Voter Turnout

STAT 615: Linear Regression

Caleb Skinner

Introduction

In modern democracies, governments are built on the often razor thin margins of contentious elections. Political parties and partisan groups in countries like the United States spend billions of dollars in hopes of convincing voters to support their candidate. These organizations are certainly interested in flipping votes from their opponents, but, typically, a more reasonable goal is to coax likely supporters from their couches and persuade probable adversaries to stay home. In fact, throughout the United States history, much of the rules and regulations surrounding elections have been mired in this political drama. From Jim Crow laws to Voter ID laws (or lack thereof) to Rice's election day holiday, we cannot escape the political struggle over voter turnout. For these reasons, many groups have considerable interest in predicting and understanding the behavior of voter turnout. Even apolitical citizens can appreciate the general relationship that voter turnout has with the health of a democracy.

In this paper, I give a brief and holistic statistical analysis on the relationship of voter turnout with several demographic and political factors. I assess state-level voter turnout from each presidential election from 2000 to 2024. While other methods may certainly prove more accurate and useful, for the purposes of this class, I will use this data to employ some of the most popular Linear Regression methods.

I compute the voter turnout ratio as the percentage of the voting eligible population that voted in each state. I am mainly interested in three effects: linear temporal trend, perceived close race, and perceived vote impact. I measure the linear temporal trend with the year of the election. If turnout is generally increasing, this should be positive. I measure how close the race is perceived to be by the national polling average leading up to the election. I hypothesize that closer elections incentive voters to vote regardless of their state. Last, I measure the perceived vote impact with the voting margin in the previous presidential election in the state. I suggest that voters associate close past races in their state with greater overall vote impact.

I also include demograhic information like income per capita, age, sex proportions, and ethnicity. For many of these demographic features, my source only included information for 2001 to 2019. For this reason I imposed a simple linear extrapolation on these features. That is, the demographic features of 2000 are estimated with the data for 2001 to 2005 for each state. My source lacked demographic information for five of the states, so I perform the analysis for the remaining 46. Altogether, there are 319 observations and 17 covariates. Overall, the relationship for state i and election j can be expressed

$$\text{turnout}_{i,j} = \beta_0 + \beta_1 \text{year}_j + \beta_2 |\text{nat-polling-margin}_{i,j}| + \beta_3 |\text{previous-margins}_{i,j}| + \beta_4 \text{demographic}_{i,j} + \epsilon_{i,j}$$

Simple OLS

First, I assess the relationship using simple linear regression. I place the full results in Table 1 and the estimates without the demographic features in Table 2. The initial OLS results suggest that voter turnout of a state has a negative relationship with the previous margin in that state, but that voter turnout has

Table 1: OLS Estimates

TERM	ESTIMATE	STANDARD ERROR	STATISTIC	P VALUE
Year	0.001	0.001	1.071	0.285
National Polling Margin	0.424	0.109	3.886	0.000
State Previous Margin	-0.191	0.034	-5.609	0.000

Table 2: OLS Estimates (Demographics removed)

TERM	ESTIMATE	STANDARD ERROR	STATISTIC	P VALUE
Year	0.003	0.000	7.567	0.000
National Polling Margin	0.396	0.124	3.186	0.002
State Previous Margin	-0.171	0.035	-4.902	0.000

an overall positive relationship with the national polling margin. The linear temporal relationship does not produce a statistically significant result. The different estimates in the result without demographic information reflects the importance of including demographic features in a model.

The national polling margin estimate is unexpected, but it is important to remember that this estimate relies on only 7 distinct elections. This makes it very difficult to draw inference on this scale. A more complex hierarchical model is likely be better suited to handle the group structure and estimate this effect than ols. I plot the relationship between national polling margin and voter turnout with a loess smoothing curve in Figure 1 in illustrate my point. In fact, at first, voter turnout decreases as national polling margin increases. Perhaps a better approached would be to categorizing elections as “close” or “not close”.

