U.S. presidential election prediction, an econometrics and machine learning techniques test

Luca Graziano, David Fernandez, Guido Giacomo Mussini, Matteo Tarenzi
grazianoluca99@gmail.com

December 2021

Abstract

The paper is an attempt at improving 2018 Fair's model forecasts results for the 2020 US elections by introducing new explanatory predictors, defining a State quarterly panel dataset and using the Ordinary Least Squares and Lasso regression estimation techniques. The enriched models successfully performed better than Fair's model and the forecasts are congruous with the real elections outcomes'. Furthermore the paper setup an analysis in an attempt to disentangle causation from correlation with regards to elections' inference results. Finally, we suggest additional studies that could further enrich the analysis studied in this paper.

Microeconometrics

As studied by Fair, economic events measured in real economic activity in the year of election has important effects on the votes for president [1]. These effects can be measured through the growth rate of real GDP per capita or by the change in the unemployment rate. This is complemented by the fact that voters tend to have a very high discount rate (i.e., they do not look very far in the past when analyzing economic results of the incumbent party).

With this in mind, in this text we are predicting the outcome of the United States' 2020 presidential elections between candidates Joe Biden (Democratic) and Donald Trump (Republican), held the third of November of 2020. That time, the elected president was Biden, with 306 electoral votes, against 232 for Trump. It was the first election since 1992, in which the incumbent president failed to win a second term.

To compute our predictions, we are implementing econometric and machine learning techniques based on the model proposed by Fair [2], which seeks to predict the outcome of vote shares in the presidential election given certain economic variables as explanatory. Main differences between our work and the one of Fair's, are that we are computing data at state level, thus we introduce State fixed effects to correct for heterogeneity, we have constructed slightly differently some variables and added new ones we think may improve prediction.

Variables

The model considers the presidential elections from 1980 to 2016, a shorter period frame compared to the one used by Fair. This choice is made to balance the use of fairly accurate variables and a sufficiently long-time interval. Indeed, it allows to find a satisfactory number of variables that, together with those of the previous model (see appendix for details), helps to create a more articulated study. Moreover, differently from Fair's work, it has been used a panel dataset composed of data collected at State quarterly level (see appendix for details), for each of the 51 states.

The new predictors used to build the improved models are the following:

- ***Tradition***: shows the tendency of a state at electing the same party, studied in the time period of 4 elections. Starting from 0, it gets 0.25 every time that democratic party won in that specific state and subtract 0.25 every time the republican party win. The result is a variable that is defined in the interval [-1, +1].
- ***Pop Density***: numbers of individuals for $kilometers^2$
- ***Unemployment***: percentage points difference between the 15^{th} quarter and the 12^{th} quarter of an administration's unemployment rate per state.
- ***Per Income***: computed with the same formula as G, it's the personal income percentage growth in the first three quarters of election years (at annual rate).
- ***Military Exp***: computed with the same formula as G, is the government military expenditure percentage growth in the first three quarters of election years (at annual rate).
- ***TurnOut***: is the percentage of population voting in relation to the total number of persons entitled to vote.

Models

The next part of the report will be dedicated to fitting data to have a statistical understanding of how much explanatory power do the chosen independent variables have in relation to our dependent variable, the *Democratic Vote Share*. All models, except for the 2012 model, are fitted to the data from 1980 up to 2016 and tested against the 2020 data as the purpose of these models is to primarily provide a good forecast of the 2020 U.S elections' outcome. However, as previously suggested some focus will also be briefly placed on the forecast for the 2012 U.S. elections'

as a mean for comparison. In this case, the in-sample data goes from 1980 up to 2008, while the out-of-sample data is that of 2012

A series of statistical models and their results will be presented:

1. A Baseline State fixed effects linear regression model for the 2020
2. A Baseline State fixed effects linear regression model for the 2012
3. An improved State fixed effects linear regression model
4. A double post model selection lasso State fixed effects regression model
5. An improved State fixed effects non-endogenous linear regression model
6. A double post model selection lasso State fixed effects non-endogenous regression model

For all models we are adopting a State fixed effects strategy to solve for the introduction of omitted variable bias derived from using a panel dataset. This variation of the regression technique exploits a series of dummy variables, which can be applied to time and/or to the statistical units, with the purpose to shift the intercept and hold constant time invariant intrinsic peculiar characteristics. The following models present state fixed effects introduced thanks to a dummy variable for each state, for a total of 51 dummy variables. Time fixed effects have been excluded from the model on basis of economic theory, as we expect variation in heterogeneity, for example in economic factors, to be more meaningful across States rather than time.

