Final Report

Hannah, Jeff, Josh Lin, Pauline

12/10/2020

```
library(tidyverse)
library(readxl)
library(janitor)
library(ggplot2)
library(sandwich)
library(lmtest)
library(stargazer)
library(patchwork)
```

```
total <- read_csv("../data/interim/covid19+cusp.csv", col_names=TRUE)
```

Introduction

Since the onset of covid-19 last December, the disease has fundamentally shaken up the United States, as the US now leads the world in total number of cases and deaths. By the end of October, there were a total of 8,852,730 recorded cases and a total of 227,178 deaths. Due to the far reaching effects in the US, there have been many questions regarding what policies and individual practices help limit the spread of the disease. However, such broad questions fail to account for the unique characteristics of each state, especially when politics come into play. This discrepancy brings us to our research question.

Considering the impact of policies and politics on covid-19 transmission in the US, our research question aims to understand if the length of sheltering in place and political party affiliation in each state is associated with the spread of covid-19. For cases of covid-19 data, we examined state data provided by the Kaiser Family Foundation and covid-cases tracking data provided by the CDC. We operationalized the spread of covid-19 using positive cases per 100,000 people in each state. This counts for the varied population sizes of each state and provides a measurement for the relative success of preventing transmission of the virus. Throughout the pandemic, there has been a strong emphasis on staying at home and avoiding unnecessary contact with others to limit the transmission of the disease. We decided to focus on sheltering in place policies because we believe they are a strong indicator of the population's pandemic behavior. We operationalized this by finding the duration of each state's shelter in place policies using the initial date the stay at home order was announced and the end date of the order. Further investigating the discrepancy in policies between states and how it affected transmission, we considered which political party led the governing body in each state. This was operationalized using the current governor's political party affiliation.

We also considered the impact of race and economic status. In our second model, we operationalized economic status using the percent of the population below the poverty line, and the percent of the population that was white as an indicator of race. Because healthcare is effected by variables such as socioeconomic status and race, low income households and minorities are historically disproportionately affected by national emergencies. We further delved into the role that race played in Covid-19 transmission and operationalized race in more detail using the percentage of the population that is white, black, hispanic, or other. We operationalized "other" as: Asian, American Indian/Alaska Native, and Native Hawaiian/Other Pacific Islander.

Taking into consideration our research question and our approach to answer it, we understand that there are some limitations in our data and our work. The pandemic is ongoing, and so is the data collection

process. However, our study only looks at data up until October 2, and so the findings we report may become invalidated by incoming data, especially as we head into the winter season and cases begin to pick up once again.

We also realize that there are differences in behavior among individuals. For instance, sheltering in place might mean absolutely staying at home and going outside only for necessary errands to one individual, while another individual might be open to dining outdoors or meeting with relatives or friends outdoors in socially distanced conditions. And these different practices can be influenced by a variety of factors: political stance, religious beliefs, personal ideologies, racial/ethnic culture, etc. Therefore, we acknowledge that simply looking at the length of sheltering in place, the party affiliation of the governor, and the mandate of facemasks among other factors does not capture this variation among individual citizens.

Given the timescope of this project and the limited data available regarding mobility, state policies, and individual behavior, the following is the statistical analysis we were able to produce. There will likely be differences in levels of cases among states due to differing state policies, and since our country runs on a representative democratic system, we can argue that individual preferences and beliefs are represented by these statewide policies and politics. In addition, our question poses some implications for other issues. If some states have higher instances of cases, this could signal a larger drain on financial and healthcare resources within those states, and this information can be useful for estimating economic downturns and the national usage of healthcare resources. Using this rationale, we now turn to our model building process.

EDA

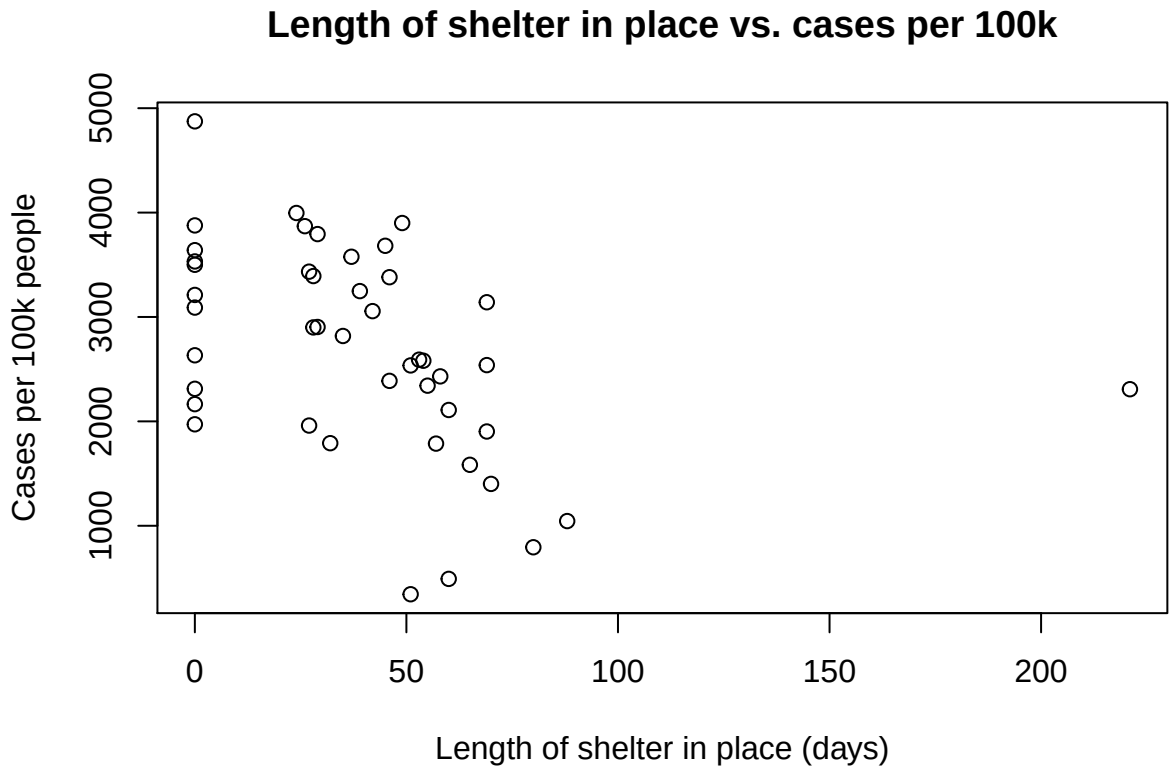
In our initial data cleaning we combined data from state policies on reopening and the general state information regarding COVID-19. It should be noted that there are multiple columns with the names: "white_percent_pop," "black_percent_pop," "hispanic_percent_pop," and "other_percent_pop." The first set of these columns represents the percentage of COVID-19 cases from the corresponding race/ethnic group. The second set of these columns that is followed by a number value in the column represents the racial composition of the total population. We calculated the length of shelter in place period for each state because this is our main indicator variable.

Since we were moving the data from excel sheets to R we had to change the date format, and convert from LINUX. We found each column with dates and used the function "excel_numeric_to_date" to make these changes.

However, there were several instances of data missing from 6 states about COVID-19 cases of specific races/ethnicities. We found the states (Hawaii, Louisiana, New Mexico, New York, North Dakota, West Virginia) with missing data, and instead of assuming values we dropped them from the data set.

Once we decided to look at the length of shelter in place as one of our main independent variables, we created a scatterplot of length of shelter in place and the number of cases per 100,000 in order to observe if there is a relationship between these two variables.

```
plot(total$`length_shelter_in_place`,
     total$cases_per_100k,
     xlab = 'Length of shelter in place (days)',
     ylab = 'Cases per 100k people',
     main = 'Length of shelter in place vs. cases per 100k')
```




```
party_model <- lm(cases_per_100k ~ length_shelter_in_place + political_party_governor,
                  data=total)
summary(party_model)
```

```
##
## Call:
## lm(formula = cases_per_100k ~ length_shelter_in_place + political_party_governor,
##     data = total)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2462.1  -473.9   180.9   624.9  1637.1
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       2785.850    289.548   9.621 3.51e-12 ***
## length_shelter_in_place             -8.447      3.870  -2.182   0.0347 *
## political_party_governorRepublican 451.062    288.319   1.564   0.1252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 895.8 on 42 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.1747
## F-statistic: 5.656 on 2 and 42 DF,  p-value: 0.006683
```

Baseline Model

We found that for each additional day of shelter in place, there were around 8 fewer Covid-19 cases per 100k people. This relationship was found to be statistically significant as indicated by the *** in our model results. With no sheltering in place policy in a Democratic state, there would be a baseline of 2785 cases per 100k people, as shown in our intercept. Even though the political party variable is not statistically significant, we found that Republican states have an additional 451 Covid-19 cases per 100k.

