PS405 - Final Problem Set

Nicholas R. Gonzalez

Question: Estimate the true conditional expecation function of repub_change_1216, given the provided explanatory variables.

Step 1: Setting up my enviroment, loading the package I will use / used, and importing the data. Also examining the data to begin thinking about creating the model.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(sandwich)
library(modelsummary)
```

```
`modelsummary` 2.0.0 now uses `tinytable` as its default table-drawing
  backend. Learn more at: https://vincentarelbundock.github.io/tinytable/

Revert to `kableExtra` for one session:

  options(modelsummary_factory_default = 'kableExtra')
  options(modelsummary_factory_latex = 'kableExtra')
  options(modelsummary_factory_html = 'kableExtra')

Silence this message forever:

  config_modelsummary(startup_message = FALSE)
```

```r
library(ggeffects)
library(broom)
library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':
```

```
    recode
```

```
The following object is masked from 'package:purrr':

    some
```

```r
library(kableExtra)
```

```
Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```r
library(estimatr)
library(lmtest)
```

```
Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```r
load("~/Documents/PS405-Linear-Models/final_problem_set/pres_elec16_data.RData")

#rename the datsa

election <- exam_data

# view(election)
# str(election)
```

Step 2: Clustering the data by state.

Generally, with all the county level data exists, it is probably a good idea to cluster the data by each state. Clustering the county data by each state also helps account of spatial similarities at the intra-state level, and potentially even the national level. In general, clustering will improve our accuracy, but it will also increase our standard errors why is a known byproduct of clustering.

```r
#named the new column state_v2 as it is the second itteration of state data in the data set.

election$state_v2 <- as.factor(factor(election$state))
```

Step 3: Creation of new variables.

There is a lot of demographic information within the data set, and after analyzing it for a bit, I think it would be wise to create some new variables that help us analyze our interest which is the change if republican vote share in the 2016 to 2012 election.

The new variables I am going to creat are as a follows: - Change in unemployment - Change in median income - Change in poverty rate - total non-white VAP

The three new variables pertaining to change, are to tease out if economic standing had anything to do with the change in who people might vote for. The total non-white VAP is to make it easier to examine if having a larger amount of non-white VAP's is associated with a change. This new variable does raise some concern however as hispanics have increasingly voted for republicans are at a much higher rate than blacks, who they will be grouped with in this new variable, but this has been a more recent development, compared to the time of this data.

For all these new variables, I am doing $2016value - 2012value$, to get the difference. For this, a positive result would mean an increase in for example median income from 2012 to 2016, and a negative result would mean a decrease.

The result in the dataframe below show a slight increase in median income, which is $2,270, and decreases in the poverty rate, and unemployment rate.

```
# doing the math


election$unempl_diff <- election$unempl_rate15 - election$unempl_rate12
election$median_diff <- election$median_income16 - election$median_income12
election$Pov_diff <- election$poverty16 - election$poverty12




# view(election)


#storing the values to examine
```

```r
mean_un <- mean(election$unempl_diff)

mean_median <- mean(election$median_diff)

mean_pov <- mean(election$Pov_diff)


#making a df to easily examine


means <- data.frame(mean_un, mean_median, mean_pov)


print(means)
```

```
    mean_un mean_median   mean_pov

1 -2.331855    2270.256 -0.3619736
```

```r
## now going to create the nonwhite variable


election$total_nonwhite_vap <- election$black_vap + election$asian_vap + election$asian_vap
```

Step 5: Making models

Now that I have some new variables created, I am going to start making some models. As someone who's research interests generally fall in the lines of urban politics, race, class and political economy, I am going to focus on indicators related to those areas first. I will bring more stuff in and or make changes if those do not net any results.

For my first models, I am going to just put all of our new variables in, and see what we get.

```r
#basic lm/ols model to start, without clusters.

basic_model <- lm(repub_change1216 ~ unempl_diff + median_diff + Pov_diff + total_nonwhite_va

# next making a model with the clusters

model_clustered <- lm_robust(repub_change1216 ~ unempl_diff + median_diff + Pov_diff + total_

# Now let's compare

modelsummary(list("Basic" = basic_model, "Clustered" = model_clustered),
gof_omit = "AIC|BIC|Log.Lik",
title = "Comparing Models",
stars = TRUE,
digits = 5)
```

Step 6: Improving the models

The models created do not explain that much. So the the next step may be to remove or add or interact variables.

```r
# going to interact the variables first to see that gives us anything. Just redoing the same

basic_model_interactions <- lm(repub_change1216 ~ unempl_diff * median_diff * Pov_diff * tota

model_clustered_interactions <- lm_robust(repub_change1216 ~ unempl_diff * median_diff * Pov_
```

Table 1: Comparing Models

	Basic	Clustered
(Intercept)	4.236***	4.236***
	(0.222)	(0.931)
unempl_diff	−0.161*	−0.161
	(0.072)	(0.286)
median_diff	0.000***	0.000**
	(0.000)	(0.000)
Pov_diff	0.047	0.047
	(0.056)	(0.068)
total_nonwhite_vap	0.000***	0.000**
	(0.000)	(0.000)
Num.Obs.	3111	3111
R2	0.059	0.059
R2 Adj.	0.058	0.058
F	48.475	
RMSE	5.40	5.40
Std.Errors		by: state_v2

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

```
                              data = election,
                              clusters = election$state_v2)

both <- list("Basic" = basic_model_interactions, "Clustered" = model_clustered_interactions)

modelsummary(basic_model_interactions, stars = TRUE)
```

```
summary(model_clustered_interactions)
```

```
Call:
lm_robust(formula = repub_change1216 ~ unempl_diff * median_diff *
    Pov_diff * total_nonwhite_vap, data = election, clusters = election$state_v2)

Standard error type:  CR2

Coefficients:
                                                  Estimate Std. Error
(Intercept)                                      4.445e+00  9.748e-01
unempl_diff                                     -1.289e-01  2.892e-01
median_diff                                     -2.345e-04  1.363e-04
Pov_diff                                         1.642e-01  1.953e-01
total_nonwhite_vap                              -6.614e-05  1.513e-05
unempl_diff:median_diff                          5.945e-06  4.882e-05
unempl_diff:Pov_diff                             1.162e-02  5.790e-02
median_diff:Pov_diff                            -5.695e-05  3.514e-05
unempl_diff:total_nonwhite_vap                  -1.265e-05  4.440e-06
median_diff:total_nonwhite_vap                  -2.461e-10  4.187e-09
Pov_diff:total_nonwhite_vap                     -1.400e-06  1.040e-05
unempl_diff:median_diff:Pov_diff                -1.341e-05  1.254e-05
unempl_diff:median_diff:total_nonwhite_vap      -4.057e-10  1.382e-09
unempl_diff:Pov_diff:total_nonwhite_vap          1.837e-06  2.645e-06
median_diff:Pov_diff:total_nonwhite_vap         -2.935e-09  3.401e-09
unempl_diff:median_diff:Pov_diff:total_nonwhite_vap -1.217e-09  8.975e-10
                                                 t value Pr(>|t|)
(Intercept)                                      4.55963 0.000155
```

```
unempl_diff                                         -0.44583 0.659636
median_diff                                         -1.72049 0.107735
Pov_diff                                             0.84075 0.414567
total_nonwhite_vap                                  -4.37199 0.001663
unempl_diff:median_diff                              0.12178 0.904258
unempl_diff:Pov_diff                                 0.20070 0.843513
median_diff:Pov_diff                                -1.62079 0.148803
unempl_diff:total_nonwhite_vap                      -2.84929 0.029218
median_diff:total_nonwhite_vap                      -0.05878 0.954563
Pov_diff:total_nonwhite_vap                         -0.13464 0.898277
unempl_diff:median_diff:Pov_diff                    -1.07002 0.302527
unempl_diff:median_diff:total_nonwhite_vap          -0.29354 0.778321
unempl_diff:Pov_diff:total_nonwhite_vap              0.69454 0.516269
median_diff:Pov_diff:total_nonwhite_vap             -0.86304 0.418432
unempl_diff:median_diff:Pov_diff:total_nonwhite_vap -1.35572 0.218548
                                                      CI Lower   CI Upper
(Intercept)                                          2.423e+00  6.467e+00
unempl_diff                                         -7.251e-01  4.673e-01
median_diff                                         -5.274e-04  5.832e-05
Pov_diff                                            -2.545e-01  5.829e-01
total_nonwhite_vap                                  -1.002e-04 -3.208e-05
unempl_diff:median_diff                             -9.574e-05  1.076e-04
unempl_diff:Pov_diff                                -1.113e-01  1.346e-01
median_diff:Pov_diff                                -1.399e-04  2.602e-05
unempl_diff:total_nonwhite_vap                      -2.352e-05 -1.786e-06
median_diff:total_nonwhite_vap                      -9.897e-09  9.405e-09
```

```
Pov_diff:total_nonwhite_vap                        -2.833e-05  2.553e-05
unempl_diff:median_diff:Pov_diff                   -4.028e-05  1.345e-05
unempl_diff:median_diff:total_nonwhite_vap         -3.731e-09  2.919e-09
unempl_diff:Pov_diff:total_nonwhite_vap            -4.825e-06  8.498e-06
median_diff:Pov_diff:total_nonwhite_vap            -1.108e-08  5.211e-09
unempl_diff:median_diff:Pov_diff:total_nonwhite_vap -3.352e-09  9.188e-10
                                                        DF
(Intercept)                                        21.920
unempl_diff                                        24.477
median_diff                                        13.763
Pov_diff                                           14.051
total_nonwhite_vap                                  9.287
unempl_diff:median_diff                            20.482
unempl_diff:Pov_diff                               15.677
median_diff:Pov_diff                                7.048
unempl_diff:total_nonwhite_vap                      5.998
median_diff:total_nonwhite_vap                      8.018
Pov_diff:total_nonwhite_vap                         4.876
unempl_diff:median_diff:Pov_diff                   14.148
unempl_diff:median_diff:total_nonwhite_vap          6.455
unempl_diff:Pov_diff:total_nonwhite_vap             5.365
median_diff:Pov_diff:total_nonwhite_vap             6.586
unempl_diff:median_diff:Pov_diff:total_nonwhite_vap  6.790

Multiple R-squared:  0.08429 ,  Adjusted R-squared:  0.07985
F-statistic: 31.33 on 15 and 49 DF,  p-value: < 2.2e-16
```

```r
# again the models pretty much explain nothing.
```

Interacting did not do anything.

Now let's bring urban population percentage into the mix.

