# What Drives Voter Registration in One of America's Most Diverse States?

## A case study on county level election data in California

## Contents

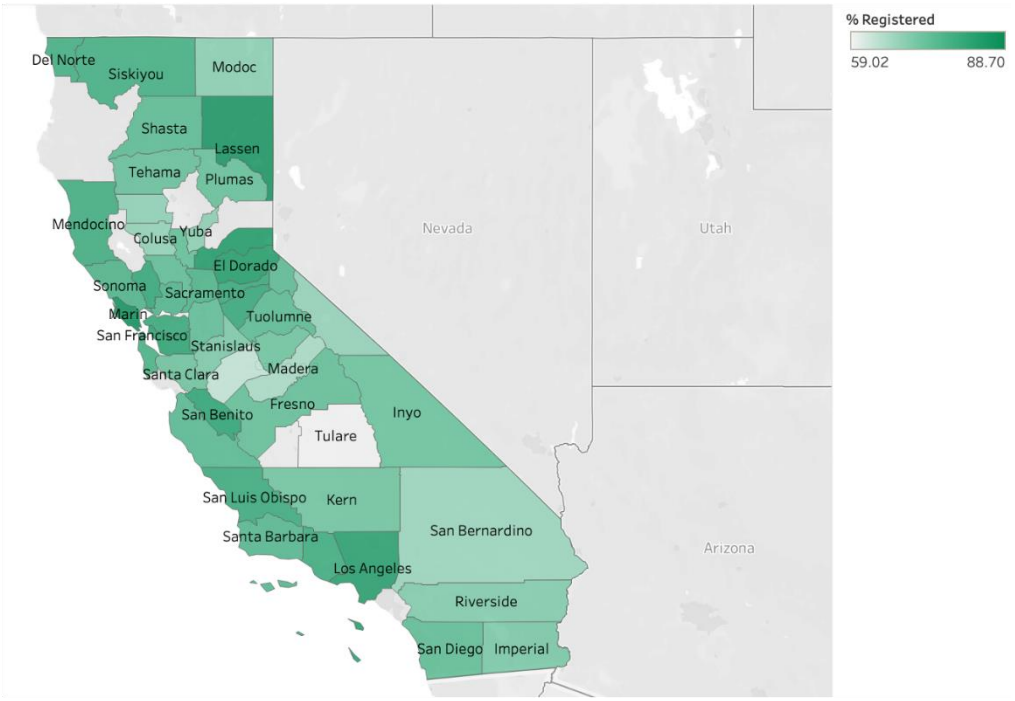
## Introduction and Literature Review

The recent US presidential election and the impact of Covid-19 have shone a new light on US politics and the state of the nation, socially, economically and ethically. The election had record voter turnout, both in terms of percentages and absolute numbers, and highlighted how divided the country is today. As a result, the question of what influences a voter to cast a vote comes to the forefront. Voter turnout, not just in the United States, but around the world, is one of the biggest questions in political economics. Oftentimes, in the world of developed democratic countries, citizens tend to not exercise their biggest right as members of democratic societies - the right to vote. This is true especially in the United States, which historically has lagged behind most other developed countries in terms of voter turnout. This is usually not seen as a favorable outcome - for instance, a low voter turnout, skewed in favor of a certain section of society, tends to bring about increasingly partisan policy in government.

It is for this reason that anomalous elections, for example, the one we just had, and the presidential election of 2008 are subject to increased scrutiny by the research community, in an attempt to explain voters' willingness, or unwillingness, to vote. One obvious category to use in exploring this question is income. Several studies (Verba and Nie 1972; Wolfinger and Rosenstone 1980; Leighley and Nagler 1992; Verba, Schlozman, and Brady 1995) have shown that people from low-income backgrounds are less likely to vote, but also more likely to vote for the Democrat party. The extent to how these predicted outcomes are likely to affect actual voter turnout is uncertain however, that is to say, it is ambiguous how much of an effect increased voter turnout amongst low income households would change an election outcome.

Race seems to be a factor in voter turnout in the United States as well. As Fowler (year) notes, people that vote tend to be "significantly wealthier, more likely to be white, and more likely to attend church" than nonvoters. Gender and education levels have also been attributed as factors that could potentially affect voter turnout, along with other factors such as marital status and age. Moreover, there seems to be a definite difference between the preferences and driving factors behind a regular voter as compared to irregular/non-voters. As mentioned above however, most of these studies tend to be correlational. Causation is much harder to tackle, as is the case with many Economics problems. This is especially exacerbated by the fact that when conducting experiments and surveys, respondents tend to not follow what they had answered in the survey when actually voting.

