Empirical models to forecast U.S. Presidential Elections

Andrea Pio Cutrera (965591)

Fellows: Margherita Maroni, Roberto Staino, Luca Donghi

Abstract

U.S. Presidential Elections are one of the most important occurrences in the world. We started from the model for U.S. popular vote forecast estimated by Fair (2009) trying to enrich his analysis with state level fixed effects in order to capture heterogeneity across States. In this attempt we wanted to discover how this heterogeneity affect the variables under study and see whether these variables are still significative or not. In another step we tried to use a machine learning algorithm (Lasso) to select variables that are important to include in a regression model in order to make forecasts. Comparing the forecasts made with these two models, the ones that turns out to be the best are the ones made with Fair's extended model. At least we are going to discuss how quantitative models of analysis like simple OLS estimates are not sufficient to attest causal link relationship between variables, and how we can use techniques to discover causation rather than just correlation.

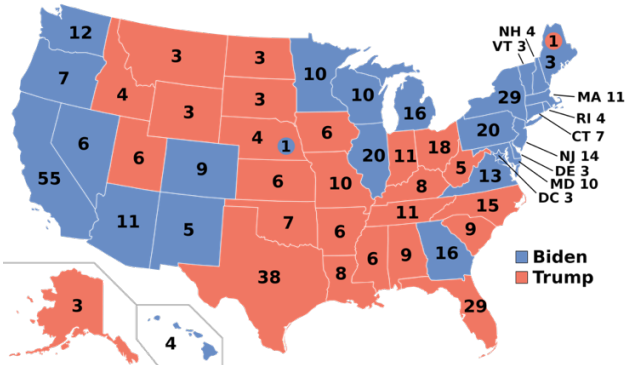
1. Introduction

United States of America Presidential Elections are big events that affect the sort of not only the people living in USA but also the entirety of the world; they have been studied by many researchers, and one of the most famous is Ray C. Fair. Professor Ray C. Fair in (May) 1978 published a paper for "*Review of Economics and Statistics*" which title was "*The Effects of Economic Events on Votes for President*". In that paper Fair tried to find out which were the main political, social and economic events affecting the vote share for one party election, and in this effort of understanding which were (and still are) the main causes of gain or loss of votes for one party, came out with the *Vote-Share Equations*.

One of the last updates has been published on the American Journal of Political Science, with the title "Presidential and Congressional Vote-Share Equations"; Fair starts from the theoretical assumptions that vote share for the 2-party vote is affected mainly by 2 sets of variables: economic conditions and incumbency of a party. The former are composed by short and long term indicators of economic performance (e.g. growth rate of real per capita GDP, absolute value of inflation rate in the first 15th quarters of administration, number of quarters in first 15th of administration in which there was a GDP growth exceeding 3.2%), the latter instead are variables that take into account the presence of an incumbent party by time and space dimensions (e.g. duration of incumbency, incumbency in the White House and in the Congress).

Starting from this empirical model developed by Fair (2009) we decided to pass from the National level to the lower State level and see whether it could be useful to take into account any state fixed effect or not. Nonetheless economy of USA is not homogeneous in all the states and what we want to check is whether the variables and the indicators found as significative by Fair still work at state level. Then in another section I'm going to present briefly how we thought to other plausible variables that can be useful to predict the vote share of 2-party vote.

Then the last passage we make is the translation of popular vote into electoral vote by putting the filter of the Great Electors since the winner is at least decreed by the candidate which takes the absolute majority in electoral vote.



2020 Presidential Elections – Great Electors gained by state

2. Data and methods

A strongly balanced panel data has been constructed by collecting data for all the tuples year-state as observations. Years are spanning a period from 1980 to 2020 (t = 11, there have been 11 elections in that period since they are taken every 4 years) for all the states (s = 50 + District of Columbia - that has its own Great Electors). So, our sample is composed by 561 observations (t * s = 11*51 = 561).

The Dependent variable called **share_d** has been computed by the percentage of votes (popular vote) for Democratic party out of the 2-party vote (Democratic and Republican).
It has been taken from MIT Election Data and Science Lab, 2017, "U.S. President 1976– 2016", link, Harvard Dataverse, and below the formula:

$$share\_d = \frac{(absolute\ number\ of\ votes\ for\ democratic\ party)}{(absolute\ n.of\ votes\ for\ democratic\ party + absolute\ n.of\ votes\ for\ rep)} * 100$$

The other variables can be grouped into sets of independent variables: namely *Economic*, *Political Incumbency, Social-Demographic and Composed* variables.