```r
basic_model_urban <- lm(repub_change1216 ~ unempl_diff + pct_urban + median_diff + Pov_diff +

# next making a model with the clusters

model_clustered_urban <- lm_robust(repub_change1216 ~ unempl_diff + pct_urban + median_diff +

# Now let's compare

modelsummary(list("Basic" = basic_model_urban, "Clustered" = model_clustered_urban),
gof_omit = "AIC|BIC|Log.Lik",
title = "Comparing Models",
stars = TRUE,
digits = 5)
```

This begins to get us somewhere. The added coefficient of percentage of population that is in an urban center, higlights that every 1 percent increase in the urban population within a county changes the amount of votes casted for the Republican presidential candidate by a decrease of 0.077. So Donald Trump made successful inroads with rural voters.

I also want to explore college educated voters, and their relatinoship to the change in republican vote share.

```r
basic_model_college <- lm(repub_change1216 ~ unempl_diff + pct_urban + prop_noncollege + medi

# next making a model with the clusters

model_clustered_college <- lm_robust(repub_change1216 ~ unempl_diff + pct_urban + prop_noncol

# Now let's compare

modelsummary(list("Basic" = basic_model_college, "Clustered" = model_clustered_college),
gof_omit = "AIC|BIC|Log.Lik",
title = "Comparing Models",
stars = TRUE,
digits = 5)
```

The results were pretty significant. The increase of non-college graduates in a county, and in a state, drastically increase the change in republican vote shares. The next demographic area I am going to check, and explore is occupation.

```r
basic_model_occupation <- lm(repub_change1216 ~ unempl_diff + pct_urban + prop_manufacturing

# next making a model with the clusters

model_clustered_occupation <- lm_robust(repub_change1216 ~ unempl_diff + prop_manufacturing +

# Now let's compare

modelsummary(list("Basic" = basic_model_occupation, "Clustered" = model_clustered_occupation)
```

```
gof_omit = "AIC|BIC|Log.Lik",
title = "Comparing Models",
stars = TRUE,
digits = 5)
```

Again, we get some change. Trump made big gains with manufacturing industring, but also made big losses in agricutlure and construction.

Let's re-run the latest models but with just blacks, instead of all nonwhites.

```
basic_model_occupation <- lm(repub_change1216 ~ unempl_diff + pct_urban + prop_manufacturing

# next making a model with the clusters

model_clustered_occupation <- lm_robust(repub_change1216 ~ unempl_diff + prop_manufacturing +

# Now let's compare

modelsummary(list("Basic" = basic_model_occupation, "Clustered" = model_clustered_occupation)
gof_omit = "AIC|BIC|Log.Lik",
title = "Comparing Models",
stars = TRUE,
digits = 5)
```

Nothing really changes.

Step 7: Piecing some stuff together.

So we know from our myrid of models so far that our new variables of difference in median income generally has not been helpful in any model, the same with total_nonwhite_vap. The difference in poverty rate has varied, but has somewhat explanatory in our models. However, throughout all of our models, the most salient variables have been the ones pertaining to education, industry, unemployment, and urban population.

Therefore, let's make a *focused* model with those varibales.

```
improved_model <- lm(repub_change1216 ~ unempl_diff + prop_manufacturing + prop_agr + prop_co

# next making a model with the clusters

improved_model_clustered <- lm_robust(repub_change1216 ~ unempl_diff + prop_manufacturing + p

# Now let's compare

modelsummary(list("Basic" = improved_model, "Clustered" = improved_model_clustered),
gof_omit = "AIC|BIC|Log.Lik",
title = "Comparing Models",
stars = TRUE,
digits = 5)
```

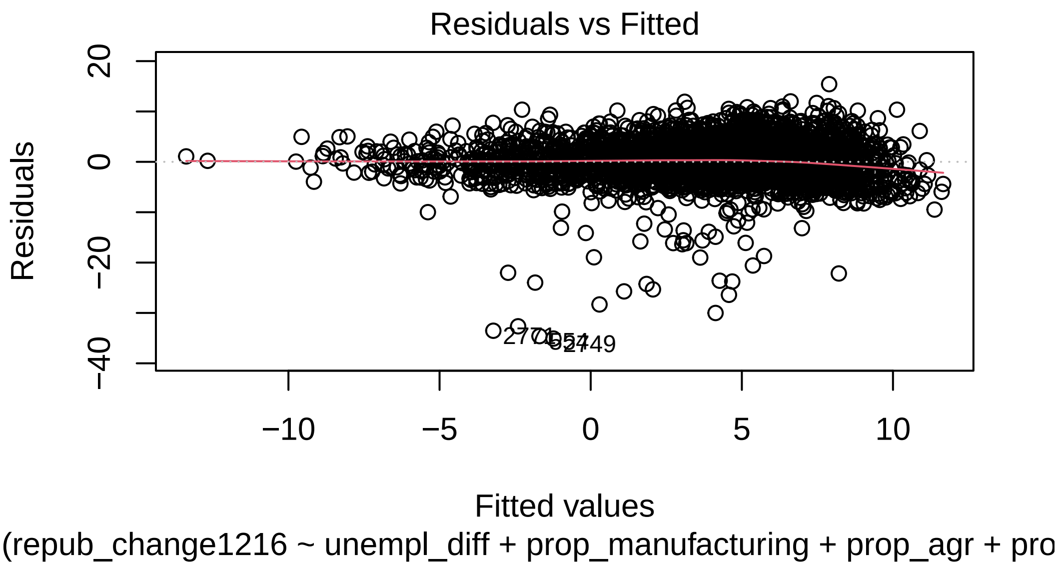
Step 8: Analyizing our model

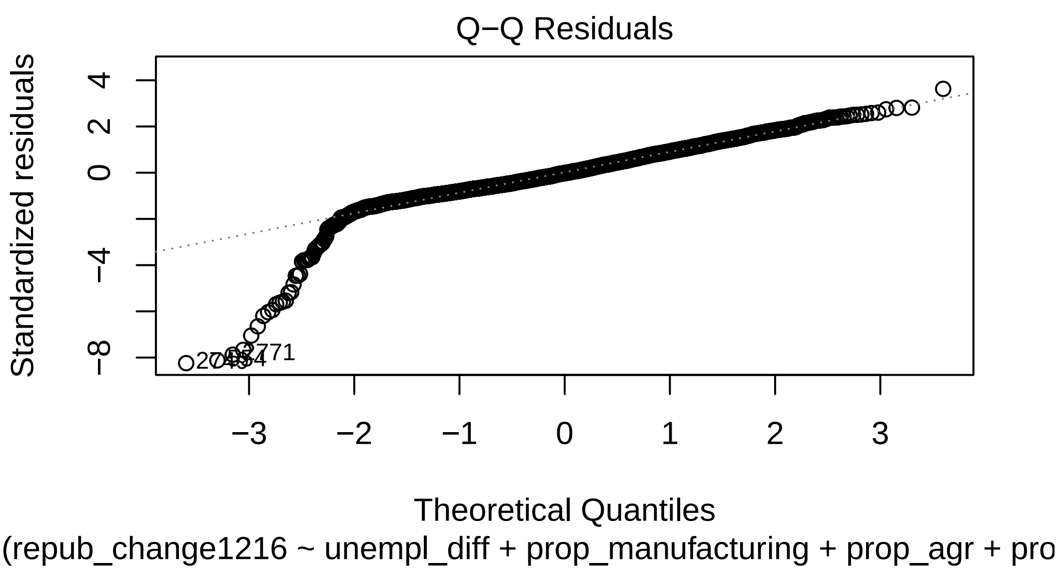
For now, these models are what we want to go forward with.

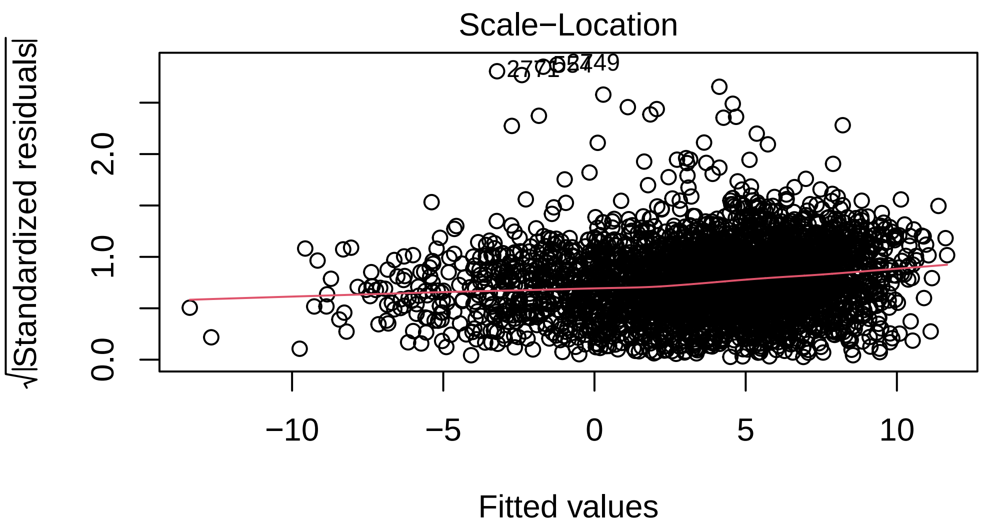
Let's analyze them to make sure they uphold with our OLS assumptions.

Checking distribution

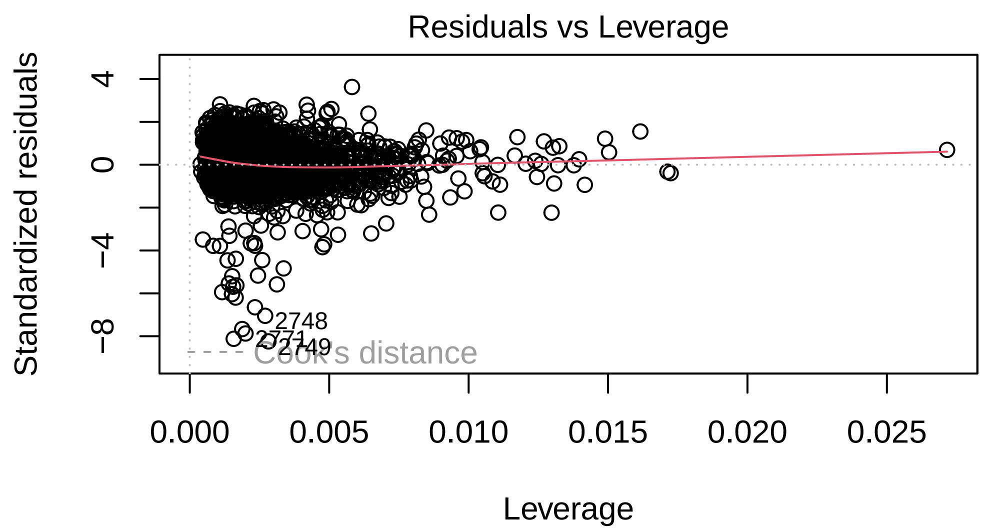
```r
plot(improved_model)
```





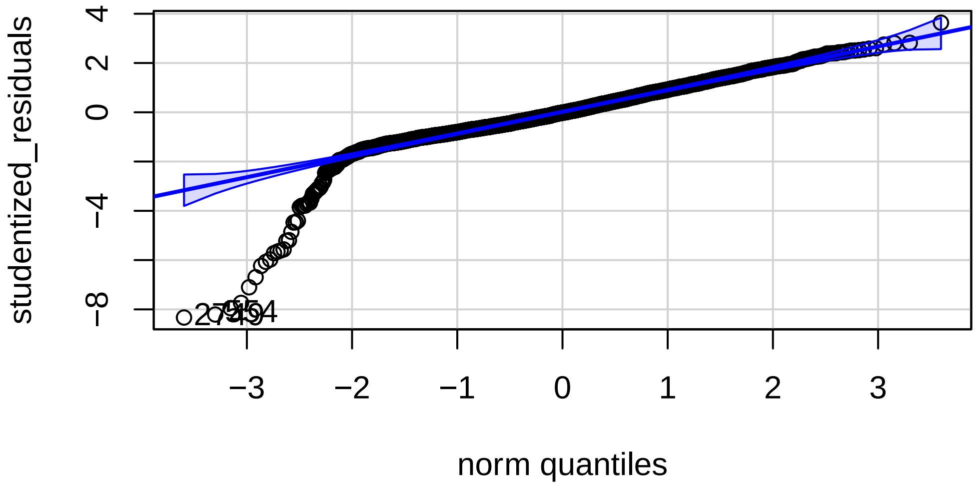


l(repub_change1216 ~ unempl_diff + prop_manufacturing + prop_agr + pro



l(repub_change1216 ~ unempl_diff + prop_manufacturing + prop_agr + pro

