Yingyi Zhu

ECON3720 Introduction to Econometrics

Professor Stefan Ruediger

3 May 2022

A Study of Factors Affecting Crime Rates in the United States During the 2010s

Introduction

The crime rate is an essential indicator of the safety level in a region. Studying the crime rate will generate a better plan for allocating resources and adjusting current policies, indirectly boosting economic growth in general. However, the violent crime rate during the pandemic is challenging to compile, so I decided to examine the decade before COVID. Therefore, the research question that I intend to answer in this paper is the relationship among different explanatory variables and the crime rates in the United States between 2010 and 2019. In a paper published by NBER, the author explored the relationship between gun ownership and the homicide rate. His constructed model considered control variables, such as per capita personal income and unemployment rate, which inspired me to include these two as independent variables in my multivariate regression. The data is taken from multiple primary sources, including the FBI, BEA, BLS, and FRED. I also incorporated the OLS model with the White and RESET tests to prove and evaluate my hypothesis.

Hypothesis

I hypothesize that all explanatory variables affect the crime rate, with personal income per capita and education attainment negatively correlated with the crime rate: the unemployment rate, population density, and marijuana legitimization are positively associated with the dependent variable.

Literature Review

To understand the topic in-depth, I have looked at several other published peer-reviewed journals and found four journal articles that provide intriguing ideas for this project. Duggan's paper used a log-log regression model to analyze the relationship between gun ownership and homicide rate. Interestingly, this article gives an idea of what could be included in the error term, as gun ownership is positively correlated with the homicide rate. In another paper, Dr. Abhijeet utilized the F-test to test if urban areas have higher crime rates than rural areas. The test result indicates that the difference in districts and unemployment correlate with the crime rate. This corresponds to my hypothesis that unemployment affects crime rates. Therefore, it sparked me to add a demographics variable (population density) to test whether it would yield similar results.

I also found articles discussing the relationship between one of the explanatory variables and the standard response variable crime rate. For the variable marijuana legitimization, I read a paper by Maier that reveals a lack of relationship between crime rates and recreational and medical marijuana legal status. It is interesting to see whether my model will support the conclusion of this article or not since I suspect involvement with the marijuana legitimization variable will have a positive effect on the crime rate. Another paper by Lochner analyzes the relationship between education and participation in criminal activity. The author starts with his hypothesis similar to mine, in which schooling reduces the crime rate with justification. The study incorporates various interaction terms when calculating the effect of years of education on incarceration, which inspires me to use the interaction of education and unemployment in my research. The result shows that the OLS estimates indicate that white high school graduates have a 0.76-percentage point lower probability of incarceration than do dropouts.

Description of the Data

I used panel data collected from 50 states for this model.

Time period: annual data of 2010-2019

Dependent Variable: Crime rate of each state per 100,000 inhabitants

Independent Variables: Personal income per capita, Unemployment rate, Marijuana legitimization date, Population Density, and Educational attainment (high school degree and above)

The complete dataset will be compiled from multiple primary sources, with 500 observations. The crime rate data intended to use is mostly abstracted from the Federal Bureau of Investigation's Crime Data Explorer. The FBI collects this data through the Uniform Crime Reporting (UCR) Program, managed via the Summary Reporting System (SRS). "Crime rate" is specifically referring to the violent crime in units of incidents per 100,000 individuals per year; thus, a violent crime rate of 300 (per 100,000 inhabitants) in a population of 100,000 would mean 300 incidents of violent crime per year in that entire population or 0.3% out of the total. Personal Income per capita (an area's income in dollars divided by its people) is collected from the Bureau of Economic Analysis; the unemployment rate (the number of employed people as a percentage of the labor force) is collected from the Bureau of Labor Statistics; the Marijuana legalization year is compiled from *mjbizdaily* which gave the information in terms of when and where marijuana was legitimized in each state. Marijuana legalization is expressed as a dummy variable, where 1 represents marijuana is legitimized, and 0 represents marijuana is illegal. Education attainment data is sorted from FRED. It indicates the percentage of high school graduates or above in each state. The variable population density is calculated by hand by dividing the total population by the total area of each state.

Since the dataset is compiled from authoritative primary sources, the reliability is ensured, but structural measurement errors still might exist, as they are unavoidable. One exception would be population density. Since it is calculated by dividing state population by area, and the state population is an estimated value, the population density might not show the actual values. On the other hand, I do not suppose there are sample selection errors since the primary sources collected the data randomly based on large sample sizes and are used universally in many areas.

Planned Analysis

Multiple linear regression was conducted to examine the research question to determine if the independent variables predict the dependent variable since a multiple linear regression model can assess the relationship between a set of independent variables and one dependent variable.

