Statistical Analysis of US Census Data

Akash Idnani, Ankit Sabharwal, Parth Limbachiya, Shrinand Thakkar
CSE 544 Probability and Statistics
Prof. Anshul Gandhi

May 10, 2018

1 Introduction

Our project is based on the US census data along with other data sets relating to crime and population numbers. The main purpose to take census data is to analyze and study different kinds of demographic factors that can affect the population in the United States. Given the wide variety of attributes that are available to us, we believe we can better and more holistically model how different variables affect each other. Using these variables for statistical analysis, not only gave us facts but also lead to interesting insights and provided some justification for the causes of phenomena. Moreover in the midst of charged current day politics around race, elections, gun control and population, we were keen to see what data has to say and what kind of inferences we can draw from it. Finally we can compare these inferences to what the prevalent news is in the media. Thus we set out to select the following 4 themes as our main areas of focus since they form a set of factors who's combination have the most "effect" on the citizens of the United States:

1. Electoral votes of counties: Statistical Evidence of effectiveness of Trump's electoral strategy
2. Analyzing race stereotypes through data
3. More law enforcement staff less crime? Deep dive into crime analysis from census data
4. Time Series Analysis of the Demographics of the US

We also wanted to analyze themes in as current a context as possible since we were all most aware of the recent socio-political occurrences in the USA. Therefore we chose the census dataset from 2016 which included electoral votes, education levels, race information, crime attributes, different income attributes et.al at county and state level. We chose this as our data center piece and selected the following to complement our analysis:

- crime data set for years from 2014-2016,
- a dataset which elucidated percentage of rural and urban population within counties
- dataset that contained population of the US from 1990 till 2016

Further details of the dataset and cleaning steps are described in the following section. To summarize our results we found that: Trump's strategy of pulling in the "blue collar" and "rural" county votes had proved effective. Another major finding was that the African American, Native American and Hispanic races did suffer significantly on economic, medical and educational fronts staying true to the stereotype. It was also seen that states with stringent gun laws need not have any lesser crime as compared to states with weaker gun laws. The population proportion in the older ages also seems to be increasing with a higher growth rate in the United States.

2 Dataset and Cleaning

2.1 Dataset Description

The main dataset we used was census data of 2016 along with election result of 2008, 2012 and 2016.Link of the same is here. It is county wise(3112 rows, one for each county) data which covers attributes such as race breakup, education level, median earnings, age break up, profession, crime rate, birth rate, mental health, unemployment, rural population etc.

- For first theme of electoral data, we have all the data in this dataset.
- For second theme of racial stereotype, we have all the data in this dataset.
- For third theme of crime, we only had county wise crime rate. To get law enforcement officer numbers, we looked at FBI data repository. It has county wise number of law enforcement officer over 1995 - 2016. In main data, we had only crime rate

for 2016. To get meaningful hypothesis, we got the crime rate over 1995-2016 from FBI data repository.

- For fourth theme of predicting population through time series data, we took population data over 1990-2016 from Federal Government Source.It has absolute population over these years, age wise break up and gender wise breakup.

2.2 Cleaning Steps Description

The following steps were followed to clean the data:

- Filling in missing values based on median
- Adjusting the names of the different counties while using two datasets using scripts
- Mapping different data sets on the basis of counties using scripts and excel
- We eliminated a couple of counties using Tukey's Outlier Analysis since a few attributes belonging for these counties was not plausible and unrealistic and was noise.
- Parsing scripts for different data types and converting to required format. This was used for time-series data in particular
- Manual filtering of words and changes which scripts could not handle. An example of this would be editing differently spelled county names

3 Themes and Hypotheses

3.1 Theme 1: Evidence of effectiveness of Donald Trump's electoral strategy

In this section, we try to identify if there is enough statistical evidence of effectiveness of Donald Trump's electoral strategy. This would help us to find whether Trump was able to positively influence particular groups of people like blue collar population or rural population or otherwise.

