Empirical models to forecast U.S. Presidential Elections

Marwan Agourram (967349)

December 2021

Abstract

Starting from the model for U.S. popular vote forecast estimated by Ray C. Fair, we firstly try to rearrange his analysis with state level fixed effects in order to capture the heterogeneity across the U. S. States.

Then, we enriched the original model by adding some economic, politic and social variables and, using a machine learning algorithm (LASSO), we selected the variables that are significant and we implemented them in the model in order to forecast the results of the 2016 and 2020 U.S. Presidential election.

Finally, we compared the forecasts made with these two approaches.

Contents

1 Introduction

In the last 50 years the United Stated of America Presidential Elections has been a world wide event due to their political relevance, not only for the US citizens, but also for all the world population. Given this importance, many researchers have been studiyng these elections in order to comprehend which factors affects the most President and the Vice-President nominees. One of these researchers is Ray C. Fair, an Economic professor at Yale University

(Connecticut) whose been studying it from 1978. Considering only the 2-party vote (Democrat *versus* Republican), Fair starts from the theoritacl assumption that the vote share is affected by mainly 2 sets of variable: economic ones and the political/social ones. The economic variables are composed by short and long-term indicators of the economic performance (e.g. growth rate of real per capita GDp, absolute value of inflation rate, number of quarters in which there was a GDP growth exceeding a certain threshold), while the latter ones consider the presence of an incumbent party by time and space dimensions.

Starting from the empirical model developed in 2009, we substituted the National level data to the lower State level in order to see whether it could be useful considering any state fixed effects in the model. An important consideration in this empirical project is given by the fact that the U.S.A is a federal republic of 50 states, and a federal district (District of Columbia), each one having its own economic and social conditions. By including the State level fixed effects we are focusing on verifying if the main indicators used by Fair still work.

In another section I am going to introduce some new economic and social/demographic variables that could be useful to predict the presidential outcome. Finally, I will briefly explain how the Presidential nominee is properly executed, and then proceed to translate the predictions in Great Electors vote.

2 Data and Methodology

In this report we focus on the Presidential elections starting from 1976 to 2020, collecting annual data for each U.S. State from various sources, and aggregating them in order to obtain a longitudinal dataset (from now on, we will refer to it as panel data). The panel data results to be *strongly balanced*, that is, it has been constructed by collecting dat for all the tuples year-state as observations. Regarding the dimension of the panel data, given our period of observations ($t = 12$, since there have been 12 elections in this period) and the statistical individuals ($i = 51$, since there are 50 Federal States and one Federal District), the sample results to be composed by 612 observations ($t * i = 12 * 51 = 612$).

Following Fairs' approach we constructed the Dependent variable of the model, called V_p, computed by the percentage of votes for the Democrats out of the 2-party vote, excluding the so called Third parties. The Popular Votes has been taken from the MIT Election Data and Science Lab, and the formula used to construct the variable is the following one:

$$V_p = \frac{N.\ of\ votes\ for\ Democrat}{N.\ of\ votes\ for\ Democrat + N.\ of\ votes\ for\ Republican}$$

The main difference with the original empirical model in terms of the observations regards the time-frequency: Fair was able to collect quarterly data, while in my project I was able to collect only annual data. This leads to an adaptation of the economic variables used by Fair considering annual data, reported in the following formulas:

- $G = [\frac{Y_i}{Y_{i-1}} - 1] * 100$: growth rate of real per capita GDP in percentual in the first three years of the on-term election year (annual rate).

- $P = [\frac{GDPD_{i-1}}{GDPD_i * (-1)} - 1]$: absolute value of the growth rate of the GDP deflator.

- Z: number of years of the administration in which the gk is greater than 2. 7 %.

- gk: growth rate given in a certain year.

The variable Y and $GDPD$, respectively real per capita GDP at annual rate and GDP deflator at annual rate, are computed as in the original work by Fair. In order to capture the voter's point of view, we used *Interactions* variables instead of the economic variables, obtained simply by multiplicating the economic variables with the Incumbency variable I, resulting in $G_i,\ P_i,\ Z_i$. The main aim in using interactions is to comprehend the correlation between the economic current situation with the political party in charge at the same time.

For the incumbency variables I kept the same ones used by Fair, that is:

- I: *Incumbent* variable, is 1 if the Democrats are in charge at the time of the eleciton, -1 otherwise.

- $DPER$: *President running again*, equal to 1 [-1] if a Democratic [Republican] president is running again, 0 otherwise.

- DUR: *Duration*, indicates how long the incumbent party has been in power in terms of consecutive mandates.