Economic variables from which we can measure long and short-term economic performance. They have been taken from Bureau of Economic Analysis link:

- **gdp**: is the nominal GDP for the year of reference in current dollars;
- **real_gdp**: it is the real GDP in chained dollars (2012) for the years of the elections by state;
- **real_pc_gdp**: it is the real per capita GDP computed as the real_gdp divided by the absolute number of inhabitants in that country (in that year);
- **deflator**: defined as the ratio between the nominal and the real GDP;
- **growth**: for this variable we followed the approach of Fair translating it in an annual base. We computed the rate of change of real GDP between the first 3 years out of 4 of the mandate (for example the value of growth for 2020 has been computed as the rate of change between the value of real_gdp of 2017 and 2019);
- **good_news**: it has been computed as Fair did but in a yearly base. It's a counter of how many years in the 3 out of the 4 years of mandate, there is a real gdp growth greater than 2.9% (imposed as a threshold);

Political Incumbency variables from which we try to measure how incumbency of a party in House, Congress and in Presidency strong is:

- **new_midterm:** It has been computed as the dependent variable of popular vote and it represents the share of votes for democrats in the midterm election. Midterm elections take place two year before the presidential election for the House of Representatives, and in our case, it represents two party share of votes between Democratic Party and Republican Party only. The share of democrats include votes for the Democratic party and his affiliates state parties (North Dakota Democratic–Nonpartisan League Party and the Minnesota Democratic–Farmer–Labor Party just because even though the different name they are indeed part of the same Democratic party). The midterm election data are taken from 1978 (meaningful for 1980) to 2018 (meaningful for 2020).
  In the House Election the vote is by district and for this reason the votes of all the districts of a state are added up. In this way it is possible to compute the votes by state. Finally, the last step is to divide the total number of votes for democrats in a state by the sum of votes for democrats and for republicans in the same state.
  If the democrats are unopposed in all the districts of a state, then the midterm variable takes value 100% for that observation. On the contrary, when they don't participate in any district of a state the midterm variable takes value 0%; It has been taken from MIT Election LAB;
- **incumbent**: as Fair did, it takes value 1 if Democrat is in white House at time of elections or -1 if Republican is in White House; (Fair, 2009)
- **dper**: as Fair did, it takes value 1 if a Democratic incumbent running again, -1 if a Republican incumbent running again, 0 otherwise; (Fair, 2009)
- **dur**: as Fair did, it takes values 0 if incumbent party has been in charge for only 1 or 2 consecutive terms, 1 if Democratic party has been in charge for 3 consecutive terms, 1.25 if for 4 consecutive terms, 1.5 if for 5 consecutive terms; the same values with the opposite sign for the Republican party; (Fair, 2009)
- **governor_dem**: it is a variable that check the political color of the governor in the election year for each state; it has been coded as 1 if governor in charge in that state was democrat; -1 if it was republican; 0 otherwise (just some cases of independent parties); It has been taken from National Governors Association;
- **Great_El**: it is the number of great electors for each state; it has been taken from National Archives of electoral college.

Social-Demographic variables from which we measure some characteristics of citizens; we decided to take into account educational level of citizens, unemployment rate, age and density of population:

- **population**: It's the population of each state from 1980 to 2019 (for 2020 we used the proxy value of 2019); it has been taken from Census Bureau;
- **density**: It's the density of population as a ratio between inhabitants and squared kilometers of land; it could be a good proxy to measure level of urbanization; Census Bureau;
- **old_voters**: it's the number, in thousands, of over 65 years old voters (which voted at the ballot) of the previous presidential election; These data have been collected from the Census Reports of each election from 1976 to 2016.
- **perc_over**: it is the percentage of votes out of all the votes coming from over 65 years old citizens;
- **highlev_educ**: it's the percentage of people who have at least a bachelor's degree;
  the data have been collected from the USDA (United States Department of Agriculture Economic Research Service) from 1980 to 2020. The data regarding the education are collected every 10 years; Educational level of citizens is one of the features that distinguish individuals in a community that are able to discern between Politics (what really matter for the country) from populism. With an increase in high level education in time, we expect an increase in volatility in popular vote since voters are more and more in the center of the political competition with less people in the far wings of parties. So more uncertainty brough by more educated citizens.

- **unemployment**: average annual unemployment rates by state; it from Us Bureau of Labor Statistics. It's a very important indicator since labor is one of the main concern of citizens and also one of the most important promise that candidate president make.