Figure 1: National Polling Margin vs Voter Turnout

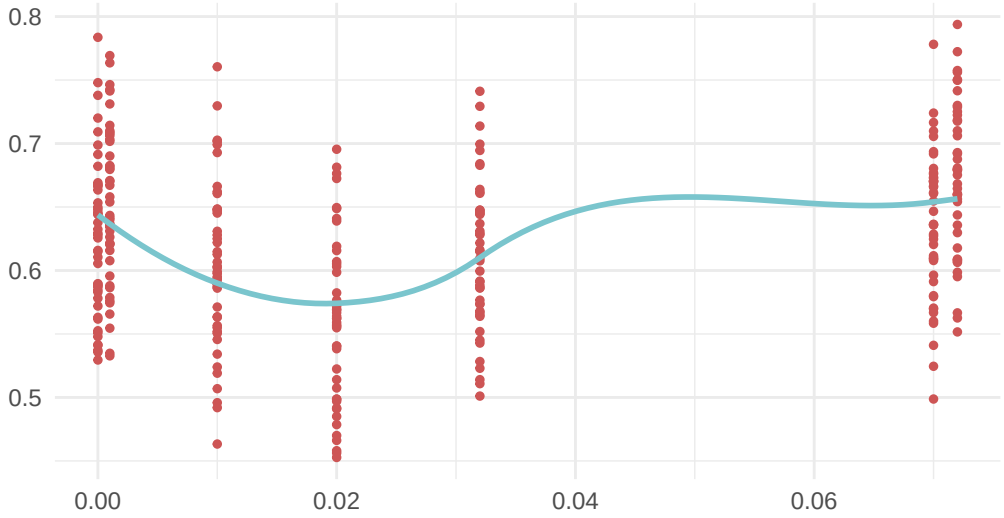


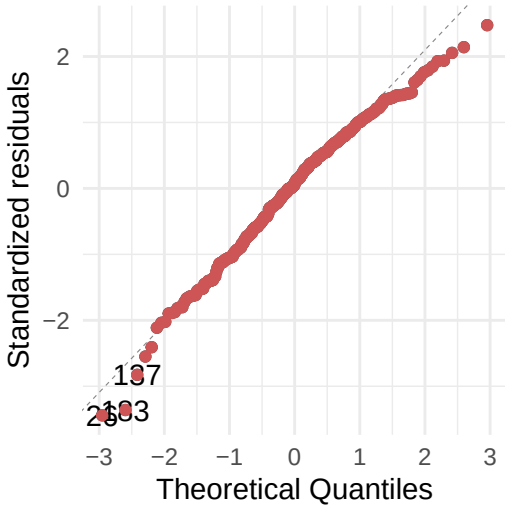
Table 3: 95% Confidence Intervals

	Normality Assumption		Non-Parametric BCa		Residual BCa	
TERM	LOWER	UPPER	LOWER	UPPER	LOWER	UPPER
Year	-0.001	0.003	-0.001	0.004	-0.001	0.003
National Poll Margin	0.209	0.639	0.218	0.620	0.214	0.635
State Previous Margin	-0.258	-0.124	-0.258	-0.116	-0.256	-0.127

Uncertainty Quantification

Next, I create confidence intervals to express uncertainty around my estimates. To assess the assumption of normality, I plot the standardized residuals with the theoretical quantiles in Figure 2 and I conduct a Shapiro-Wilks Test of Normality. The qq-plot appears to be non-normal and the test rejects the null hypothesis of normality (p-value: .0043). For this reason, I compute nonparametric and residual BCa bootstrap intervals to compare with the normal confidence intervals. I place the intervals for the three quantities of interest in Table 3. Overall, the three intervals are very similar and confirm the same conclusions as before.

Figure 2: National Polling Margin vs Voter Turnout



Weighted Regression

Weighted Regression is typically used when the model possesses heteroskedastic errors. I plot the residuals against the fitted values in Figure 3. The errors appear relatively constant, but weighted regression is fun, so let's compute it anyway.

