Team Project 2 Part 1: Estrogen Bioassay

Guillem Amat, Calvin Dong, Jose Moscoso, Varun Prasad, Anshupriya Srivastava

November 4, 2019

Summary

This report investigates two different applications of hierarchical or multilevel models. In the first part, a varying intercept model was developed to determine whether the results of a rat uterotrophic bioassay test were consistent across international multi-laboratories studies. The model used laboratory as the random intercept. The final model showed that the uterotrophic bioassay was successful at identifying estrogenic effects of EE and anti-estrogenic effects of ZM. The response trends showed that uterus weight exhibited an increasing dose response trend for EE and a decreasing dose response trend for ZM. In addition, the dose response varied across labs with following labs identified as outliers: Poulenc, Basf, Huntingd, KoreaPar and ChungKor. Finally, protocol B showed the greatest sensitivity to EE and ZM on rat's uterus weight and is the most recommended. Although the model does provide effective answers to the presented questions, it is important to note that it does not fully satisfy all of the assumptions for linear regression, particularly normality. Therefore, further analysis or an alternate model may better explain the results of the rat uterotrophic bioassay across these labs.

Introduction

Estrogens are a group of hormones that are commonly associated with female biological processes such as puberty, menstruation, and pregnancy. However, since they also help regulate the growth of tissues in the body and are key in the endocrine system, any modification of their activity has potentially significant results. In recent years, the increased presence of compounds that affect estrogen activity has increased the need to screen for chemicals that may cause endocrine disruption. To identify estrogen agonists and antagonists, a rat uterotrophic bioassay was conducted as part of an international multi-laboratory study. In this assay, compounds that affect estrogen activity are administered to rats, and the rats' uterus weight is measured as the response. The rats in this study have had their ovaries removed or are immature, so they do not produce estrogen. Rats were randomized to treatment groups that included a control group and several groups in which either a known estrogen agonist (EE) or antagonist (ZM) were administered. The focus of the study was to determine if results were consistent across all the laboratories.

Data

The data used for this analysis consisted of a groups of rats randomized across different treatment groups and tested across 19 laboratories. Each of them applied at least one protocol to test the dose response of EE and ZM. Each treatment group received different doses of EE and ZM, measured in mg/kg/day, as shown in Table 1.

Group	1	2	3	4	5	6	7	8	9	10	11
EE (mg/kg/day)	0	0	.01	.03	0.1	0.3	1	3	10	3	3
ZM (mg/kg/day)	0	0	0	0	0	0	0	0	0	0.1	1

Table 1: EE and ZM Doses by Treatment Group

In group 1, the rats received no doses of EE or ZM while in group 2 they received a vehicle control of EE, such as saline. The laboratories also performed at least one of four different protocols as part of the treatment.

Protocol A used immature female rats dosed by oral gavage (3 days). Protocol B used immature female rats dosed by injection (3 days). Protocol C used adult ovariectomized female rats dosed by injection (3 days). Protocol D used adult ovariectomized female rats dosed by injection (7 days).

Initial inspection of the data indicated that the presence of several missing values, specifically four in the "uterus weight" column and two in the "weight" column. These values were removed from the dataset to simplify the analysis. These values were also not concentrated to one protocol or one lab, so removing them is unlikely to have a significant effect on the results.

Exploratory data analysis first investigated the distribution of the response variable of uterus weight. As shown in the following plots, the original distribution was right skewed. Several different transformations, including logarithmic and square root, were applied on the data, but none truly normalized the distribution. However, the log transformation did provide the best normalization of the uterus weight and is the easiest to interpret. Thus, models were developed using the log transformation of uterus weight as the response variable.

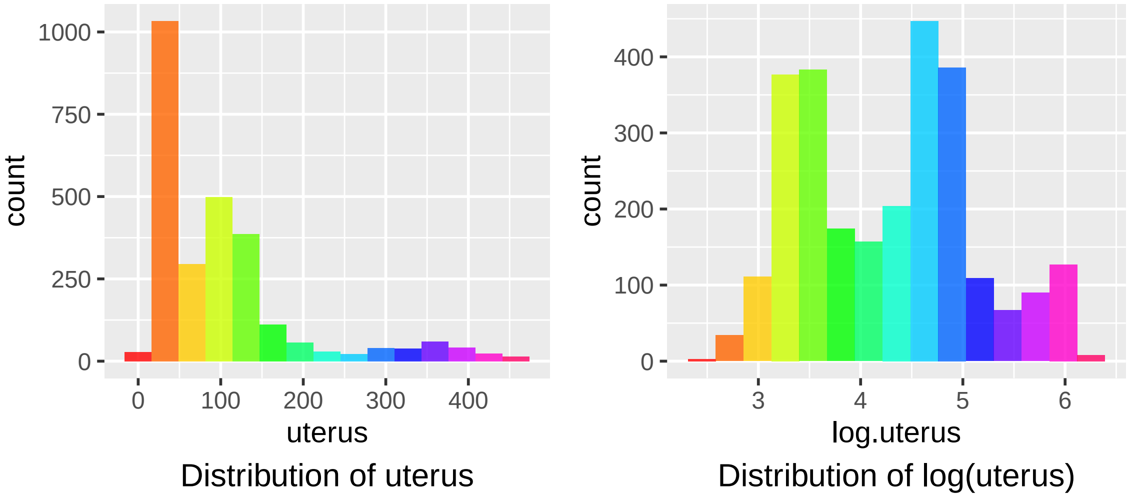


Figure 1: Distributions of Uterus Weight

Additional EDA focused on the output of the response variable (uterus weight) against the following potential predictors: protocol; body weight; dose of EE; dose of ZM; and lab. Group was not explored in EDA because of its high correlation with EE and ZM, as shown in Table 1.

Body weight is linearly related with log uterus based on EDA, but it clearly distributes into two clusters, which is consistent with the design of the four protocols (two of them use immature rats and the other two use adults rats). Therefore, body weight may not be considered as a predictor because it is probably highly correlated with protocol. However, this can only be truly concluded after model selection. The distributions of log uterus are clearly different in the labs based on respective boxplots for each lab. The box plot of protocol vs. log uterus shows there are obvious differences among four protocols. Protocols seem to differ in these relationships (EE vs. log uterus and ZM vs. log uterus) as shown in Figure 2. The slopes of EE vs. log(uterus) by protocol appear to be similar while the slopes of ZM vs. log(uterus) by protocol are clearly different for A compared to the other three protocols. This means there are probably interactions between EE and protocols as well as ZM and protocols. See additional plots describing these trends in Appendix 1.1.

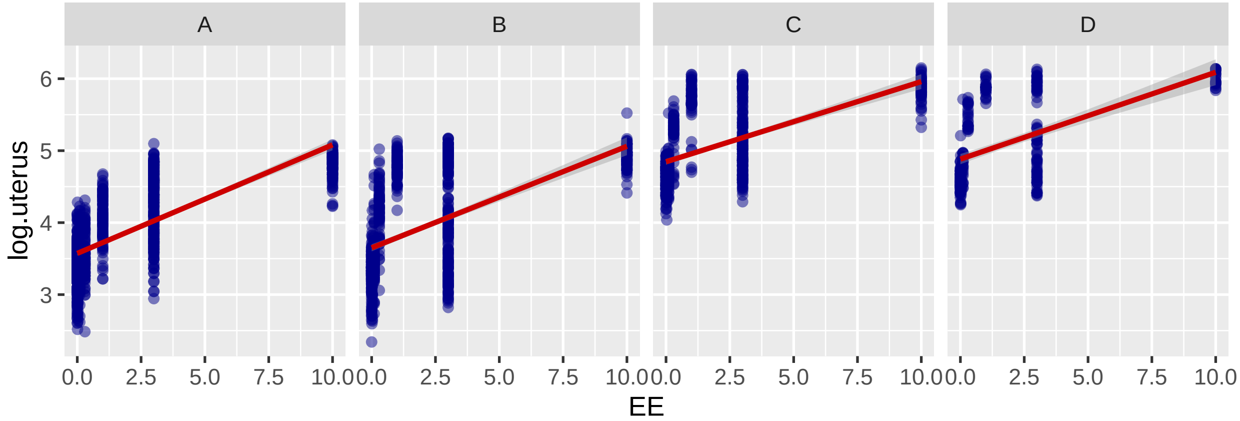


Figure 2: Effect of EE on Log(Uterus) by Protocol

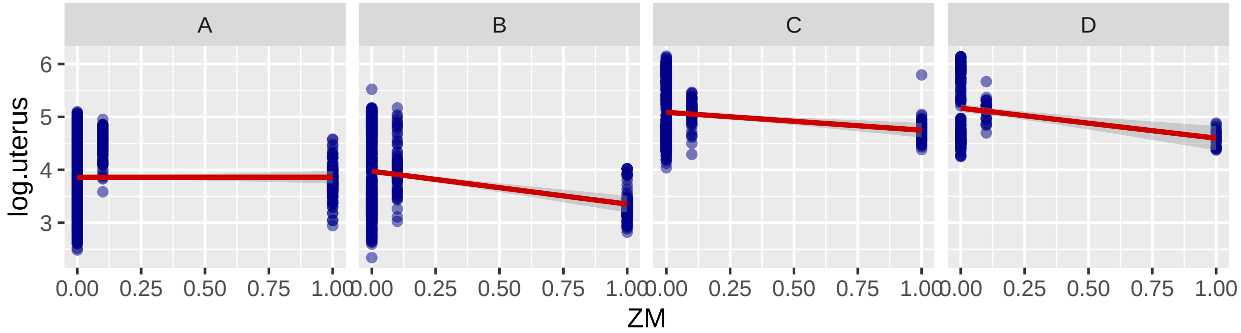


Figure 3: Effect of ZM on Log(Uterus) by Protocol

Model

A random-intercepts model was used to fit the data. The general formula for a random-intercepts model is the following:

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + ... + \beta_n x_{ij} + \epsilon_{ij};\ i = 1, ..., n_j;\ j = 1, ..., J$$

$$\epsilon_{ij} \sim (0, \sigma^2); \beta_{0j} \sim (\beta_0, \tau^2)$$

where i indexes observations, j indexes groups, σ^2 is the within group (residual) variance, and τ^2 is the between group (intercept) variance.

The final model was fit with the following predictor variables: EE; ZM; protocol; EE:protocol; and ZM:protocol. Labs were used as the level since the problem is focused on the consistency of bioassay results across labs. Therefore, it is most appropriate to generate the varying-intercepts model using labs as the random effect variable. The model was developed through a manual selection process and using F-tests for comparison. Weight was found to be not significant when removed from the model since protocol contains the necessary information about weight. Furthermore, group was not included since knowing the EE and ZM doses is enough to identify the treatment group. This intermediate model only had the variables of EE, ZM, and protocol. Interaction terms between EE and protocol and ZM and protocol were added to this model and an F-test was performed to see if there was a significant difference between the models. The F-test resulted in a p-value of 2.2e-16, indicating that adding these interaction terms was very significant.

The final model had an AIC of 3228.354, a residual variance of 0.18532, and an intercept variance of 0.03225. The coefficients of this model are shown in the following table.

	Estimate	Std. Error	df	t value	Pr(>|t|)
Intercept	3.561e+00	4.463e-02	2.141e+01	79.785	< 2e-16
EE	1.539e-01	4.698e-03	2.647e+03	32.754	< 2e-16
ZM	-2.044e-01	4.682e-02	2.647e+03	-4.366	1.31e-05
Protocol B	1.156e-01	2.718e-02	2.624e+03	4.255	2.16e-05
Protocol C	1.355e+00	3.049e-02	2.592e+03	44.457	< 2e-16
Protocol D	1.357e+00	3.957e-02	2.663e+03	34.308	< 2e-16
EE:Protocol B	-9.070e-04	7.178e-03	2.647e+03	-0.126	0.899456
EE:Protocol C	-3.547e-02	7.805e-03	2.647e+03	-4.545	5.75e-06
EE:Protocol D	-2.288e-02	1.047e-02	2.647e+03	-2.186	0.028904
ZM:Protocol B	-6.316e-01	7.149e-02	2.646e+03	-8.835	< 2e-16
ZM:Protocol C	-2.987e-01	7.798e-02	2.646e+03	-3.830	0.000131
ZM:Protocol D	-5.440e-01	1.046e-01	2.646e+03	-5.200	2.15e-07

Table 2: Coefficients of Multilevel Model

The model thus has the final following formula:

$$y_{ij} = \beta_{0,j} + \beta_1 EE_{ij} + \beta_2 ZM_{ij} + \beta_3 Protocol_{ij} + \beta_4 EE : Protocol_{ij} + \beta_5 ZM : Protocol_{ij} + \epsilon_{ij};\ i = 1, ..., n_j;\ j = 1, ..., J$$

$$\epsilon_{ij} \sim (0, \sigma^2); \beta_{0j} \sim (\beta_0, \tau^2)$$

where i indexes observations, j indexes labs, σ^2 is the within lab variance, and τ^2 is the between lab variance.