Baseline Fixed Effects Linear Regression Model

The Baseline State fixed effects linear regression model is an Ordinary Least Squares regression adjusted for heteroskedasticity applied on a panel dataset with six independent variables. This model is an adapted replica of Ray Fair's ordinary least squares model originally fitted and tested against a dataset of historical annual data. With the aim of comparing the forecast results of the 2020 against those of 2012, a State fixed effects regression was run to estimate the coefficients and proceed for the predictions. Other than the use of different in-sample data, no changes have been made between the model of 2020 and 2012.

$$Votes_{t,j} = \beta_j D_j + \beta_1 GxI_{t,j} + \beta_2 PxI_{t,j} + \beta_3 ZxI_{t,j} + \beta_4 DPER_{t,j} + \beta_5 DUR_{t,j} + \beta_6 I_{t,j} \tag{1}$$

The estimates of the presidential equation presented in Table 1 show that only one economic variable (G_i) is significant in both estimations (2012 and 2020) at 99% significance level. Meaning that an increase of one percentage point in real per capita GDP leads to a 0.415 increase in vote share in the 2020 elections and a 0.453 increase in the 2012 elections. As expected, the variables that affects mostly the vote share are the economic growth, as individuals mainly look at economic factors when voting.

P_i, which is a measure of inflation, is showing the desired sign, but not statistical significance. Moreover, it's not adding valuable information to the model and its low coefficient seem to point out that individuals, when voting, are not considering how prices have moved throughout the administration.

On the other hand, Z_i, an overall evaluation of economic growth, is significant only at 90% confidence level and only for the 2020 estimation, meaning that an additional quarter of strong growth leads to an increase of 0.260 percentage points in the vote share. This is proof to the fact that voters tend to have a high discount rate and give higher value at recent results than results shown through larger periods.

Lastly, *DPER* and *DUR* are significant in both models at 99% confidence, while I is significant in both models at 95% confidence. It is worth noticing that in our estimation, *DPER* got an opposite sign as in Fair's estimation (Fair, 2009), meaning that if an incumbent Democratic president is running again, he will have a 2.379 (2.476 for

Table 1: Comparison of 2020 and 2012 baseline models coefficients

Variables	(1) 2020 Model	(2) 2012 Model
G_i	0.415***	0.453***
	(0.067)	(0.072)
P_i	-0.077	-0.051
	(0.048)	(0.054)
Z_i	0.260*	0.237
	(0.141)	(0.155)
DPER	-2.379***	-2.476***
	(0.653)	(0.665)
DUR	-5.552***	-5.601***
	(0.529)	(0.492)
i	3.314**	3.058**
	(1.364)	(1.540)
Constant	35.346***	33.121***
	(2.465)	(2.833)
Observations	510	408
Adjusted R-squared	0.78	0.79
Root MSE	5.0682	4.8619
Endogeneity	-0.034	-0.072
State Fixed Effects	Yes	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

2012) percentage point decrease in the vote share at state level in 2020. For *DUR*, we obtained the same sign as Fair's, meaning that and additional term in the White House reduces the Democrat vote share in 5.552 percentage points (5.601 for 2012). People seem to prefer variability of candidates in hope for more economic growth

On the 2020 model, main object of this analysis, has been first carried out an F-test to check whether the independent variables had joint conditional statistical significance

$$F(6, 448) = 17.33$$
$$\text{Prob} > F = 0.0000$$

showing how the results are statistically distinguishable from zero, thus meaning they are relevant for the model.

Then, a State poolability test was performed to evaluate the joint conditional significance of the State dummy variables and confirm the economic intuition that they are useful to make inferences on the dependent variable.

$$F(50, 453) = 60.04$$
$$\text{Prob} > F = 0.0000$$

The P-values of these two joint significance tests are exactly 0.000 underlining their strong explanatory power on the dependent variable and in particular for the State fixed effects confirm the presence of variation in heterogeneity across States.

These two models represent a great Baseline to further improve on since they do not have omitted variable bias, thanks to the State fixed effects, and do not miss any important predictor that could explain the *Democratic Vote Share*, a statement proven by the visual check that shows lack of correlation between the residuals and the predictions in figure 1 for what concerns the 2020 model and figure 2 for what regards the 2012 model (see Appendix).

Improved State Fixed Effects Linear Regression Model

The second model in Equation 2 is an improved version of the Baseline model that uses the additional variables in an attempt to perform a better forecast. Here, we added variables which we think may explain better the outcome of the elections and thus allow us to have a more accurate prediction of the party shares at State level.

$$Votes_{t,j} = \beta_j D_j + \beta_1 GxI_{t,j} + \beta_2 PxI_{t,j} + \beta_3 ZxI_{t,j} + \beta_4 DPER_{t,j} + \beta_5 DUR_{t,j} + \beta_6 I_{t,j}$$
$$+ \beta_7 Tradition_{t,j} + \beta_8 PopDensity_{t,j} + \beta_9 UxI_{t,j} + \beta_{10} MxI_{t,j} + \beta_{11} TurnOut_{t,j} (2)$$

The results of this estimations can be seen in Table 2. While the model appear to be great, after testing for endogeneity, we found that the model's independent variables had a -0.42 correlation with the error term. Therefore, we decided to drop the endogenous variables *Population Density* and *Tradition*, since any inference from an endogenous model result in an inconsistent estimate. The Improved model adjusted for endogeneity is shown in Equation 3 and had a -0.003 correlation with the error term.

$$Votes_{t,j} = \beta_j D_j + \beta_1 GxI_{t,j} + \beta_2 PxI_{t,j} + \beta_3 ZxI_{t,j} + \beta_4 DPER_{t,j} + \beta_5 DUR_{t,j} + \beta_6 I_{t,j}$$
$$+ \beta_7 UxI_{t,j} + + \beta_8 MxI_{t,j} + \beta_9 TurnOut_{t,j} (3)$$

