Socioeconomic Determinants of 2020 U.S. Presidential Election County-Level Voter Turnout

Yuen Ler Chow, John Rho, and Henry Wu

December 20, 2024

Contents

1 Introduction and Motivation

2 Data Description and EDA

There are a few different data sources joined together to make this dataset. The turnout rate data is calculating by dividing the voter turnout for the 2020 presidential election in each county (from the MIT Election Lab) by the voting-eligible population (U.S. citizens age 18 and up) according to the 2020 5-year American Community Survey released by the U.S. Census Bureau. The resulting turnout rate should be a proportion between 0 and 1. The exception for the voter turnout data is Alaska, whose voter turnout data is organized by election districts instead of borough and Census areas (Alaska's county equivalents). To have this data be consistent with the predictor variables, I got estimates for Alaska voter turnout data by borough and Census area from a blog post.

The predictors (county-level demographic and socioeconomic characteristics) are from Opportunity Insights, a Harvard-based research lab studying economic opportunity in the United States. Descriptions of the variables can be found here. Datasets for FIPS state and county codes are also used to merge the data sources.

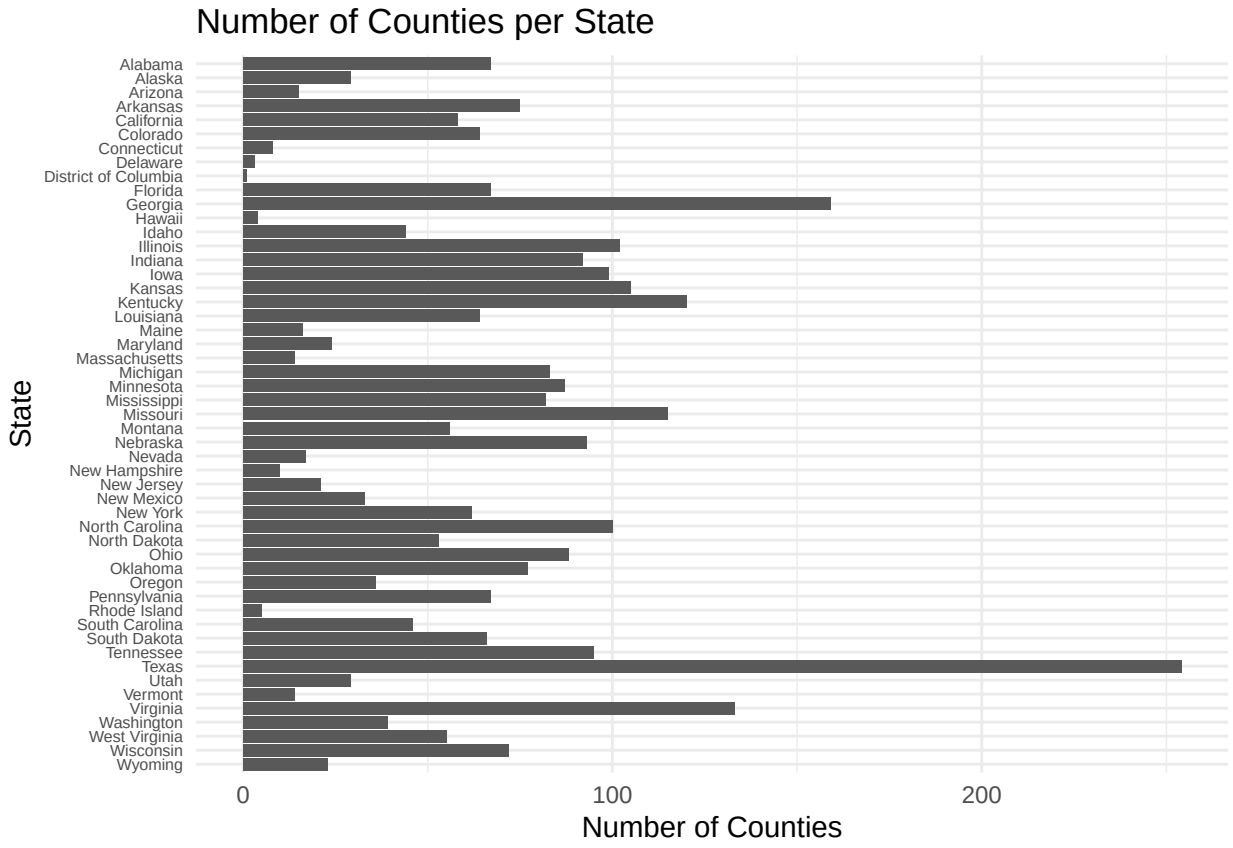
2.1 Descriptive Statistics

The dataset contains 3,141 observations and 21 variables, 19 of which (all except county and FIPS code) will be used as predictors. All predictors except state are continuous. For each of our continuous predictors, we summarize the number of missing values, the mean, median, standard deviation, interquartile range, minimum value, and maximum value (Table 1).