Composed variables in which we take into account the interaction between incumbent and the economic variables. It makes all the economic variables reasonable because for example we can expect a good economic performance affects positively the share of democratic vote when the incumbent is a democrat, and negatively (more votes for Republican) if the incumbent is a Republican (i.e. **growthinc** = growth*incumbent; **deflatorinc** = deflator*incumbent; **g_ninc** = good_news*incumbent; **gdp_i** = gdp*incumbent; and the same for real and per capita GDP – **real_gdp_i** and **real_pc_gdp_i**). So, from this moment on when I will talk about economic variables I'm referring to the ones multiplied by the incumbent since they are the ones that make sense.

3. Empirical strategy

We are going to start from the estimation of a fixed effects model that contains the same variables used by Fair (2009). We will concentrate our analysis on the last three years of elections that are 2012, 2016 and 2020 and we are going to estimate the coefficients with a simple OLS method of regression in order to detect the effect of independent variables on the popular vote for democratic party.
Then, for these three models of regression (2012, 2016 and 2020) we are going to perform a test for the joint significance of economic variables to see whether the variables should be included in our regression model. Thereafter, we will do the same for the state level fixed effects.
Using these models of regression, we can make predictions about the popular vote for each state and then we will translate this popular vote into the electoral vote putting the filter of the great electors. After that, we will compare our results with the real results of the last three years of Presidential elections. In particular, we will firstly compare the MSE of the popular vote (the square of the difference between real values and fitted values) to see how the models are performing in forecasting; and then, we will observe how the electoral vote we computed is different from the real one.
The last effort will be put in the estimation of the same regression enriched with additional variables (described above) that we thought they would be useful to predict the share of democratic vote. In particular, we took into consideration the following variables:

- the presence of a democratic or republican Governor in each state, expected to influence the vote for the same party of him/herself;
- the midterm elections that are very relevant for the eventual re-election of an incumbent President, expected to be a good benchmark for an evaluation of the job done by the President in charge; a good score in that election could be a sort of confirmation for the next Presidential elections;
- the density that suggests the level of urbanization of the states, could be a good indicator to distinguish country from city life, that are very different and have different needs and priorities;
- the age of the voters because older voters are less prone to change, and they are fond of one party, it is very difficult to make them change idea;
- the level of education that it is a good indicator of the awareness of people towards the distinction between politics and populism;
- the unemployment because having a job is one the main concern of common citizens and it could be useful to predict the party they will support.

With these additional variables we will apply the Lasso technique to make the machine select the variables that are really useful to predict the outcome of elections. When we will apply the Lasso we will always keep real per capita GDP (real_pc_gdp_i). In particular, we are going to use a double-post-model selection approach in order to obtain a consistent estimate of the parameters. Instead, if we would have used the simple naïve-post-model selection we could have obtained inconsistent estimates (incurring maybe in an endogeneity problem).

First, we created a macro that stores all the states fixed effects calling it X_effects.
Secondly, we generate other two new macros with all our variables from which we wanted the Lasso to select:

- *X_lasso* = it includes all the variables except our dependent one and real gdp per capita (real_pc_gdp_i);
- *X_lasso_pc* = it is a subset of the previous macro in which we exclude the economic variables (just because we are going to use this macro for real per capita gdp and these economic variables are by construction dependent on it).

At this point we will apply the Lasso to our dependent variable and the macro *X_lasso*; then we will do the same for real per capita GDP and the macro *X_lasso_pc*. The two subset of variables we obtain from this two-stage selection Lasso are the variables that we are going to include in our regression model (always keeping state level fixed effects).
At least, we will make predictions with that model and compare the forecasts obtained with the real values regarding the three elections.

4. Empirical results

4.1 Regression model with Fair's variables and state fixed effects

In this section, we start discussing the results of the regression only with the variables used by Fair and the state fixed effects done for the three years under consideration.

```
. reg share_d incumbent growthinc deflatorinc g_ninc dper dur i.state if year < 2012
```

Source	SS	df	MS	Number of obs	=	408
				F(56, 351)	=	26.81
Model	37138.31	56	663.184108	Prob > F	=	0.0000
Residual	8681.53602	351	24.7337209	R-squared	=	0.8105
				Adj R-squared	=	0.7803
Total	45819.8461	407	112.579474	Root MSE	=	4.9733

share_d	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
incumbent	-.9796427	1.763584	-0.56	0.579	-4.448163	2.488878
growthinc	.6928663	.0845816	8.19	0.000	.5265158	.8592168
deflatorinc	5.95662	2.030099	2.93	0.004	1.963933	9.949307
g_ninc	.0174783	.3199235	0.05	0.956	-.6117297	.6466864
dper	-3.645022	.7441099	-4.90	0.000	-5.108497	-2.181547
dur	-6.698553	.6847024	-9.78	0.000	-8.045189	-5.351918

