Microeconometrics and Causal Inference Empirical Project:

Analyzing and Predicting US Elections

Abstract

In this paper, we analyzed US presidential elections focusing on economic and incumbency variables. We used the equations in Fair's research (2009) as the base of our analyses. The equations are updated incorporating new data available as of December 2020, additional variables are introduced, and different model approaches are compared in order to use a reliable method to predict US elections as well as to make inferences about economic and political issues. Double-Selection Lasso is individuated as the best model approach among selected others for its reduced errors characteristic. Lastly, its capacity to provide uniform causal inferences increases the statistical power of our analysis.

1. Introduction

In Fair's research, three vote-share equations are estimated using different timing criteria. For the purposes of our paper, we utilized only the first equation for predicting Vp, the Democratic share of the two-party presidential vote. The original paper equation is as follows:

$$V^{p} = k + \alpha \cdot G + \beta \cdot P + \delta \cdot Z$$

Where *k* is the constant term including all the non-economic values such as *DPER*, whether presidential incumbent is running again, *DUR*, whether a party has been in power for one term or more, and *I*, which presidential incumbent party is in power at the time of election. The symbols α, β, δ represent the estimated coefficients respectively of *G*, growth rate of real per capita GDP in the first three quarters of the on-term election year (annual rate), *P*, absolute value of the growth rate of the GDP deflator in the first 15 quarters of the administration (annual rate), and *Z*, the number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2 percent at an annual rate. Searching for alternative and possibly better models, some variables such as *WAR* which accounted for whether a war was initiated during the election year were excluded, but other variables such as unemployment growth rate were added in order to understand other possible connections with the democratic share prediction.

Considering the sample period range from 1976 to 2019, the equation is modified and re-evaluated to gain substantial insights about the inherent capacity of economic variables to forecast US Elections, in particular the variables of interest are derived from GDP such as growth rate. If proved true, this would cast a dilemma on whether the Elections are affected by the economy of the country or the opposite. In addition, what role do public policies play in this causal chain? Are public policies really chosen and caused by the voters persuaded by the economic circumstances in that precise period? Are they chosen and caused by voters at all? On the other hand, how is the economy impacted by public polices of one party instead of another?

The selection, cleaning, transformation and exploration of the data is presented in *Section 2*; the methodology behind our analysis is presented in *Section 3*; the search for alternative models and the findings are presented in *Section 4*, while the conclusions are presented in *Section 5*.

2. Data Selection and Exploration

As starting point, we collected the data to rebuild the equation described above from Fair's original paper. Every dataset was gathered from official sources of the US government and other reliable affiliates. Having this in mind, we must say that Fair uses quarterly data that is annualized to compute the coefficient as well as he considered the US as a whole country. In our analysis, we consider each state of the US separately (since we are studying presidential elections, we treated Washington D.C as separated country like in the electoral system). This approach didn't allow us to deal with quarterly data since it was available only for a small part of the considered period. Furthermore, datasets such as unemployment rate and demographics started to be collected stately only from recent years as well. Subsequently, the data was re-modeled using panel data layout as this allowed us to detect more information more accurately (i.e., more degrees of freedom, more variability and more statistical effects). After merging the datasets from different years and removing superfluous variables, we computed the original equation estimators *DPER, DUR, I, G, P* and *Z*, but using different time criteria (quarters were\ interpreted as years, i.e., 15 quarters is equal to 4 years). Having imported the dataset of the original variables, a closer assessment can be executed to search for insights before the analysis.

3. Methodology

Once the data was collected, transformed and imported correctly, we started with descriptive analytics tools, mainly graphical, as we were looking for important information that could have been useful before modelling such as distributions, patterns, heterogeneity and signs of possible fixed effects. We proceeded with the multiple regression model, which is an extension of the simple linear model, however with combined variables such as *G*I, P*I* and *Z*I.* We plotted the residuals to verify whether our model should have been adjusted for heteroskedasticity though a robust regression, and as a further confirmation, we run the Breusch-Pagan/Cook-Weisberg test. Our main goal was to create a model with the least MSE coefficient, with a fairly high R-squared value and unbiased. Following this philosophy, we decided to test the model for fixed effects which gave a great boost to the predicting power of the multiple regression. After performing a joint significance test of the economic variables and the poolability test, we confirmed fixed effects in our data and predicted consequently the dependent variable. We compared the forecast of the democratic share of the votes, which was translated into electoral votes, with the results of the 2012 and 2016 elections. Nevertheless, we wanted to be cautious towards the findings, so we applied the Least Absolute Shrinkage and Selection Operator, also LASSO which is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and the inferential interpretability of the model. We performed both the Naïve Post-Selection Lasso and the Double-Selection Lasso, the latest being the most satisfying one considering our goals.