Include descriptions of variables!!

Variable	Missing	Mean	Median	SD	IQR	Min	Max
frac_coll_plus2010	0	0.19	0.17	0.09	0.09	0.04	0.71
foreign_share2010	0	0.04	0.02	0.06	0.04	0	0.72
med_hhinc2016	1	4.9e+04	4.71e+04	1.34e+04	1.47e+04	2.02e+04	1.29e+05
poor_share2010	0	0.16	0.15	0.06	0.08	0	0.53
share_white2010	0	0.78	0.86	0.2	0.27	0.03	0.99
share_black2010	0	0.09	0.02	0.15	0.1	0	0.86
share_hisp2010	0	0.08	0.03	0.13	0.07	0	0.96
share_asian2010	21	0.01	0	0.02	0.01	0	0.43
gsmn_math_g3_2013	73	3.21	3.24	0.78	0.98	−0.66	6.58
rent_twobed2015	76	692	643	205	196	236	2.09e+03
singleparent_share2010	0	0.31	0.3	0.09	0.1	0	0.81
traveltime15_2010	0	0.4	0.38	0.14	0.19	0.1	0.99
emp2000	0	0.57	0.58	0.08	0.1	0.24	0.84
ln_wage_growth_hs_grad	684	0.08	0.07	0.14	0.13	−0.72	0.91
popdensity2010	1	263	45.3	1.77e+03	96.7	0.04	7.06e+04
ann_avg_job_growth_2004_2013	5	0	0	0.01	0.02	−0.08	0.12
job_density_2013	2	124	18.5	863	43.3	0.02	3.67e+04
turnout.rate	0	0.66	0.66	0.11	0.14	0.19	1.58

Table 1: Descriptive statistics of the continuous predictors in the dataset.



2.2 Pre-Processing

Most variables have either zero or a small fraction of observations missing. The exception is `ln_wage_growth_hs_grad`, which has 21.8% of its observations missing. To handle the missing data, we drop the `ln_wage_growth_hs_grad` variable altogether and drop the counties that have missing data in at least one of the remaining variables, leaving 2,999 observations and 18 predictors.

There is one invalid value for turnout rate greater than 1, so we set it equal to 1. There are a few invalid values for mean math scores less than 0, so we set them equal to 0.

2.3 Exploratory Graphs

2.3.1 Turnout Rate

The histogram of turnout rate (Figure 1) shows an approximately normal distribution.

2.3.2 Poverty Rate and Turnout Rate

To see the relationship between voter turnout and one predictor variable hypothesized to be associated with it, we plot the 2010 poverty rate against the 2020 turnout rate for each county (Figure 2). There is a moderate to strong negative association between the variables ($r = -0.571$).