Using this information, we take a closer look at the voter turnout in our state of California, exploring based on regions and one level deeper into county level voter registration data. Looking at data at the county level also helps circumvent some of the problems associated with surveys mentioned above.



(Figure 1, % voter registration in California for each county)

Apart from choosing California due to its relevance to us, it also happens to be an ideal state for such a study. While it is known to be a democrat-dominated state on the whole, like many other states, a large percentage of this population is concentrated in urban areas, such as San Francisco and Los Angeles. A large number of counties in the countryside tend to support the Republican/other parties. Moreover, California also enjoys diversity on many fronts. Racially speaking, California is home to all of the biggest groups and has been getting continually more diverse - 39% are Hispanic, 37% are White, 15% are Asian American and 6% are African American. This parity can be observed even when looking at age and gender demographics. (Census, 2019)

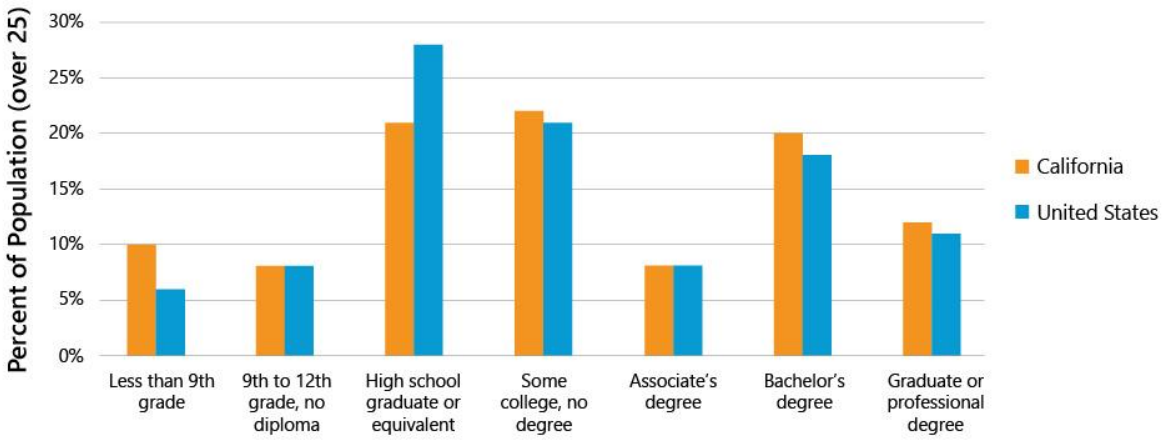
By utilizing county and region level voter registration data, we aim to better understand voter-turnout in the state using recent voter registration data and gain insight into what factors are involved in predicting voter patterns, and voter registration numbers. By understanding and focusing on factors that are highly correlated with voter turnout, we can try and better gauge turnout in California for future elections. We also hope that these are results that could be extrapolated to other states. Perhaps not at the level of an entire state, but perhaps at the county level - in terms of demographics, for example, one could find a county on the East Coast that is similar to San Francisco and extrapolate the results we provide with reasonable confidence. In this study, we analyze the significance of some of the factors touted above in a more restricted environment, and check if some of the claims made hold water at a state and county level.

## Model Specification

In order to understand the landscape of voters in California who played a role in the 2020 election we are assuming that 2019 data on voter registration reflects voter turnout for this year's

election. In our model, the y variable is % of people registered to vote per county in the state of California. In a regression model, all the influences on the dependent variable that are not captured by a model are collected in the error term which we assume is uncorrelated with the regressor. In order to stay true to this assumption as much as possible, we have a few x variables that will help minimize estimation bias and the error term. We also conduct a standard Breusch Pagan test to test this assumption. Our x variables per county include median age, median income, race of voters as well as party affiliations. We include these factors based on the literature review conducted earlier - these are factors that one would expect to be correlated with voter turnout, and are also backed by research in the field - and also based on our own initial analysis of the data, which is detailed in the Results section.