Double Post Model Selection Lasso State Fixed Effects Regression model

This final model is a double selection lasso State fixed effects regression estimate technique that helps in identifying the most relevant covariates to explain the *Democratic Vote Share*. Its starting point is the Improved model we built.

$$L_{Lasso}(\hat{\beta}) = \sum_{i=1}^{n} (y_i - x_i^j \hat{\beta})^2 + \lambda \sum_{j=1}^{m} |\hat{\beta}_j| \quad (4)$$

The penalty term, whose optimal value has been found via a k-fold cross validation, applied by the lasso regression helps in avoiding over-fitting by shrinking the estimated coefficients toward zero and eventually setting them to zero for some variables. The lasso penalty has been applied to all regressors except our variable of interest GxI which has always been retained in the model.

A variation to the naive application lasso regression has been chosen to avoid introducing omitted variable bias, which is due to the tendency of the lasso to set to zero all variables with modest ability in predicting the dependent variable, despite being highly relevant predictors of the variable of interest. The double selection technique requires two stages. Firstly, it looks for the variables useful to explain the dependent variable (Votes) and secondly the variables relevant to predict the variable of interests, in this case GxI.

The second step is of extreme importance, because exclusion of a modest predictor of the dependent variable but a strong predictor of the independent variable can create a substantial omitted variable bias.

Statistically insignificant variables might have been selected as relevant by the lasso to explain either the dependent variable or the variable of interest in the different stages, but once the union of the meaningfully explanatory variables is taken into account it is possible that the independent variables has failed to add further information than those already provided by the GxI.

Estimation Results

On Table 2 are displayed the results of the all the estimated models for the 2020 elections. Due to the fact that the Lasso regression did not exclude any variable from the Improved model, we ended up running the same regression (both for the endogenous and non-endogenous models). Although we tested different Lasso selection methods, including the plugin method, to select variables and compute estimates, cross validation yielded the most accurate predictions of the 2020 elections, as well as matched with our economic expectations instead of dropping most variables.

In the Improved, non-endogenous model (model 4 in the table), Z is now significant with 99% confidence level, also resulting in a higher coefficient than in the Baseline model (model 1). Regarding *DPER*, we obtained a slightly lower coefficient compared to the Baseline estimation, meaning that for an incumbent president, running again would decrease the vote share in the election. About *DUR*, the coefficient estimated is also lower than in the Baseline model, meaning that the more time a party is in the White House, the lower the vote share for the upcoming election. This can be interpreted as that the voters prefer changing the party in power. Confirming the hypothesis that economic growth is the main aspect that individuals look at when voting

On the other hand, variable I became an insignificant variable, meaning that the incumbent president party does not have an effect on *Democratic Vote Share*.

Military Expenditure also resulted to be statistically significant, meaning that a one percentage point increase, would decrease by 0.240 the *Democratic Vote Share* when the Democrats are in power (i=1), and would increase it when the Republicans are in power (i=-1). This is aligned with the fact that Democrats tend to give less support to military endeavor, therefore people in favour of this expenditure probably would not vote for the Democrat party

Lastly, the *Turnout* does not seem to add information to the model. Increasing the percentage of voters with respect to the potential total votes, does not weight the balance towards any party.

Table 2: Comparison of the models' coefficients - 2020

	(1)	(2)	(3)	(4)	(5)
Variables	Baseline model	Improved model	Lasso model	Improved non-endogenous model	Lasso non-endogenous model
G_i	0.415***	0.360***	0.360***	0.411***	0.411***
	(0.067)	(0.072)	(0.072)	(0.071)	(0.071)
P_i	-0.077	-0.055	-0.055	-0.041	-0.041
	(0.048)	(0.044)	(0.044)	(0.045)	(0.045)
Z_i	0.260*	0.445***	0.445***	0.385***	0.385***
	(0.141)	(0.128)	(0.128)	(0.130)	(0.130)
DPER	-2.379***	-2.621***	-2.621***	-2.426***	-2.426***
	(0.653)	(0.670)	(0.670)	(0.675)	(0.675)
DUR	-5.552***	-6.258***	-6.258***	-5.777***	-5.777***
	(0.529)	(0.596)	(0.596)	(0.590)	(0.590)
i	3.314**	1.808	1.808	1.850	1.850
	(1.364)	(1.324)	(1.324)	(1.328)	(1.328)
tradition		2.043***	2.043***		
		(0.447)	(0.447)		
popdensity		0.019***	0.019***		
		(0.006)	(0.006)		
U_i		-0.815	-0.815	-0.892*	-0.892*
		(0.509)	(0.509)	(0.509)	(0.509)
M_i		-0.199***	-0.199***	-0.240***	-0.240***
		(0.052)	(0.052)	(0.053)	(0.053)
TurnOut		-0.034	-0.034	-0.028	-0.028
		(0.050)	(0.050)	(0.052)	(0.052)
Constant	35.346***	40.307***	40.307***	37.912***	37.912***
	(2.465)	(3.510)	(3.510)	(3.574)	(3.574)
Observations	510	510	510	510	510
Adjusted R-squared	0.787	0.819	0.819	0.808	0.808
Root MSE	5.068	4.673	4.673	4.812	4.812
Endogeneity	-0.034	-0.420	-0.420	-0.003	-0.003
State Fixed Effects	Yes	Yes	Yes	Yes	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Forecast Results

The U.S electoral system is based on the majority system that was not considered by Fair who used the popular vote in his work. The model has been adapted then in order to predict the number of electoral votes taken by the 2 main parties in each State, for each election from 1980 to 2020.