4. Alternative Models and Findings

4.1 *Exploring Panel Data*

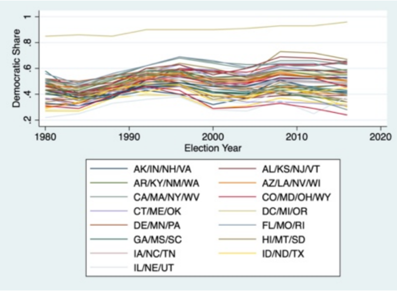


Figure 1. Vp - TwoWay Scatter, overlay

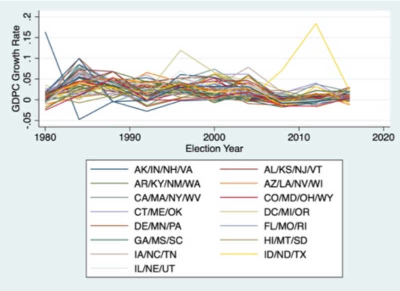


Figure 2. G - TwoWay Scatter, overlay

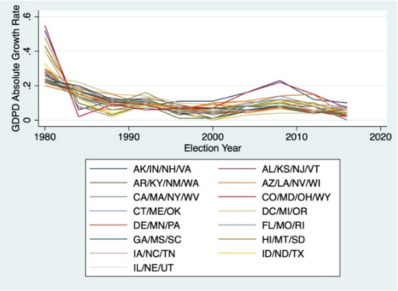


Figure 3. P - TwoWay Scatter, overlay

As we can see from the graphs above, there is not a particular pattern except for the positively growing trend through the years. But we can confirm that there is good heterogeneity across our dataset as it is confirmed also by figure 4 below.

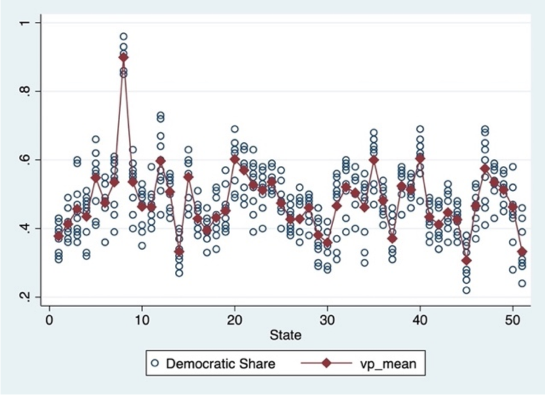
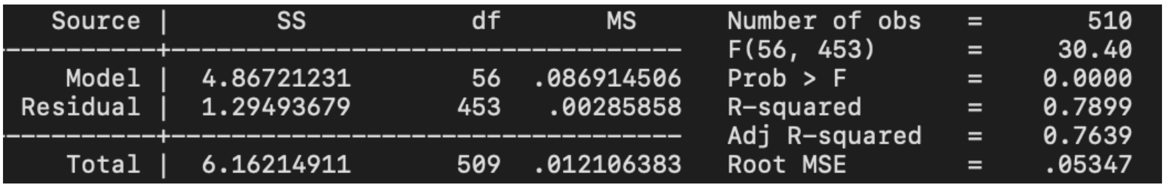


Figure 4. Fixed Effects: Heterogeneity across States

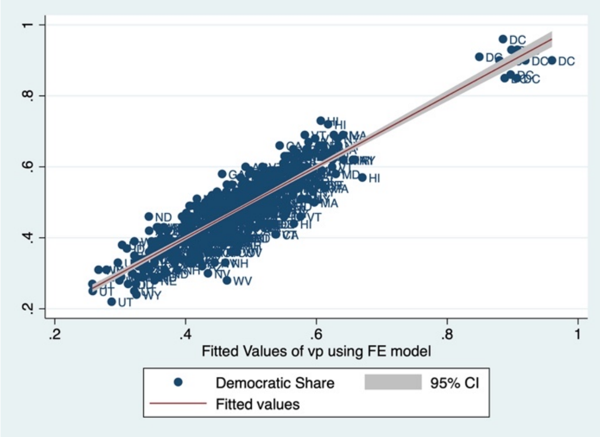
4.2 *Fixed-Effects Multiple Regression Model*

