Télécom Paris

Data analysis in Economics 2 Applied Econometrics

Realized By :
Narjes Haouala
Mootez Dakhlaoui
Jorick Defraine

Supervised By :
Laurie CIARAMELLA

April 4, 2021

Contents

1 Introduction

1.1 Motivation

In today's world where sicknesses, poverty and wars terrorise the entire planet, politics obviously has a long way to go. Maybe the first step for improvement is to better understand the people, especially for the interested parties.

It's important for the politicians to get closer to creating a "common good" that is for the people, settling issues and ensuring social welfare and integrity, one way to reach that is by understanding the factors that influence voters' preferences.

One of the most important factors is unemployment since it is related to many other economic indexes and it continues to be a challenging problem in most countries.

This project aims to explain the effect of unemployment on the election. To do so, we will study the different variables and the correlation between them, we will identify the omitted variable and do multiple regressions to determine the significance of the variable : unemployment and its influence (coefficient) on the results of the elections.

1.2 Research Question

In this context, we will answer, the following question:

What is the effect of unemployment on the share of the votes for Donald Trump?

1.3 Process

In order to answer this question, we used data containing many features that may impact the results of the election, like unemployment,Poverty,Income,Covid-19 and many others like it.

In the first part we did an analytical analysis of the data to see the factors that had a huge impact compared to the others. This helped to have a more detailed look for our data.

Our strategy was first to identify our variable of interest which was in this case the Unemployment rate. Then we tried to find the best regression model that contains the maximum control variables that are correlated with our X (unemployment) and had a causal effect on Y(share of the votes for Donald Trump) in order to unbias the coefficient of the estimator.

2 State of the art: previous literature

Looking at elections across history, we can deduce that there are substantial grounds to believe that economic distress played a significant role in the outcome of the 2020 election. Numerous studies on the effect of the economic situation on election results show that dissatisfied voters tend to punish the incumbent and his party by voting for the opposite party (Lewis-Beck and Paldam 2000).

The economic crisis has also been associated with the people's will to vote, (McCartney 2017) which directly affects the outcome of the election. Research, conducted on trade shocks, has further supported the argument that vote shifts are driven by economic forces.(Dippel, Gold, and Heblich 2016; Autor et al. 2017; Malgouyres 2017).

According to a study from M. Incantapulo called The Effects of Unemployment on Voter Turnout in U.S. National Elections, People that lose their job near Election Day are more likely to change their vote. Moreover, data supports the hypothesis that unemployment promotes political mobilization when the unemployment rate is high and results in political withdrawal when it is low.

"In general, unemployed Americans' political behavior is meaningfully influenced by unemployment context to an extent that we do not observe among gainfully employed Americans."

3 Data description

3.1 Data Source

The data set is taken from American statistic websites. The FiveThirtyEight is the source of the election data and the United States Census has the demographic features for each county and state.

FiveThirtyEight, sometimes rendered as 538, is an American website that focuses on opinion poll analysis, politics, economics, and sports blogging.

The United States Census Bureau (USCB), officially the Bureau of the Census, is a principal agency of the U.S.

Federal Statistical System, responsible for producing data about the American people and economy.

3.2 Data Description

The data consist of information of 50 states that are physically on the American soil.each state is decomposed into many counties. The data is available for each county so we have grouped by the state in order to have a global view of the impact of each state on the election. each state is identified by its position (latitude and longitude),the number of votes for both Electors and many other features determining the economic situation of the states. the main features that we will use are :

- Unemployment Unemployment rate (percent) per county.
- Income : Median household income per county.
- Poverty : percent under poverty level per county.
- Cases : number of positive Covid cases per county.
- deaths : number of deaths confirmed by the county.
- Professional : percent employed in management, business, science, and arts per county.
- Construction percent employed in natural resources, construction, and maintenance per county.
- Production percent employed in production, transportation, and material movement per county.

4 Descriptive analysis data

4.1 Comparison between the 2016 and 2020 election

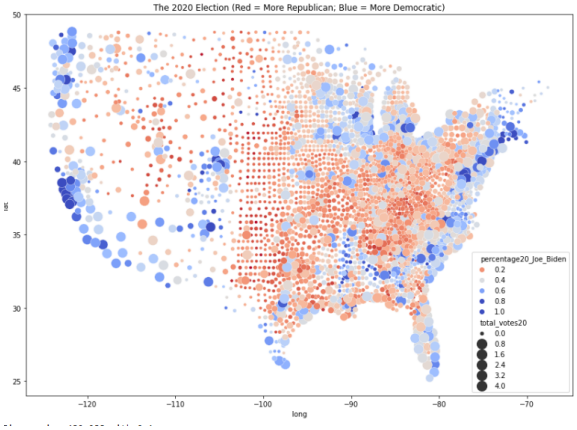


Figure 1: The 2020 Results Compared to the 2016

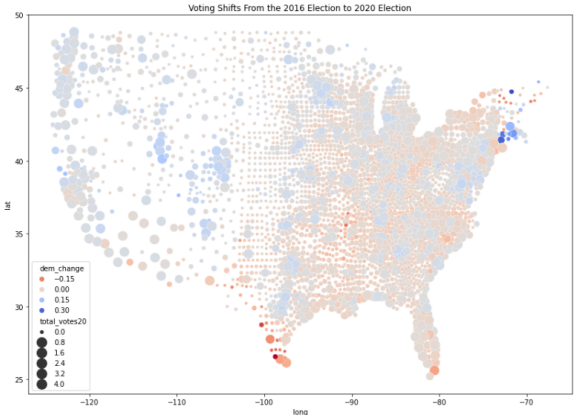


Figure 2: Voting shifts from 2016 to 2020 Election.

We can see clearly that the big cities are more in favor of Democrats whereas the rural areas are in favor of Republicans. There is a clear increase in the gap between the big cities and the more rural areas. Almost all major cities have become more democratic than they were in 2016(right picture blue dots). The rural areas, represented by the smallest data points, are almost all red in colour, especially in Central and Eastern America. Overall, this map tells the story of a country where the areas in continuous red have become even redder, while the areas in continuous blue have become even bluer, illustrating the partisan conviction of many parts of the country, even in relation to the 2016 election.

4.2 Influence of economic factors and the pandemic on US election results

In this part we'll tackle mainly the covid and the major economic factors that each state had. The 2020 election was one of the remarkable elections in the United States because of the presence of a new factor: the new Pandemic.we analyzed this factor in order to find out if it has a huge impact or not on the results of the election.

4.2.1 Covid 19

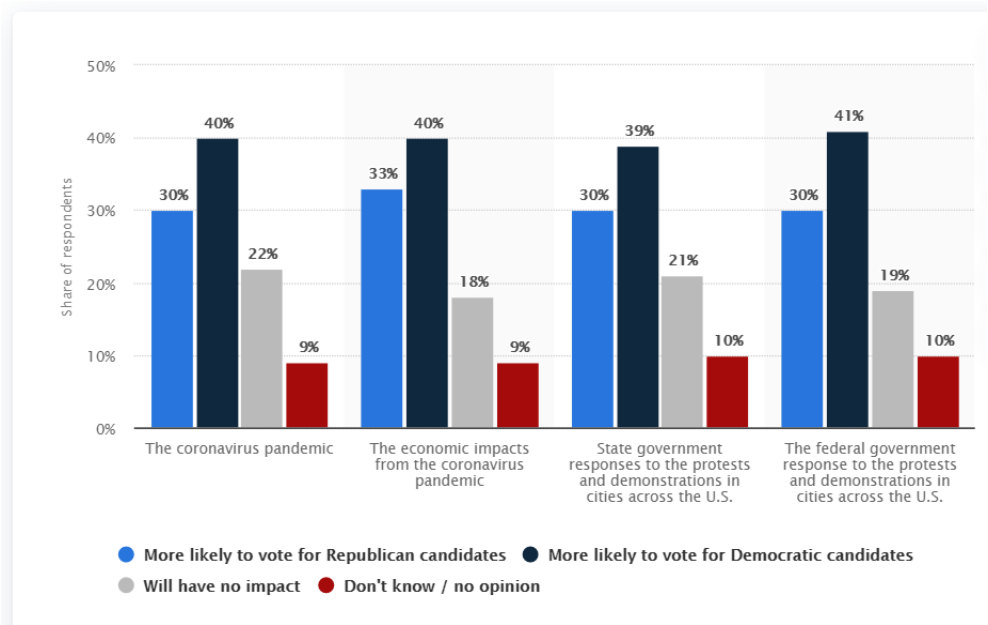


Figure 3: effect of current events on voter intentions in the U.S. 2020 election

According to statista, In a survey conducted mid-June 2020, 40 percent of U.S. adults stated that the coronavirus (COVID-19) pandemic made them more likely to vote for Democratic candidates in the U.S. election to be held in that year. A similar number - 39 percent - also said that state governmental responses to the widespread protests made them more likely to vote for the Democrats. Conversely, for both of these current events, 30 percent of respondents stated they were more likely to vote Republican.

