Forecasts and Inferences of U.S. Presidential Elections

Matteo Biglioli - 938199 - matteo.biglioli@studenti.unimi.it

Gaspare Mattarella, Sara Gironi

Abstract

The aim of this project is to analyze the US presidential elections, focusing both on economic and incumbency variables. We use the research by Ray Fair[1] as a starting point and try to enrich his model with a panel data analysis. We then collect other explanatory variables and leverage LASSO to select the best ones to forecast the outcome of the presidential elections. We then compare the result obtained with the model proposed by LASSO to the one leveraging Fairs' study. In conlusion we discuss the empirical challenges of estimating the effect of the political colour of the winner on the public policies.

Data and methods

In order to enrich the model by Fair we collected a panel dataset, for the period 1976-2016, containing both economic and incumbency variables by state. Because our goal was to construct a complete, balanced and flexible dataset, we leveraged a Postgresql* database that allowed us both to easily match data from different datasources and fastly compute different aggregations of our variables.
Apart from the variables proposed by Fair himself, we collected a few more economic variables, as shown below:

Name	Description	Frequency	By state
p	Per capita personal income	Quarterly	By state
e	Employment	Annualy	By state
oil	Oil price	Quarterly	Nationwide
nasdaq	NASDAQ index	Quarterly	Nationwide
i_o	Import Export index	Quarterly	Nationwide
h	Housing price	Annualy	Nationwide

Even if we noticed that Fair explicitly work just with quarterly data, in order to remove data for the last quarter of administration (elections are held in November), we did not manage to obtain all the variables with quarterly frequency.
Leveraging the flexibility of SQL queries:

- For the variables with annual frequency we computed the percentual growth in both the 4 years and the last year of administration.
- For the variables with quarter frequency we computed the percentual growth in both the 15 quarters and the last 3 quarters of administration.
- For all variables we computed the number of years or quarters with values over the state mean during the presidency.

Obviously all the additional variables are multiplied by the incumbency variable at the first stage of our analysis, in order to give credit for the variation of each explanatory variable to the party in charge.

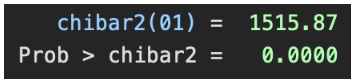
We also highlight that from a reverse engineering of the Fair equation we noticed that the value used to compute the Z_t explanatory variable (3.2%) seemed to be the mean of GDP growth in the U.S. for the study period. In order to compute an analogous variable, our Z_t is actually the number of year (in the president term) where the GDP growth was above the state mean computed in the 1967-2016 period.

Descriptive statistics and Empirical results

Breush-Pagan Test

The first step of our study is a test for heteroskedasticity, specifically a Breush-Pagan test on the residual. Results are shown below:

*https://www.postgresql.org/



```
    chibar2(01) =   1515.87
Prob > chibar2 =    0.0000
```

As expected we have to account for heteroskedasticity by leveraging a robust covariance estimator, that will ensure the reliability of the standard errors of our estimates. We also notice that, because we are working with a panel dataset, we need to use the clustered robust standard errors: this will allow to have a different variance for each state, but also take into account that observations whitin each state could be correlated.

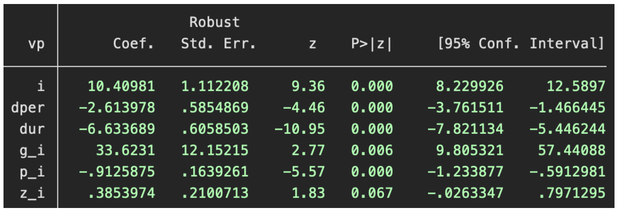
Fair model estimates

The first model we estimate is a state-level fixed effects model that includes the same explanaroty variables used by fair:

$$V_t = \beta_0 + \beta_1(G_t \times I_t) + \beta_2(P_t \times I_t) + \beta_3(Z_t \times I_t) + \beta_4 I_t + \beta_5\text{DPER} + \beta_6\text{DUR} + S + u_t$$

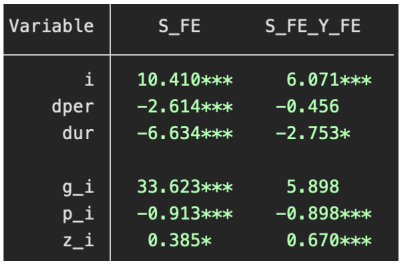
Where S is the component related to fixed effect at state-level.

