**Final Project-**

**Guns in 51 US States Analysis**

Khushboo Rathore

# **Contents**

1. [Objective 3](#_bookmark0)
2. [Variable definitions 3](#_bookmark1)
3. [Summary statistic and distributions 4](#_bookmark2)

[Summary statistic 4](#_bookmark3)

[Distributions 5](#_bookmark4)

1. [Correlation matrix 11](#_bookmark5)

[Violent crime rate across different years 12](#_bookmark6)

[Average avginc V.S average of vio across different years 12](#_bookmark7)

[Shall-carry law V.S average of vio across different years 13](#_bookmark8)

[Violent crime rate across different states 14](#_bookmark9)

1. Explonatory Variable [Expectation 15](#_bookmark10)
2. [Approach 15](#_bookmark11)
3. [Models 16](#_bookmark12)
   1. [Checking for Heteroskedasticity: 16](#_bookmark13)
   2. [Model 1 - Pooled OLS model (without cluster robust errors) 18](#_bookmark14)
   3. [Model 2 - Pooled OLS model (adjusted for cluster robust errors) 19](#_bookmark15)
   4. [Hausman test for Endogeneity 20](#_bookmark16)
   5. [Model 3 - Fixed Effects Model – Entity Fixed (adjusted for cluster robust](#_bookmark17)

[errors) 21](#_bookmark17)

* 1. [Model 4 - Fixed Effects Model – Entity and Time Fixed 22](#_bookmark18)
  2. [F-Test for significance of time variables 23](#_bookmark19)

1. [Conclusion 24](#_bookmark20)
2. [Limitations of analysis 24](#_bookmark21)

# **Problem Statement**

The main objective of our project is to analyze the data to gain insights about the understanding of how shall-issues laws affect crime rate across 51 states in U.S from 1977 to 1999.

# **Variable Analysis**

* The total number of observations in our dataset are 1173 which is spread across 51 states within 23 years.
* Violent is the dependent variable which is a qualitative variable and we have 12 independent variables which are- rob, mur, shall, incarc\_rate, density, avginc, pop, pm1029, pw1064, pb1064, stateid, year
* There is no missing data and so we can conclude that this is a balanced panel data.

|  |  |  |
| --- | --- | --- |
| **Variable** | **type** | **Definition** |
| ***Vio*** | *numerical* | violent crime rate (incidents per 100,000 members of the population) |
| ***Rob*** | *numerical* | robbery rate (incidents per 100,000) |
| ***Mur*** | *numerical* | murder rate (incidents per 100,000) |
| ***Shall*** | *Binary* | = 1 if the state has a shall-carry law in effect in that year  = 0 otherwise |
| ***incarc\_rate*** | *numerical* | incarceration rate in the state in the previous year (sentenced  prisoners per 100,000 residents; value for the previous year) |
| ***Density*** | *numerical* | population per square mile of land area, divided by 1000 |
| ***Avginc*** | *numerical* | real per capita personal income in the state, in thousands of dollars |
| ***Pop*** | *numerical* | state population, in millions of people |
| ***pm1029*** | *numerical* | percent of state population that is male, ages 10 to 29 |
| ***pw1064*** | *Numerical* | percent of state population that is white, ages 10  to 64 |
| ***pb1064*** | *numerical* | percent of state population that is black, ages 10 to 64 |
| ***Stated*** | *Character* | ID number of states (Alabama = 1, Alaska = 2, etc.) |
| ***Year*** | *Character* | Year (1977-1999) |

# 

# Summary Statistics and Distributions

## 

## **Summary Statistic-**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Median** | **Std** | **Min** | **Max** |
| ***Vio*** | 503.07 | 439.45 | 334.28 | 47.00 | 2,921.80 |
| ***Rob*** | 161.82 | 107.45 | 170.51 | 6.40 | 1,635.10 |
| ***Mur*** | 7.67 | 11.25 | 7.52 | 0.20 | 80.60 |
| ***incarc\_rate*** | 226.58 | 166.00 | 178.89 | 19.00 | 1,913 |
| ***Density*** | 0.35 | 0.08 | 1.36 | 0.00 | 11.10 |
| ***Avginc*** | 13.72 | 9.83 | 2.56 | 8.55 | 23.65 |
| ***Pop*** | 4.82 | 3.92 | 5.25 | 0.40 | 33.15 |
| ***pm1029*** | 16.08 | 17.59 | 1.73 | 12.21 | 22.35 |
| ***pw1064*** | 62.95 | 54.90 | 9.76 | 21.78 | 76.53 |
| ***pb1064*** | 5.34 | 8.50 | 4.89 | 0.25 | 26.98 |

Frequency

Frequency 100

0

100

200

300

0

50

**Distributions-**

***rob***

***vio***

**Distribution**

**Variable**

Robbery Crime Rate per 100,000 population (BJS)

1500

1000

500

0

Violent Crime Rate per 100,000 population (BJS)

3000

2000

1000

0

150

200

***shall***

***mur***

1,000

800

600

frequency

400

200

0

1

0

Murder Crime Rate per 100,000 population (BJS)