4.2.1.a Covid Cases

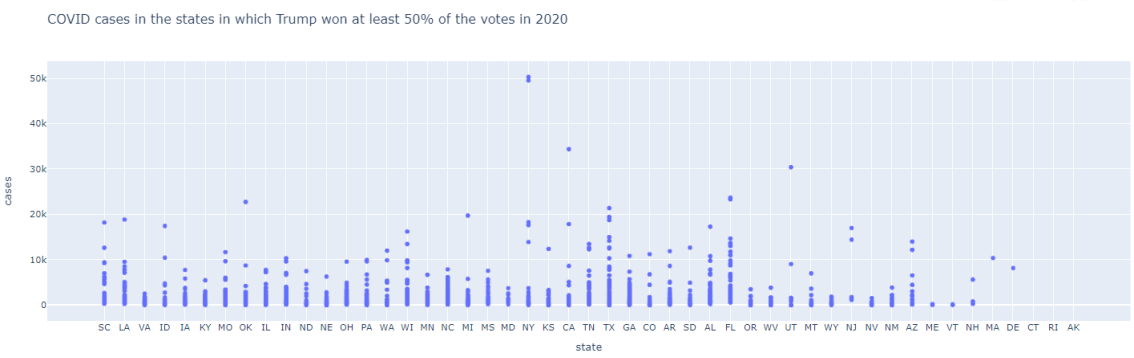


Figure 4: COVID cases in the states in which Trump won at least 50% of the votes in 2020

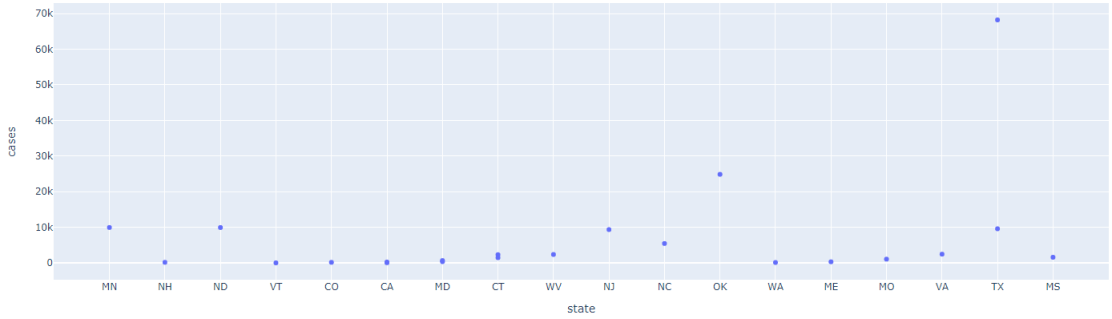


Figure 5: COVID cases in the states in which Trump won at most 50% of the votes in 2020

So here you can see the number of covid cases in the states which trump has at least 50% of the votes. More than 10,000 COVID cases were found in 25 of the states where Trump got the majority of the votes. on the right picture, Only 6 states where Trump did not earn at least 50% of the vote had more than ten thousand COVID events. so this is not decisive ! What about deaths caused by COVID?

4.2.1.b Covid Deaths

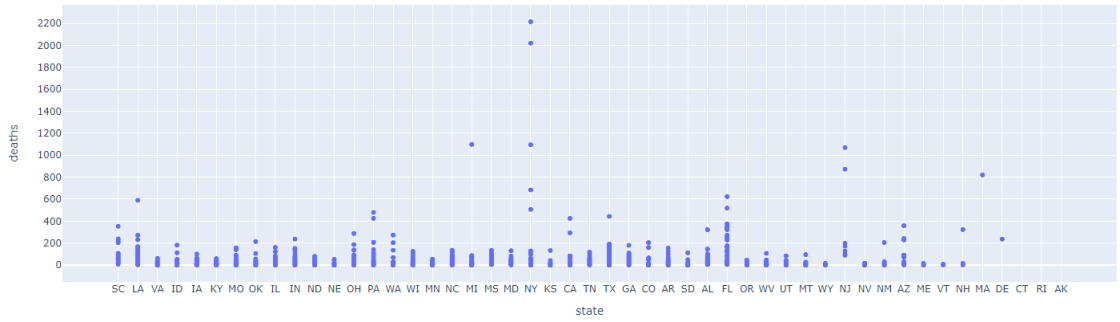


Figure 6: COVID cases in the states in which Trump won at least 50% of the votes in 2020

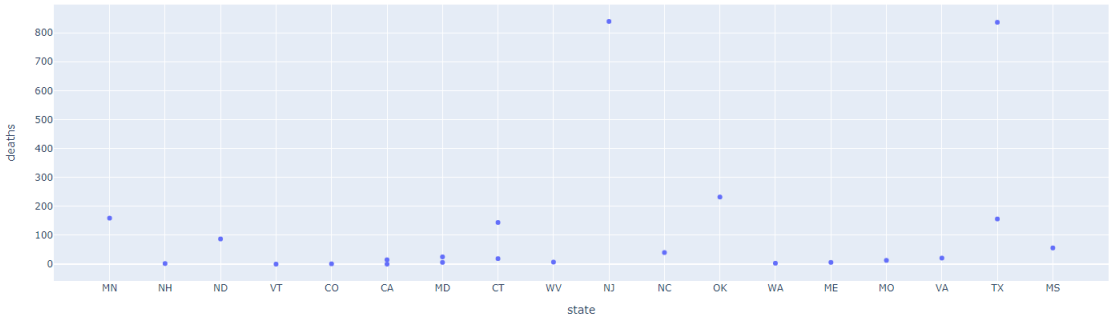


Figure 7: COVID cases in the states in which Trump won at least 50% of the votes in 2020

More than 400 COVID deaths were reported in many of the states where Trump got the majority of the vote in 2020. On the other hand, there were more than 400 COVID deaths in just two of the 18 states for which data is available and in which Trump did not obtain at least 50% of the vote. Should we rule out the possibility of a pandemic affecting the 2020 presidential election in the United States? Definitely not.

4.2.2 Economic factors

4.2.2.a Unemployment

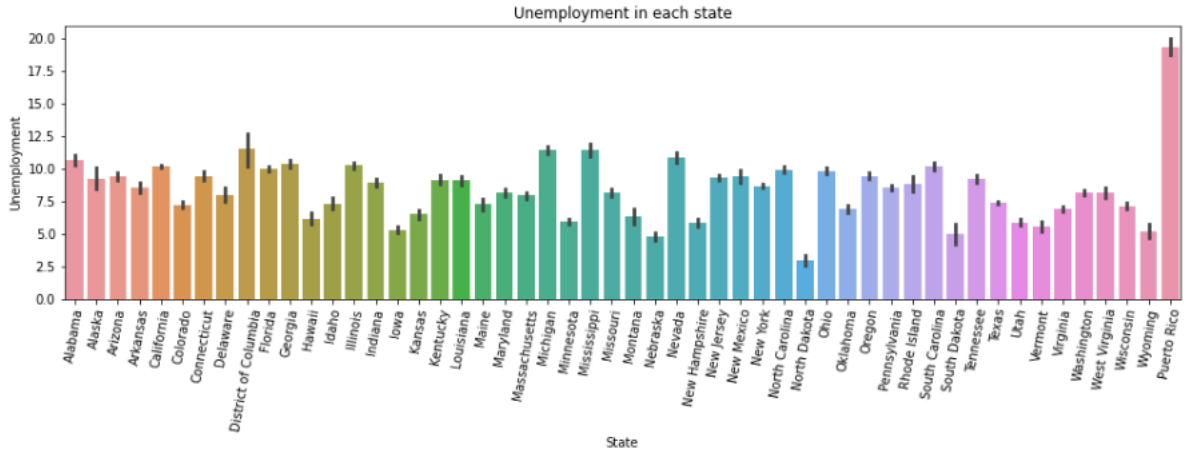


Figure 8: Unemployment rate per State

in Puerto Rico as an example where the unemployment rate is the highest in the whole USA, the democratic bureau has won the election with more than 60% of the votes. We can discard the results we got from Puerto Rico regarding it's not a part of the American soil physically so we'll take as example the second highest unemployment rate which is in District of Columbia. The democratic bureau has won with 92.15% of the votes so the picture in right confirms the result we just had. On the other hand,In North Dakota where we have the lowest unemployment rate Trump campaign allowed him to win with a whopping 65,1%.