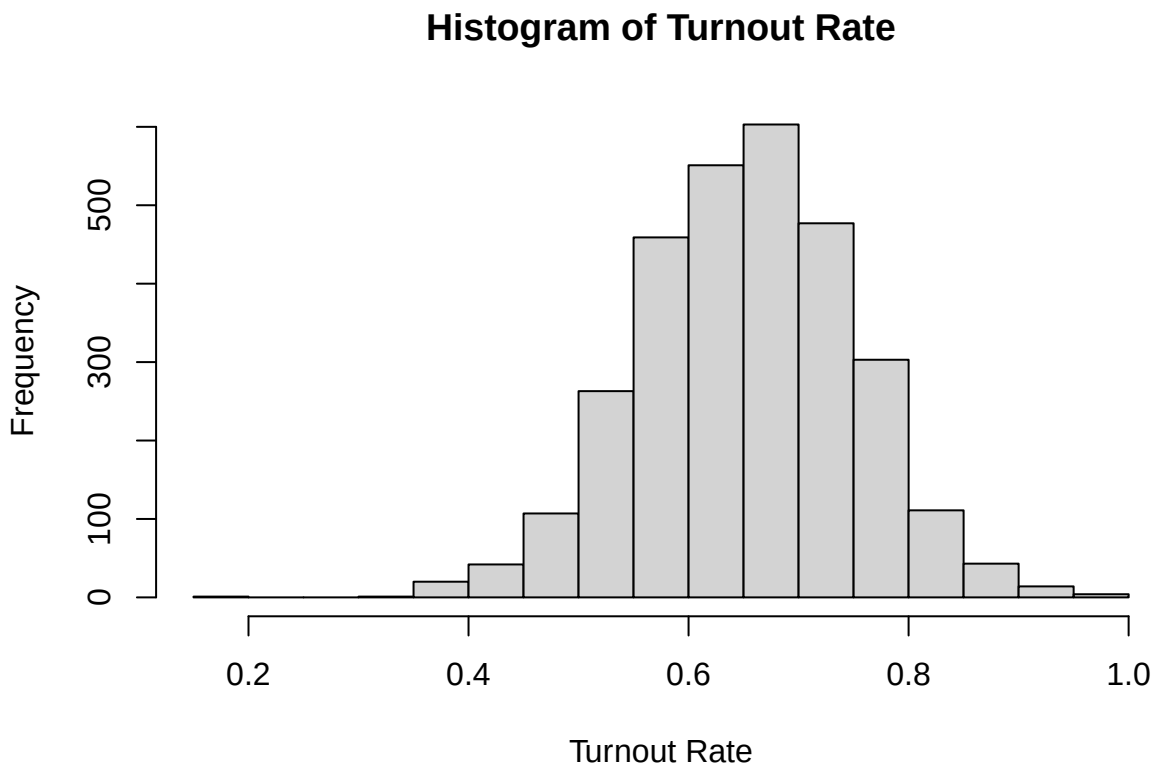


Figure 1: Histogram of turnout rate for counties shows an approximately normal distribution.

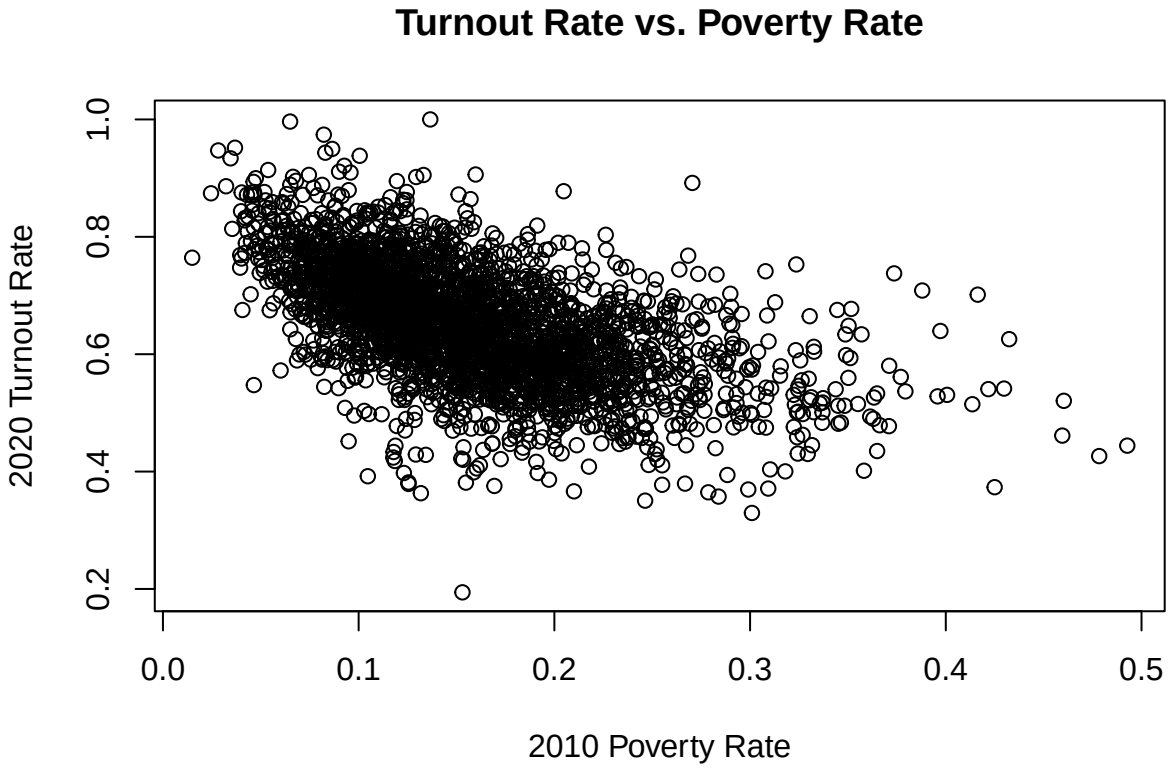


Figure 2: Poverty rate vs. turnout rate.

3 Methods

4 Results

4.1 Baseline Model

We check that our hypothesis that the turnout rate can be predicted from county demographics is reasonable by fitting a baseline linear regression model containing all predictors except state and no interaction terms. The model output is shown in Table 2.