80

60

40

20

0

Frequency 200

0

100

300

400

***density***

***incarc\_rate***

10

8

6

density

4

2

0

72-99 ONLY - Lagged Rate per 100,000 resident pop of sentenced prisoners in Stat

2000

1500

1000

500

0

Frequency

Frequency

0

200

400

600

800

1000

0

100

200

300

***pop***

***avginc***

40

30

20

pop

10

0

25

20

15

avginc

10

5

Frequency 200

Frequency 50

0

100

300

400

0

100

***pw1064***

***pm1029***

pw1064

80

60

40

20

pm1029

22

20

18

16

14

12

Frequency 100

Frequency

60

0

50

150

200

0

20

40

80

100

Frequency 100

0

50

150

200

***pb1064***

0

5

10

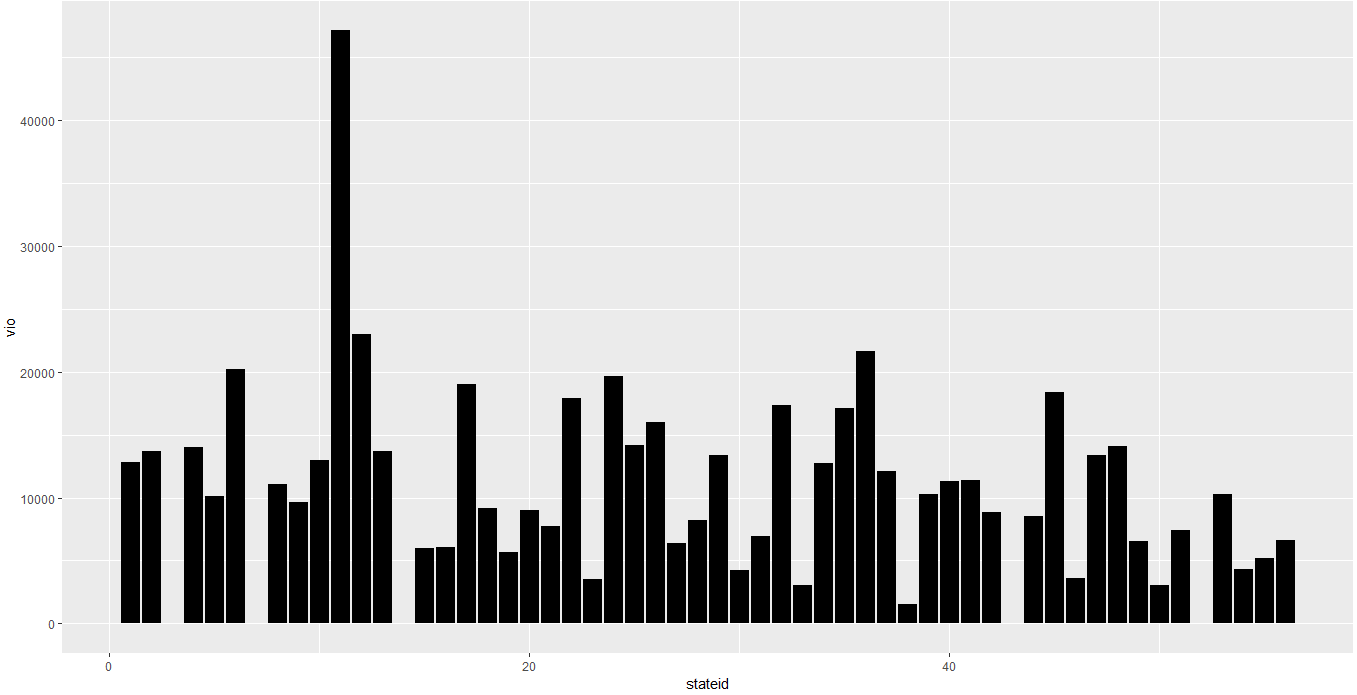
15

pb1064

20

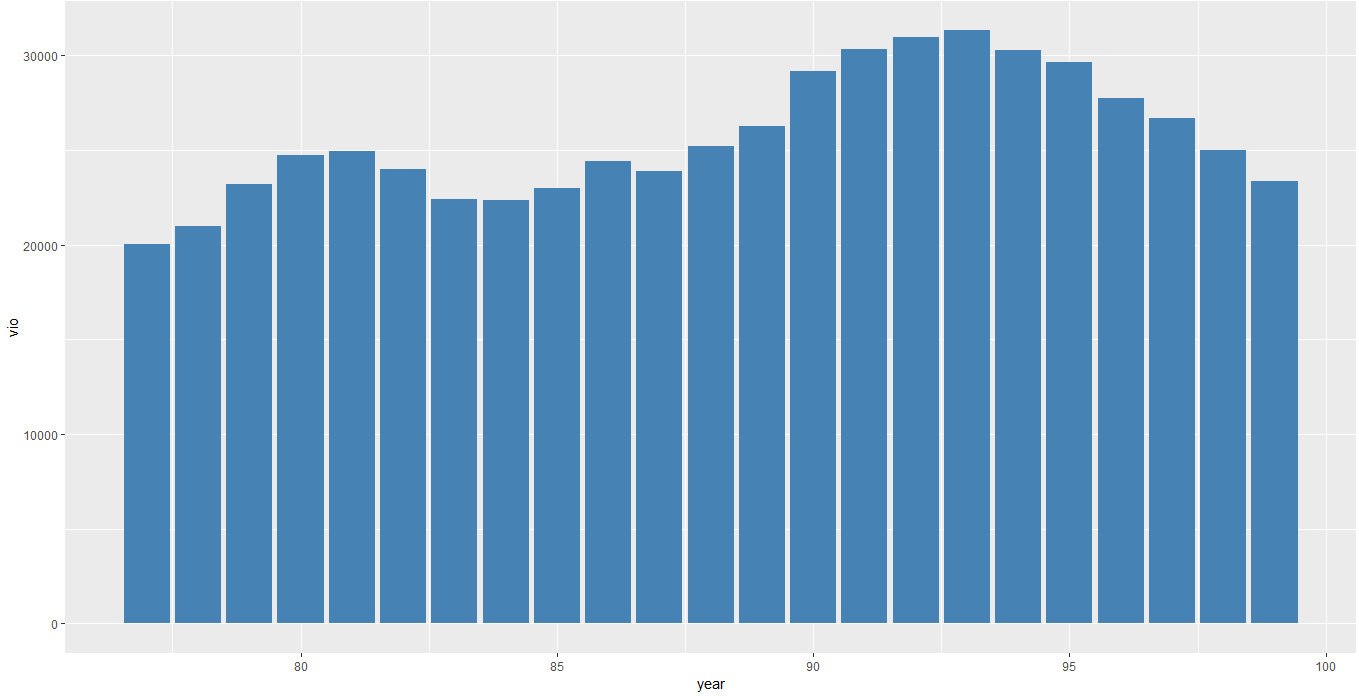
25