4.2.2.b Population

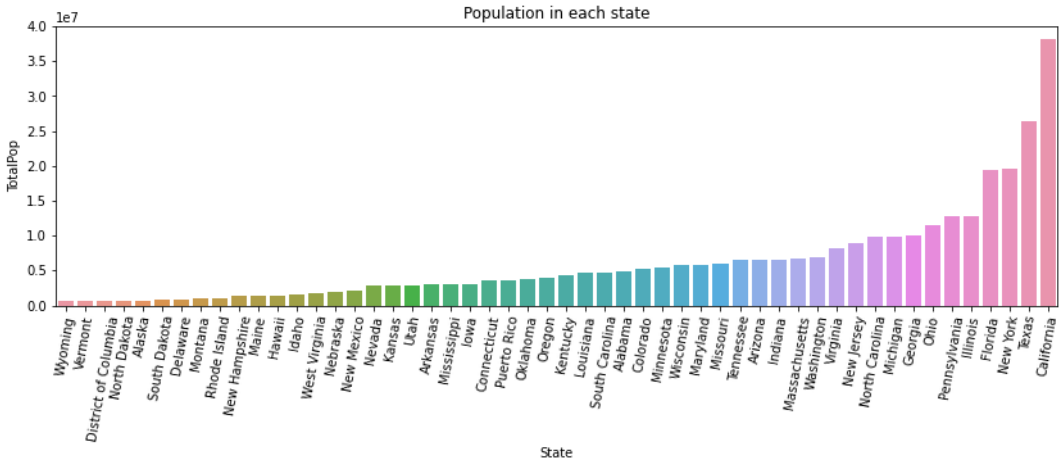


Figure 11: Population per State

As we can see in most populated states like California most of the people voted for Biden as shown in the right picture. but using only the number of population doesn't explain the difference in the votes because in Texas the second highest states when talking about population Trump won with 52% . So population is not really a crucial factor when talking about this election.

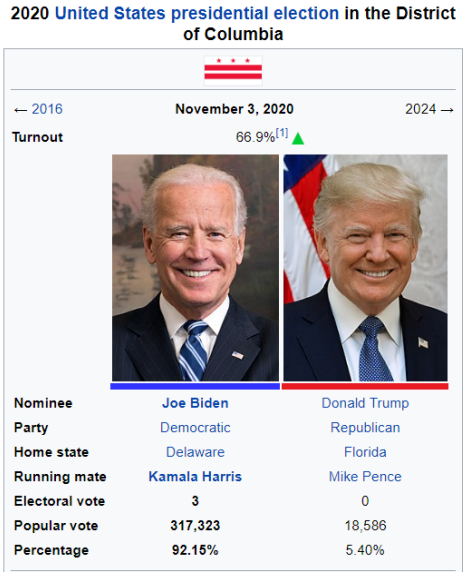


Figure 9: Results in District Of Columbia

Figure 10: Results in North Dakota

4.2.2.c Poverty

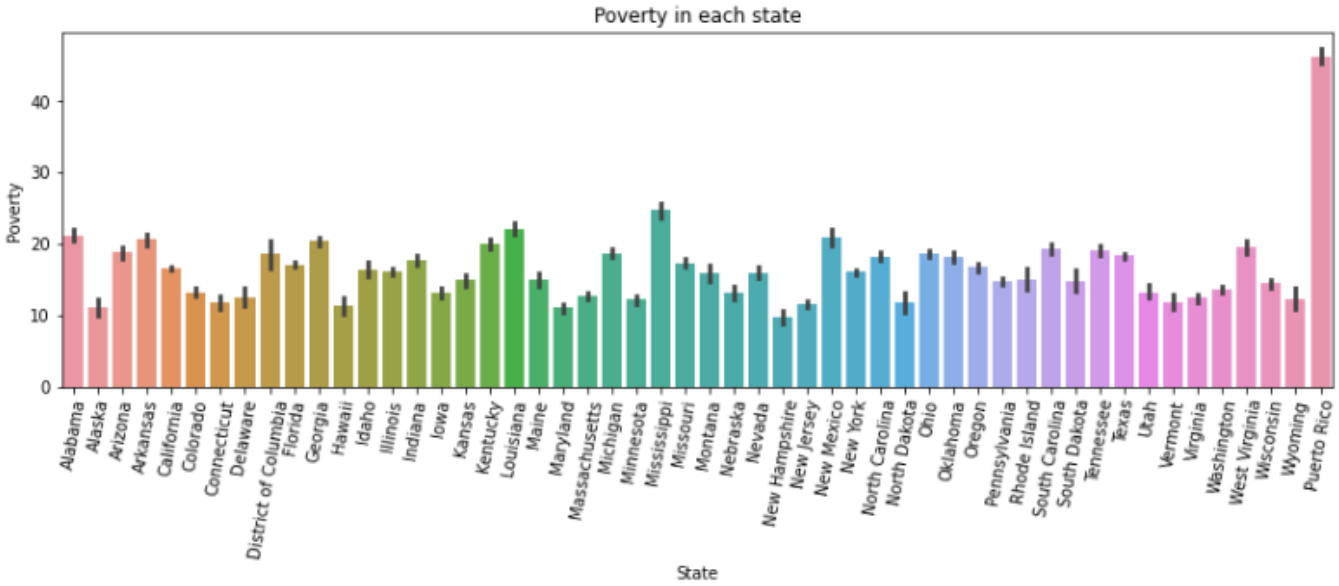
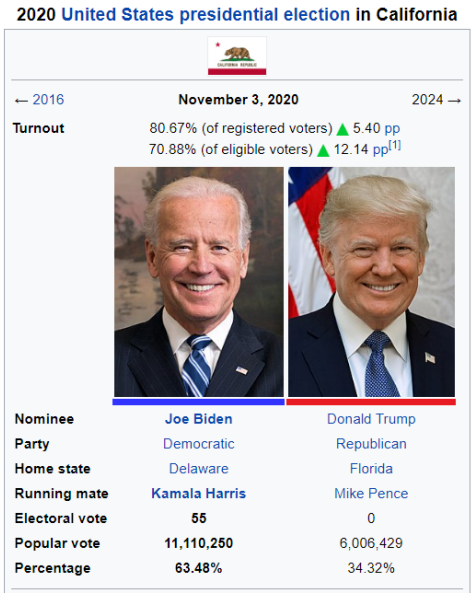


Figure 14: Population per State

Poverty will represent the most meaningful indicator to the reason for the difference in the number of the votes between the 2 competitors compared to the 2016 election. As we said before Puerto Rico has the highest unemployment rate so naturally it will have the highest poverty rate across the USA which was a positive factor for the democratic bureau that won the votes in this state. Alaska which has one of the lowest poverty rates was in favor of republicans with 52.83% So to conclude this first part we can say that Poverty and unemployment have a remarkable impact on the results of the election.



2020 United States presidential election in California

← 2016	November 3, 2020	2024 →
Turnout	80.67% (of registered voters) ▲ 5.40 pp 70.88% (of eligible voters) ▲ 12.14 pp[1]	
Nominee	**Joe Biden**	Donald Trump
Party	Democratic	Republican
Home state	Delaware	Florida
Running mate	**Kamala Harris**	Mike Pence
Electoral vote	**55**	0
Popular vote	**11,110,250**	6,006,429
Percentage	**63.48%**	34.32%

Figure 12: Results in California

2020 United States presidential election in Texas

← 2016	November 3, 2020	2024 →
Turnout	66.73%[1] ▲	
Nominee	**Donald Trump**	Joe Biden
Party	Republican	Democratic
Home state	Florida	Delaware
Running mate	**Mike Pence**	Kamala Harris
Electoral vote	**38**	0
Popular vote	**5,890,347**	5,259,126
Percentage	**52.06%**	46.48%