Assumptions and Definitions

- Definition: Rural County (RC): Rural county is a county in which rural population is greater than 50%.
- Definition: Blue Collar States (BC): The blue collar states can be defined as a set of people involved in farming, production and transportation and whose population in the county is greater than 25%.
- Definition: Aged counties are counties having median age greater than 40.
- Assumption: The distributions of both RC and BC counties is normally distributed
- Assumption: For paired t-test to work properly, the number of samples should be less but the data here is bit large so we overlooked that assumption and applied t-test.
- Assumption: The threshold value for deviation of given distribution from a "theoretical" normal is taken as 0.15

Verifying the Assumptions

1. Verify that the distribution of county with major rural population and distribution of counties with major Blue Collar Population, we took following steps
   (a) Plot KDE to verify that whether our data follows normal distribution or not and if it does finding the parameters of normal distribution using MLE.
      i. Plot of percentage of votes to republican by rural counties in 2016

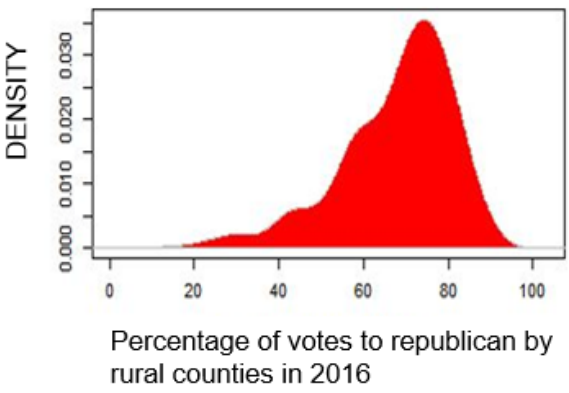

Figure 1: Percentage of votes to republican by rural counties in 2016

Since the graph represents a figure similar to a normal distribution, so we can assume that the data comes from normal distribution but we don't know it's parameters. Thus, in order to find parameters, **we applied MLE and found that the data follows normal with mean = 67.77309795 and variance = 189.858528**.

ii. Percentage difference of Republican votes in rural counties between 2012 and 2016

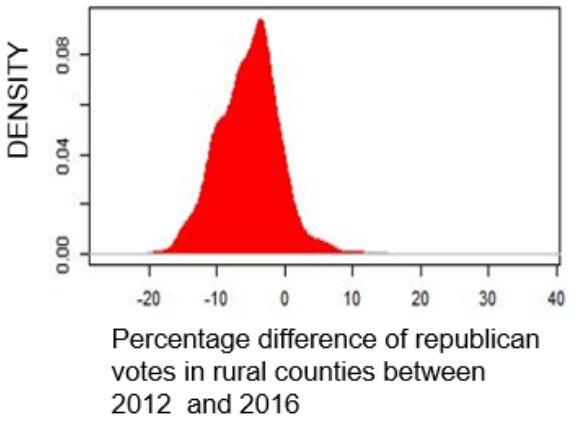

Figure 2: Percentage difference of republican votes in rural counties in 2012 and 2016

Since the graph represents a figure similar to a normal distribution, so we can assume that the data comes from normal distribution. So in order to find parameters we applied MLE to find parameters of normal distribution. Thus data follows normal with mean = -3.01366 and variance = 2.67351.

iii. Percentage of republican votes in blue collar counties in 2016

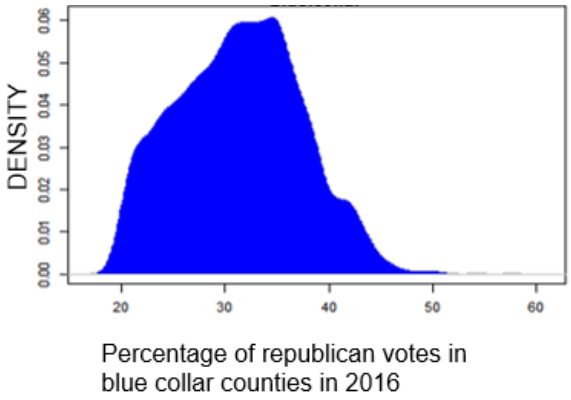

Figure 3: Percentage of votes to republican by blue collar population in 2016

Since the graph looks like normal, so the data comes from normal distribution but we don't know it's parameters. So in order to find parameters we applied MLE to find parameters of normal distribution. Thus data follows normal with mean = 64.04561686 and variance = 232.6460365.