**Plot of stateid vs violence**



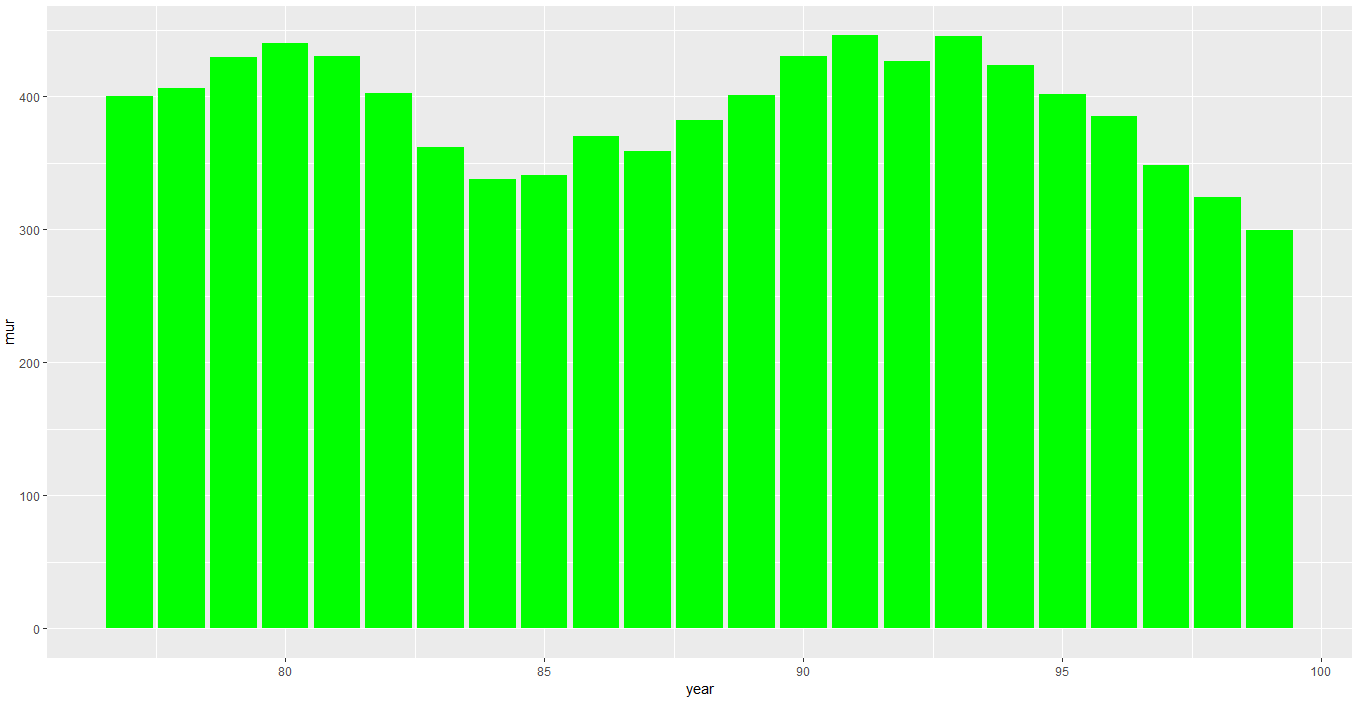
On plotting sum of violence against each stateid, we find that stateid 11 has the maximum violence compared to all the other states. Also, state with id 38 has the least violence rate.

**Plot of Year vs Violence**



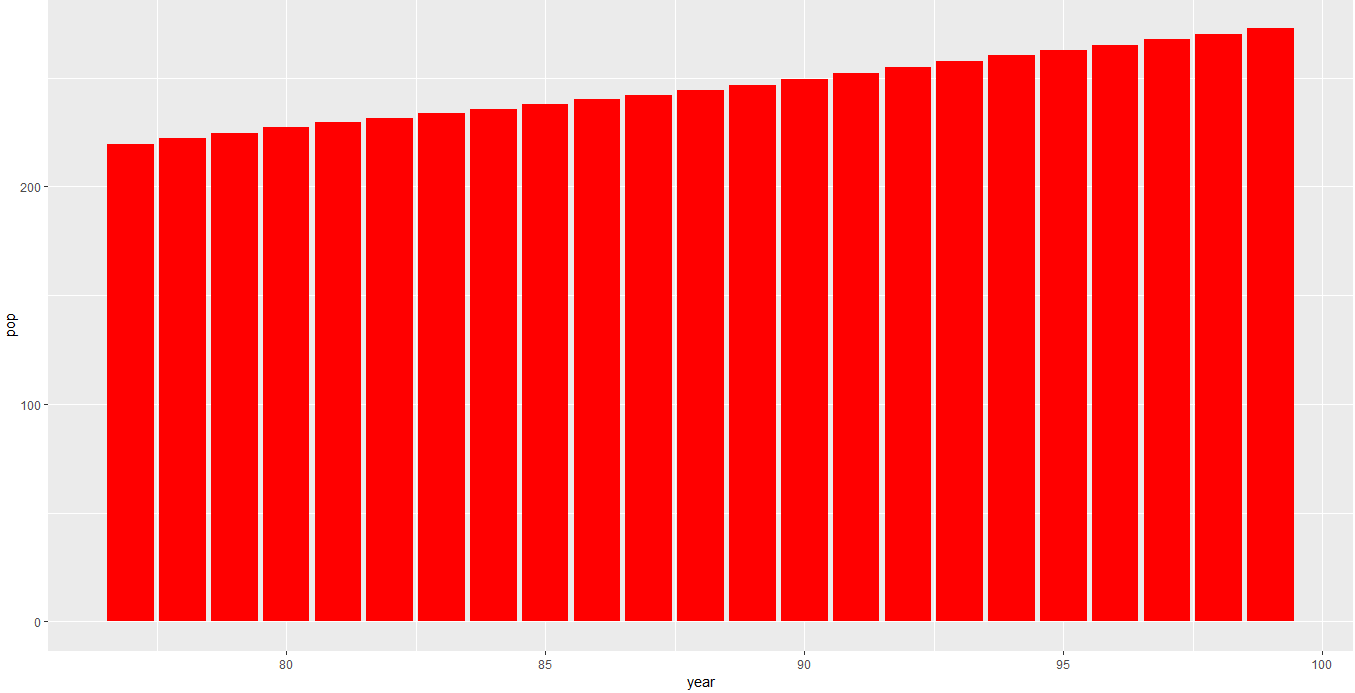
We can see from the figure above that year 1993 had the highest number of violence cases and year 1977 has the least violent cases.

**Plot of Year Vs Murder**



On plotting sum of murder against each year, we find that year 1991 had the highest number of Murder cases and lowest was in 1999.

**Plot for Year Vs Population**



On plotting sum of population against year we see that population has increased with every passing year.