After conducting this initial exploratory analysis, we build our main Multi-Linear Regression model, and hence, adopt all the assumptions of the model. For this model, our initial hypothesis for the coefficients was as follows: we expect income and the percentage of white-identifying individuals in a county to be positively correlated with the voter turnout in a county. We also expected median age to be positively correlated with voter turnout, but we expected this to be a slightly weaker relationship compared to the other variables. Party preference was another ambiguous variable - it encodes information/is associated with other factors like education levels, but then the question becomes whether or not education levels tend to increase or decrease voter registration in California. There is literature that supports both sides of the answer, but for our hypothesis, we believe that the correlation is positive, purely from the fact that we were analyzing the state of California, which tends to have moderate to high education levels in several counties.




(Figure 2: Educational Attainment in California)

We expect race and income to both have a strong relation with voter turnout. Those from minority races with lower incomes would be more likely to vote for every election as they have more at stake. We expect democratic voters to mainly be based in the larger counties that cover cities where most working people reside, as well as government and educational institution areas.

Despite including these control variables, we also expect to have omitted variables, some of which are discussed later in the paper. We believe the relationship is linear in parameters, and while we expect relationships between the regressors in the model (like age and income, income tends to be positively associated with age), we expect these relationships to not cause the problem of perfect multicollinearity. We explore this assumption further in the Results section, before getting into the main model. After including the other control variables, we expect the overall coefficient of determination value to be quite high (i.e., a significant portion of the variance in our dependent variable should be explained by our model). We will also conduct tests of significance, including t-tests and f-tests later in the paper to validate these expectations.

## Data

Our data comes from the United States Census Bureau and the CA government website. First, we took data for voter turnout at a county-based level for all 58 counties in the state for the year 2019 (we chose the most recently available data, which was October 1st, 2019). The dataset also included information on party preferences, which we encoded into three major categories, 'Democrat,' 'Republican,' and 'Other' (which includes all other parties and Independents). We merged this information with data from the census on population, median income, median age, and race percentages. The census data was available only for 42 counties. To account for this lack of data, we first split our data into 10 sub-regions based on the California census system and filled in the missing data for a given county with the sub-region average. We did this to ensure that we filled the missing values for a given county with a value that represented similar counties (based on location).

To reiterate, the dependent variable we test is the percentage of the population registered to vote in a county (number between 0 and 100). The independent variables we hypothesize to be significant are median income (in absolute $ terms), median age (in absolute terms), party preference (i.e., percentage of voters registered democrat/republican/other) and a variable on race (i.e., percentage of population that is white or black). We also include entity fixed effects for the sub-region to discount any inter-region effects on our regression, based on our initial exploratory analysis of the trends in the data.

## Results

To start off our analysis and understanding of the various regressors we chose to investigate, we ran a simple Ordinary Least Squares (OLS) regression for each independent variable against our dependent variable – percentage of registered voters in CA. The following analyses were used to deduce significance of the variables of our data and educate our final model. Note that however, this set of regressions is potentially subject to Omitted Variable Bias, especially because we do not add region-based fixed effects to these regions. We do this for our final MLR model.

### Race

One of the key factors we had in our initial hypothesis was race. The following table details the results of the regressing registered voter percentage on firstly, percentage of population that is

white, and secondly, percentage of population that is white. We do not regress on both of these variables to avoid multicollinearity, which would be a violation of our regression assumptions.

**Results**

| | *Dependent variable:* | |
|---|---|---|
| | % of registered voters in CA | |
| | (1) | (2) |
| White | 0.069 | |
| | (0.068) | |
| Black | | -0.067 |
| | | (0.317) |
| Constant | 71.619*** | 76.700*** |
| | (4.844) | (1.397) |
| Observations | 58 | 58 |
| $R^2$ | 0.018 | 0.001 |
| Adjusted $R^2$ | 0.001 | -0.017 |
| F Statistic (df = 1; 56) | 1.030 | 0.045 |
| *Note:* | * Significant at the 10% level | |
| | ** Significant at the 5% level | |
| | *** Significant at the 1% level | |
| | + Significant at the 10% level | |

As seen in the table, our $R^2$ values are quite small, which might hint at the fact that these variables are not viable predictors of voter registration. Moreover, a two-sided t-test of both of these predictors (equivalent to an F test in the case of one regressor) shows that the coefficients are not significant. We explore this further in our Multi-Linear Regression model later. Of these two variables, the percentage of population that is White is the more significant predictor, because it has a higher t-statistic value. In standard form, the above regressions are represented as follows, where *B* and *W* denote Black and White population in percentage terms, respectively:

$$\hat{y} = 0.069 * W + 71.619$$
$$(0.068) \qquad (4.844)$$

$$\hat{y} = -0.067 * B + 76.7$$
$$(0.317) \qquad (1.397)$$

It is interesting to note that the coefficient of *B* is negative: this seems to imply that as the percentage of Black individuals in the population goes up, the percentage of registered voters is expected to drop. While this is in line with what we saw earlier about the voting patterns of socio-economic minorities, it might also be an affect of the OVB mentioned earlier. We explore this further in our final model.

