**Data II Final Project Write-up**

Topic: Governor Polls and its impact on President’s approval rate

By Denny Ziyang Deng

The Topic that I picked to finish up the quarter is regarding the relationships between governor’s poll in pivotal swing states and how that might introduce an impact on the up-coming midterm election in 2022. As we all know, Trump’s defeat in the 2020 election has marked an end to his era while also leaving many legacy behind. 72 million votes for Trump certainly means something intriguing for the political battle in the US. As the first 200 days of Biden’s administration almost pass, President Biden is experiencing one of the most significant approval rate drops since WWII.

What follows up next is the Virginia governor and New Jersey Senate President Election. While Edward Durr Jr, a commercial truck driver who only advertised his campaign through social media posts and tweets, beat incumbent New Jersey Senate President Steve Sweeney in the election, an astounding similarity can be found between Durr’s journey of success and that of Former president Donald Trump. They both held no elective experience; both heavily used social media to campaign. Moreover, Virginia recently had welcomed its new republican governor Glenn Youngkin who also held no elective office experience. Youngkin, who served as Private Equity CEO in his past career again worked his way up to head of State through personal wealth contributions like that of Donald Trump’s. And Virginia, used to be a light-blue state that Biden beatTrump in the 2020 presidential election by over 10% margins. All these observations add up to imply something interesting for researchers to contemplate: is there a particular relationship between democratic candidates’ defeat in the state campaign versus people’s disapproval rating for president Biden?

To analyze the problem, I first downloaded the database of president's approval poll and state governor’s poll from source website *FiveThirtyEight.* And I planned and separated the research into following steps:

* Data Cleaning and Merging
* Linear Regression and Muti-linear Regression
* Supervised Regression Machine Learning
* Plotting of Residuals, regression, facetgrid, and Pie Chart

**Data Cleaning and Merging:**

First I downloaded two databases and data-cleaned them for further usage. For both databases, I dropped unnecessary columns like poll\_id, population, candidate\_id, sponsor\_name, date\_time, etc. Other than that, in order to filter the most desired database related to our research topic, I selected states in the governor database where states are equal to Virginia and New Jersey. Thirdly, I merged the databases in line 54 with:

pd.merge(df\_gov, df\_psd, how='right', on = ['pollster\_id','pollster','display\_name','pollster\_rating\_id','pollster\_rating\_name','fte\_grade','methodology']

Where I matched the data with seven keys, basically the pollster id and names, to complete the merge in a right merge. **The reason we are doing a right merge is because we are more interested in the endogenous variable President approval rate, which lies in the right dataframe president.** Fourthly, I renamed a couple columns for easier understanding. I renamed sample sizesX and Y to be the sample sizes of poll governors and poll presidents. Also, I named a politician column to be president for further clarification. Fifthly, I divided the data frame into two with democratic governor’s dataframe and republican’s dataframe. Lastly, I quantify the fte\_grade column which gives different ratings to pollsters based on FiveThirtyEight’s own grading standard. While the grades are ranked from A+ to D, I give them numerical values referenced from the UChicago GPA metrics, a 4.0 for A, 3.0 for B, etc.

Pivotal parameter definition:

pct: percentage of winning poll rate for governor

Methodology: method of survey from pollsters

Yes: approval rate for president Biden

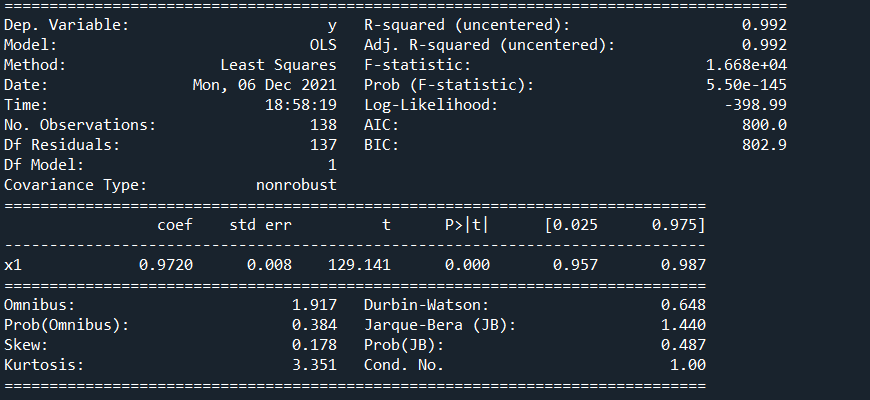
No: Disapproval rate for president Biden