Figure 13: Results in Texas

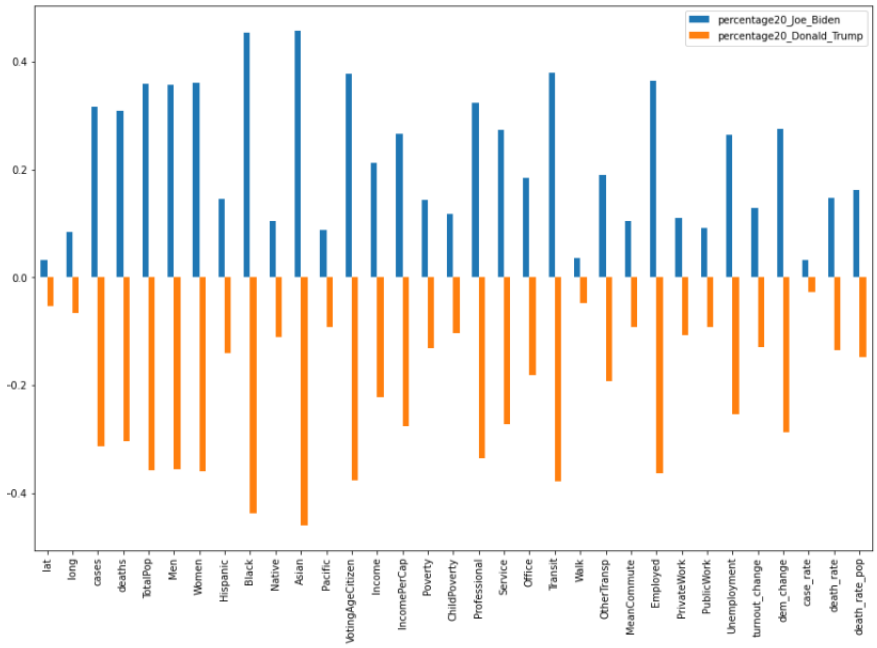
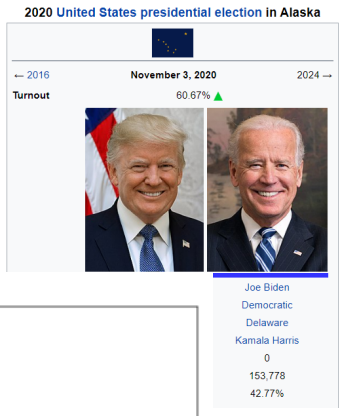


Figure 15: The factors in the 2020 presidential election that had a negative correlation with Trump's votes but a positive correlation with Biden's votes

The variables that were negatively correlated with Trump's votes were also positively correlated with Biden's votes. In the 2020 presidential race, what went against Trump was exactly what worked in Biden's favor. we can see clearly that some features like ethnicity,sex,income and Employement have great impact on the result of the election . we will try to confirm this result when working with our empirical strategy.
This means that we'll work only on one share of the votes (of Trump/ or Biden) because this will lead to the same result because Shares of Trump=1-Shares of Biden.



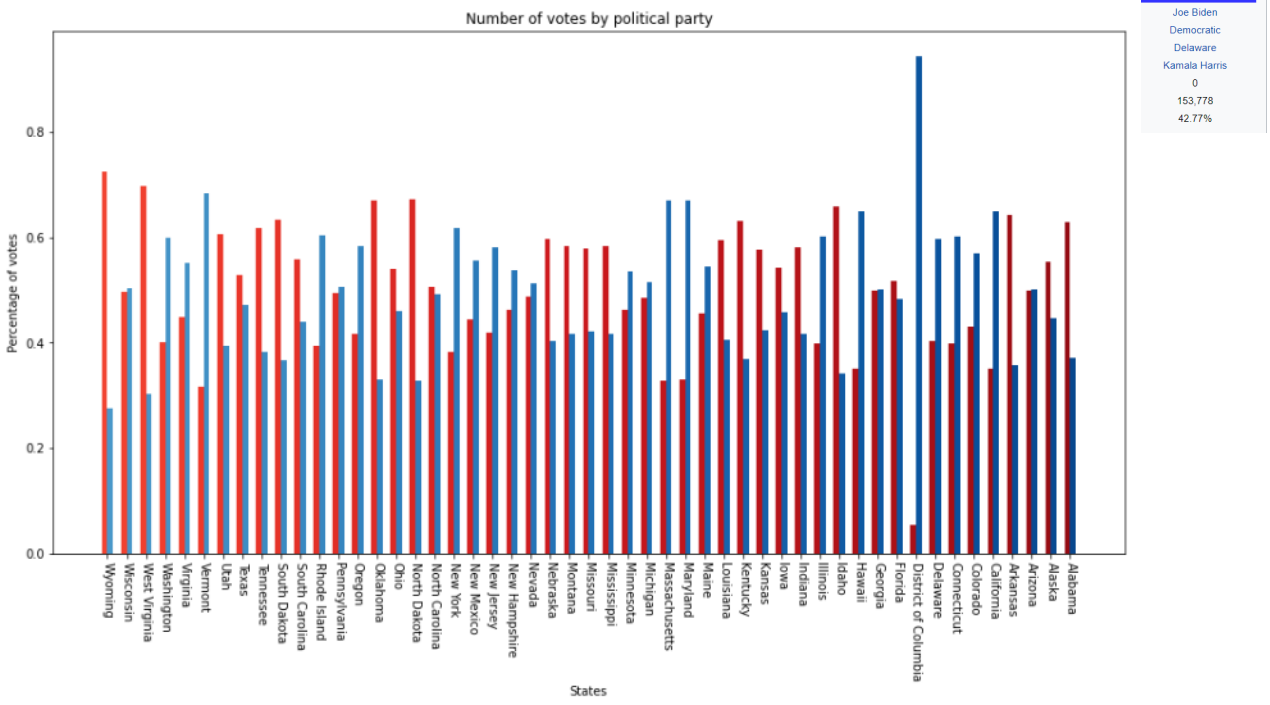


Figure 16: Comparaison between votes in each states for both political parties

Interesting observations:

Republican president's vote are really correlated with race. He receives votes mostly from white people. Moreover his votes come from people with low income, non self-employed. Democratic president has majority of correlation perfectly opposite.

5 Empirical strategy

5.1 Benchmark model

Our variable of interest : explanatory variable is X='unemployment'and the variable we want to explain is Y="The share of votes for Trump"

$$PercentageDonaldTrump = \beta 0 + \beta 1 unemployement + U \tag{1}$$

In this case, the $\beta 1$ can't explain the relation between unemployment and the results of election. In fact, there are many other variables that may influence these results and can also impact unemployment. We say that β is biased and inconsistent.

5.2 Omitted variables

When the model doesn't take into account some variables that are considered to be relevant, a bias will be introduced and will attribute the impact of these variables to those who are included in the model. In this case, our model will not allow us to respond efficiently to our research question. How can we in our case determine what are the factors having an effect on the share of Trump and that can be considered as omitted variables. Even though many factors may influence the share of Trump on the votes such as the number of tweets, the campaign finance, the popularity of the candidate's party... These factors are not really correlated to the variable of interest so they can not be considered as omitted variables. The omitted variables should verify the two criteria : being correlated with X and having impact on Y.

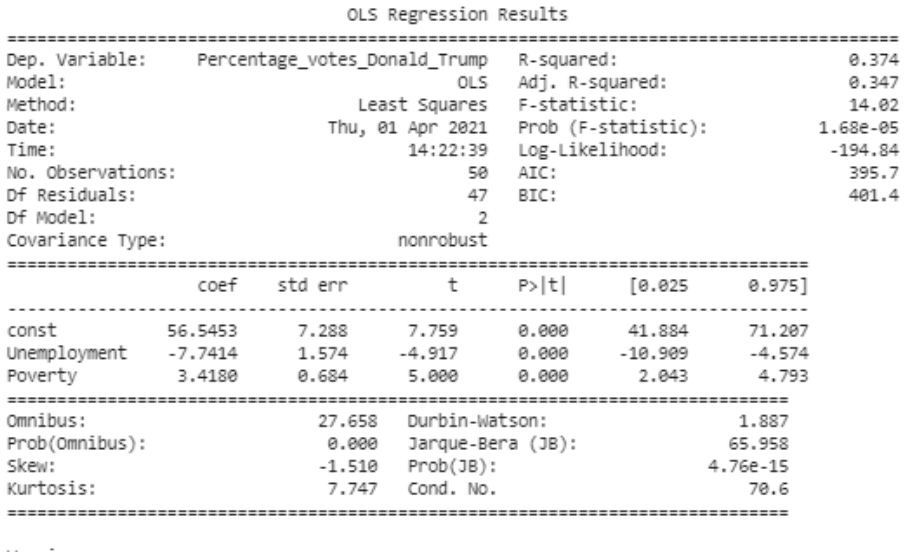
For example, Poverty,Income,Covid(deaths/cases),State wealth (construction/Production..), Type of Population... All these variables can be considered as omitted variable bias and we can use them as control variables