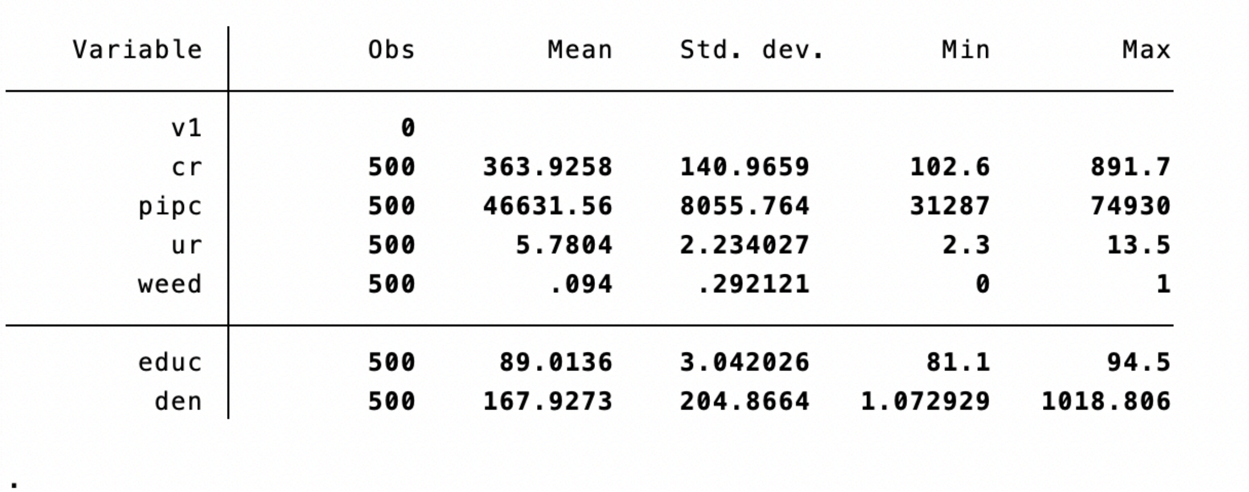
The proposed model:

Crime Rate = $\beta0$ + $\beta1$Personal Income per capita + $\beta2$Unemployment Rate + $\beta3$Weed + $\beta4$Education+ $\beta5$Population Density + u

Methodology and Model

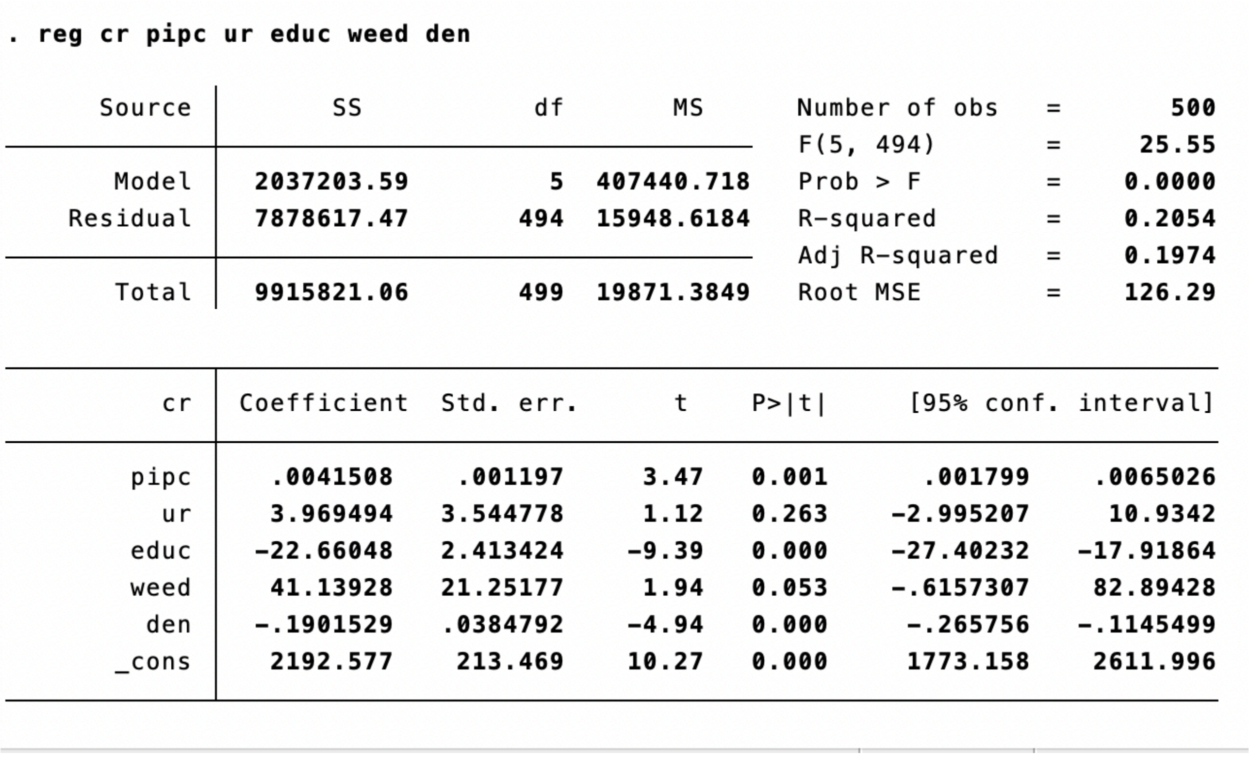
For the whole project, I used the significance level of 0.05.

Data summary:



```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
          v1 |          0
          cr |        500    363.9258    140.9659      102.6      891.7
        pipc |        500    46631.56    8055.764      31287      74930
          ur |        500      5.7804    2.234027        2.3       13.5
        weed |        500        .094     .292121          0          1
-------------+---------------------------------------------------------
        educ |        500     89.0136    3.042026       81.1       94.5
         den |        500    167.9273    204.8664   1.072929   1018.806

.
```

By running the planned model:

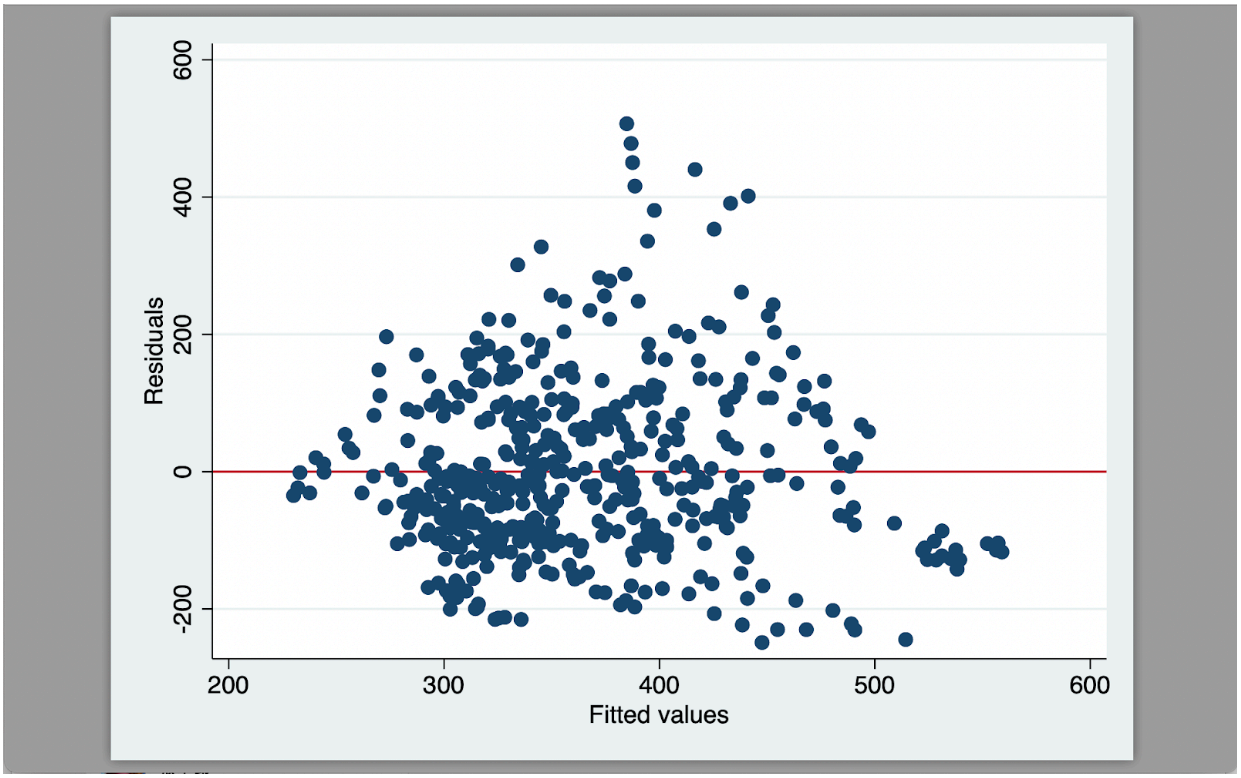


```
. reg cr pipc ur educ weed den

      Source |       SS           df       MS      Number of obs   =       500
-------------+----------------------------------   F(5, 494)       =     25.55
       Model |  2037203.59         5  407440.718   Prob > F        =    0.0000
    Residual |  7878617.47       494  15948.6184   R-squared       =    0.2054
-------------+----------------------------------   Adj R-squared   =    0.1974
       Total |  9915821.06       499  19871.3849   Root MSE        =    126.29

------------------------------------------------------------------------------
          cr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        pipc |   .0041508    .001197     3.47   0.001      .001799    .0065026
          ur |   3.969494   3.544778     1.12   0.263    -2.995207     10.9342
        educ |  -22.66048   2.413424    -9.39   0.000    -27.40232   -17.91864
        weed |   41.13928   21.25177     1.94   0.053    -.6157307    82.89428
         den |  -.1901529   .0384792    -4.94   0.000     -.265756   -.1145499
       _cons |   2192.577    213.469    10.27   0.000     1773.158    2611.996
------------------------------------------------------------------------------
```

(cr for violent crime rate, pipc for personal income per capita, ur for the unemployment rate, educ for education level, weed for recreational use of marijuana, den for population density)

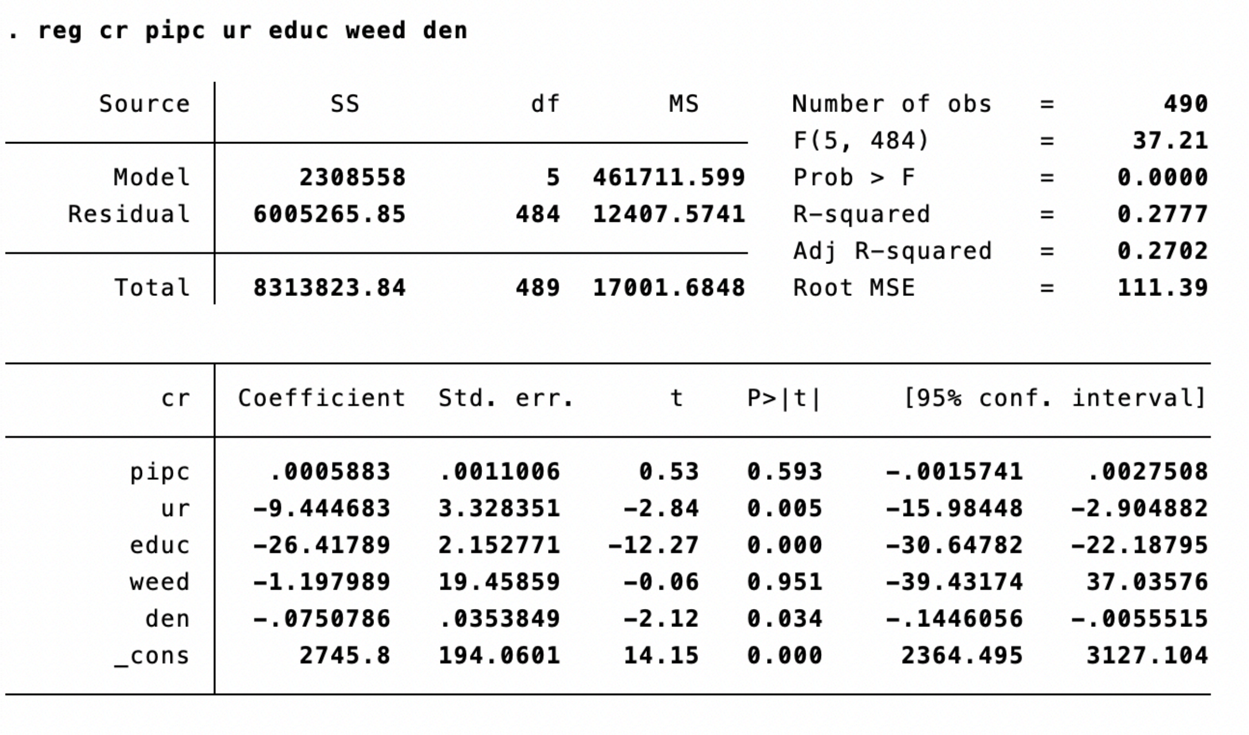
The adjusted R-square is 19.74%. Three independent variables are statistically significant, and two independent variables are not statistically significant.