The last variable included in the work done by Fair is a control variable, WAR, that flags the elections during which there has been the World War I or II. Since the aim of this project is forecasting the presidential election during 2016 and 2020, I decided to include the control variable $COVID$, which results in flagging only the last two years.

In the second part of the project, we inclued some other indicators that could possibly enrich the original work and represent other aspects for each State. These indicators are the following ones:

- *Unemployment rate* is a social-economic variable that measures the percent of the labor force that results to be jobless for each State.

- *Civilian labor* is a social-economic variable that measures the portion of the US citizens per State that could be possibly employed or unemployed.

- *Educational level* is social variable that measures the percent of the population that has achieved at least a Bachelor degree.

- *Personal Income in USD* is a social-economic variable that measures the average personal income.

- *Per capita Personal Income in USD* is a social-economic variable that measures the average per capita personal income.

- *Per capita Disposable Income in USD* is a social-economic variable measures the disposable income, that is, the Gross Annual Income minus Payable taxes and other deductions.

- *Personal Current Taxes in USD* is a social-economic variable that measures the amount of Payable taxes that the US citizens for each State has to pay based on their personal income.

- *Federal Government* is a social-economic variables that measures the total expense generated by the Government of each U.S. State.

- *Government and Government Enterprises* is a social-economic variable that measures the state federal expensise in a business enterprise where the government or state has significant control through full, majority, or significant minority ownership.
- *Healthcare and Health Services* is a social-economic variable that measures the government cost in sustaining health related services.
- *House Pricing* is a social-economic variable that measures the price changes of residential housing as a percentage change from some specific start date (1980=100).
- *Density* is a demographic variable that measures the resident State's population per square Km.

To normalize these added indicators, at least for the social-economic ones, I decided to compute the percentage change with respect of the precedent year: the variables in the model that were modified are all indicated by the *Delta* term preceeding their original name:

$$Delta\ Variable\ Name = \frac{(Variable\ Name)_i}{(Variable\ Name)_{i-1}} - 1$$

3 Empirical strategy

As stated before, the main aim of these project is to analyse the effectiveness of Fair's presidential equation in modelling the Democratic vote share in the presidential elections using panel data and State data, considering V_p as the dependent variable and all the other variables used by Fair as the policy variables.

Considering the last two years of elections that are 2016 and 2020, we are going to estimate the coefficients using a simple OLS regression in order to the effect of the independent variables on the dependent variable. After obtaining the coefficients of the independent variables that should be considered in our model, we will perform a test for joint significance of the Fair's economic variables in order to capture their significance in the model and, therefore, be able to exclude them without compromising the results. Once the OLS model is completed we proceed to forecast the Presidential election in the two considered years and, therefore, translate the popular vote into the electoral vote by adding the filter of the Great Electors. Obtained the forecast of the presidential election votes in 2016 and 2020, we proceed to compute the residuals and the squared residuals, obtained by measuring the difference between real true outcome of the popular vote in each state and the forecasted one.

Last but not least, we will add the additional indicators that might provide a better forecast of V_p. The variables that we took in consideration are the ones for which we computed the percentage change, that is:

- Δ *Personal and Disposable Income*: the variation within a year of the per capita personal income and disposable income in each State. The former indicator represent the average amount of economic value that each citizens had, while the latter is states the residual amount of money after taxes. Low-level incomes represents a bad economic situation, while low-level in disposable incomes can be perceived as "bad" by the citizens, mostly if it is due to high tax-level.

- Δ *House Pricing*: is a broad measure of the movement of single-family house prices. It is a weighted, repeat-sales index, meaning that it measures average price changes in repeat sales or refinancings on the same properties. It serves as a timely, accurate indicator of house price trends at various geographic levels. [HPI]

- Δ *Government Expenditure*: considering different source of costs sustained by the federal government, they mainly explain the kind of political behaviour pursued by the Party in charge in each State. This could reflect the affection/disaffection of the citizens towards the main political themes.

- *Unemployment rate*: since one of the main concerns of all the US citizens is related to having a job, the unemployment rate could be related to some of the political promises made by the candidates and, therefore, represent a possible predictable variable.

- *Educational Level*: it is a good indicator of the awareness of people towards the distinction between various politics.

- *Population Density*: this indicator is usefull in order to distinguish various areas in the U.S., that is, urban areas against rural areas, leading to different interpretarions since each area has its own needs and priorities.

Considering these indicators, I included in the model also their lagged values $(n - 1)$, since it is possibly explanatory that the voters start becoming more aware of some factors during last year before the elections.

To these additional variables we applied a machine learning technique, the LASSO (Least Absolute Shrinkage and Selection Operator), that is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and the interpretability of the resulting statistical model. Using this technique enable us to select only the variables that results to be truly useful to predict the outcome of the Presidential elections. While applying the LASSO, we keep growth rate of real per capita GDP, adjusted with the interaction (G_i).