6 Results

6.1 Simple OLS Regression on 2020 Elections

We started by processing a simple OLS Regression in which we didn't take into account the omitted variables. The figure 17 represents the result of a regression that corresponds to the following simple model:

$$PercentageDonaldTrump = \beta_0 + \beta_1 unemployement + U$$



OLS Regression Results

Dep. Variable:	Percentage_votes_Donald_Trump	R-squared:	0.374
Model:	OLS	Adj. R-squared:	0.347
Method:	Least Squares	F-statistic:	14.02
Date:	Thu, 01 Apr 2021	Prob (F-statistic):	1.68e-05
Time:	14:22:39	Log-Likelihood:	-194.84
No. Observations:	50	AIC:	395.7
Df Residuals:	47	BIC:	401.4
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	56.5453	7.288	7.759	0.000	41.884	71.207
Unemployment	-7.7414	1.574	-4.917	0.000	-10.909	-4.574
Poverty	3.4180	0.684	5.000	0.000	2.043	4.793

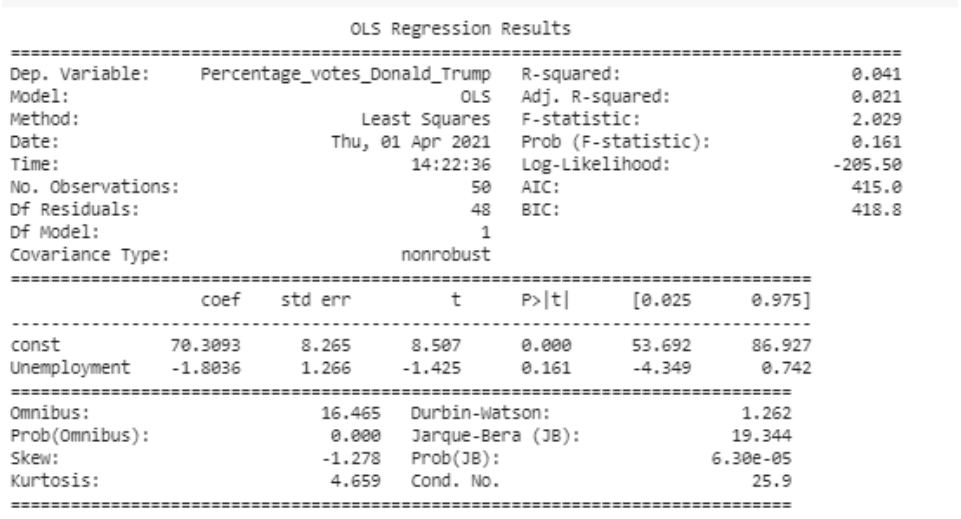
Omnibus:	27.658	Durbin-Watson:	1.887
Prob(Omnibus):	0.000	Jarque-Bera (JB):	65.958
Skew:	-1.510	Prob(JB):	4.76e-15
Kurtosis:	7.747	Cond. No.	70.6

Figure 17: Simple OLS regression for the model: Percentage Donald

$$Trump = \beta_0 + \beta_1 * Unemployement + U$$

As mentioned in the previous section, a highly relevant variable could be poverty: it is plausible that the people facing financial difficulties while Trump was president of the US, they won't vote for him in the 2020's election. In addition, the high level of unemployment will certainly lead to the increase of poverty in the country. From where, the positive correlation between poverty and unemployment. By adding the variable Poverty considered to be significant in our study, we obtain the second model

$$PercentageDonaldTrump = \beta_0 + \beta_1 Unemployement + \beta_2 Poverty$$



OLS Regression Results

Dep. Variable:	Percentage_votes_Donald_Trump	R-squared:	0.041
Model:	OLS	Adj. R-squared:	0.021
Method:	Least Squares	F-statistic:	2.029
Date:	Thu, 01 Apr 2021	Prob (F-statistic):	0.161
Time:	14:22:36	Log-Likelihood:	-205.50
No. Observations:	50	AIC:	415.0
Df Residuals:	48	BIC:	418.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	70.3093	8.265	8.507	0.000	53.692	86.927
Unemployment	-1.8036	1.266	-1.425	0.161	-4.349	0.742

Omnibus:	16.465	Durbin-Watson:	1.262
Prob(Omnibus):	0.000	Jarque-Bera (JB):	19.344
Skew:	-1.278	Prob(JB):	6.30e-05
Kurtosis:	4.659	Cond. No.	25.9

Figure 18: Simple OLS regression by taking into account the variable Poverty

In this case, we can notice that the coefficient associated with the Unemployment variable is reduced from -1.8 in the first simple regression without taking into account the poverty to -7.7 in the second case. In fact, the first model didn't explain correctly the share of Trump in the election since poverty was implicitly explained in the variable unemployment. However, in the second model, the results were improved and the bias was reduced.

It is undeniable that there are other factors that can explain the results of the elections. Hence, we made the same reasoning as before and this allowed us to add other omitted variables. In fact, we can not deny the effect of the new pandemic on the 2020 election. Let's consider the variable "cases" related to the number of cases of Covid 19 per State.

- There is a relation between the number of cases and unemployment/income since rich people will take all possible measures to protect themselves. This high precaution may need a lot of money to avoid getting infected by the virus.
- Covid affects directly Y: share of the votes (direct causal effect). In fact, the higher the number of cases is, the more people will think that the actual president didn't take the right decisions concerning the pandemic. Therefore, it will have a negative effect in our case on the number of votes for Trump.

More control variables should be taken into consideration in the model such as Covid (cases/deaths) or both of them in order to have an accurate study.

When taking these features into account, we obtained the results below:

OLS Regression Results

Dep. Variable:	Percentage_votes_Donald_Trump	R-squared:	0.865
Model:	OLS	Adj. R-squared:	0.842
Method:	Least Squares	F-statistic:	38.44
Date:	Sat, 03 Apr 2021	Prob (F-statistic):	2.87e-16
Time:	21:29:30	Log-Likelihood:	-156.48
No. Observations:	50	AIC:	329.0
Df Residuals:	42	BIC:	344.3
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.5615	27.088	0.242	0.810	-48.104	61.227
Unemployment	-2.6640	0.992	-2.686	0.010	-4.666	-0.662
Poverty	0.4975	0.634	0.785	0.437	-0.782	1.777
deaths	0.0005	0.000	2.023	0.049	1.19e-06	0.001
cases	-0.4791	0.646	-0.742	0.462	-1.782	0.824
Construction	3.3983	0.510	6.660	0.000	2.369	4.428
Production	1.7411	0.332	5.248	0.000	1.072	2.411
Income	-9.789e-05	0.000	-0.387	0.701	-0.001	0.000

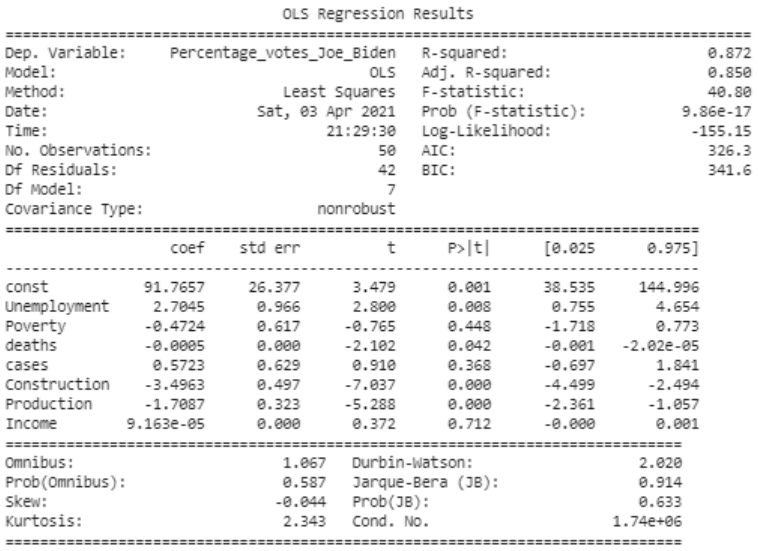
Omnibus:	0.605	Durbin-Watson:	2.036
Prob(Omnibus):	0.739	Jarque-Bera (JB):	0.682
Skew:	-0.045	Prob(JB):	0.711
Kurtosis:	2.435	Cond. No.	1.74e+06

Figure 19: OLS regression with additional control variables to explain the share of votes of Trump

The number of COVID cases is positively correlated with the Unemployment / Income but it has a negative impact on the outcome (Share of the votes for Trump). Thus, we have a Downward bias. Indeed, the coefficient associated with Unemployment decreased when adding the appropriate control variables from -1.8 to -2.6.