Checking if the fitted residuals were uncorrelated with independent variables (potential heteroskedasticity):



There are fitted residuals with values of around 400 outliers in the upper part of the residual plot. These outliers are observations in the area of Alaska. They can be seen as outliers because Alaska is outside the continental United States and is located near the north pole. The weather condition is fatal, and residents are exceptionally sparsely distributed. The attributes of these observations are different from other continental states and are not the primary investigation focus, so I decided to remove them from the samples.

After the removal of 10 observations from Alaska, the model gives the following result:



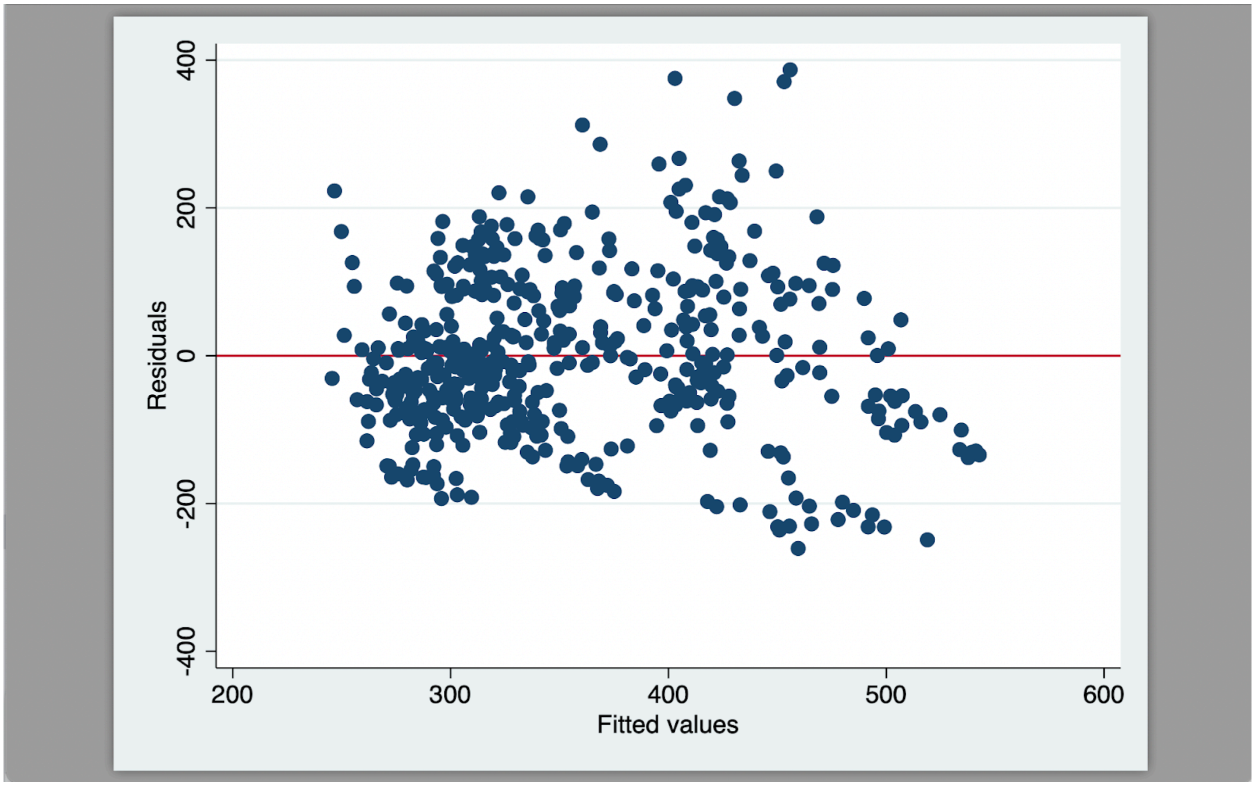
```
. reg cr pipc ur educ weed den

      Source |       SS           df       MS      Number of obs   =       490
-------------+----------------------------------   F(5, 484)       =     37.21
       Model |     2308558         5  461711.599   Prob > F        =    0.0000
    Residual |  6005265.85       484  12407.5741   R-squared       =    0.2777
-------------+----------------------------------   Adj R-squared   =    0.2702
       Total |  8313823.84       489  17001.6848   Root MSE        =    111.39

------------------------------------------------------------------------------
          cr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        pipc |   .0005883   .0011006     0.53   0.593    -.0015741    .0027508
          ur |  -9.444683   3.328351    -2.84   0.005    -15.98448   -2.904882
        educ |  -26.41789   2.152771   -12.27   0.000    -30.64782   -22.18795
        weed |  -1.197989   19.45859    -0.06   0.951    -39.43174    37.03576
         den |  -.0750786   .0353849    -2.12   0.034    -.1446056   -.0055515
       _cons |     2745.8   194.0601    14.15   0.000     2364.495    3127.104
------------------------------------------------------------------------------
```