Focusing on the LASSO, in this project we decided to apply two of the main regularization techniques: the Naive-post selection and the Double-post model selection (DS). In particular, we based our LASSO forecasts on the DS selection technique, since we are interested in obtaining a consistent estimate of the parameters; Instead, using the Naive selection technique could lead to inconsistent parameters estimation, possibly generating an endogeneity problem.

In order to correctly perform the LASSO technique we created three *macros*:

- *X_effects*: this macro stores all the state fixed effects and is always included in the LASSO model.

- *X_lasso*: this macro includes all the variables used in the model, except the dependent variable V_p and G_i.

- *X_lasso_pc*: is a subset of the *X_lasso*, in which we exclude all the economic variables.

We then proceed to apply the LASSO to the dependent variable and the macro *X_lasso*; then, we do the same for G_i and the macro *X_lasso_pc*. Each time we perform the LASSO we obtain as result a subset of variables (that were included in the macro of reference) that results to be the most significant and that will be included the regression model. At least, we make predictions with the model containing the variables selected from the lasso and, applying the same reasoning that we used before, we compare the forecasted results with the real ones.

4 Empirical results

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

This section is dedicated to interpreting the results obtained regressing the variables in the original proect and the state dixed effects in the years 2016 and 2020.

For the 2016 regression we found that the variables $DPER, DUR, G_i, Z_i$ are the only ones with significative coefficients. Growth G_i and $DPER$ are the only variables having positive coefficient,s while the other variables presents negative ones. (see Figure 1 & 2). The reasons for that signs could be explained by the fact that good economic performances can increase the confidence in the government, and also by the fact that voters trusts the political nominee in charge and is willing to confirm it in order to keep the situation stable. The reasons for the negative signs relies in the fact that citizens would like to see a change in the political party in charge. Running the Poolability test on the economic variables used by Fair, we notice that all of them should be taken into account in the regression. That is, we reject the null hypothesis that the three coefficients are jointly zero. At least, we performed the poolability test also on the state fixed effects and all of them should also be accounted, since we reject the null hypothesis again.

For the 2020 regression we reach the same results as in 2016 (see Figure 3 & 4).

We can state that the significative variables remain the same for both the models we estimated. Therefore, enlarging the sample by including state level data does not while including the same explanantory variables does not change the outcome of Fair's analysis.

Finally, we computed also the residuals and the squared residuals of the two models. We can easily notice that the squared errors between 2016 and 2020 presents a strong difference, since in 2020 the MSE went down by approximately 11 points (see Figure 5). This reduction in error could be driven by many factors, such as the various news affecting Donald J. Trump persona, as also the spread through social media of the "fake-news detecting techniques".

The results obtained by our prediction model are then translated into Great Electors vote, and then compared with the true results obtained in the two years in consideration. The following table shows the results:

	2016	2020
Prediction	226	298
Real	227	306

Table 1: Prediction comparison with reality.

4.2 Regression with our additional variables selected by LASSO

This section is dedicated to discussing the forecasts obtained using variables that are selected through the LASSO method.

First, we use the LASSO for our dependent variable V_p, while keeping the states fixed effects, and the algorithm selected $Delta\_HHS\_1\_i\_2$, $Delta\_PCDI\_1\_i\_2$, $Delta\_WWS\_i\_2$, that are the changes over one year (squared) of, respectively, Healthcare service expenditure, per capita disposable income and Wages and salaries all selected from the *X_lasso* macro. Then, we applied the LASSO for G_i and the subset of variables chosen by the machine is DUR, $DPER$, and I, all selected from the *X_lasso_pc*.

With the selected variables, we run the regression (see Figure 6). All the selected variable expect for I and the change over year of the Health expenditure results to be significative, with G_i and $DPER$ being the only two variables with positive coefficients. Also, since using the squared values of some variables, we can easily notice that the change in WWS, $PCDI$, and HHS have, in absolute values, large coefficients.

Continuing our analysis in 2020, we applied once again the LASSO on the dependent variable with the *X_lasso* macro resulting in the following selected variables: $\Delta\_HHS\_1\_i\_2$, $\Delta\_WWS\_i\_2$,. Then, we applied LASSO on G_i on the *X_lasso_pc* macro, that selects:

$$Z\_i,\ Delta\_PCI\_i,\ Delta\_PCT\_i,\ Delta\_PCI\_i,\ Delta\_HHS\_1\_i\_2,\ Delta\_PCDI\_1\_i\_2$$

The only statistically significant variables are G_i, $Delta\_WWS\_i\_2$ and $Delta\_PCDI\_1\_i\_2$, with the latter one being the only one with negative coefficient (see Figure 7).