I compute the weights with the following formula

$$w_i = \frac{1}{\hat{y_i}^2}$$

where $\hat{y_i}$ are the estimated residuals of the ols solution by the fitted values with a smoothing spline. In

Figure 3: Absolute Errors vs Fitted Values

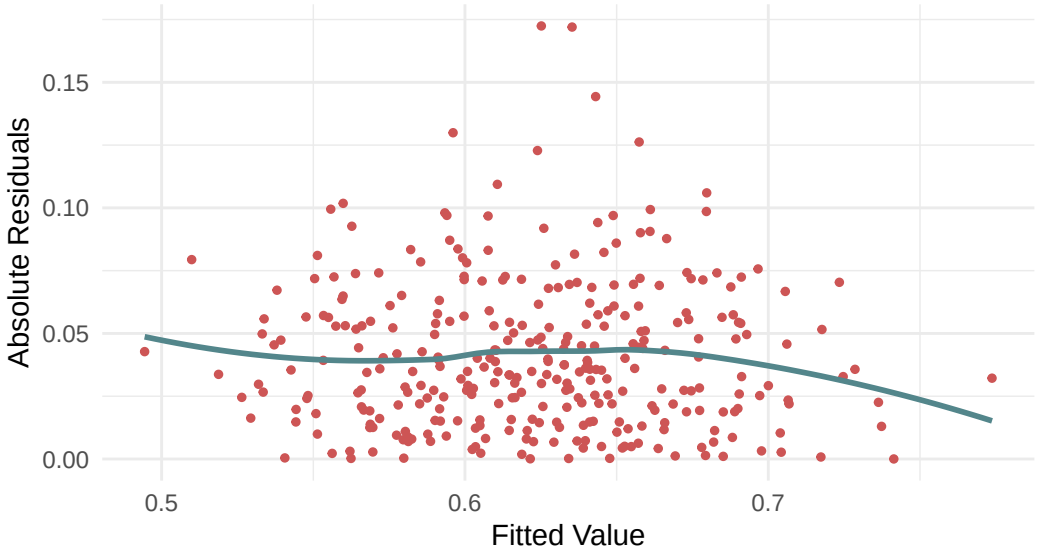


Table 4: Weighted OLS

TERM	ESTIMATE	STANDARD ERROR	STATISTIC	P VALUE
Year	0.001	0.001	1.217	0.224
National Polling Margin	0.392	0.113	3.460	0.001
State Previous Margin	-0.170	0.035	-4.857	0.000

simple terms, the weighted regression organizes the weights so more influence is given to the observations with a larger residual. This should create more homoskedastic errors.

I compute the weighted least squares regression in Table 4 and find similar results as before.

I check for homoskedasticity again in Figure 4, and it appears not much has changed. This is likely because the variance was already fairly homoscedastic to begin with.

Robust Regression

Robust Regression is one method to reduce the effect of outliers on estimates. The model does not appear to have many strong outliers, but a few observations do have high leverage. I compute robust regression by minimizing mean absolute error instead of mean squared error. This is a specific form of quantile regression (with $\tau = .5$). Interestingly, The estimate is an unbiased estimate of the median instead of the mean. I compute the results and place the estimates in Table 5. The estimates and confidence intervals are similar to OLS.

Prediction

Lastly, I am interested in measuring Linear Regression's ability to predict the voter turnout in the 2024 election. I train the data on the 2000-2020 presidential elections and test the results on 2024. For comparison, I use ols, ridge regression, and lasso regression. The penalty term of both ridge and lasso are estimated with