### Party Preference

The second set of variables we consider are party preferences. We divide preferences into three categories, Democrat, Republican and Other (which includes smaller parties, independent candidates, etc.). The results are as follows:

**Results**

| | *Dependent variable:* | | |
|---|---|---|---|
| | % of registered voters in CA | | |
| | (1) | (2) | (3) |
| Democratic | $0.145^{+}$ | | |
| | (0.087) | | |
| Republican | | $-0.143^{+}$ | |
| | | (0.078) | |
| Other | | | 0.344 |
| | | | (0.228) |
| Constant | $70.823^{***}$ | $81.180^{***}$ | $74.326^{***}$ |
| | (3.464) | (2.699) | (1.636) |
| Observations | 58 | 58 | 58 |
| $R^2$ | 0.048 | 0.057 | 0.039 |
| Adjusted $R^2$ | 0.031 | 0.040 | 0.022 |
| F Statistic (df = 1; 56) | $2.805^{+}$ | $3.359^{+}$ | 2.276 |
| *Note:* | * Significant at the 10% level | | |
| | ** Significant at the 5% level | | |

*** Significant at the 1% level
+ Significant at the 10% level

(Table 2: Regression results for party preference variables)

Our $R^2$ values for category are much better than the previous category, which again points to what we will potentially see in the MLR case. Note that for the Republican category, the slope coefficient does not make perfect sense intuitively: the sign is negative, i.e., according to this regression, as the percentage of Republican leaning population increases, the number of registered voters decrease. This could be a sign of the OVB mentioned earlier – there might be some other variable correlated to the Republican variable in the error term. However, based on the t-statistic values, it is the most significant of the three. The regression equations for the Democrat and Republican variables (*D* and *R*) are as follows:

$$\hat{y} = 0.145 * D + 70.823$$
$$(0.087) \quad (3.464)$$

$$\hat{y} = -0.143 * R + 81.180$$
$$(0.078) \quad (2.699)$$

## Age and Income

The last set of variables we look at are age and income. We look at median age and median income. The results are as follows:

**Results**

| | *Dependent variable:* | |
|---|---|---|
| | % of registered voters in CA | |
| | (1) | (2) |
| Median Age | 0.410*** | |
| | (0.118) | |
| Median Income | | 0.0001+ |
| | | (0.00004) |
| Constant | 59.955*** | 73.130*** |
| | (4.810) | (1.958) |
| Observations | 58 | 58 |
| $R^2$ | 0.177 | 0.059 |
| Adjusted $R^2$ | 0.163 | 0.042 |
| F Statistic (df = 1; 56) | 12.080*** | 3.496+ |

*Note:* * Significant at the 10% level
** Significant at the 5% level
*** Significant at the 1% level
+ Significant at the 10% level

(Table 3: Regression results for age and income variables)

Median age seems to be a significant predictor, and median income, while not as strong as age, still carries some predictive power. The $R^2$ also denote the significance of these variables, along with the t-statistic values for both. In fact, age seems to be our most significant predictor so far! The equations are as follows:

$$\hat{y} = 0.410 * A + 59.955$$
$$(0.118) \qquad (4.810)$$

$$\hat{y} = 0.0001 * I + 73.130$$
$$(0.00004) \qquad (1.958)$$

## Multi-Linear Regression Model

Using our findings from the previous OLS analyses, we choose the following independent variables for our MLR model – median age, percentage of Black population, median income, percentage of registered Republicans. Our approach was to choose the most significant variable from our initial OLS regressions, with the exception of percentage of Black population (*B*) – *W* was the more significant predictor, but we decided to go ahead with *B* because the negative sign for the coefficient was of interest, based on earlier literature review as well.[1] Our dependent variable, as stated earlier, is the percentage of people registered to vote. We also include unit Fixed Effects to tackle any OVB due to geographic differences. We do this by including a dummy variable for each region.

