

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	1
2	Data Description and Exploratory Data Analysis	1
2.1	Descriptive Statistics	2
2.2	Pre-Processing	4
2.3	Exploratory Graphs	4
3	Methods	6
4	Results	6
4.1	Baseline Model	6
4.2	Regularization	8
4.3	Correlation Structure	10
4.4	Interaction Terms	11
4.5	Final Model Including State Fixed Effects	12
5	Discussion and Conclusion	12
6	Bibliography	13
7	Appendix	13

1 Introduction and Motivation

2 Data Description and Exploratory Data Analysis

The dataset combines voter turnout data from the MIT Election Lab, population data from the U.S. Census Bureau, and socioeconomic and demographic predictors from Opportunity Insights. The turnout rate data is calculating by dividing the voter turnout for the 2020 presidential election in each county (MIT Election Data and Science Lab 2018) by the voting-eligible population (U.S. citizens age 18 and up) (US Census Bureau 2020). The resulting turnout rate should be a proportion between 0 and 1. The exception for the voter turnout data is Alaska, whose data is organized by election districts. Estimates for Alaska voter turnout data by county equivalent (borough and Census area) are derived from another source (cinyc 2021).

The predictors (county-level demographic and socioeconomic characteristics) are from Opportunity Insights, a Harvard-based research lab studying economic opportunity in the United States (Chetty et al. 2022). For predictors labeled with years, the data is for the labeled year(s) or the 5-year period ending in the labeled year. Basic descriptions of the predictors can be found in Table 1 and more detailed descriptions can be

found [here](#). Datasets for FIPS state and county codes are also used to merge the data sources (US Census Bureau 2023).

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 variables, we summarize the number of missing values, mean, median, standard deviation, interquartile range, minimum value, and maximum value (Table 1).

Variable	Description	Missing	Mean	Median	SD	IQR	Min	Max
frac_coll_plus2010	Proportion of people age 25 or older with a bachelor's degree or higher	0	0.19	0.17	0.09	0.09	0.04	0.71
foreign_share2010	Proportion of residents that are foreign-born	0	0.04	0.02	0.06	0.04	0	0.72
med_hhinc2016	Median household income	1	48,981	47,127	13,398	14,687	20,171	129,150
poor_share2010	Proportion of residents below the federal poverty line	0	0.16	0.15	0.06	0.08	0	0.53
share_white2010	Proportion of residents that are White non-Hispanic	0	0.78	0.86	0.2	0.27	0.03	0.99
share_black2010	Proportion of residents that are Black non-Hispanic	0	0.09	0.02	0.15	0.1	0	0.86
share_hisp2010	Proportion of residents that are Hispanic	0	0.08	0.03	0.13	0.07	0	0.96
share_asian2010	Proportion of residents that are Asian non-Hispanic	21	0.01	0	0.02	0.01	0	0.43
gsmn_math_g3_2013	Mean 3rd grade math test scores (grade level)	73	3.21	3.24	0.78	0.98	-0.66	6.58
rent_twobed2015	Median gross rent for two-bedroom housing units	76	692.34	642.51	205.04	195.93	236	2,085.23
singleparent_share2010	Proportion of households with children that have a single parent	0	0.31	0.3	0.09	0.1	0	0.81
traveltime15_2010	Proportion of workers age 16 or older who do not work at home that have a commute shorter than 15 minutes	0	0.4	0.38	0.14	0.19	0.1	0.99
emp2000	Proportion of residents age 16 or older that are employed	0	0.57	0.58	0.08	0.1	0.24	0.84
ln_wage_growth_hs_grad	Difference in log average hourly wage for high school graduates between 2010–2014 and 2005–2009	684	0.08	0.07	0.14	0.13	-0.72	0.91
popdensity2010	Residents per square mile	1	262.67	45.3	1,774.99	96.74	0.04	70,583.6
ann_avg_job_growth_2004_2013	Average annualized job growth rate	5	0	0	0.01	0.02	-0.08	0.12
job_density_2013	Jobs per square mile	2	124.24	18.47	862.85	43.3	0.02	36,663.2
turnout.rate	Proportion of voting-eligible residents who voted in the 2020 presidential election	0	0.66	0.66	0.11	0.14	0.19	1.58

Table 1: Descriptions and 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

Figure 1 shows the number of counties or county equivalents for each state in the dataset. The counts range from 1 in the District of Columbia to about 225 in Texas.