Figure 4: Absolute Errors vs Fitted Values

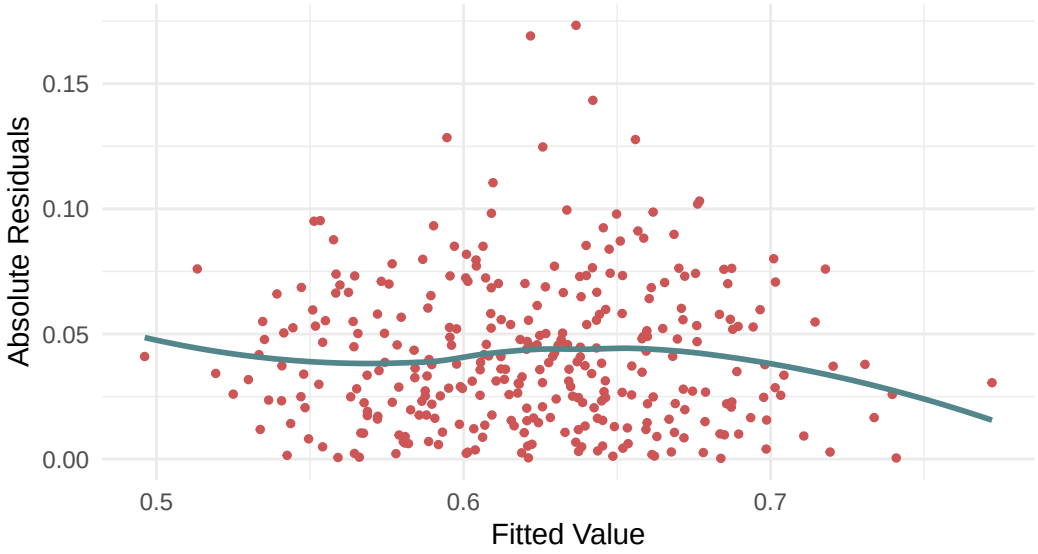


Table 5: Robust Regression

TERM	ESTIMATE	LOWER	UPPER
Year	0.000	-0.002	0.003
National Polling Margin	0.526	0.222	0.781
State Previous Margin	-0.177	-0.243	-0.046

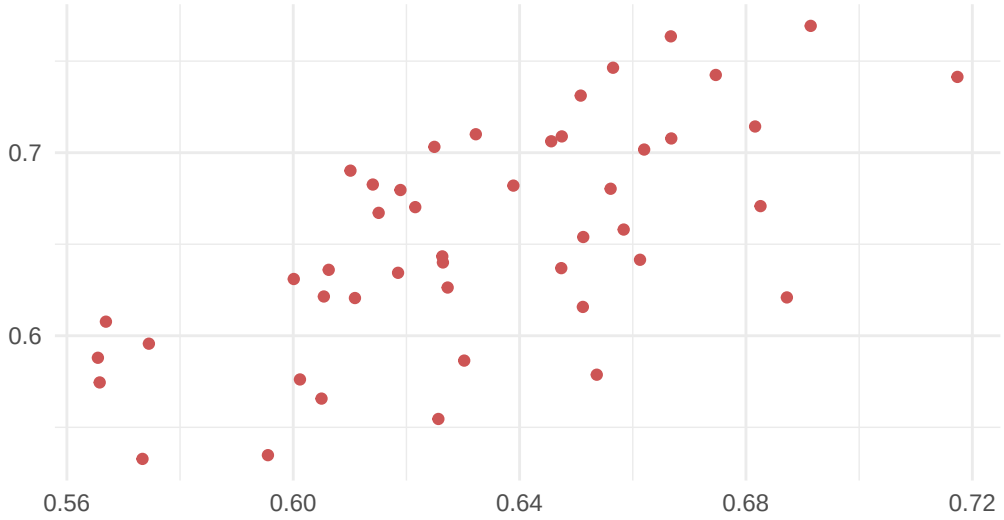
Table 6: Prediction RMSE

METHOD	ESTIMATE
ols	0.052
ridge	0.050
lasso	0.050

cross validation on the training data. I choose the penalty term that minimizes the mean squared error of the prediction.

Ultimately, all three methods perform very similar (Table 6). I plot the relationship between predicted and observed turnout in Figure 5 to demonstrate the relaitvely weak linear predictive power.

Figure 5: Predicted Turnout vs Observed Turnout (Ridge)



Limitations

One major limitation of this method is the unaccounted dependence of the voter turnout of the same state. Clearly, the voter turnout in Alabama in 2020 is highly correlated with the voter turnout in Alabama in 2024. However, outside of general demographic trends, I was unable to capture that relationship in the model. A second limitation is the multicollinearity of some of the demographic variables. More work could be done to accurately account for these important demographic traits.

Future Work

Overall, it would be very interesting to continue this work on voter turnout to smaller scales. It is important to understand what traits or beliefs lead an individual to vote. State-level data makes it difficult to capture these nuanced trends. Moreover, future methods could account for the grouped structure of states, counties, or cities with a hierarchical model.

References

1. Election project (Voter Turnout)
2. Florida Election Lab (Voter Turnout)
3. American Presidency Project (national polling data)
4. Harvard Presidential Election Data Base
5. Stats America (demographic information)