The adjusted R-squared improved from 19.74% to 27.02%. Two variables are not statistically significant. However, I believe that the variable personal income per capita is economically significant since violent crimes are highly correlated with properties and personal income levels. The recreational use of marijuana, in contrast, is economically insignificant, also proved by other sources, which contradicts my original belief. "a federally funded study found that legalizing marijuana has little to no impact on rates of violent or property crime." (Jaeger, 2021) Thus, I decided to remove this variable from the model.

A one-unit increase in personal income per capita results in a 0.00058 unit increase in the crime rate. A one percent increase in the unemployment rate results in a 9.44 unit decrease in the crime rate. This result contradicts the initial assumption that a higher unemployment rate would lead to higher crime rates. This negative correlation is also proved in "The Contemporaneous Effect of

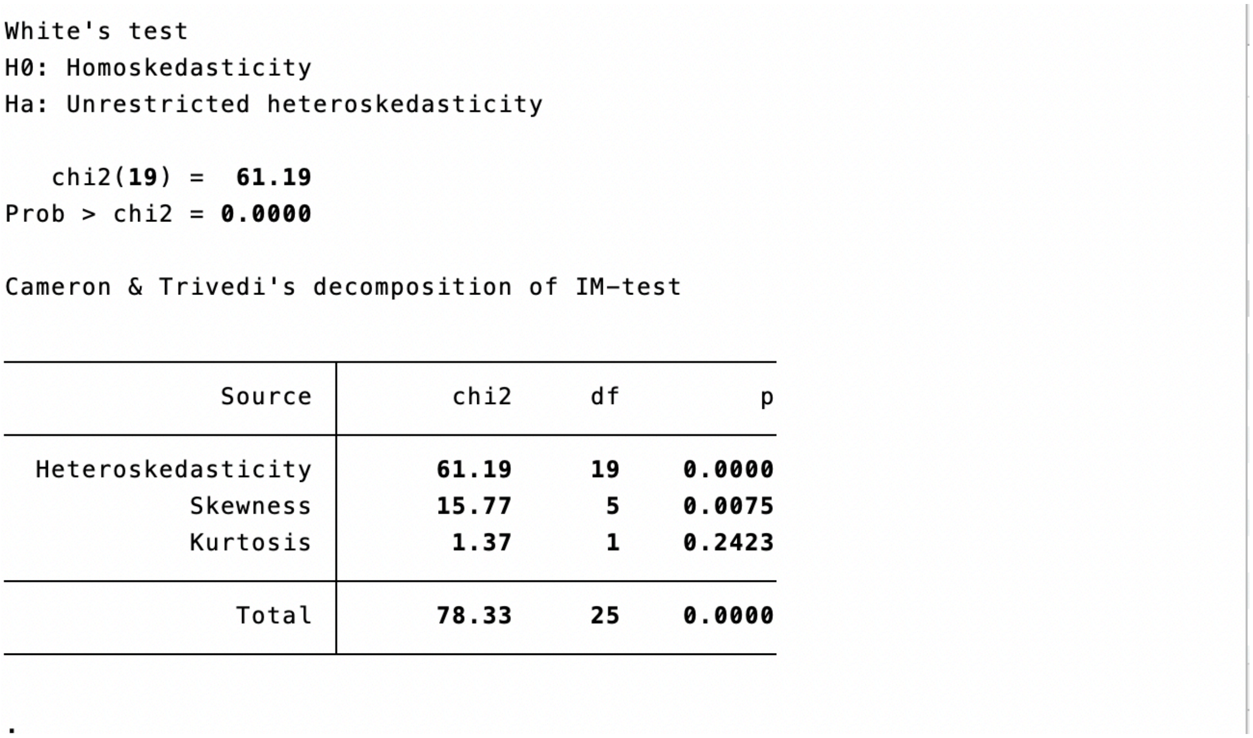
Unemployment on Crime Rates" by Guanlin Gao. If a state has a higher percentage of people with a high school degree or higher, this state, on average, has 26.41 fewer cases. With one more person per square mile increase in population density, the violent crime rate will likely reduce by 0.07 cases on average, which contradicts the initial hypothesis. Due to a posited surveillance effect, crimes of violence will be inversely related to density.