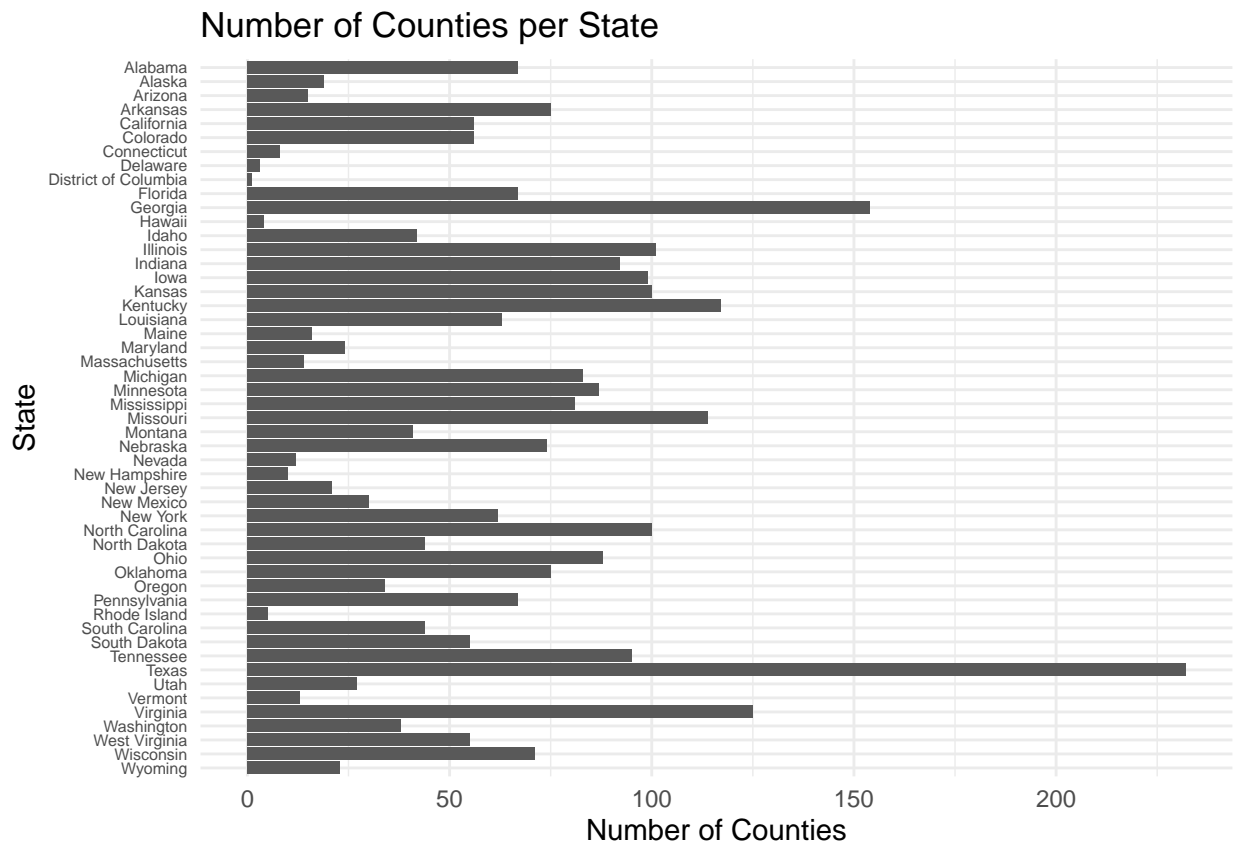


Figure 1: Bar chart showing the number of counties or county equivalents per state in the dataset.

The histogram of turnout rate (Figure 2) shows an approximately normal distribution. Most counties have turnout rates between 0.45 and 0.85, with few extreme values.

