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

Exploratory Data Analysis

Yuen Ler Chow, John Rho, and Henry Wu

Data Description

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.

Setup

```r
rm(list = ls())
require(readr)
require(tidyr)
require(dplyr)
require(knitr)
```

```r
data <- read.csv("../data/processed/data.csv")
head(data)
```

```
##     State         County fips frac_coll_plus2010 foreign_share2010
## 1 Alabama Autauga County 1001         0.22199036       0.020154603
## 2 Alabama Baldwin County 1003         0.26071036       0.037591625
## 3 Alabama Barbour County 1005         0.13349621       0.028143950
## 4 Alabama    Bibb County 1007         0.09924053       0.006859188
## 5 Alabama  Blount County 1009         0.12633450       0.047343444
## 6 Alabama Bullock County 1011         0.10972187       0.013493270
##   med_hhinc2016 poor_share2010 share_white2010 share_black2010 share_hisp2010
## 1      54052.80      0.1059177       0.7724616      0.18134174     0.02400542
## 2      52003.09      0.1229422       0.8350479      0.09752284     0.04384824
## 3      33114.85      0.2506308       0.4675311      0.47190151     0.05051535
## 4      39846.45      0.1268499       0.7502073      0.22282349     0.01771765
## 5      46361.12      0.1331379       0.8888734      0.01500297     0.08070200
## 6      31304.78      0.2804486       0.2191680      0.70221734     0.07119296
```

```
##   share_asian2010 gsmn_math_g3_2013 rent_twobed2015 singleparent_share2010
## 1    0.0078302799          2.759864        739.3654              0.2833759
## 2    0.0059535136          2.792510        816.8452              0.2778664
## 3    0.0036882064          1.600009        527.2908              0.4680706
## 4    0.0007418721          1.531674        604.2776              0.3201363
## 5    0.0018735955          2.815403        567.6959              0.2589052
## 6    0.0017932489          1.039439        266.0000              0.5778636
##   traveltime15_2010   emp2000 ln_wage_growth_hs_grad popdensity2010
## 1         0.2041625 0.6095865            -0.06331379       91.80268
## 2         0.2753262 0.5770263             0.03009291      114.64751
## 3         0.3760492 0.4532710             0.18936642       31.02921
## 4         0.2526830 0.4942406            -0.02007263       36.80634
## 5         0.1943438 0.5778096             0.09646260       88.90219
## 6         0.3921350 0.3746639             0.36383346       17.52395
##   ann_avg_job_growth_2004_2013 job_density_2013 turnout.rate
## 1                  0.010145103        40.719135    0.6618366
## 2                  0.012950056        50.085987    0.6529056
## 3                 -0.020755908         9.230672    0.5402712
## 4                 -0.004644653        12.875392    0.5456975
## 5                 -0.008120399        36.175354    0.6419098
## 6                  0.026254078         6.954023    0.5908043
```

Descriptive Statistics

We have no categorical variables. For each of our continuous variables, we summarize the number of missing values, the mean, median, standard deviation, interquartile range, minimum value, and maximum value.

```r
predictors <- names(data)[!(names(data) %in% c('State', 'County', 'fips'))]
summary_table <- data.frame()

for (predictor in predictors) {
  column <- data[[predictor]]
  num_missing <- sum(is.na(column))
  mean_var <- mean(column, na.rm = TRUE)
  median_var <- median(column, na.rm = TRUE)
  sd_var <- sd(column, na.rm = TRUE)
  iqr_var <- IQR(column, na.rm = TRUE)
  min_var <- min(column, na.rm = TRUE)
  max_var <- max(column, na.rm = TRUE)

  summary_table <- rbind(summary_table, data.frame(
    Variable = predictor,
    Missing = num_missing,
    Mean = round(mean_var, 2),
    Median = round(median_var, 2),
    SD = round(sd_var, 2),
    IQR = round(iqr_var, 2),
    Min = round(min_var, 2),
    Max = round(max_var, 2)
  ))
}

kable(summary_table)
```