iv. Percentage difference of number of republican votes of blue collar from '12 to '16

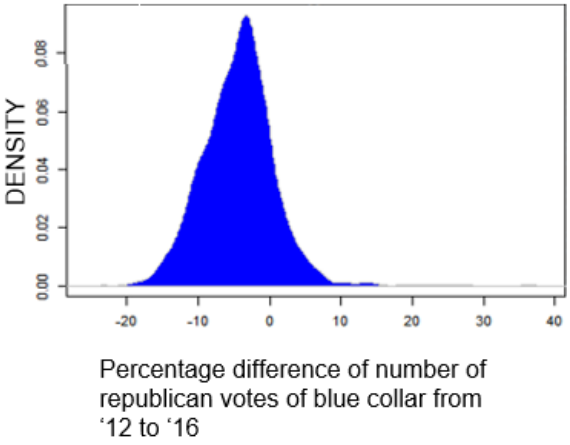

Figure 4: Percentage difference of votes to republican by blue collar population in 2012 and 2016

(b) Apply KS test to check how much our distribution varied from the "theoretical" distribution. We obtained the maximum difference D as: $\mathbf{D = 0.0283 < 0.15}$(Threshold value) for RP and $\mathbf{D = 0.0197}$ Which is less than 0.15(Threshold value) for Blue Collar.

Thus our distribution could be considered normal.

2. Paired-T test work properly for small data samples but we have huge data so we ignored this assumption and apply Paired T test.

3. T Critical One tail for $\alpha = 0.05$ is 1.645828485

3.1.1 Hypothesis 1: Votes by rural population in 2016 for Republican party is same as in 2012 i.e, Trump has no effects on Rural Population

H_0: Votes by rural population in 2016 for Republican party is same as in 2012 i.e, Trump has no effects on Rural Population
H_1: Votes by rural population in 2016 for Republican party is greater than in 2012 i.e, Trump has positive effects on Rural Population

Since we proved that both (Votes for republican in 2012 and in 2016) distributions are asymptotically normal, we can safely apply Paired T-Test

Result
T Statistic= -41.75203031 , and |-41.75203031| > |1.645828485| (Given t critical value for $\alpha = 0.005$).
Thus we a very confident result since p-value tends to 0.

Conclusion Republican party receives greater votes in 2016 than in 2012 i.e, there is strong statical evidence that Trump has positive effects on Rural Population.

3.1.2 Hypothesis 2: Votes by Blue Collar population in 2016 for Republican party is same as in 2012 i.e, Trump has no effects on Blue Collar Population

H_0=Votes by Blue Collar population in 2016 for Republican party is same as in 2012 i.e, Trump has no effects on Blue Collar Population
H_1=Votes by Blue Collar population in 2016 for Republican party is greater than in 2012 i.e, Trump has positive effects on Blue Collar Population

Since we proved that both (Votes for republican in 2012 and in 2016) distributions are asymptotically normal, we can safely apply Paired T-Test

Result
T Statistic= -40.27671115 , and |-40.27671115| > |1.645828485| (Given t critical value for $\alpha = 0.005$). Thus we a very confident result since p-value tends to 0.

Conclusion Republican party receives greater number of votes by blue collar in 2016 than in 2012 i.e, there is statical evidence that Trump has positive effects on Blue Collar Population.

3.1.3 Hypothesis 3: Votes by Aged Counties in 2016 for Republican party is same as in 2012 i.e, Trump has no effects on Blue Collar Population

H_0=Votes by Aged Counties in 2016 for Republican party is same as in 2012 i.e, Trump has no effects on Blue Collar Population
H_1=Votes by Aged Counties in 2016 for Republican party is greater than in 2012 i.e, Trump has positive effects on Blue Collar Population

Since we proved that both (Votes for republican in 2012 and in 2016) distributions are asymptotically normal, we can safely apply Paired T-Test

Result
T Statistic= -41.9401242547459 , and |-41.9401242547459| > |1.645828485| (Given t critical value for $\alpha = 0.005$). Thus we a very confident result since p-value tends to 0.