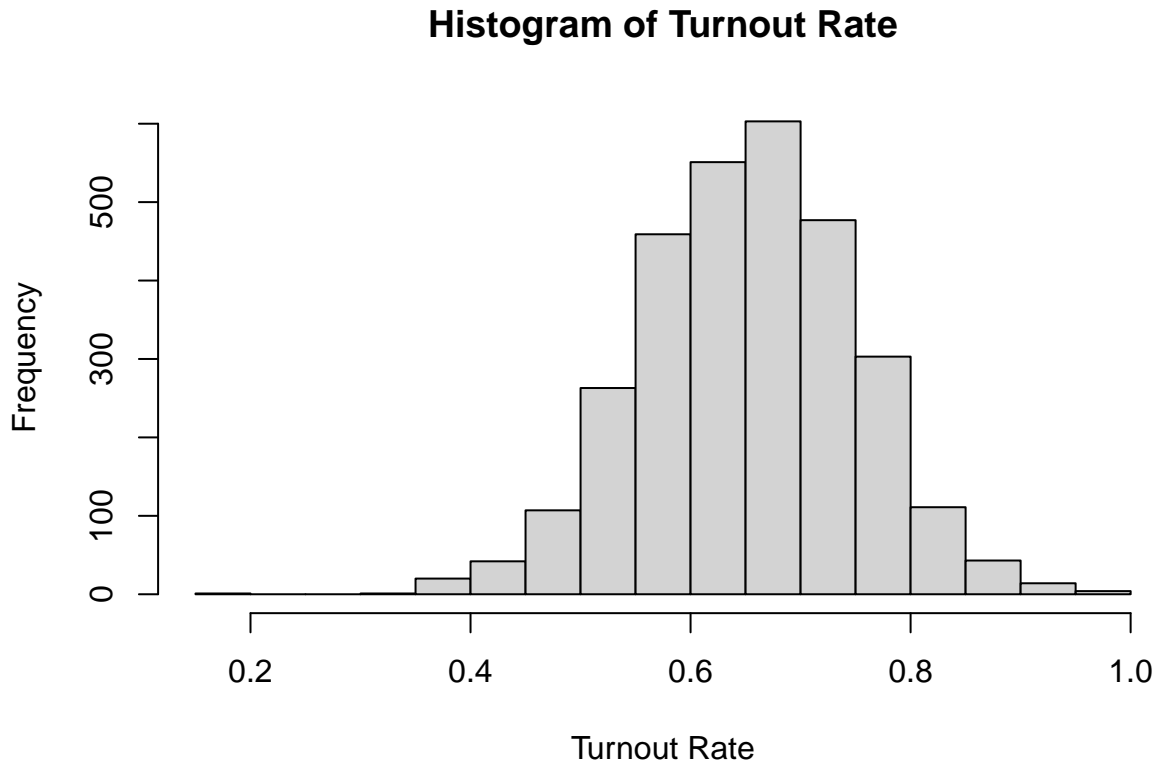


Figure 2: Histogram of turnout rate for counties shows an approximately normal distribution.

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 3). There is a moderate negative association between the variables ($r = -0.571$).