The variables selected by the lasso from the two different years denotes an important role covered by the observations regarding the last year before the presidential election. This could be explained by the importance related to the event and by how the US citizens are affected by the last two years of the mandate.

As we did before, we compute the fitted values of the two models and then obtain the errors. The MSE obtained using the LASSO method between the years 2016 and 2020 are quite similar. We can observe the comparison of the MSE obtained using the OLS with Fairs variables and the ones obtained using the LASSO technique in Figure 8.

Finally, we proceed to compute the Great Electors vote, translating the popular vote. The results are summarized in the following table, that includes also the values forecasted with the OLS model:

	2016	2020
Prediction with Fair	226	298
Prediction with LASSO	216	353
Real	227	306

Table 2: OLS and LASSO model prediction comparison with reality.

5 Conclusions

Our first attempt in forecasting the Presidential outcome using Fair's variable and state level fixed effects turns out to be quite accurate, in particular for the 2016 Presidential election (-1 from the true value). Since the main aim of this project was to improve the original model in order to consider the heterogeneity across different U.S. States, we can state that we were good in resembling Fair's methodology, and with a certain degree of freedom, we were also able to improve it without compromising the predictions. Even if the predictions are exactly equal to the true values, they predict well election of the candidate in both the years under consideration.

When we added our variables (social-economic and demographic one) we noticed quite easily a changed in the results. In fact, in both the considered years our predictions results to be underestimating in 2016 and overestimating in 2020. These results, that are actually capable to predict the election of the candidate, may be due to the heterogeneous variables that we selected. In 2016 the LASSO selected only one variable that is related to the government expenses (Healthcare services)and and two variables that impact directly the day-by-day US citizen (Wages and Personal Disposable Income). By integrating in the model the lagged values of the variables, we tried to prove the importance of the last two years of the mandate in terms of "possibility to be re-elected". Resembling the 2016 election, also the 2020 model selected mainly lagged variables. Moreover, we can notice that in 2020 the LASSO selected more variables related to each US citizen, such as Disposable Income and Current Taxes, may be due to the Corona-virus Pandemic. At least, the MSE obtained with the LASSO model are quite high and this could actually explain the less accuracy that we obtained.

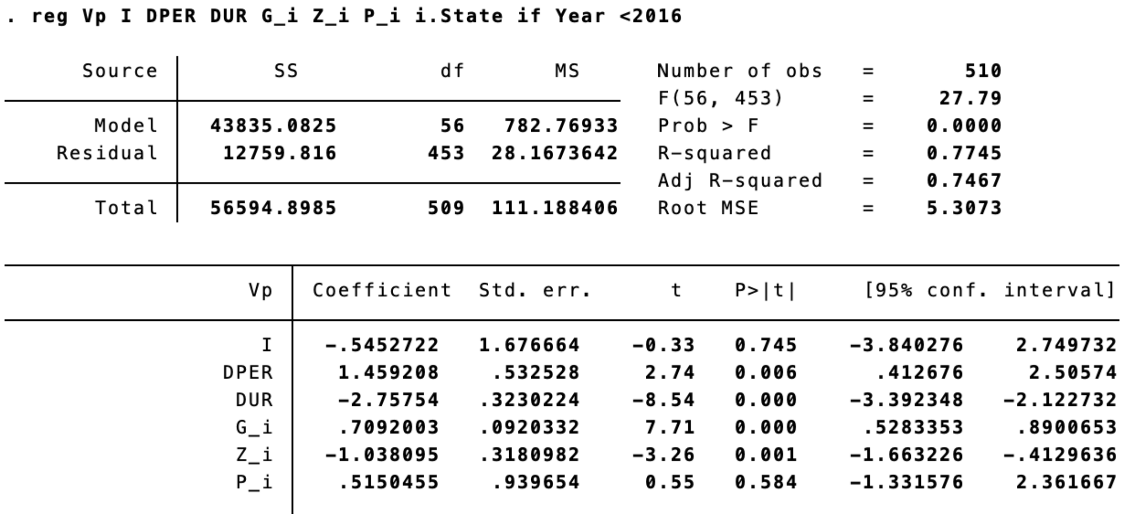
In conclusion, our model can be used to predict if the Democratic party will win the presidential equation, but it is not good enough to forecast the win of the Republicans. It is possible to improve it by adding more important indicators that we had missed while designing our empirical research.