The model was validated by checking the following four key assumptions for linear models: linearity, equal variance, independence, and variance. For this model, linearity is difficult to assess since there are only a few values for the continuous variables of EE and ZM. However, for some labs, there is a generally equal spread of points about 0 when plotting the variable against the residuals. A plot of the fitted values against the predictors shows a general even spread about 0, though this is not true near the right end of the plot. However, equal variance and independence can generally be satisfied from this plot. Normality is clearly violated as shown in a q-q plot since the points deviate from the 45 degree line, thus weakening the model. This may be due to the more bimodal nature of the response variable distribution. Regardless, for the purpose of this analysis, conclusions will be made assuming the model is valid. Plots for assumptions are shown in Appendix 1.2.

Results

Based on the results shown in Table 2, several key insights can be determined. The intercept indicates the baseline log of uterus weight for a control group rat treated via protocol A. First, EE and ZM are significant in the model and their coefficients show the expected trend of an estrogen agonist and antagonist. With all else being equal, each 1 mg/kg/day dose of EE will increase the log of uterus weight by 0.1539 mg i.e. the uterus weight by 1.1663536 mg. With all else being equal, each 1 mg/kg/day dose of ZM will decrease the log of uterus weight by -0.2044 mg i.e. the uterus weight by 0.8151228 mg. The individual protocol terms are also significant, but for protocols C and D, this may be partially due to the fact that rats in these protocols are adults and will have a larger uterus by default. The interaction terms of EE and ZM with protocol provide insight into the sensitivity of the effects of EE and ZM. Compared to protocol A, all three protocols have a weaker effect when EE is administered as indicated by the negative coefficients. However, this value is not significant for protocol B. Furthermore, the interaction between ZM and protocol B shows the largest negative value coefficient and largest t-value by magnitude for these interaction terms. Thus, protocol B has the largest effect when ZM is administered and has a negligible effect compared to protocol A when EE is administered. Therefore, protocol B should be recommended for best determining sensitivity of the estrogen compounds.

A dotplot of the multilevel model can be used to identify outliers and thus further highlight the variation in results across labs. From the results of the dotplot shown below, it is clear that dose response results vary across labs, and there are five clear outliers: Chungkor, KoreaPar, Huntingd, Basf, and Poulenc.

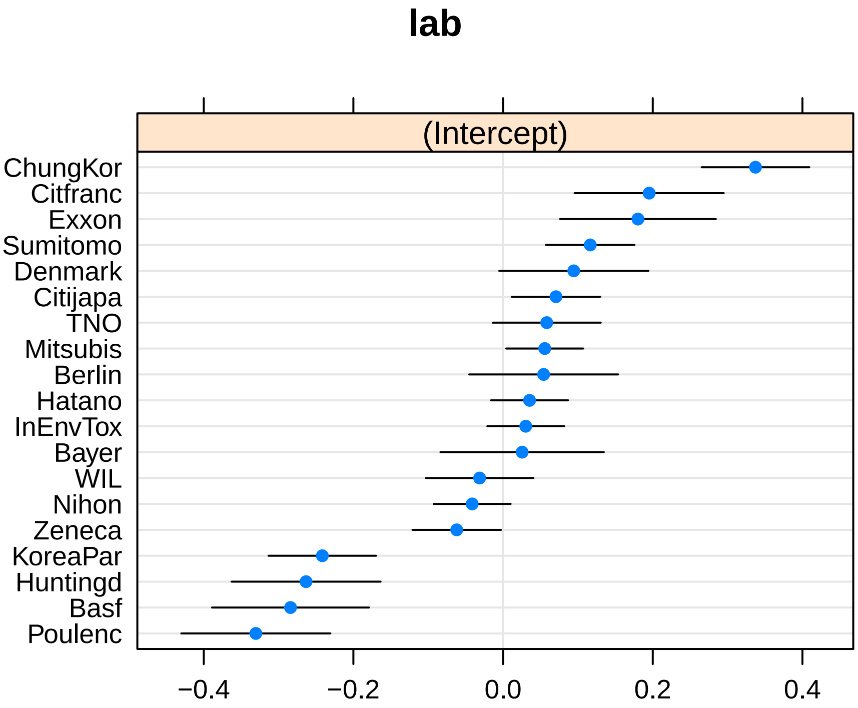


Figure 4: Dotplot of Labs from Multilevel Model

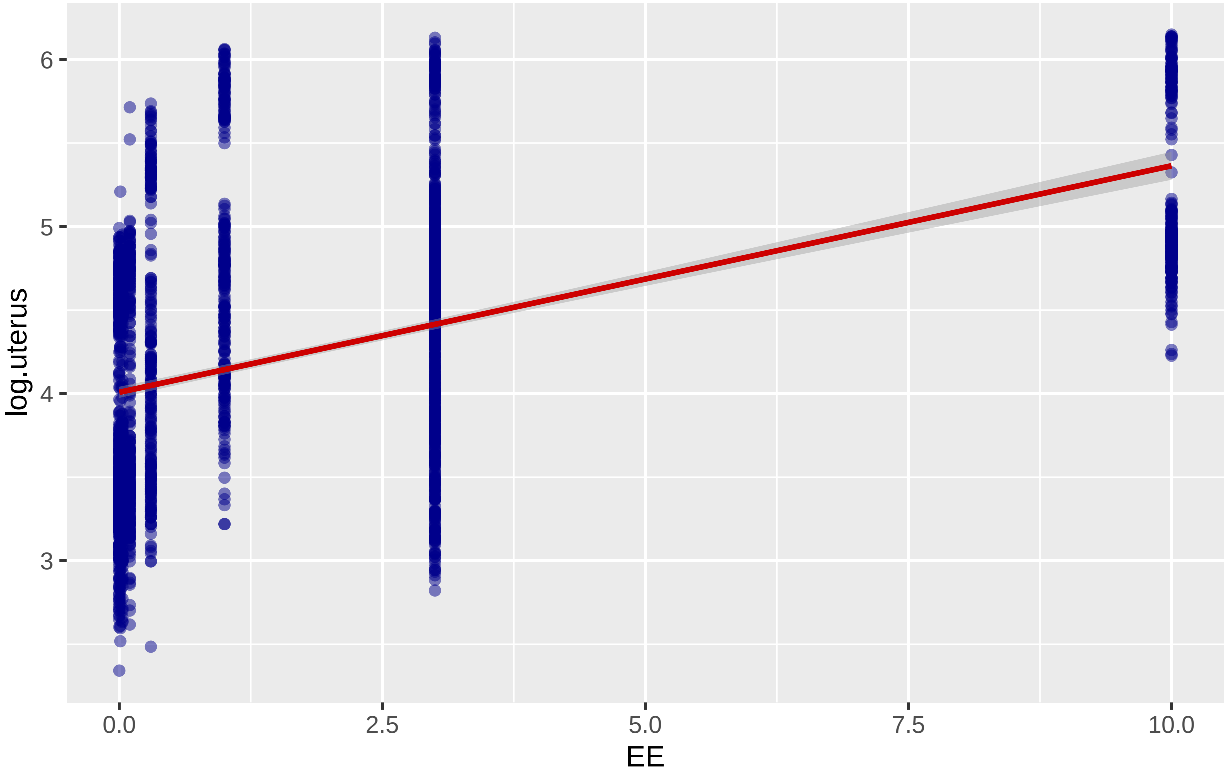
Conclusions

Overall, while the model was not fully valid due to the violation of normality, many insights about the bioassay data could still be generated. Specifically, the results showed that there is an increasing dose in estrogen agonist EE increased uterus weight while an increasing dose of the antagonist ZM decreased uterus weight. The dose response trends varied across labs as shown by the dotplot, and there were five clear outliers. Finally, based on the interactions between protocol and EE or ZM, it was clear that protocol B showed the greatest sensitivity to the effects of EE and ZM. However, there are several limitations in this model. Since the uterus weight was not normally distributed even after a logarithmic transformation, a multilevel linear regression is not the optimal method for fitting this data. This is further supported by the violation of the normality assumption. Furthermore, not all the labs tested all groups or all protocols. As a result, there is less uniformity amongst the labs, so generalizability of the results may be weakened. The chance of errors affecting the data is higher when only a few labs are testing certain protocols or groups. It would also be beneficial to have more data on ZM doses, particularly when EE is not administered, to better see its effects. Having fewer doses of ZM compared to EE makes it more difficult to assess linearity and reduces some certainty in the results. Further exploration should be done with a model that can account for a non-normal distribution to see if the insights generated from this hierarchical model will hold true for a more appropriate model.