Working with the majority system, however, brings some difficulties: firstly, the threshold that gives all the votes to one party if takes more than 50% can affect in a large way the winner prediction. Small errors around the 50% can have a big impact on the electoral votes if the prediction assign the state to the wrong party. Indeed, a 2% mistake in popular votes between 49% and 51% will have a much greater effect on the winner prediction than making a 10% mistake between 51% and 61%.

Additionally, each State has a different number of Electoral Votes (E.V.): for example an error in Texas that has 40 E.V. will be more relevant on the prediction than an error in Virginia that has only 3 E.V.

Moreover, the errors could randomly correct themselves. Wrong electoral prediction can compensate each other creating correct final results. This mean that it is important, for the purpose of evaluating the forecasting abilities of the various models, considering one by one all the States, which is presented in table 4 and not only in the final elections' outcome

For the Baseline model, we show the results of the 2012 and 2020 predictions in Table 3, it appears that the 2012 has predicted more accurately the democratic share than the 2020 model did. This result is rather unexpected since the predictions of the 2012 model are based on a smaller number of observations and the same number of parameters, and thus should present an higher standard error. We believe the higher error in prediction for the 2020 model is due to the breaking out of the Covid19 pandemic and the inability for historical data to reflect the rather exceptional social circumstances the elections have been carried out in.

In conclusion, it can be seen how while the 2020 model failed to predict correctly the elections' outcome, the 2012 model due to the unexpected lower MSE successfully forecasted the results.

Table 3: Comparison between the 2012 and the 2020 predictions of electoral votes

Year	Real Votes	Threshold to win	Predicted Votes	Error (in Votes)
2020	308	270	243	-65
2012	332	270	282	-50

The negative sign in the error relates to an under-prediction;
while a positive sign relates to an over-prediction

When comparing the results produced by the five estimated models for the 2020 elections in table 4, it can be seen that the Baseline model has performed worse than the rest. The predictions from the improved and lasso model are nearly perfect with an over-prediction of merely 3 votes, also reflected by the lower Root MSE of 0.046. Nevertheless, as mentioned before, these are endogenous models, so we prefer to stick with the non-endogenous models even though they did not had as good predictive power.

The double selection lasso regression has not been able to further improve on the already accurate Improved models. Once again, the matching predictions are due to the fact that the lasso regression has not modified the Improved models at all.

In table 5 we show the descriptive statistics of the predictions of all estimated models for the 2020 elections. It turns out that the Baseline model under-predicted the actual values by a rough 2 percentage points, while the non-endogenous models over-predicted by also 2 percentage points.

In addition, it can also be seen by looking at the standard deviation that the non-endogenous models, that they are more capable to capture the real variation than the Baseline model. Finally, by looking at table 6 it is shown that the non-endogenous models have a lower mean error, lower variation in error and lower maximum error than the Baseline model, meaning it over-predicts, but with less intensity.

Table 4: Comparison of the predictions of all 2020 models

Model	Real Votes	Threshold to win	Predicted Votes	Error (in Votes)
Baseline Model	306	270	243	-63
Improved Model	306	270	309	+3
Lasso Model	306	270	309	+3
Improved non-endogenous Model	306	270	291	-15
Lasso non-endogenous Model	306	270	291	-15

The negative sign in the error relates to an under-prediction;
while a positive sign relates to an over-prediction

Table 5: Comparison of descriptive statistics of predictions between all 2020 models

Model	mean	std	min	max
True Dem Share	49.698775	12.313016	27.519566	94.466954
Baseline Predictions	47.653862	9.398993	27.922493	89.378677
Improved Predictions	51.438530	10.787266	32.574490	100.768440
Lasso Predictions	51.438530	10.787266	32.574490	100.768440
Improved non-endogenous Predictions	51.65237	10.32142	31.43247	92.442
Lasso non-endogenous Predictions	51.65237	10.32142	31.43247	92.442

Table 6: Comparison of descriptive statistics of residuals between all 2020 models

Model	mean	std	min	max
Baseline Residuals	2.044	6.343	-17.911	12.617
Improved Residuals	-1.739	5.321	-18.140	6.998
Lasso Residuals	-1.739	5.321	-18.140	6.998
Improved non-endogenous Residuals	-1.953	5.867	-20.320	8.097
Lasso non-endogenous Residuals	-1.953	5.867	-20.320	8.097

Causal Inference

The different ideologies of Rep and Dem are represented by the policies pursued by the respective candidates for the presidency. We will expect, for example, a growth in Military expenditure during the office of a republican president. It may therefore be of interest to investigate whether there is a causal relationship between presidential alternation and its impact on the country's main socio-economic indicators.