The table of results can be found in the appendix (Table 4), as it is too large to be included in the main document. The regression equation is as follows (we omit the 9 dummy variables used to encode the Fixed Effects). The omitted group for our fixed effects was the Central Coast region:

$$\hat{y} = 0.299 * B - 0.118 * R + 0.541 * A + 0.001 * I + 59.417$$

(The standard errors can be found in the Appendix, they were omitted due to formatting issues)

Since our model is a linear-linear model, we can expect a 1-unit change in our independent variables to bring out a change in the dependent variable proportional to the coefficient of the particular independent variable. For instance, if we consider the median age variable, we expect an increase in median age of 1 to cause an increase in percentage of registered voters by 0.541 * 1%. This similar analysis can be done for all other independent variables as well. The dummy

[1]

variables can be interpreted as showing the average difference in voter registration percentage for a given region as compared to the Central Coast region.

To validate our findings, we conduct a Breusch-Pagan test to test for homoskedasticity of errors, which is one of the MLR model assumptions. The null hypothesis of the BP test is that the variances of the residuals are constant when measured against the independent variables (i.e., homoscedastic errors), and the alternative hypothesis is that we have heteroskedasticity in our residuals. We regress residuals$^2$ against our original combination of independent variables and run an overall regression F-test – the F stat value we get is 1.66, which is lesser than the critical value of 2.34 – this means that we fail to reject the null hypothesis of the BP test, which means we cannot say that our model violates the MLR assumption! We also analyzed the residual plots of our model for each of the main independent variables, which can be found in the Appendix – we found no patter in these plots, i.e., the MLR assumption of linearity in parameters holds.

The $R^2$ of the MLR model was 0.516, which means that our model explains more than 50% of the variance in voter registration percentage in California (for a given county). Note that however, the adjusted $R^2$ is 0.373, which implies that some of our independent variables are unnecessary – this may be due to the dummy variables we included in the estimation procedure. We also ran an overall F-test to measure the significance of all our independent variables (including the dummy variables); i.e., we tested for the null hypothesis that all the coefficients of our independent variables are equal to 0. The alternative hypothesis is that at least one of the coefficients is non-zero, which means that our model offers has at least more predictive power than a baseline constant predictor.

We measured our F-statistic to be 3.60423, which when measured against the critical value at a 1% confidence level (3.43, using 13 and 44 degrees of freedom) implies that we reject the above null hypothesis.

## Conclusion and Summary

This project sets out to answer an important question on the minds of many citizens who live in a democracy today – what prompts one to vote, or in our particular case, register to vote?
Based on our initial literature review and overarching political theory, we hypothesized that voter registration percentage in California would be well explained by factors such as race, income, age and party preferences. Other factors that seemed to be important were education, religion, and gender, however, we decided to not use these factors due to a lack of available data. Before jumping into the Multi-Linear Regression model, we first tested each of our variables by running a series of simple Ordinary Least Squares regressions to gain some initial validation about our hypothesis. While it held true for the most part, some details surprised us, for example - the relative insignificance of race as a predictor of voter turnout. We expected this to have a larger effect, but the results we got showed otherwise, both in the initial OLS and the final MLR model - this could potentially be because our data was limited in terms of the time period it covered (we only looked at the year 2019). Apart from this however, we found that our final predicted model performed quite well in terms of predictive power, as seen in our analysis of the $R^2$ values above, which means that our initial hypothesis was indeed correct. (We had an $R^2$ value of around 0.5 for the MLR model)

In terms of what our paper provides to existing literature, the county and region-based approach we followed allowed us to control for geographical differences and instead analyze voters at the county/region level (as opposed to a country wide analysis). Our dummy variables highlight the difference between the Central Coast region in California (which contains tourism heavy cities like Santa Barbara and San Luis Obispo) and other regions in California, for example, heavily metropolitan regions like the San Francisco Bay area. The coefficient of the dummy variable for SF Bay Area is negative, while the regression did not classify this coefficient as being statistically significant, the negative value is interesting because it shows how voter registration is higher (albeit marginally) in a heavily urbanized area like San Francisco, as compared to coastal cities like San Luis Obispo. While this may not be especially useful for all regions, it is interesting to note how geographic regions differ in terms of their registration patterns. It would be interesting to note how this analysis could be extended to other states in the US, and how the patterns change there.