Comparison between Democratic and Republican States

We wanted to find the number of Democratic and Republican states to see if there was an equal number of states under each political party. To further create a visual with the variable of the political party of governors, we wanted to create a barplot to demonstrate the total number of cases in Democratic states versus Republican states. We first filtered the data to create a new data frame with information only pertaining to Democrats, and one only pertaining to Republicans. We created a new variable and summed the cases per 100,000 for Democrats and Republicans, and used this to create a barplot.

```
num_dem <- sum(total$political_party_governor == 'Democratic')
num_repub <- sum(total$political_party_governor == 'Republican')

num_dem
```

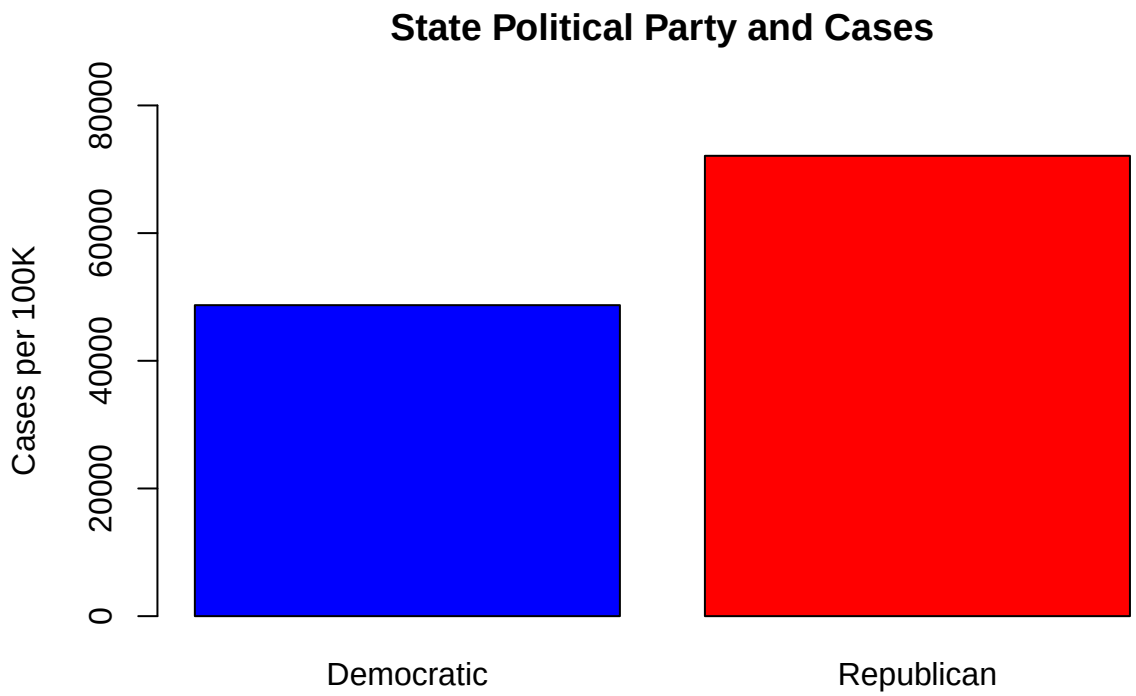
```
## [1] 21
```

```
num_repub
```

```
## [1] 24
```

In order to visualize the difference in the number of Covid-19 cases between states that had governors who were Democratic or Republic, we created a bar plot. We created a new dataframe isolating the variables of case count and political party, and found the number of cases for each political party by summing the counts of each. The number of Covid-19 cases in Democratic states was 61221, while the number of cases in Republican states was 79000. We chose the color blue to represent Democratic states and the color red to represent Republican states because these are the parties respective color.

```
total %>%
  group_by(political_party_governor) %>%
  summarise(total_pol_cases = sum(cases_per_100k)) %>%
  with(barplot(total_pol_cases, main="State Political Party and Cases",
               ylab = 'Cases per 100K',
               names.arg=c('Democratic', 'Republican'),
               col = c('blue','red'), ylim=c(0, 80000)))
```




```
poverty_model <- lm(cases_per_100k ~ length_shelter_in_place + political_party_governor +
                      percent_povertyline_2018 + as.numeric(white_percent_cases), data=total)
summary(poverty_model)
```

```
##
## Call:
## lm(formula = cases_per_100k ~ length_shelter_in_place + political_party_governor +
##     percent_povertyline_2018 + as.numeric(white_percent_cases),
##     data = total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1995.56  -617.60    10.14   551.92  1905.42
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1660.463    873.265   1.901   0.0645 .
## length_shelter_in_place             -9.062      3.817  -2.374   0.0225 *
## political_party_governorRepublican  395.490    271.534   1.457   0.1531
## percent_povertyline_2018            129.629     51.159   2.534   0.0153 *
## as.numeric(white_percent_cases)    -809.667    725.836  -1.115   0.2713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 832.9 on 40 degrees of freedom
## Multiple R-squared:  0.3513, Adjusted R-squared:  0.2864
```

```
## F-statistic: 5.415 on 4 and 40 DF,  p-value: 0.001399
```

For our Improvement v1 model, we added 2 additional features: the percent of the population under the poverty line in 2018 and the percentage of cases in the entire state population that were white. We added these additional features to control for socioeconomic status and race because based on historical trends, low income individuals and BIPOC are impacted the most by inequities in policy. For each additional day of shelter in place there is a decrease in 9 Covid-19 cases per 100k people, which is both statistically and practically significant. For each additional percent of the population living under the poverty line, there are an additional 129 cases per 100k people, which is also statistically significant. This is also practically significant because people living under the poverty line have limited access to healthcare and are more vulnerable to epidemics. Although our variable of political party is not statistically significant, there is an increase in 395 cases per 100k people for Republican states. Lastly, the variable of percentage of white cases indicates that there is a decrease in 809 cases for every 1% increase in the white population. This is likely to be practically significant because historically white Americans have not been marginalized by nationwide crises.

```r
#Improvement v2

poverty_model_v2 <- lm(cases_per_100k ~ length_shelter_in_place + political_party_governor +
                         percent_povertyline_2018 + as.numeric(white_percent_cases) +
                         as.numeric(black_percent_cases) + as.numeric(hispanic_percent_cases) +
                         as.numeric(other_percent_cases), data=total)
summary(poverty_model_v2)
```

```
##
## Call:
## lm(formula = cases_per_100k ~ length_shelter_in_place + political_party_governor +
##     percent_povertyline_2018 + as.numeric(white_percent_cases) +
##     as.numeric(black_percent_cases) + as.numeric(hispanic_percent_cases) +
##     as.numeric(other_percent_cases), data = total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1873.94  -605.82    46.65   390.73  1940.87
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         401.627   1281.530   0.313  0.75574
## length_shelter_in_place              -9.228      3.896  -2.369  0.02318 *
## political_party_governorRepublican   448.651    273.394   1.641  0.10926
## percent_povertyline_2018             175.203     60.313   2.905  0.00617 **
## as.numeric(white_percent_cases)     -107.877   1138.072  -0.095  0.92499
## as.numeric(black_percent_cases)     -454.185   1578.008  -0.288  0.77509
## as.numeric(hispanic_percent_cases)  2011.164   1487.426   1.352  0.18455
## as.numeric(other_percent_cases)    -1667.042   1967.591  -0.847  0.40230
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 824.3 on 37 degrees of freedom
## Multiple R-squared:  0.4123, Adjusted R-squared:  0.3011
## F-statistic: 3.709 on 7 and 37 DF,  p-value: 0.0039
```

For our Improvement v2 model, we added 3 additional features: the percentage of cases in the entire state population that were black, hispanic, and other. We included these features because we wanted to observe how the rates of Covid-19 affect each different racial/ethnic group. These 3 additional features are covariates of the percentage of white cases because the standard error in our Improvement v1 model accounts for these additional features. By adding these 3 variables, there is a lack of statistical significance for each

race/ethnicity feature because there is collinearity. However, our model is still robust because length of sheltering in place and percent under the poverty line are still significant. For each additional day of shelter in place there is a decrease in 9 Covid-19 cases per 100k people, which is both statistically and practically significant. For each additional percent of the population living under the poverty line, there are an additional 175 cases per 100k people, which is also statistically significant. Although our variable of political party is not statistically significant, there is an increase in 448 cases per 100k people for Republican states.

Comparing Covid-19 Cases by Ethnicity

When looking at the percentage of cases by ethnicity, there were 6 states that had NA values for certain ethnicities. We removed these states from our dataset in order to assess the percentage cases by ethnicity as comprehensively as possible.

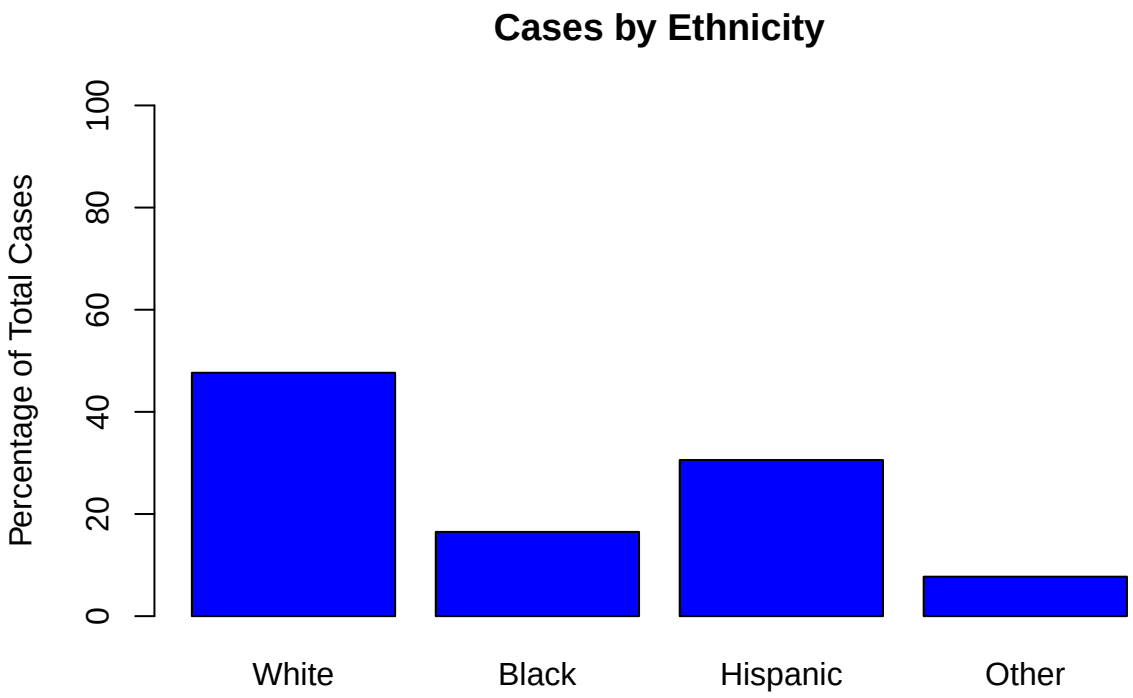
There were 2 states, Montana and Texas, which had values of "<0.1" for the percentage cases of the black population and the percentage cases of the "other" population respectively. In order to represent this value more accurately, we pulled external data from (https://dphhs.mt.gov/Portals/85/publichealth/documents/CDEpi/DiseasesAtoZ/2019-nCoV/COVID%20EPI%20PROFILE%2010162020.pdf https://www.kff.org/other/state-indicator/covid-19-cases-by-race-ethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D) and found that the percentage of cases in the black population was 0.007 in Montana and the percentage of cases in the "other" population was 0.02 in Texas.