```r
studentized_residuals <- rstudent(improved_model)

qqPlot(studentized_residuals)
```




```
[1] 2749  554
```

QQ plot results for the regular model is normally distributed to a point, but the lower tail has some serious outliers and is pretty extreme. This impacts some of the validity of our model. Our model is skewed negatively to the left, in statistical terms.

Let's check for homoskedasticity.

```r
bptest(improved_model)
```

```
	studentized Breusch-Pagan test

data:  improved_model
BP = 11.868, df = 7, p-value = 0.105
```

```r
bptest(improved_model_clustered)
```

```
	studentized Breusch-Pagan test

data:  improved_model_clustered
BP = 11.868, df = 7, p-value = 0.105
```

The models prove to be homoskedasticity, so because of this, and the size of our data, which is over 3,000 observations, the skewed tail of our QQ plot is less of a concern.

Step 8 Addressing QQ plot, non-normal residuals concern.

Even with the concern not being as high, I want to see see if there is something we can do to get a normal distribution.

First, let's log the percent of people living in an urban center.

```r
election$log_pct_urban <- log1p(election$pct_urban / 100)  # transforming the variable, I am
```

Nothing really change. While my model did pass the Breusch-Pagan test, I will still use HC3 for one of the models, just to be *safe* regarding my results, since my qqplot show some concerns.

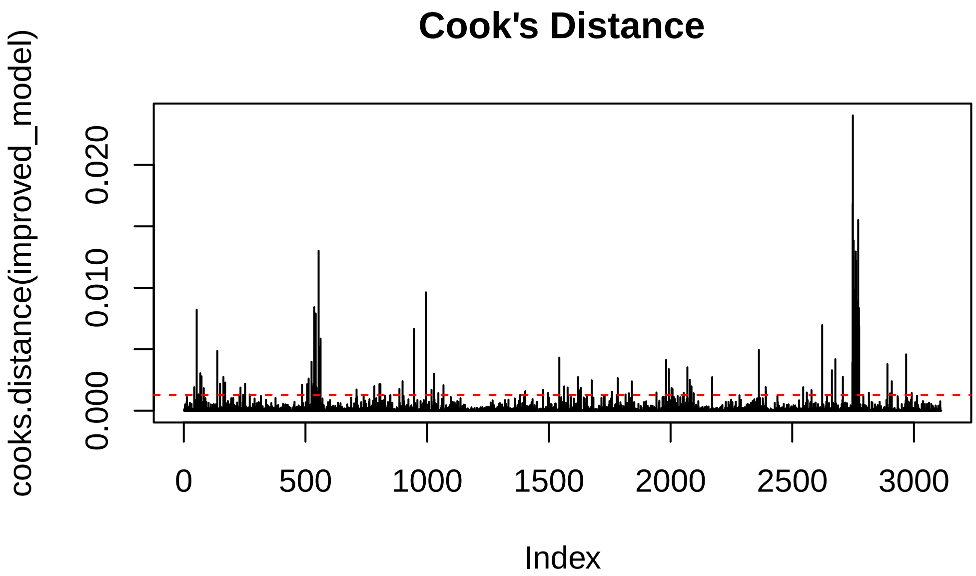
```r
coeftest(improved_model, vcov = vcovHC(improved_model, type = "HC3"))
```

```
t test of coefficients:

                    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)       -14.1690890   0.8357754 -16.9532 < 2.2e-16 ***
unempl_diff         0.3034771   0.0600382   5.0547 4.560e-07 ***
prop_manufacturing  9.2656043   1.4408837   6.4305 1.468e-10 ***
prop_agr          -12.2491194   1.4148551  -8.6575 < 2.2e-16 ***
prop_constr       -15.3131783   3.8001294  -4.0296 5.720e-05 ***
pct_urban          -0.0559566   0.0031915 -17.5328 < 2.2e-16 ***
prop_noncollege    27.4301236   0.9891446  27.7312 < 2.2e-16 ***
Pov_diff            0.1637868   0.0411630   3.9790 7.079e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

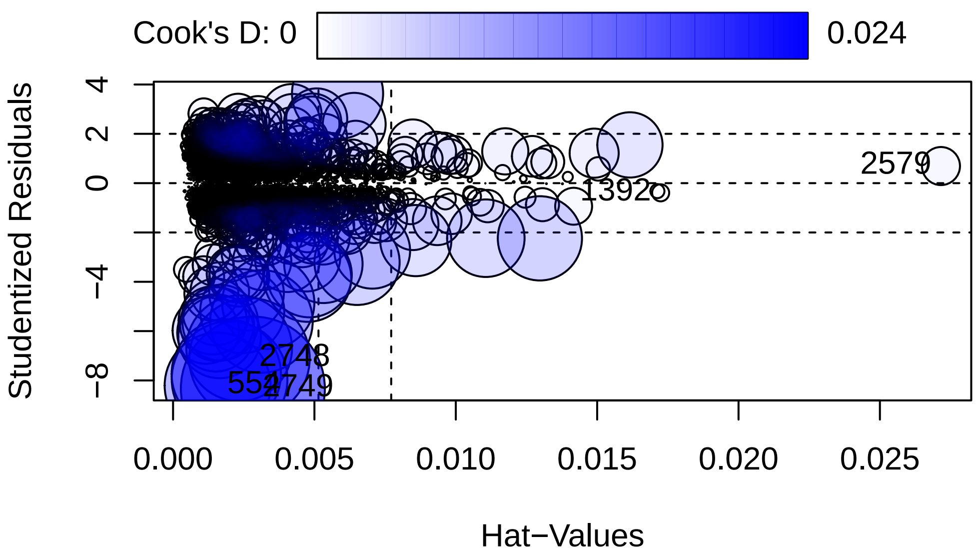
Step 9: Influential Points

The last thing I want to check to maybe correct the model, is to examine influential points.

```r
# checking influential points

plot(cooks.distance(improved_model), type = "h", main = "Cook's Distance")
abline(h = 4/nrow(election), col = "red", lty = 2)
```




```
#influence plot to look at cooks distance

influencePlot(improved_model)
```




```
       StudRes         Hat        CookD
554  -8.2083345 0.001576079 0.0130164048
```

```
1392 -0.3920671 0.017245724 0.0003372762
2579  0.6930919 0.027159511 0.0016766591
2748 -7.1038084 0.002703059 0.0168288772
2749 -8.3316986 0.002822431 0.0240301445
```

```r
# I asked ChatGPT how to get what points were influential in a specific dataframe, and get tl

influential_points <- which(cooks.distance(improved_model) > (4/nrow(election)))

election[influential_points, ] %>%
  summarise(
    mean_median_income16 = mean(median_income16, na.rm = TRUE),
    mean_median_income12 = mean(median_income12, na.rm = TRUE),
    mean_poverty_2012 = mean(poverty12, na.rm = TRUE),
    mean_poverty_2016 = mean(poverty16, na.rm = TRUE)
  )
```

```
  mean_median_income16 mean_median_income12 mean_poverty_2012 mean_poverty_2016
1             46165.89             43789.29          18.00965           17.7807
```