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	0.00	0.72
med_hhinc2016	1	48980.92	47127.10	13398.03	14687.30	20170.89	129150.34
poor_share2010	0	0.16	0.15	0.06	0.08	0.00	0.53
share_white2010	0	0.78	0.86	0.20	0.27	0.03	0.99
share_black2010	0	0.09	0.02	0.15	0.10	0.00	0.86
share_hisp2010	0	0.08	0.03	0.13	0.07	0.00	0.96
share_asian2010	21	0.01	0.00	0.02	0.01	0.00	0.43
gsmn_math_g3_2013	73	3.21	3.24	0.78	0.98	-0.66	6.58
rent_twobed2015	76	692.34	642.51	205.04	195.93	236.00	2085.23
singleparent_share2010	0	0.31	0.30	0.09	0.10	0.00	0.81
traveltime15_2010	0	0.40	0.38	0.14	0.19	0.10	0.99
emp2000	0	0.57	0.58	0.08	0.10	0.24	0.84
ln_wage_growth_hs_grad	684	0.08	0.07	0.14	0.13	-0.72	0.91
popdensity2010	1	262.67	45.30	1774.99	96.74	0.04	70583.63
ann_avg_job_growth_2004_2013	5	0.00	0.00	0.01	0.02	-0.08	0.12
job_density_2013	2	124.24	18.47	862.85	43.30	0.02	36663.16
turnout.rate	0	0.66	0.66	0.11	0.14	0.19	1.58

```
dim(data)
```

```
## [1] 3141   21
```

Missingness

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.

```
data <- select(data, -ln_wage_growth_hs_grad)
data <- subset(data, apply(data, 1, FUN = function(x) {!any(is.na(x))}))
dim(data)
```

```
## [1] 2999   20
```

Exploratory Graphs

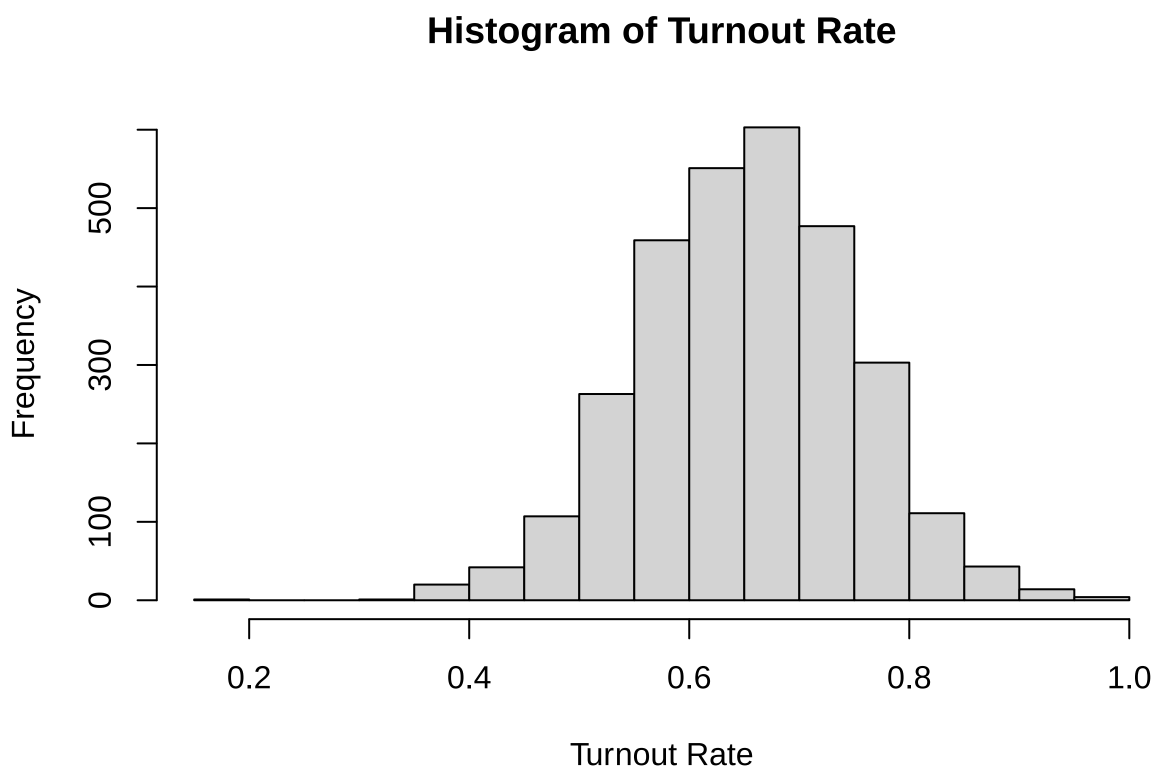
Turnout Rate

There is one invalid value for turnout rate greater than 1, so we set it equal to 1. The histogram shows that the turnout rates are approximately normally distributed.

```
subset(data, turnout.rate > 1)[c('State', 'County', 'turnout.rate')]
```

```
##              State         County turnout.rate
## 2391 South Dakota Hanson County     1.006321
```

```
data <- data %>%
  mutate(turnout.rate = case_when(
    turnout.rate > 1 ~ 1,
    .default = turnout.rate
  ))
hist(data$turnout.rate, main = 'Histogram of Turnout Rate', xlab = 'Turnout Rate')
```