After adjusting the observations, the residual plot is as following:



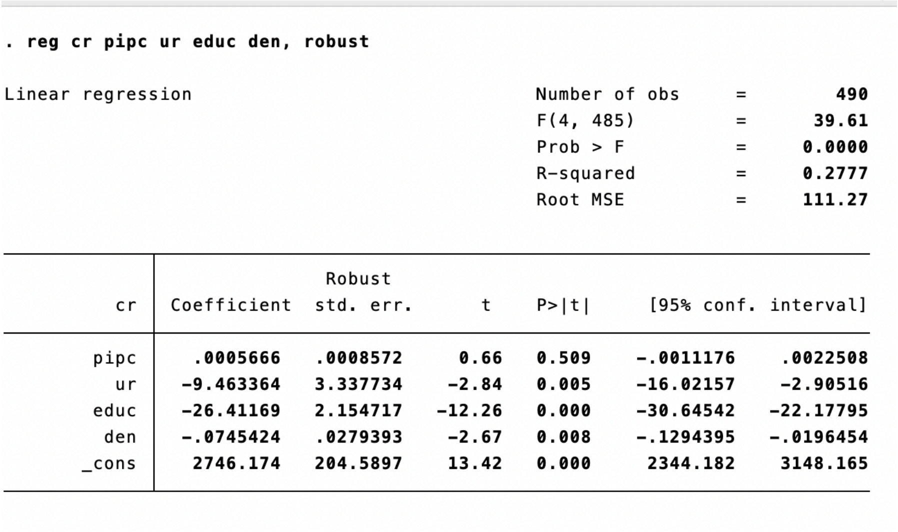
The model has improved as solid support of the previous argument on Alaska. Observations are more randomly distributed after outliers are taken out than in the last plot. However, there is still a trend of fanning out which needs further explanation. This trend is contributed mainly by observations from the state of New Mexico, which has a high violent crime rate as it borders Mexico with illegal drug and immigration problems.

Then, I suspect there is heteroskedasticity in the model. People with lower education levels tend to have a more considerable variance in committing a crime. In contrast, people with higher education levels have a lower variance in choosing to commit a crime. Thus, given different values of the percentage of people with at least high school degrees, the conflict of crime rate tends to differ. After running the white test for heteroskedasticity, I obtain the following result:



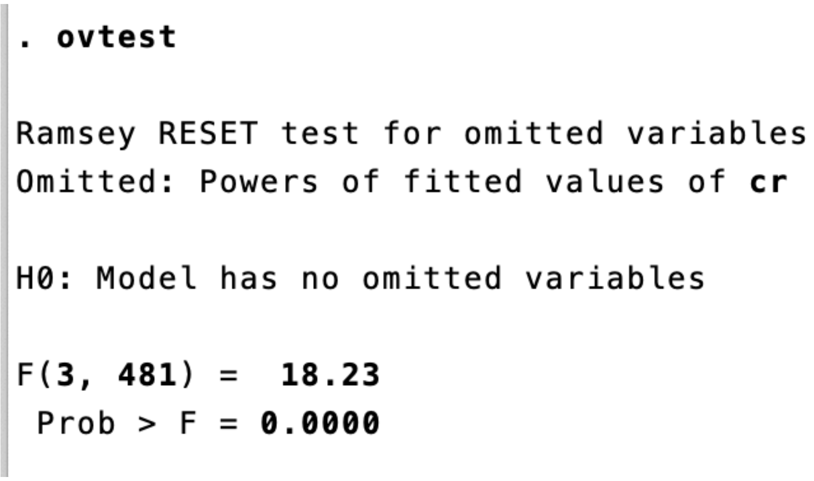
```
White's test
H0: Homoskedasticity
Ha: Unrestricted heteroskedasticity

   chi2(19) =  61.19
Prob > chi2 = 0.0000

Cameron & Trivedi's decomposition of IM-test

-------------------------------------------------
              Source |       chi2     df        p
---------------------+---------------------------
  Heteroskedasticity |      61.19     19   0.0000
            Skewness |      15.77      5   0.0075
            Kurtosis |       1.37      1   0.2423
---------------------+---------------------------
               Total |      78.33     25   0.0000
-------------------------------------------------

.
```

The null hypothesis is rejected with a low p-value of 0.0000, meaning heteroskedasticity in the model. Then, I use a heteroskedasticity-robust model to obtain the estimators. The sample size of 490 is large enough for the heteroskedasticity-robust model, and the result gives more unbiased variances. The following STATA output shows statistics of the heteroskedasticity-robust model.



```
. reg cr pipc ur educ den, robust

Linear regression                               Number of obs     =        490
                                                F(4, 485)         =      39.61
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2777
                                                Root MSE          =     111.27

------------------------------------------------------------------------------
             |               Robust
          cr | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        pipc |   .0005666   .0008572     0.66   0.509    -.0011176    .0022508
          ur |  -9.463364   3.337734    -2.84   0.005    -16.02157    -2.90516
        educ |  -26.41169   2.154717   -12.26   0.000    -30.64542   -22.17795
         den |  -.0745424   .0279393    -2.67   0.008    -.1294395   -.0196454
       _cons |   2746.174   204.5897    13.42   0.000     2344.182    3148.165
------------------------------------------------------------------------------
```