For 2012 we found that the variables growth, deflator, dper and dur are the only ones with significative coefficients. Growth and deflator have positive coefficients as expected, while dper and dur have negative coefficients. The reasons for that signs could be explained by the fact that good economic performance can make stakeholders, companies and also citizens feel good about the way in which the mandate has been carried out by the President, instead the reason of their negative sign probably lies in the fact that people want to see a change in the party of the President.
Now we run a joint test of significance for the economic variables and we see that all of them should be taken into account for the regression. The null hypothesis that the three coefficients of economic variables (for growth, deflator, good news) are jointly zero is rejected.
Then, we do the poolability test for the state fixed effects and outcome suggests also that all of them should be taken into account since the null hypothesis is rejected again.

```
 ( 1)  growthinc = 0
 ( 2)  deflatorinc = 0
 ( 3)  g_ninc = 0

       F(  3,    351) =   32.24                F( 50,    351) =   26.26
            Prob > F =    0.0000                     Prob > F =    0.0000
```

For 2016 we reach the very same results in coefficients significance and sign.

```
. reg share_d incumbent growthinc deflatorinc g_ninc dper dur i.state if year < 2016
```

Source	SS	df	MS	Number of obs	=	459
				F(56, 402)	=	29.95
Model	42997.1749	56	767.806696	Prob > F	=	0.0000
Residual	10307.3958	402	25.640288	R-squared	=	0.8066
				Adj R-squared	=	0.7797
Total	53304.5707	458	116.385526	Root MSE	=	5.0636

share_d	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
incumbent	-1.106153	1.716779	-0.64	0.520	-4.481139	2.268833
growthinc	.5965291	.0815686	7.31	0.000	.4361747	.7568835
deflatorinc	6.081467	1.802034	3.37	0.001	2.53888	9.624054
g_ninc	.1904117	.3002862	0.63	0.526	-.3999159	.7807392
dper	-3.480426	.7535819	-4.62	0.000	-4.96188	-1.998973
dur	-6.744239	.6838094	-9.86	0.000	-8.088528	-5.39995

Here again joint test of significance on economic variables and the poolability test on state fixed effects suggest us to take into account all of them in the regression model.

```
( 1)  growthinc = 0
( 2)  deflatorinc = 0
( 3)  g_ninc = 0

      F(  3,    402) =   27.84                F( 50,    402) =   29.93
           Prob > F =    0.0000                     Prob > F =    0.0000
```

We repeated the OLS procedure again for 2020, reaching the same results.

```
. reg share_d incumbent growthinc deflatorinc g_ninc dper dur i.state if year < 2020
```

Source	SS	df	MS	Number of obs	=	510
				F(56, 453)	=	33.47
Model	49571.035	56	885.197054	Prob > F	=	0.0000
Residual	11979.5511	453	26.4449251	R-squared	=	0.8054
				Adj R-squared	=	0.7813
Total	61550.5861	509	120.924531	Root MSE	=	5.1425

share_d	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
incumbent	-2.246738	1.709056	-1.31	0.189	-5.605399	1.111923
growthinc	.5835099	.0784451	7.44	0.000	.4293484	.7376715
deflatorinc	7.448893	1.7247	4.32	0.000	4.059487	10.8383
g_ninc	.3540129	.2725116	1.30	0.195	-.1815308	.8895566
dper	-3.615119	.7262395	-4.98	0.000	-5.042335	-2.187902
dur	-6.726564	.6927627	-9.71	0.000	-8.087991	-5.365137

Joint test of significance for economic variables, and poolability test on state fixed effects:

```
( 1)  growthinc = 0
( 2)  deflatorinc = 0
( 3)  g_ninc = 0

       F(  3,    453) =    31.36          F( 50,    453) =    33.58
            Prob > F =     0.0000               Prob > F =     0.0000
```

At the end, we can say that the significative variables remain the same for all three models we estimated and that at least the coefficients of these variables have quite the same magnitude. Therefore, enlarging the sample to more recent years, keeping the same regressors, does not change the outcome of this first analysis.