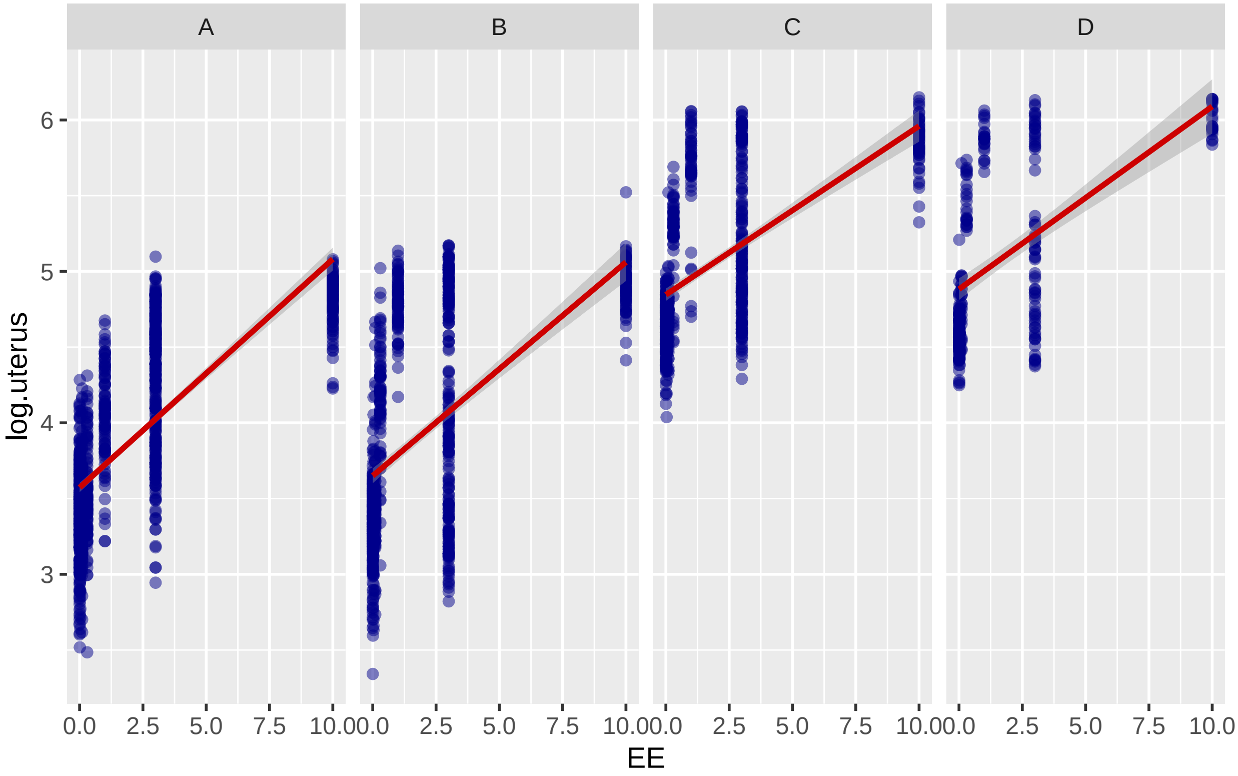
Appendices

Appendix 1: Estrogen Bioassay

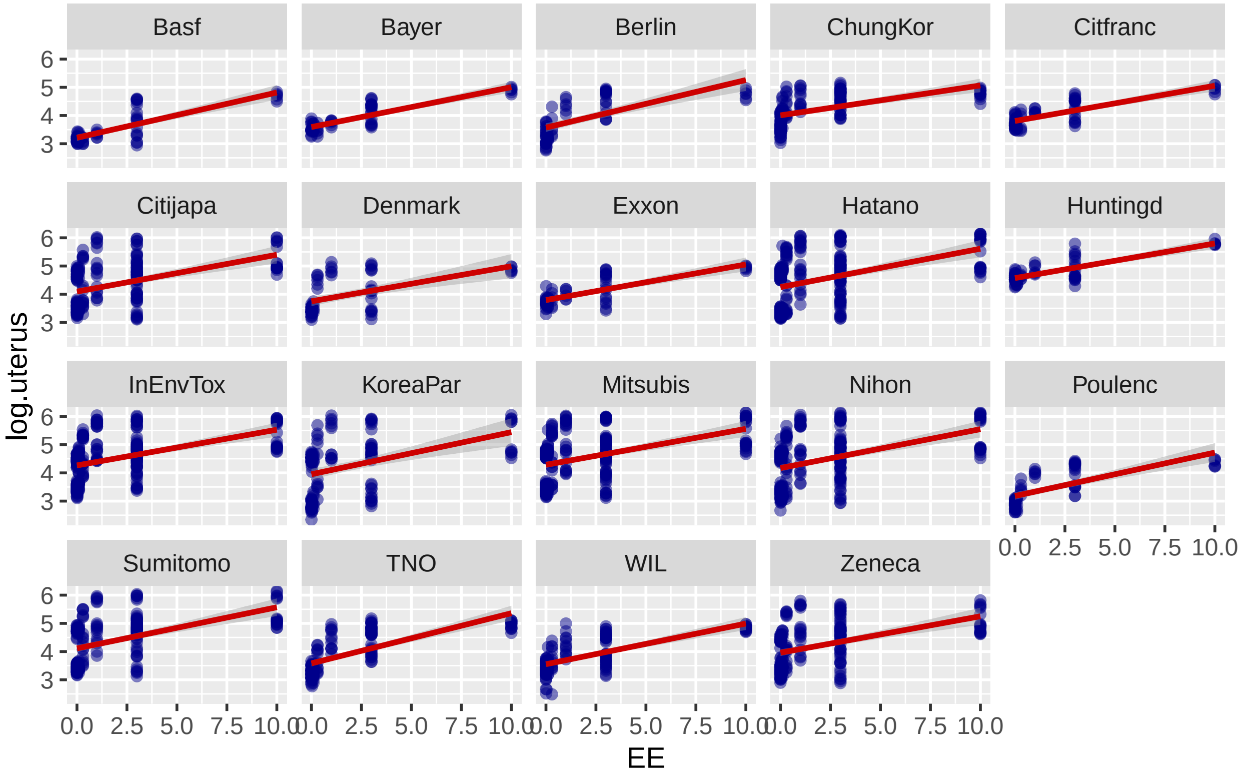
Appendix 1.1: EDA



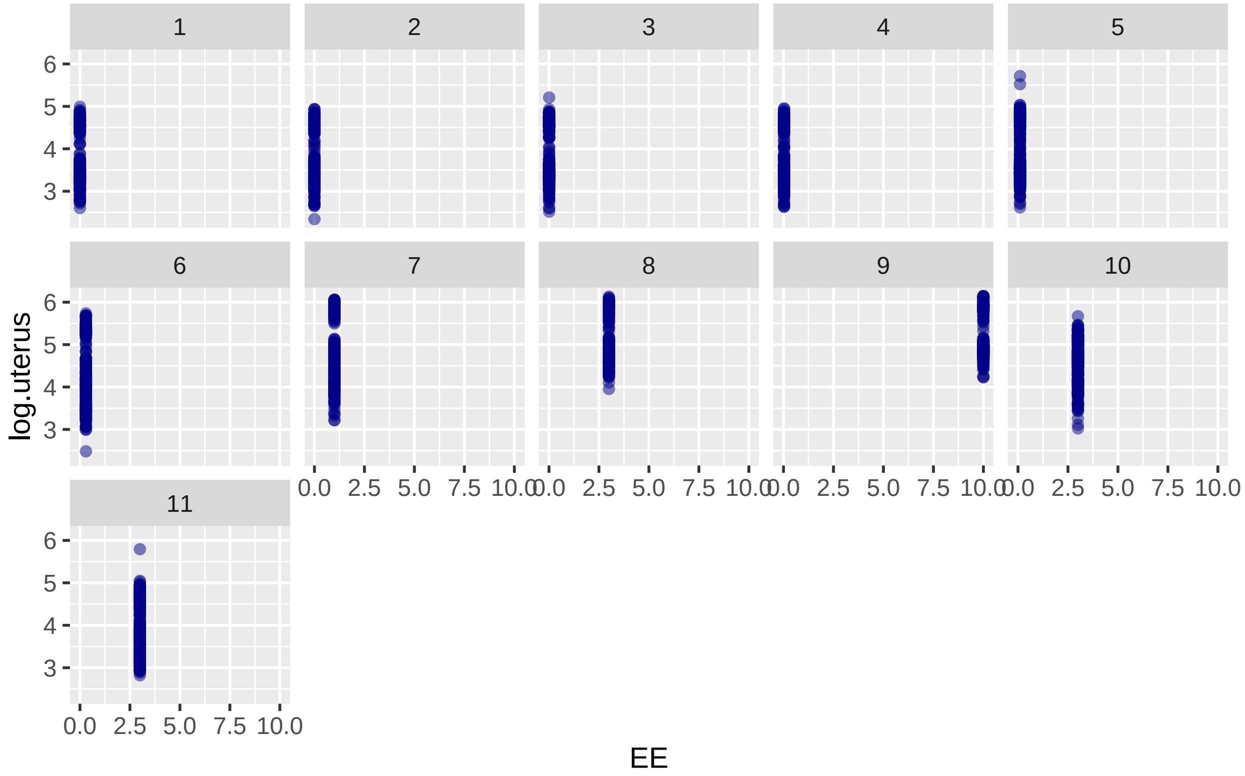
EE vs. log.uterus



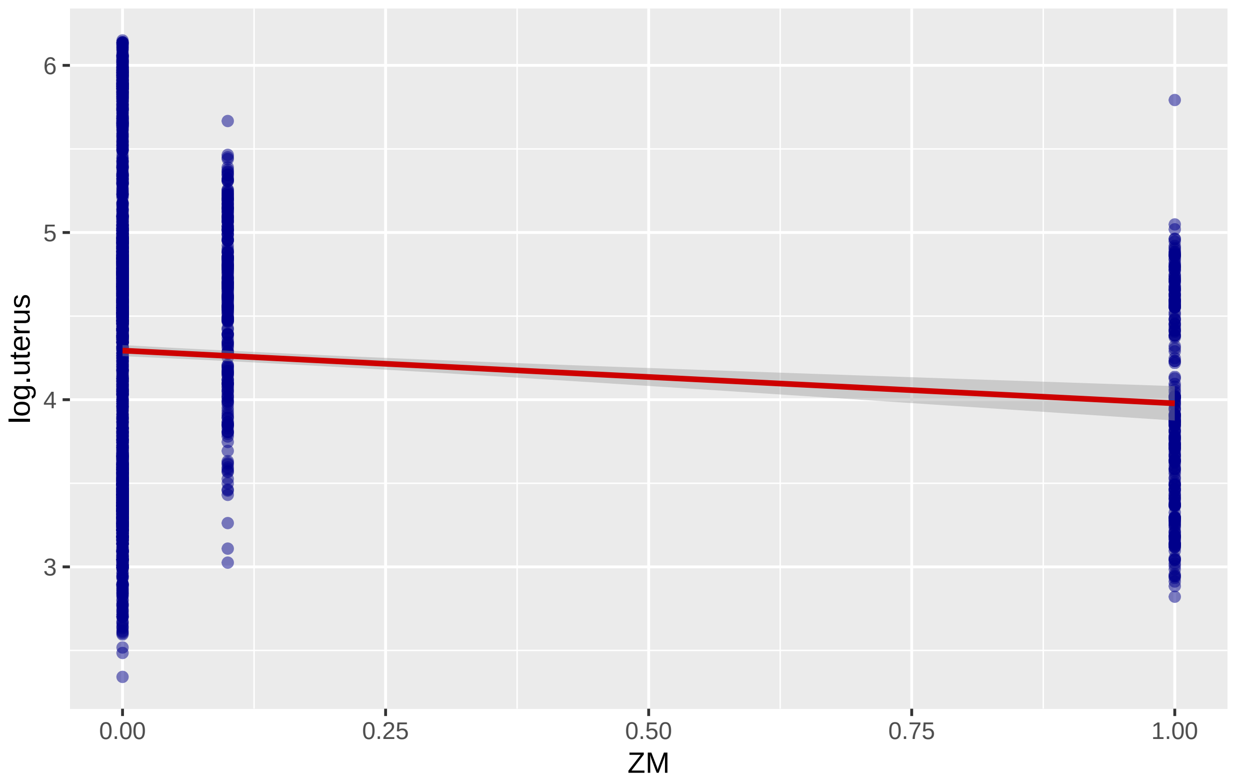
EE vs. log.uterus by protocol



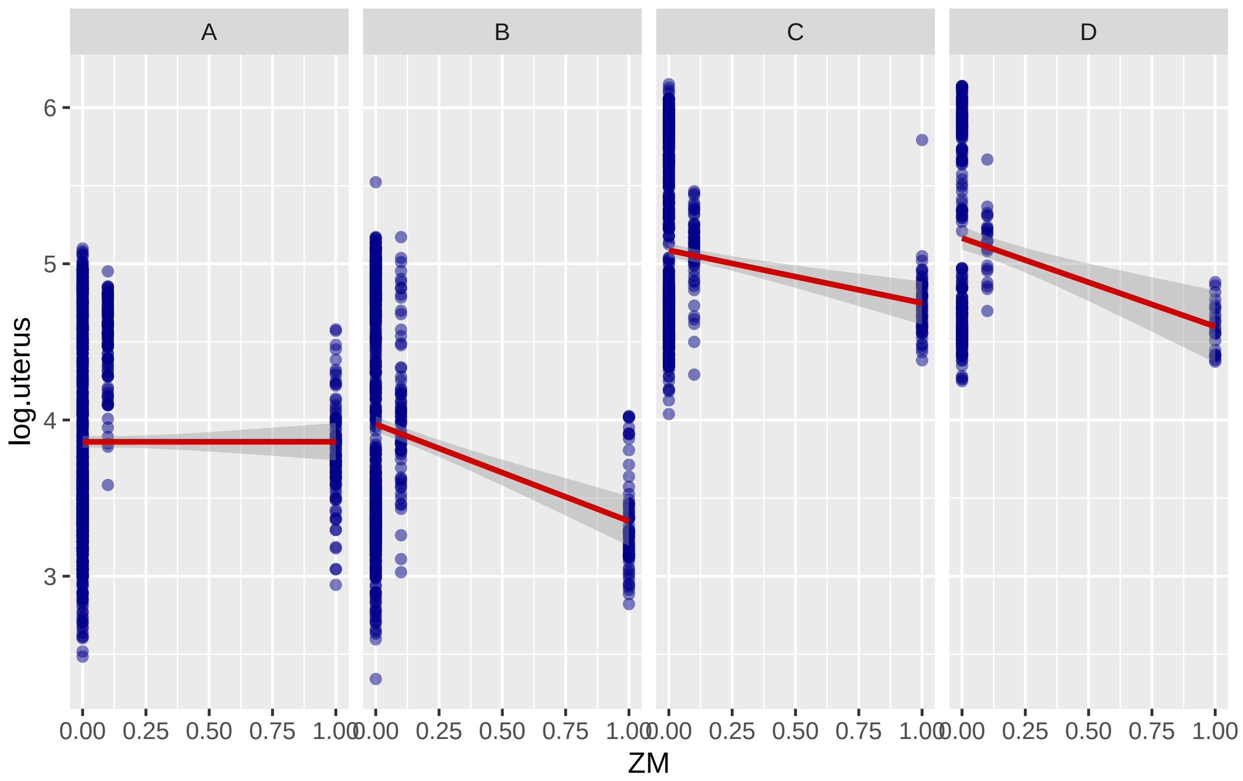
EE vs. log.uterus by protocol



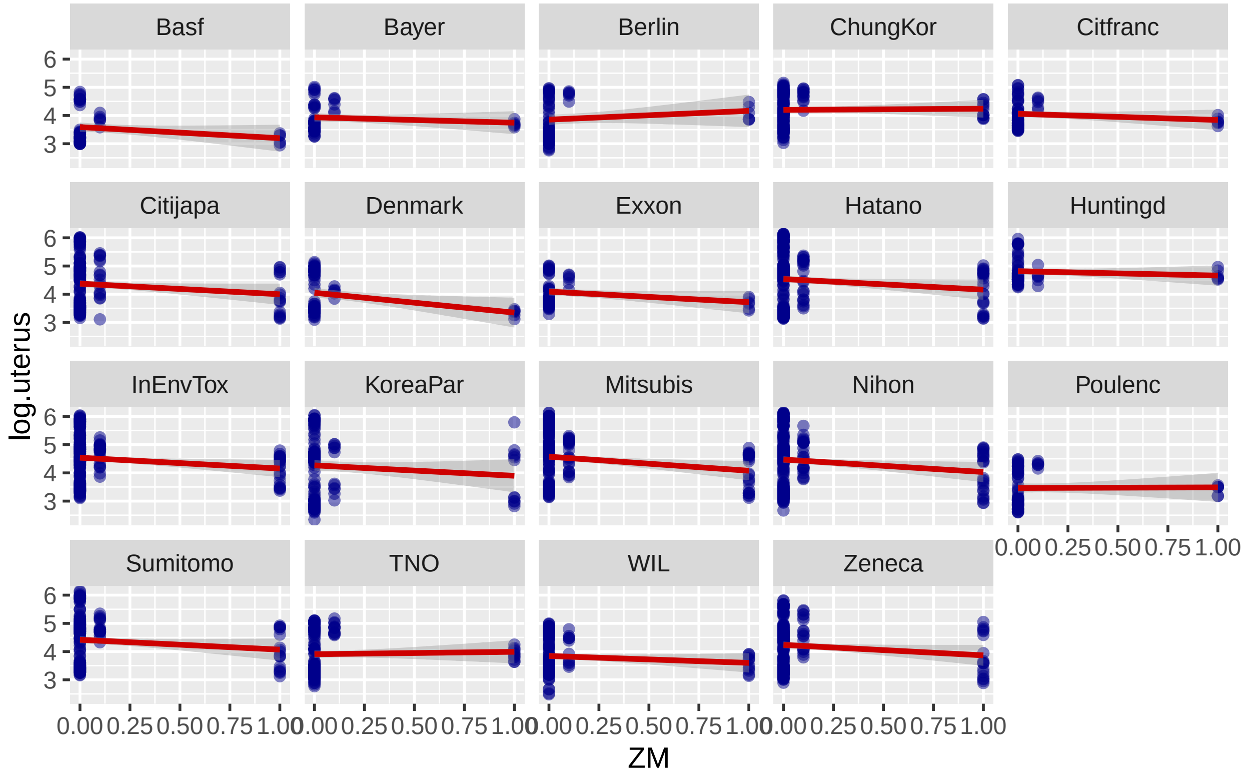
EE vs. log.uterus by group



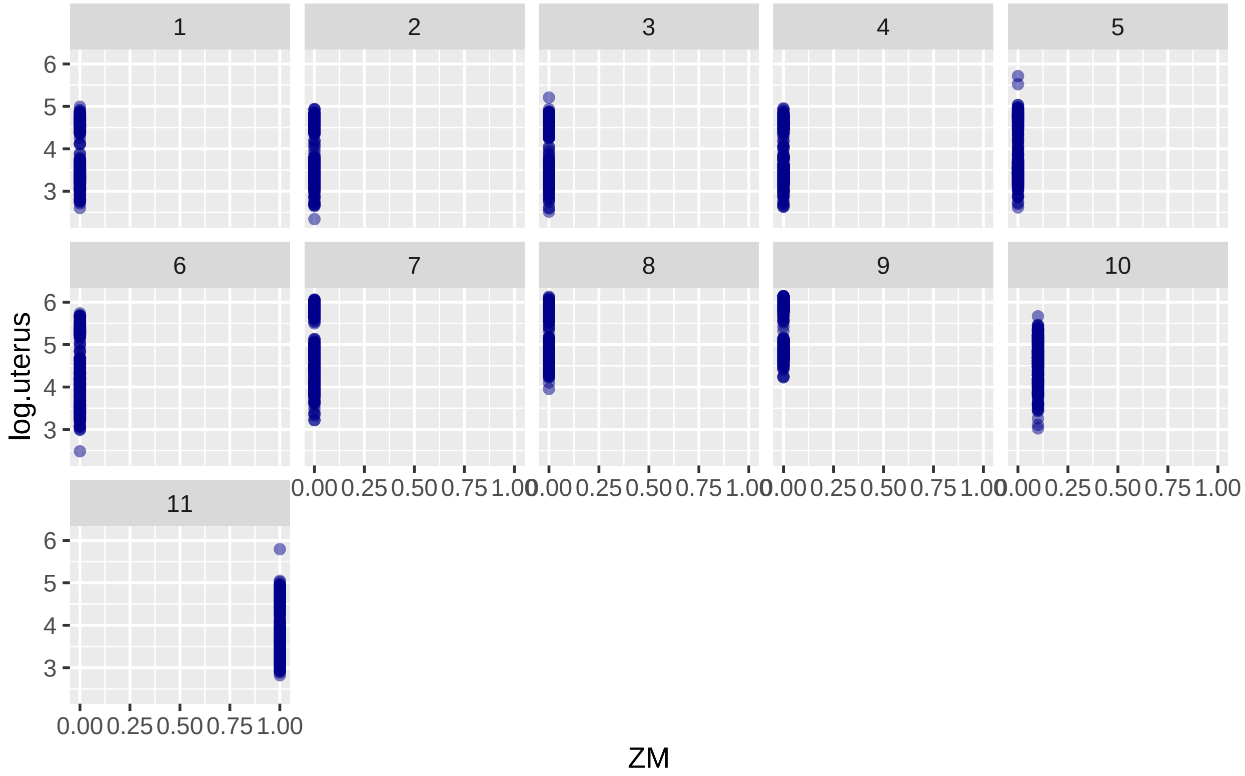
ZM vs. log.uterus



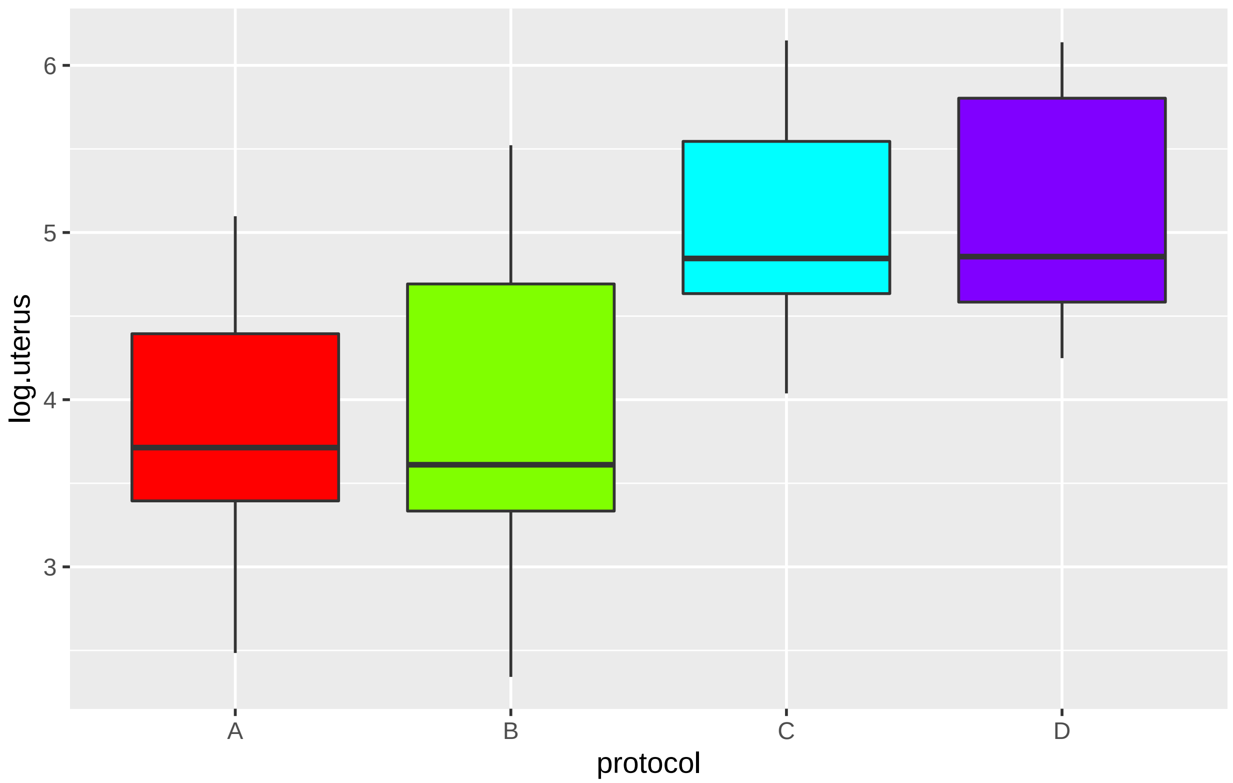
ZM vs. log.uterus by protocol



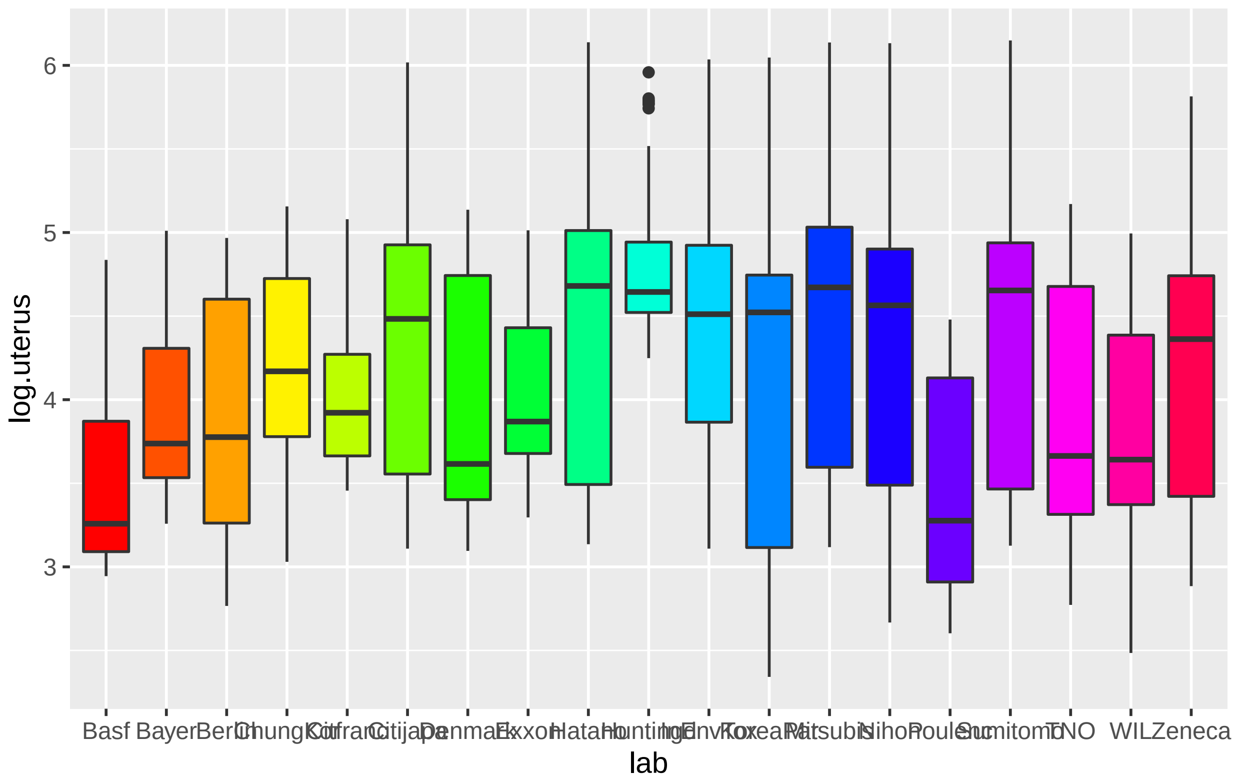
ZM vs. log.uterus by lab



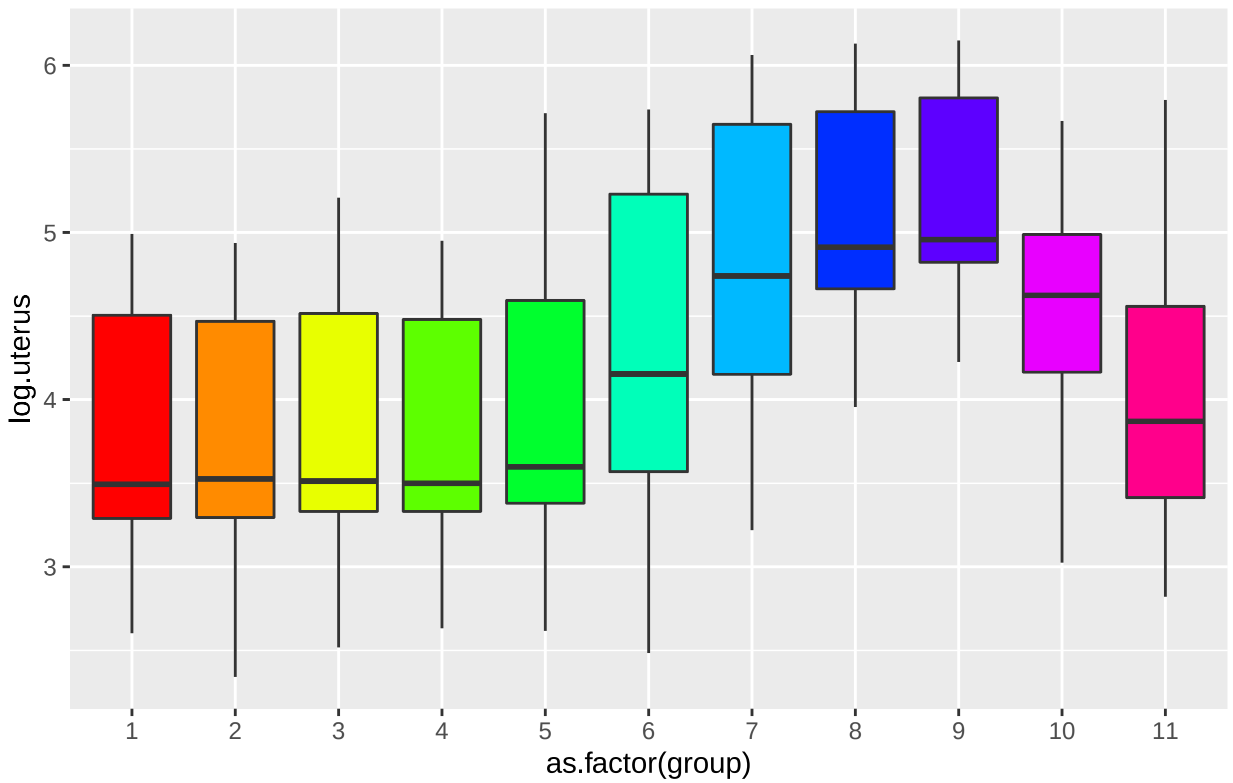
ZM vs. log.uterus by group



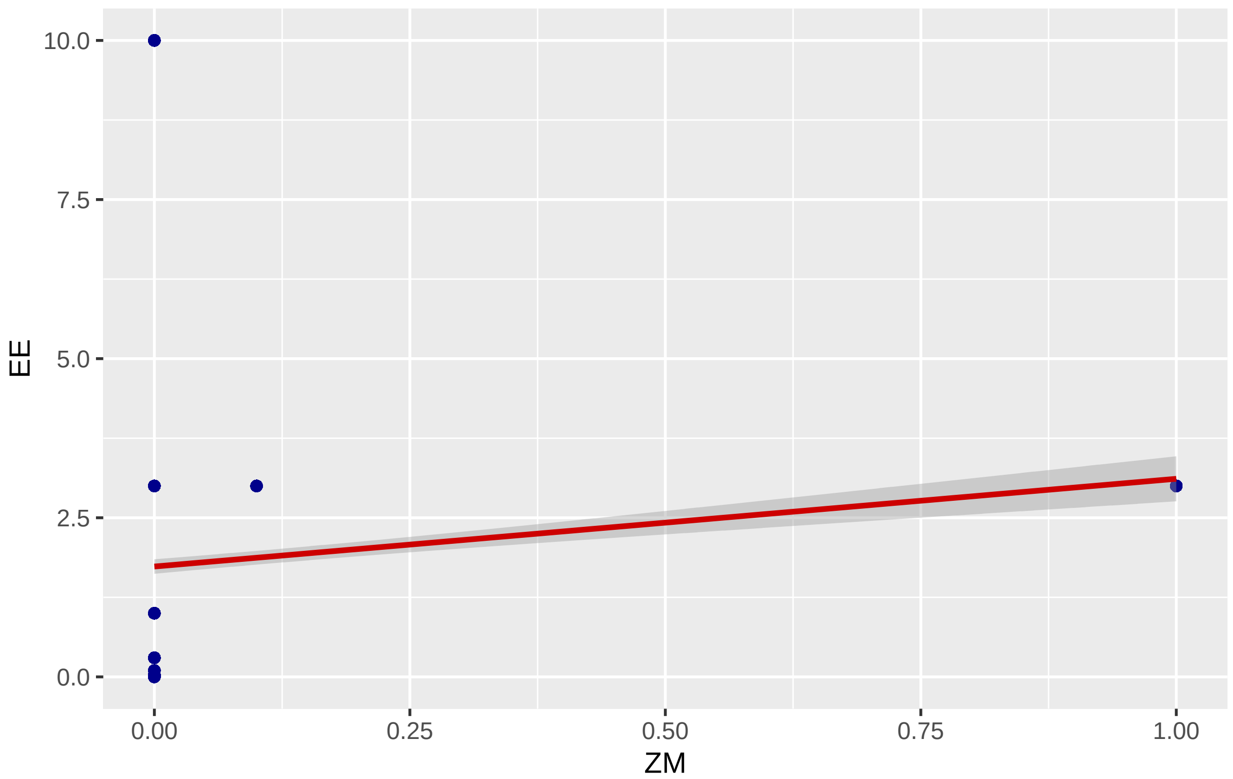
protocol vs. log.uterus



lab vs. log.uterus

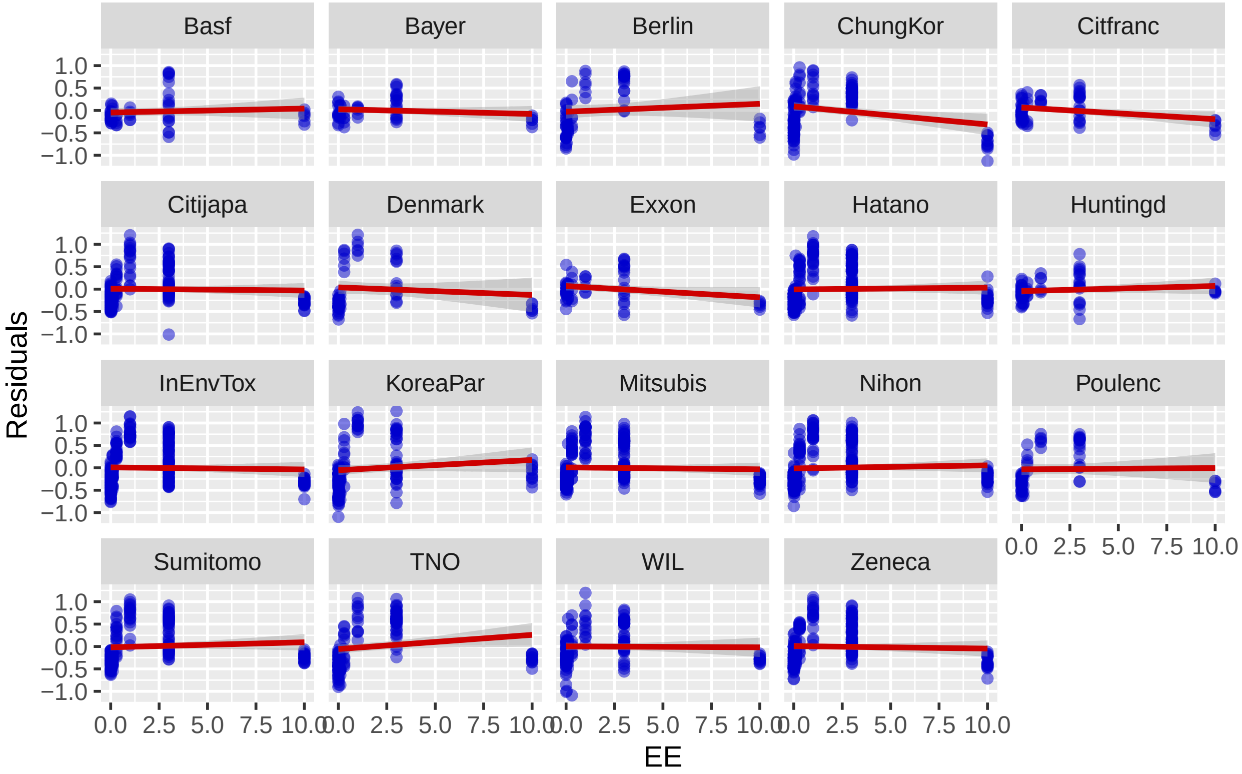


group vs. log.uterus

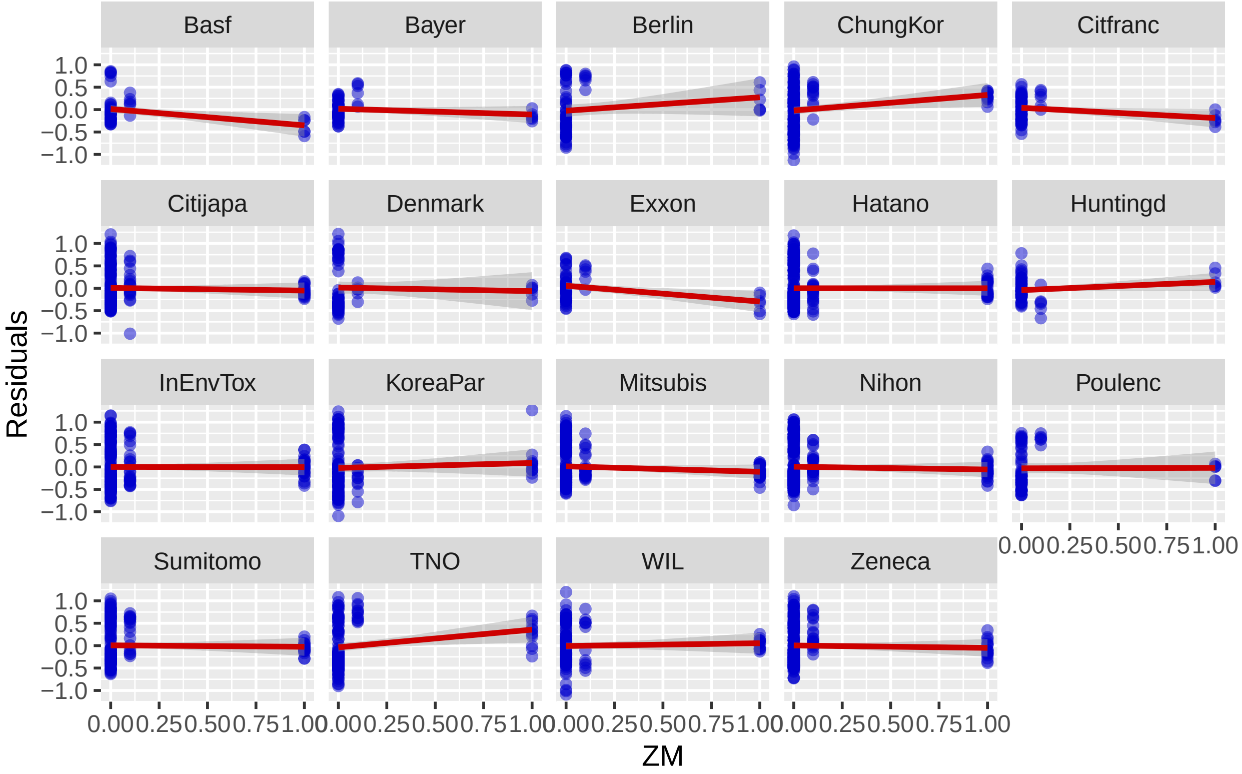


EE vs. ZM

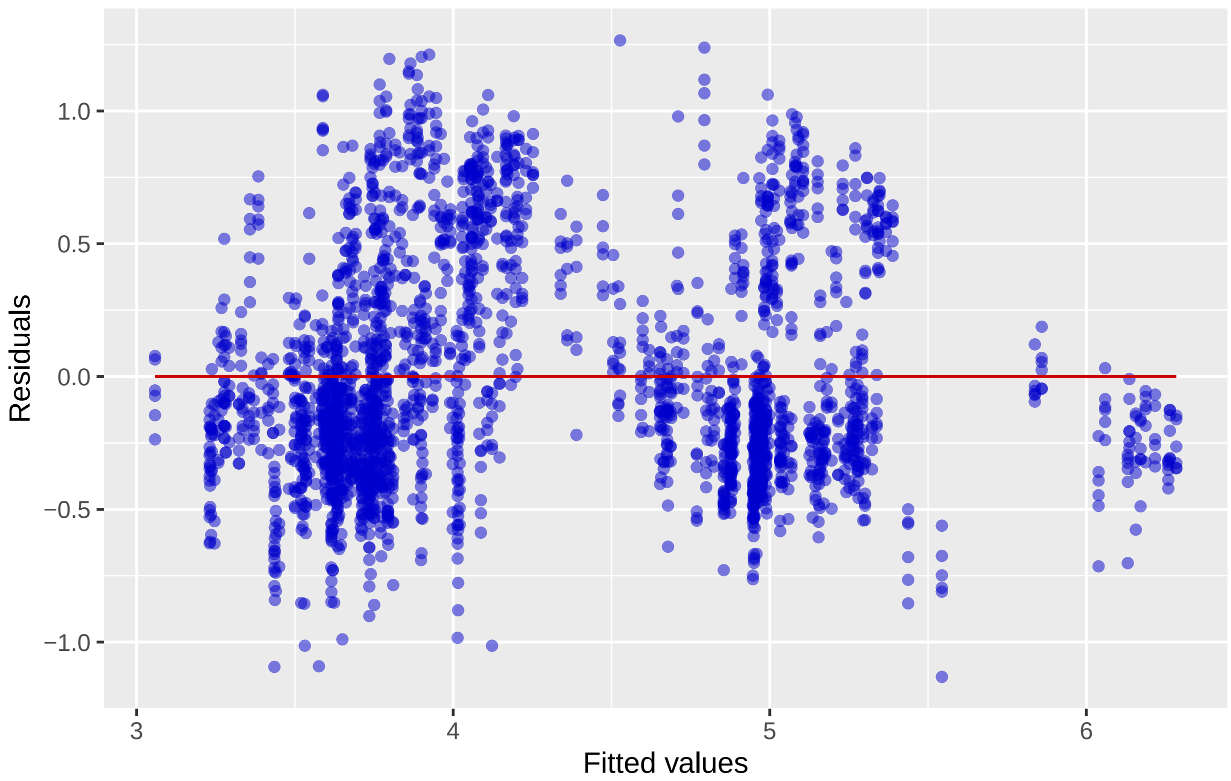
Appendix 1.2: Model Validation



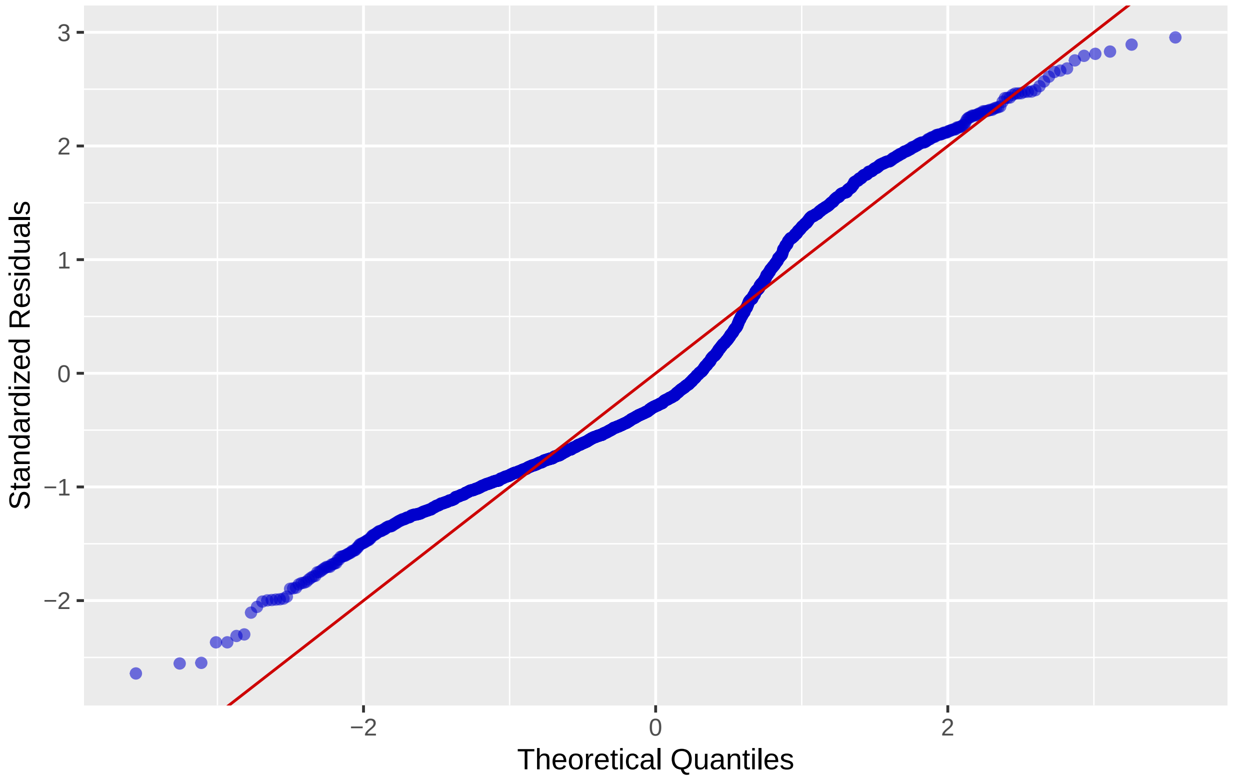
Residual vs EE



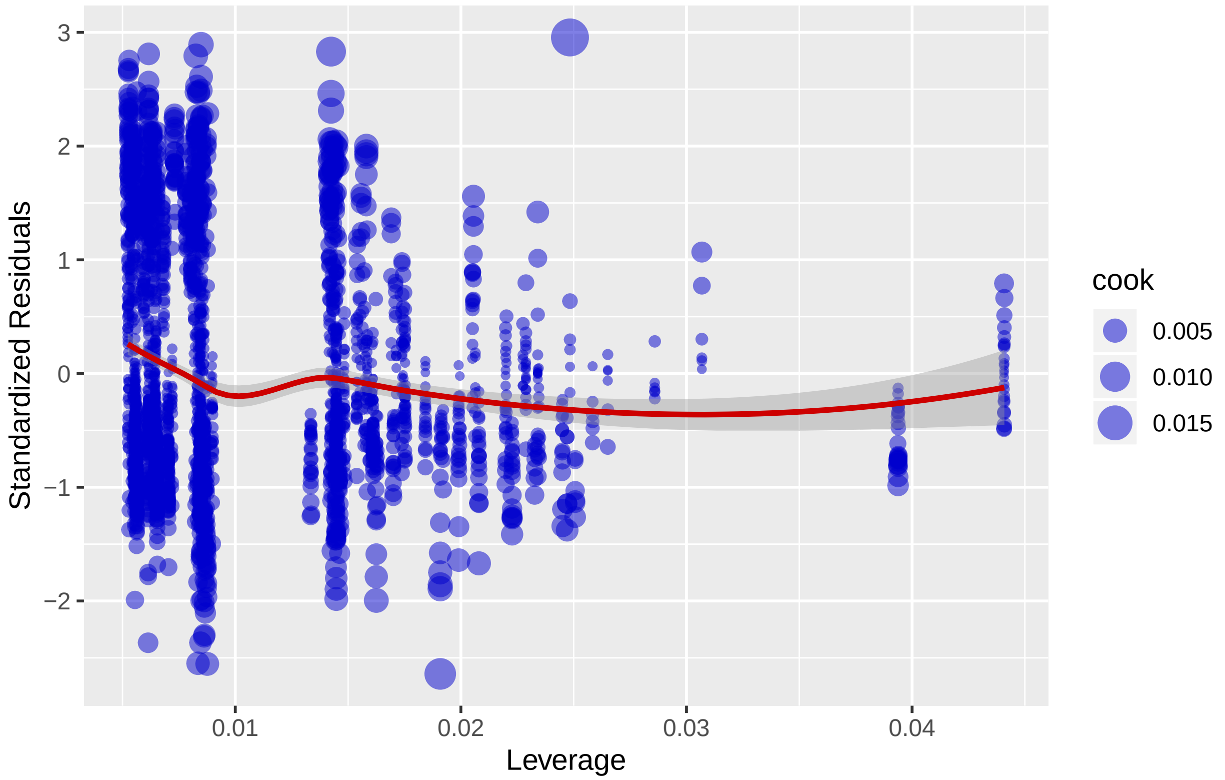
Residual vs ZM



Residuals vs Fitted values


3
2
1
0
−1
−2
Standardized Residuals
−2
0
2
Theoretical Quantiles

Normal Q−Q



Standardized Residuals vs Leverage

Appendix 2: Code

```
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE,
                      fig.pos = "H")
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(gridExtra)
library(lme4)
library(lmerTest)
library(lattice)
library(xtable)
library(tidyverse)
library(caret)
library(e1071)
library(pROC)
library(arm)
library(rms)
# Import data and convert "." in row 6 to NA - change directory as needed
#bioassay <- read.table("/Users/Calvin/Desktop/2019Fall/IDS_702_Modeling/TP_2/bioassay.txt", sep = ' ',

bioassay <- read.table("C:/Users/Varun/Documents/MIDS/Fall 2019/IDS 702 - Modeling and Representation o
                       header = TRUE, na.strings = c("NA","."))

# Remove NA rows
```

```r
bioassay <-na.omit(bioassay)

# Create numeric and transformed variables
bioassay$log.uterus <- log(bioassay$uterus)
# bioassay$sqrt.uterus <- sqrt(bioassay$uterus)
# bioassay$sqr.uterus <- bioassay$uterus^2
# bioassay$log.sqr.uterus <- bioassay$log.uterus^2
# bioassay$log.weight <- log(bioassay$weight)
# bioassay$EE.c <- bioassay$EE - mean(bioassay$EE)
# bioassay$ZM.c <- bioassay$ZM - mean(bioassay$ZM)

# Summaries
summary(bioassay)
#count(bioassay$lab)
#count(bioassay$protocol)
#count(bioassay$group)
# Determine distribution of response variable uterus weight
ut_plot <- ggplot(bioassay,aes(uterus)) +
  geom_histogram(alpha=.8,fill=rainbow(15),bins=15) +
  labs(caption="Distribution of uterus") +
  theme(plot.caption = element_text(hjust = 0.5, size = 12))

# # ggplot(bioassay,aes(uterus)) +
#   geom_histogram(alpha=.8) +
#   labs(caption="Distribution of uterus by protocol") +
#   facet_wrap(~protocol,ncol=4) +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))

# Log transformation
ut_log_plot <- ggplot(bioassay,aes(log.uterus)) +
  geom_histogram(alpha=.8,fill=rainbow(15),bins=15) +
  labs(caption="Distribution of log(uterus)") +
  theme(plot.caption = element_text(hjust = 0.5, size = 12))

grid.arrange(ut_plot, ut_log_plot,ncol = 2)