All values in the percentage cases by ethnicity variables had to be converted to numeric values, as they were in character format from the Excel sheet where the data was pulled from.

```
#sum number of cases for each ethnicity
white <- (sum(total$total_cases*as.numeric(total$white_percent_cases)) /
            sum(total$total_cases))*100
black <- (sum(total$total_cases*as.numeric(total$black_percent_cases)) /
            sum(total$total_cases))*100
hispanic <- (sum(total$total_cases*as.numeric(total$hispanic_percent_cases)) /
               sum(total$total_cases))*100
other <- (sum(total$total_cases*as.numeric(total$other_percent_cases)) /
            sum(total$total_cases))*100
```

In order to visualize the difference in percentage of cases by each ethnicity, we created a bar plot comparing each group. To visualize the total number of cases by ethnicity, we multiplied the number of total cases per state by the percentage of cases per state for each ethnicity group. The number of cases for the white, black, hispanic, and "other" populations were 4467323, 1323546, 2536099, and 745397.2 respectively. The white population had over 3 times the number of cases that the black population had, as well as 6 times the "other" population. Although white Americans have the highest percentage of cases, our models are still practically significant because they are the racial majority in the country.

```
#making a barplot of cases for cases by ethnicity

barplot(c(white, black, hispanic, other), main="Cases by Ethnicity",
        ylab = 'Percentage of Total Cases',
        names.arg=c('White', 'Black', 'Hispanic', 'Other'),
        col = ("blue"), ylim=c(0, 100))
```



CLM Assumptions

IID Sampling (Assumption 1)

The data is not IID because the data was collected by state which clusters data geographically and different regions of the country will have similar policies, political standings, etc. Although we cannot change how data collection occurred, we can acknowledge that our data is not IID and appropriately assess results based on this information.

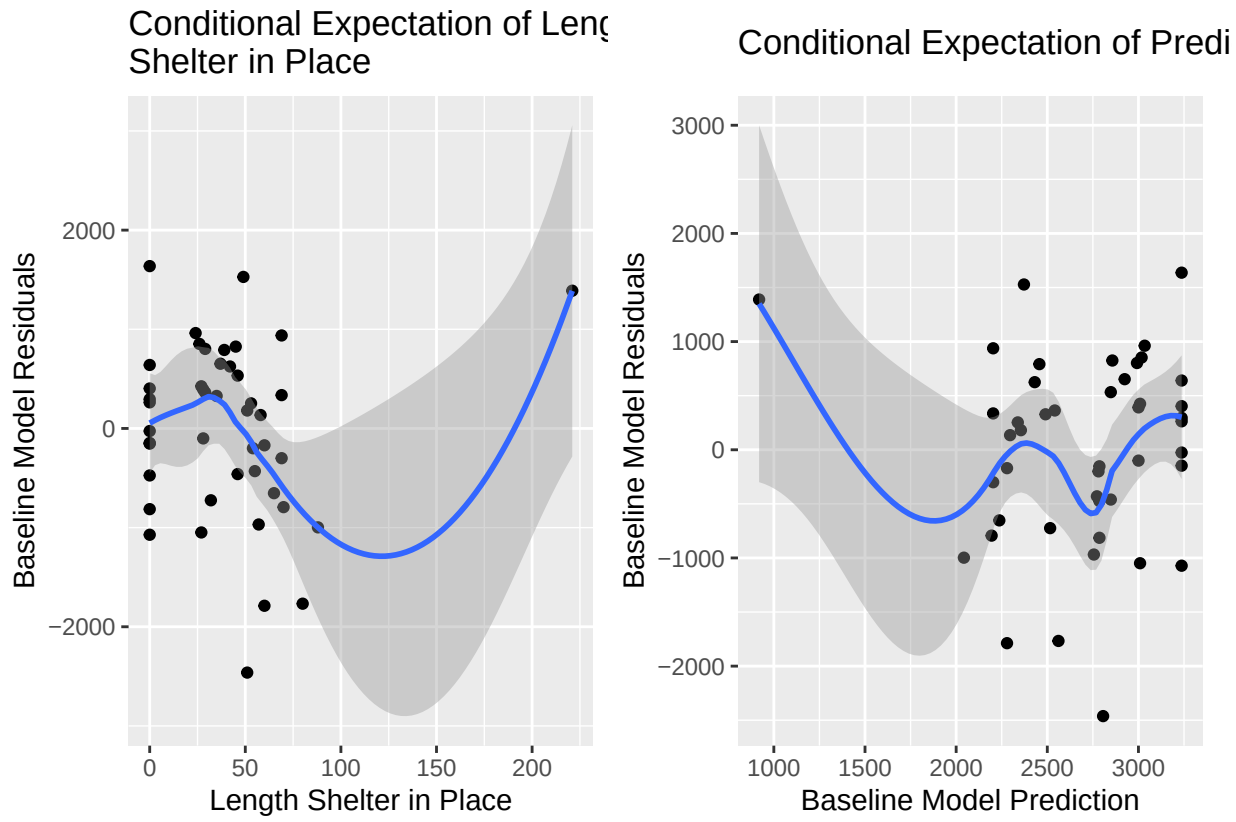
Linear Conditional Expectation (Assumption 2)

```
x_resid <- total %>%
  mutate(party_model_resid = resid(party_model),
         party_model_prediction = predict(party_model),
         poverty_model_resid = resid(poverty_model),
         poverty_model_prediction = predict(poverty_model),
         poverty_model_v2_resid = resid(poverty_model_v2),
         poverty_model_v2_prediction = predict(poverty_model_v2))

model1_plot_1 <- x_resid %>%
  ggplot(aes(x = length_shelter_in_place, y = party_model_resid)) +
  geom_point() + stat_smooth() +
  labs(x ='Length Shelter in Place', y = 'Baseline Model Residuals',
       title = 'Conditional Expectation of Length\nShelter in Place')
```

```
model1_plot_3 <- x_resid %>%
  ggplot(aes(x = party_model_prediction, y = party_model_resid)) +
  geom_point() + stat_smooth() +
  labs(x ='Baseline Model Prediction', y = 'Baseline Model Residuals',
       title = 'Conditional Expectation of Prediction')
```

```
model1_plot_1 | model1_plot_3
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



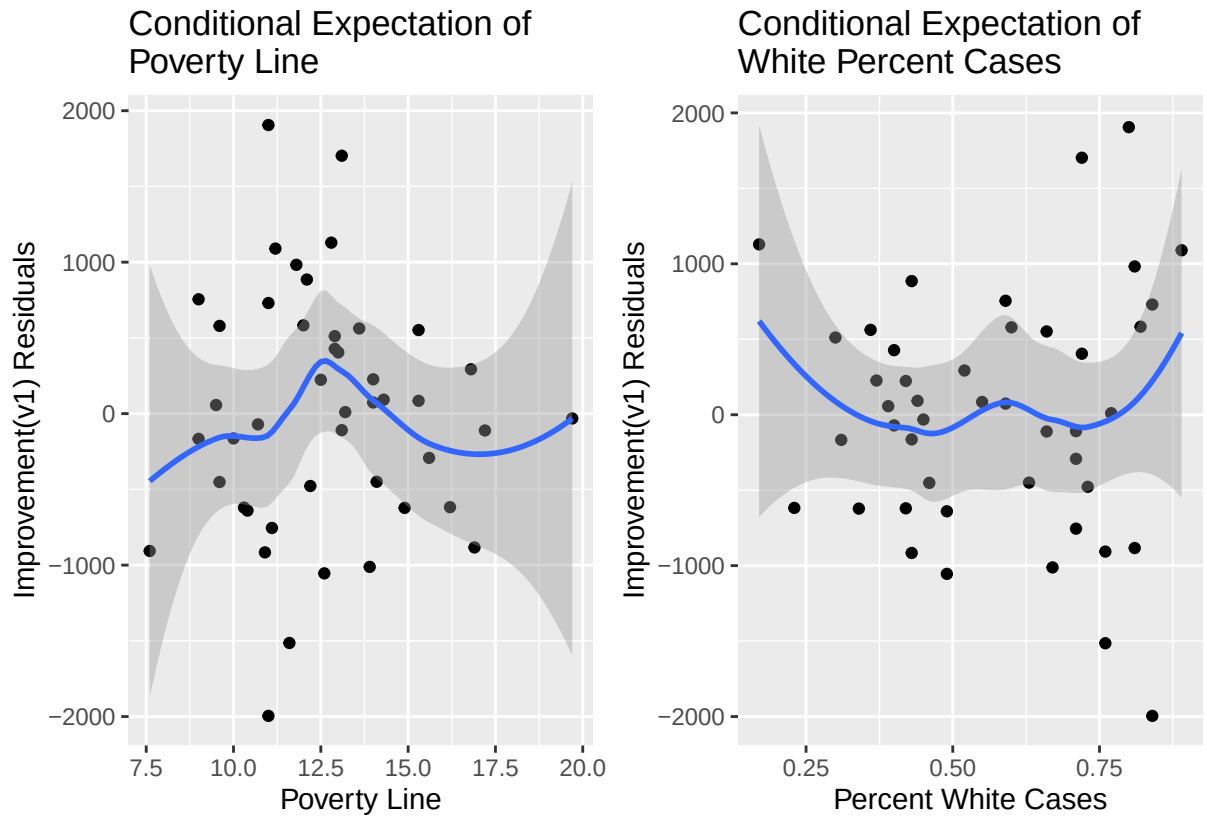
For our first model, it seems that our model may have nonlinear conditional expectation. Looking at the plots for each feature as well as the predictions, it seems that our model tends to underestimate the actual values in the dataset. We did not analyze linear conditional expectation for the party affiliation feature due to the variable's descriptive nature. However, as the baseline model only includes our key variables and does not account for additional covariates that could help improve the model, we feel that we can still learn from our model, particularly the relationship between the length of the sheltering in place mandate in each state and the rate of cases per 100,000 individuals in each state.