6.2 Comparison to The Share of Biden

Since we have two candidates, we expect that the reason that makes a voter choose Biden, would be the reason that makes him doesn't choose Trump. For example, if a high unemployment influenced negatively the share of votes of Trump, it will normally influence positively the share of votes of Trump.



OLS Regression Results

Dep. Variable:	Percentage_votes_Joe_Biden	R-squared:	0.872
Model:	OLS	Adj. R-squared:	0.850
Method:	Least Squares	F-statistic:	40.80
Date:	Sat, 03 Apr 2021	Prob (F-statistic):	9.86e-17
Time:	21:29:30	Log-Likelihood:	-155.15
No. Observations:	50	AIC:	326.3
Df Residuals:	42	BIC:	341.6
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	91.7657	26.377	3.479	0.001	38.535	144.996
Unemployment	2.7045	0.966	2.800	0.008	0.755	4.654
Poverty	-0.4724	0.617	-0.765	0.448	-1.718	0.773
deaths	-0.0005	0.000	-2.102	0.042	-0.001	-2.02e-05
cases	0.5723	0.629	0.910	0.368	-0.697	1.841
Construction	-3.4963	0.497	-7.037	0.000	-4.499	-2.494
Production	-1.7087	0.323	-5.288	0.000	-2.361	-1.057
Income	9.163e-05	0.000	0.372	0.712	-0.000	0.001

Omnibus:	1.067	Durbin-Watson:	2.020
Prob(Omnibus):	0.587	Jarque-Bera (JB):	0.914
Skew:	-0.044	Prob(JB):	0.633
Kurtosis:	2.343	Cond. No.	1.74e+06

Figure 20: OLS regression with additional control variables to explain the share of votes of Biden

We notice here that the coefficient associated to each variable explaining the share of Biden in the votes is exactly the opposite to the coefficient in the case of Trump. Consequently, the results correspond exactly to our expectancy: Factors that went negative for Trump went totally positive for Biden.

6.3 Studying The Correlation Between Variables

Multicollinearity is a problem because the variables in the regression should be independent. If the degree of correlation between variables is high enough, it can cause problems when fitting the model and interpreting the results.

	const	Unemployment	Poverty	deaths	cases	Construction	Production	Income
const	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Unemployment	NaN	1.000000	0.754247	0.277919	0.293714	-0.177328	0.101459	-0.339925
Poverty	NaN	0.754247	1.000000	0.057873	0.198636	0.178321	0.361196	-0.762514
deaths	NaN	0.277919	0.057873	1.000000	0.798578	-0.148241	-0.034283	0.085230
cases	NaN	0.293714	0.198636	0.798578	1.000000	0.059020	0.101361	-0.118474
Construction	NaN	-0.177328	0.178321	-0.148241	0.059020	1.000000	0.187808	-0.552942
Production	NaN	0.101459	0.361196	-0.034283	0.101361	0.187808	1.000000	-0.652412
Income	NaN	-0.339925	-0.762514	0.085230	-0.118474	-0.552942	-0.652412	1.000000

Figure 21: Correlation between explanatory variables

Deaths and cases are much correlated (0.798)>0.5 so there is no additional effect to add deaths while we have the variable cases; that's why we can use only one of them in the model. The variable "Income" is highly correlated with Production, Construction and poverty so we need to drop it.

6.4 Significance of Our Variable of Interest : Unemployment

According to the figure 13: OLS regression with additional control variables to explain the share of votes of Trump. We can conclude about the significance of our variable of interest. In fact, an econometric study gives us the following results:

$$TstatisticUnemployment = -2.4$$

$$PValueUnemployment = 0.016 > 1.6\%$$

We reject the null hypothesis at 5% and 10% level We accept the null hypothesis at 1% level We can consider that in more than 95% of the times $\beta_{unemployment} \neq 0$. From where, we can admit the importance of the variable Unemployment in explaining the results of elections.

OLS Regression Results

Dep. Variable:	Percentage_votes_Donald_Trump	R-squared:	0.851
Model:	OLS	Adj. R-squared:	0.834
Method:	Least Squares	F-statistic:	50.36
Date:	Sat, 03 Apr 2021	Prob (F-statistic):	4.17e-17
Time:	21:29:30	Log-Likelihood:	-158.90
No. Observations:	50	AIC:	329.8
Df Residuals:	44	BIC:	341.3
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.9937	6.411	-0.155	0.878	-13.915	11.927
Unemployment	-2.4327	0.975	-2.495	0.016	-4.397	-0.468
Poverty	0.5210	0.431	1.208	0.234	-0.348	1.390
cases	0.5555	0.393	1.414	0.164	-0.236	1.347
Construction	3.3830	0.388	8.714	0.000	2.601	4.165
Production	1.8003	0.243	7.397	0.000	1.310	2.291

Omnibus:	1.552	Durbin-Watson:	1.988
Prob(Omnibus):	0.460	Jarque-Bera (JB):	1.289
Skew:	-0.209	Prob(JB):	0.525
Kurtosis:	2.334	Cond. No.	186.

Figure 22: OLS regression

$$PercentageDonaldTrump = -0,9937 - 2,43 * Unemployement + 0,52 * Poverty$$

$$+0.55 * cases + 3,38 * Construction + 1,8 * Production$$

The confidence interval for $\beta_{unemployment}$ at 95%: $(\beta_1 - 1.96SE\beta_1\ ,\ \beta_1 + 1,96SE\beta_1) = \{-4.397, -0.468\}$
The 95% interval does not contain 0 so we can confirm with the T-statistic test that

$$\beta_{unemployment} \neq 0$$

.

6.5 Heteroskedastic or Homoscedastic

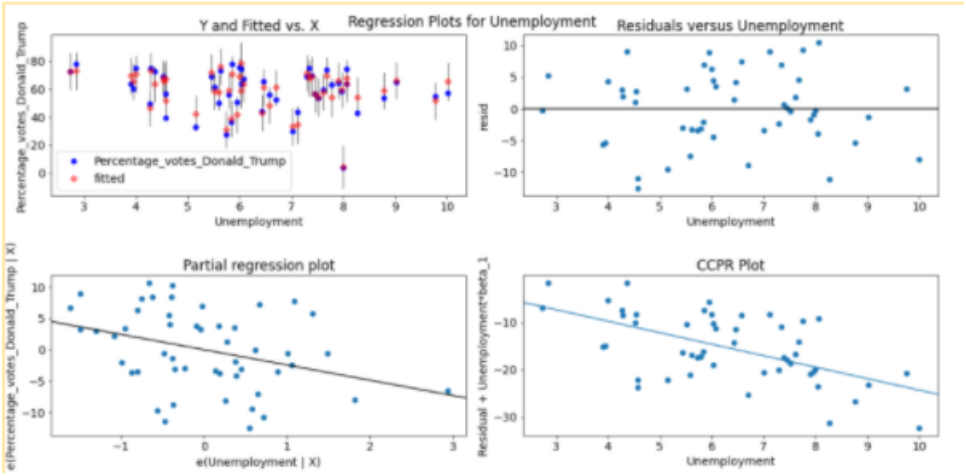


Figure 23: Regression plots for unemployment

On the figure number 2 on the upper-right we can confirm that we don't have Heteroskedasticity in our data because the variance of the error term doesn't not vary or depend too much on x (here unemployment). for many values of x the var(u|x) is pretty constant.

Confirming the homoscedasticity hypothesis:

In order to confirm our results obtained by visualisation (figure 23), we opted for robust regression which is a form of regression analysis designed to overcome some limitations of traditional parametric

and non-parametric methods. In fact, when the data contains outliers, the variance will be high. Least squares estimation is inefficient and can be biased . If it is the case in our data, we can say that we have Heteroskedasticity. Robust regression can mask these outliers and would give a different result than the classical regression.

Robust linear Model Regression Results

Dep. Variable:	Percentage_votes_Donald_Trump	No. Observations:	50
Model:	RLM	Df Residuals:	44
Method:	IRLS	Df Model:	5
Norm:	HuberT		
Scale Est.:	mad		
Cov Type:	H1		
Date:	Sat, 03 Apr 2021		
Time:	21:29:32		
No. Iterations:	25		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.8297	7.183	0.116	0.908	-13.249	14.908
Unemployment	-2.5742	1.092	-2.357	0.018	-4.715	-0.433
Poverty	0.5107	0.483	1.057	0.291	-0.436	1.458
cases	0.7006	0.440	1.592	0.111	-0.162	1.563
Construction	3.3336	0.435	7.664	0.000	2.481	4.186
Production	1.7828	0.273	6.538	0.000	1.248	2.317

If the model instance has been used for another fit with different fit parameters, then the fit options might not be the correct ones anymore .