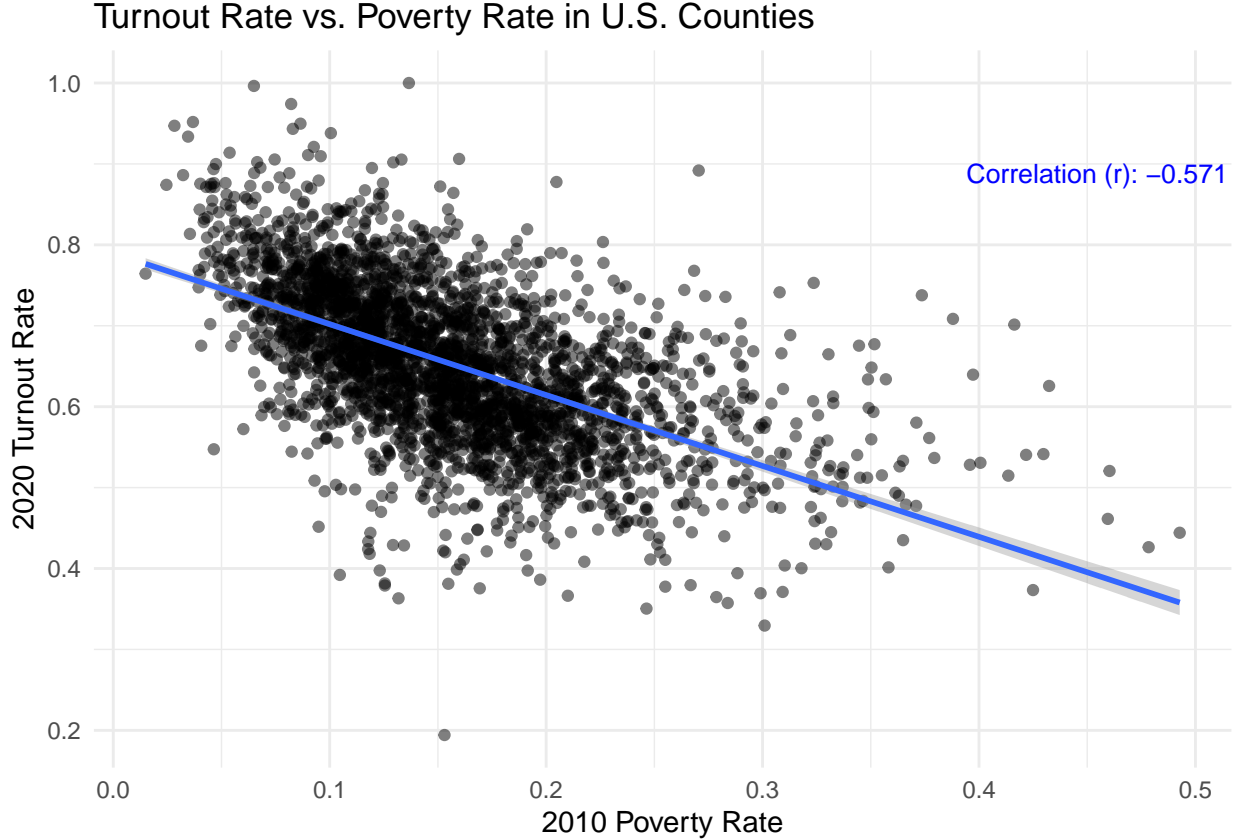


Figure 3: Scatterplot of poverty rate vs. turnout rate. The negative association between the two variables aligns with the theory that socioeconomic disadvantage is associated with lower electoral participation.

3 Methods

4 Results

4.1 Baseline Model

We first fit a simple linear regression model containing all predictors except state and no interaction terms. The model output is shown in Table 2.

The baseline model demonstrates several significant relationships. The model explains approximately 44% of the variance in turnout rates ($R^2 = 0.442$, adjusted $R^2 = 0.439$). 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, are not significantly related to turnout at the $\alpha = 0.05$ threshold. This initial exploration suggests that socioeconomic conditions significantly shape local electoral participation.

Variable	Estimate	SE	<i>t</i> -value	<i>p</i> -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.

4.1.1 Diagnostics

We conduct diagnostics for the baseline model to assess its suitability for linear regression. Below we assess the assumptions required for OLS linear regression:

- **Existence of Variance:** Residuals are reasonably dispersed, confirming the existence of variation in the dependent variable.
- **Linearity:** The residuals vs. fitted values plot (Figure 4a) does not show pronounced curvature, suggesting linearity is generally satisfied.
- **Independence:** Spatial correlation may exist between neighboring counties; external tests (like Moran’s I) should be considered for future analyses.
- **Homogeneity (Homoscedasticity):** Some fanning in the residuals vs. fitted values plot (Figure 4a) suggests heteroscedasticity. Robust standard errors or alternative modeling approaches may be warranted.
- **Normality:** The Q-Q plot (Figure 4b) shows mostly normal residuals, with minor deviations in the tails.

Overall, while the model is a decent fit, improvements—such as using robust errors or accounting for spatial autocorrelation—could refine our inference.

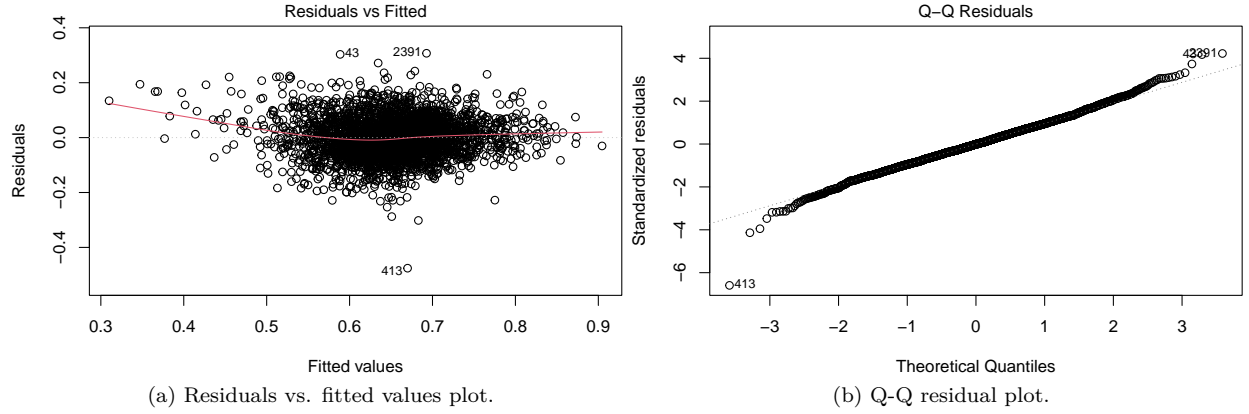


Figure 4: Diagnostic plots for baseline model.