# 

# **4. Correlation Matrix**

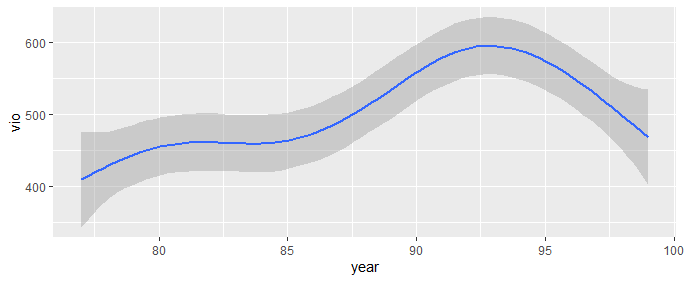
# 

**INFERENCE-**

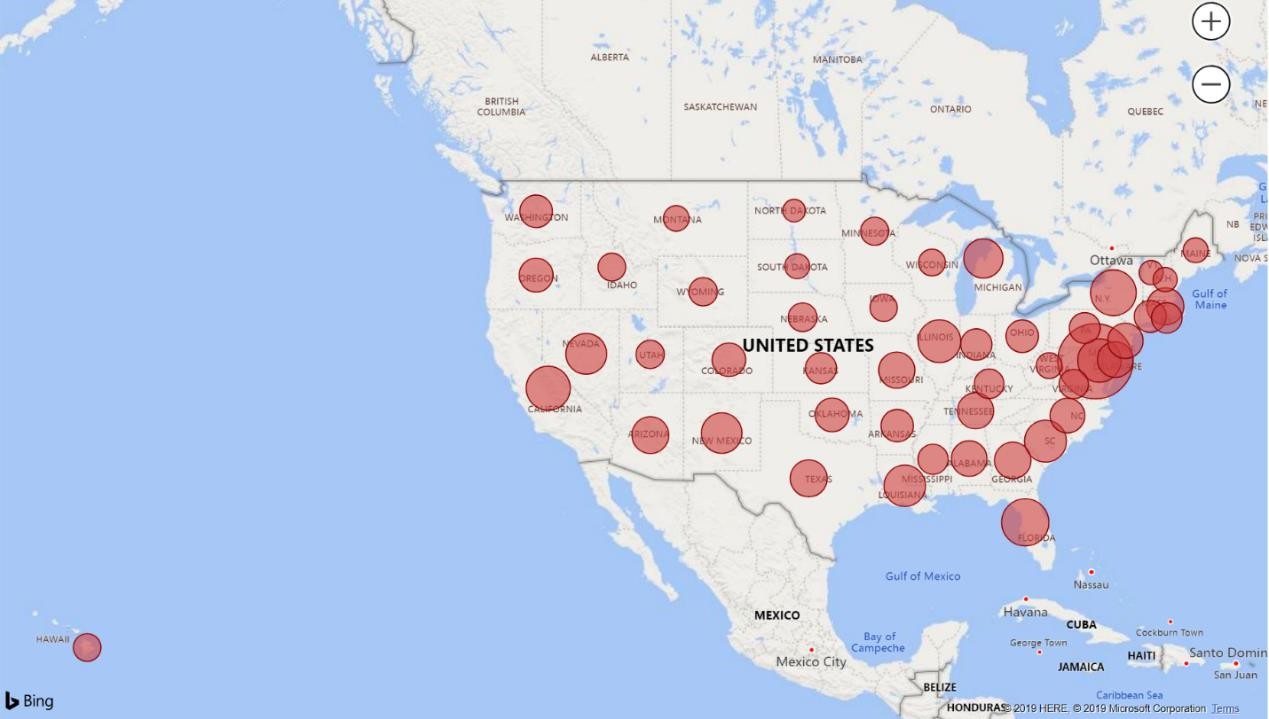
We have plotted a correlation matrix to find the correlation between explanatory variable (Violent Crime Rate) and quantitative explanatory variables if they have a positive or negative or no association. In some case, the increase in some variables can lead to increase in the crime rate leading to a positive correlation, similarly if the crime rate decreases with the rise in some variables then they have a negative correlation. We can come to know which the variables that influence our dependent variable are and how strongly which can help us draw important conclusion. We have found the following inferences about our variables after studying the correlation matrix-

* Areas that have higher density of population tend to have a higher crime rate.
* Higher the population of black people, higher the crime rate.
* Higher percentage of white populations are more likely to live in areas which tend to have lower population density.
* Higher the population of white people, lower the crime rate.
* When the percentage of number of males increase, crime rate decreases.

**-Violent crime rate across different years**



**-Violent crime rate across different states**





# **Explanatory Variables Expectation-**

* 1. **Rob**: It is natural for the crime rate to increase as the number of robberies increase as robbery is a crime which would be exponentially increasing and hence we expect “Rob” is positive.
  2. **Mur:** As the number of murders increases, the crime rate would also increase as murder is a crime by itself and hence we can expect “Mur” to be positive.
  3. **Shall:** This is the variable on which we have an argument claim that if the state allow citizens to carry concealed guns, the violent crime rate would decrease. As there is a negative association between “crime rates” and “shall” we expect “shall” to be negative.
  4. **incarc\_rate:** As the new law will be enforced, people will be less likely to involve in criminal activity and so we can expect this variable to be negative.
  5. **Density**: If the population of an area is higher, they are more likely to have a higher criminal rate as there are more chances of criminals as compared to areas with less density of population and so we can expect “Density” to be positive.
  6. **Avginc:** People who have low income area more likely to involve themselves in crime to get some more money, hence “Avginc” can be expected to be a negative as when the income of people in an area is low the crime rate is expected to be higher.
  7. **Pop**: As the population increases, there can be a chance for the crime rate to increase or decrease deepening on the type of people, hence we cannot expect anything about “Pop” variable with much confidence.
  8. **Pm1029:** The “Males” are more likely to be involving in illegal/criminal activities and hence we can expect this variable to be positive.
  9. **Pw1064:** As the percentage of state population (white) goes up, we expect the violence rate to go up. However the expected signs of variables are debatable.
  10. **Pb1064:** The Blacks are more likely to be involved in illegal/criminal activities and hence we expect this variable to be positive.
  11. **Year**-After observing the trend of crime rate over the span of years we can infer that though more number of states have enforced the shall-law with time but the overall crime rate has slightly increased over the years i.e. from 1977 to 1999

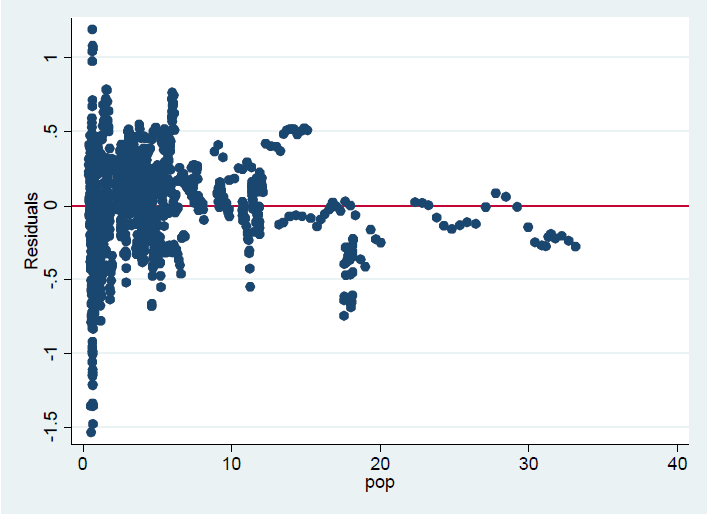
# 

# **6. Problem Statement Approach**

1. We conduct explanatory data analysis which is scatterplots, heatmaps, correlation matrix and other statistical and distributions analysis.
2. Check for the presence of Heteroscedasticity using the “**White Test for Heteroscedasticity”.**
3. Check for the presence of Heteroscedasticity by plotting the **“Residuals vs Fitted”** plot.
4. Applied the Pooled ordinary least squares on our dataset without cluster robust error.
5. Run Pooled ordinary least squares with cluster robust error on our dataset.
6. Use the **Hausman** test to check for the **“Endogeneity Problem”** in our dataset**.**
7. Run the Entity Fixed Effect model with cluster robust error on the dataset.
8. Run the Entity and Time Fixed Effect model with cluster robust error on the dataset.
9. Compare the outcomes after running each of the models to come to the conclusion of which is the best model for our dataset.