Variable	Estimate	SE	t-value	p-value
(Intercept)	0.608	0.0329	18.5	2.4e−72
frac_coll_plus2010	0.372	0.026	14.3	5.44e−45
foreign_share2010	0.11	0.0496	2.23	0.026
med_hhinc2016	1.3e−07	2.54e−07	0.51	0.61
poor_share2010	−0.576	0.0404	−14.3	1.08e−44
share_white2010	0.0423	0.0208	2.03	0.0421
share_black2010	0.0591	0.0208	2.83	0.00462
share_hisp2010	−0.0511	0.0249	−2.06	0.0398
share_asian2010	−0.506	0.0917	−5.52	3.71e−08
gsmn_math_g3_2013	−0.000993	0.00214	−0.464	0.643
rent_twobed2015	−7.02e−06	1.43e−05	−0.492	0.623
singleparent_share2010	−0.0617	0.0246	−2.51	0.012
traveltime15_2010	−0.0425	0.0117	−3.63	0.000283
emp2000	0.114	0.0273	4.17	3.17e−05
popdensity2010	−2.26e−07	5.51e−06	−0.041	0.967
ann_avg_job_growth_2004_2013	−0.707	0.107	−6.62	4.12e−11
job_density_2013	−4.92e−06	1.13e−05	−0.434	0.665

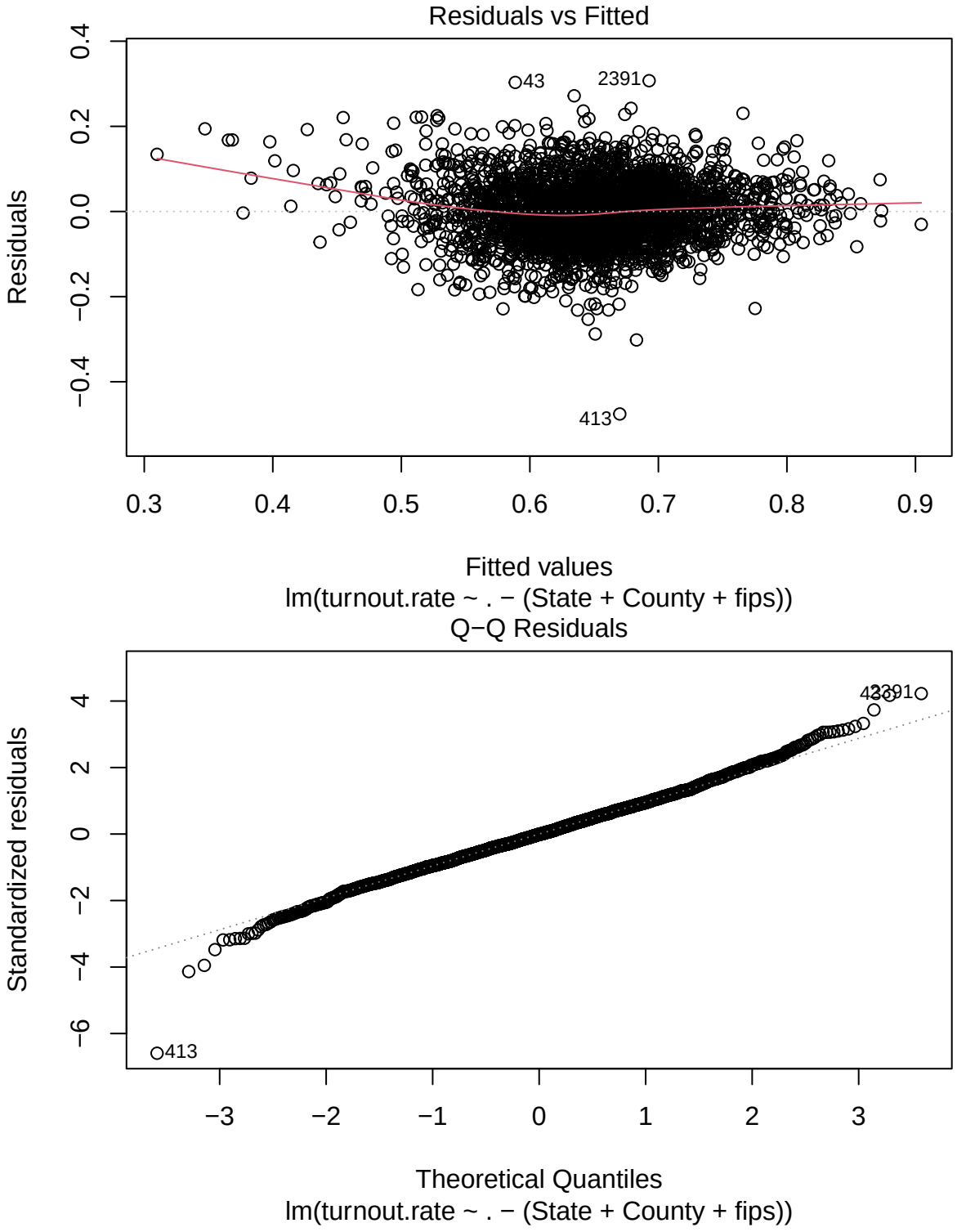
Table 2: Baseline model output.