4.2 Regularization

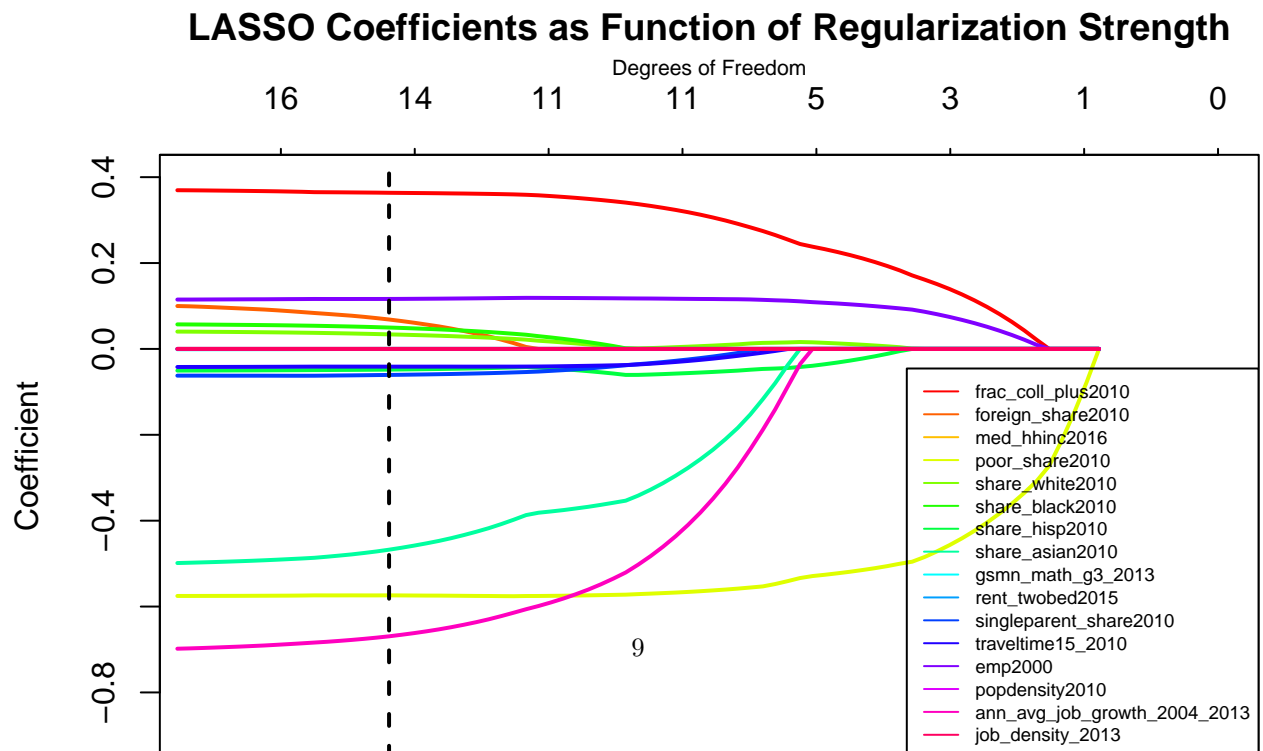
To address potential overfitting and identify the most influential variables, we employ LASSO regularization, which introduces a penalty that can shrink some coefficients to zero.

Elaborate on LASSO

The optimal $\lambda = 2.769\text{e-}4$ was selected through ten-fold cross-validation to achieve a penalty level that balances predictive accuracy and model complexity. This approach ensures we do not overfit by including too many small-effect predictors.

Variable	Coefficient
(Intercept)	0.612
frac_coll_plus2010	0.363
foreign_share2010	0.0702
med_hhinc2016	4.44e−08
poor_share2010	−0.574
share_white2010	0.0339
share_black2010	0.0499
share_hisp2010	−0.0486
share_asian2010	−0.47
gsmn_math_g3_2013	0
rent_twobed2015	0
singleparent_share2010	−0.0605
traveltime15_2010	−0.0412
emp2000	0.116
popdensity2010	−8.43e−07
ann_avg_job_growth_2004_2013	−0.67
job_density_2013	−3.17e−06

Table 3: Coefficients after LASSO regularization with $\lambda = 2.769\text{e-}4$.