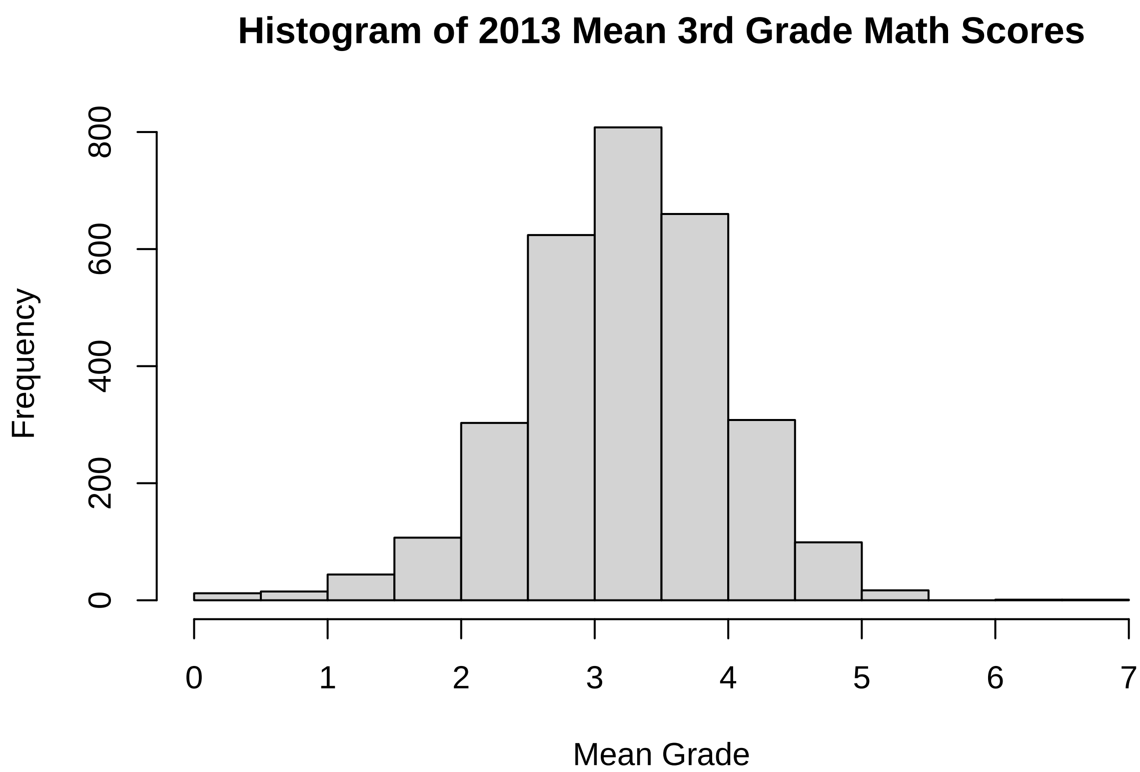
Math Scores

There are a few invalid values for mean math scores less than 0, so we set them equal to 0. The histogram shows that mean math scores are approximately normally distributed.

```r
subset(data, gsmn_math_g3_2013 < 0)[c('State', 'County', 'gsmn_math_g3_2013')]
```

```
##          State              County gsmn_math_g3_2013
## 71      Alaska Bethel Census Area        -0.6613265
## 1145 Louisiana      Madison Parish        -0.5496646
## 1158 Louisiana  St. Helena Parish        -0.4451542
```