Fte\_grade: rating grades in number for different pollsters

**Linear Regression and Multi-Linear Regression:**

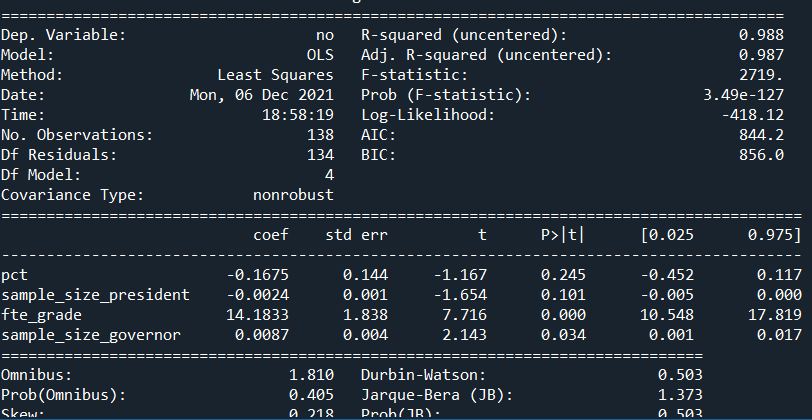
With the merged database, we first runned the regression of

and obtained:



We are seeing strong coefficient determination R2 for the OLS. And we are seeing Solid P values to satiate our significance confidence level at 95%. This can be interpreted as holding everything constant (ceteris paribus), for every unit of increase in pct supporting rate of democratic governor, Biden’s approval rate will likely go up for 0.97%.

Next we are looking at our multi-linear regression. With the model of

We are obtaining 

And strong R2. But we are seeing statistical insignificance for pct and sample\_size\_president this time not passing confidence level at 95%. We hypothesize its due to insufficient sample sizes of the data frame since we are only having 134 observations after narrowing filtering.

**Supervised Machine Learning & Plots**

With train\_test\_split from the Sklearn package, we split and pct and yes into train and test groups to do the prediction and calculate residuals between them. After three different models, we still stick with LinearRegression because other models like RandomForestRegressor only raise the score by 0.01. With Y\_pred calculated based on the LinearRegression class, we then use the mean\_square\_error to calculate the MSE which shows up to be 15.57. This is a relatively high error between the training and test result for Y, maybe more samples and different models will do a better prediction on this.

Further, we also did a MSE on the Supervised Machine Learning Multi-linear Regression model and obtained a MSE of roughly 17.63. This suggests something significantly inaccurate may lie within our model but we will leave it here at the current stage.

Then regarding the plotting, we employed seaborn as well as other generalized functions to graph the regression, residuals, facetgrid as well as pie chart.

Residuals are plotted in the scatterplot that graphs the trained and test sets. Regression simply marks the line through the scatter points of governor’s pct and Biden’s yes points. Facetgrid graphs the pct vs yes in terms of three different parties:Democrats, Republican, Others. And the pie charts are generalized in function to graph the ingredients of online survey types, and party types.

**Conclusion**

Overall, the correlation between the governor’s percentage poll rate for winning the Virginia Governor position does have strong correlation with Biden’s approval rate. We can see this through the Strong R2 and further positive correlation coefficient to state the fact. Yet the causality remains uncertain and other factors like education, state\_unemployment, mortgage\_rate, and interest\_rate might conduct further interesting questions on the research. In a word, the current stage of research serves to represent the general idea and toll the bell on the Biden Administration's policy direction as more critics grow chastising toward his various political actions.