +
```

```
    geom_smooth(method="lm", se=F) +
    theme_classic() +
    labs(x="EE (mg/kg/day)",
         y="log of uterus (mg)") +
    theme(legend.position="none") +
    facet_grid(~lab)
bio.sample %>%
    ggplot(aes(x=ZM, y=log(uterus), colour=lab)) +
    geom_point(alpha=0.4, shape=19) +
    geom_smooth(method="lm", se=F) +
    theme_classic() +
    labs(x="ZM (mg/kg/day)",
         y="log of uterus (mg)") +
    theme(legend.position="none") +
    facet_grid(~lab)
bioassay %>%
    ggplot(aes(x=EE, y=log(uterus), colour=protocol)) +
    geom_point(alpha=0.4, shape=19) +
    geom_smooth(method="lm", aes(x=EE, y=log(uterus)), se=F) +
    labs(x="EE (mg/kg/day)",
         y="log of uterus (mg)") +
    facet_grid(~protocol) +
    theme_classic()
bioassay %>%
    ggplot(aes(x=ZM, y=log(uterus), colour=protocol)) +
    geom_point(alpha=0.4, shape=19) +
    geom_smooth(method="lm", aes(x=ZM, y=log(uterus)), se=F) +
    labs(x="ZM (mg/kg/day)",
         y="log of uterus (mg)") +
    facet_grid(~protocol) +
    theme_classic()
final_model_with_weight_c = lmer(log(uterus) ~ (1 | lab) +
                  weight_c + EE + ZM + protocol +
                  EE:protocol + ZM:protocol,
              bioassay)

final_model = lmer(log(uterus) ~ (1 | lab) +
                  EE + ZM + protocol +
                  EE:protocol + ZM:protocol,
              bioassay)
# car::vif(model4)
anova(final_model_with_weight_c, final_model, test.statistic=c("LR")) %>%
    kable()
# summary(final_model)
final_model %>%
    summ(confint=TRUE,
         ci.width=.95, digit=5,
         model.info=getOption("summ-model.info", FALSE),
         model.fit=getOption("summ-model.fit", FALSE),
         groups.table = getOption("summ-groups.table", FALSE))

rand.effect = ranef(final_model, condVar=TRUE)
qq = attr(ranef(final_model, condVar = TRUE)[[1]], "postVar")
df = data.frame(Intercepts=rand.effect$lab[,1],
```

```
                sd.interc=2*sqrt(qq[,,1:length(qq)]),
                lev.names=rownames(rand.effect$lab))
df$lev.names = factor(df$lev.names,
                      levels=df$lev.names[order(df$Intercepts)])
ggplot(df, aes(lev.names, Intercepts, shape=lev.names)) +
    geom_hline(yintercept=0) +
    geom_errorbar(aes(ymin=Intercepts-sd.interc,
                      ymax=Intercepts+sd.interc), width=0, color="black") +
    geom_point() +
    guides(size=FALSE, shape=FALSE) +
    scale_shape_manual(values=rep(2, 19)) +
    theme_classic() +
    labs(x="Lab", y="Standard normal quartiles") +
    theme(axis.text.x=element_text(size=rel(1)),
            axis.title.x=element_text(size=rel(1.2)),
            axis.text.y=element_text(size=rel(1)),
          axis.title.y=element_text(size=rel(1.2)),
          panel.grid.minor=element_blank(),
            panel.grid.major.x=element_blank()) +
    coord_flip()
```

Appendix - Code

```
library(tidyverse)
library(lme4)
library(lmerTest)
library(caret)
library(rlist)
library(kableExtra)
library(usdm)
hist_votes_all = read.delim("history_stats.txt", stringsAsFactors = FALSE)
voters_all = read.table("voter_stats.txt", header = TRUE, comment.char = "", stringsAsFactors = FALSE)
rand_counties =c("WATAUGA", "ANSON", "CUMBERLAND", "GREENE", "PERQUIMANS", "PAMLICO", "RICHMOND", "COLU
names(hist_votes_all) <- c("county_desc", "precinct_abbrv", "vtd_abbrv", "age", "party_cd", "race_code"
comb_all_hist = hist_votes_all %>% group_by(county_desc, party_cd, race_code, ethnic_code, sex_code, ag
                                   summarise(total_voted = sum(total_voted))
comb_all_votes = voters_all %>% group_by(county_desc, party_cd, race_code, ethnic_code, sex_code, age)
                                   summarise(total_voters = sum(total_voters))

comb_all = comb_all_hist %>% left_join(comb_all_votes,by =c("county_desc", "party_cd", "race_code", "et
#comb_all = left_join(voters_all,hist_votes_all)
voters_sample = comb_all %>% filter(county_desc %in% rand_counties) %>%
                             filter(!is.na(total_voters)) %>%
                               filter(total_voters>=total_voted) %>%
                               mutate(voted_proportion = total_voted/pmax(total_voters,total_voted))

variblesInterested <- c("county_desc", "age", "party_cd", "race_code", "ethnic_code", "sex_code", "vote

voters_sample$race_code<- as.factor(voters_sample$race_code)
voters_sample$race_code = relevel(voters_sample$race_code, "W")
voters_sample$sex_code<- as.factor(voters_sample$sex_code)
voters_sample$sex_code = relevel(voters_sample$sex_code, "F")
voters_sample$ethnic_code<- as.factor(voters_sample$ethnic_code)
voters_sample$ethnic_code = relevel(voters_sample$ethnic_code, "NL")
mod3 = glmer(cbind(total_voted,total_voters-total_voted) ~ sex_code:party_cd +
    age + age:party_cd + sex_code + ethnic_code + race_code +
    party_cd + party_cd:race_code + (1 | county_desc), data = voters_sample, family = binomial)