Conclusion Republican party receives greater number of votes by Aged Counties in 2016 than in 2012 i.e, there is statical evidence that Trump has positive effects on Aged Population.

3.2 Theme 2: Analyzing race stereotypes through data

In this section, we try to identify if and how different races are affected poverty, education and health. This would help us to find whether stereotypes related to race are indeed valid and also if there are alarming differences between them.

Assumptions and Definitions

- Definition: Historically Disadvantaged Group(HDG) includes African American, Native American and Hispanic races.
- Definition: White/HDG dominated state refers to any county with White/HDG population greater than fifty percent respectively.
- Assumption: The distributions of both HDG and White dominated counties is asymptotically normally
- Assumption: The threshold value for deviation of empirical distribution from a "theoretical" normal is taken as 0.15

Verifying the Assumptions

To verify that the White dominated county distribution, we first calculated the mean and standard deviation of the White dominated county distribution based on our data and we generated a normal distribution using the computed mean and variance as parameters. We applied the KS test to check how much our distribution varied from the "theoretical" normal distribution. We obtained the maximum difference D as: $\mathbf{D = 0.051437 < 0.15}$(Threshold value). Thus our distribution could be considered asymptotically normal.

Similarly to verify HDG dominated states, we applied the Permutation test. We applied it by first calculating the mean and standard deviation of the HDG dominated county distribution

based on our data and we generated a normal distribution using the computed mean and variance as parameters. We obtained the maximum difference D as: $\mathbf{D = 0.13 < 0.15}$(Threshold value)

For each of the following hypothesis relating to race, the title of each subsection refers to the null hypothesis which we try to disprove.

3.2.1 Hypothesis 1: Mean of poverty index is same for counties with majority population of HDG, and counties with majority white population

H_0= Mean of poverty index is same for counties with majority population of HDG, and counties with majority white population.
H_1= Mean of poverty index for counties with majority of HDG is more than counties with majority white population
Since we proved that both distributions are asymptotically normal, we can safely apply One tail Wald's two population test(since we are only concerned if HDG dominated states have higher poverty index)

Result

Wald Statistic	95% CI
-17.44	-1.96

Since |-17.44| > |-1.96|, we have a very confident result since p-value tends to 0.

Thus we have a very confident result since p-value tends to 0.

Conclusion Historically disadvantaged group are poorer than white population

3.2.2 Hypothesis 2: Mean of education(Less than high school population) is same for counties with majority of HDG and white dominated counties

H_0= Mean of education(Less than high school population) is same for counties with majority of HDG and white dominated counties.
H_1= Mean of education(Less than high school population) index for counties with majority of HDG is more than white majority county

Since we proved that both distributions are asymptotically normal, we can safely apply One tail Wald's two population test(since we are only concerned if HDG dominated states have higher percentage of less than high school educated population)

Result

Wald Statistic	95% CI
-17.57	-1.96

Since |-17.57| > |-1.96|, we have a very confident result since p-value tends to 0.

Conclusion Historically disadvantaged group have less education than white population

3.2.3 Hypothesis 3: Mean of HIV positive rate is same for counties with majority of HDG and white dominated counties

H_0= Mean of HIV positive rate is same for counties with majority of HDG and white dominated counties.
H_1= Mean of HIV positive rate for counties with majority of HDG is more than white majority county.

Since we proved that both distributions are asymptotically normal, we can safely apply One tail Wald's two population test(since we are only concerned if HDG dominated states have higher HIV positive rate)

Result

Wald Statistic	95% CI
-26.38	-1.96

Since |-26.38| > |-1.96|, we have a very confident result since p-value tends to 0.

Conclusion Historically disadvantaged group have significantly higher HIV positive rate

3.3 Theme 3: Law enforcement and intensity of crime

In this section, we are trying to find if state intervention is effective in curbing crime rates or not.

Assumptions and Definitions We have considered total crimes and number of law enforcement officers over 1995 to 2016. We have assumed that other factors remain constant but which might not be the case. Like over the years well being of people have increased with increase in GDP etc. And it could have negatively affected crime rate.