Okay, so this is sort of telling. The mean median income changes, but the poverty does not. What I am observing for this is that it may have been wrong to create variables that looked at the difference in poverty, and income or even unemployment. The reason for this is places that already had high poverty, that remained with high poverty, may have changed their vote. They could be unhappy with that fact the party they voted for last time, did not improve their material circumstance. This is not something that the rate of change would showcase.

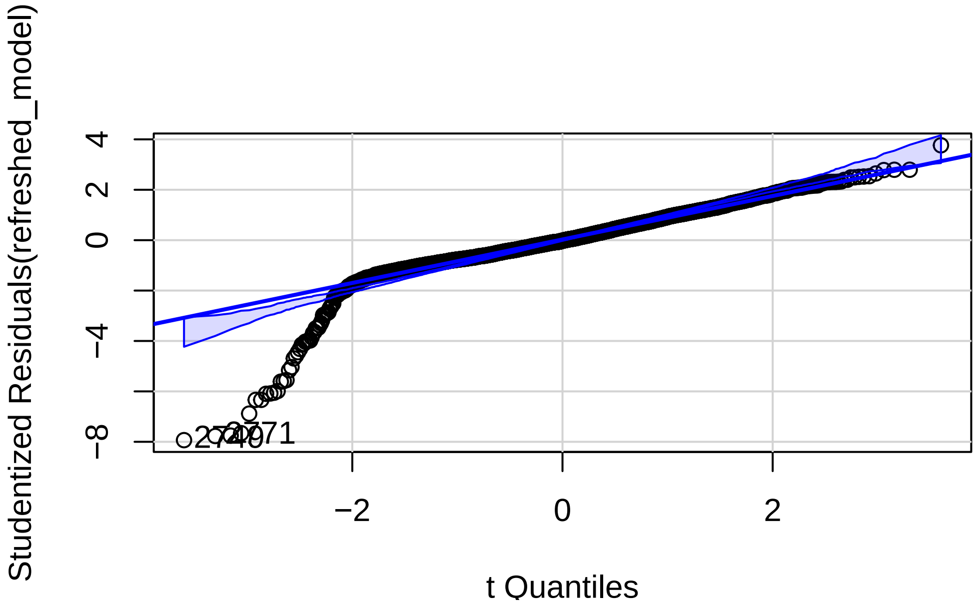
Step 10: Recreating models, with not new variables created.

I am just going to now create two new models, with all the variables are of interest, but not combine any or make any new ones.

```
refreshed_model <- lm(repub_change1216 ~ poverty12 + poverty16 + unempl_rate12 + unempl_rate

refreshed_model_clustered <- lm_robust(repub_change1216 ~ poverty12 + poverty16 + unempl_rate



modelsummary(list("Refreshed Basic" = refreshed_model, "Refreshed Clustered" = refreshed_mode
gof_omit = "AIC|BIC|Log.Lik",
title = "Comparing Models",
stars = TRUE,
digits = 5)
```

I will come back to the results in a bit but let's check if things are more "normal".

```
qqPlot(refreshed_model)
```




```
[1] 2749 2771
```

```
bptest(refreshed_model)
```

```
	studentized Breusch-Pagan test

data:  refreshed_model
BP = 23.147, df = 9, p-value = 0.005875
```

The qqPlot hardly changes, but now the model fails the homoskedastic check.

Using each unique year would violate the multi-collinearity OLS assumption, but the results do not really improve anything either. So this test was somewhat irrelevant, but it was worth checking.

Step 10: Models with seperate years

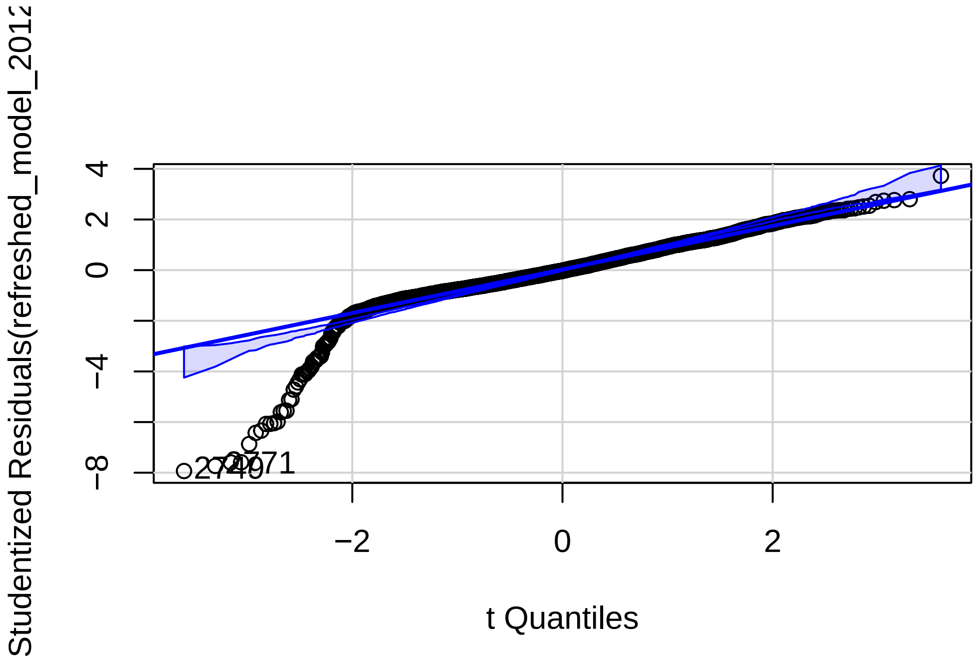
The last way we can examine the relationship(s) with the change or lack of in poverty rates, median income, and unemployment, is by examining them all individually.

```
# model for 2012

refreshed_model_2012 <- lm(repub_change1216 ~ poverty12 + unempl_rate12 + median_income12 + p

refreshed_model_clustered_2012 <- lm_robust(repub_change1216 ~ poverty12 + unempl_rate12 + me

#models for 2016

refreshed_model_2016 <- lm(repub_change1216 ~ poverty16 + unempl_rate15 + median_income16 + p

refreshed_model_clustered_2016 <- lm_robust(repub_change1216 ~ poverty16 + unempl_rate15 + me


modelsummary(list("Refreshed Basic 2012 " = refreshed_model, "Refreshed Clustered 2016 " = re
gof_omit = "AIC|BIC|Log.Lik",
title = "Comparing Models",
stars = TRUE,
digits = 5)
```

```
# going to leave results for now
```

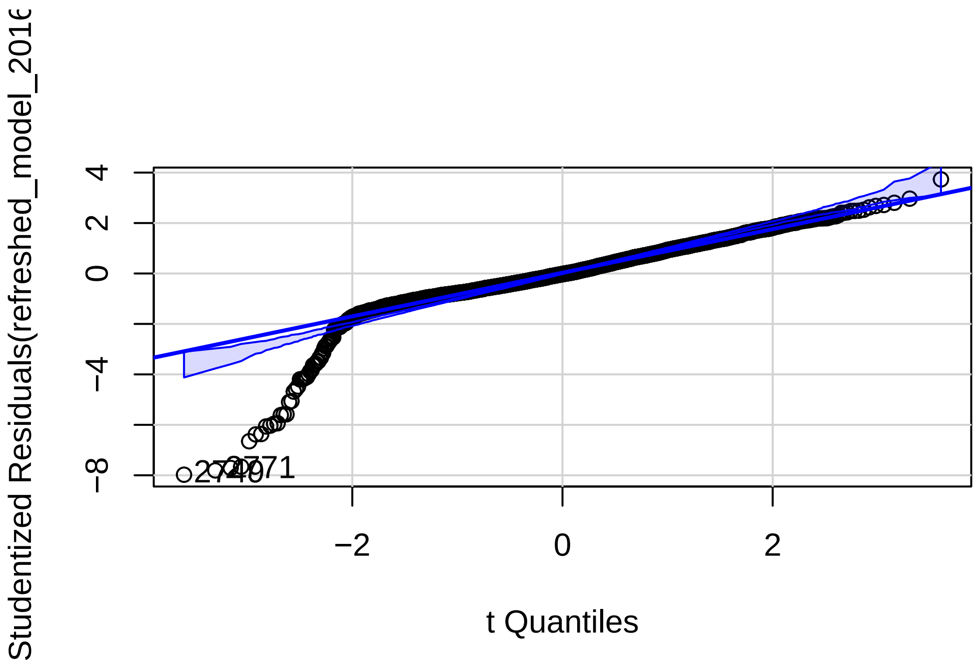
Now testing assumptions

```r
qqPlot(refreshed_model_2012)
```




```
[1] 2749 2771
```

```r
qqPlot(refreshed_model_2016)
```




```
[1] 2749 2771
```

```r
bptest(refreshed_model_2012)
```

```
	studentized Breusch-Pagan test

data:  refreshed_model_2012
BP = 22.687, df = 6, p-value = 0.0009085
```

```r
bptest(refreshed_model_2016)
```

```
	studentized Breusch-Pagan test
```

```
data:  refreshed_model_2016
BP = 22.715, df = 6, p-value = 0.0008976
```

Again, not much has changed.