To study this type of causal relation, it may be appropriate to use Difference in Difference method. However, implement this analysis bring some challenges.

- It is difficult to identify what the treatment is. In fact, if we consider it as the election of a democratic president, it would be hard to identify a control group, since each state would be affected in the same way.
- Although the policies are implemented at federal level, the states have a very strong heterogeneity that can influence the impact of national policies and thus invalidate the analysis. In particular, the states differ in ideology, socioeconomic characteristics, geography, and climate. This leads state governments to have different priorities and legislate on the basis of the needs of their state

To limit the influence of external variables, it is possible to study similar territories, for example, states geographically close and with comparable socio-economic characteristics. However, it would be impossible to generalize the results on national scale because of the heterogeneity of the US.
Given the difficulties in identifying a control and a treatment group, it may be useful to use a technique such as Regression Discontinuity Design on different Variables. For example, it could be verified if there was a significant increase in military spending in the transition between Obama and Trump in 2016. This type of analysis, carried out for individual states, would largely eliminate the difficulties mentioned above. Since this analysis is carried out on individual states in consecutive periods, it would make it possible to eliminate the noise caused by structural differences between states. In addition, after having obtained all the studies at the state level, it would be possible to aggregate the obtained data and to draw conclusions at the national level.

However, this method also has its challenges:

- Defining the threshold correctly could be complex. It would be difficult to identify when the policies of the new president start to show their effect
- There may be external events that could invalidate the analysis. For example, the impact of possible wars, pandemics or economic crises would affect both the policies implemented by presidents and the variables that are tested.

Summarising, given the differences between the States, particularly at the cultural and socio-economic level, it would be better to carry out a study on a state-by-state basis than to compare them. It is thus considered that the RDD can provide more consistent results.

In conclusion, it is considered that carrying out this type of analysis could be useful in order to decode which are actually the effects of the various policies on the territory, empirically studying which are those that show best results -think, for example, of the recent Nobel Prize in Economics for a study on the effectiveness of the minimum wage-. These analyses can also help to outline the effective power that a party in power has, that is, how its proposed policies can concretely affect the social and economic dimensions of the country and what would be its influence once the party in power is alternated.

Conclusions

Starting from Fair's model, we managed to build a model that is able to provide more accurate forecasts by adding new variables and adapting the estimation strategy.

Concerning the encountered challenges, the main obstacle to the analysis has been finding relevant quantitative and explanatory variables on quarterly State data that could help in reaching better forecasts. Having better data availability would improve the estimation and thus, the predictions, as well as including predictors that relate to social themes. Furthermore, we favoured a linear model to ensure better interpretation of the results but a non-linear model would probably ensure improved predictions.

To enrich the analysis, it would be interesting performing additional studies, for example a sentiment analysis on political social networks, including Twitter.

Finally, empirical models to disentangle the causation from correlation issue could be prepared to align forecasts results with an expectation of policies' effects on the particular States in the long run.

Bibliography

References

[1] Ray C Fair. The effect of economic events on votes for president. *The review of economics and statistics*, pages 159–173, 1978.

[2] Ray C Fair. Presidential and congressional vote-share equations. *American Journal of Political Science*, 53(1):55–72, 2009.

Appendix

1 Data and Fair's Variables

The votes data used to compute the dependent variable was downloaded from the Bureau of Economic Analysis (BEA) website. The *Democratic Vote Share* has been constructed using only the votes obtained by democrats and republicans, eliminating other parties who ran for the White House.

Here's the explanation of all the variables used in the model:

Variables in Fair's model (Roy C.Fair , 2018):

- ***Votes***: as in Fair's work, is the "democratic share of the two-party presidential vote"
- ***G***: is the Real GDP per capita percentual growth in the first three quarters of election year (at annual rate).
- ***i***: is 1 if the Democrats are in the White House at the time of the election, -1 otherwise.
- ***DUR***: is 0 if either party has been in the White House for one term, 1 [-1] if the Democratic [Republican] party has been in the White House for two consecutive terms, 1.25 [-1.25] if the Democratic [Republican] party has been in the White House for three consecutive terms, 1.50 [-1.50] if the Democratic [Republican] party has been in the White House for four consecutive terms, and so on.
- ***DPER***: is 1 if a Democratic presidential incumbent is running again, -1 if a Republican presidential incumbent is running again, and 0 otherwise.
- ***P***: is absolute value of the growth rate of the GDP deflator in the first 3 years of the administration.
- ***Z***: Number of quarters of the first 3 years which registered a growth rate of real per capita GDP greater than 3.2% (see Appendix for details).

To compute Fair's variables, we downloaded quarterly data on nominal Gross Domestic Product (GDP), Consumer Price Index, and population at State level.

The quarterly GDP data by state was found only from 2005 to 2020 from the Bureau of Economics and Analytics (BEA) website. To extend our time span, we divided the annual GDP from 1980 to 2005 by 4, for each state. Although this doesn't allow to capture quarterly peaks or falls in the GDP growth between 1980 and 2004, especially around the election period, by joining this approximation with the real data between 2005 and 2020, the prediction ability of the model is improved.