3.3.1 Hypothesis 1: Is there relation between number of law enforcement officers and number of crimes?

H_0= There is not a significant relationship between total number of law enforcement officers and total number of crimes(Correlation coefficient is not significantly different from zero)
H_1= There is a significant relationship between total number of law enforcement officers and total number of crimes(Correlation coefficient is significantly different from zero)

We are applying correlation test and using threshold value at |0.8|. Because in real life scenario, values greater than |0.8| suggests strong relationship.

Result
Pearson coefficient= -0.81 Thus we have very

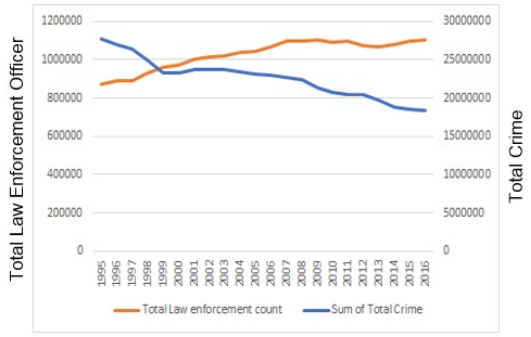

Figure 5:

strong negative correlation.

Conclusion Increase in number of law enforcement officer reduces number of crimes.

3.3.2 Hypothesis 2: Do more gun laws means less crime?

H_0= Mean of crime per thousand for state with strict gun laws is same for state with weak gun laws
H_1= Mean of crime per thousand for state with strict gun laws lesser than states with strict gun laws
We applied one tail wald's two population test.

Verifying the Assumptions Similar to previous wald's test, we have checked normality of data of two population with theoretical normal distribution by KS test. We have used parametric inference technique MLE for deriving parameter of normal distribution before running KS test.

Result for KS test:
Population 1: Crime rate for state with strict gun laws - D = 0.2138
Population 2: Crime rate for state with strict gun laws - D = 0.2157
It doesn't follow the threshold we set of 0.15 but 0.21 is not too big and we can say that it is not too much deviated from the normal distribution.

Result

Wald Statistic	95% CI
-0.7099	-1.96

|-0.7099| < |-1.96|, and p-value = 0.26.

So, we have to accept null hypothesis that there is no significant difference in crime rate for states with strict and weak gun laws.

Conclusion We can conclude that strict gun laws do not necessarily translate into lesser crime rate. That means strict gun law is not panacea to everything and federal government should look at other ways to curb crimes.

3.3.3 Hypothesis 3: Prediction of crime rate based on number of law enforcement officer

We are applying regression over data of crime rate per thousand and law enforcement officer per thousand over 1995-2016 data.

Verifying the Assumptions For regression, data has to be homoskedastic. So to check that we have plotted the graph and result is as below: Clearly data are not homoskedastic and

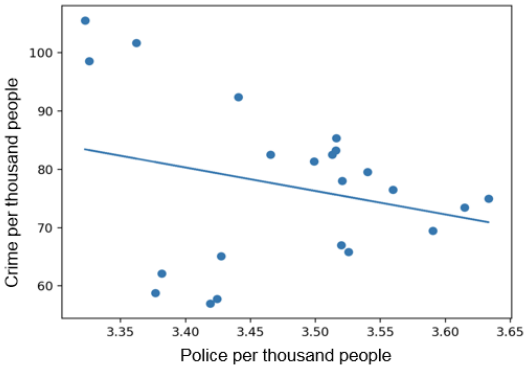

Figure 6: Data is heteroskedastic

we can't apply regression reliably.

Result
Result for regression:

MAPE - 15.04 SSE - 3813.57

As expected, results are not good and error is high. So, we cant use law enforcement officer variable to predict crime rate individually.

Conclusion We can't reliably predict crime rate from number of law enforcement officers.

3.4 Theme 4: Time Series Analysis of the Demographics of the US

In this section, we have used time series analysis techniques to model, and estimate the US population. In particular, we have used AR (Auto Regression) and EWMA (Exponentially Weighted Moving Average) techniques to model the US population (1990-2016) and to draw estimates for coming years from the trained model.