```
model2_plot_3 <- x_resid %>%
  ggplot(aes(x = percent_povertyline_2018, y = poverty_model_resid)) +
  geom_point() + stat_smooth() +
  labs(x ='Poverty Line', y = 'Improvement(v1) Residuals',
       title = 'Conditional Expectation of\nPoverty Line')

model2_plot_4 <- x_resid %>%
  ggplot(aes(x = as.numeric(white_percent_cases), y = poverty_model_resid)) +
```

```
  geom_point() + stat_smooth() +
  labs(x ='Percent White Cases', y = 'Improvement(v1) Residuals',
       title = 'Conditional Expectation of\nWhite Percent Cases')

model2_plot_5 <- x_resid %>%
  ggplot(aes(x = poverty_model_prediction, y = poverty_model_resid)) +
  geom_point() + stat_smooth() +
  labs(x ='Improvement(v1) Prediction', y = 'Improvement(v1) Residuals',
       title = 'Conditional Expectation of Prediction')
```

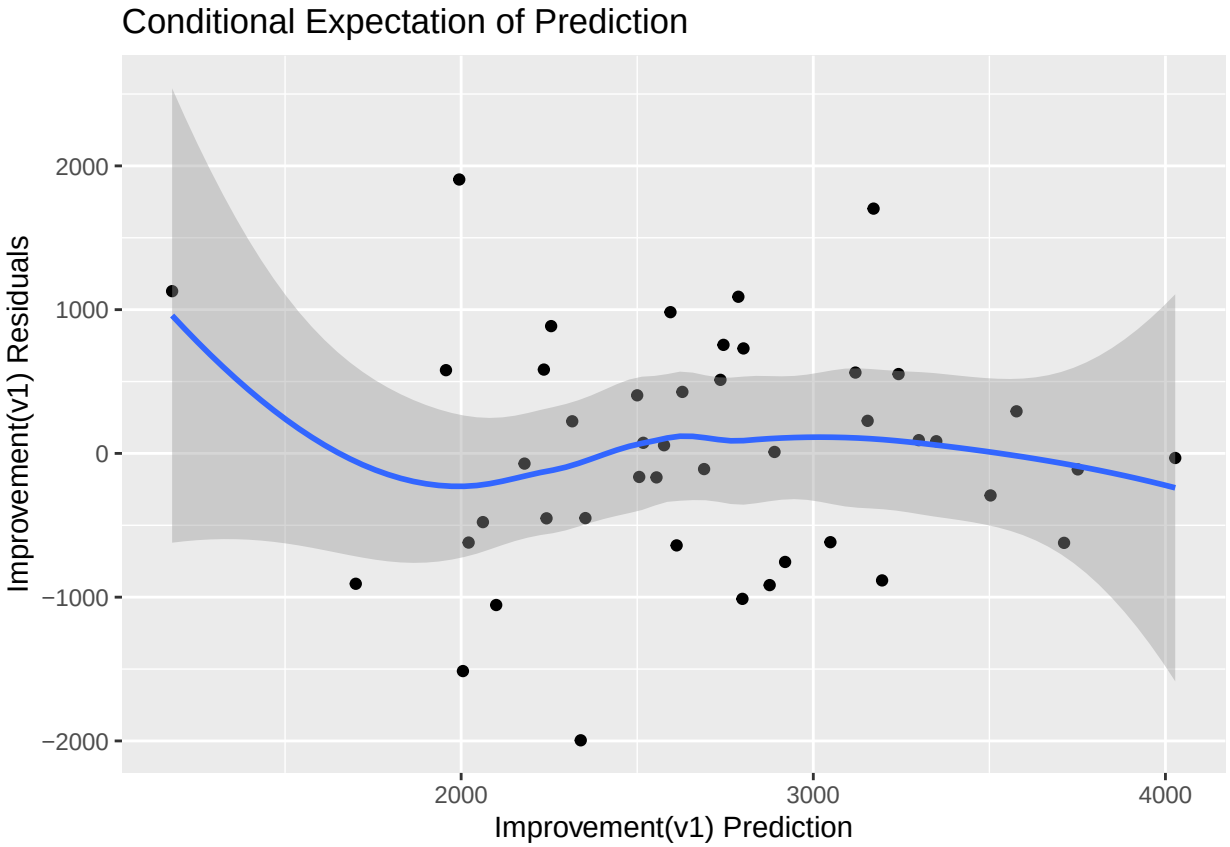
```
model2_plot_3 | model2_plot_4
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
model2_plot_5
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



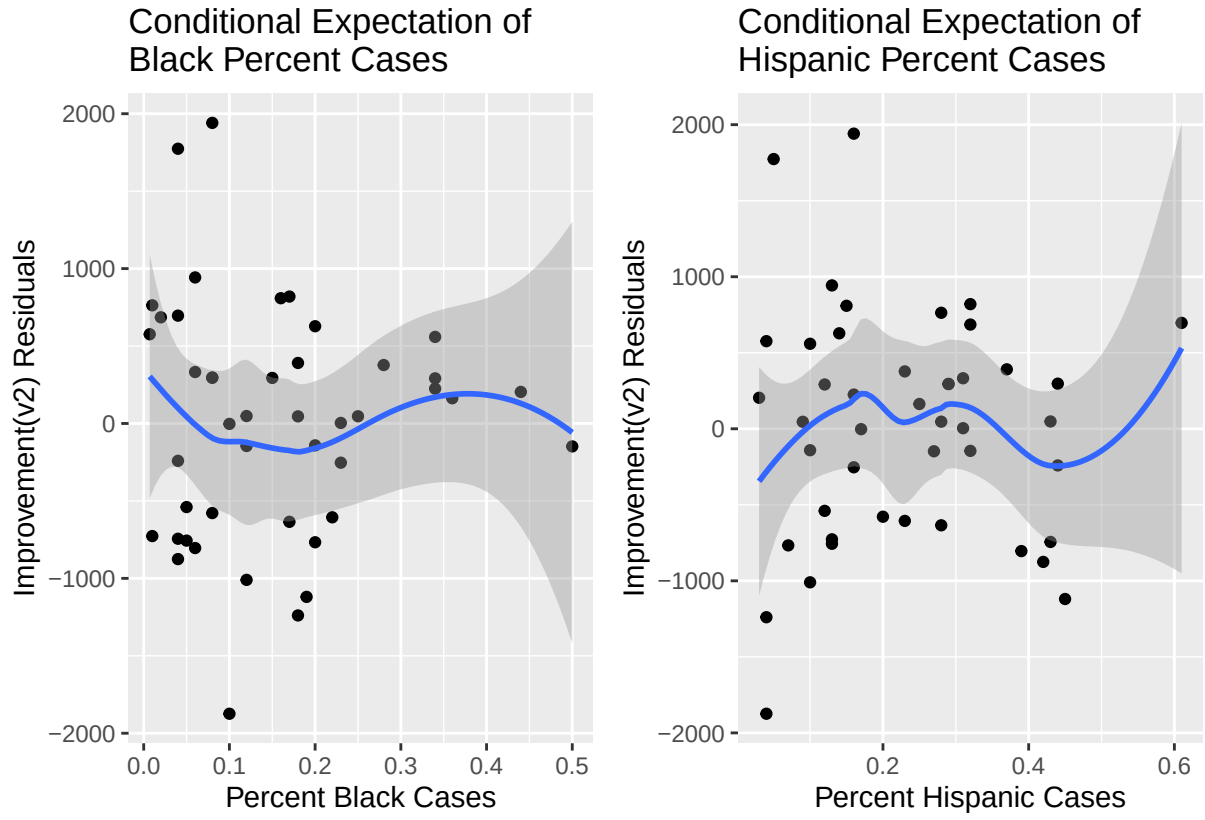
For our second model, we appear to approach linear conditional expectation with the inclusion of new variables, percentage of the population under the poverty line and the rate of covid-19 cases among the white population across each state. The new features appear relatively linear, and the distribution of the residuals versus the predictions also appears relatively linear. This indicates an improvement in our model compared to the baseline model.

```
model3_plot_5 <- x_resid %>%
  ggplot(aes(x = as.numeric(black_percent_cases), y = poverty_model_v2_resid)) +
  geom_point() + stat_smooth() +
  labs(x ='Percent Black Cases', y = 'Improvement(v2) Residuals',
       title = 'Conditional Expectation of\nBlack Percent Cases')

model3_plot_6 <- x_resid %>%
  ggplot(aes(x = as.numeric(hispanic_percent_cases), y = poverty_model_v2_resid)) +
  geom_point() + stat_smooth() +
  labs(x ='Percent Hispanic Cases', y = 'Improvement(v2) Residuals',
       title = 'Conditional Expectation of\nHispanic Percent Cases')

model3_plot_7 <- x_resid %>%
  ggplot(aes(x = as.numeric(other_percent_cases), y = poverty_model_v2_resid)) +
  geom_point() + stat_smooth() +
  labs(x ='Percent Other Cases', y = 'Improvement(v2) Residuals',
       title = 'Conditional Expectation of\nOther Percent Cases')

model3_plot_8 <- x_resid %>%
  ggplot(aes(x = poverty_model_v2_prediction, y = poverty_model_v2_resid)) +
  geom_point() + stat_smooth() +
```

```
  labs(x ='Improvement(v2) Prediction', y = 'Improvement(v2) Residuals',
       title = 'Conditional Expectation of\nPrediction')
```