# ggplot(bioassay,aes(sqrt.uterus)) +
#   geom_histogram(alpha=.8,fill=rainbow(15),bins=15) +
#   labs(caption="Distribution of sqrt.uterus") +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))
#
# ggplot(bioassay,aes(sqr.uterus)) +
#   geom_histogram(alpha=.8,fill=rainbow(15),bins=15) +
#   labs(caption="Distribution of sq.uterus") +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))
#
# ggplot(bioassay,aes(log.sqr.uterus)) +
#   geom_histogram(alpha=.8,fill=rainbow(15),bins=15) +
#   labs(caption="Distribution of log.sq.uterus") +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))
#
# ggplot(bioassay,aes(log.uterus)) +
#   geom_histogram(alpha=.8) +
#   labs(caption="Distribution of log.uterus by protocol") +
#   facet_wrap(~protocol,ncol=4) +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))
## EE by protocol
EE_P_trend <- ggplot(bioassay,aes(x=EE, y=log.uterus)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") +
  facet_wrap(~protocol,ncol=4) +
  # labs(caption="EE vs. log.uterus by protocol") +
  theme(plot.caption = element_text(hjust = 0.5, size = 12))
EE_P_trend
# ZM by protocol
ZM_P_trend <- ggplot(bioassay,aes(x=ZM, y=log.uterus)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") +
  facet_wrap(~protocol,ncol=4) +
  # labs(caption="ZM vs. log.uterus by protocol") +
  theme(plot.caption = element_text(hjust = 0.5, size = 12))
ZM_P_trend
##weight vs uterus
# ggplot(bioassay,aes(x=weight, y=log.uterus)) +
#   geom_point(alpha = .5,colour="blue4") +
#   geom_smooth(method="lm",col="red3") +
#   labs(caption="weight vs. log.uterus") +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))
#
# ggplot(bioassay,aes(x=weight, y=log.uterus)) +
#   geom_point(alpha = .5,colour="blue4") +
#   geom_smooth(method="lm",col="red3") +
#   facet_wrap(~protocol,ncol=4) +
#   labs(caption="weight vs. log.uterus by protocol") +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))
#
# # ggplot(bioassay,aes(x=log.weight, y=log.uterus)) +
# #   geom_point(alpha = .5,colour="blue4") +
# #   geom_smooth(method="lm",col="red3") +
# #   labs(caption="log.weight vs. log.uterus") +
# #   theme(plot.caption = element_text(hjust = 0.5, size = 20))
#
# ggplot(bioassay,aes(x=weight, y=log.uterus)) +
#   geom_point(alpha = .5,colour="blue4") +
#   geom_smooth(method="lm",col="red3") +
#   facet_wrap(~protocol,ncol=4) +
#   labs(caption="weight vs. log.uterus by protocol") +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))
#
# ggplot(bioassay,aes(x=weight, y=log.uterus)) +
#   geom_point(alpha = .5,colour="blue4") +
#   geom_smooth(method="lm",col="red3") +
#   facet_wrap(~lab,ncol=5) +
#   labs(caption="weight vs. log.uterus by lab") +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))
#
```

```
# ggplot(bioassay,aes(x=weight, y=log.uterus)) +
#   geom_point(alpha = .5,colour="blue4") +
#   geom_smooth(method="lm",col="red3") +
#   facet_wrap(~group,ncol=5) +
#   labs(caption="weight vs. log.uterus by group") +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))

##EE vs. uterus
ggplot(bioassay,aes(x=EE, y=log.uterus)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") +
  labs(caption="EE vs. log.uterus") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

ggplot(bioassay,aes(x=EE, y=log.uterus)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") +
  facet_wrap(~protocol,ncol=4) +
  labs(caption="EE vs. log.uterus by protocol") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

ggplot(bioassay,aes(x=EE, y=log.uterus)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") +
  facet_wrap(~lab,ncol=5) +
  labs(caption="EE vs. log.uterus by protocol") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