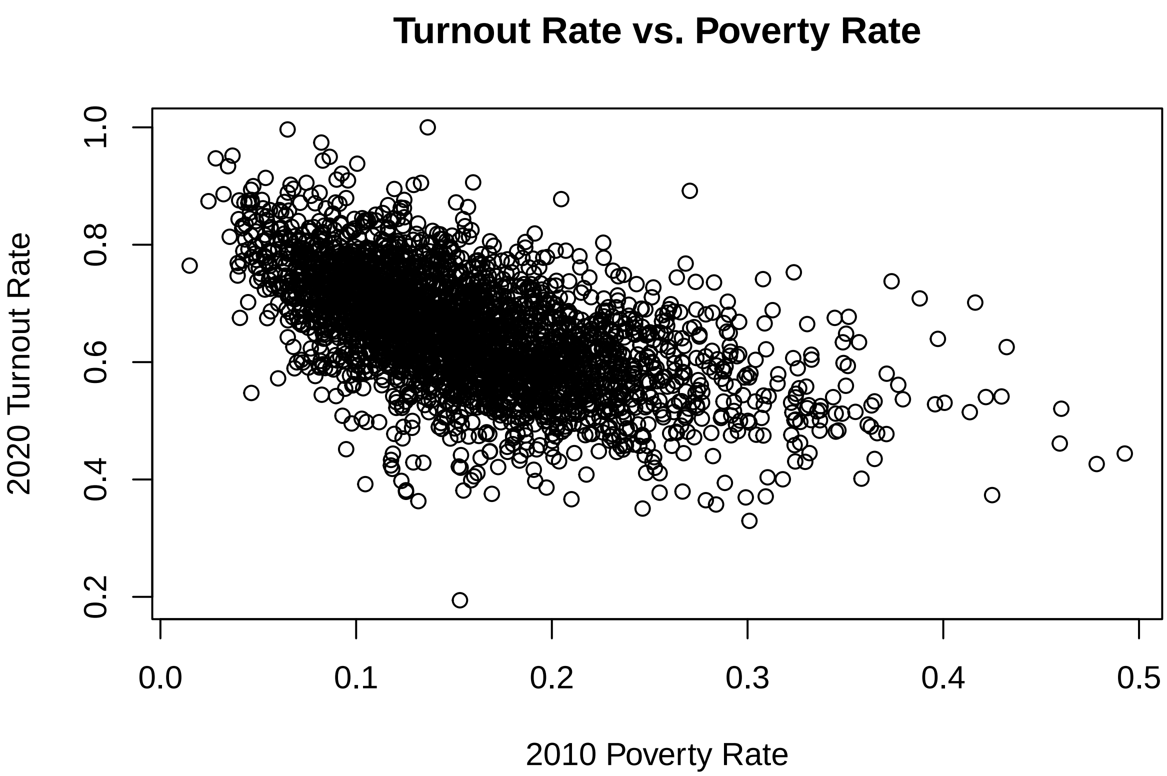
```r
data <- data %>%
  mutate(gsmn_math_g3_2013 = case_when(
    gsmn_math_g3_2013 < 0 ~ 0,
    .default = gsmn_math_g3_2013
  ))
hist(data$gsmn_math_g3_2013, main = 'Histogram of 2013 Mean 3rd Grade Math Scores', xlab = 'Mean Grade')
```



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. There is a strong negative trend in the plot.

```
plot(data$poor_share2010, data$turnout.rate, main = 'Turnout Rate vs. Poverty Rate', xlab = '2010 Pover
```



Preliminary Model

We also check that our hypothesis that the turnout rate can be predicted from county demographics is reasonable by fitting a linear regression model.

```r
lm_model <- lm(turnout.rate ~ . - (State + County + fips), data = data)
summary(lm_model)
```