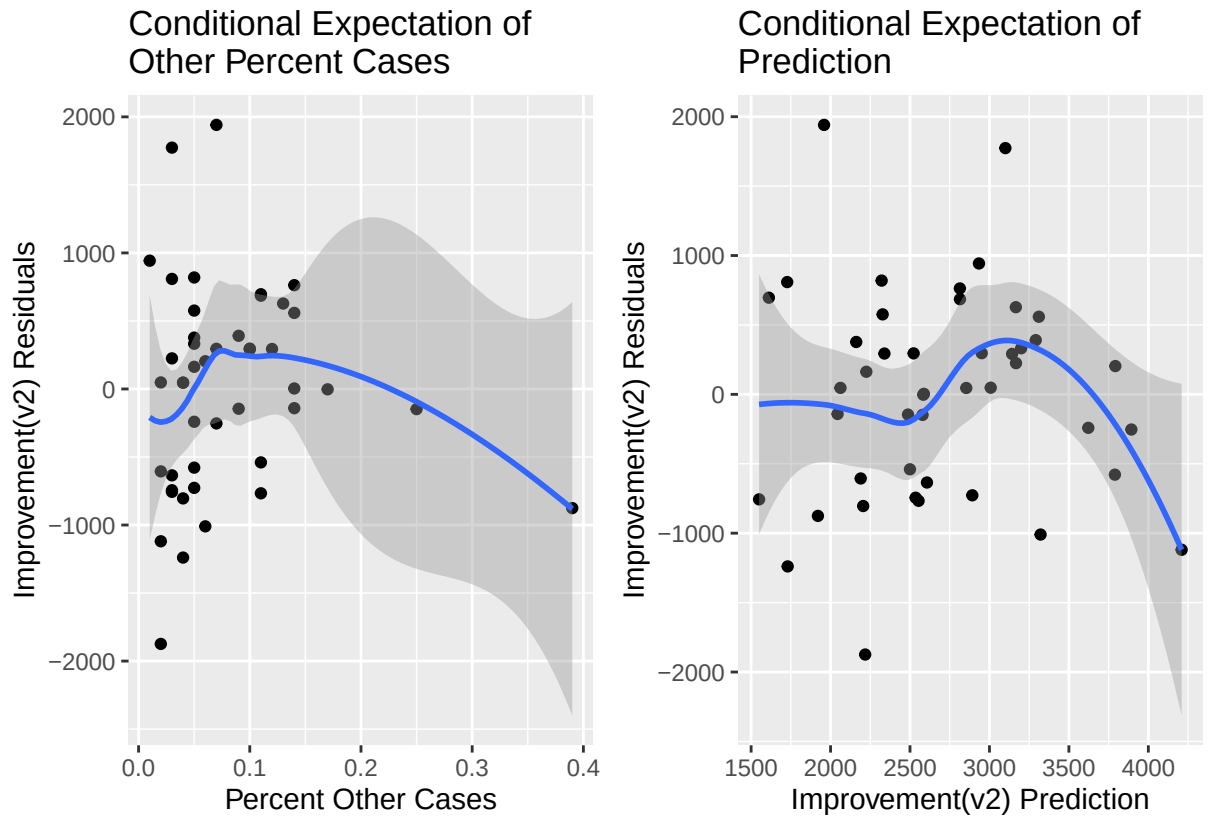
```
model3_plot_5 | model3_plot_6
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
model3_plot_7 | model3_plot_8
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



For our last model, we appear to deviate from linear conditional expectation. In this model, we added features for the percentage of cases for each ethnic group within each state: blacks, hispanics, and other races/ethnicities. However, looking at the linear conditional expectation for these additional features, it seems that the residuals are conditionally dependent on the location of the x values for each feature. This means that we do not have linear conditional expectation. Despite this deviation, this model is only testing the robustness of our second model, and it might indicate that our second model better captures the relationships within the real data.

Perfect Collinearity (Assumption 3)

```
summary(party_model)
```

```
##
## Call:
## lm(formula = cases_per_100k ~ length_shelter_in_place + political_party_governor,
##     data = total)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2462.1  -473.9   180.9   624.9  1637.1
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      2785.850    289.548   9.621 3.51e-12 ***
## length_shelter_in_place            -8.447      3.870  -2.182   0.0347 *
## political_party_governorRepublican 451.062    288.319   1.564   0.1252
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 895.8 on 42 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.1747
## F-statistic: 5.656 on 2 and 42 DF,  p-value: 0.006683
```

```
summary(poverty_model)
```

```
##
## Call:
## lm(formula = cases_per_100k ~ length_shelter_in_place + political_party_governor +
##     percent_povertyline_2018 + as.numeric(white_percent_cases),
##     data = total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1995.56  -617.60    10.14   551.92  1905.42
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1660.463    873.265   1.901   0.0645 .
## length_shelter_in_place             -9.062      3.817  -2.374   0.0225 *
## political_party_governorRepublican  395.490    271.534   1.457   0.1531
## percent_povertyline_2018            129.629     51.159   2.534   0.0153 *
## as.numeric(white_percent_cases)    -809.667    725.836  -1.115   0.2713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 832.9 on 40 degrees of freedom
## Multiple R-squared:  0.3513, Adjusted R-squared:  0.2864
## F-statistic: 5.415 on 4 and 40 DF,  p-value: 0.001399
```

```
summary(poverty_model_v2)
```

```
##
## Call:
## lm(formula = cases_per_100k ~ length_shelter_in_place + political_party_governor +
##     percent_povertyline_2018 + as.numeric(white_percent_cases) +
##     as.numeric(black_percent_cases) + as.numeric(hispanic_percent_cases) +
##     as.numeric(other_percent_cases), data = total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1873.94  -605.82    46.65   390.73  1940.87
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         401.627   1281.530   0.313  0.75574
## length_shelter_in_place              -9.228      3.896  -2.369  0.02318 *
## political_party_governorRepublican   448.651    273.394   1.641  0.10926
## percent_povertyline_2018             175.203     60.313   2.905  0.00617 **
## as.numeric(white_percent_cases)     -107.877   1138.072  -0.095  0.92499
## as.numeric(black_percent_cases)     -454.185   1578.008  -0.288  0.77509
## as.numeric(hispanic_percent_cases)  2011.164   1487.426   1.352  0.18455
## as.numeric(other_percent_cases)    -1667.042   1967.591  -0.847  0.40230
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 824.3 on 37 degrees of freedom
## Multiple R-squared:  0.4123, Adjusted R-squared:  0.3011
## F-statistic: 3.709 on 7 and 37 DF,  p-value: 0.0039
```

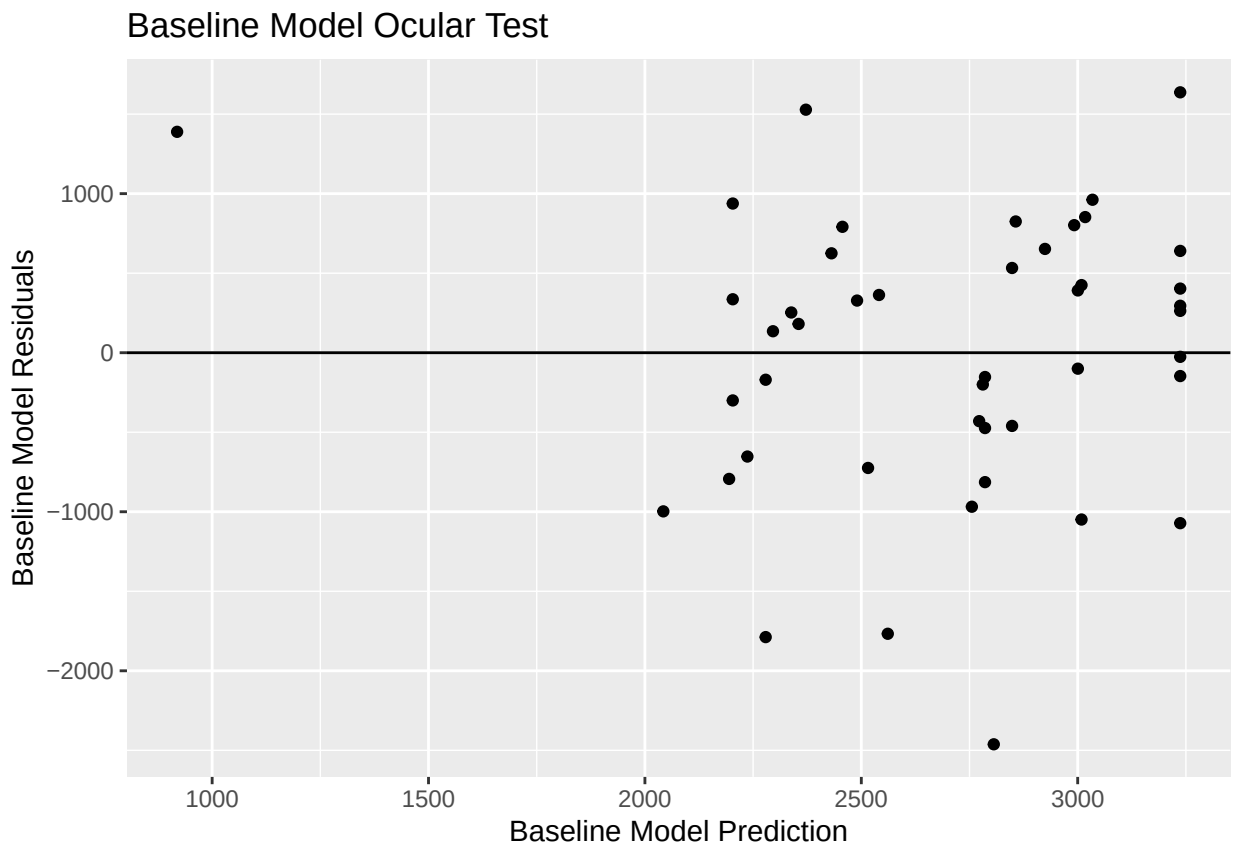
There does not appear to have colinearity or near colinearity from our data which means we do not violate the assumption. The varibles we chose are appropriate for building our models.

Homoskedastic Conditional Variance (Assumption 4)

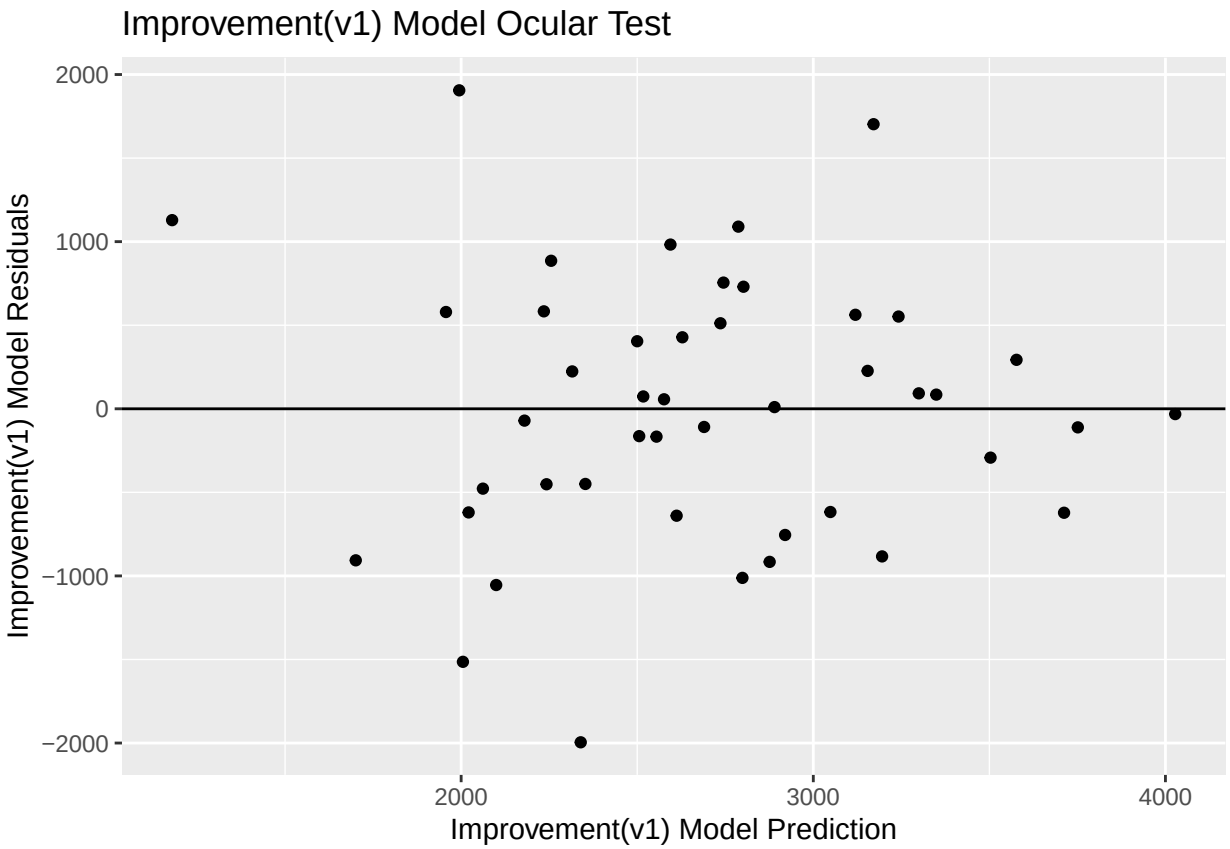
Ocular test