ggplot(bioassay,aes(x=EE, y=log.uterus)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") +
  facet_wrap(~group,ncol=5) +
  labs(caption="EE vs. log.uterus by group") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

##ZM vs. uterus
ggplot(bioassay,aes(x=ZM, y=log.uterus)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") +
  labs(caption="ZM vs. log.uterus") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

ggplot(bioassay,aes(x=ZM, y=log.uterus)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") +
  facet_wrap(~protocol,ncol=4) +
  labs(caption="ZM vs. log.uterus by protocol") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

ggplot(bioassay,aes(x=ZM, y=log.uterus)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") +
  facet_wrap(~lab,ncol=5) +
  labs(caption="ZM vs. log.uterus by lab") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

ggplot(bioassay,aes(x=ZM, y=log.uterus)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") +
  facet_wrap(~group,ncol=5) +
  labs(caption="ZM vs. log.uterus by group") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

#protocol vs. log.uterus
ggplot(bioassay,aes(x=protocol, y=log.uterus)) +
  geom_boxplot(fill=rainbow(4)) +
  labs(caption="protocol vs. log.uterus") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

#lab vs. log.uterus
ggplot(bioassay,aes(x=lab, y=log.uterus)) +
  geom_boxplot(fill=rainbow(19)) +
  labs(caption="lab vs. log.uterus") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

#group vs. log.uterus
ggplot(bioassay,aes(x=as.factor(group), y=log.uterus)) +
  geom_boxplot(fill=rainbow(11)) +
  labs(caption="group vs. log.uterus") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

# ##Test interactions
# ggplot(bioassay,aes(x=EE, y=weight)) +
#   geom_point(alpha = .5,colour="blue4") +
#   geom_smooth(method="lm",col="red3") +
#   labs(caption="weight vs. EE") +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))
#
# ggplot(bioassay,aes(x=ZM, y=weight)) +
#   geom_point(alpha = .5,colour="blue4") +
#   geom_smooth(method="lm",col="red3") +
#   labs(caption="weight vs. ZM") +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))

ggplot(bioassay,aes(x=ZM, y=EE)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(method="lm",col="red3") +
  labs(caption="EE vs. ZM") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))
# Model 1
#model1 <- lmer(log.uterus ~ weight + EE + ZM + (1 + weight + ZM | protocol) + (1 + weight + ZM | lab)
#summary(model1)
#AIC(model1)
#BIC(model1)

# Model 2
```

```
#model2 <- lmer(log.uterus ~ weight + EE + ZM + weight:EE +(1 + weight + ZM | protocol) + (1 + weight +
#summary(model2)
#AIC(model2)
#BIC(model2)

# Compare models 1 and 2
#anova(model1,model2)

# Model 3
#model3 <- lmer(log.uterus ~ weight + EE + ZM + weight:EE + weight:ZM + (1 + weight + ZM | protocol) +
#summary(model3)
#AIC(model3)
#BIC(model3)

# Compare models 2 and 3
#anova(model2,model3)

# Test models
#model4 <- lmer(log.uterus ~ weight + protocol + EE + ZM + (1|lab), data = bioassay)
#summary(model4)
#AIC(model4)
#BIC(model4)

#model5 <- lmer(log.uterus ~ weight + EE + ZM + (1|protocol) + (1|lab), data = bioassay)
#summary(model5)
#AIC(model5)
#BIC(model5)

#model6 <- lmer(log.uterus~ weight + EE + ZM + protocol + (1|lab), data = bioassay)
#summary(model6)
#AIC(model6)
#BIC(model6)

#model7 <- lmer(log.uterus ~ EE + ZM + protocol + (1|lab), data = bioassay)
#summary(model7)
#AIC(model7)
#BIC(model7)

#anova(model6,model7)