As λ increases (moving right on the x -axis of Figure 5), the regularization strength increases and more predictors drop out of the model with coefficients of 0. The vertical dashed black line shows the chosen λ . Predictors with coefficients of 0 at the vertical line are dropped from the final model. This graph visually demonstrates which variables are “important enough” to survive the LASSO penalty.

Not much regularization for most variables

LASSO retained most predictors but excluded `gsmn_math_g3_2013` and `rent_twobed2015` (Table 3). This aligns with earlier findings that these variables were not statistically significant in the preliminary linear model.

4.3 Correlation Structure

Examining correlations among predictors helps identify potential multicollinearity and structure in the data.

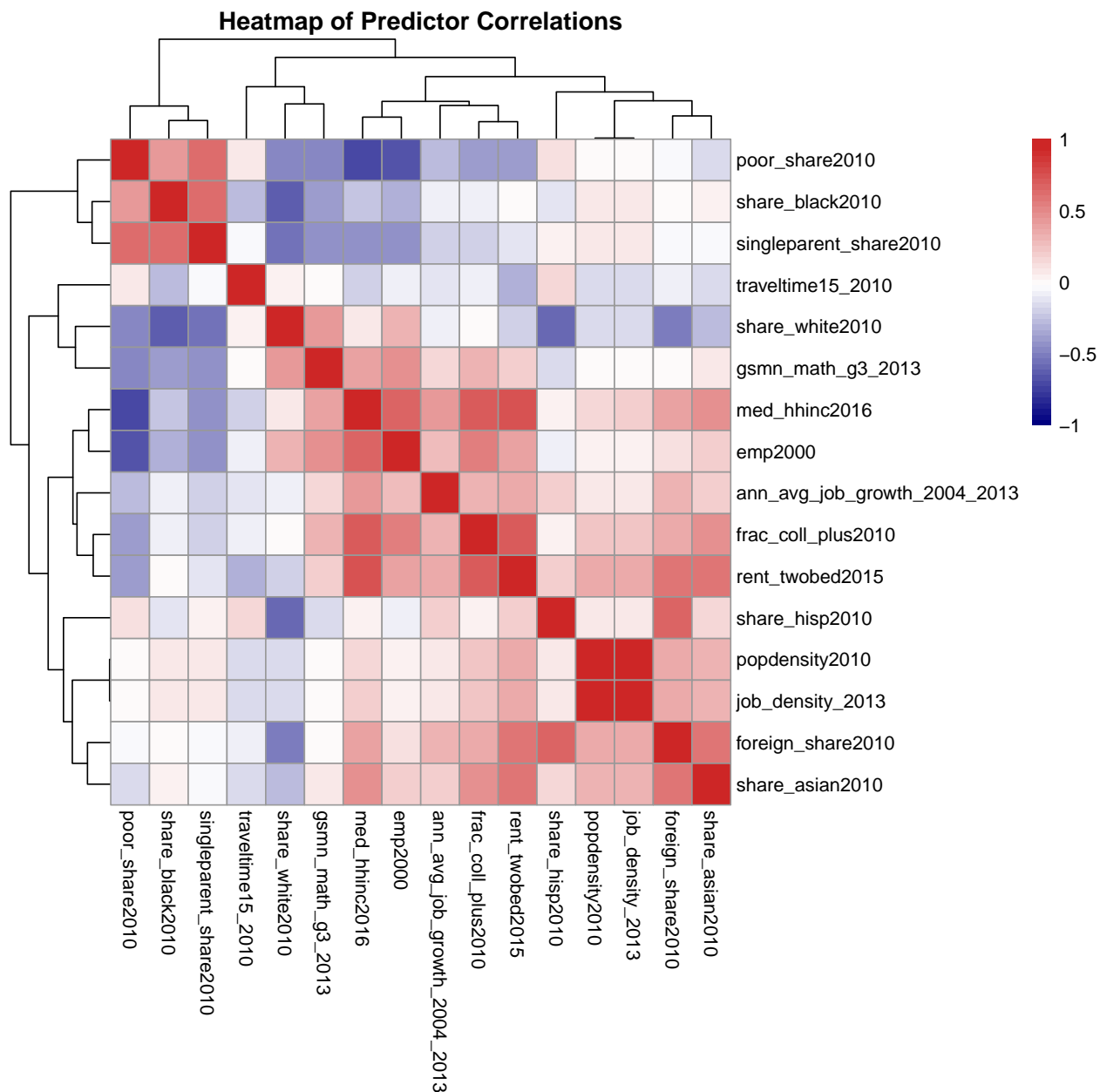


Figure 6: CAPTION!