Here we display the estimates of the regression:



```
             |               Robust
          vp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           i |   10.40981   1.112208     9.36   0.000     8.229926     12.5897
        dper |  -2.613978   .5854869    -4.46   0.000    -3.761511   -1.466445
         dur |  -6.633689   .6058503   -10.95   0.000    -7.821134   -5.446244
         g_i |    33.6231   12.15215     2.77   0.006     9.805321    57.44088
         p_i |  -.9125875   .1639261    -5.57   0.000    -1.233877   -.5912981
         z_i |   .3853974   .2100713     1.83   0.067    -.0263347    .7971295
```

Before interpreting the outcome, we explain why we did not used year-level fixed effects. As shown below, adding year-level fixed effects makes DPER, DUR and, most importantly, $(G_t \times I_t)$ statistically insgnificant.



```
Variable |     S_FE       S_FE_Y_FE
---------+-------------------------
       i |   10.410***     6.071***
    dper |   -2.614***    -0.456
     dur |   -6.634***    -2.753*
         |
     g_i |   33.623***     5.898
     p_i |   -0.913***    -0.898***
     z_i |    0.385*       0.670***
```

The reason for this could be the following: because GDP growth in the election year is an important explanatory variable, if not the most important, and because it is usually percieved equally across all the states[†] , it seems that the year fixed effect are mimicking the effect of this variable, making it statistically insignificant. We can then conclude that year-level fixed effect tend to incorporate the explanatory power of our variables, therefore, because we know that the variables used by Fair are actually statistically significant, we will exclude year-level fixed effect from our models now on.

From the regression estimates we notice that the highest coefficent is the one related to the GDP growth in the election year , followed by the one related to the incumbency variable. These two estimates tell us that an incumbent party has a 10.4% advantage on the state-level elections, that is to add to a 33.6% advantage for every percentage point of GDP growth in the election year. The last positive coefficent is related to $Z_t \times I$, which tells us that every year, during the presidency, that a state had a GDP growth over his mean, gives the incumbent party a .385% advantage.

We can then easily interpret the coefficent related to $P_t \times I$, which tells us that every percentage point of inflation gives a disadvantage to the party in charge of -0.9%. We finally notice that the lowest coefficent is the one related to DUR, that is the explanatory variable regarding the duration of the party in charge. As reported by Fair this coefficents models the desire for change in the political framework: as can be seen in a lot of other states, the population usually grows tired of having the same party in charge for a long period, for this reason being at the power for a lot of consecutive terms has actually a negative effect on the party share.

We kept the DPER explanatory variable at last because it seems to have strange behaviour: we would expect its coefficent to be positive, because it is related to the familiarity of the president in charge, but we got a strong negative value. This could be due to the fact that his effect is actually incorporated in the relationship between the two other incumbency variables discussed above. We know that, when the current party has been in charge for just one term, $DUR = 0$; in that case we would have that $\beta_4 I_t + \beta_6\text{DUR} = \beta_4 I_t$. We can then theorize that, because in the 1976-2016 sample almost all president were elected for two consecutive terms, the effect of a president running again is basically the same as having the incumbent party at the power for just one term. Bearing this in mind we can not actually then interpret the incumbency variables in the same way as Fair did, because in our case their effect seems to

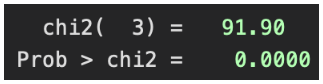
[†] Of course GDP vary for each state, but we will usually see all of the states GDPs growing and declining together, as the U.S. GDP grows and declines.

be redistribuited following a strange behaviour.