Step 11: Rounding up what has been learned and observed, and some new variables.

This section is dedicated to working through putting together all the things I would like to present, and showcase in the "final write up".

The first thing I would like to go over, and discuss is the utility of my create variables. Those being "pov_diff", which speaks to the difference in poverty rate btween 2016 and 2012, and "median_diff" as well as "unempl_diff" which also look at the difference in that same temporal period. While these are useful as potential descriptors, they cause mutlicollinearity issues. The reason this OLS assumption is volated is because the two poverty rates are highly correlated with eachother, and the same is true for median income and unemployment rate.

```r
cor(election$poverty12, election$poverty16)
```

```
[1] 0.9607138
```

```r
cor(election$median_income16, election$median_income12)
```

```
[1] 0.9787943
```

```
cor(election$unempl_rate15, election$unempl_rate12)
```

```
[1] 0.8840874
```

It is a point of interest to potentially measure how change might have affected the repulican vote share, but that is probably for a separate model, or could be fixed with further attempts to log variables, or even use a polynomial model. Generally however, the models with my new created variables is extreme on the left side, and is negatively skewed because it is skewed to the left. The last variable I want to explore is a state-based variable, to tease out some of the results from one of the previous models. When we looked at industry, proportion of manufacturing workers was salient. Because of this, and knowing that Trump made in-roads with non-urban voters from other models as well, I am going to create a rust belt variable.

```
# defining these states the rust belt
rust_belt_states <- c("Ohio", "Michigan", "Pennsylvania", "Kentucly", "West Virginia", "Wisc

# New York is a questionable rust belt state, but I will leave it in for now.

election$rust_belt <- ifelse(election$state %in% rust_belt_states, 1, 0)

rust_belt_model <- lm(repub_change1216 ~ poverty16 + unempl_rate15 + median_income16 + pct_u

modelsummary(rust_belt_model, stars = TRUE)
```

So the results of the rust belt variable are statistically significant. Let's make a model with the 2012 econonomic indicators, manufacturing, and rust belt.

```r
holisitc_model <- lm(repub_change1216 ~ rust_belt + prop_manufacturing + pct_urban + poverty1

modelsummary(holisitc_model, stars = TRUE)
```

So this model looks good, but unemployment rate does not seem to explain anything, the same with median_income. So let's remove it.

```r
revised_model <- lm(repub_change1216 ~ rust_belt + prop_manufacturing + pct_urban + poverty1

modelsummary(revised_model, stars = TRUE)
```

In this most recent model, all of our results are significant at the .0001 level. Our R^2 also does not change despite removing some variables so this show they did not really explain much, or add much to our model.

The last thing I want to explore is rural-ness a bit deeper. So I am going to make a variable for count with a sub 50% urban percentage.

```r
# variable creation

election$rural <- ifelse(election$pct_urban < 50, 1, 0)
```

```r
#creating new model

rural_model <- lm(repub_change1216 ~ rust_belt + prop_manufacturing + pct_urban + poverty16 +

modelsummary(rural_model, stars = TRUE)
```

'

Step 12: Putting together final materials

Based on all we have uncovered, we want to check some assumptions of our final models

Our final models are going to be as follows, and why:

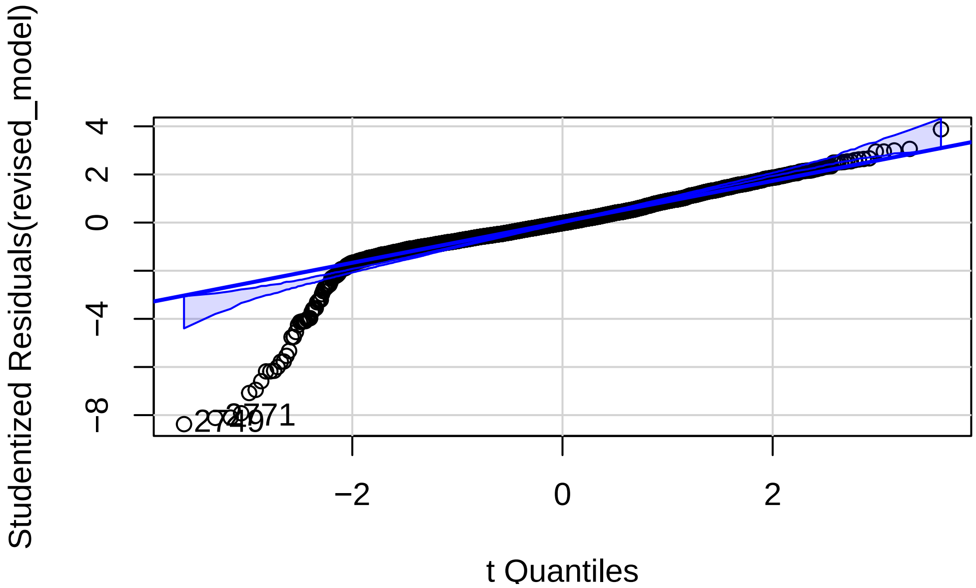
Model 1: repub_change126, prop_noncollege, pct_urban, prop_manufacturing, rust_belt, poverty16,

```
# at this point I explored using poverty12 because that would make more sense theoretically,



final_rust_belt <- lm(repub_change1216 ~ poverty16 + prop_noncollege + prop_manufacturing + r


modelsummary(final_rust_belt)
```

Why?: This model contains the most salient results, with the most explanatory power. As all these results were statistically significant.

Our model is heteroskedastic however, as scene below, so we will use HC3 standard errors. Also known as robust standard errors.

```
qqPlot(revised_model)
```




```
[1] 2749 2771
```

```r
bptest(revised_model)
```

```
	studentized Breusch-Pagan test

data:  revised_model
BP = 37.19, df = 5, p-value = 5.488e-07
```

```r
# using hc3

final_rust_belt_hc3 <- lm_robust(repub_change1216 ~ poverty16 + rust_belt + pct_urban + prop_
                                 data = election, se_type = "HC3")
```

```r
modelsummary(final_rust_belt_hc3, stars = TRUE)
```

While this model may violate the multicollinearity assumption with how its variables are created, our original "imrpoved" model passed the BP test. So it is also of use. This can be seen below, and the model.

```r
improved_model <- lm(repub_change1216 ~ unempl_diff + prop_manufacturing + prop_agr + prop_co

bptest(improved_model)
```

```
	studentized Breusch-Pagan test

data:  improved_model
BP = 11.868, df = 7, p-value = 0.105
```

Then lastly, we have our clustered model, of what is our best model, without the rust-belt variable as this would create issues with our state based cluster.

```r
final_cluster <- lm_robust(repub_change1216 ~ unempl_diff + prop_manufacturing + prop_agr + p

modelsummary(final_cluster, stars =  TRUE)
```

```r
# for the final presenation we should prob present this model unclustered just for references

final_uncluster <- lm_robust(repub_change1216 ~ unempl_diff + prop_manufacturing + prop_agr +

modelsummary(final_uncluster, stars = TRUE)
```

Final Model Summary Table to compare is:

```r
coef_names <- c(
  "(Intercept)" = "Intercept",
  "poverty16" = "Poverty Rate 2016",
  "unempl_rate15" = "Unemployment Rate 2015",
  "median_income16" = "Median Income 2016",
  "pct_urban" = "Urban Pop. Percentage",
  "prop_noncollege" = "Proportion Non-College",
  "rust_belt" = "Rust Belt States",
  "prop_agr" = "Proportion Agricultural Workers",
  "prop_manufacturing" = "Proportion Manufacturing Workers",
  "prop_constr" = "Proportion Construction Workers",
  "unempl_dff" = "Difference in 2015 and 2012 Unemployment Rate",
  "Pov_diff" = "Difference in Povery Rate in 2016 and 2012"
)
```

```r
# Create a modelsummary table comparing all models

modelsummary(
  list(
    "HC3 Robust" = final_rust_belt_hc3,
    "Unclustered" = final_uncluster,
    "Clustered" = final_cluster,
    "Improved Model" = improved_model,
    "Rust Belt" = final_rust_belt
  ),
  stars = TRUE,
  coef_map = coef_names
)
```

I think ultimately our HC3 model is the best. It has the highest R^2, it corrects for our hetero/normality issues. It also has the highest confidence for our coefficients, and best highlights the urban-rural, and manufacturing divides in American Politics.

```r
#finalizing sutff


final <- modelsummary(
  final_rust_belt_hc3,
  stars = TRUE,
  coef_map = coef_names,
  title = "Table 1: HC3 Robust Regression Results",
  output = "latex_tabular",
```

```
)
```