# **Models**

## **Step 1-Checking for Heteroskedasticity using Residual Plot:**



**INFERENCE-**

1. After plotting all the dependent variables against the residual values, we found that the variable “population” has heteroskedasticity.
2. Therefore, the standard errors of the pooled OLS without robust standard errors are incorrect.
3. Since pooled OLS model ignores the panel nature of the data, we cannot trust the results of this model.
4. Coefficient sign: Density, pm1029 – expected to be positive but model shows negative. Incarrc\_rate expected to be negative, but model shows positive.

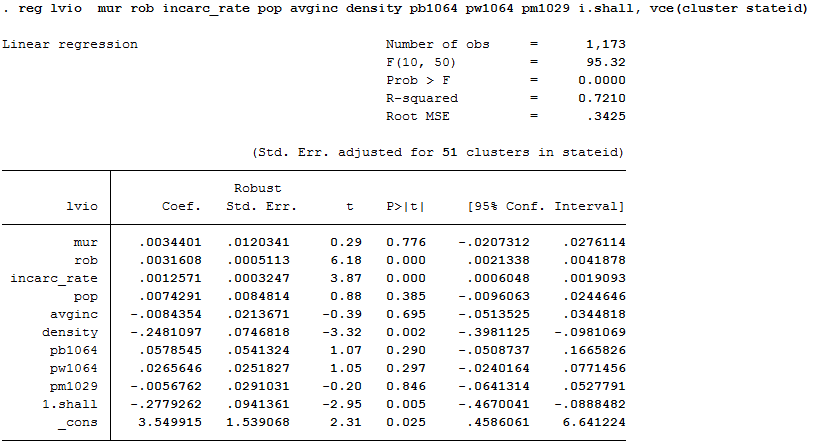
## **Model 1- Pooled OLS model Without Cluster Robust Standard Errors**

## 

### INFERENCE:

* + 1. The coefficient of Shall equals -0.277, which suggests that shall-issue laws reduce the violent crime rate by 28%, for every percent increase in Shall.
    2. This is a huge decline in the crime rate and the result is highly significant result as well and we can infer this by the P value.
    3. As expected, all the other variables are positively contributing to crime except pm1029, density, avginc etc. All the variables are statistically significant at 5% significance level except pm1029, avginc and mur
    4. We will now run a regression with the Robust Cluster Standard error to remove the problem of heteroskedasticity to get accurate standard errors, inefficient, unbiased and consistent but our estimates will still be inefficient

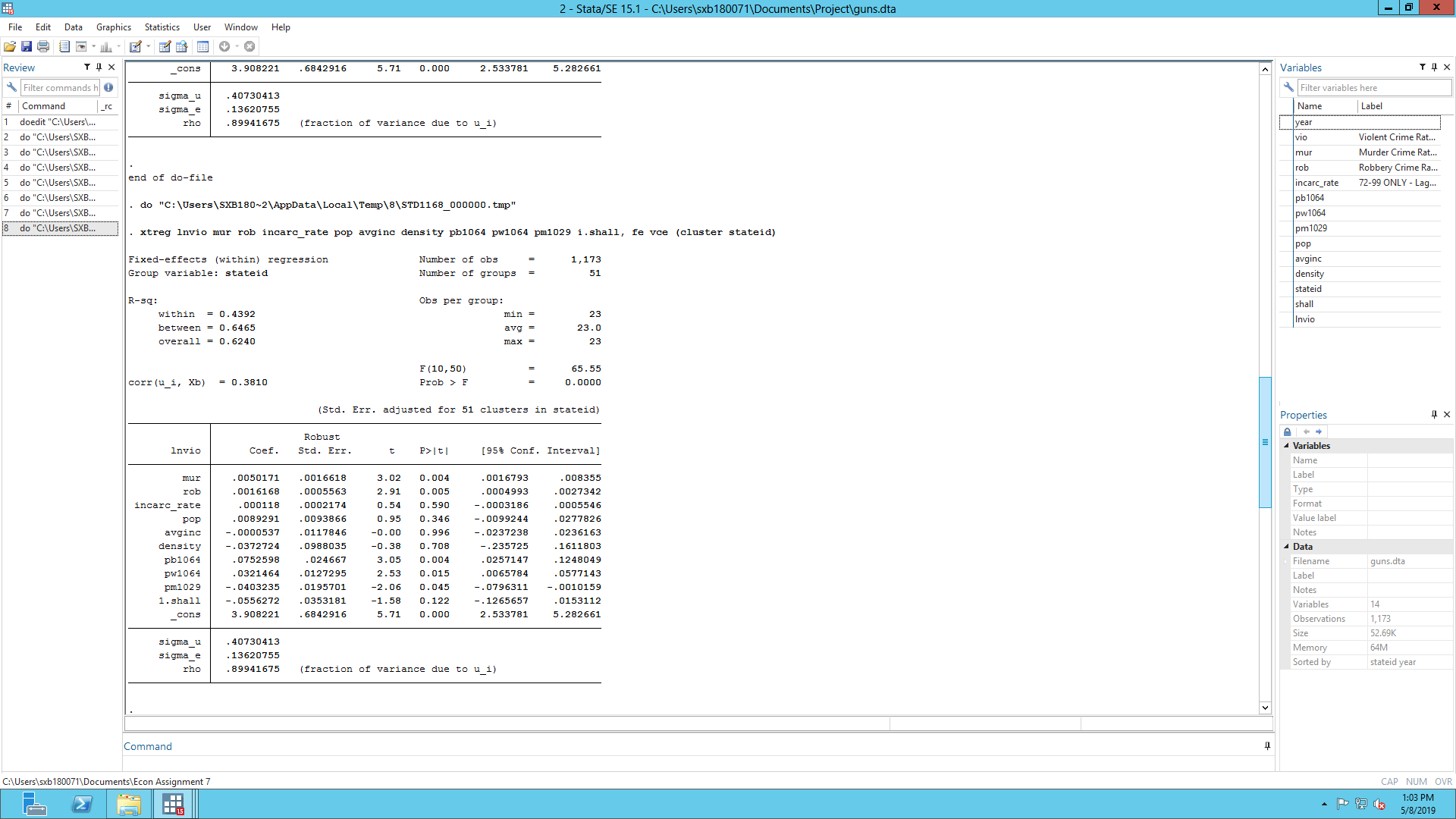
Model 2- Pooled OLS model with Robust Cluster Standard Error