Now we compute the fitted values for each of three models just estimated and we make a comparison of the means of the Squared Error (MSE). We computed the MSE as the mean of the squared difference between the real and the fitted values.

Variable	Obs	Mean	Std. Dev.	Min	Max
e_fair_12sq	51	36.66	53.46185	.0249783	227.5895
e_fair_16sq	51	37.30578	52.40706	.0737126	321.6577
e_fair_20sq	51	43.48491	65.03748	.0001657	410.7105

We can notice that for 2012 and 2016, the MSE presents quite similar values whereas for 2020 we notice an increase up to 43.5. We can maybe justify this increment by saying that the spread of pandemics brought a very high stream of uncertainty.

At this stage, we put the filter of great electors to the percentage of share of democrats to see which state has taken more than 50% of popular vote. Then, summing up all the great electors taken in each state we get the predictions of electoral votes gained by Democrats at the time of elections. Below a comparison between the predicted and the real number of great electors gained by Democrats:

	2012	**2016**	**2020**
Prediction with Fair	300	239	286
Real	332	227	306

4.2 Regressions with our additional variables selected by Lasso

For the sake of the analysis, we decided to restrict the access of the Lasso to a subset of our dataset to observations of the years that precede our predicted year (e.g. Lasso can use observations from 1 to 408 for 2012).

In the first step, we make the Lasso for the dependent variable share_d and the machine select the real GDP growth, the squared deflator and the squared high level education from the macro *X_lasso*. Then, we make again the lasso for the real per capita GDP variable and the machine selects only the state fixed effects from the macro *X_lasso_pc*.

With the results obtained we run the regression between our dependent variable and the two subsets of variables selected by Lasso.

```
. reg share_d real_pc_gdp_i growthinc deflatorinc2 highlev_educ2 i.state if year < 2012
```

Source	SS	df	MS	Number of obs	=	408
				F(54, 353)	=	40.21
Model	39412.3178	54	729.857737	Prob > F	=	0.0000
Residual	6407.52824	353	18.1516381	R-squared	=	0.8602
				Adj R-squared	=	0.8388
Total	45819.8461	407	112.579474	Root MSE	=	4.2605

share_d	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
real_pc_gdp_i	.0001033	.0005621	0.18	0.854	-.0010021	.0012087
growthinc	.2638371	.0490723	5.38	0.000	.1673264	.3603479
deflatorinc2	16.33839	1.166533	14.01	0.000	14.04417	18.63262
highlev_educ2	.0062221	.0015713	3.96	0.000	.0031319	.0093123

From the output above, we can say that growth, deflator squared, and high-level education squared are significative and they have all positive coefficients. Among the three, the deflator indicator has a very large coefficient equal to 16.3.

Continuing our analysis on a larger subset of our original dataset with observations spanning from 1 to 459 that are the ones before year 2016. We make again the Lasso double selection and running the regression with the selected variables (i.e. dur, growth, deflator squared, dper). The output is the following:

```
. reg share_d real_pc_gdp_i dur growthinc deflatorinc2 dper i.state if year < 2016
```

Source	SS	df	MS	Number of obs	=	459
				F(55, 403)	=	39.26
Model	44920.2009	55	816.730926	Prob > F	=	0.0000
Residual	8384.36979	403	20.8048878	R-squared	=	0.8427
				Adj R-squared	=	0.8212
Total	53304.5707	458	116.385526	Root MSE	=	4.5612

share_d	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
real_pc_gdp_i	.0005212	.0006015	0.87	0.387	-.0006613	.0017037
dur	-1.787775	.3923094	-4.56	0.000	-2.559004	-1.016547
growthinc	.437525	.0679361	6.44	0.000	.3039715	.5710785
deflatorinc2	13.1011	1.133818	11.55	0.000	10.87216	15.33004
dper	-.3990912	.3323984	-1.20	0.231	-1.052543	.2543601

The significative variables are dur, and again growth and deflator squared since their p-values are below the threshold at 5% significance level.
Both growth and deflator have positive coefficients and the latter has a very high magnitude on share_d. Instead, the duration of the presence of the incumbent party seems to have a negative effect on the popular vote for Democrats as expected.