The baseline model demonstrates several significant relationships. The model explains approximately 44% of the variance in turnout rates ($R^2 = 0.442$) and is highly significant ($F = 147.4$, $p < 2.2 \times 10^{-16}$). Education emerges as a strong positive predictor, with a one percentage point increase in college education associated with a 0.372 percentage point increase in turnout ($p < 0.001$). Other significant positive predictors include foreign-born share ($\hat{\beta} = 0.110$), white population share ($\hat{\beta} = 0.042$), black population share ($\hat{\beta} = 0.059$), and employment ($\hat{\beta} = 0.114$). Several factors show significant negative associations with turnout: poverty rate exhibits a strong negative effect ($\hat{\beta} = -0.576$), as do Hispanic population share ($\hat{\beta} = -0.051$), Asian population share ($\hat{\beta} = -0.506$), single parent share ($\hat{\beta} = -0.062$), travel time ($\hat{\beta} = -0.043$), and job growth ($\hat{\beta} = -0.707$). Lastly, several variables, including median household income, math scores, two-bedroom rent,

population density, and job density, did not show significant relationships with turnout ($p > 0.05$).

4.1.1 Diagnostics

We conduct diagnostics for the baseline model to assess its suitability for linear regression.



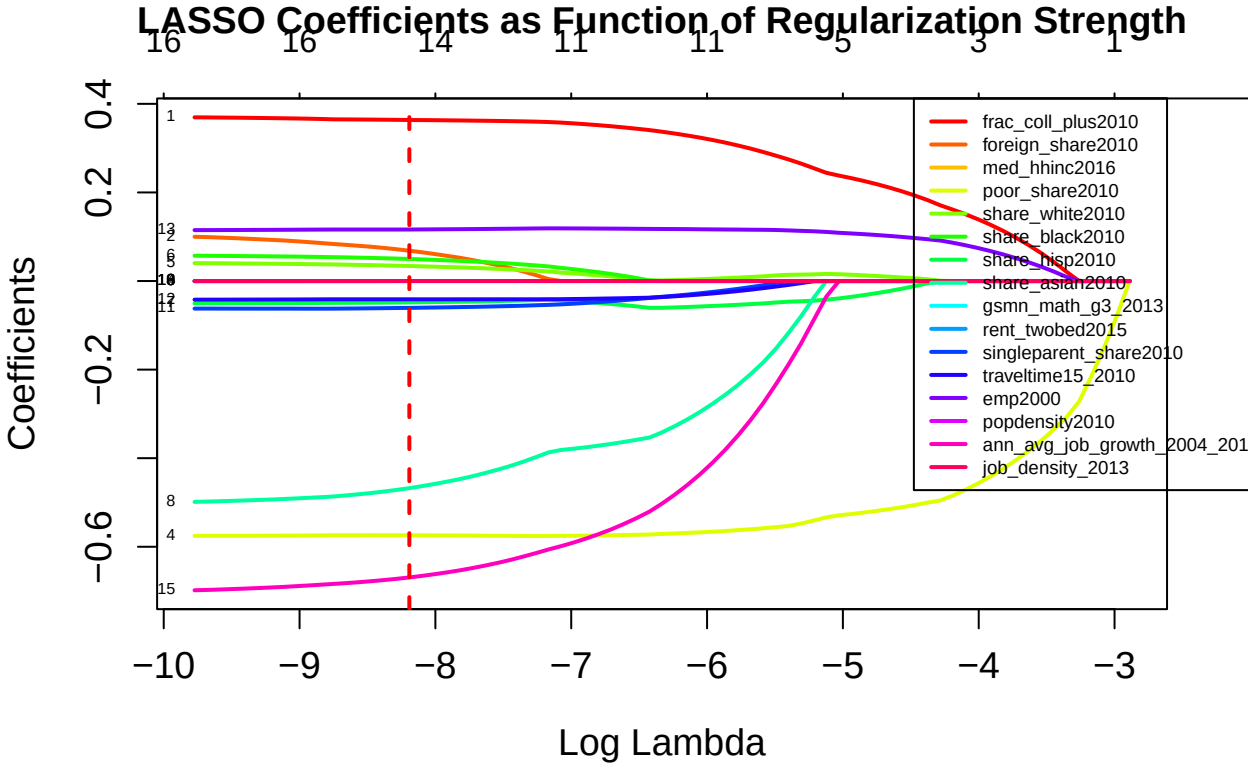
- **Existence of Variance:** The spread of residuals in the plots demonstrates clear variation in our

dependent variable, confirming the existence of variance. The residuals show a reasonable spread around zero, with most falling between -0.2 and 0.2, indicating that our model has captured meaningful variation in the data while maintaining reasonable error terms.