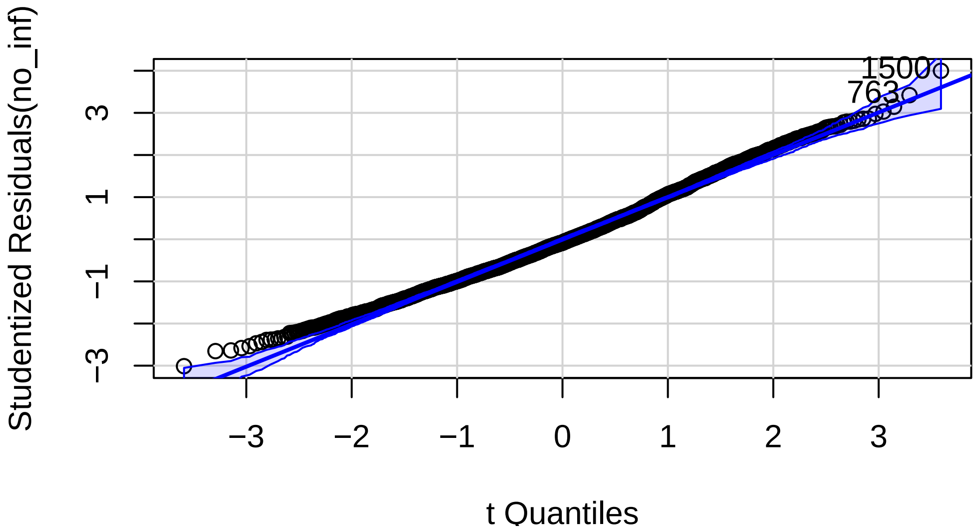
```
Warning: To compile a LaTeX document with this table, the following commands must be placed i

\usepackage{tabularray}
\usepackage{float}
\usepackage{graphicx}
\usepackage{codehigh}
\usepackage[normalem]{ulem}
\UseTblrLibrary{booktabs}
\UseTblrLibrary{siunitx}
\newcommand{\tinytableTabularrayUnderline}[1]{\underline{#1}}
\newcommand{\tinytableTabularrayStrikeout}[1]{\sout{#1}}
\NewTableCommand{\tinytableDefineColor}[3]{\definecolor{#1}{#2}{#3}}

To disable `siunitx` and prevent `modelsummary` from wrapping numeric entries in `\num{}`, ca

options("modelsummary_format_numeric_latex" = "plain")
 This warning appears once per session.
```

```
df <- tidy(final_rust_belt_hc3, conf.int = TRUE)

df <- df[df$term != "(Intercept)", ]

final_plot <- ggplot(df, aes(x = estimate, y = reorder(term, estimate))) +
  geom_point(color = "blue", size = 3) +  # Point for estimate
```

```
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high), height = 0.2, color = "black") +
  labs(
    title = "Figure 1: Plot for Final Robust Model",
    x = "Estimate",
    y = "Variables"
  ) +
  theme_minimal()
```

I just realized at this point I examined the influential points earlier in the project, but never explored removed them. So I am going to do that now for what was my "final" model I was going to present.

```
influential_points <- which(cooks.distance(final_rust_belt) > (4/nrow(election)))

election_no_influential <- election[-influential_points, ]

no_inf <- lm(repub_change1216 ~ poverty16 + rust_belt + prop_manufacturing + pct_urban + prop

qqPlot(no_inf)
```




```
 763 1500
 726 1440
```

```r
bptest(no_inf)
```

```
    studentized Breusch-Pagan test

data:  no_inf
BP = 111.85, df = 5, p-value < 2.2e-16
```

```r
modelsummary(no_inf)
```

```r
influence <- cooks.distance(final_rust_belt)

threshold <- 4 / nrow(election)
```

```r
influential_points <- which(influence > threshold)

influential_df <- election[influential_points, ]

head(influential_df)
```

```
    st_fips cnty_fips fips_code state_abb      state          county black_vap
53        1       105     01105        AL    Alabama    Perry County      5055
68        4       001     04001        AZ    Arizona   Apache County       212
82        4       027     04027        AZ    Arizona     Yuma County      3188
138       5       111     05111        AR   Arkansas Poinsett County      1327
149       5       133     05133        AR   Arkansas   Sevier County       597
163       6       011     06011        CA California   Colusa County       179
    citizen_vap hispanic_vap noncitizen_vap white_vap asian_vap prop_noncollege
53         7783           88             36      2609        18       0.9091491
68        49439         2119            881     11014       120       0.8989875
82       117225        55196          29926     54961      1353       0.8602607
138       18347          231            191     16649         0       0.9198527
149       10220         1142           1987      8104        31       0.9174753
163       11955         4516           3300      6722       205       0.8568408
    prop_noncollege_white prop_manufacturing   prop_agr prop_constr
53              0.8734638         0.23305085 0.09167951  0.07087827
68              0.7903820         0.01957603 0.02784988  0.09382436
82              0.8024147         0.04086061 0.10093890  0.05297271
138             0.9129967         0.16172507 0.08269542  0.06857143
149             0.8997812         0.30074971 0.06084198  0.07771050
```

163	0.7742866	0.10256996	0.25699600	0.04386065

	unempl_rate15	unempl_rate12	poverty16	median_income16	poverty12
53	10.0	13.4	46.9	27403	36.6
68	13.4	19.0	33.1	32366	33.2
82	21.8	23.9	22.6	39700	21.9
138	5.7	8.7	21.9	35851	27.9
149	6.4	8.8	20.4	37014	24.8
163	15.3	20.7	14.3	48006	13.4

	median_income12	pct_urban	evang_pct	mormon_pct	total_2008	dem_2008	gop_2008
53	25215	0.00	50.875657	0.000000	6159	4457	1679
68	32886	25.94	10.977495	42.804787	24262	15390	8551
82	39072	89.57	25.053957	8.605852	43615	18559	24577
138	32248	28.86	88.963718	0.000000	7928	2742	4903
149	32662	36.40	76.978417	0.000000	4580	1291	3125
163	49871	68.28	9.702605	3.263444	6429	2569	3733

	oth_2008	total_2012	dem_2012	gop_2012	oth_2012	total_2016	dem_2016	gop_2016
53	23	6065	4536	1504	25	5255	3823	1403
68	321	25465	16833	8198	434	18659	12196	5315
82	479	33935	14025	19500	410	40759	18336	20586
138	283	7550	2387	4967	196	5283	1878	3149
149	164	4328	1042	3135	151	4517	1075	3281
163	127	5865	2248	3521	96	6260	2484	3425

	oth_2016	repub_change1216	state_v2	unempl_diff	median_diff	Pov_diff
53	29	1.9003611	Alabama	-3.4	2188	10.3
68	1148	-3.7082929	Arizona	-5.6	-520	-0.1
82	1837	-6.9561600	Arizona	-2.1	628	0.7

```
138       256        -6.1817952   Arkansas         -3.0         3603     -6.0
149       161         0.2014008   Arkansas         -2.4         4352     -4.4
163       351        -5.3216405 California         -5.4        -1865      0.9
    total_nonwhite_vap log_pct_urban rust_belt rural
53                5091     0.0000000         0     1
68                 452     0.2306354         0     1
82                5894     0.6395882         0     0
138               1327     0.2535564         0     1
149                659     0.3104216         0     1
163                589     0.5204591         0     0
```

The influence points that were removed, we places that were heavy in poverty I think. Removing those, we now have a normalized qqPlot, or distribution.

```r
modelsummary(
  list("No Influential Points" = no_inf,
       "Original Model" = final_rust_belt, "HC3" = final_rust_belt_hc3),
  gof_map = c("r.squared", "adj.r.squared", "AIC", "BIC", "RMSE"), stars = TRUE
)
```

Okay, so the influence points do not change the results really at all besides making our R^2 higher.

Final Write Up

This project attempted to uncover an answer to the question what caused the change in republican vote share in the 2016, and 2012 elections. There are a lot of ways to engage with this question contextually, and theoretically. The theoretical foundations for the models, and the final model for this project, beyond the statistical skills learned in the course, are taken from existing American Politics research (cite Tessler, Hacker, Jardina, Skocpol, rust belt shit). However, this project avoided an overly in-depth analysis with them to stay within the confines of the assignment, and focus specifically on the statistical modeling part of the problem set.

```r
print(final)
```

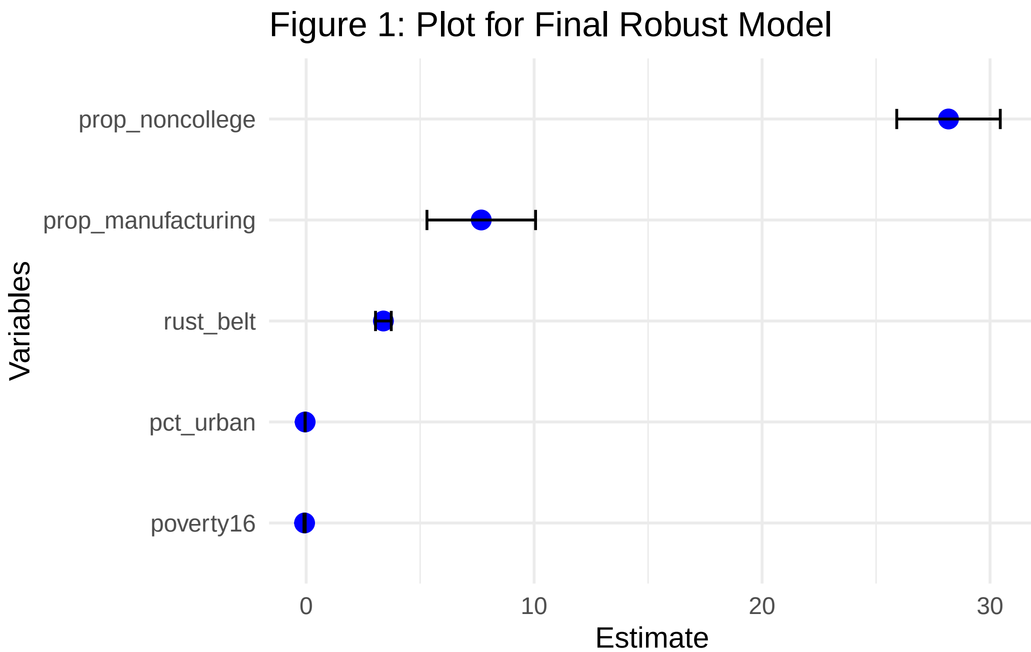
```
\begin{tabular}{ll}
\hline
& (1) \\ \hline
Intercept                         & \num{-17.252}*** \\
& (\num{0.830})    \\
Poverty Rate 2016                 & \num{-0.070}***  \\
& (\num{0.014})    \\
Urban Pop. Percentage             & \num{-0.045}***  \\
& (\num{0.003})    \\
Proportion Non-College            & \num{28.177}***  \\
& (\num{1.158})    \\
Rust Belt States                  & \num{3.388}***   \\
& (\num{0.176})    \\
Proportion Manufacturing Workers & \num{7.682}***   \\
& (\num{1.215})    \\
Num.Obs.                          & \num{3111}       \\
R2                                & \num{0.461}      \\
R2 Adj.                           & \num{0.460}      \\
AIC                               & \num{17605.3}    \\
```

```
BIC                                          & \num{17647.6}    \\
RMSE                                         & \num{4.09}       \\
\hline
\end{tabular}
```