Since the standard Multiple Regression has a very poor R-squared value and considering our previous assumption, we used the Least Square Dummy Variable model for each state (we exclude the first state as base state) in order to implement our fixed-effects model. The Adjusted R-squared is fairly high because it is strongly increased by the state fixed-effects modelling process. The results are shown in Figure 5, 6 and 7 below.



```
      Source |       SS           df       MS      Number of obs   =       510
-------------+----------------------------------   F(56, 453)      =     30.40
       Model |  4.86721231        56  .086914506   Prob > F        =    0.0000
    Residual |  1.29493679       453   .00285858   R-squared       =    0.7899
-------------+----------------------------------   Adj R-squared   =    0.7639
       Total |  6.16214911       509  .012106383   Root MSE        =    .05347
```

Figure 5. Fixed-Effects Multiple Regression Model



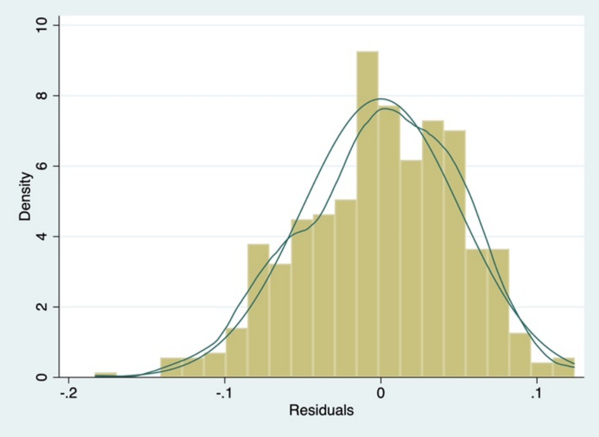
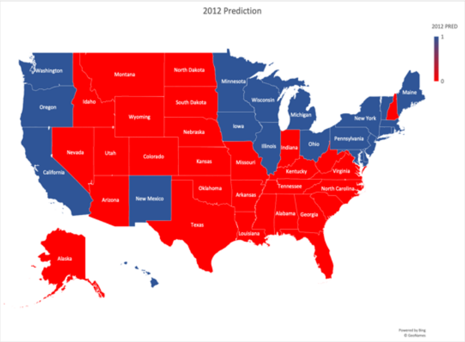


Figure 6. FE Model - Fitting Power

Figure 7. FE Model - Residuals

Except for z_i, all other estimators are statistically significant as the p-value is less than 0.05. As we can see, our model is fitting values quite accurately. Washington D. C. can be considered an outlier since it has an abnormal GDP per capita since it is not a legal state. From the residuals plot we can infer that our assumptions are correct and valid under the Central Limit Theorem as our sample is large enough to converge into normal distribution. Although, our model residuals are almost normally distributed as the histogram shows.

4.3 *Computing Fitted Values and Forecasting US Elections*



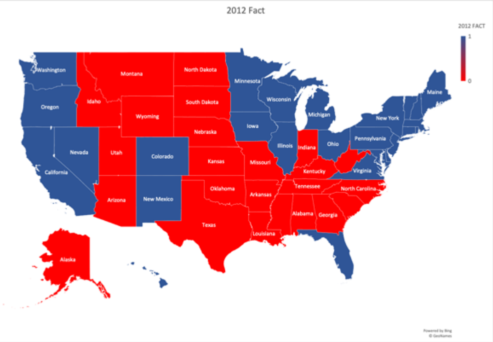
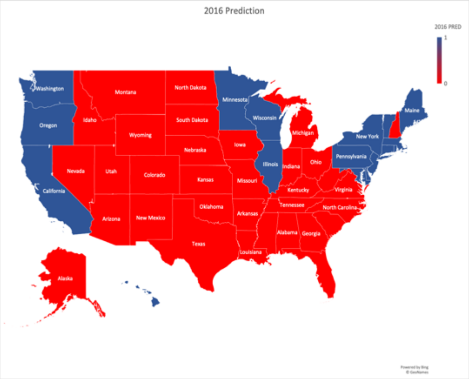