The Consumer Price Index was obtained from the Federal Reserve Economic Data (FRED) webpage. We used the annual, not seasonally adjusted series from 1980 to 2020 to discount the inflation effect on the nominal variables (GDP, Income and Military Expenditure) of the model. We computed the real variables in 2020 prices.

Quarterly population by state was not available. Instead, the model uses then the annual population estimates for each state made by the American Census. Data was downloaded from their website. The annual estimates were made every ten years using the decade population as base.

As shown, in the model we have considered all the variables introduced by Fair, except for *WAR*. These variables have been constructed as in the previous study, using the nominal GDP, the real GDP, and the population for each state.

For Z, it has been applied a correction for the period from 1980 to 2004. As we only had annual GDP data at state level for this period, we divided it by 4 so it would match the rest of the dataset. When computing Z for this time frame, the result wouldn't be comparable to the computation of the 2004 to 2020 period (for which we did had quarterly data). Therefore, we empirically corrected this issue in the following three steps:

1. Add to each value between 1980 and 2004 a number X.
   With X = AVG(2005:2020) - AVG(1980:2004)
2. Round down the results of step 1
3. Add to these data a random value between [-2:2]

Apart from the use of a state fixed effect strategy, our model proposal differs from Fair's original model as we added new variables to improve it's predictive power.

First, *Tradition* has been created to study a certain trend that does not depend on the economic performance, but on the frequency that a given party wins a state during elections. It was built using the data of the winners of the previous 4 elections. The second added variable is *Population Density*. It was defined by dividing the state population by the extension in km squared.

Another new variable is the annual growth of the unemployment rate. We computed it as the difference between the level of unemployment in the last quarter before the election and the level at the last quarter of the third year in office. It tests the effects of a possible increase in unemployment in the year preceding the elections. The variable has been constructed in this way for two reasons: 1) in the first year of the mandate the new president can implement his policies, but it will probably take some time before they show a significant effect. 2) As Fair's studied it (Fair, 1978), population tend to have a high discount rate (short memory) so they judge a president's performance based on the most recent year of it's mandate rather than over the entire term in office.

Military Expenditure has been included as well to try to determine if it has an influence on the elections, as the amount of money spent on defense is one of the most significant democratic-republican contrasts.

Finally, through Turnout we want to determine whether there is a correlation between the number of votes that Democrats get and the percentage of people who vote (out of the total potential voters).