```
plot_4 <- x_resid %>%
  ggplot(aes(x = party_model_prediction, y = party_model_resid)) +
  geom_point() + geom_hline(yintercept = mean(x_resid$party_model_resid)) +
  labs(x ='Baseline Model Prediction', y = 'Baseline Model Residuals',
       title = 'Baseline Model Ocular Test')

plot_5 <- x_resid %>%
  ggplot(aes(x = poverty_model_prediction, y = poverty_model_resid)) +
  geom_point() + geom_hline(yintercept = mean(x_resid$poverty_model_resid)) +
  labs(x ='Improvement(v1) Model Prediction', y = 'Improvement(v1) Model Residuals',
       title = 'Improvement(v1) Model Ocular Test')

plot_6 <- x_resid %>%
  ggplot(aes(x = poverty_model_v2_prediction, y = poverty_model_v2_resid)) +
  geom_point() + geom_hline(yintercept = mean(x_resid$poverty_model_v2_resid)) +
  labs(x ='Imrpovement(v2) Model Prediction', y = 'Improvement(v2) Model Residuals',
       title = 'Improvement(v2) Model Ocular Test')
```

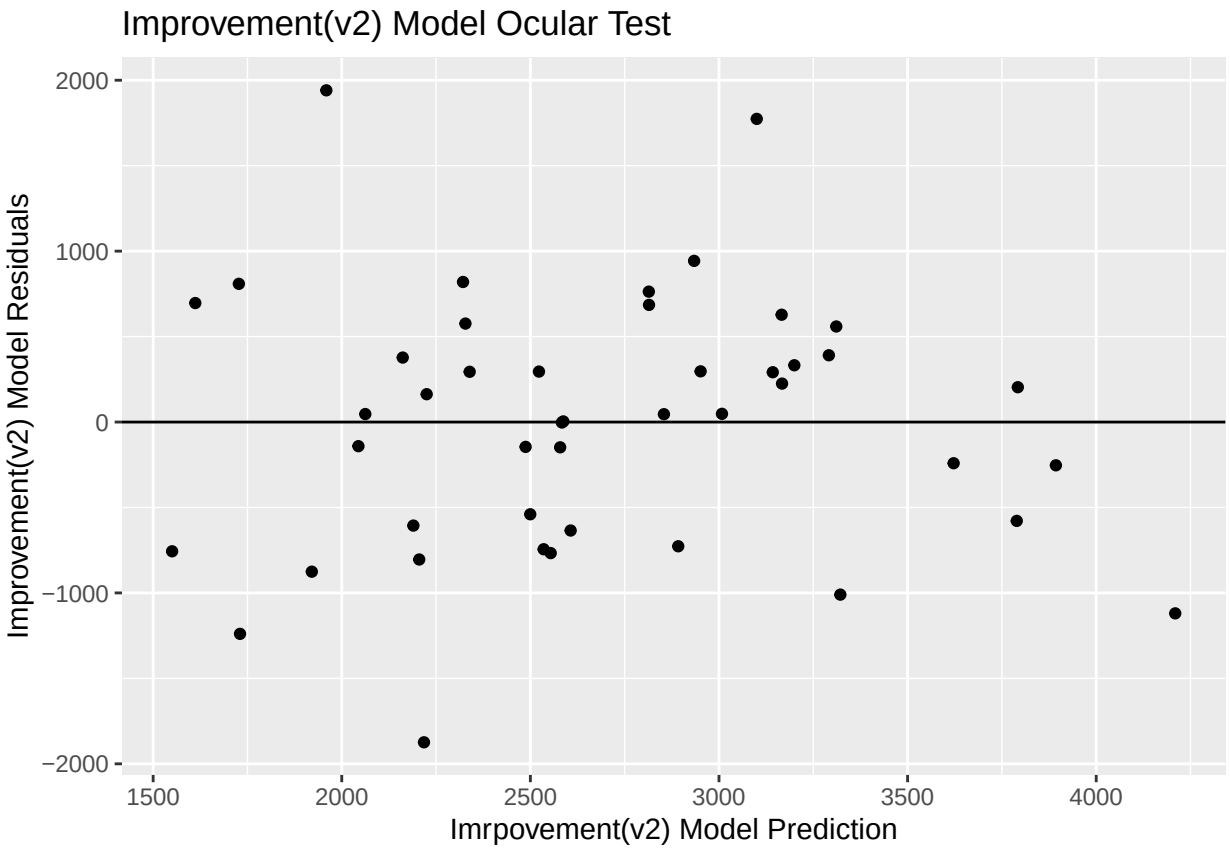
```
plot_4
```


Baseline Model Ocular Test
Baseline Model Residuals
1000
0
−1000
−2000
1000
1500
2000
2500
3000
Baseline Model Prediction