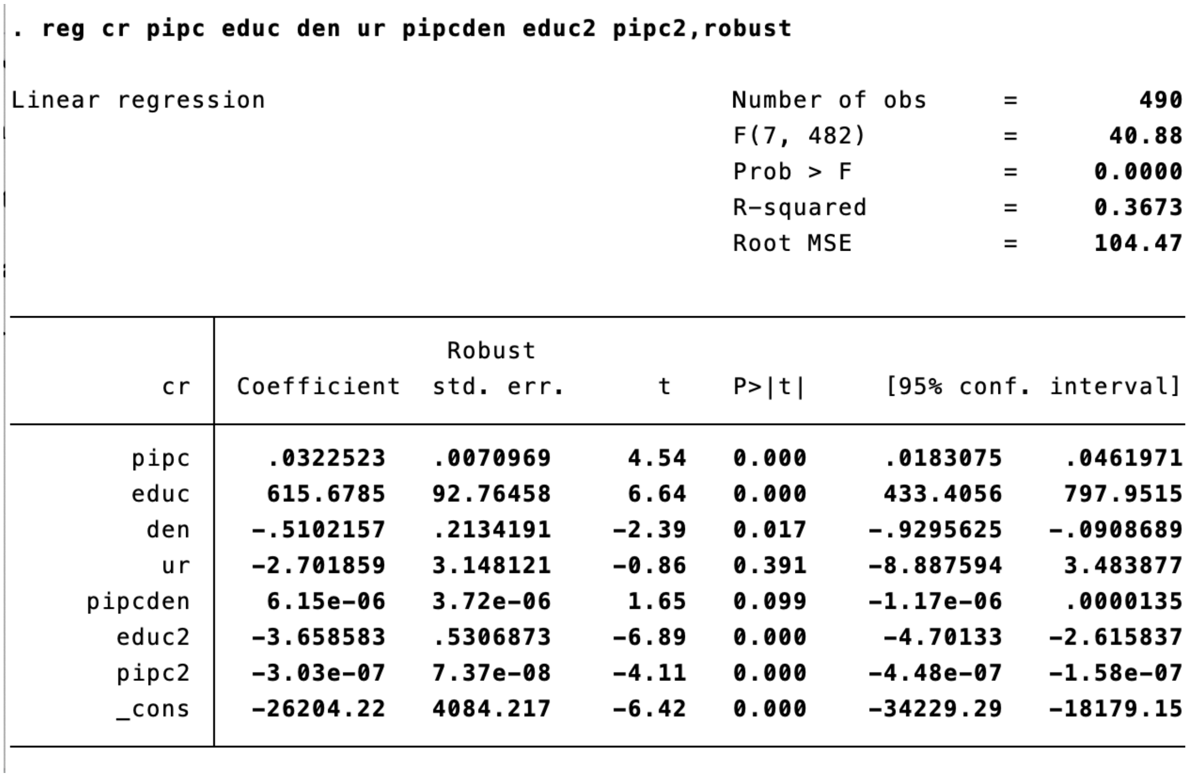
After running the planned model, I want to check if the original model has an omitted variable bias or/and functional form misspecification since I suspect that there might be relationships between independent variables.



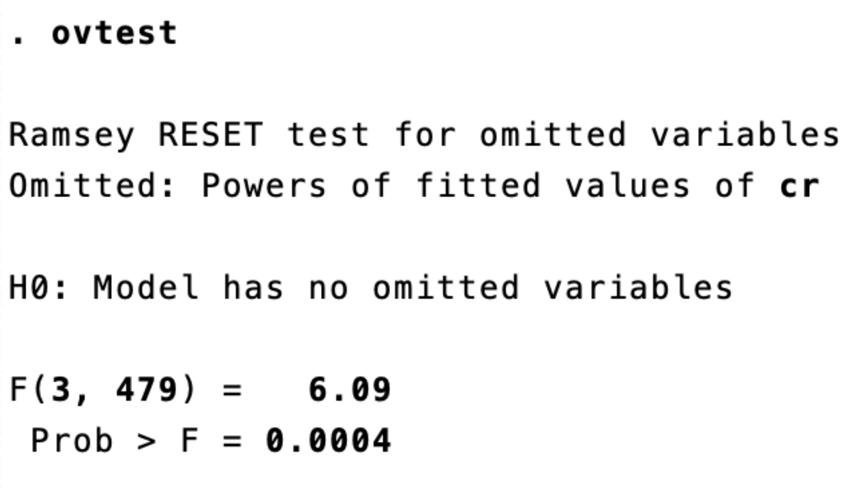
```
. ovtest

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of cr

H0: Model has no omitted variables

F(3, 481) =  18.23
 Prob > F = 0.0000
```

The null hypothesis for the RESET Test is that the model does not contain functional form misspecification, meaning that there are no omitted variables and functional form misspecification. Since the p-value (0.0000) is smaller than 0.05, the null hypothesis is rejected, suggesting that the model omitted variables and contained functional form misspecification. I decided not to use a log model since most of the variables are measured in rates and percentages. As the personal income per capita is different given different population densities, especially in urban areas, so I add an interaction variable pipc*den.

Moreover, I decide to square pipc since as personal income increases initially, there might be an envy effect since some people earn more income than others, resulting in a higher crime rate. However, once personal income rises to a certain point, it changes as the income of the entire population tends to increase, so I chose to square education since the effect of educ^2 might have a more negative effect (lowering crime rate) on crime rate compared to educ.



```
. reg cr pipc educ den ur pipcden educ2 pipc2,robust

Linear regression                               Number of obs     =        490
                                                F(7, 482)         =      40.88
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3673
                                                Root MSE          =     104.47

------------------------------------------------------------------------------
             |               Robust
          cr | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        pipc |   .0322523   .0070969     4.54   0.000     .0183075    .0461971
        educ |   615.6785   92.76458     6.64   0.000     433.4056    797.9515
         den |  -.5102157   .2134191    -2.39   0.017    -.9295625   -.0908689
          ur |  -2.701859   3.148121    -0.86   0.391    -8.887594    3.483877
     pipcden |   6.15e-06   3.72e-06     1.65   0.099    -1.17e-06    .0000135
       educ2 |  -3.658583   .5306873    -6.89   0.000     -4.70133   -2.615837
       pipc2 |  -3.03e-07   7.37e-08    -4.11   0.000    -4.48e-07   -1.58e-07
       _cons |  -26204.22   4084.217    -6.42   0.000    -34229.29   -18179.15
------------------------------------------------------------------------------
```



```
. ovtest

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of cr

H0: Model has no omitted variables

F(3, 479) =   6.09
 Prob > F = 0.0004
```