3.4.1 Hypothesis 1: Modeling the US population with an Auto Regressive model.

Test

- We have used 1990-2012, year wise total US population data, to train the AR model, and the data from 2013-2016 was used as validation data to estimate how well the model has been trained.
- Then we calculated the Average Error for **various orders (parameter p)** of the AR model to minimize the errors on predicted values, calculated on the validation data.
- We then used our trained model to project population for years, 2017-2020.

Result

Order	Error
AR(3)	**0.0445138810136**
AR(4)	**0.0346934561381**
AR(5)	**0.0381118874578**
AR(6)	**0.0366419231027**

We tried modeling AR on the data with various parameters(orders), and of which p=4 had the minimal error on the validation data.

Conclusion The Auto Regressive model projects U.S. population in 2020 at around 332,421,097. This represents an increase of 9314524, or 0.7130 percent (Growth rate) from 2016.

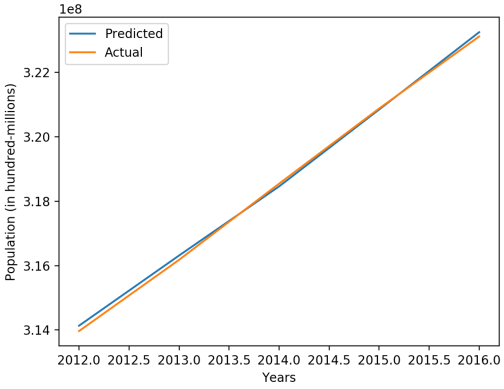

Figure 7: AR(4) on Total U.S. Population (2013-2016)

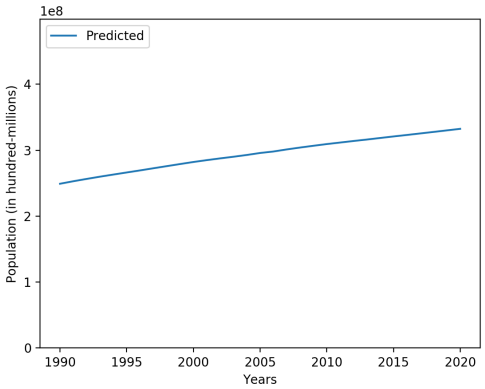

Figure 8: Total U.S. Population Projection

3.4.2 Hypothesis 2: Using EWMA to model and estimate the U.S. population of 2017.

Test

- Here also we have used 22 years-data (1990-2012), U.S. population data to train the EWMA model and validated the model with 4 years-validation data (2013-2016).
- We have computed the Average Error for various degrees of weighting decrease (parameter alpha) of the EWMA model to minimize the errors of its performance on the validation data.
- Then we used our model to project population for the year, 2017.

Result

Degree of weighting decrease	Error
EWMA(0.5)	**1.16085844871**
EWMA(0.8)	**0.855413873021**
EWMA(1.0)	**0.720175683776**

We tried modeling EWMA on the data with **various parameters(degrees of weighting decrease)**, and of which alpha=1 had the minimal error on the validation data.

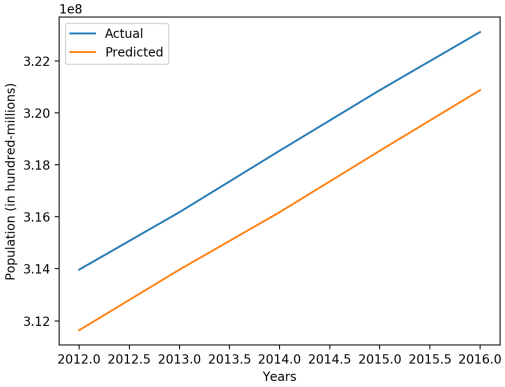

Figure 9: EWMA(1) on Total U.S. Population (2013-2016)

Conclusion The EWMA model projects U.S. population in 2017 at around 320,871,983. This represents an increase with a constant Growth rate (alpha=1) from 2016.

3.4.3 Hypothesis 3: Trend analysis of U.S. population by Age brackets

Test

- We have used the same dataset but now on the year wise U.S. population data by age brackets, to train the AR model. After that we tried fitting it on the validation data.
- We then used our model to analyze the trends of demographics of U.S. by age brackets for the years, 2017-2020.