```
plot_5
```


Improvement(v1) Model Ocular Test
Improvement(v1) Model Residuals
2000
1000
0
−1000
−2000
2000
3000
4000
Improvement(v1) Model Prediction

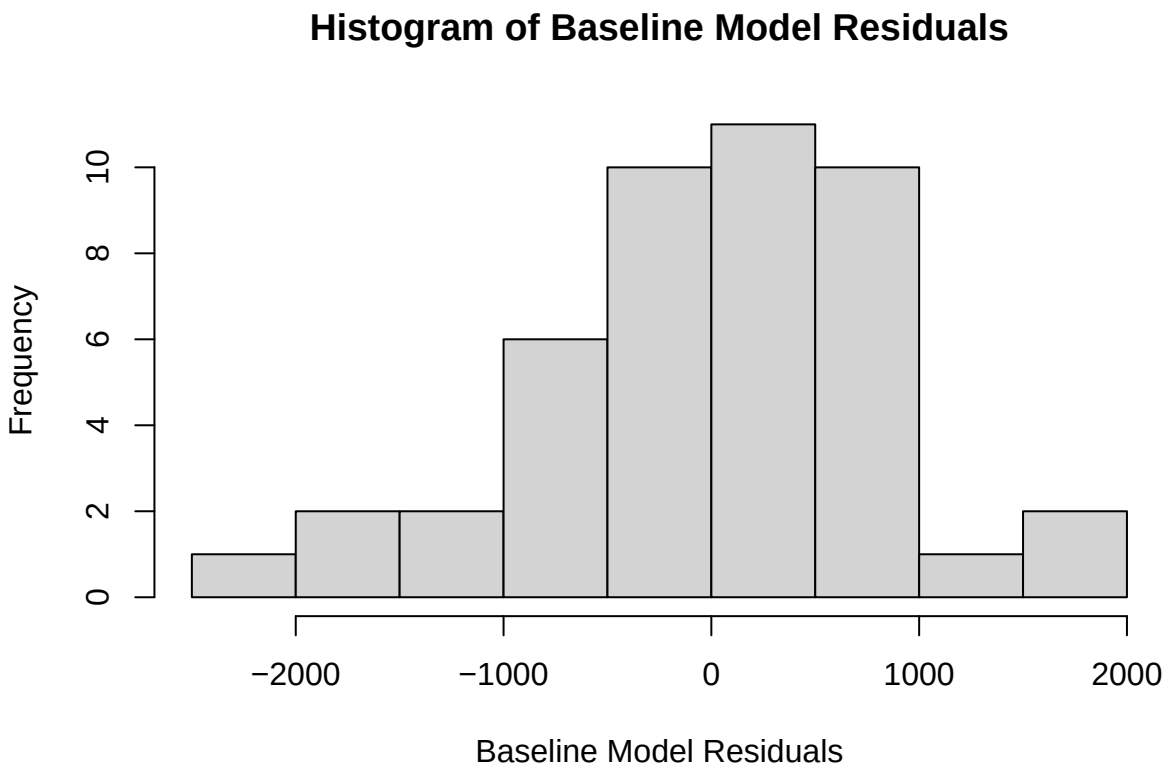
```
plot_6
```



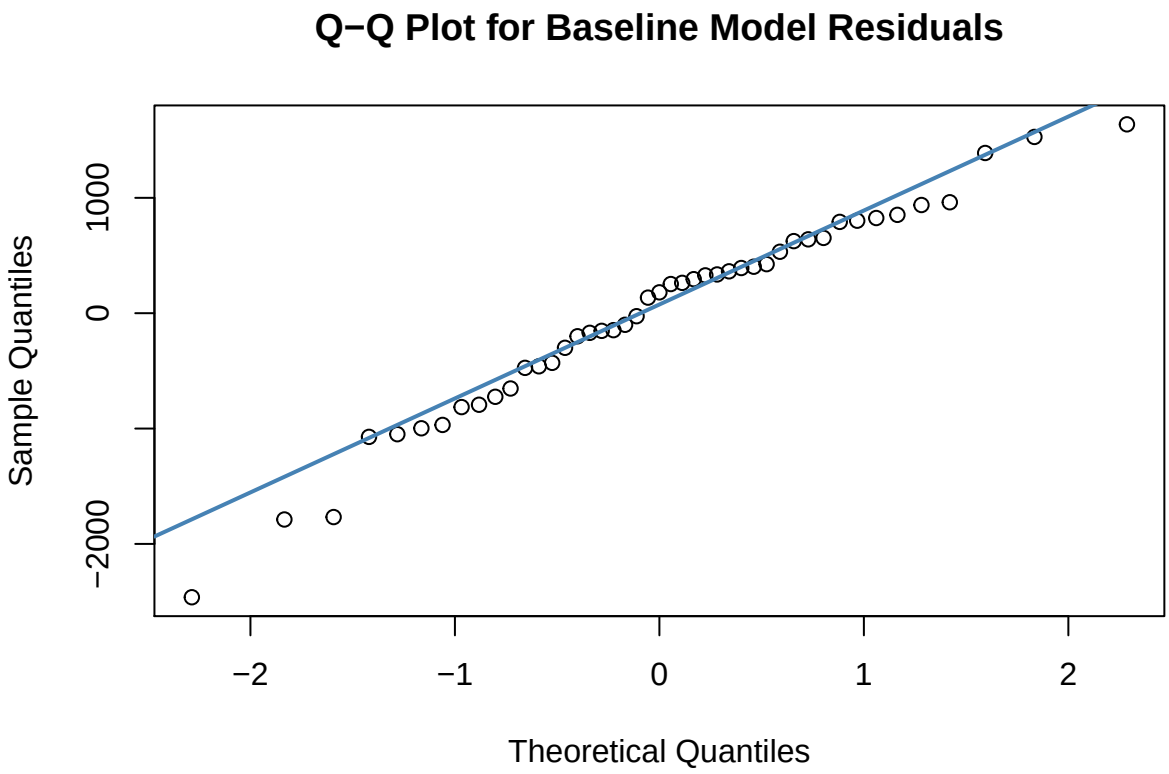
We decided to use the ocular test for homoskedastic conditional variance. In our plots to help us see if there is an even number of data points on both sides of the mean we used "geom_hline" to create a horizontal line at the mean of all the data. For each plot there is visually an even number of points above and below the line, a fanning out of data across the predicted values. There is homoskedasticity with points distributed in a distance fairly evenly from the regression line.

Normally Distributed Errors (Assumption 5)

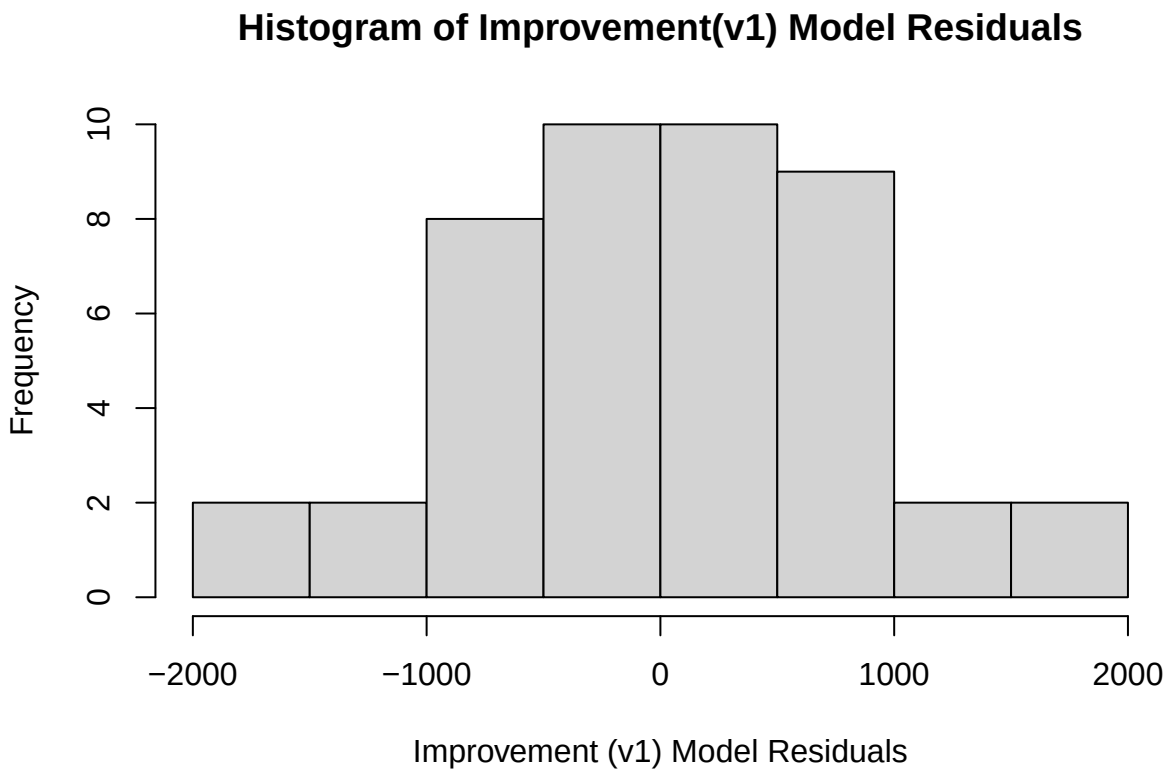
```
hist(x_resid$party_model_resid, main = "Histogram of Baseline Model Residuals",
     xlab = 'Baseline Model Residuals')
```



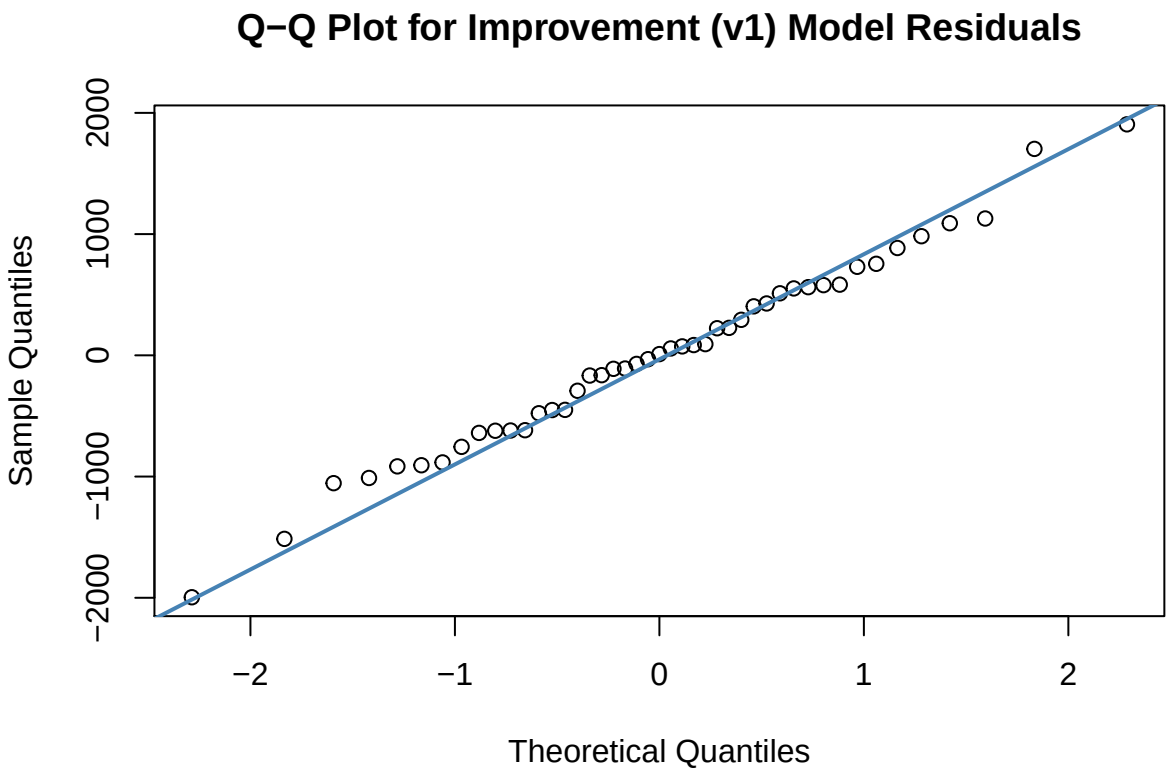