6 Causal Inference

One of the most relecant challenges that researchers encounter concerns the definition of control and treatment groups. In natural experiments it is difficult to clearly distinguish between the two groups. Projecting this issue in our model, the problem is that even thought the presidential election affects all the states at the same time, the federal structure of the U.S.A imposes a certain degree of freedom across states that maked the evaluation of controls and treatement groups either problematic or handy. Since not all the policies and laws happens to be simulateneously affecting the 50 states, it is possible to invoke methods such as Difference-in-Difference (DD). The DD methods is used to estimate the effect of a certain treatment by comparing the changes in outcome over time between the treated group and the non-treated group. Therefore, the estimator comes from a combination of two dimensions (space and time) and it looks like this:

$$DD = [\bar{Y}(Partecipant^{post}) - \bar{Y}(Partecipant^{pre})] - [\bar{Y}(Comparison^{post}) - \bar{Y}(Comparison^{pre})]$$

Recalling the fact that our model is based on the 50+1 states in the U.S., we should specify that models like DD could actually perform bad, since the estimator require the two groups not to be on different trajectories before the program. A possible solution to this problem could be the Geographic Regression Discontinuity Design, where the analyst compares units close to a border that separates adjacent treated and control areas. Once again, the GRDD can face difficulties while separating the effect of the treatment of interest from other features of the geographic unit.



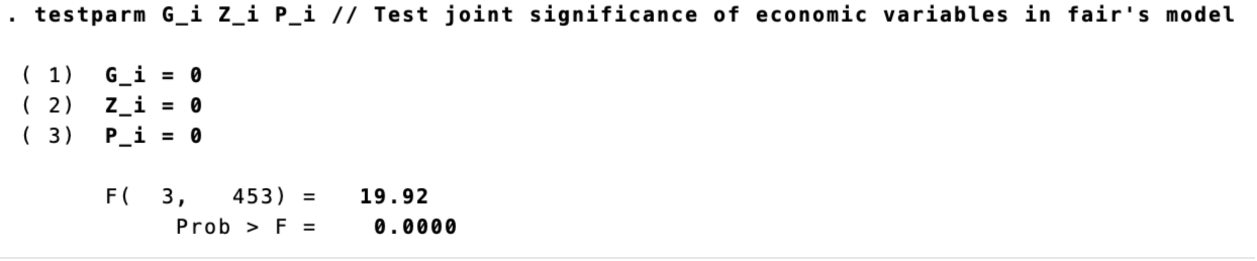
```
. reg Vp I DPER DUR G_i Z_i P_i i.State if Year <2016

      Source |       SS           df       MS      Number of obs   =       510
-------------+----------------------------------   F(56, 453)      =     27.79
       Model |  43835.0825        56  782.76933   Prob > F        =    0.0000
    Residual |   12759.816       453 28.1673642   R-squared       =    0.7745
-------------+----------------------------------   Adj R-squared   =    0.7467
       Total |  56594.8985       509 111.188406   Root MSE        =    5.3073

------------------------------------------------------------------------------
          Vp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           I |  -.5452722   1.676664    -0.33   0.745    -3.840276    2.749732
        DPER |   1.459208    .532528     2.74   0.006      .412676     2.50574
         DUR |   -2.75754   .3230224    -8.54   0.000    -3.392348   -2.122732
         G_i |   .7092003   .0920332     7.71   0.000     .5283353    .8900653
         Z_i |  -1.038095   .3180982    -3.26   0.001    -1.663226   -.4129636
         P_i |   .5150455    .939654     0.55   0.584    -1.331576    2.361667
             |
```

Figure 1: Regression model with Fair's variables and state fixed effect for 2016

Another challenge could be related to the disentanglement of causation from correlation. In our case we would want to distinguish the effect of the political colour of the winner from the effect given by the different aspects that affected the population political preferences and consequently the winner of the elections.