Figure 24: Robust regression

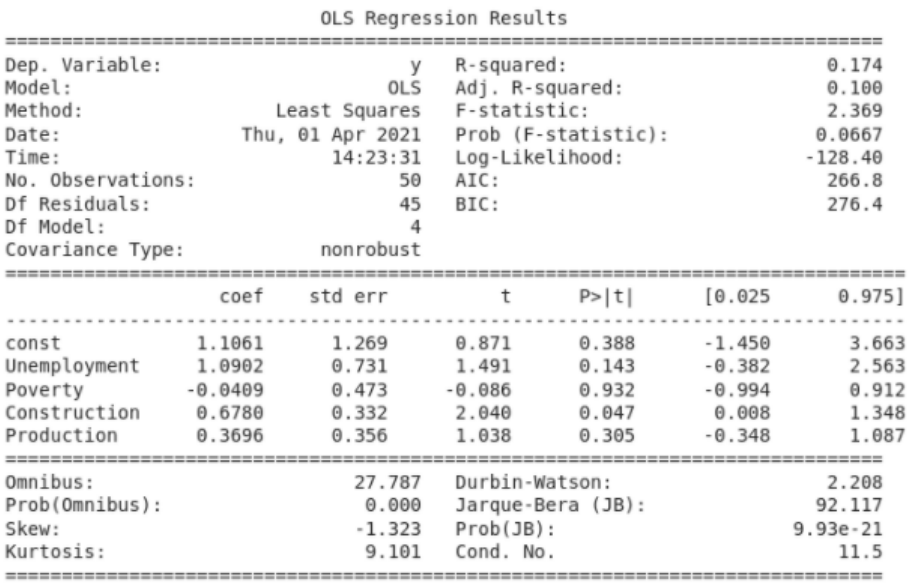
The results from this Robust OLS (figure 22) confirms that we don't have Heteroskedasticity, regarding we got the same coefficients when using OLS with the same features

6.6 Change of The Impact of Unemployment Between 2016 and 2020 Elections

Panel Data Without Covid Cases
We would like to know if the number of covid cases has an impact on the result of the elections. In order to have an idea about it, we process an OLS Regression with different panel data. We create a panel data that contains no information about covid cases as a witness.

This OLS regression takes into account the omitted variables between the elections of 2016 and 2020. In order to achieve it, we subtracted the share of vote for Donald Trump and the Unemployment, Construction, Production and Poverty variables by states.

Here is the results of the OLS Regression with this panel data:



OLS Regression Results

Dep. Variable:	y	R-squared:	0.174
Model:	OLS	Adj. R-squared:	0.100
Method:	Least Squares	F-statistic:	2.369
Date:	Thu, 01 Apr 2021	Prob (F-statistic):	0.0667
Time:	14:23:31	Log-Likelihood:	-128.40
No. Observations:	50	AIC:	266.8
Df Residuals:	45	BIC:	276.4
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.1061	1.269	0.871	0.388	-1.450	3.663
Unemployment	1.0902	0.731	1.491	0.143	-0.382	2.563
Poverty	-0.0409	0.473	-0.086	0.932	-0.994	0.912
Construction	0.6780	0.332	2.040	0.047	0.008	1.348
Production	0.3696	0.356	1.038	0.305	-0.348	1.087

Omnibus:	27.787	Durbin-Watson:	2.208
Prob(Omnibus):	0.000	Jarque-Bera (JB):	92.117
Skew:	-1.323	Prob(JB):	9.93e-21
Kurtosis:	9.101	Cond. No.	11.5

Figure 25: OLS Regression from panel data without the covid cases variable

We obtained an R-squared of 0.174. It means that only 17.4% of variation in Y is explained by the variable Unemployment, Poverty, Construction and Production. We can also see that the omitted variable has a lot of impact on Y with a coefficient of 1.1061 and a t-test of 0.871. We can conclude that this OLS regression is not sufficient to explain the share of votes for Donald Trump. However, it is very useful to have this result to conclude then on the next regressions.

Panel Data With Covid Cases variable

We would like to know if the number of covid cases has an impact on the result of the elections. In order to have an idea about it, we process an OLS Regression with different panel data. We create a panel data that contains no information about covid cases as a witness.

This OLS regression takes into account the omitted variables between the elections of 2016 and 2020. In order to achieve it, we subtracted the share of vote for Donald Trump and the Unemployment, Construction, Production and Poverty variables by states.

Here is the results of the OLS Regression with this panel data:

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.247
Model:	OLS	Adj. R-squared (uncentered):	0.163
Method:	Least Squares	F-statistic:	2.946
Date:	Sat, 03 Apr 2021	Prob (F-statistic):	0.0220
Time:	21:29:32	Log-Likelihood:	-128.78
No. Observations:	50	AIC:	267.6
Df Residuals:	45	BIC:	277.1
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0	0	nan	nan	0	0
Unemployment	0.6526	0.576	1.134	0.263	-0.507	1.812
Poverty	0.0678	0.464	0.146	0.884	-0.866	1.002
cases	-0.0539	0.210	-0.257	0.798	-0.476	0.368
Construction	0.8077	0.320	2.527	0.015	0.164	1.452
Production	0.4152	0.359	1.157	0.254	-0.308	1.138

Omnibus:	26.147	Durbin-Watson:	2.253
Prob(Omnibus):	0.000	Jarque-Bera (JB):	80.107
Skew:	-1.263	Prob(JB):	4.03e-18
Kurtosis:	8.663	Cond. No.	inf

Figure 26: OLS Regression from panel data with the covid cases variable

According to the table above, an augmentation by 1% in the change in the unemployment rate between 2016 and 2020 will result in a 0.651% increase in the share of votes of Donald Trump. We notice that the coefficients of the variables: Unemployment, Poverty, Production and construction changed significantly between the first (figure25) and the regression above (figure26). In addition, we have a significant coefficient for the variable Cases. In fact, the pandemic had an important influence on all variables and specifically on the variable unemployment. Besides, the more the number of Covid cases are high in a state the less the share of votes for Trump is.

6.7 Non linear regression

In this part we are going to check if the impact of a change in the employment rate had the same results on the shares of the votes or it depends on the level of the change. so we'll test the linearity of not unemployment on Y.

6.7.1 Visualisation

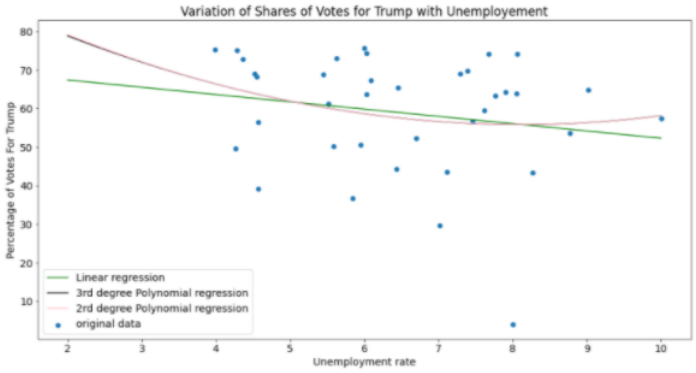


Figure 27: Linear vs nonlinear regression model of Unemployment

From this plot, we notice a big similarity between the different models. We can't conclude if a nonlinear model will perform better than a linear one. In fact, the polynomial fits better in the head and tail of data but in the middle we can see that the three models are almost superimposed.

In addition, according to the figure, the 3rd degree and 2^{nd} degree polynomial also give the same curve.