```
qqnorm(x_resid$party_model_resid, main = "Q-Q Plot for Baseline Model Residuals")
qqline(x_resid$party_model_resid, col = "steelblue", lwd = 2)
```



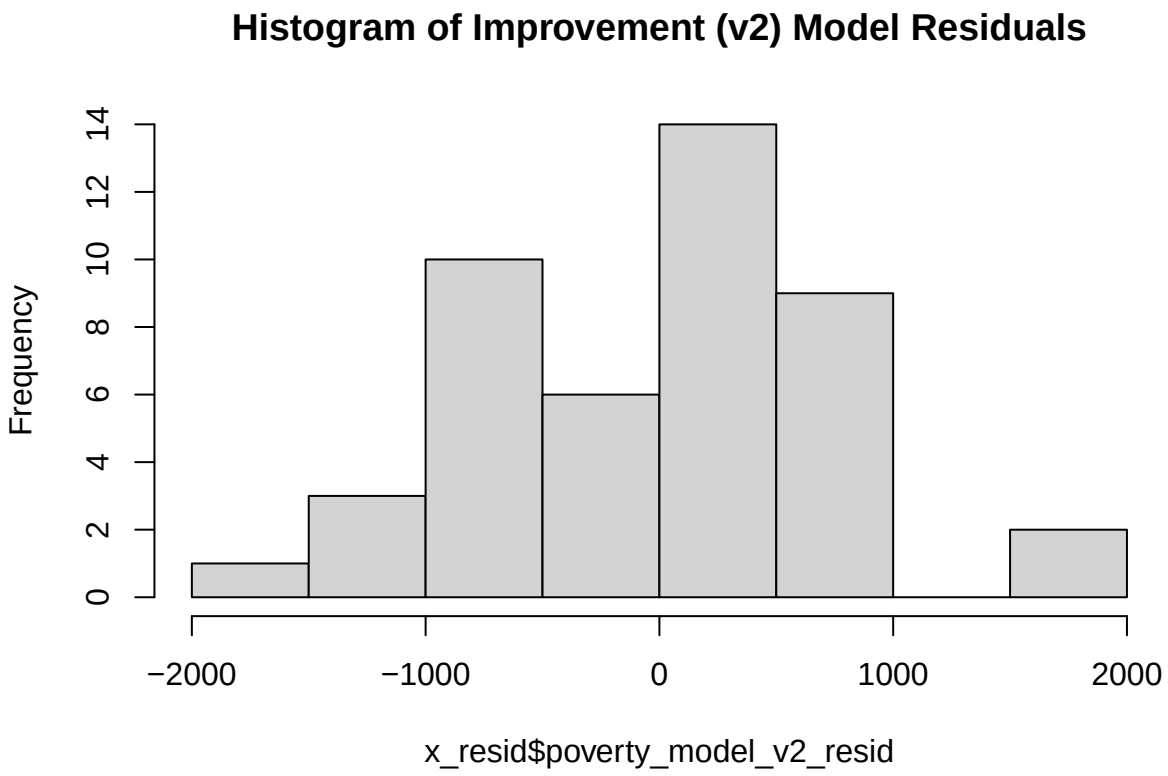

```
hist(x_resid$poverty_model_resid, main = "Histogram of Improvement(v1) Model Residuals",
     xlab = 'Improvement (v1) Model Residuals')
```



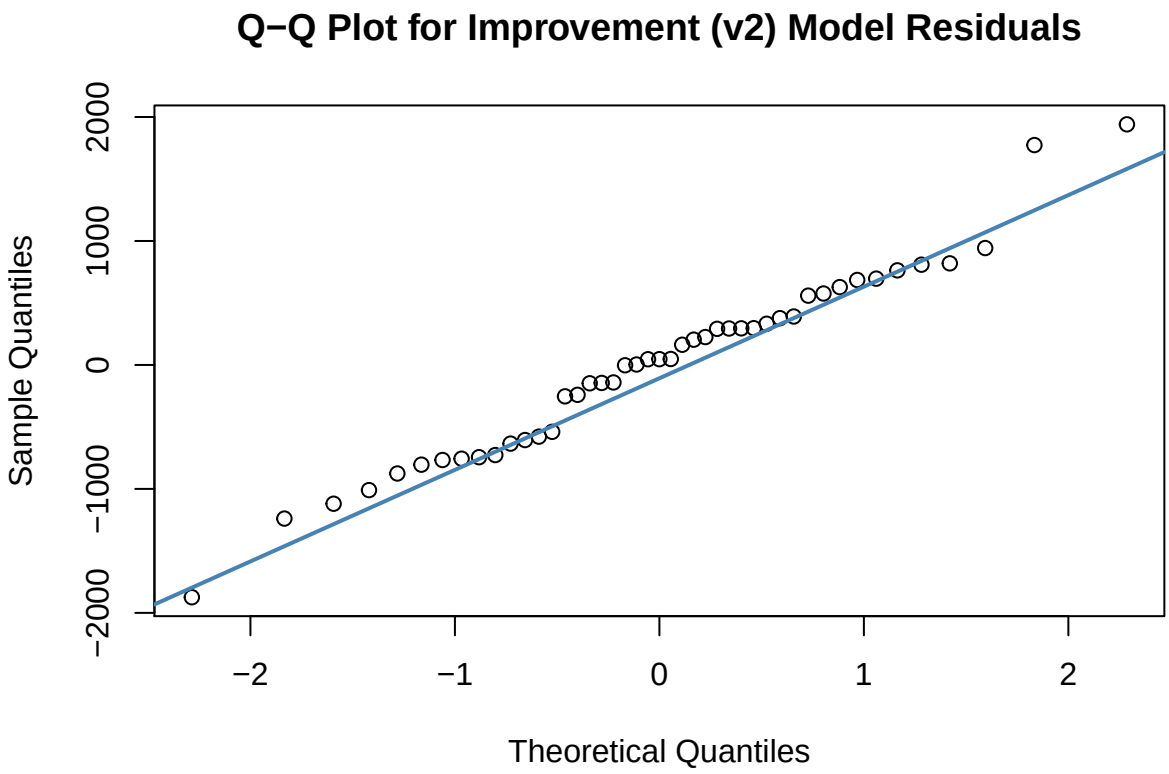

```r
qqnorm(x_resid$poverty_model_resid, main = "Q-Q Plot for Improvement (v1) Model Residuals")
qqline(x_resid$poverty_model_resid, col = "steelblue", lwd = 2)
```




```
hist(x_resid$poverty_model_v2_resid, main = "Histogram of Improvement (v2) Model Residuals")
```




```
qqnorm(x_resid$poverty_model_v2_resid, main = "Q-Q Plot for Improvement (v2) Model Residuals")
qqline(x_resid$poverty_model_v2_resid, col = "steelblue", lwd = 2)
```



We created histograms and QQ plots of the residuals for each model that we created (baseline, improvement (v1), and improvement (v2) in order to determine if the residuals were normally distributed. The baseline model residuals appeared to be very close to normally distributed, although there are a few points of data that stray from the QQ line. The improvement (v1) model residuals were the most normally distributed of the 3 models, with a normal distribution depicted in the histogram as well as data points that fit the QQ line well. Lastly, the improvement (v2) model residuals were the furthest from normally distributed of the 3 models, with data points that strayed from the QQ line the most. However, even the improvement (v2) model was still relatively normally distributed. Because the assumption that the errors are normally distributed is satisfied, this allows us to determine whether the independent variables and the entire model are statistically significant, as well as perform statistical hypothesis testing and generate reliable confidence intervals and prediction intervals.

```
se.model1 = coeftest(party_model, vcov = vcovHC)[ , "Std. Error"]
se.model2 = coeftest(poverty_model, vcov = vcovHC)[ , "Std. Error"]
se.model3 = coeftest(poverty_model_v2, vcov = vcovHC)[ , "Std. Error"]

stargazer(party_model, poverty_model, poverty_model_v2,
          type = "text", omit.stat = "f", se = list(se.model3),
          star.cutoffs = c(0.05, 0.01, 0.001),
          model.numbers = FALSE,
          column.labels = c('Baseline', 'Improvement (v1)', 'Improvement (v2)'),
          dep.var.labels = c('Cases per 100000 Individuals'),
          covariate.labels = c('Sheltering in Place Period', 'Political Party (Republican)',
                               'Percent under Poverty Line',
                              'Percent Covid Cases (White)', "Percent Covid Cases (Black)",
                              'Percent Covid Cases (Hispanic)', 'Percent Covid Cases (Other)'),
          title = "Table 1: Relationship between cases per 100000 and state political party")
```

```
##
## Table 1: Relationship between cases per 100000 and state political party
## ================================================================================
##                                          Dependent variable:
##                                 ---------------------------------------------
##                                          Cases per 100000 Individuals
##                                    Baseline      Improvement (v1)  Improvement (v2)
## --------------------------------------------------------------------------------
## Sheltering in Place Period          -8.447            -9.062*           -9.228*
##                                     (7.420)           (3.817)           (3.896)
##
## Political Party (Republican)        451.062           395.490           448.651
##                                    (304.204)         (271.534)         (273.394)
##
## Percent under Poverty Line                           129.629*          175.203**
##                                                      (51.159)          (60.313)
##
## Percent Covid Cases (White)                          -809.667          -107.877
##                                                      (725.836)        (1,138.072)
##
## Percent Covid Cases (Black)                                            -454.185
##                                                                       (1,578.008)
##
## Percent Covid Cases (Hispanic)                                         2,011.164
##                                                                       (1,487.426)
##
## Percent Covid Cases (Other)                                           -1,667.042
##                                                                       (1,967.591)
##
## Constant                          2,785.850*        1,660.463          401.627
##                                   (1,285.019)       (873.265)         (1,281.530)
##
## --------------------------------------------------------------------------------
## Observations                          45                45                45
## R2                                   0.212             0.351             0.412
## Adjusted R2                          0.175             0.286             0.301
## Residual Std. Error            895.759 (df = 42) 832.925 (df = 40) 824.273 (df = 37)
## ================================================================================
## Note:                                              *p<0.05; **p<0.01; ***p<0.001
```

Overall, based on our regression analysis, we found that the number of covid-19 cases per 100,000 individuals decreases by about 9 cases for every additional day of sheltering in place but increases if the state is Republican (based on the party affiliation of the governor). There is no stasticial significance for the baseline model. In our improvement v1 model there is statistical significance for the period of time of sheltering in place and the percent of the population living under the poverty line. The improvement v2 model also has statistical significance for the period of time of sheltering in place and the percent of the population living under the poverty line. Although we do not have statistical significance for political party status we believe our model is still practically significant because higher cases per 100,000 in Republican states appear to match the current trends in covid-19 cases as the pandemic continues to affect states with more lax policies.

Conclusion

Piecing all our work together, it appears that as of right now, states with longer sheltering in place periods and a largely Democratic population are faring better compared to their Republican counterparts. States with a higher proportion of citizens under the poverty line also have higher rates of Covid-19 cases. The

relationship between the number of cases and race is still unclear. However, it is important to note that our model is descriptive, not explanatory. Therefore, our regression models are only capable at describing the trends within the data; in other words, we can only look at our models to see the effects of the covid-19 pandemic across the country. Another limitation is that we wanted to look at other factors such as age, closures of institutions (i.e. restaurants, gyms, etc.), unemployment rate, etc. but due to high collinearity between variables, we were unable to address those features. With that in mind, though, we can still move to a closer understanding at the different relationships between covid-19, state policies, and even socioeconomic and racial factors.