2 Additional graphs and tables

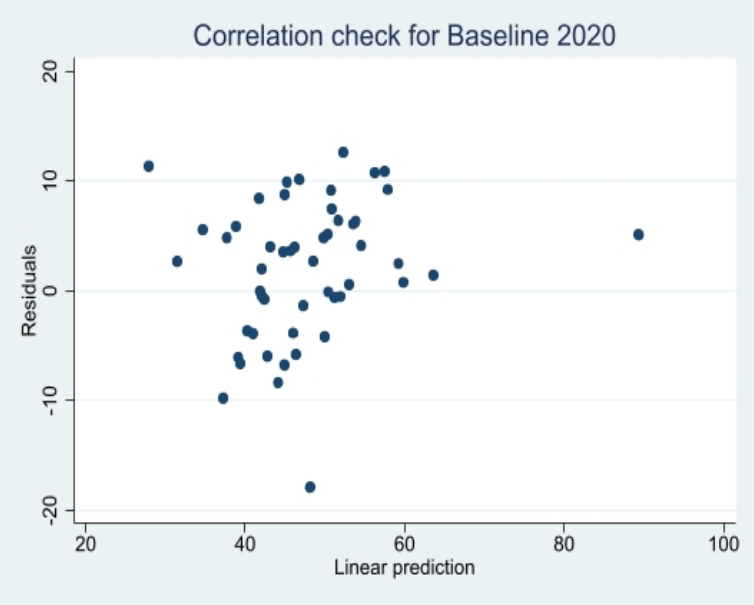

Figure 1: Correlation between the residuals and the predictions of 2020

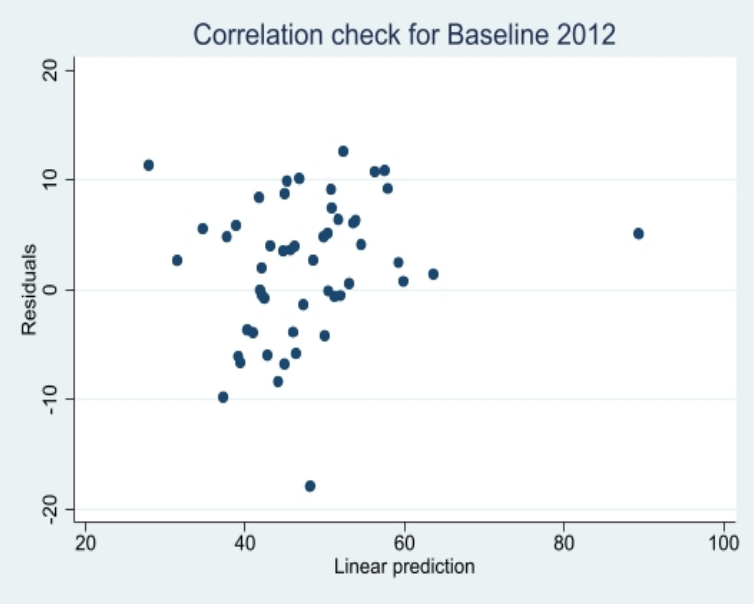

Figure 2: Correlation between the residuals and the predictions of 2012

Table 7: 2020 Votes prediction, comparison by state

State	True Dem Votes	Baseline Votes	Improved Votes	Lasso Votes	Improved non-end Votes	Lasso non-end Votes
AK	0	0	0	0	0	0
AL	0	0	0	0	0	0
AR	0	0	0	0	0	0
AZ	11	0	0	0	0	0
CA	55	55	55	55	55	55
CO	9	0	9	9	9	9
CT	7	7	7	7	7	7
DC	3	3	3	3	3	3
DE	3	3	3	3	3	3
FL	0	0	0	0	0	0
GA	16	0	16	16	0	0
HI	4	4	4	4	4	4
IA	0	0	6	6	6	6
ID	0	0	0	0	0	0
IL	20	20	20	20	20	20
IN	0	0	0	0	0	0
KS	0	0	0	0	0	0
KY	0	0	0	0	0	0
LA	0	0	8	8	0	0
MA	11	11	11	11	11	11
MD	10	10	10	10	10	10
ME	4	0	4	4	4	4
MI	16	16	16	16	16	16
MN	10	10	10	10	10	10
MO	0	0	0	0	0	0
MS	0	0	0	0	0	0
MT	0	0	0	0	0	0
NC	0	0	0	0	0	0
ND	0	0	0	0	0	0
NE	0	0	0	0	0	0
NH	4	0	4	4	0	0
NJ	14	14	14	14	14	14
NM	5	5	5	5	5	5
NV	6	0	6	6	6	6
NY	29	29	29	29	29	29
OH	0	0	0	0	18	18
OK	0	0	0	0	0	0
OR	7	7	7	7	7	7
PA	20	20	20	20	20	20
RI	4	4	4	4	4	4
SC	0	0	0	0	0	0
SD	0	0	0	0	0	0
TN	0	0	0	0	0	0
TX	0	0	0	0	0	0
UT	0	0	0	0	0	0
VA	13	0	13	13	0	0
VT	3	3	3	3	3	3
WA	12	12	12	12	12	12
WI	10	10	10	10	10	10
WV	0	0	0	0	5	5
WY	0	0	0	0	0	0
Total	306	243	309	309	291	291

Table 8: 2020 vote share prediction, comparison by state

state	True Dem Share	Baseline Pred	Improved Pred	Lasso Pred	Improved non-end Pred	Lasso non-end Pred
AK	44.738	38.866	40.651	40.651	41.542	41.542
AL	37.089	40.942	43.355	43.355	44.825	44.825
AR	35.788	44.132	48.665	48.665	47.303	47.303
AZ	50.157	41.765	44.286	44.286	45.490	45.490
CA	64.909	52.292	59.738	59.738	59.339	59.339
CO	56.938	46.819	53.061	53.061	51.143	51.143
CT	60.195	53.841	59.252	59.252	58.274	58.274
DC	94.467	89.379	100.768	100.768	92.442	92.442
DE	59.627	53.535	57.802	57.802	56.519	56.519
FL	48.305	44.770	48.510	48.510	49.972	49.972
GA	50.119	46.178	50.762	50.762	49.096	49.096
HI	65.033	63.586	68.382	68.382	73.230	73.230
IA	45.817	49.966	50.386	50.386	52.789	52.789
ID	34.123	31.461	38.117	38.117	35.520	35.520
IL	58.659	54.547	59.306	59.306	61.166	61.166
IN	41.805	41.831	46.437	46.437	45.658	45.658
KS	42.507	37.683	40.524	40.524	41.191	41.191
KY	36.800	42.779	41.546	41.546	43.557	43.557
LA	40.536	46.341	50.124	50.124	48.896	48.896
MA	67.116	57.878	63.876	63.876	63.812	63.812
MD	67.029	56.246	60.486	60.486	59.957	59.957
ME	54.670	49.864	53.242	53.242	53.892	53.892
MI	51.414	51.935	55.544	55.544	56.517	56.517
MN	53.640	53.032	53.524	53.524	56.167	56.167
MO	42.164	46.000	46.949	46.949	48.444	48.444
MS	41.615	42.135	43.256	43.256	44.393	44.393
MT	41.603	42.385	45.021	45.021	45.235	45.235
NC	49.316	45.676	49.395	49.396	49.605	49.605
ND	32.783	39.378	41.345	41.345	41.610	41.610
NE	40.216	34.647	37.849	37.849	36.546	36.546
NH	53.748	44.954	50.774	50.774	49.355	49.355
NJ	58.071	51.682	57.680	57.680	57.133	57.133
NM	55.518	50.352	54.697	54.697	54.700	54.700
NV	51.223	48.516	59.126	59.126	57.872	57.872
NY	61.717	59.225	64.865	64.865	66.687	66.687
OH	45.923	47.293	49.160	49.160	50.618	50.618
OK	33.060	39.117	41.274	41.274	40.523	40.523
OR	58.307	50.831	56.808	56.808	56.125	56.125
PA	50.589	51.175	54.282	54.282	55.139	55.139
RI	60.599	59.828	65.552	65.552	68.014	68.014
SC	44.073	42.060	44.966	44.966	45.977	45.977
SD	36.565	40.227	42.550	42.550	43.110	43.110
TN	38.172	44.899	46.606	46.606	48.842	48.842
TX	47.169	43.141	45.970	45.970	46.792	46.792
UT	39.306	27.922	32.574	32.574	31.432	31.432
VA	55.155	45.221	50.731	50.731	49.429	49.429
VT	68.299	57.450	61.300	61.300	60.202	60.202
WA	59.926	50.777	56.514	56.514	56.765	56.765
WI	50.319	50.417	50.590	50.590	53.174	53.174
WV	30.201	48.113	48.342	48.342	50.522	50.522
WY	27.520	37.259	36.843	36.843	37.727	37.727