- **Linearity:** The Residuals vs Fitted plot reveals a relatively flat red line hovering around zero, suggesting the linearity assumption is reasonably met. While there is some pattern in the spread of residuals, the scatter appears generally random. The plot identifies points 43, 2391, and 413 as potential outliers that warrant further investigation. Overall, the linearity assumption appears to be satisfied, though with some potential concerns that might need additional examination.
- **Independence:** Independence cannot be directly assessed from these diagnostic plots alone. Given that this analysis uses county-level data, there is likely spatial correlation present between neighboring counties. Additional specific tests would be necessary to evaluate this assumption, such as Moran's I for spatial autocorrelation. We hope to ask a Teaching Fellow about further analysis regarding this possible violation of our assumptions.
- **Homogeneity (Homoscedasticity):** Examining the Residuals vs Fitted plot, we observe a fanning pattern where the spread of residuals is wider in the middle range of fitted values. This pattern suggests the presence of heteroscedasticity, meaning the variance of residuals is not constant across all fitted values. This violation of the homoscedasticity assumption suggests we should consider using robust standard errors or weighted least squares estimation methods to address this issue.
- **Normality:** The Q-Q plot provides a visual assessment of normality by comparing the standardized residuals against theoretical normal quantiles. The majority of points follow the diagonal line, suggesting approximate normality in the central region of the distribution. However, we observe some deviation at both tails, particularly with point 413 showing as a significant lower outlier and points near 43 & 2391 deviating at the upper tail. Given our large sample size, the Central Limit Theorem suggests that these deviations from normality are less concerning for inference purposes.

4.2 Regularization

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                                          s0
## (Intercept)                    6.121987e-01
## frac_coll_plus2010             3.632965e-01
## foreign_share2010              7.023135e-02
## med_hhinc2016                  4.442856e-08
## poor_share2010                -5.736737e-01
## share_white2010                3.394170e-02
## share_black2010                4.990823e-02
## share_hisp2010                -4.857310e-02
## share_asian2010               -4.695689e-01
## gsmn_math_g3_2013              .
## rent_twobed2015                .
## singleparent_share2010        -6.054149e-02
## traveltime15_2010             -4.123835e-02
## emp2000                        1.160911e-01
## popdensity2010                -8.432251e-07
## ann_avg_job_growth_2004_2013 -6.699229e-01
## job_density_2013              -3.165965e-06
```



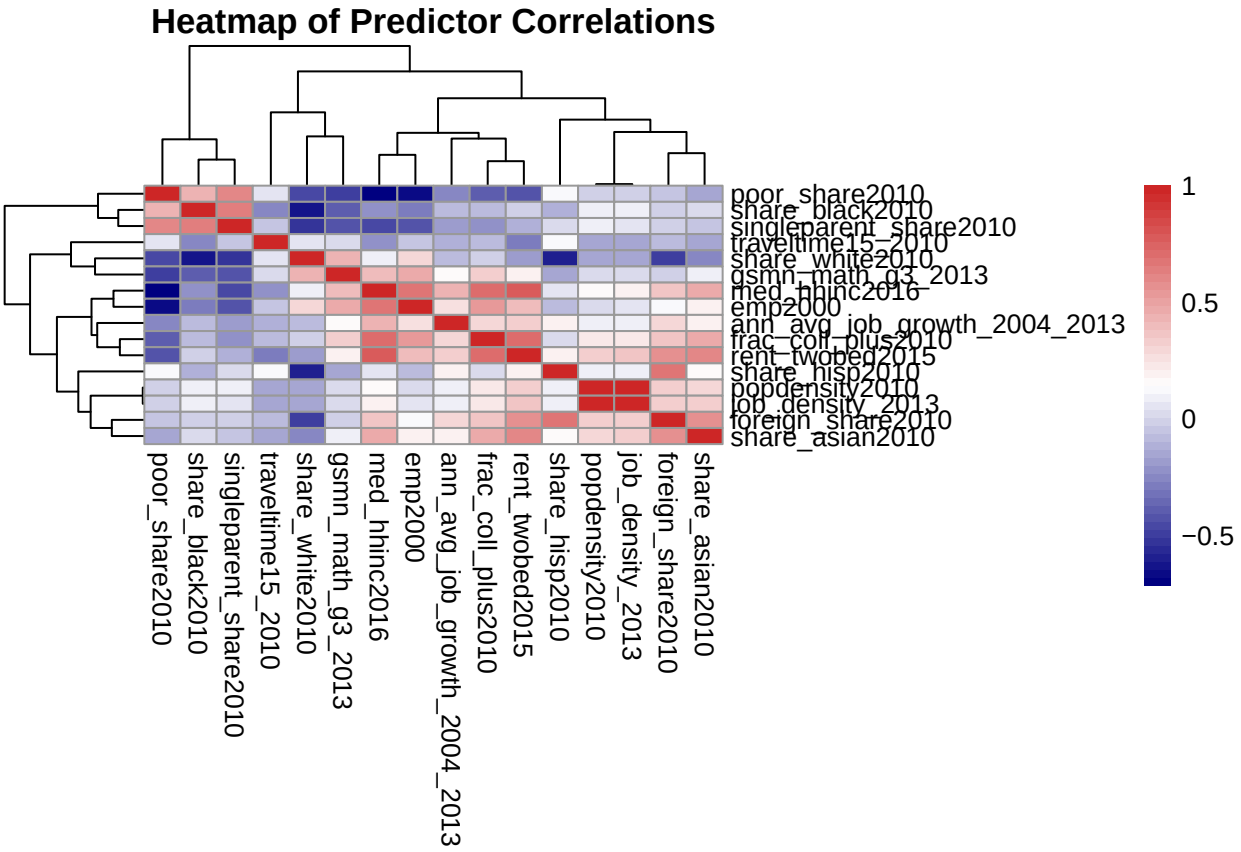