The final model this project puts forth can be seen above. This model using the variables urban population percentage, if the state is part of the rust belt or not, the 2016 poverty rate, the proportion of manufacturing workers as well as the proportion of non college graduates. Since the model struggles to satisfy the assumptions of homoskedasciticty, and normally distributed residuals, this is attempted to be corrected by using robust standard errors. These variables were chosen because through the construction of numerous models, they were routinely most salient. They also theoretically make sense when employing the knowledge we know about American politics, from previously mentioned literature.

The results of this model indicate salient factors in the change of the republican share. Our poverty rate coefficient has a small relationship with our intercept, but this relationship is statistically significant. This relationship proves to be somewhat important, and having a larger sample size like we do likely helps confirm this. Another salient indicator, although not a tremendously large relationship, is that of urban population percentage. When holding other variables constant, for each 1 percentage point increase in urban population, the change in Republican vote share decreases by 0.043 percentage points. The size of the effects, and their confidence intervals can be seen below.

```r
print(final_plot)
```



The coefficients that were most salient, which we can see in our model table as well as our plot that showcases the estimate and the confidence intervals, are rust belt states, and proportion of non-college graduates. The proportion of non-college-educated voters has a large and positive effect. This means that counties with a higher share of non-college-educated voters saw significantly greater increases in the Republican vote share. While not as large of a relationship, the states within our rust belt variable, also have a strong, positive and statistically significant result. These results highlight the electoral shifts made in the 2016 election, and the areas Trump had success in.

While our final model violates some OLS assumptions, as stated already, these were attempted to be corrected using robust standard errors. As demonstrated throughout the project as well, solutions were tried such as logging and transforming variables, and using diagnostic tools. The project also attempted to remove any influence points, and when doing this did this produce a more normal distribution, however the final model presented is just the model with the influence points, that used robust standard errors. The reason for this is because the primary conclusions remain the same, so removing the data seemed unnecessary. The additional models without the influence points however would likely be in the appendix if this was a full article. Interactions were also explored in this project, but did result in anything novel. Also, some of the models developed were more successful than others, and again, they would be in the final paper, or appendix were this a manuscript for an article. However, the model chosen had the best overall results all things considered.

In short, I hope I have demonstrated successfully, the analytical research process used to develop a linear regression model, that aims to answer this problem sets question. To do this, I employed tools learned throughout the quarter, such exploring the transformation of variables, running diagnostics to explore my models, and general coding skills for data presentation and visualization. I also hope that have been successful in creating a useful model, that helps presents answers to the question itself. Ultimately, after all of this, this project finds that Trump made in-roads with citizens who were not college degree holders in the rust belt, more than anything else. And the nature of American federalism winning these unique counties, was worth more than making national in-roads, which is why he won the election but lost the popular vote.

Addendum

As mentioned periodically, this project employed Large-Language Models, specifically Open AI's ChatGPT at various points in this project. Specifically, when exploring the utility of log-transforming the variables. As when doing so by myself, I ran into some issues with some of the rows being "-inf", or N/A, which led to models not running. I also asked for some help on how to store the influence points ina. data-frame, to get a rough idea what the influence points were. I did this twice. Then a few times I asked questions related to wording, defining terms, or just attempting to get a better understanding of something I was trying to talk about, or use.

	(1)
(Intercept)	4.445***
	(0.263)
unempl_diff	−0.129
	(0.096)
median_diff	0.000***
	(0.000)
Pov_diff	0.164
	(0.147)
total_nonwhite_vap	0.000***
	(0.000)
unempl_diff × median_diff	0.000
	(0.000)
unempl_diff × Pov_diff	0.012
	(0.051)
median_diff × Pov_diff	0.000
	(0.000)
unempl_diff × total_nonwhite_vap	0.000***
	(0.000)
median_diff × total_nonwhite_vap	0.000
	(0.000)
Pov_diff × total_nonwhite_vap	0.000
	(0.000)
unempl_diff × median_diff × Pov_diff	0.000
	(0.000)
unempl_diff × median_diff × total_nonwhite_vap	0.000
	(0.000)
unempl_diff × Pov_diff × total_nonwhite_vap	0.000
	(0.000)
median_diff × Pov_diff × total_nonwhite_vap	0.000
	(0.000)
unempl_diff × median_diff × Pov_diff × total_nonwhite_vap	0.000
	(0.000)
Num.Obs.	3111
R2	0.084
R2 Adj.	0.080
AIC	19 271.5
BIC	19 374.2
Log.Lik.	−9618.744
RMSE	5.33

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 2: Comparing Models

	Basic	Clustered
(Intercept)	7.012***	7.012***
	(0.227)	(0.931)
unempl_diff	−0.263***	−0.263
	(0.065)	(0.256)
pct_urban	−0.077***	−0.077***
	(0.003)	(0.007)
median_diff	0.000***	0.000***
	(0.000)	(0.000)
Pov_diff	0.087+	0.087
	(0.050)	(0.067)
total_nonwhite_vap	0.000***	0.000**
	(0.000)	(0.000)
Num.Obs.	3111	3111
R2	0.230	0.230
R2 Adj.	0.229	0.229
F	185.466	
RMSE	4.89	4.89
Std.Errors		by: state_v2

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 3: Comparing Models

	Basic	Clustered
(Intercept)	−16.894***	−16.894***
	(0.927)	(2.320)
unempl_diff	−0.072	−0.072
	(0.059)	(0.273)
pct_urban	−0.040***	−0.040***
	(0.003)	(0.007)
prop_noncollege	28.059***	28.059***
	(1.061)	(3.021)
median_diff	0.000	0.000
	(0.000)	(0.000)
Pov_diff	0.160***	0.160*
	(0.046)	(0.070)
total_nonwhite_vap	0.000+	0.000+
	(0.000)	(0.000)
Num.Obs.	3111	3111
R2	0.372	0.372
R2 Adj.	0.370	0.370
F	305.914	
RMSE	4.41	4.41
Std.Errors		by: state_v2

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 4: Comparing Models

	Basic	Clustered
(Intercept)	−14.118***	−14.118***
	(0.943)	(2.046)
unempl_diff	0.294***	0.294
	(0.064)	(0.241)
pct_urban	−0.055***	−0.055***
	(0.003)	(0.008)
prop_manufacturing	9.104***	9.104*
	(1.383)	(3.603)
prop_agr	−12.392***	−12.392**
	(1.481)	(3.931)
prop_constr	−15.625***	−15.625**
	(3.747)	(5.742)
prop_noncollege	27.356***	27.356***
	(1.126)	(2.817)
median_diff	0.000	0.000
	(0.000)	(0.000)
Pov_diff	0.168***	0.168**
	(0.044)	(0.064)
total_nonwhite_vap	0.000	0.000
	(0.000)	(0.000)
Num.Obs.	3111	3111
R2	0.416	0.416
R2 Adj.	0.414	0.414
F	244.968	
RMSE	4.26	4.26
Std.Errors		by: state_v2

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Comparing Models

	Basic	Clustered
(Intercept)	−14.141***	−14.141***
	(0.941)	(2.043)
unempl_diff	0.295***	0.295
	(0.063)	(0.242)
pct_urban	−0.055***	−0.055***
	(0.003)	(0.008)
prop_manufacturing	9.042***	9.042*
	(1.385)	(3.600)
prop_agr	−12.451***	−12.451**
	(1.482)	(3.926)
prop_constr	−15.707***	−15.707**
	(3.749)	(5.702)
prop_noncollege	27.415***	27.415***
	(1.122)	(2.804)
median_diff	0.000	0.000
	(0.000)	(0.000)
Pov_diff	0.167***	0.167**
	(0.044)	(0.065)
black_vap	0.000	0.000
	(0.000)	(0.000)
Num.Obs.	3111	3111
R2	0.416	0.416
R2 Adj.	0.414	0.414
F	244.997	
RMSE	4.26	4.26
Std.Errors		by: state_v2

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 6: Comparing Models

	Basic	Clustered
(Intercept)	−14.169***	−14.169***
	(0.914)	(1.987)
unempl_diff	0.303***	0.303
	(0.063)	(0.244)
prop_manufacturing	9.266***	9.266**
	(1.373)	(3.571)
prop_agr	−12.249***	−12.249**
	(1.446)	(3.940)
prop_constr	−15.313***	−15.313**
	(3.730)	(5.917)
pct_urban	−0.056***	−0.056***
	(0.003)	(0.007)
prop_noncollege	27.430***	27.430***
	(1.090)	(2.724)
Pov_diff	0.164***	0.164**
	(0.042)	(0.061)
Num.Obs.	3111	3111
R2	0.415	0.415
R2 Adj.	0.414	0.414
F	314.607	
RMSE	4.26	4.26
Std.Errors		by: state_v2