For 2020 we are interested in the observations spanning from 1 to 511 because we want the lasso selects observations until 2016. Running the regression, we reach these results:

```
. reg share_d real_pc_gdp_i dur growthinc deflatorinc2 dper i.state if year < 2020
```

Source	SS	df	MS	Number of obs	=	510
				F(55, 454)	=	40.77
Model	51187.9787	55	930.690522	Prob > F	=	0.0000
Residual	10362.6074	454	22.8251264	R-squared	=	0.8316
				Adj R-squared	=	0.8112
Total	61550.5861	509	120.924531	Root MSE	=	4.7776

share_d	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
real_pc_gdp_i	.0006401	.0006258	1.02	0.307	-.0005898	.00187
dur	-2.783142	.319719	-8.70	0.000	-3.411455	-2.15483
growthinc	.5067635	.0673141	7.53	0.000	.3744777	.6390494
deflatorinc2	10.3156	.9292359	11.10	0.000	8.489464	12.14174
dper	-.3095573	.3434371	-0.90	0.368	-.984481	.3653664

Here we got similar results of those of 2016. The only significative variables are dur (with negative coefficient), deflator squared (with high positive coefficient) and growth (with positive coefficient).

After this analysis, we can compute the fitted values for the three models described above. From the fitted values we calculated the squared error, and we compare their means in order to see which one is performing better in forecasting popular vote.

In the following picture you will find a comparison between the forecasts made with Fair variables and state fixed effect, and the forecast predicted with our additional variables. The prediction made with variables selected by Lasso procedure seem not to have improved the results got by Fair's extended models. The means of the squared errors are very high, and from the Great electors we are going to forecast we will see how these estimates are not reliable at all.

```
. summarize e_fair_12sq e_fair_16sq e_fair_20sq e_l_12sq e_l_16sq e_l_20sq
```

Variable	Obs	Mean	Std. Dev.	Min	Max
e_fair_12sq	51	36.66	53.46185	.0249783	227.5895
e_fair_16sq	51	37.30578	52.40706	.0737126	321.6577
e_fair_20sq	51	43.48491	65.03748	.0001657	410.7105
e_l_12sq	51	85.44543	105.3397	.0023596	564.6041
e_l_16sq	51	52.1443	86.768	.0030115	511.2197
e_l_20sq	51	117.576	132.1982	.0074868	694.6714

Putting the filter of great electors in order to translate the popular vote into the electoral vote, we will select the states where the Democratic Party obtain the absolute majority of popular vote and we will assign all the great electors of that state to Democrats. Then, we will sum all them up to obtain the number of great electors gained by the Democratic President who is running for the White House.

Our results are summarized in the following table that contains the comparison between the values forecasted by us and real great electors gained by Democrats.

	2012	**2016**	**2020**
Prediction with Lasso	504	395	501
Prediction with Fair	300	239	286
Real	332	227	306

5. Conclusions and description of the results

In the first attempt in which we estimated a model using Fair's variables and state level fixed effects, the forecasts are quite accurate because the mean squared errors are not large and not very different from the realized. If we take a look also at the analysis Fair did in his paper, we can find similar results, therefore we were good in resembling his methodology. We attempted to improve a bit his methodology because we were able to take into account the heterogeneity across all the 50 different states of U.S. (plus the District of Columbia). This is the reason why our output is not identical to the one of Fair because he used aggregated data whereas we used state level data. The number of great electors we computed using the fitted values is not far away from the real number of electors gained by Democrats. Even though there is a small difference between the two values, they predict well election of the candidate for the three years under consideration (namely, 2012, 2016 and 2020).

When we add our social and demographic variables the results obviously changed. Our purpose was to take into account other aspects that matter in everyday life of the voters. In particular we posed our interest on the education of voters, on their age, on the level of unemployment by state. Moreover, we were interested in other political elections that usually take place in U.S., such as the Administrative elections of each State and the midterm elections that take place two years before the Presidential ones. Unfortunately, the only in 2012 Lasso selects the high-level education (squared, maybe because there is a nonlinear relation) that probably confirm our expectation. We thought that an increase in the level of education could have been a sufficiently strong to bring political battle more towards the center, and especially the Democratic party has been able to capture that votes.
In 2012, the lasso double selection mechanism selects also growth and deflator squared; for 2016 and 2020 it selects these two variables and also two of the variables suggested by Fair, namely dper and dur, always with negative coefficients. When we run the regressions, the only variables that are significative are growth and deflator and dur, except for 2012 when the dur variable is not even selected.
When we compute the MSE of our forecasts we got higher results than the ones obtained using Fair's variables. This obviously means that our new models are less precise in forecasting and they are not very much reliable. Observing the number of great electors estimated with our new model, we can say that we always overestimate the number of electoral voters gained by Democrats. In addition, although for 2012 and 2020 we predict the win of the Democratic Party; for 2016 we erroneously forecast the election of the Democratic candidate, while the winner was indeed a Republican (Donald Trump).
In conclusion, our model could be used only to predict if the Democratic party will win but it is not good enough to forecast the win of the Republicans. Therefore, this type of model is not reliable at all. We would maybe improve it adding crucial factors that we are probably missing in designing our empirical research.
The last remark should be done on the 2020 forecast that has been obviously shaped by the uncertainty brought by the spread of the pandemic that is still ongoing.