```
## Predictors kept by LASSO:

##  [1] "frac_coll_plus2010"           "foreign_share2010"
##  [3] "med_hhinc2016"                "poor_share2010"
##  [5] "share_white2010"              "share_black2010"
##  [7] "share_hisp2010"               "share_asian2010"
##  [9] "singleparent_share2010"       "traveltime15_2010"
## [11] "emp2000"                      "popdensity2010"
## [13] "ann_avg_job_growth_2004_2013" "job_density_2013"

##
## Predictors removed by LASSO (coefficients = 0):

## [1] "gsmn_math_g3_2013" "rent_twobed2015"
```

Running the LASSO regularization found an optimal lambda that zeroed out the features gsmn_math_g3_2013 as well as rent_twobed2015. This seems to align with our preliminary analysis that showed that both of these predictors had p-values that were far above our p=0.05 threshold, both at p>0.60.



Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2.98e+03	15.9				
2.98e+03	15.4	1	0.491	94.9	4.27e−22

P value 2.2e-16 signifcant.. means that interaction term is significant and helps to explain the variance in the model.

```
##
## Call:
## lm(formula = turnout.rate ~ . - State - County - fips - gsmn_math_g3_2013 -
##     rent_twobed2015 + factor(State), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42729 -0.03049 -0.00122  0.03066  0.29353
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.336e-01  2.979e-02  17.912  < 2e-16 ***
## frac_coll_plus2010             2.398e-01  2.109e-02  11.373  < 2e-16 ***
## foreign_share2010             -5.603e-02  4.458e-02  -1.257 0.208914
## med_hhinc2016                  9.682e-07  2.088e-07   4.636 3.70e-06 ***
## poor_share2010                -3.474e-01  3.486e-02  -9.966  < 2e-16 ***
## share_white2010                9.735e-02  1.979e-02   4.918 9.21e-07 ***
## share_black2010                1.042e-01  2.123e-02   4.908 9.73e-07 ***
```