Tests for statistical significance

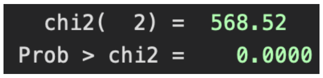
To actually check if the explanatory variables introduced in our model are statistically significant we can perform two test:

1. We test for the joint significance of the economic variables and notice, from the results below, that we can reject the null hypothesis $H_0 : \{(G_t \times I_t) = 0,\ \ (P_t \times I_t) = 0,\ \ (Z_t \times I_t) = 0\}$



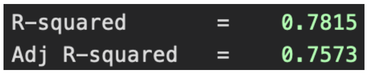
```
  chi2(  3) =   91.90
Prob > chi2 =    0.0000
```

2. We test for the joint significance of the state-level fixed effects (**poolability test**), and we can reject the null hypothesis as well:



```
  chi2(  2) =  568.52
Prob > chi2 =    0.0000
```

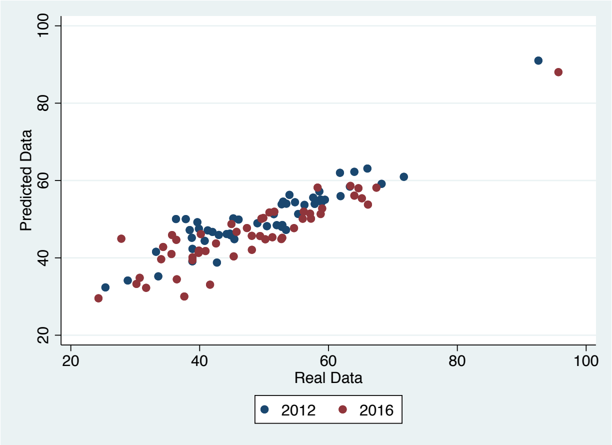
Given the two tests above we can conclude that the model we estimated is built from statistically significant variables. This conclusion can be further confirmed by the Adjusted-R^2 statistical measure, that represents the proportion of the variance for a dependent variable that is explained by the explanatory variables, and is displayed below:



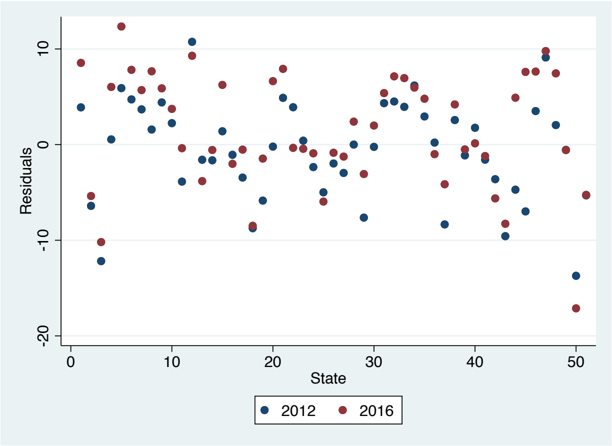
```
R-squared       =    0.7815
Adj R-squared   =    0.7573
```

Predictions

We can now leverage the model we built to compute fitted values. Before computing the actual prediction about electoral votes, we can compare the forecasts with the real data of 2012 and 2016, as shown in the following graph:



Even if from the graph above we can see that our model seems to behave as expected, we can't actually grasp the full extent of the errors due to the fact that we can't identify the difference between real and predicted share for each state. For this reason we display the residuals of our predictions in order to better understand the behaviour of the model:



From the graph above we can notice that our residuals are pretty high, but we expected this because we are using a model thought for the whole country at a state level. We can also highlight that the residuals seems to be simmetrically scattered around zero, as it should be. We ca even compute the MSE for 2012 and 2016:

Year	MSE
2012	27.084
2016	37.578

Finally we can assign every state to the candidate with the majority of votes and sum them up to forecast the actual elections. Below we display the forecasted electoral votes given to the democratic party in the 2012 and 2016 elections:

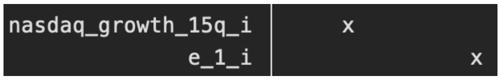
Year	Real	Forecast	Error
2012	332	292	40
2016	227	242	-15

From this table we can see that we would have actually rightfully predicted the outcome of both election. We also highlight that in this case we are cheating, in fact we used as training dataset to compute our estimates the whole period 1976-2016 (including 2012 and 2016), which is the reason why our predicions are close to the actual outcomes.