The heatmap shows groups of variables that cluster together, indicating underlying socioeconomic dimensions (e.g., poverty and race-related measures clustered together, economic growth and density measures in another cluster). These clusters may reflect underlying latent factors that shape voter turnout.

One concern shown in the correlation matrix is the very high correlation between population density and job density

4.4 Interaction Terms

We test whether adding an interaction term between education and poverty (`frac_coll_plus2010:poor_share2010`) improves model performance.

A highly significant test result ($p < 2.2e-16$) suggests the interaction between college education fraction and

poverty share is crucial, indicating that the effect of education on turnout may depend on the poverty context of the county (and vice versa).

4.5 Final Model Including State Fixed Effects

We run a final model excluding the variables zeroed out by LASSO and adding state fixed effects to control for unobserved state-level heterogeneity.

Variable	Estimate	SE	<i>t</i> -value	<i>p</i> -value
(Intercept)	0.534	0.0298	17.9	3.6e−68
frac_coll_plus2010	0.24	0.0211	11.4	2.33e−29
foreign_share2010	−0.056	0.0446	−1.26	0.209
med_hhinc2016	9.68e−07	2.09e−07	4.64	3.7e−06
poor_share2010	−0.347	0.0349	−9.97	4.95e−23
share_white2010	0.0974	0.0198	4.92	9.21e−07
share_black2010	0.104	0.0212	4.91	9.73e−07
share_hisp2010	0.0031	0.0249	0.125	0.901
share_asian2010	−0.52	0.095	−5.48	4.72e−08
singleparent_share2010	−0.0894	0.0205	−4.36	1.32e−05
traveltime15_2010	−0.091	0.0116	−7.85	5.76e−15
emp2000	0.11	0.0252	4.38	1.22e−05
popdensity2010	5.03e−06	4.51e−06	1.11	0.265
ann_avg_job_growth_2004_2013	−0.698	0.0918	−7.6	3.88e−14
job_density_2013	−1.07e−05	9.24e−06	−1.16	0.247

Table 4: Output of final model except for state fixed effects.

With state fixed effects, the adjusted R^2 improves to approximately 0.646, suggesting that differences between states explain a significant portion of turnout variation. After controlling for state-level factors, education, poverty, and various demographic characteristics remain significant. Notably, poverty and time-to-work remain negatively associated with turnout, while educational attainment consistently shows a positive association. This final specification suggests that while local socioeconomic conditions are important, broader state-level contexts also shape the electoral participation landscape.

5 Discussion and Conclusion

Our analysis shows that socioeconomic and demographic factors strongly influence county-level voter turnout. Education and certain demographic features (e.g., Black population share) are robust, positive predictors of turnout, while higher poverty rates, longer travel times, and certain population characteristics (e.g., Asian population share) are negatively associated. LASSO regularization supports the exclusion of non-influential predictors, refining the model and reinforcing the significance of key variables. Incorporating interaction

terms and state fixed effects further refines our understanding, revealing that the influence of education on turnout may be contingent on the economic context and that state-level factors account for substantial variation across the U.S. counties.

There are some limitations to our findings. One important caveat is that these associations between county-wide factors and voter turnout do not imply anything about the individuals within counties. Concluding that, for instance, highly educated people are more likely to vote solely based on the data presented here would be an example of the ecological fallacy, where inferences about individuals are made based on inferences about groups containing those individuals (Freedman 1999). Another limitation is that there is a somewhat arbitrary time lag between the measurement of the predictors and the outcome (the socioeconomic and demographic factors are measured between 2000 and 2016, but the voter turnout rate is measured in 2020). Regardless, these findings have implications for policymakers and organizations interested in increasing voter participation. Interventions that improve socioeconomic conditions, reduce poverty, enhance education, and consider unique state-level political climates could foster higher electoral engagement, but more research should be done at the individual level or experimental level to reach more statistically sound conclusions.