### INFERENCE:

1. As we can observe there is a significant increase in the standard errors of the Shall variable which have increased to 0.094 from 0. 0264.This is a strong indication for the presence of heteroskedasticity in the data.
2. Other variables also show a significant increase in standard errors.
3. There can be a problem of endogeneity in our model as there could be some unobserved heterogeneity i.e. omitted variables bias or simultaneous causality bias which got added in the error term and will lead to biased and inconsistent estimators. One such omitted variable could be attitude of people towards driving.
4. There is large difference in standard errors of the pooled models with and without using robust standard error. This implies that there may be heteroskedasticity and/or auto-correlation. This can be tested using White-test.
5. We will proceed with fixed effects model as that will confirm the existence of endogeneity, however, we show endogeneity using Hausman test as well. We reject the null hypothesis of no endogeneity.

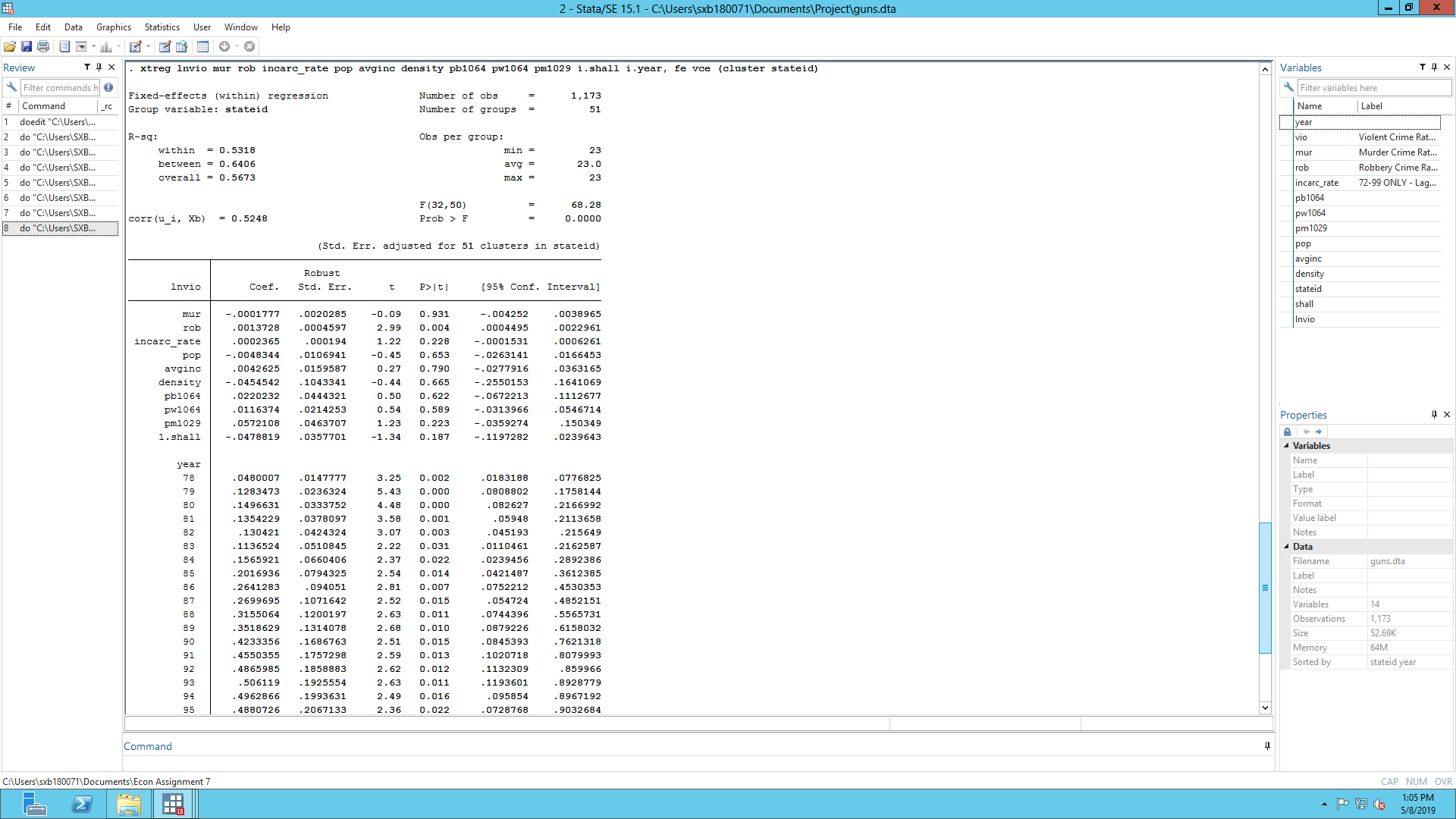
Model 3- Fixed Effects Model – Entity Fixed (adjusted for cluster robust errors)