#model8 <- lmer(log.uterus ~ EE + ZM + (1|lab) + (1|protocol), data = bioassay)
#summary(model8)
#AIC(model8)
#BIC(model8)

#model9 <- lmer(log.uterus ~ EE.c + ZM.c + (1 + ZM.c |lab) + (1 + ZM.c |protocol), data = bioassay)
#summary(model9)
#AIC(model9)
#BIC(model9)

#model10 <- lmer(log.uterus ~ EE.c + ZM.c + (1 | lab) + (1 + ZM.c |protocol), data = bioassay)
#summary(model10)
#AIC(model10)
```

```
#BIC(model10)

##Final Model
model11 <- lmer(log.uterus ~ EE + ZM + protocol + EE:protocol + ZM:protocol + (1|lab), data = bioassay)
summary(model11)
AIC(model11)
BIC(model11)

# Model 11 without interactions
model12 <- lmer(log.uterus ~ EE + ZM + protocol + (1|lab), data = bioassay)
summary(model12)
AIC(model12)
BIC(model12)

# Test interactions: significant, so use model 11
anova(model11, model12)
res <- residuals(model11)
pred <- predict(model11)

# Lab subset for linearity
#bioassay$res <- res
#bioassay_labs <- bioassay[bioassay$lab == c("Basf","Denmark"),]

#Linearity
# weight_res <- data.frame(bioassay$weight, res)
# ggplot(weight_res, aes(bioassay.weight, y=res)) +
#   geom_point(alpha = .5,colour="blue3") +
#   geom_line(y=0, col="red3") +
#   xlab("Weight") +
#   ylab("Residuals") +
#   labs(caption="Residuals vs Weight") +
#   theme(plot.caption = element_text(hjust = 0.5, size = 20))

EE_plot <- ggplot(bioassay, aes(EE, y=res)) +
  geom_point(alpha = .5,colour="blue3") +
  #geom_line(y=0, col="red3") +
  geom_smooth(method = "lm", col = "red3") +
  xlab("EE") +
  ylab("Residuals") +
  labs(caption="Residual vs EE") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))+
  facet_wrap(~lab,ncol = 5)
EE_plot

#ZM_res <- data.frame(bioassay$ZM.c, res)
ZM_plot <- ggplot(bioassay, aes(ZM, y=res)) +
  geom_point(alpha = .5,colour="blue3") +
  # geom_line(y=0, col="red3") +
  geom_smooth(method = "lm", col = "red3") +
  xlab("ZM") +
  ylab("Residuals") +
  labs(caption=("Residual vs ZM")) +
  theme(plot.caption = element_text(hjust = 0.5, size = 20)) +
  facet_wrap(~lab,ncol = 5)
ZM_plot

# Independence and Equality of Variance
pred_res <- data.frame(pred, res)
var_plot <- ggplot(pred_res, aes(pred, y=res)) +
  geom_point(alpha = .5,colour="blue3") +
  geom_line(y = 0, col = "red3") +
  #geom_smooth(method="loess",col="red3") +
  xlab("Fitted values") +
  ylab("Residuals") +
  labs(caption="Residuals vs Fitted values") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))
var_plot

# Normality
std_res <- (res - mean(res)) / sd(res)
std_res_df <- data.frame(std_res)
q_plot <- qplot(sample = std_res, data = std_res_df, color=I("blue3"), alpha=.5) +
  geom_abline(intercept = 0, slope = 1, col="red3") +
  xlab("Theoretical Quantiles") +
  ylab("Standardized Residuals") +
  labs(caption="Normal Q-Q") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20), legend.position = "none")
q_plot

# Check outlier with high influencial
cook <- cooks.distance(model11)
lev <- hatvalues(model11)
cookd <- data.frame(lev, std_res, cook)

ggplot(cookd, aes(lev, std_res)) +
  geom_point(aes(size=cook), col="blue3", alpha=.5) +
  geom_smooth(method="loess", col="red3") +
  xlab("Leverage") +
  ylab("Standardized Residuals") +
  labs(caption="Standardized Residuals vs Leverage") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

# Multicollinearity
#vif(model11)
#find outliers of labs

ranef <- ranef(model11)

lab_outlier <- data.frame(ranef)
colnames(lab_outlier)[colnames(lab_outlier)=="grp"] <- "Labs"

ggplot(lab_outlier, aes(x=Labs, y=condval)) +