Figure 8. Prediction for 2012 - 271 Votes

Figure 9. Actual Votes 2012 – 332



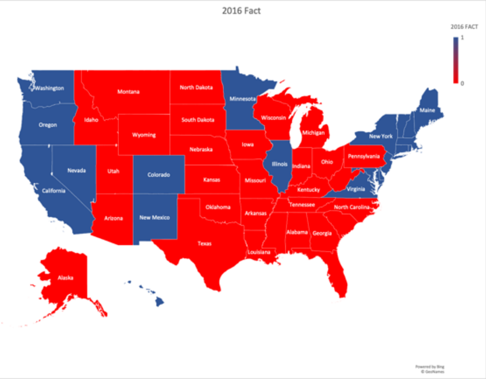


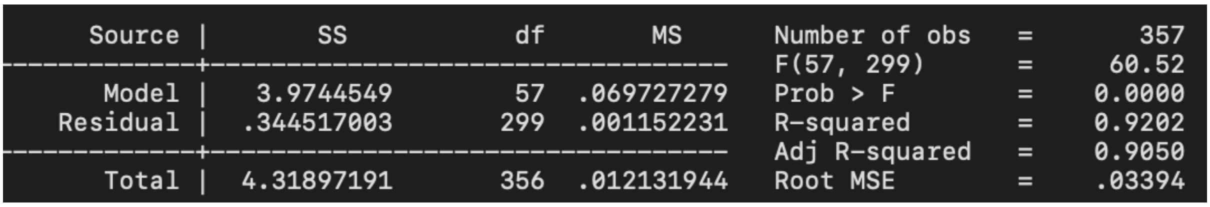
Figure 10. Prediction for 2016 – 226

Figure 11. Actual Votes 2016 - 227

Our model performed well in forecasting the in-sample values for Vp although it still predicted for 2012 the winning of the Democratic Party, the actual votes deviates a lot (even if in the confidence interval) while for 2016, the model predicted almost 100% that the Democratic Party would lose (226 votes against 227). However, if we take a closer look at each state, we can see that there is still much difference.

4.4 *Naïve Post-Selection Lasso Model*

The OLS regression has some problems concerning reliability of prediction power. In other words, it has not very good out-of-sample prediction, it doesn't deal well with strong dependence among variables, and it doesn't handle noise in the independent variables. Also, it might magnify the omitted variables bias. This is confirmed also by the results plots above. After generating squared values for *P, UG,* and *Gender_M*, we performed the Lasso one time (Post-Selection) using the Adaptive selection criteria. Regressing Vp on the selected variables by the lasso gave the following residuals results represented in Figures 12 and 13 below.



```
      Source |       SS           df       MS      Number of obs   =       357
-------------+----------------------------------   F(57, 299)      =     60.52
       Model |   3.9744549        57  .069727279   Prob > F        =    0.0000
    Residual |  .344517003       299  .001152231   R-squared       =    0.9202
-------------+----------------------------------   Adj R-squared   =    0.9050
       Total |  4.31897191       356  .012131944   Root MSE        =    .03394
```

Figure 12. Lasso Regression Results

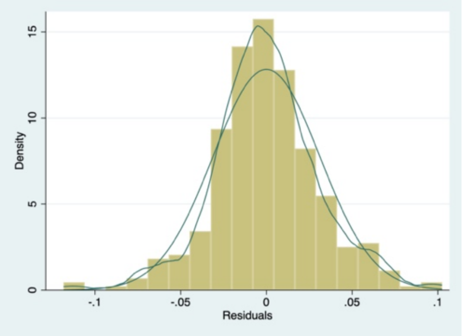


Figure 13. Lasso - Residual Scatter

4.5 *Double-Selection Lasso Model*

From Figures 12 and 13, we can see that the residuals are more normally distributed. However, the Naïve Post-Selection Lasso has some problems at its roots. First of all, as we work with real data, there is the likely possibility that the model selection makes mistakes about small coefficients of the estimators as they are set to 0 to improve the fit. In addition, the actual distribution of α is not concentrated. In the end, although the model has more predicting power than the fixed-effects OLS as the R-squared is greater, the model is not realistically usable to make inferences. To remedy, we should use the Double-Selection Lasso Model. This model selection approach consists in selecting to subsets based on the dependent variable, Vp, and the variable of interest, g_i. Then regress the two variables on the respective selected variables from the lasso. In this way, we may resolve post-estimation problems (uniformly consistent estimators) and thus, the inferences we make would be more reliable (also because the noise of the variables would be reduced even more). The residuals of the Double Lasso are presented in Figure 14 below.

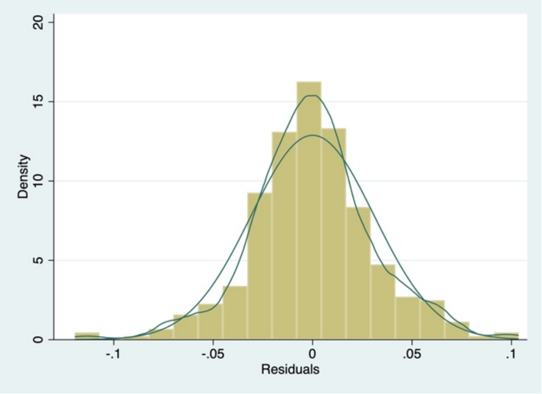
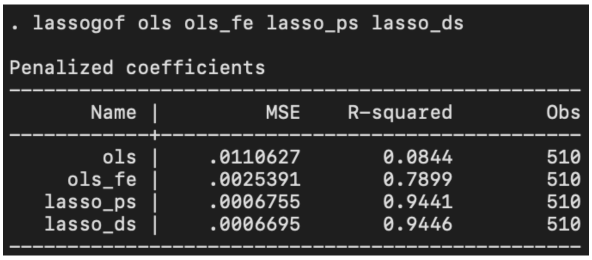