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 7: Comparing Models

	Refreshed Basic	Refreshed Clustered
(Intercept)	−8.718***	−8.718**
	(1.664)	(3.278)
poverty12	−0.255***	−0.255***
	(0.045)	(0.067)
poverty16	−0.033	−0.033
	(0.047)	(0.068)
unempl_rate12	0.189**	0.189
	(0.061)	(0.222)
unempl_rate15	−0.074	−0.074
	(0.091)	(0.287)
median_income12	0.000	0.000
	(0.000)	(0.000)
median_income16	0.000***	0.000**
	(0.000)	(0.000)
pct_urban	−0.034***	−0.034***
	(0.003)	(0.007)
prop_noncollege	28.406***	28.406***
	(1.325)	(3.440)
black_vap	0.000	0.000
	(0.000)	(0.000)
Num.Obs.	3111	3111
R2	0.405	0.405
R2 Adj.	0.403	0.403
F	234.626	
RMSE	4.29	4.29
Std.Errors		by: state_v2

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 8: Comparing Models

	Refreshed Basic 2012	Refreshed Clustered 2016	Refreshed Basic 2016	Refreshed Clu
(Intercept)	−8.718***	−8.718**	−9.585***	−9.10
	(1.664)	(3.278)	(1.632)	(1.64
poverty12	−0.255***	−0.255***	−0.284***	
	(0.045)	(0.067)	(0.022)	
poverty16	−0.033	−0.033		−0.26
	(0.047)	(0.068)		(0.02
unempl_rate12	0.189**	0.189	0.165***	
	(0.061)	(0.222)	(0.035)	
unempl_rate15	−0.074	−0.074		0.155
	(0.091)	(0.287)		(0.05
median_income12	0.000	0.000	0.000***	
	(0.000)	(0.000)	(0.000)	
median_income16	0.000***	0.000**		0.000
	(0.000)	(0.000)		(0.00
pct_urban	−0.034***	−0.034***	−0.033***	−0.03
	(0.003)	(0.007)	(0.003)	(0.00
prop_noncollege	28.406***	28.406***	28.696***	28.254
	(1.325)	(3.440)	(1.310)	(1.32
black_vap	0.000	0.000	0.000	0.00
	(0.000)	(0.000)	(0.000)	(0.00
Num.Obs.	3111	3111	3111	311
R2	0.405	0.405	0.402	0.39
R2 Adj.	0.403	0.403	0.401	0.39
F	234.626		348.080	339.4
RMSE	4.29	4.29	4.30	4.3
Std.Errors		by: state_v2		

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

	(1)
(Intercept)	−10.061***
	(1.551)
poverty16	−0.169***
	(0.022)
unempl_rate15	−0.032
	(0.051)
median_income16	0.000***
	(0.000)
pct_urban	−0.039***
	(0.003)
prop_noncollege	27.020***
	(1.244)
rust_belt	3.685***
	(0.192)
Num.Obs.	3111
R2	0.460
R2 Adj.	0.459
AIC	17 608.4
BIC	17 656.7
Log.Lik.	−8796.188
F	441.340
RMSE	4.09

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

	(1)
(Intercept)	−9.185***
	(1.547)
rust_belt	3.347***
	(0.198)
prop_manufacturing	7.503***
	(1.166)
pct_urban	−0.041***
	(0.003)
poverty16	−0.158***
	(0.022)
unempl_rate15	−0.020
	(0.050)
median_income16	0.000***
	(0.000)
prop_noncollege	24.589***
	(1.293)
Num.Obs.	3111
R2	0.467
R2 Adj.	0.466
AIC	17 569.1
BIC	17 623.5
Log.Lik.	−8775.558
F	389.137
RMSE	4.06

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

	(1)
(Intercept)	−17.252***
	(0.858)
rust_belt	3.388***
	(0.195)
prop_manufacturing	7.682***
	(1.172)
pct_urban	−0.045***
	(0.003)
poverty16	−0.070***
	(0.013)
prop_noncollege	28.177***
	(1.138)
Num.Obs.	3111
R2	0.461
R2 Adj.	0.460
AIC	17 605.3
BIC	17 647.6
Log.Lik.	−8795.632
F	530.190
RMSE	4.09

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

	(1)
(Intercept)	−17.087***
	(0.909)
rust_belt	3.389***
	(0.195)
prop_manufacturing	7.780***
	(1.185)
pct_urban	−0.047***
	(0.005)
poverty16	−0.070***
	(0.013)
prop_noncollege	28.173***
	(1.138)
rural	−0.157
	(0.285)
Num.Obs.	3111
R2	0.461
R2 Adj.	0.460
AIC	17 607.0
BIC	17 655.3
Log.Lik.	−8795.481
F	441.776
RMSE	4.09

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

	(1)
(Intercept)	−17.252
	(0.858)
poverty16	−0.070
	(0.013)
prop_noncollege	28.177
	(1.138)
prop_manufacturing	7.682
	(1.172)
rust_belt	3.388
	(0.195)
pct_urban	−0.045
	(0.003)
Num.Obs.	3111
R2	0.461
R2 Adj.	0.460
AIC	17 605.3
BIC	17 647.6
Log.Lik.	−8795.632
F	530.190
RMSE	4.09

	(1)
(Intercept)	−17.252***
	(0.830)
poverty16	−0.070***
	(0.014)
rust_belt	3.388***
	(0.176)
pct_urban	−0.045***
	(0.003)
prop_manufacturing	7.682***
	(1.215)
prop_noncollege	28.177***
	(1.158)
Num.Obs.	3111
R2	0.461
R2 Adj.	0.460
AIC	17 605.3
BIC	17 647.6
RMSE	4.09

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

	(1)
(Intercept)	−14.169***
	(1.987)
unempl_diff	0.303
	(0.244)
prop_manufacturing	9.266**
	(3.571)
prop_agr	−12.249**
	(3.940)
prop_constr	−15.313**
	(5.917)
pct_urban	−0.056***
	(0.007)
prop_noncollege	27.430***
	(2.724)
Pov_diff	0.164**
	(0.061)
Num.Obs.	3111
R2	0.415
R2 Adj.	0.414
AIC	17 860.9
BIC	17 915.3
RMSE	4.26
Std.Errors	by: state_v2

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

	(1)
(Intercept)	−14.169***
	(0.834)
unempl_diff	0.303***
	(0.060)
prop_manufacturing	9.266***
	(1.438)
prop_agr	−12.249***
	(1.411)
prop_constr	−15.313***
	(3.792)
pct_urban	−0.056***
	(0.003)
prop_noncollege	27.430***
	(0.987)
Pov_diff	0.164***
	(0.041)
Num.Obs.	3111
R2	0.415
R2 Adj.	0.414
AIC	17 860.9
BIC	17 915.3
RMSE	4.26

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

	HC3 Robust	Unclustered	Clustered	Improved Model
Intercept	−17.252***	−14.169***	−14.169***	−14.169***
	(0.830)	(0.834)	(1.987)	(0.914)
Poverty Rate 2016	−0.070***			
	(0.014)			
Urban Pop. Percentage	−0.045***	−0.056***	−0.056***	−0.056***
	(0.003)	(0.003)	(0.007)	(0.003)
Proportion Non-College	28.177***	27.430***	27.430***	27.430***
	(1.158)	(0.987)	(2.724)	(1.090)
Rust Belt States	3.388***			
	(0.176)			
Proportion Agricultural Workers		−12.249***	−12.249**	−12.249***
		(1.411)	(3.940)	(1.446)
Proportion Manufacturing Workers	7.682***	9.266***	9.266**	9.266***
	(1.215)	(1.438)	(3.571)	(1.373)
Proportion Construction Workers		−15.313***	−15.313**	−15.313***
		(3.792)	(5.917)	(3.730)
Difference in Povery Rate in 2016 and 2012		0.164***	0.164**	0.164***
		(0.041)	(0.061)	(0.042)
Num.Obs.	3111	3111	3111	3111
R2	0.461	0.415	0.415	0.415
R2 Adj.	0.460	0.414	0.414	0.414
AIC	17 605.3	17 860.9	17 860.9	17 860.9
BIC	17 647.6	17 915.3	17 915.3	17 915.3
Log.Lik.				−8921.461
F				314.607
RMSE	4.09	4.26	4.26	4.26
Std.Errors			by: state_v2	

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

	(1)
(Intercept)	−16.560
	(0.681)
poverty16	−0.103
	(0.011)
rust_belt	3.105
	(0.152)
prop_manufacturing	7.756
	(0.930)
pct_urban	−0.042
	(0.002)
prop_noncollege	28.202
	(0.909)
Num.Obs.	2987
R2	0.574
R2 Adj.	0.573
AIC	15 288.1
BIC	15 330.1
Log.Lik.	−7637.065
F	803.212
RMSE	3.12

	No Influential Points	Original Model	HC3
(Intercept)	−16.560***	−17.252***	−17.252***
	(0.681)	(0.858)	(0.830)
poverty16	−0.103***	−0.070***	−0.070***
	(0.011)	(0.013)	(0.014)
rust_belt	3.105***	3.388***	3.388***
	(0.152)	(0.195)	(0.176)
prop_manufacturing	7.756***	7.682***	7.682***
	(0.930)	(1.172)	(1.215)
pct_urban	−0.042***	−0.045***	−0.045***
	(0.002)	(0.003)	(0.003)
prop_noncollege	28.202***	28.177***	28.177***
	(0.909)	(1.138)	(1.158)
R2	0.574	0.461	0.461
R2 Adj.	0.573	0.460	0.460

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001