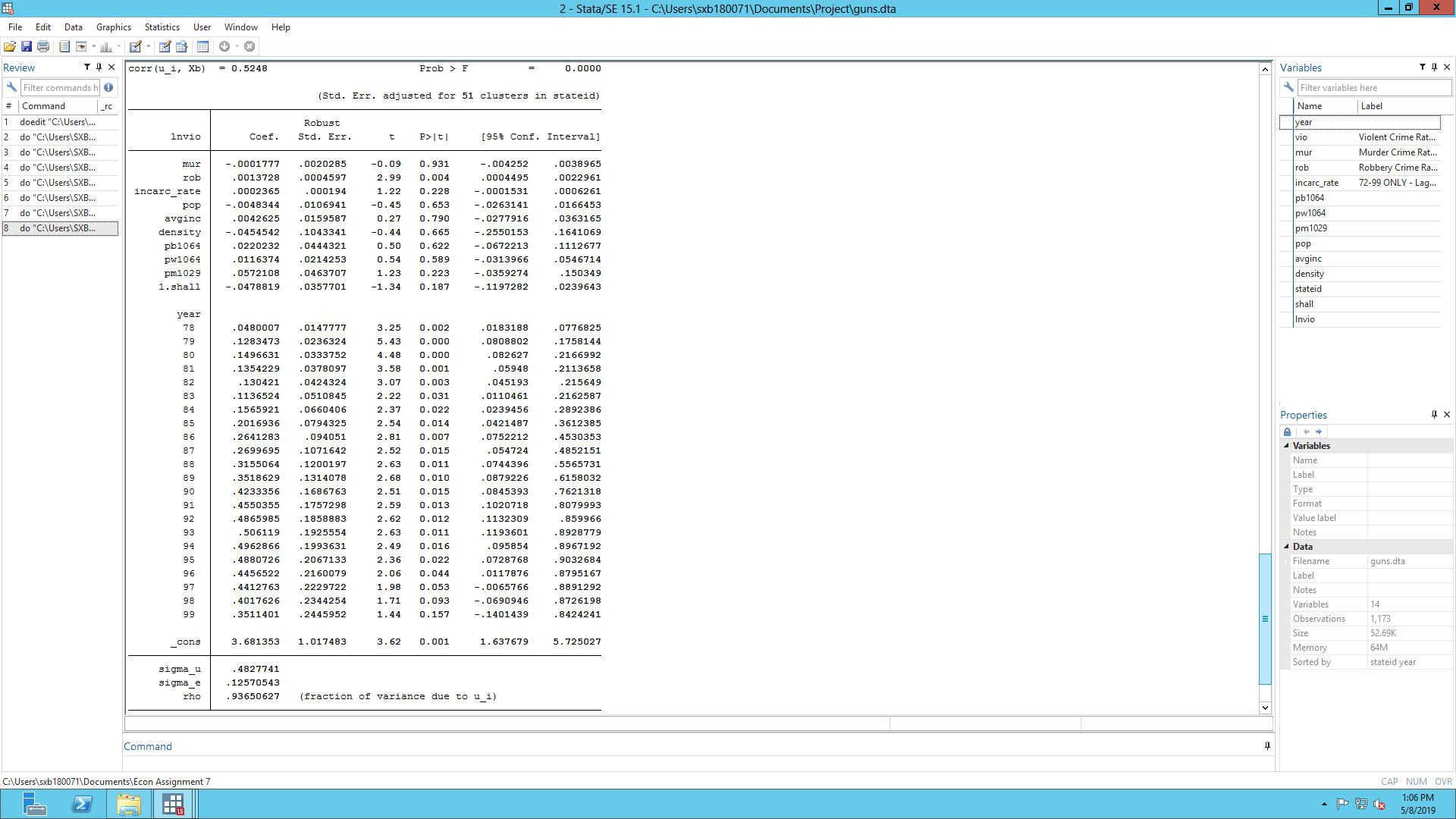


**INFERENCE-**

* The results change when we run the fixed effects model with fixed states.
* The absolute effect of Shall decreases to 5.6% from 27.79% which is a significant drop. The effect of shall issue laws on the violent crime rate is no longer statistically significantly different from zero due to insignificant p values and this makes our estimate not very reliable.
* This can be due to unobserved heterogeneity i.e. omitted variable bias which could have caused to overstate the effect of State in the previous model
* The regression model with fixed effects is more credible because this controls for unobserved heterogeneity that vary between states but that are constant over time.

## **Model 4- Fixed Effects Model – Entity and Time**



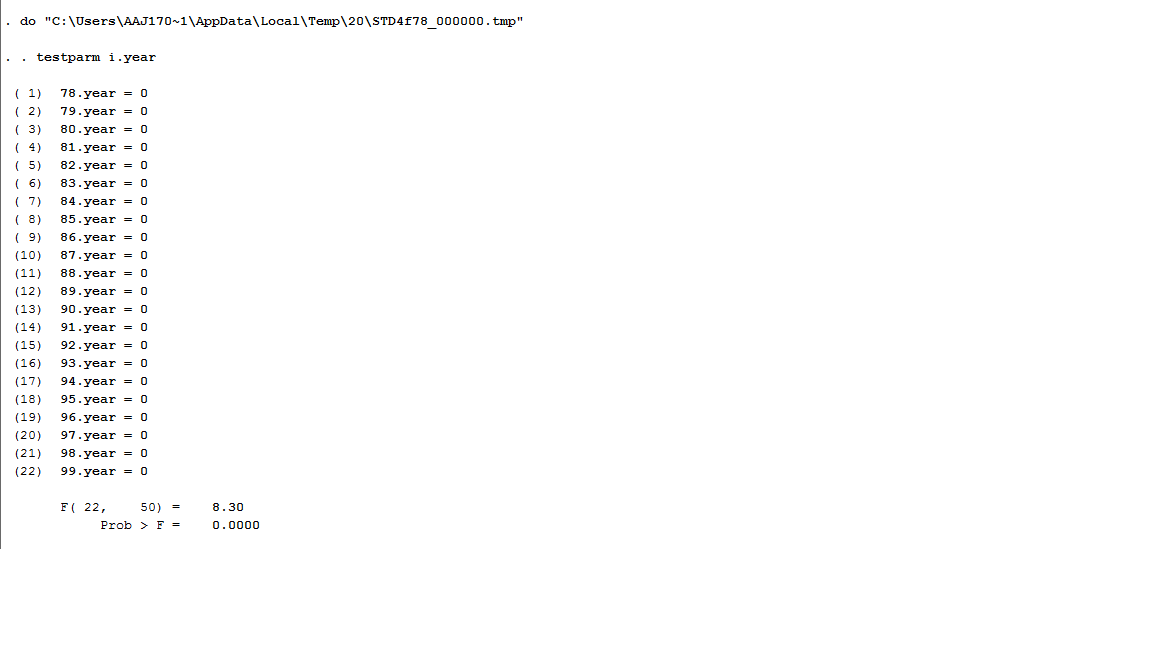


**INFERENCE-**

* The results change slightly when we run the fixed effects model with both states and time fixed as compared to our previous model.
* There is a slight change in the effect of shall-carry law on violence rate further reduced from 5.5% to 4.8% in the entity and time fixed effect model.
* The coefficient is still not significantly different from zero as and we can infer this from the P values.
* After running the test to check if the time variables are statistically significant, we can infer that the time variables are significant. Hence we can infer that the time and state fixed model is a better fit than the only state fixed model.