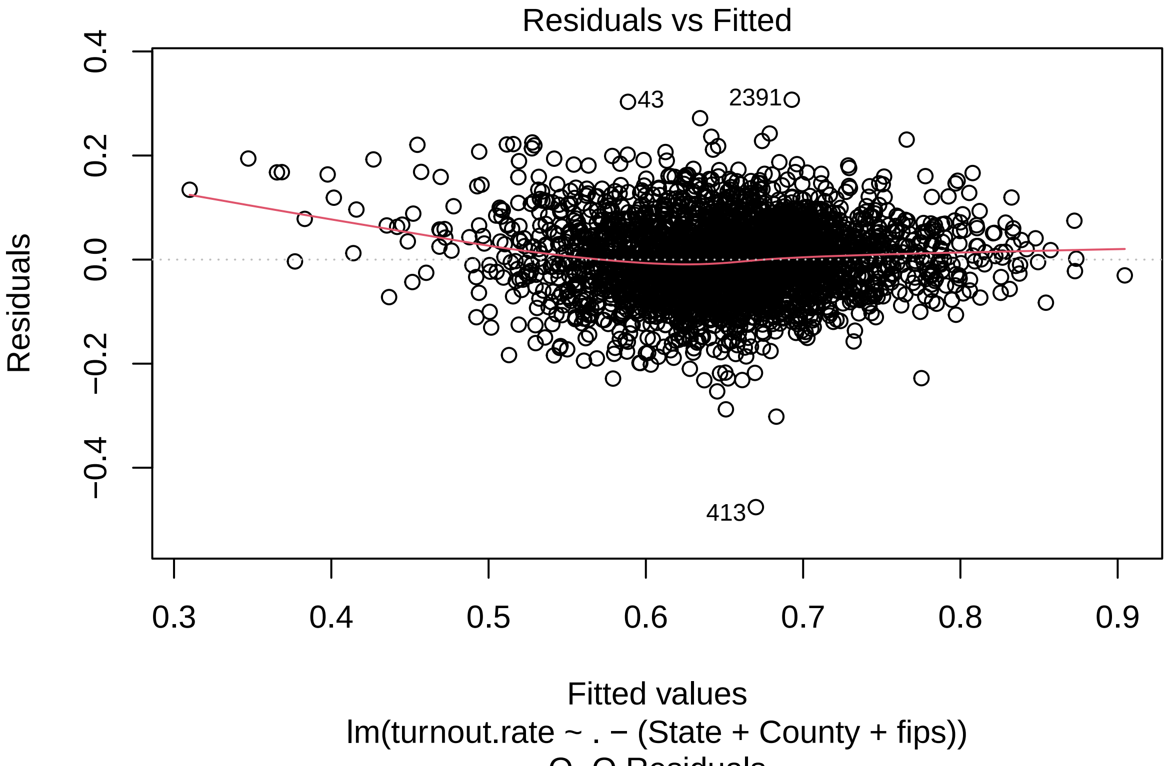
```
##
## Call:
## lm(formula = turnout.rate ~ . - (State + County + fips), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47572 -0.04720 -0.00065  0.04700  0.30720
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   6.083e-01  3.290e-02  18.489  < 2e-16 ***
## frac_coll_plus2010            3.718e-01  2.598e-02  14.313  < 2e-16 ***
## foreign_share2010             1.105e-01  4.960e-02   2.227 0.026033 *
## med_hhinc2016                 1.296e-07  2.543e-07   0.510 0.610193
## poor_share2010               -5.756e-01  4.036e-02 -14.262  < 2e-16 ***
## share_white2010               4.233e-02  2.082e-02   2.033 0.042122 *
## share_black2010               5.908e-02  2.084e-02   2.835 0.004615 **
## share_hisp2010               -5.112e-02  2.486e-02  -2.056 0.039840 *
## share_asian2010              -5.062e-01  9.173e-02  -5.519 3.71e-08 ***
## gsmn_math_g3_2013            -9.931e-04  2.142e-03  -0.464 0.642954
## rent_twobed2015              -7.016e-06  1.425e-05  -0.492 0.622587
```