6. Causal inference for our models

In the previous sections we have estimated some models with simple OLS method of regression trying to find out which where the causes leading popular vote in the last three U.S. Presidential elections. According to the results we obtained there are some variables that are quite significative in explaining the votes gained by Democrats during elections, but we do not really know anything about the real causal link that connects each independent variable with votes given to Democrats (our dependent variable); indeed, the only thing we know is that there exist a statistical correlation among the variables that turns out to be significative, but it is not a sufficient condition for *causation* and this is why we are going to deeply investigate the real relation between variables.
At this point, in order to make causal inference on agents that really affect our variable of interest, we could use some techniques that have been designed for this purpose.
Among the techniques we have studied, *Difference-in-Difference* estimation seems to be the one that better fits our models.
Difference in Difference is based on the assumption of common trend in time and it is developed in two dimensions:

1. before and after comparison: it compares the same individual or the same community before and after the program.
2. participant and non-participant comparison: it compares participant in the program with those individuals who do not take part in.

Therefore, the estimator comes from a combination of these two dimensions and it looks like this:

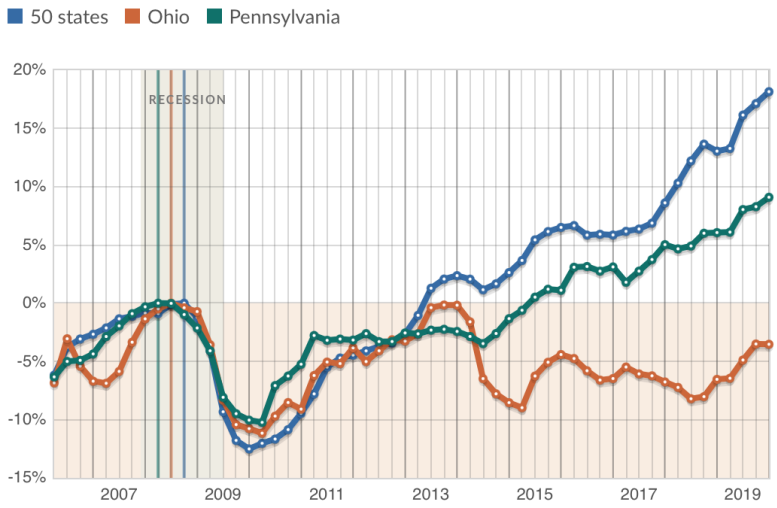
$$DD = \left[(participant\ post) - \overline{Y}(participant\ pre)\right] - [\overline{Y}(comparison\ post) - \overline{Y}(comparison\ pre)]$$

To use this type of estimation in our model, we should pose our attention towards the public policies designed by the Presidents during their mandates. Of course, the public policies are very different depending on the political color of the winner. We then decided to take into consideration the Presidential election of 2016 when the Republican President Donald Trump was elected after eight years of Democratic incumbency. So, we can consider the 2016 as the turning point for public policies in the last ten years of U.S. history. We choose two very similar States of U.S. in order to verify *the difference in the difference* in the implementation of public policies.
The two states chosen, Pennsylvania and Ohio, have a quite similar number of inhabitants and they also have similar population density. Moreover, they are neighboring in the North-West part of America and their wealth per capita is quite comparable.
Below a synthesis of the main factors characterizing the two states:

year	states	population	real_pc_gdp	governor_dem	highlev_educ	land	density	unemployment
2016	Ohio	11634370	50,2858943	-1	27,8	116096	100,213358	5
2016	Pennsylvania	12782275	54,0839717	1	30,8	119283	107,159235	5,4