## 

## **F-Test for significance of time variables**

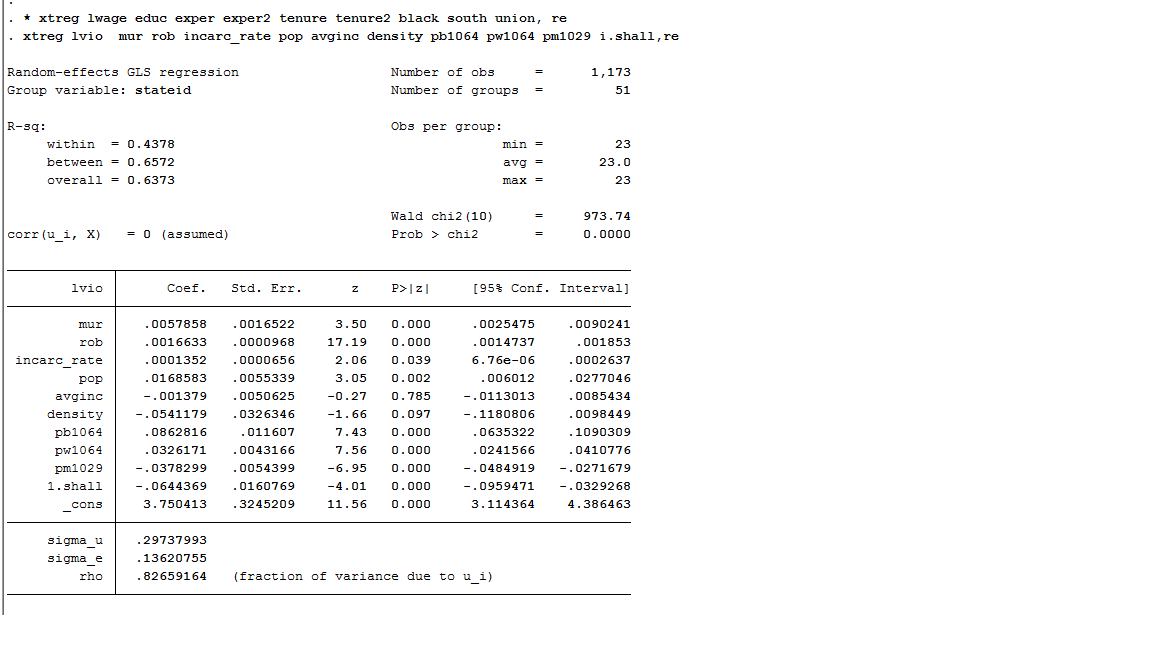


**INFERENCE-**

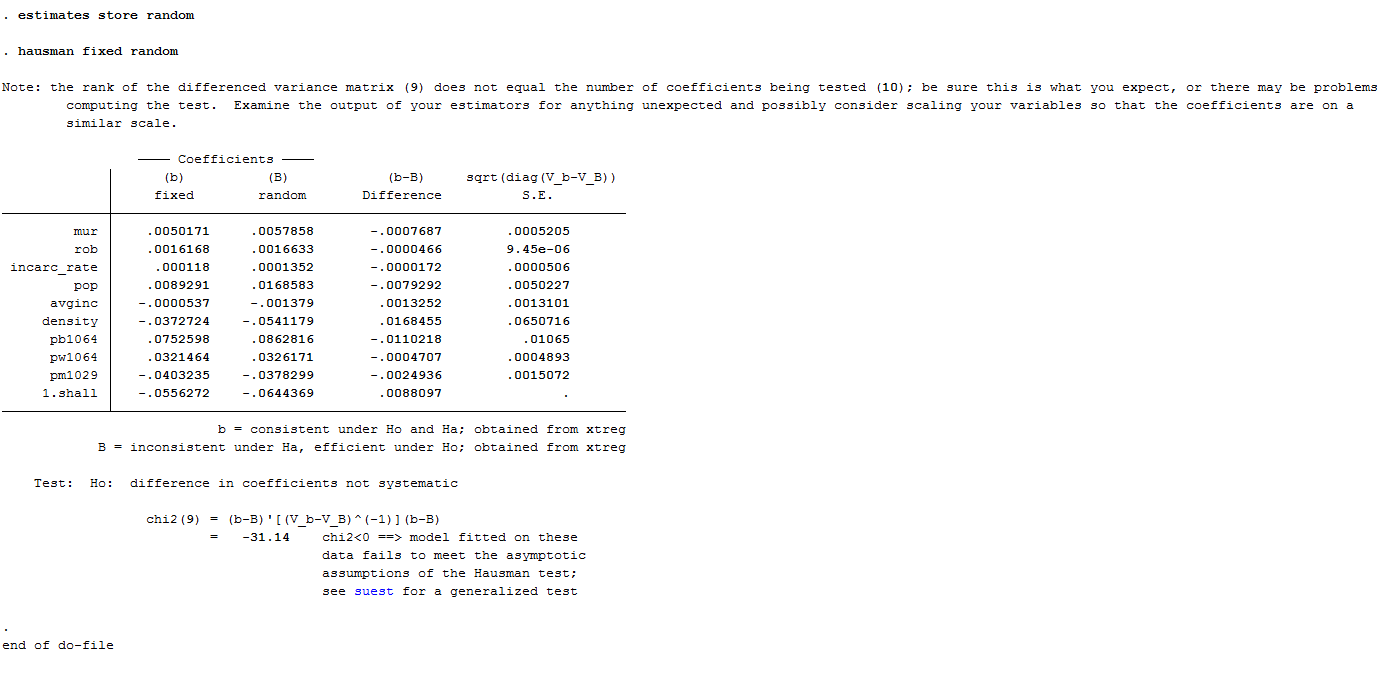
As we can see that the F stat-value is very high, hence we can infer that time variables are statistically significant and should be kept in our model for better results.

**Model 5- Random Effect Model**

The data set used includes data from all the states therefore it is not random data. However, we can prove that random effect model is not appropriate for this data using Hausman Test.



**Hausman Test**



**OBSERVAION-**

We can see from the above table that there are significant differences in co-efficient between the fixed effect and random effect models. Therefore, we can conclude that random effect model is not appropriate for this data.

**CONCLUSION-**

Based on the above analysis, we can conclude that entity and time fixed effect model is the best.

**LIMITATION-** However, there are some limitations which are:

* Simultaneously causality bias: the variable incarc\_rate (incarceration rate) and violence rate effect each other causing bias in the estimate of incarc\_rate.
* Inefficient: estimates are computed only using variations within entities.
* Time variant factors: Unobserved heterogeneity which is time variant is not controlled in the fixed effect model. This causes biased results.

# **Conclusion**

* There is a large estimated effect of concealed weapons laws in pooled OLS models.
* This effect is however due to omitted variable bias and unobserved characteristics because the effect disappears when state and time effects are added.
* So the model with both time and state fixed effects is the best model.
* We can conclude that there is no significant effect of concealed weapon laws on the violent crime rate.

# **Limitations of analysis**

* Even the FE estimate could have following bias:
  + It captures only within variation
  + It captures only variables which are constant
* FE estimator is still biased if the unobserved heterogeneity changing over time, and correlated with the regressors.
* This means that if some variable such as attitude of people towards driving changes over time then it will not be captured here.
* We will need more data to check for that effect but it is really difficult to get such data

**THANK YOU**