Lasso

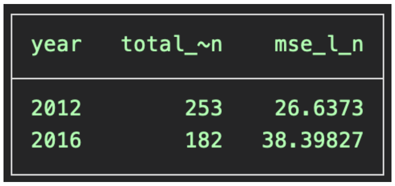
To further reduce our omitted variable bias, we can add more explanatory varables leveraging Lasso: a Machine learning algorithm whose goal is to select the right number of variables that should be included in a model. We know that, when working with machine learning in econometrics, we can not just blindly apply algorithms; that is because ML focuses most on performance and flexibility rather than interpretability, which is one of the most important econometrics aspect.
Because we know that $(G_t \times I_t)$ is the most important explanatory variable in Fair model, we will always keep it in the regression, but we can not just run Lasso excluding $(G_t \times I_t)$ from its penalty: that would still give us omitted variables problems due to the fact that Lasso would not select other explanatory variables correlated with GDP. The right way to apply lasso is to use double-selection Lasso, where we apply the algorithm both to our dependent variable and to $(G_t \times I_t)$ and then merge the two results together.
Notice that, because we know that the variables used in Fair's model are statistically significant, we will always keep them outside of the lasso penalty. We highlight that, because of the nature of our additional variables, we do not expect them to have a non-costant partial effect on the democratic share, that is because they represent basic "good news" and "bad news", therefore we will only introduce them in levels.

Below we display the outcome of DS-Lasso:



```
nasdaq_growth_15q_i |      x
              e_1_i |                x
```

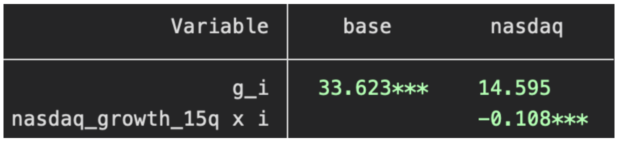
We notice that the algorithm selects both NASDAQ growth whitin the 15 quarters of the president term and employment growth in the last year. The problem with these explanatory variables is that, as shown below, our MSE and consequently our predictions are worsened by their addition:



```
year   total_~n    mse_l_n

2012        253    26.6373
2016        182   38.39827
```

The explanation of this might lie in the nature of the NASDAQ related variable. We know, from Fair's model, that GDP growth is a strong explanatory variable when trying to forecast election outcomes, and we obviously know that the NASDAQ index is strongly correlated with the GDP of the U.S.. Because in the Double-Selection procedure the first computation is done on the dependent variable, without taking into account $(G_t \times I_t)$, Lasso is selecting the NASDAQ explanatory variable that acts as a proxy for GDP growth. The problem is that the NASDAQ index is computed at the federal level and has a unique coefficent in our regression model, even if each state is affected in a different way by its growth[‡]. This interpretation can be confirmed by comparing the estimates for $(G_t \times I_t)$ when adding the NASDAQ explanatory variable in the model[§]:



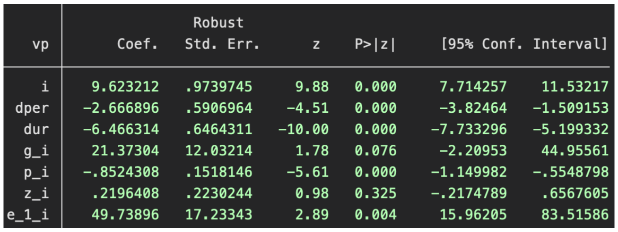
```
             Variable |     base        nasdaq

                  g_i |  33.623***     14.595
nasdaq_growth_15q x i |               -0.108***
```

After this brief discussion we can then exclude the NASDAQ index from the variables that can be selected by the Lasso algorithm, and computing DS-Lasso a second time we get as output just the explanatory variable related to growth in employment. Below we display the estimates outcome:

[‡] Because companies in NASDAQ are concentrated in a few key states (e.g. California).