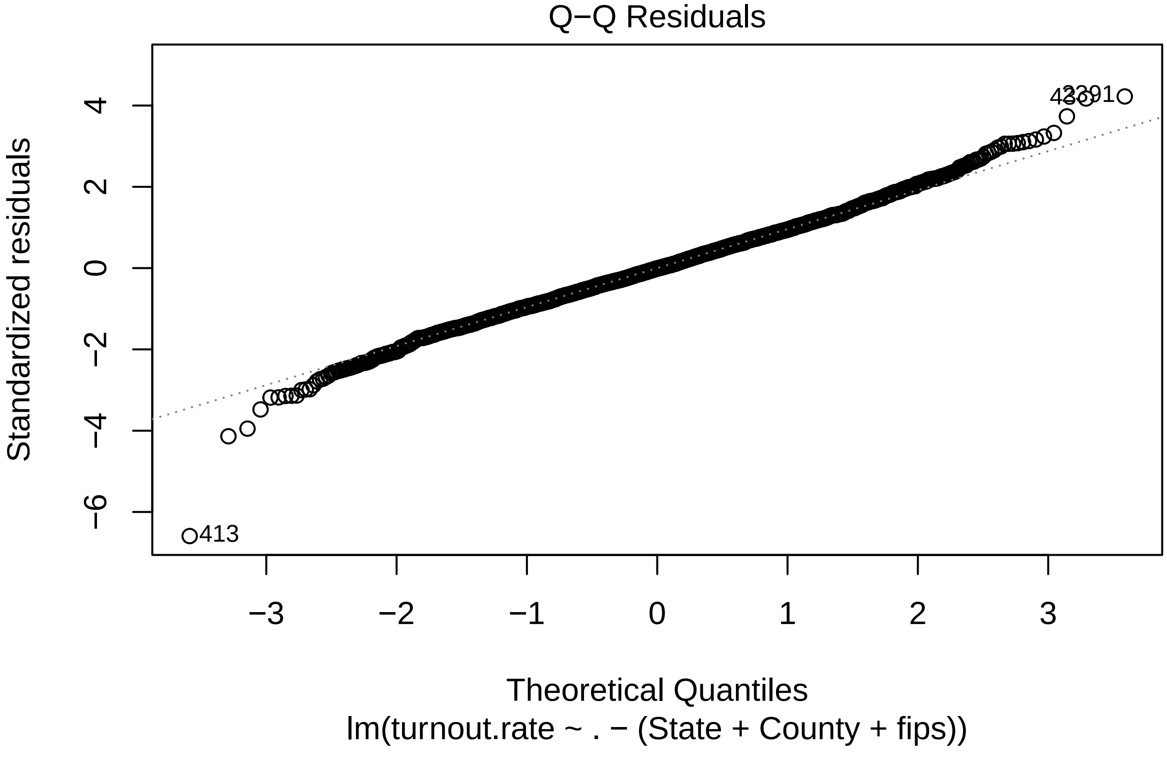
```
## singleparent_share2010       -6.174e-02  2.455e-02  -2.515 0.011971 *
## traveltime15_2010            -4.253e-02  1.170e-02  -3.634 0.000283 ***
## emp2000                       1.137e-01  2.729e-02   4.167 3.17e-05 ***
## popdensity2010               -2.256e-07  5.510e-06  -0.041 0.967336
## ann_avg_job_growth_2004_2013 -7.074e-01  1.068e-01  -6.624 4.12e-11 ***
## job_density_2013             -4.916e-06  1.134e-05  -0.434 0.664669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07304 on 2982 degrees of freedom
## Multiple R-squared:  0.4416, Adjusted R-squared:  0.4386
## F-statistic: 147.4 on 16 and 2982 DF,  p-value: < 2.2e-16
```

The linear regression model examining voter turnout demonstrates several significant relationships while controlling for State, County, and FIPS fixed effects. The model explains approximately 44% of the variance in turnout rates (Adjusted R-squared = 0.4386) and is highly significant (F = 147.4, $p < 2.2 \times 10^{-16}$). Education emerges as a strong positive predictor, with a one-unit increase in college education associated with a 0.371 increase in turnout ($p < 0.001$). Other significant positive predictors include foreign-born share ($\beta = 0.110$, $p < 0.05$), white population share ($\beta = 0.042$, $p < 0.05$), black population share ($\beta = 0.059$, $p < 0.01$), and employment ($\beta = 0.137$, $p < 0.001$). Conversely, several factors show significant negative associations with turnout: poverty rate exhibits a strong negative effect ($\beta = -0.576$, $p < 0.001$), as do Asian population share ($\beta = -0.506$, $p < 0.001$), single parent share ($\beta = -0.062$, $p < 0.05$), travel time ($\beta = -0.043$, $p < 0.001$), and job growth ($\beta = -0.707$, $p < 0.001$). Notably, several variables including median household income, Hispanic population share, math scores, two-bedroom rent, population density, and job density did not show significant relationships with turnout ($p > 0.05$). The residual standard error of 0.073 on 2982 degrees of freedom suggests relatively precise estimates, while the overall model significance ($p < 2.2 \times 10^{-16}$) indicates strong explanatory power in predicting voter turnout rates.

Diagnostics

```r
plot(lm_model, c(1, 2))
```





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.

Overall Recommendations and Next Steps

Based on these diagnostics, several actions are recommended. First, investigate points 413, 43, and 2391 for potential data issues or substantive influence. Second, implement robust standard errors to address the observed possible heteroscedasticity. Third, consider spatial correlation adjustments given the county-level nature of the data.

Regarding the model, we will test interaction terms to see how different factors affect each other. We will also observe how applying regularization impacts our regression.