As mentioned earlier, our regression seems quite solid in fitting the data, however we expected more variables to be statistically significant. Given the lack thereof, it is important to recognize the presence of OVB and include more variables in the future such as gender or religion as mentioned above. The point of MLR is to include more than one variable that influences our dependent variable. Influences that are not captured by the model are collected in the error term, which we assume are uncorrelated with the regressor as an assumption of our linear regression models. However, this assumption can be violated if we do not do a good job of including variables that would influence the dependent variable and that are independent of the other regressors to avoid multicollinearity. When this assumption is violated the coefficient-estimations do not reflect the true coefficients (i.e., they will be biased) and hence our overall predictions will be incorrect due to the Omitted Variable Bias. (OVB) Some of these variables could be religion, education, and gender – it may indeed be the case that a variable like education is, as mentioned in the introduction, correlated with party preference.

In addition to including more variables for a better model, another area of expansion for this project is in comparing our findings to the 2016 election. 2016 was a polarizing election year, one that shined light on the deepening partisan divide in America, highlighted the impact of social media and brought many emotionally charged topics to the table. Comparing the influence of our independent variables on the percentage of registered voters would be an interesting direction for this work - especially the age variable as it seemed to be the most significant variable influencing the 2020 election registration. It is nice to see that the model does indeed represent the changes we saw in political participation this year.

## Sources


Leighley, J. E., & Nagler, J. (1992). Socioeconomic Class Bias in Turnout, 1964-1988: The Voters Remain the Same. *The American Political Science Review*. https://www-jstor-org.libproxy.berkeley.edu/stable/pdf/1964134.pdf

*American Community Survey*. (2019a). [Government]. US Census Bureau. https://data.census.gov/cedsci/table?q=race&g=0400000US06.050000&y=2019&tid=ACSDT1Y2019.B02001&hidePreview=true

*American Community Survey*. (2019b). [Government]. US Census Bureau. https://data.census.gov/cedsci/table?q=income&g=0400000US06.050000&tid=ACSST1Y2019.S1901&hidePreview=false

Brady, H. E., Verba, S., & Schlozman, K. L. (1995). Beyond ses: A resource model of political participation. *American Political Science Review*, *89*(2), 271–294. https://doi.org/10.2307/2082425

*Demographics | ca @50 million*. (n.d.). https://ca50million.ca.gov/demographics/

Fowler, A. G. (2013). *Citable link Five Studies on the Causes and Consequences of Voter Turnout* [Doctoral dissertation]. Harvard University.

*Report of registration: California secretary of state*. (2019). [Government ]. Election and Voter Information; California Secretary of State. https://www.sos.ca.gov/elections/report-registration

*U. S. Census bureau quickfacts: California*. (n.d.). https://www.census.gov/quickfacts/CA

VERBA, S. (n.d.). *Participation in america*. https://press.uchicago.edu/ucp/books/book/chicago/P/bo3637096.html

*Voter Turnout*. (n.d.). Massachusetts Institute of Technology. https://electionlab.mit.edu/research/voter-turnout

## Appendix

### MLR Results

**Results**

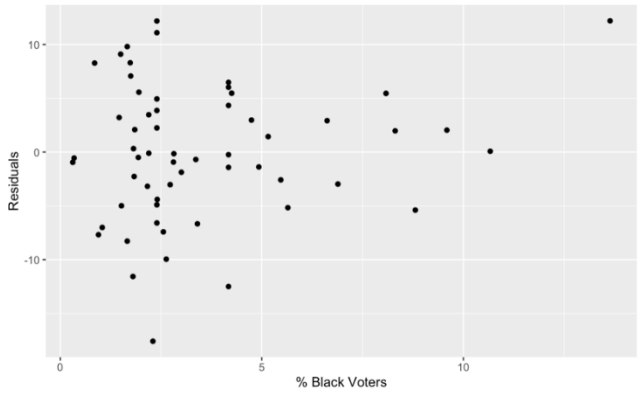
| | *Dependent variable:* |
|---|---|
| | X..Registered |
| X..Black | 0.299 |
| | (0.347) |
| X..Republican | -0.118 |
| | (0.123) |
| Median_age_total | 0.541*** |
| | (0.133) |
| Median_income | 0.0001 |
| | (0.0001) |
| factor(Region)Inland Empire | -9.810* |
| | (4.582) |
| factor(Region)Los Angeles County | 1.502 |
| | (5.839) |
| factor(Region)North Coast | -5.580+ |
| | (2.957) |
| factor(Region)Northern San Joaquin Valley | -7.369* |
| | (3.158) |
| factor(Region)Orange County | -3.865 |
| | (5.648) |
| factor(Region)San Diego - Imperial | -4.954 |
| | (4.211) |
| factor(Region)San Francisco Bay Area | -5.358 |
| | (3.813) |
| factor(Region)Southern San Joaquin Valley | -7.580* |
| | (3.620) |
| factor(Region)Superior California | -2.296 |
| | (2.828) |
| Constant | 59.417*** |
| | (8.051) |