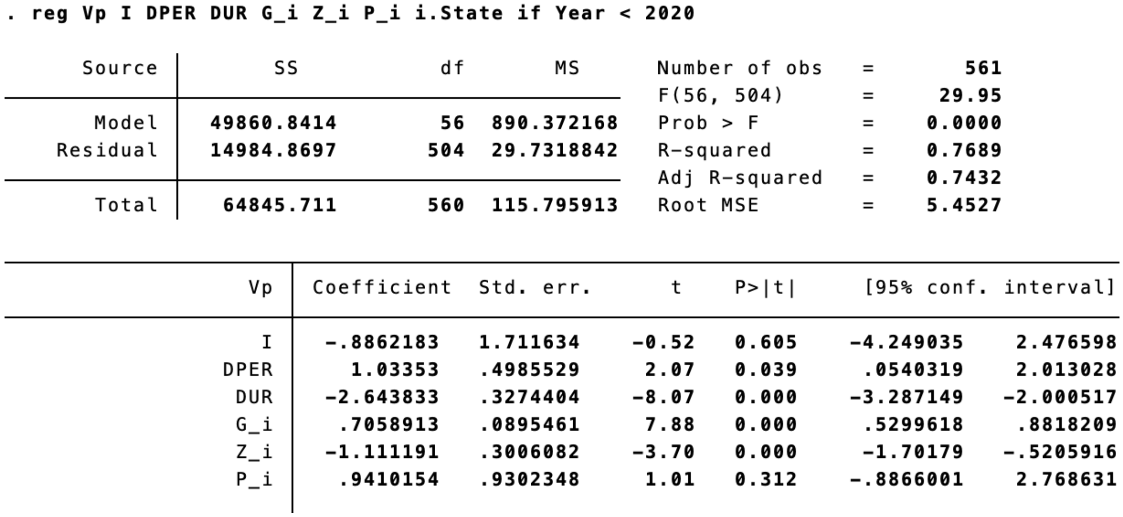


```
. testparm G_i Z_i P_i // Test joint significance of economic variables in fair's model

 ( 1)  G_i = 0
 ( 2)  Z_i = 0
 ( 3)  P_i = 0

       F(  3,   453) =   19.92
            Prob > F =    0.0000
```

Figure 2: Poolability test for state fixed effect for 2016

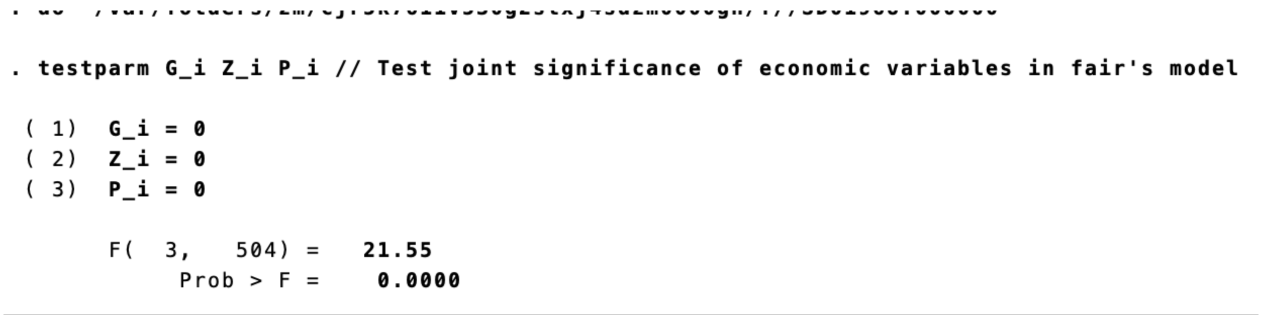


```
. reg Vp I DPER DUR G_i Z_i P_i i.State if Year < 2020

      Source |       SS           df       MS      Number of obs   =       561
-------------+----------------------------------   F(56, 504)      =     29.95
       Model |  49860.8414        56  890.372168   Prob > F        =    0.0000
    Residual |  14984.8697       504  29.7318842   R-squared       =    0.7689
-------------+----------------------------------   Adj R-squared   =    0.7432
       Total |   64845.711       560  115.795913   Root MSE        =    5.4527

------------------------------------------------------------------------------
          Vp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           I |  -.8862183   1.711634    -0.52   0.605    -4.249035    2.476598
        DPER |    1.03353   .4985529     2.07   0.039     .0540319    2.013028
         DUR |  -2.643833   .3274404    -8.07   0.000    -3.287149   -2.000517
         G_i |   .7058913   .0895461     7.88   0.000     .5299618    .8818209
         Z_i |  -1.111191   .3006082    -3.70   0.000     -1.70179   -.5205916
         P_i |   .9410154   .9302348     1.01   0.312    -.8866001    2.768631
             |
```

Figure 3: Regression model with Fair’s variables and state fixed effect for 2020



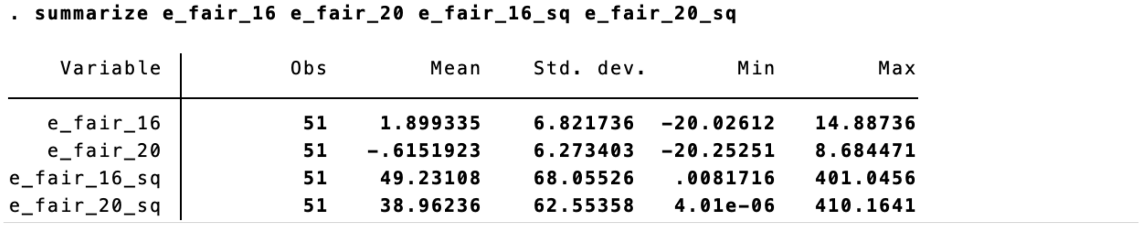
```
. do "/var/folders/zm/cj!3k/611v350g25txj45d2m0000gn/T//SD01568.000000

. testparm G_i Z_i P_i // Test joint significance of economic variables in fair's model

 ( 1)  G_i = 0
 ( 2)  Z_i = 0
 ( 3)  P_i = 0

       F(  3,   504) =   21.55
            Prob > F =    0.0000
```