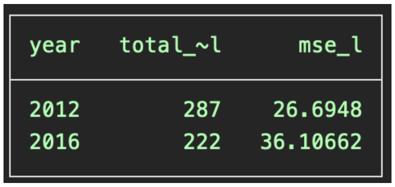
[§] The difference in the absolute value of the two coefficents is explained by the fact that g_i is related to **one year** GDP growth **per capita** wheather nasdaq_growth_15q_i is related to **15 quarters** NASDAQ growth.



```
                      Robust
      vp      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

       i   9.623212   .9739745     9.88   0.000     7.714257    11.53217
    dper  -2.666896   .5906964    -4.51   0.000     -3.82464   -1.509153
     dur  -6.466314   .6464311   -10.00   0.000    -7.733296   -5.199332
     g_i   21.37304   12.03214     1.78   0.076     -2.20953    44.95561
     p_i  -.8524308   .1518146    -5.61   0.000    -1.149982   -.5548798
     z_i   .2196408   .2230244     0.98   0.325    -.2174789    .6567605
   e_1_i   49.73896   17.23343     2.89   0.004     15.96205    83.51586
```

From which we notice that Fair's explanatory variables show similar results as before while the coefficent relative to $E_t \times I_t$ tells us that a 1% growth in employment during the last year of the mandate gives the incumbent party a 49% advantage on its share[***]. We can then compute fitted values and we get the following forecast:

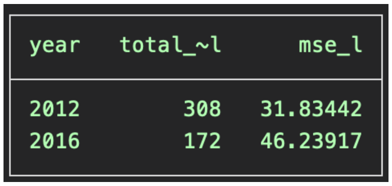


```
year   total_~l      mse_l

2012        287    26.6948
2016        222   36.10662
```

We can notice that even though the MSE are lower, we actually got a worse prediction for the 2012 elections compared to the one of Fair model. We can interpret this result in the following way: as we know the 2016 elections were quite a surprise for professional forecasters, because, even if the majority of statistical models predicted a Democratic win, we now know that it did not went that way. This kind of strange behaviour makes 2016 an outlier, and, because we are working with a small dataset, a single outlier with this power can easily mess up the actual estimates of our model.

To confirm this hypotesis we estimated the same model using only data before 2016 and computed fitted values. As shown below the model performs drastically better in forecasting the 2012 elections, even if its MSE is higher than before:[¶]



```
year   total_~l      mse_l

2012        308   31.83442
2016        172   46.23917
```

We can finally merge these results and display the forecast for the 2012 elections, computed with the model that excludes 2016 data from its training dataset, and the forecast for the 2016 elections computed with the whole dataset as training dataset.

Year	Real	Forecast	Error
2012	332	308	24
2016	227	222	5

Discussion of Results

We can now briefly discuss the results obtained in the previous section.

As stated before, the closeness of our predictions with the actual outcome should not be a surprise; that is a consequence of the fact that we are actually forecasting with a model that "already knows the answers" because we gave them to it in the training dataset. We also highlight that, after reviewing the results, our set of explanatory variables was not that useful as expected. The reason for this is that, when collecting them, we were biased by the framework of Fair's study that was at a federal level; because of that we mainly focused on nationwide general economic aspects, such as Import/Export growth, while electors actually vote taking into account different and more basic aspects, such as the working condition of their local community (modeled by state-level employment growth).

Causal Inference

As defined by Rinfret et al.[2], public policy is a course of action created and/or enacted, typically by a government, in response to public, real-world problems. Even if the U.S. is a federal republic, we oviously expect the political colour of the presidential party to have a drastical impact regarding public policies both at federal and state level. Before addressing the empirical challenges of estimating the effect that the colour of the winner party might have on public policies, we must be sure to understand the full picture of a natural experiment like this one.

As citizens and electors we are definitely interested in understanding how our votes will have an impact on the public policies of our contries. Focusing our discussion on the U.S., we know that, as a consequence of the *first-past-the-post* electoral system, there are basically two main parties from which the president can come from: the Democratic Party and the Repubblican Party. For this reason to fully

[***] Even if it seems a pretty high value we must notice that the employment growth has $max = 0.09$ and $mean = 0.019$.

[¶] Easily explained by the fact that we are reducing the dimension of our training dataset.

understand the impact of our votes we must understand the effect that each of the two party wuold have on the public policies.

Now that we know the reason why we are discussing this, keeping that in mind we can proceed to analyze the empirical challenges of this kind of setting.