  geom_point(aes(colour = Labs), size=10) +
  ylab("The random effect on the fixed intercept") +
  labs(caption="Intercepts of different labs", position="middle") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20),
        axis.title.x = element_blank())

# Dotplot
dotplot(ranef(model11, condVar = TRUE))
```

Team Project 2 Part 2: Who Voted in 2016 & Who didn't in NC

Anshupriya Srivastava, Guillem Amat, Calvin Dong, Jose Moscoso, Varun Prasad

11/3/2019

Executive Summary

- Our analysis identified that all demographic indicators in the dataset (age, ethnicity, race, gender, party) were very significant predictors of voter turnout.
- Males were less likely to vote than Females, Whites had a higher voter turnout, Hispanic & Latinos were less likely to vote than their counterparts, Republicans enjoyed the highest voter turnout across parties and Older people were more politically engaged.
- The odds and probability of voting differed significantly by county.
- The turnout differed between Males and Females across party affiliatons. Democratic Females were the most likely people to vote.

Introduction

The North Carolina State Election Board (NCSBE) is the agency responsible for overseeing the election process and reporting and compliance with campaign financing. They have online data on voter registration and turnout, among other items. The aim of this analysis is to use the NC voting records for the November 2016 general election, and try to identify/estimate classes of registered voters who voted in 2016. Also, the following questions are to be answered:

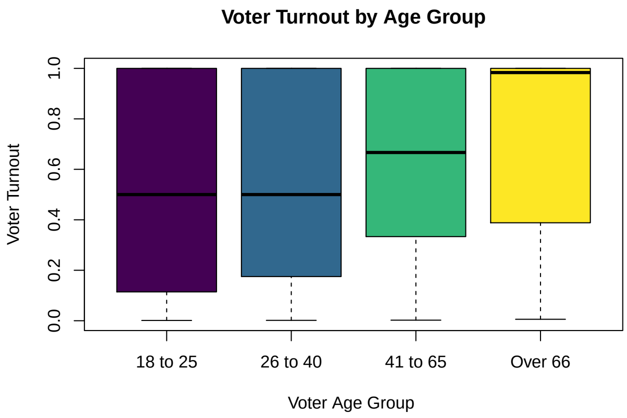
```
*How did demographic subgroups vote in 2016? For example, how did the turnout
for males compare to the turnout for females after controlling for other potential
predictors?
*Did the overall probability or odds of voting differ by county in 2016?
*How did the turnout rates differ between females and males for the different
party affiliations?
```

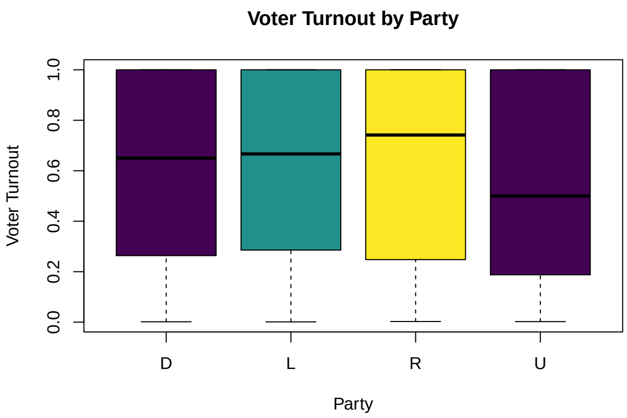
Data

The data for this exercise was made available on Sakai. The files "history_stats_20161108.txt", "voter_stats_20161108.txt," and "DataDictionaryForVoterStats.txt" have been used to answer the questions mentioned above.

Transformations have been performed on the two datasets. All duplicate rows have been dropped and only complete cases have been used. Certains columns like election date, stats_type, precinct_abbrv and vtd_abbrv have been dropped. Election Date is irrelevant for this analysis so it is dropped. Stats_type is a redundant column for this analysis since people who are registered are the only people who can vote. precinct_abbrv and vtd_abbrv have a lot of factors and so they are dropped as well. Total_voters are aggregated on county_desc, party_cd, race_code, ethnic_code, and age to get a rough idea of the demographic of the voters. There were some rows that had some anomalies. The number of people who voted seemed to be more than the number of people who were registered to vote. In these cases, the value of the total votes has been substituted by the values of the registered voters. A new column has been added to calculate voter turnout ratio. Finally, the two datasets are merged into one for further evaluation.

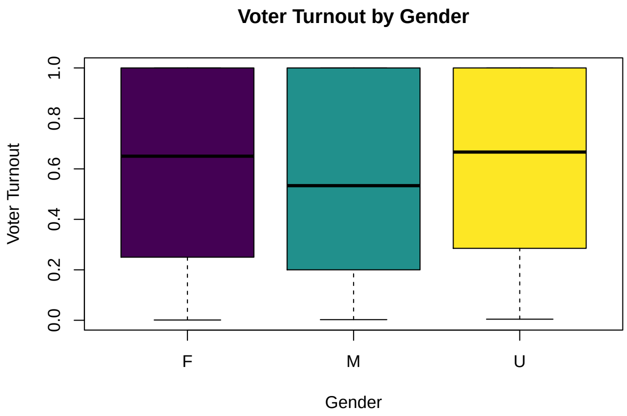
We started exploring our data for potential trends by visualizing the relationship between voter turnout among the main demographic indicators.