6.7.2 Computation

In order to confirm what we observed in the figure above (figure27) , we computed the different error and compared the four models:

```
For the linear regression we obtained
Mean Absolute Error: 10.302369113816177
Mean Squared Error: 212.5250586531978
Root Mean Squared Error: 14.578239216489685

********************
For the 2nd degree polynomial regression we obtained
Mean Absolute Error: 9.688848156057105
Mean Squared Error: 202.14403630975707
Root Mean Squared Error: 14.217736680279216

********************
For the 3nd degree polynomial regression we obtained
Mean Absolute Error: 9.689332352948277
Mean Squared Error: 201.97594714905682
Root Mean Squared Error: 14.211824202017727
```

Figure 28: Error computation for the different regressions

To handle this problem and make an accurate choice we opted finally for the hypothesis testing

6.7.3 Hypothesis Testing

if we consider

$$H0 : \beta_{Unemployment^2} = 0$$

$$H1 : \beta_{Unemployment^2} \neq 0$$

OLS Regression Results

Dep. Variable:	Percentage_votes_Donald_Trump	R-squared:	0.860
Model:	OLS	Adj. R-squared:	0.837
Method:	Least Squares	F-statistic:	36.97
Date:	Sat, 03 Apr 2021	Prob (F-statistic):	5.75e-16
Time:	21:29:33	Log-Likelihood:	-157.32
No. Observations:	50	AIC:	330.6
Df Residuals:	42	BIC:	345.9
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	26.7569	27.545	0.971	0.337	-28.832	82.346
x1	-19.3685	13.682	-1.416	0.164	-46.980	8.243
x2	3.0516	2.239	1.363	0.180	-1.467	7.570
x3	-0.1707	0.116	-1.466	0.150	-0.406	0.064
x4	0.5482	0.439	1.248	0.219	-0.338	1.434
x5	0.4438	0.395	1.122	0.268	-0.354	1.242
x6	3.4146	0.386	8.842	0.000	2.635	4.194
x7	1.8302	0.242	7.558	0.000	1.342	2.319

Omnibus:	2.213	Durbin-Watson:	2.010
Prob(Omnibus):	0.331	Jarque-Bera (JB):	1.368
Skew:	-0.083	Prob(JB):	0.505
Kurtosis:	2.207	Cond. No.	1.34e+04

Figure 29: Polynomial regression of degree 3

From a first look, we can say that we have a concave function because $\beta_1 = -19$ so we have a negative impact but β_2 associated to Unemployment^2 is positive (equal to 3.05). As a result, we'll have a negative effect until a certain level than there will be no effect but The P-values of β_2 and β_3 associated with the Unemployment^2 and $Unemployment^3$ are equal respectively to β_2= 0.18 and with β_3=0.15. At a 5% level we accept the hypothesis thatβ_2 =0 Same thing for β_3=0

Test both coefficient together
Now, let's test the joint hypotheses

$$H0 : \beta_{Unemployment2} = 0 and \beta_{Unemployment3} = 0$$

$$H1 : \beta_{Unemployment2} \neq 0 or \beta_{Unemployment3} \neq 0$$

We can conclude that the F statistic is equal to 1.37 with a P value of 0.264.
Hence, we can't reject the hypothesis H0 at 5% of times.
=> The hypothesis that the regression for the variable unemployment is not linear(polynomial of degree up to 3) is rejected at the 5% significance level against the alternative that it is a linear regression.

6.7.4 Interaction:

6.7.4.a Unemployment and poverty

Are less unemployment rates more effective to gather more votes when there is less Poverty? Are there any nonlinear interactions between Unemployment and poverty ? Before Computing, our reasoning consisted in considering the people that are rich and are unemployed(heirs/drug dealers). Those persons are not too much concerned by the election. As a consequence, that we can assume that there is a nonlinear relation between the two variables Let us confirm our expectations by considering the following model:

$$PercentageDonaldTrump = \beta_0 + \beta_1 Unemployement + \beta 2 Poverty + \beta 3 (Unemployment * Poverty) +$$

$$\beta_4 cases + \beta_5 Construction + \beta_6 Production + U$$

$$\frac{\Delta unemployment}{\Delta Poverty} = \beta_1 + \beta 3 Poverty$$

With β_3 is the effect of Unemployment from unit change in Poverty

OLS Regression Results

Dep. Variable:	Percentage_votes_Donald_Trump	R-squared:	0.860
Model:	OLS	Adj. R-squared:	0.840
Method:	Least Squares	F-statistic:	43.86
Date:	Sat, 03 Apr 2021	Prob (F-statistic):	9.30e-17
Time:	21:29:33	Log-Likelihood:	-157.47
No. Observations:	50	AIC:	328.9
Df Residuals:	43	BIC:	342.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-24.3484	15.941	-1.527	0.134	-56.497	7.800
Unemployment	0.8610	2.277	0.378	0.707	-3.730	5.452
Poverty	2.2561	1.168	1.932	0.060	-0.098	4.611
UnemploymentPoverty	-0.2335	0.146	-1.595	0.118	-0.529	0.062
cases	0.4233	0.395	1.072	0.290	-0.373	1.220
Construction	3.3432	0.382	8.742	0.000	2.572	4.114
Production	1.8385	0.240	7.647	0.000	1.354	2.323

Omnibus:	1.238	Durbin-Watson:	1.912
Prob(Omnibus):	0.538	Jarque-Bera (JB):	1.042
Skew:	-0.125	Prob(JB):	0.594
Kurtosis:	2.338	Cond. No.	2.13e+03

Figure 30: OLS for the interaction between unemployment and poverty

Does the coefficient of Unemployment * Poverty = 0?
P value = 0.118 which means at 5% of the cases we can not reject the Hypothesis null at those levels.
Does the coefficient of Unemployment = 0?
P-value = 0.7 which means at 10% of the cases we can not reject the Hypothesis null at those levels.
Do the coefficient of both Unemployment and Unemployment * Poverty = 0?
The F statistic = 4.49450697 with a p-value of 0.0116 so at 5% we reject the null Hypothesis

$$\frac{\Delta unemployment}{\Delta Poverty} = 0.861 - 0.2335 Poverty$$

Poverty=0 => 0.861 and Poverty=20% => $0.861 - 20 * 0.2335 = -3.809$

6.7.4.b Unemployment and number of Covid cases

Let's see if there is a non linear relation between the number of cases and Unemployment. Following the same method as before, we obtained the following results

OLS Regression Results

Dep. Variable:	Percentage_votes_Donald_Trump	R-squared:	0.853
Model:	OLS	Adj. R-squared:	0.832
Method:	Least Squares	F-statistic:	41.54
Date:	Sat, 03 Apr 2021	Prob (F-statistic):	2.49e-16
Time:	21:29:33	Log-Likelihood:	-158.63
No. Observations:	50	AIC:	331.3
Df Residuals:	43	BIC:	344.6
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.0662	7.123	-0.430	0.669	-17.430	11.298
Unemployment	-2.0638	1.119	-1.845	0.072	-4.320	0.192
Poverty	0.5141	0.434	1.185	0.243	-0.361	1.389
Unemploymentcases	-0.2379	0.347	-0.686	0.496	-0.937	0.462
cases	2.2353	2.481	0.901	0.373	-2.767	7.238
Construction	3.4173	0.394	8.678	0.000	2.623	4.211
Production	1.7550	0.254	6.920	0.000	1.244	2.266

Omnibus:	2.313	Durbin-Watson:	2.020
Prob(Omnibus):	0.315	Jarque-Bera (JB):	1.510
Skew:	-0.170	Prob(JB):	0.470
Kurtosis:	2.219	Cond. No.	253.

Figure 31: OLS for the interaction between unemployment and number of Covid cases

Does the coefficient of Unemployment * cases = 0?
P-value = 0.496 which means at 5% of the cases we can not reject the Hypothesis null at those levels.
Does the coefficient of Unemployment = 0?
P-value = 0.072 which means at 5% of the cases we can not reject the Hypothesis null at those levels.
Do the coefficients of both Unemployment and Unemployment * Poverty = 0?
The F-statistic =3.31 with a p-value of 0.0459 so at 5% we reject the null Hypothesis

$$\frac{\Delta unemployment}{\Delta cases} = -2.06 - 0.23 cases$$

Poverty=0 => -2.06 and Poverty=20% => $-2.06 - 20 * 0.23 = -6.66$ So we can decide based on the T-statistic and P-value that are not significant at 5% level that we don't have a non linear effect between Cases and Unemployment => no interaction

7 Conclusion

In summary, this study provides empirical evidence that unemployment have a significant impact on the share of Trump in the election, not only in 2016 but also in 2020. Furthermore, many other factors are related to the unemployment and turn to influence significantly the elections results especially the pandemic.

8 References

Site de documentation Python : https://docs.python.org/3/
Articles:
*)Instrumental Variables and Causal Mechanisms: Unpacking The Effect of Trade on Workers and Voters (Dippel, Gold, and Heblich 2016; Autor et al. 2017; Malgouyres 2017)
*)Economic voting: an introduction (Lewis-Beck and Paldam 2000)

*)The Economic Vote: How Political and Economic Institutions Condition Election Results, Raymond M. Duch and Randolph T. Stevenson.

https://www.greelane.com/fr/science-technologie-math%C3%A9matiques/sciences-sociales/defining-omitted-variables-bias-1146179/–

https://www.statista.com/statistics/1122759/presidential-election-effect-current-issues/statisticContainer

Sites for the Data:
https://data.world/datasets/census
https://data.fivethirtyeight.com/
https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/