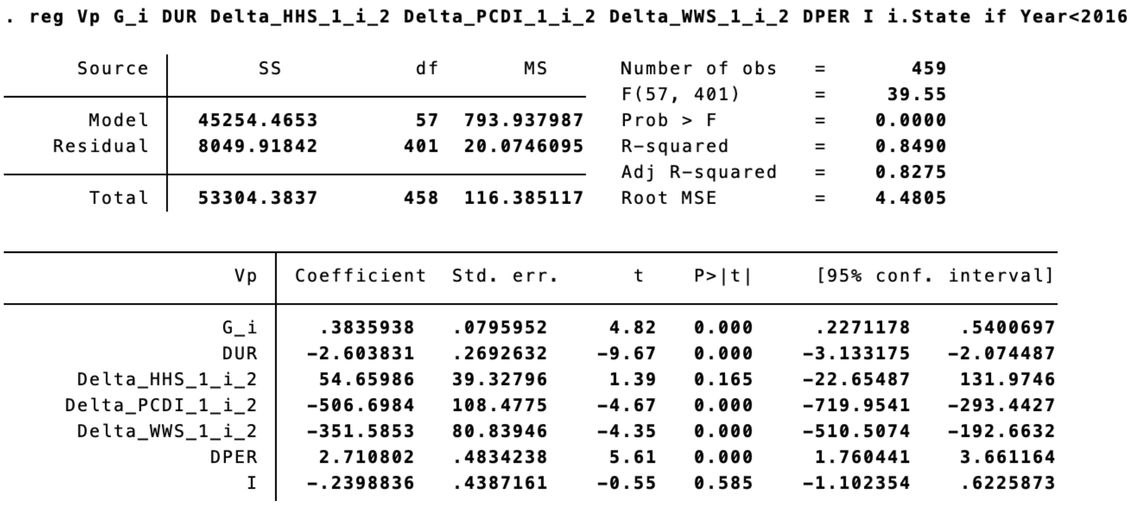
Figure 4: Poolability test for state fixed effect for 2020



```
. summarize e_fair_16 e_fair_20 e_fair_16_sq e_fair_20_sq

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
   e_fair_16 |         51    1.899335    6.821736  -20.02612   14.88736
   e_fair_20 |         51   -.6151923    6.273403  -20.25251   8.684471
e_fair_16_sq |         51    49.23108    68.05526   .0081716   401.0456
e_fair_20_sq |         51    38.96236    62.55358   4.01e-06   410.1641
```

Figure 5: Mean squared errors comparison 2016-2020 in Fair's model.

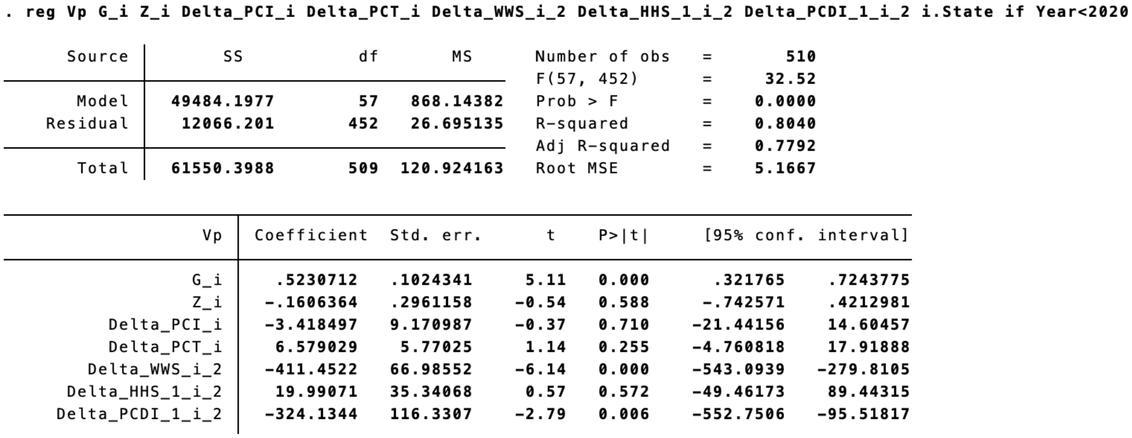


```
. reg Vp G_i DUR Delta_HHS_1_i_2 Delta_PCDI_1_i_2 Delta_WWS_1_i_2 DPER I i.State if Year<2016

      Source |       SS           df       MS      Number of obs   =       459
-------------+----------------------------------   F(57, 401)      =     39.55
       Model |  45254.4653        57  793.937987   Prob > F        =    0.0000
    Residual |  8049.91842       401  20.0746095   R-squared       =    0.8490
-------------+----------------------------------   Adj R-squared   =    0.8275
       Total |  53304.3837       458  116.385117   Root MSE        =    4.4805

------------------------------------------------------------------------------
          Vp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         G_i |   .3835938   .0795952     4.82   0.000     .2271178    .5400697
         DUR |  -2.603831   .2692632    -9.67   0.000    -3.133175   -2.074487
Delta_HHS_1_i_2 |   54.65986   39.32796     1.39   0.165    -22.65487    131.9746
Delta_PCDI_1_i_2 |  -506.6984   108.4775    -4.67   0.000    -719.9541   -293.4427
Delta_WWS_1_i_2 |  -351.5853   80.83946    -4.35   0.000    -510.5074   -192.6632
        DPER |   2.710802   .4834238     5.61   0.000     1.760441    3.661164
           I |  -.2398836   .4387161    -0.55   0.585    -1.102354    .6225873
```