6 Bibliography

- Chetty, Raj, John Friedman, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. 2022. “Replication Data for: The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility.” Harvard Dataverse. <https://doi.org/10.7910/DVN/NKCQM1>.
- cincy. 2021. “Alaska Presidential Results by County Equivalent, 1960-2020.” *RRH Elections*. <https://rrhelections.com/index.php/2021/04/13/alaska-presidential-results-by-county-equivalent-1960-2020/9/>.
- Freedman, David. 1999. “Ecological Inference and the Ecological Fallacy.” *International Encyclopedia of the Social & Behavioral Sciences* 6 (4027-4030): 1–7.
- MIT Election Data and Science Lab. 2018. “County Presidential Election Returns 2000-2020.” Harvard Dataverse. <https://doi.org/10.7910/DVN/VOQCHQ>.
- US Census Bureau. 2020. “B05003: Sex by Age by Nativity and Citizenship Status.” [https://data.census.gov/table/ACSDT5Y2020.B05003?t=Citizenship&g=010XX00US\\$0500000&y=2020&d=ACS%205-Year%20Estimates%20Detailed%20Tables&moe=false&tp=true](https://data.census.gov/table/ACSDT5Y2020.B05003?t=Citizenship&g=010XX00US$0500000&y=2020&d=ACS%205-Year%20Estimates%20Detailed%20Tables&moe=false&tp=true).
- . 2023. “American National Standards Institute (ANSI), Federal Information Processing Series (FIPS), and Other Standardized Geographic Codes.” *Census.gov*. <https://www.census.gov/library/reference/code-lists/ansi.html>.

7 Appendix

State	Estimate	SE	<i>t</i> -value	<i>p</i> -value	State	Estimate	SE	<i>t</i> -value	<i>p</i> -value
Alaska	0.0161	0.0184	0.873	0.383	Montana	0.0762	0.0124	6.13	9.75e−10
Arizona	0.0521	0.0175	2.98	0.00286	Nebraska	0.025	0.0108	2.32	0.0205
Arkansas	−0.0994	0.00992	−10	2.74e−23	Nevada	0.0537	0.0187	2.87	0.00418
California	0.0492	0.0121	4.06	5e−05	New Hampshire	0.00639	0.0201	0.318	0.751
Colorado	0.0888	0.0115	7.74	1.36e−14	New Jersey	0.0268	0.0156	1.71	0.0872
Connecticut	−0.0424	0.0223	−1.91	0.0568	New Mexico	0.0212	0.015	1.41	0.157
Delaware	0.00291	0.0343	0.0848	0.932	New York	−0.0549	0.0111	−4.94	8.3e−07
District of Columbia	−0.0488	0.0596	−0.819	0.413	North Carolina	0.0447	0.00925	4.83	1.44e−06
Florida	0.0482	0.0105	4.59	4.61e−06	North Dakota	−0.00176	0.0125	−0.141	0.888
Georgia	−0.0194	0.00855	−2.27	0.0232	Ohio	−0.0265	0.00995	−2.67	0.00767
Hawaii	0.109	0.0399	2.73	0.00644	Oklahoma	−0.0836	0.0105	−7.93	2.99e−15
Idaho	0.043	0.0121	3.57	0.000368	Oregon	0.118	0.0129	9.19	7.15e−20
Illinois	−0.0258	0.0097	−2.66	0.00789	Pennsylvania	−0.0151	0.0105	−1.43	0.152
Indiana	−0.0722	0.00995	−7.26	4.97e−13	Rhode Island	−0.0554	0.0274	−2.02	0.0432
Iowa	0.0377	0.0101	3.74	0.000189	South Carolina	0.000997	0.0113	0.0881	0.93
Kansas	−0.0158	0.0101	−1.56	0.119	South Dakota	0.00321	0.0118	0.273	0.785
Kentucky	−0.0159	0.00944	−1.68	0.0928	Tennessee	−0.0691	0.00959	−7.21	7.02e−13
Louisiana	0.00109	0.0103	0.106	0.916	Texas	−0.0202	0.0093	−2.18	0.0297
Maine	0.0753	0.0166	4.54	5.89e−06	Utah	0.0405	0.014	2.89	0.00391
Maryland	−0.0418	0.0143	−2.92	0.00355	Vermont	0.0311	0.0181	1.72	0.086
Massachusetts	0.0211	0.0178	1.18	0.238	Virginia	0.0065	0.009	0.723	0.47
Michigan	0.0559	0.0101	5.56	2.93e−08	Washington	0.111	0.0125	8.92	7.76e−19
Minnesota	0.0689	0.0103	6.68	2.81e−11	West Virginia	−0.085	0.0111	−7.67	2.35e−14
Mississippi	0.00126	0.00972	0.13	0.897	Wisconsin	0.0513	0.0106	4.84	1.36e−06
Missouri	−0.0311	0.00943	−3.29	0.000997	Wyoming	−0.00924	0.015	−0.617	0.537

Table 5: State fixed effects for final model.