Table 9: Vote share residual comparison

State	True Dem Share	Baseline Res	Improved Res	Lasso Res	Improved non-end Res	Lasso non-end Res
AK	44.738	5.873	4.087	4.087	3.196	3.196
AL	37.089	-3.854	-6.266	-6.266	-7.736	-7.736
AR	35.788	-8.344	-12.878	-12.878	-11.516	-11.516
AZ	50.157	8.391	5.871	5.871	4.667	4.667
CA	64.909	12.617	5.171	5.171	5.570	5.570
CO	56.938	10.119	3.877	3.877	5.795	5.795
CT	60.195	6.354	0.943	0.943	1.921	1.921
DC	94.467	5.088	-6.301	-6.301	2.025	2.025
DE	59.627	6.092	1.825	1.825	3.108	3.108
FL	48.305	3.535	-0.204	-0.204	-1.667	-1.667
GA	50.119	3.942	-0.642	-0.642	1.024	1.024
HI	65.033	1.447	-3.349	-3.349	-8.198	-8.198
IA	45.817	-4.150	-4.570	-4.570	-6.972	-6.972
ID	34.123	2.662	-3.994	-3.994	-1.398	-1.398
IL	58.659	4.112	-0.647	-0.647	-2.507	-2.507
IN	41.805	-0.026	-4.632	-4.632	-3.853	-3.853
KS	42.507	4.823	1.982	1.982	1.316	1.316
KY	36.800	-5.979	-4.746	-4.746	-6.757	-6.757
LA	40.536	-5.805	-9.588	-9.588	-8.360	-8.360
MA	67.116	9.238	3.239	3.239	3.304	3.304
MD	67.029	10.783	6.543	6.543	7.072	7.072
ME	54.670	4.807	1.428	1.428	0.779	0.779
MI	51.414	-0.521	-4.130	-4.130	-5.104	-5.104
MN	53.640	0.607	0.115	0.115	-2.527	-2.527
MO	42.164	-3.836	-4.785	-4.785	-6.280	-6.280
MS	41.615	-0.520	-1.641	-1.641	-2.778	-2.778
MT	41.603	-0.782	-3.418	-3.418	-3.632	-3.632
NC	49.316	3.640	-0.080	-0.080	-0.289	-0.289
ND	32.783	-6.595	-8.562	-8.562	-8.828	-8.828
NE	40.216	5.569	2.367	2.367	3.670	3.670
NH	53.748	8.795	2.975	2.975	4.393	4.393
NJ	58.071	6.390	0.392	0.392	0.938	0.938
NM	55.518	5.167	0.821	0.821	0.818	0.818
NV	51.223	2.707	-7.903	-7.903	-6.649	-6.649
NY	61.717	2.491	-3.148	-3.148	-4.971	-4.971
OH	45.923	-1.370	-3.236	-3.236	-4.695	-4.695
OK	33.060	-6.057	-8.214	-8.214	-7.463	-7.463
OR	58.307	7.477	1.500	1.500	2.182	2.182
PA	50.589	-0.586	-3.692	-3.692	-4.550	-4.550
RI	60.599	0.771	-4.953	-4.953	-7.414	-7.414
SC	44.073	2.013	-0.892	-0.892	-1.904	-1.904
SD	36.565	-3.661	-5.985	-5.985	-6.545	-6.545
TN	38.172	-6.727	-8.434	-8.434	-10.670	-10.670
TX	47.169	4.028	1.199	1.199	0.377	0.377
UT	39.306	11.384	6.732	6.732	7.874	7.874
VA	55.155	9.934	4.423	4.423	5.725	5.725
VT	68.299	10.849	6.999	6.999	8.098	8.098
WA	59.926	9.149	3.412	3.412	3.160	3.160
WI	50.319	-0.098	-0.271	-0.271	-2.855	-2.855
WV	30.201	-17.912	-18.140	-18.140	-20.320	-20.320
WY	27.520	-9.740	-9.324	-9.324	-10.207	-10.207