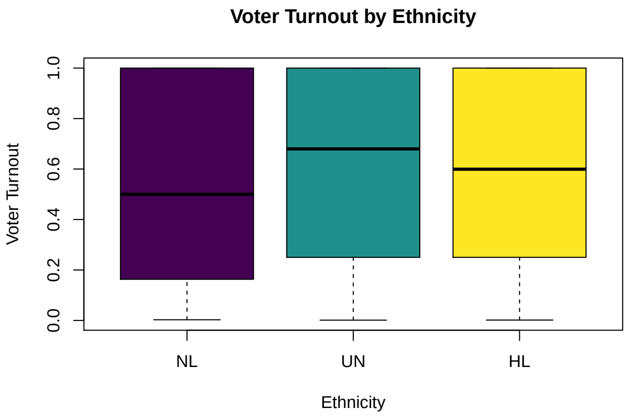




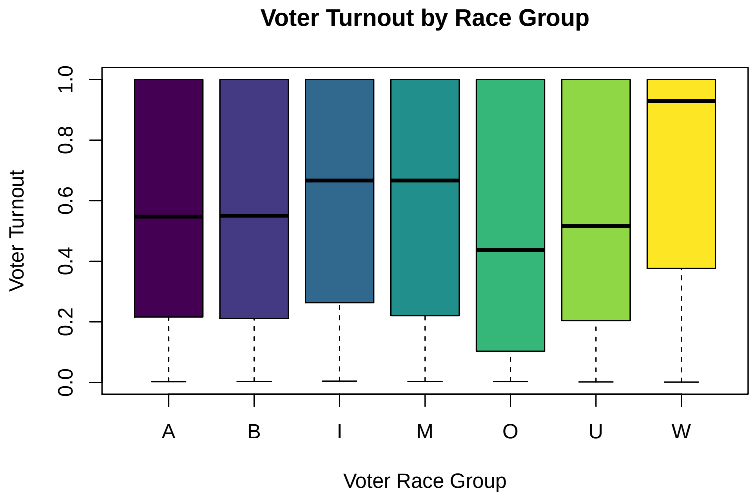
- There seems to be a significant positive relationship between voter turnout and age. The older that people get, the more engaged they become with politics.

- Voter Turnout differs a lot between the different parties. Libertarian and Republicans are much more politically than Democrats.



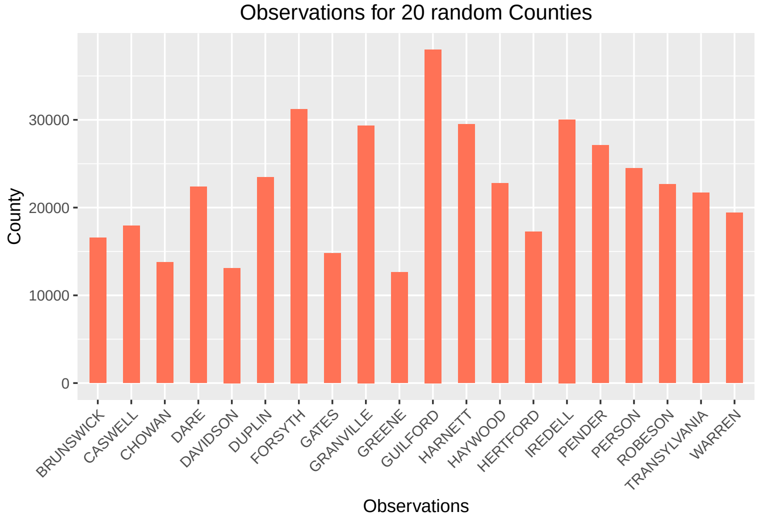


- Gender has a strong influence on voter turnout, where Females are consistently more likely to vote than their Male counterparts.
- Ethnicity visibly affects voter turnout. Hispanic and Latinos are significantly less likely to vote than other ethnicities.

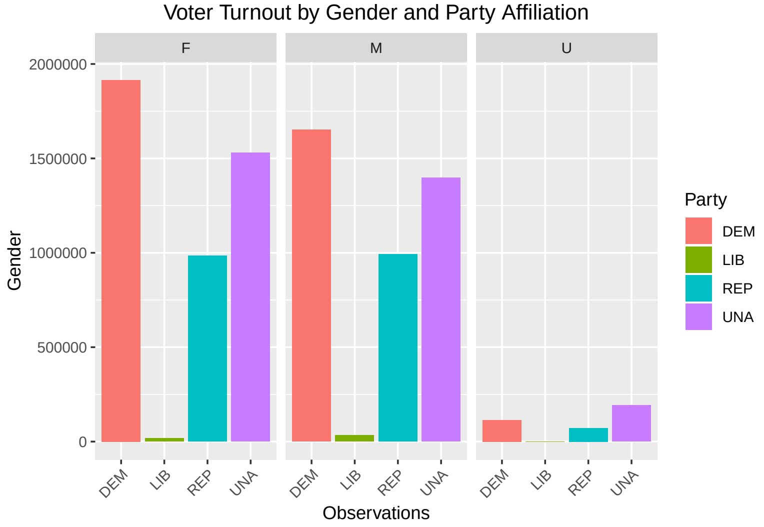


- Voting patterns among different race groups differ significantly. White people tend to vote more than other races.

The plot given below shows that there is a difference between different counties in terms of voter turnout. Guilford has the highest turnout and Greene has the lowest.



Women who are affiliated to the Democratic Party tend to vote more than the men affiliated to the same party. There doesn't seem to be much difference in terms of Libertarians. A similar trend is observed in terms of Republicans. Unaffliated female voters tend to vote more as compared to unaffliated men. People with undesignated genders have a lower overall turnout rate.



Since all of these demographic traits show different trends in voter turnout, we will consider all of them as variables when fitting our model.

Model

Because the predictor variable is categorical, we used a hierarchial logistic model that borrowed information at the counties level and from the individual voters. This model, also known as the logit function, is represented by the following equation:

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + ... + \beta_p x_{nip} + \gamma_{0j};$$

$$\gamma_{0j} \sim N(0, \sigma_0^2)$$

The logit function calculates the log-odds of the response variable due to changes in the predictor variables. Exponentiating the resulting coefficients will give the odds of a response variable depending on changes in the predictor variables. The full logistic model contained the following predictor variables: age, ethnicity, gender, race, party and interaction term between gender and party. The coefficients of this model are shown in the following table, where the "Percentage Change" column indicates the percentage change in odds of voting (after exponentiating and comparing estimates) relative to the baseline.

	Estimate	Std. Error	z value	Pr(>|z|)	Percentage Change
(Intercept)	0.1869725	0.0445044	4.2012105	0.0000265	-
sex_codeM	-0.3376945	0.0051602	-65.4417834	0.0000000	-28.6586782351945
sex_codeU	-0.2337866	0.0210141	-11.1252338	0.0000000	-20.8469317871455
party_cdLIB	-0.4222196	0.0352743	-11.9696120	0.0000000	-34.4409938618779
party_cdREP	0.0514762	0.0062232	8.2716074	0.0000000	5.28240976383483
party_cdUNA	-0.2855503	0.0056778	-50.2925933	0.0000000	-24.8399481700532
race_codeB	-0.1433579	0.0133512	-10.7374403	0.0000000	-13.355611191832
race_codeI	-0.4539728	0.0185732	-24.4423085	0.0000000	-36.4899964480443
race_codeM	-0.0359115	0.0222465	-1.6142532	0.1064725	-3.5274375640503
race_codeO	-0.3568592	0.0164538	-21.6885512	0.0000000	-30.0128975374694
race_codeU	0.0226402	0.0160345	1.4119640	0.1579606	2.28984002599202
race_codeW	0.1342595	0.0130113	10.3186943	0.0000000	14.3689532650118
ethnic_codeNL	0.0802143	0.0115724	6.9315097	0.0000000	8.35192476294624
ethnic_codeUN	-0.0073505	0.0118656	-0.6194808	0.5355997	-0.732355271871876
ageAge 26 - 40	0.2565774	0.0050696	50.6112545	0.0000000	29.2498816589416
ageAge 41 - 65	0.9849838	0.0049282	199.8667775	0.0000000	167.776846928569
ageAge Over 66	1.0354282	0.0060094	172.3017711	0.0000000	181.631180606825
sex_codeM:party_cdLIB	0.3500085	0.0458765	7.6293587	0.0000000	41.9079652974638
sex_codeU:party_cdLIB	0.6231439	0.1503697	4.1440803	0.0000341	86.4781554270042
sex_codeM:party_cdREP	0.2660440	0.0083428	31.8890672	0.0000000	30.4792522841806
sex_codeU:party_cdREP	0.2556859	0.0358324	7.1356133	0.0000000	29.1347077229631
sex_codeM:party_cdUNA	0.1931558	0.0077835	24.8161478	0.0000000	21.30717558949
sex_codeU:party_cdUNA	0.0106555	0.0246026	0.4331032	0.6649398	1.07124500819724

The model thus has the final following formula:

$$\log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \beta_{0j} + \beta_1 * sexCode_{ij} + \beta_2 * party_{ij} + \beta_3 * race_{ij} + \beta_4 * ethnicity_{ij} +$$
$$\beta_4 * age_{ij} + \beta_5 * gender : party_{ij} + \gamma_{0j}$$

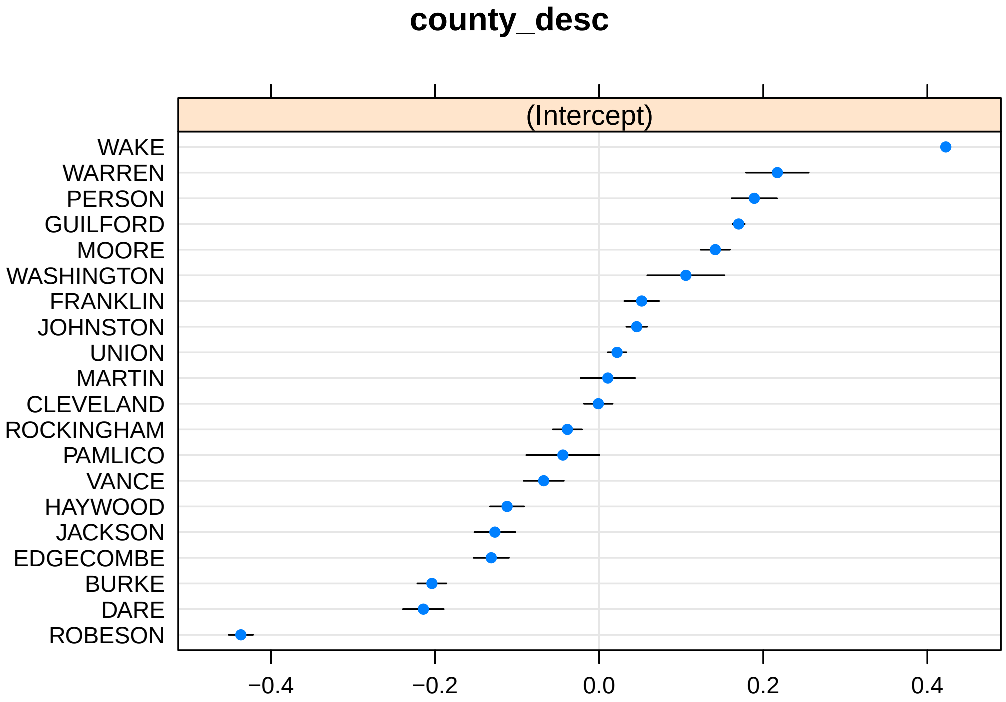
Results

From the table above, we obtain several key insights about voting odds relative to the baseline of a female, Democrat, Asian, Hispanic/Latino, 18-25 year old voter. The final model shows that men tend to vote 28.65% less than women. If we take into account the voting tendency based on party affliation it is suggested by the model that Libertarian vote 34.4% less than Democrats, Republicans vote 5% more than Democrats, and voters with no affliation tend to vote 24.83% less than voters affliated with the Democratic Party.

The confusion matrix for the model is shown below. The final model has an accuracy of **96.11%**. The sensitivity and the specificity values are 0.9322 and 0.9727 respectively.

Prediction	Reference	Freq
0	0	523659
1	0	38115
0	1	38235
1	1	1360965

The voter turnout by county varies significantly. There are two clear outliers, Wake and Robeson.



Conclusions

Overall, this valid model shows the variation of voter turnout per county and the variation in odds of voting for different demographic groups. However, there are several limitations in this model and analysis. First, only 20 counties are sampled from the entire dataset, so we are effectively assuming that they are truly representative of the entire population of North Carolina. Using a larger portion of the data would make the model more accurate. In addition, the accuracy of 96.11% indicates a high chance of overfitting, especially since we test and train on the same dataset. Therefore, the true predictive ability of this model is likely limited. Female Democrat voters have 0.8176 higher odds of voting compared to their male counterparts. Liberal Male voters have 0.023 higher odds of voting than their Female counterparts. Female Republican voters have 0.207 higher odds of voting than their Male counterparts. Unaffiliated Female voters have 0.288 higher odds of voting than their Male counterparts. This was calculated by trying different combinations of Genders based on party affiliation.

Appendix

```r
Model_voters_1 <- glmer(cbind(voters_voted, voters_registered - voters_voted) ~ age
                        + party_cd + race_code + ethnic_code + sex_code +
                          party_cd:sex_code + (1|county_desc),
                          family = binomial(link = "logit"), data = merged_data)
```