By adding this new independent variable, I obtained the highest p-value of the RESET (0.0004), which means the slightest evidence against the null hypothesis. Although it is still statistically significant to reject the null hypothesis, the p-value greatly improved from 0.0000 to 0.0004, proving that the model with the interaction term (pipc*educ), education square, and personal income square could be the most appropriate. Despite that ur and pipc*den are statistically significant, their economic significance weighs larger than their statistical insignificance.

Result

The final model's R Squared value is 0.3673, so 36.73% percent of the variation in crime rate is explained by the explanatory variables. One variable (ur) is statistically insignificant according to the individual t-test. Still, it is economically significant since unemployed people with poor economic conditions would contribute to violent crime more than people with better economic conditions. A one-unit increase in ur leads to a 2.7 decrease in the crime rate. A one-unit increase in pipc leads to (0.032 + 0.000006*den - 0.0000006*pipc) change in the crime rate. A one-unit increase in educ leads to (615.68 - 7.32*educ) change in the crime rate. A one-unit increase in den leads to a (-0.51 + 0.000006*pipc) change in the crime rate.

Acknowledgment

Due to limited ability and resources to collect data for more variables, other potential independent variables are not included in the model to better explain the crime rate variation. This limitation leads to a low p-value of the RESET even after improving the p-value with more suitable functional forms. Thus, this subject is still available for further research with more detailed social categories and data arrangements.

Conclusion

According to the results obtained from the proposed model, the variables contribute to the variation of the crime rate. What is surprising is that the legitimization of marijuana is statistically insignificant with a low coefficient estimate, which contradicts the expectations but corroborates Maier's conclusion that there lacks a relationship between crime rates and marijuana's legal status. While deleting the weed variables, the model demonstrates that personal income, population density, unemployment rate, and education level are associated with the dependent variable, violent crime rate. However, the relationship between population density and crime rate, unemployment rate, and crime rate contradicts the original hypothesis. This conclusion would suggest a shift in policy focuses on crime rates in rural areas and less attention to areas with high unemployment rates. Additionally, education remains to be the most important factor in the model, which suggests that universal state-wide education is a strong tool for social stability.

Work Cited

“Annual Unemployment Rates by State.” *Annual Unemployment Rates by State | Iowa Community Indicators Program*, https://www.icip.iastate.edu/tables/employment/unemployment-states.

“Crime Rate by State, 2010.” *Infoplease*, https://www.infoplease.com/us/crime/crime-rate-state-2010.

Crime, crime-data-explorer.app.cloud.gov/pages/explorer/crime/crime-trend.

“Educational Attainment, Annual: High School Graduate or Higher by State.” *FRED*, Federal Reserve Bank of St. Louis, https://fred.stlouisfed.org/release/tables?rid=330&eid=394766.

“List of the U.S. States and Territories by Violent Crime Rate.” *Wikipedia*, Wikimedia Foundation, 5 Apr. 2022, https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_violent_crime_rate.

Mariapushkareva. “GDP per Capita in the US States with Tableau.” *Kaggle*, Kaggle, 20 Feb. 2020, https://www.kaggle.com/code/mariapushkareva/gdp-per-capita-in-us-states-with-tableau.

“Personal Income.” *Personal Income | U.S. Bureau of Economic Analysis (BEA)*, https://www.bea.gov/data/income-saving/personal-income.

Published by Statista Research Department, and May 17. “U.S.: GDP per Capita by State in 2018.” *Statista*, 17 May. 2021, https://www.statista.com/statistics/248063/per-capita-us-real-gross-domestic-product-gdp-by-state/.

Smith, Jeff, et al. “Where Marijuana Is Legal in the United States.” *MJBizDaily*, 14 Apr. 2022, https://mjbizdaily.com/map-of-us-marijuana-legalization-by-state/.

“Unemployment Rates for States.” *U.S. Bureau of Labor Statistics*, 2 Mar. 2022, https://www.bls.gov/lau/lastrk19.htm.

“Useful Stats: Per Capita Gross State Product, 1998-2018.” *SSTI*, https://ssti.org/blog/useful-stats-capita-gross-state-product-1998-2018

“Impact of Marijuana Legalization on Crime Rates Is Underestimated,” Kyle Jaeger, 15 Oct. 2021, https://www.marijuanamoment.net/impact-of-marijuana-legalization-on-crime-reduction-is-being-underestimated-new-study-finds/

“The Contemporaneous Effect of Unemployment on Crime Rates,” Guanlin Gao, https://swer.wtamu.edu/sites/default/files/Data/Gao.pdf

Duggan, M. (2000). *More Guns, More Crime*. National Bureau of Economic Research.

Bhattacharya, A. (2020). *Analysis of the Factors Affecting Violent Crime Rates in the US.* International Journal of Engineering and Management Research.

L. Maier, S. (n.d.). *The implications of marijuana ... - sage journals*, 17 April. 2022, https://journals.sagepub.com/doi/10.1177/0091450917708790

Lochner, L., & Moretti, E. (n.d.). *The effect of education on crime: Evidence from prison inmates, arrests, and ...*, 18 April. 2022