Figure 6: Regression model with LASSO selected variables and state fixed effect for 2016.

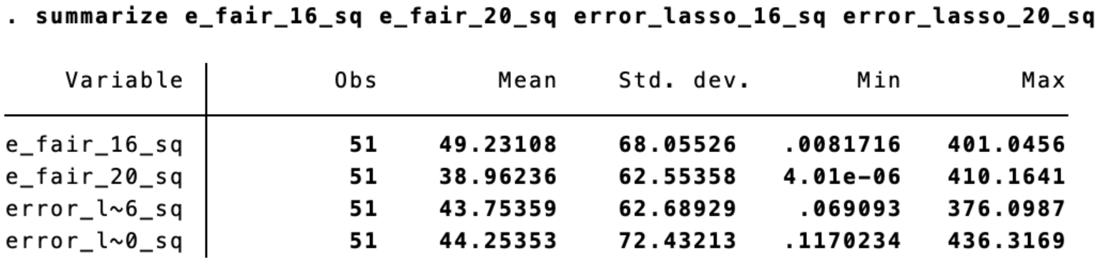


```
. reg Vp G_i Z_i Delta_PCI_i Delta_PCT_i Delta_WWS_i_2 Delta_HHS_1_i_2 Delta_PCDI_1_i_2 i.State if Year<2020

      Source |       SS           df       MS      Number of obs   =       510
-------------+----------------------------------   F(57, 452)      =     32.52
       Model |  49484.1977        57   868.14382   Prob > F        =    0.0000
    Residual |   12066.201       452   26.695135   R-squared       =    0.8040
-------------+----------------------------------   Adj R-squared   =    0.7792
       Total |  61550.3988       509  120.924163   Root MSE        =    5.1667

------------------------------------------------------------------------------
              Vp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-----------------+----------------------------------------------------------------
             G_i |   .5230712   .1024341     5.11   0.000      .321765    .7243775
             Z_i |  -.1606364   .2961158    -0.54   0.588     -.742571    .4212981
     Delta_PCI_i |  -3.418497   9.170987    -0.37   0.710    -21.44156    14.60457
     Delta_PCT_i |   6.579029    5.77025     1.14   0.255    -4.760818    17.91888
   Delta_WWS_i_2 |  -411.4522   66.98552    -6.14   0.000    -543.0939   -279.8105
 Delta_HHS_1_i_2 |   19.99071   35.34068     0.57   0.572    -49.46173    89.44315
Delta_PCDI_1_i_2 |  -324.1344   116.3307    -2.79   0.006    -552.7506   -95.51817
```

Figure 7: Regression model with LASSO selected variables and state fixed effect for 2020.



```
. summarize e_fair_16_sq e_fair_20_sq error_lasso_16_sq error_lasso_20_sq

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
e_fair_16_sq |         51    49.23108    68.05526   .0081716   401.0456
e_fair_20_sq |         51    38.96236    62.55358   4.01e-06   410.1641
error_l~6_sq |         51    43.75359    62.68929    .069093   376.0987
error_l~0_sq |         51    44.25353    72.43213   .1170234   436.3169
```

Figure 8: Mean squared errors comparison 2016-2020 in Fair's model and in the LASSO model.