```
## share_hisp2010                     3.103e-03  2.489e-02   0.125 0.900789
## share_asian2010                   -5.204e-01  9.504e-02  -5.476 4.72e-08 ***
## singleparent_share2010            -8.936e-02  2.048e-02  -4.363 1.32e-05 ***
## traveltime15_2010                 -9.101e-02  1.159e-02  -7.851 5.76e-15 ***
## emp2000                            1.104e-01  2.520e-02   4.382 1.22e-05 ***
## popdensity2010                     5.026e-06  4.507e-06   1.115 0.264946
## ann_avg_job_growth_2004_2013      -6.982e-01  9.184e-02  -7.603 3.88e-14 ***
## job_density_2013                  -1.071e-05  9.239e-06  -1.159 0.246589
## factor(State)Alaska                1.608e-02  1.842e-02   0.873 0.382787
## factor(State)Arizona               5.215e-02  1.747e-02   2.985 0.002861 **
## factor(State)Arkansas             -9.944e-02  9.917e-03 -10.027  < 2e-16 ***
## factor(State)California            4.921e-02  1.212e-02   4.062 5.00e-05 ***
## factor(State)Colorado              8.884e-02  1.148e-02   7.740 1.36e-14 ***
## factor(State)Connecticut          -4.244e-02  2.227e-02  -1.905 0.056830 .
## factor(State)Delaware              2.913e-03  3.433e-02   0.085 0.932388
## factor(State)District of Columbia -4.882e-02  5.958e-02  -0.819 0.412677
## factor(State)Florida               4.820e-02  1.050e-02   4.591 4.61e-06 ***
## factor(State)Georgia              -1.941e-02  8.546e-03  -2.271 0.023205 *
## factor(State)Hawaii                1.088e-01  3.992e-02   2.726 0.006444 **
## factor(State)Idaho                 4.300e-02  1.206e-02   3.566 0.000368 ***
## factor(State)Illinois             -2.579e-02  9.701e-03  -2.658 0.007893 **
## factor(State)Indiana              -7.224e-02  9.951e-03  -7.259 4.97e-13 ***
## factor(State)Iowa                  3.768e-02  1.008e-02   3.738 0.000189 ***
## factor(State)Kansas               -1.577e-02  1.011e-02  -1.560 0.118888
## factor(State)Kentucky             -1.587e-02  9.438e-03  -1.682 0.092771 .
## factor(State)Louisiana             1.088e-03  1.026e-02   0.106 0.915566
## factor(State)Maine                 7.530e-02  1.659e-02   4.539 5.89e-06 ***
## factor(State)Maryland             -4.177e-02  1.432e-02  -2.918 0.003554 **
## factor(State)Massachusetts         2.107e-02  1.785e-02   1.181 0.237859
## factor(State)Michigan              5.593e-02  1.006e-02   5.561 2.93e-08 ***
## factor(State)Minnesota             6.894e-02  1.032e-02   6.682 2.81e-11 ***
## factor(State)Mississippi           1.263e-03  9.723e-03   0.130 0.896657
## factor(State)Missouri             -3.106e-02  9.426e-03  -3.295 0.000997 ***
## factor(State)Montana               7.623e-02  1.243e-02   6.133 9.75e-10 ***
## factor(State)Nebraska              2.503e-02  1.080e-02   2.318 0.020507 *
## factor(State)Nevada                5.369e-02  1.873e-02   2.866 0.004183 **
## factor(State)New Hampshire         6.395e-03  2.012e-02   0.318 0.750609
## factor(State)New Jersey            2.675e-02  1.563e-02   1.711 0.087174 .
## factor(State)New Mexico            2.117e-02  1.497e-02   1.414 0.157338
## factor(State)New York             -5.491e-02  1.112e-02  -4.939 8.30e-07 ***
## factor(State)North Carolina        4.468e-02  9.253e-03   4.829 1.44e-06 ***
## factor(State)North Dakota         -1.759e-03  1.251e-02  -0.141 0.888178
## factor(State)Ohio                 -2.654e-02  9.949e-03  -2.668 0.007670 **
## factor(State)Oklahoma             -8.357e-02  1.053e-02  -7.934 2.99e-15 ***
## factor(State)Oregon                1.182e-01  1.286e-02   9.191  < 2e-16 ***
## factor(State)Pennsylvania         -1.508e-02  1.052e-02  -1.433 0.152074
## factor(State)Rhode Island         -5.543e-02  2.740e-02  -2.023 0.043192 *
## factor(State)South Carolina        9.969e-04  1.132e-02   0.088 0.929836
## factor(State)South Dakota          3.214e-03  1.180e-02   0.273 0.785250
## factor(State)Tennessee            -6.913e-02  9.586e-03  -7.211 7.02e-13 ***
## factor(State)Texas                -2.023e-02  9.300e-03  -2.175 0.029692 *
## factor(State)Utah                  4.048e-02  1.402e-02   2.888 0.003910 **
## factor(State)Vermont               3.112e-02  1.812e-02   1.718 0.085953 .
## factor(State)Virginia              6.502e-03  8.997e-03   0.723 0.469910
```

```
## factor(State)Washington            1.113e-01  1.247e-02   8.924  < 2e-16 ***
## factor(State)West Virginia        -8.496e-02  1.108e-02  -7.668 2.35e-14 ***
## factor(State)Wisconsin             5.134e-02  1.060e-02   4.841 1.36e-06 ***
## factor(State)Wyoming              -9.239e-03  1.497e-02  -0.617 0.537151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.058 on 2934 degrees of freedom
## Multiple R-squared:  0.6536, Adjusted R-squared:  0.646
## F-statistic: 86.49 on 64 and 2934 DF,  p-value: < 2.2e-16
```

5 Discussion and Conclusion

6 Bibliography

7 Appendix