| | |
|---|---|
| Observations | 58 |
| $R^2$ | 0.516 |
| Adjusted $R^2$ | 0.373 |
| F Statistic | $3.604^{***}$ (df = 13; 44) |
| *Note:* | * Significant at the 10% level |
| | ** Significant at the 5% level |
| | *** Significant at the 1% level |
| | + Significant at the 10% level |

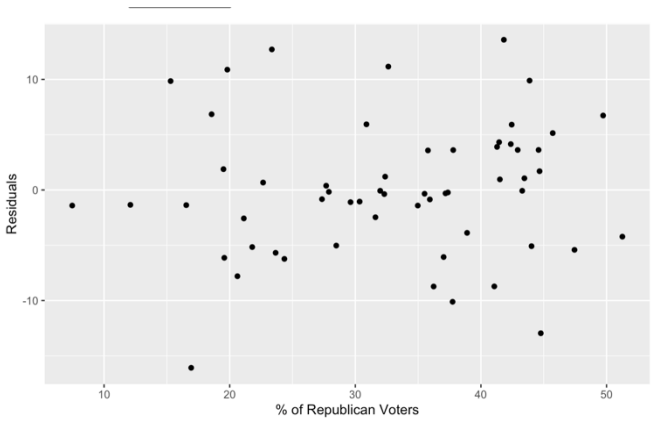
(Table 4: MLR model results)

## Residual Graphs

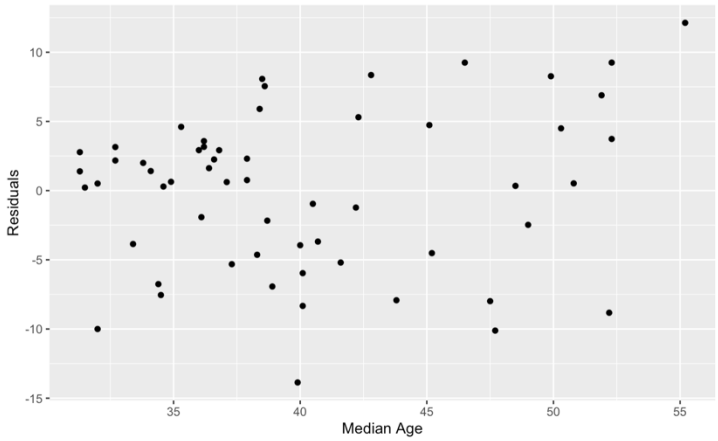
**Residual Plots for each OLS regression**
Race



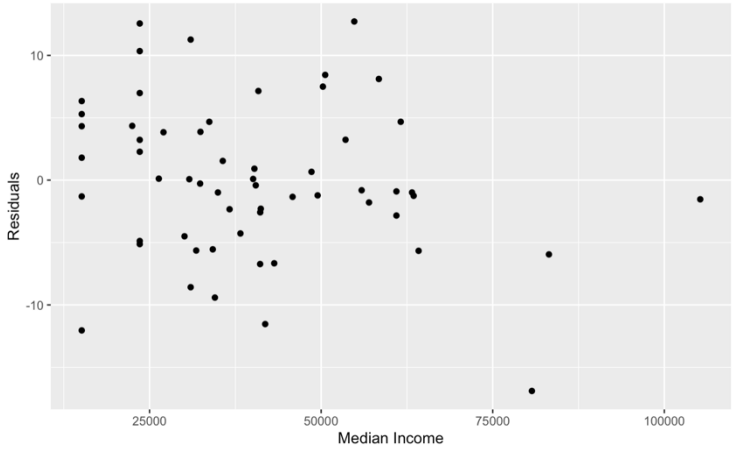

Party Preference



Age




Income




There is no noticeable pattern in the residual plots, hence our linear model seems to be valid.