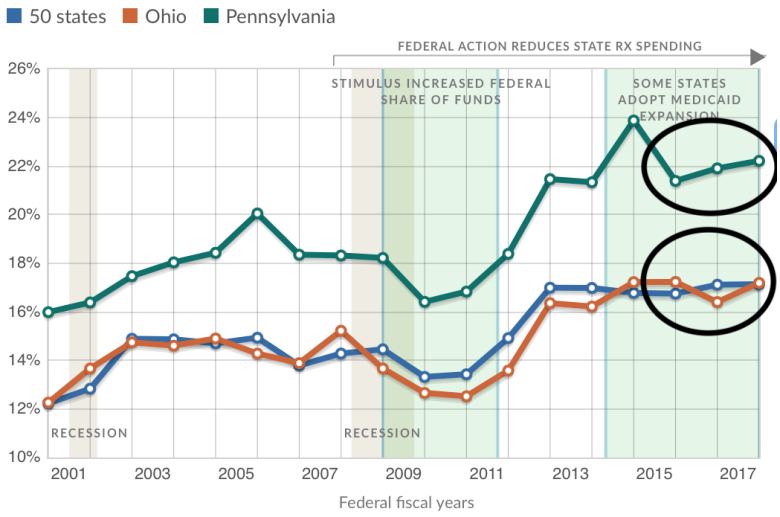
As we can see, the two states seem to have the same economic and social background (i.e., similar level of unemployment, education, wealth and urbanization).
The only thing that really distinguish Pennsylvania and Ohio is the political color of their governors. While Ohio is governed by a Republican, Pennsylvania is run by a Democrat. This important feature regarding their administration is what really distinguish their *participation* from their *non-participation* to the public policies designed by the President in charge in 2016. What we mean by *participation* is more a figurative way to say that a State is participating in the political life of the winner President. What we expect is a stronger implementation of the public policies designed by Trump in Ohio, whereas we expect a softer implementation or even an aversion towards the presidential policies from the Pennsylvania's governor.

A good example to begin with could be the economic policies enacted by Trump's administration from 2016. We know that the main aim of President Trump was to boost economic growth with trade protectionism and tax cuts. His slogan was to "Make America Great Again", therefore he promised that he would have taken care of Americans first and he would have dealt with immigration issue. About this, we searched some data regarding the fiscal pressure that was enacted in Ohio and Pennsylvania from 2016 on. For fiscal pressure we found an indicator of the tax revenue; on one hand, Pennsylvania experienced a small decrease after the elections but then it continues with its strong increasing trend that started after the period of recession (2008); on the other hand, Ohio experienced a more consistent decrease of its tax revenues after the election of Donald Trump and then it continued with a constant trend in the fiscal pressure, with a slight further increase, still remaining well below the level attained in recession period.
In the following picture we show you how the two states behaved differently. Pennsylvania follows the median tax revenue of the others 50 states, whereas Ohio is far below the median; just like Trump wanted. From the graph we can also see that the common trend assumption is also respected until more or less 2014.



Change in Tax Revenue from each State's peak quarters. Adjusted for inflation.

Bringing the attention to another issue, in particular the spending in medical health spending.
One of the most important goal of Trump's mandate was to eliminate the so called "Obamacare" (i.e., The Patient Protection and Affordable Care Act, signed by President Barack Obama in 2010). This policy was designed to increase government expenditures in health care after the last reform of the Healthcare system way back in 1965, when U.S. experienced the passage from Medicare to Medicaid. In fact, during the Trump's government, the federal spending on Medicaid grew by 1.6% and it was the smallest annual increase from 2012.
Then, what we suppose from the two states we are considering is a decrease in state spending in healthcare as a share of own source revenue for Ohio and maybe a constant trend for Pennsylvania. Our expectations are confirmed as we can see from the graph below; Ohio's governmental spending on healthcare is experiencing a decrease of some point in percent (as a % of spending coming from their own revenues – so even a point in % is a significative change) in 2016. To the contrary Pennsylvania is adverse to the cuts Trump wants to make, and increase their share of spending in healthcare. Even for this example the common trend assumption is respected.



State Medicaid spending as a share of own source revenue from 2000 to 2017

Thanks to these little examples we just show, we can highlight that a policy-wise point of view could be a good mix of both quantitative and qualitative procedures of analysis. On one hand, we need data to understand how much indicators change and vary together, how variables are correlated but, on the other hand, we should always keep in mind that a causal link must exist to confirm the meaningfulness of the relation.

We can conclude by saying that policies enacted by U.S. Presidents have different consequences on the States. When political color of the Governor is the same as the President's one there is a greater probability for a policy to be implemented firmly. To the contrary, a different color may even lead to an aversion to it. The difference-in-difference technique could be very useful in a setting like this: a big country with such big political, social and economic issues in which there are commonalities and differences, but just remember the common trend assumption.