The most obvious challenge is related to the definition of control and treatment groups. As we discussed in the lectures, in natural experiments we don't always have a clear distinction between the two groups as we would have in a laboratory experiment[‖] and we need to make the best out of the observational data at our disposal. In our case the problem is that the U.S. elections impacts at the same moment in time all fifty states, therefore we don't actually have a proper control group which is not subjected to the outcome of the elections. To cope with this lack we could try to exploit a similar approach to the one used in *Birth in Hard Times When You Belong To Minorities*[3], where, because different regions were affected by the recession with different strenght at different times, it was possible to estimate a Difference in Differences model to address the coefficent of interest.

A second challenge could be the one related to endogeneity due to omitted variables. As we were reminded once more in the last election, the different states of the federation are definetly heterogeneous; moreover the reasons behind these differences can not be explain by a small set of explanatory variables but lies in different shades that charaterize the life of their citizens, all the way from economics to traditions passing through aspects for which it is rather difficult, if not impossible, to define an explanatory variable, such as ideology. To control for omitted variables we could introduce entity-level[**] fixed effects in our model.

The last challenge we can expect is the one related to the disentanglement of causation from correlation. The situation we are facing in this setting might be similar to the one discussed in *Do Fiscal Rules Matter?*[4], where one of the challenges was to distiguish the effect of the introduction[¶¶] of fiscal rules per se from the effect given by the goodness of the authorities that needed to commit to those rules and of the citizens that needed to follow their authorities. In our case we would want to distinguish the effect of the political colour of the winner per se from the effect given by the different aspects that affected the population political preferences and consequently the winner of the elections. One example that can definetly explain this problem can be found in the evolution of the two parties. There was a time when the two parties seemed to has switched roles regarding racial integration, in the sense that the Democratic party was against the abolition of slavery[††] and any kind of integration of black people; it was actually under a Repubblican president that slavery was finally abolished [‡‡] and the integration process could begin. Comparing this with the picture we live in our days, we can see that the different approaches regarding racial integration, and also all the other aspects of public policies, cannot be thought as just caused by the colour of the party in charge but rather correlated to the subset of electors that voted the running president and party. In a way we are fortunate because strong shifts like the one descripted above do not happen overnight but as consequences of long and clearly observable processes. Nevertheless we also know that, as we grow as a specie, a lot of smaller shifts are happening all the time in completely different directions. One contemporary example could be related to the legalization of cannabis for recreational use in Alaska back in 2014, under a Repubblican governor [§§]. In this case we can surely consider this policy as correlated to the general ideology of the voters rather than caused by the colour of the party in charge. As for the other challenges we could try to apply the same approach discussed in the paper, called difference in discontinuity, even if in our case we may find the application a lot more complex. The cause of this complexity still lies in the heterogeneity between different states of the federation, where different correlations may apply based on the dissimilarities that we can found in their voters.

As stated in this brief discussion we can notice that evaluating the effect of the political colour of the running party over public policies presents a lot of different challenges that we need to control for in order to compute a reliable estimate. Our main focus, as econometricians, should therefore be not only on the practical aspect of the experiment setting but rather on the method that we need to follow in order to properly leverage all the statistical instruments that we will learn in our career.

[‖] I.e. In a medical experiment regarding the effects of a new treatment we can easily define two groups: one of people that will recieve the actual treatment and one of people that will recieve a placebo.

[**] We highlight that these entity-level fixed effect could be requested not only at the basic state-level but we may need to take into account different kind of entities such as group of states, districts, ...

[¶¶] Actually the exemption.

[††] Recall the slogan fot the 1868 Democratic National Convention: *This is a White Man's Country, Let White Men Rule.*

[‡‡] Abraham Lincoln, January 31, 1865.

[§§] The example was chosen for its simplicity even if it does not completely resamble our discussion that is related to policies at a federal level.

References

[1] Ray C. Fair. *Presidential and Congressional Vote-Share Equations*, 2007.

[2] Sara R. Rinfret, Denise Scheberle, and Michelle C. Pautz *Public Policy: A Concise Introduction*, 2018.

[3] Paola Bertoli, Veronica Grembi and The Linh Bao Nguyen. *Birth in Hard Times When You Belong To Minorities*, 2020.

[4] Veronica Grembi, Tommaso Nannicini, and Ugo Troiano. *Do Fiscal Rules Matter?*, 2016.