Result

Order	Error
AR(3)	**0.1535365081715**
AR(4)	**0.114114485368**
AR(5)	**0.105355924118**
AR(6)	**0.122228242273**

We have chosen the order of AR which minimizes the average error for all age brackets as our model parameter.

Conclusion

Age Bracket	Growth Rate(%)
(0-20)	**-0.0265**
(20-40)	**0.7565**
(40-60)	**-1.2676**
(60 and above)	**2.8763**

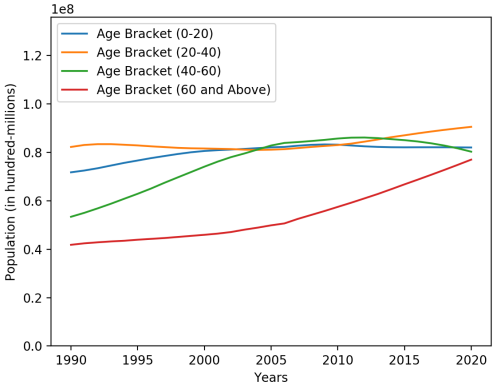

Figure 10: Total U.S. Population Projection by Age

Population in the age bracket of 60 and above is increasing rapidly which gives indication that US might be getting aged. On the other side 20-40 bracket is growing healthily which is good sign for economy.

4 Prior Work

- **Theme 1**: Various news paper agencies and think tank[1] have done research on Trump's electoral strategy. But as it is not theoretical research it was mostly opinionated and comparison of mean and visual graphs. They haven't used any statistical technique. While we used t-test and Wald's test to measure the effectiveness of Republican party's campaign statistically. Mostly, results were supporting their opinions.
- **Theme 2**: The NY Times[2] had done a study of looking at college education and race where they look at representation of the races over time at the state level. We on the other hand have looked at a county level analysis for high school level education and have combined the "minority" races into one group. Moreover we have also considered other metrics such as HIV rates and poverty to get a better understanding of the race stereotypes.
- **Theme 3**: Gifford law center[3] has done research on the relation between strength of gun law and level of crime. They are cross tabulating mean crime rate across the category(weak and strong gun law) of the states which doesn't account for standard deviation in that category. So comparing absolute mean is not good idea. We compared it through Wald's test which essentially compares their mean considering standard de-

viation. Our result contradicts their result. Moreover, they find relation only with gun crime but we tried to find relation with overall crime because accuracy of gun crime data was not good.

- **Theme 4**: ESRC Center for Population Change[4] have applied Bayesian and Monte Carlo Markov Chain time series models for population estimation of England and Wales. But we feel that population trend is fairly certain over next 4 years and easy to predict with simple models. We used models covered in class and it gave similar results.

5 Future Work

- **Theme 1**: Traditionally republican party has support base of business community who believe in free market and less social welfare scheme. But Donald Trump appealed to working class, less educated, blue collar workers. With his campaign message, he might have alienated traditional support base of republican party. We can work in that direction to check if that was the case or not.

- **Theme 2**: The future analysis in the case of education, would be to look at a time series based analysis and see if there are any trends that emerge. Another key idea would be look at the health metric's under Donald Trump's reign as president given the talks of reduction in budget. This will give an idea on how much different policies affect the counties which are dominated by different races.

- **Theme 3**: While running regression and predicting crime rate, we considered only one parameter of number of law enforcement officer. As crime depends on well being of society and many more parameters, we can include those parameters and predict crime rate more accurately.

- **Theme 4**: We only predicted population in every age bracket. It gives sign that population above 60 is increasing very fast while above 40 is also increasing. But younger population remains at constant level. We strongly feel that population in the age group of 20-40 is supported by immigrants. And if that is the case, we can make good argument against stringent immigration policies. As anti-immigrant policy could lead United States to aged country as other developed economies which were opposed to immigrants(Ex.Japan).

6 Link to Github Repository and Data Sets

Code repository link-click here

7 References

[1]PEW Research Center
[2]NY Times study
[3] Giffords Centre Research
[4]ESRC Center for Population Change