Figure 14. DS Lasso - Reiduals Scatter

Additional model comparison between the ones presented in this paper is provided in Figure 15 as it's used to decide which one to implement.



```
. lassogof ols ols_fe lasso_ps lasso_ds

Penalized coefficients
-------------------------------------------------
        Name |         MSE    R-squared        Obs
-------------+-----------------------------------
         ols |    .0110627       0.0844        510
      ols_fe |    .0025391       0.7899        510
    lasso_ps |    .0006755       0.9441        510
    lasso_ds |    .0006695       0.9446        510
-------------------------------------------------
```

Figure 15. Model Comparison

5. Conclusions

Of course, our model is very limited and there are empirical challenges that need to be addressed. First, the limitations concerning the data are multiple. For example, the data is just not available because it wasn't collected, and this is a problem also regarding the merge of the datasets as they start from different years. Ideally, we should have simulated the missing data through other techniques or restricted the considered period. Also, the criteria to estimate Vp implies that we would know in advance the economic output of the country for the year of the election. This can be redefined finding a relevant interval to predict the elections with useful anticipation. Another important point is the splitting of the data. For a much more reliable model selection, we should have split the dataset into two samples, training and validation. Unfortunately, the splitting of a panel data should be made with the use of a cut-off value which is significant for the data we are analyzing. In our case, this value could have been the 2008 economic crisis. But again, this would have required a separate more detailed study. At this point, we could have applied the difference-in-difference statistical technique which would have allowed us to compare a treatment group with the control group. We could have computed the effects of the independent variable on the dependent variable but without the time effects. The other main factor is the variables we considered as we might want to investigate more (also with the help of other studies) what influences elections both from an economic and a social-psychological perspective. This is because we can't actually know all the variables useful to forecast the elections, but we can select more variables from different reasonings. For example, if we think about the political color of the parties, red for democratic and blue for republicans, they are used to emphasize and spread the believes of the party (Adams, 2007). In this sense, the blue color is most associated with hope and liberal thoughts while red with radicals, socialism and traditional way of development. In terms of public polices, we can observe the differences between Obama's and Trump's public policies effects during their terms. Obama's implemented policies centered to benefit all people of US, like with the Obamacare, while Trump acted merely in the "interests" of Americans, for example with more conservative taxation on imports and domestic products. The last main problem is disentangling causation from correlation. In fact, correlation doesn't imply causation as one event can occur in proximity of another but without being caused by the latest, but instead the two are caused by an underlying and unobserved third event. In other words, the pillar of our analysis is that GDP (and economic variables correlated with it) is highly correlated with the winning party. At this point, we must ask ourselves whether for example the economic variables are generated by the effect of economic policies implemented by one party rather than another one, or whether the winning party is elected in order to implement policies in accordance with the current economic situation. To answer these questions, we should also answer the question we introduced in the first pages of the paper: Are public policies really chosen and caused by the voters persuaded by the economic circumstances in that precise period? Which cannot be answered in this paper. Another empirical problem is to measure the impact of such policies or better, the impact of such policies in the 4 years period after the implementation (assuming they are implemented in the first term year). This would help us understand if a party policy has a real effect or its impact is dominated by a previous policy (as well as contrasted).

In conclusion, the model presented in this paper should be enriched with additional variables extracted from several analyses of different kinds covering multiple socio-economic aspects. Furthermore, the time period criteria used in building the variables should be redefined to make it more feasible to implement in real world. And lastly, it should be re-thought using split datasets to measure its out-of-sample prediction and inferential power.

References

- Fair Original Paper and Model: https://fairmodel.econ.yale.edu/vote2020/index2.htm
- Popular vote by US State: https://doi.org/10.7910/DVN/42MVDX
- US Annual GDP by State in current and in chained dollars: https://apps.bea.gov/itable/iTable.cfm?ReqID=70&step=1#reqid=70&step=1&isuri=1
- US Population by State can be sourced from State Intercensal Datasets: https://www.census.gov/data/datasets.html
- Updated Fair Model: https://fairmodel.econ.yale.edu/RAYFAIR/PDF/2018B.pdf