decomp_votes = data.frame(Out=integer())

result <- rep(0,0)
for (i in seq(1,dim(voters_sample[, c("total_voters", "total_voted")])[1])) {
#for (i in seq(1,5)) {
ones <- (rep(1, voters_sample[i, c("total_voters", "total_voted")]$total_voted))
zeros <- (rep(0, (voters_sample[i, c("total_voters", "total_voted")]$total_voters
                    - voters_sample[i, c("total_voters", "total_voted")]$total_voted)))
 row_result <- list.append(ones,zeros)
 result <- list.append(result, row_result)
}
result
preds<-round(fitted(mod3)*voters_sample$total_voters)
preds

voters_sample$pred_voted<-preds

result_preds <- rep(0,0)
```

```
for (i in seq(1,dim(voters_sample[, c("total_voters", "pred_voted")])[1])) {
#for (i in seq(1,5)) {
ones <- (rep(1, voters_sample[i, c("total_voters", "pred_voted")]$pred_voted))
zeros <- (rep(0, (voters_sample[i, c("total_voters", "pred_voted")]$total_voters
                  - voters_sample[i, c("total_voters", "pred_voted")]$pred_voted)))
 row_result <- list.append(ones,zeros)
 result_preds <- list.append(result_preds, row_result)
}

conf_input<-data.frame(result,result_preds)
Conf_mat_iter <- confusionMatrix(as.factor(conf_input$result_preds),
                                 as.factor(conf_input$result),positive = "1")
sexcodeplt <- voters_sample %>%
  ggplot(aes(x = sex_code, y = voted_proportion,fill=sex_code)) +
  geom_violin(trim = FALSE) + geom_boxplot(width=0.1, fill="white") + scale_fill_brewer(palette = "Past

sexpartyplt <- voters_sample %>%
  ggplot(aes(x = sex_code, y = voted_proportion,fill=sex_code)) +
  geom_violin(trim = FALSE) + geom_boxplot(width=0.1, fill="white") + scale_fill_brewer(palette = "Past
  facet_wrap(~party_cd) + labs(y= "Proportion Voted", x = "Sex Code",title = "Proportion voted by Sex i

grid.arrange(sexpartyplt,sexcodeplt,nrow = 1)

ageplt <- voters_sample %>%
  ggplot(aes(x = age, y = voted_proportion,fill=age)) +
  geom_violin(trim = FALSE) + geom_boxplot(width=0.1, fill="white") + scale_fill_brewer(palette = "Past

agepartyplt <- voters_sample %>%
  ggplot(aes(x = age, y = voted_proportion,fill=age)) +
  geom_violin(trim = FALSE) + geom_boxplot(width=0.1, fill="white") + scale_fill_brewer(palette = "Past
  facet_wrap(~party_cd) + labs(y= "Proportion Voted", x = "Age Band",title = "Proportion voted by Age i

partyplt <- voters_sample %>%
  ggplot(aes(x = party_cd, y = voted_proportion,fill=party_cd)) +
  geom_violin(trim = FALSE) + geom_boxplot(width=0.1, fill="white") + scale_fill_brewer(palette = "Past

raceplt <- voters_sample %>%
  ggplot(aes(x = race_code, y = voted_proportion,fill=race_code)) +
  geom_violin(trim = FALSE) + geom_boxplot(width=0.1, fill="white") + scale_fill_brewer(palette = "Past

grid.arrange(ageplt,partyplt,nrow = 1)

library(kableExtra)
library(tidyverse)
coefficients = c("Intercept", "Age 26 - 40", "Age 41 - 65", "Age Over 66", "Sex = Male", "Ethnicity = H
log_odds = c(0.208693, 0.065287, 0.943712, 0.976090, -0.320462, -0.042770, 0.129667, 0.249391, 0.129149
odds = log_odds %>% exp() %>% round(4)
stderror = c(0.039217, 0.013752, 0.012988, 0.014254, 0.007828, 0.018285, 0.008565, 0.019184, 0.020942,
z = c(5.321, 4.747, 72.660, 68.478, -40.937, -2.339, 15.138, 13.000, 6.167, 1.948, 5.020, 21.627, -39.6
p_val = c("1.03e-07", "2.06e-06", rep("< 2e-16", 3), 0.019331, rep("< 2e-16", 2), 6.96e-10, 0.051379, 5
signif = c(rep("***", 5), "*", rep("***", 3), ".", rep("***", 3))

sm = tibble("Coefficient" = coefficients,
            "Log-Odds" = log_odds %>% round(4),
            "Odds Ratio" = odds,
```

```
            "Std. Error" = stderror,
            "z-values" = z,
            "p-value" = p_val,
            "Signif." = signif)

sm %>%
  kable("latex", booktabs = T, linesep = "",
        escape = F, caption = "Binomial Regression Summary",
        align = c('l', rep('r', 5), 'c'),
        col.names = linebreak(c(
          "Coefficients",
          "Log-Odds\nEstimate",
          "Odds Estimate",
          "Log-Odds\nStd. Error",
          "z-value",
          "Pr(>|z|)",
          "Signif."
        ))) %>%
  kable_styling(full_width = F, latex_options = "hold_position") %>%
  footnote(general = "See appendix for full summary table.")
```