Henry Bi, Haojuan He, Sven Wu

Evgeniya Duzhak

Econ 140

01 December 2020

Communities and Crime Data Analysis

Introduction:

Education plays a significant role in predicting the number of crimes committed in the population. According to many empirical researches, states with higher levels of education attainment usually resulted in crime rates that are below the national average. This is an interesting topic to explore since researchers in multiple disciplines such as social science, criminal justice studies, and political science are debating on whether or not the impact of education levels on reducing the crime rate is significant. As economists, we are interested in understanding the social return of government public investments in education that can be potentially beneficial to the overall economic output as it can generate more labor productivity. Academic research conducted by Lance Lochner and Enrico Moretti has found that "schooling significantly reduces the probability of incarceration and arrest." We hope to test their conclusion and identify possible additional factors that correlate with education such as race, income, employment rate, etc.

In this regression analysis project we aim to develop insights regarding the effectiveness of education of which policymakers or local communities can use to not only employ to help stagger crime rates in their area, but also use it as a reference to respond to the future demographic and social change of their area. Our goal is to first find data from creditworthy sources to reduce systematic errors in the data collection process and minimize any potential response biases. We will then proceed to hypothesize a list of variables that have significant impacts to changes in the local crime rates. After deciding the list of variables that are not highly correlated with each other, we would want to test if there exists any heteroskedasticity that can reduce the accuracy of our model. Lastly, we would develop insights to improve the model.

Model specification:

In this model, we used our intuition after assessing academic research papers and proposed to use the number of violent crimes per 100,000 people as the dependent variable. We have chosen to use percentage less than 9th grade, percentage with bachelor's degree or more and percentage of population that did not graduate from high school as our key independent variables which we will run an OLS multiple regression on. We would expect an inverse relationship (-) between the number of crime rates and the percentage of the population who are below 9th grade. We expect that teenagers are less likely to be committing violent crimes as it is relatively rare. A similar relationship is also expected for the percentage of the population who holds bachelor's degrees or higher since higher education levels tend to reduce the motivation for violent crimes. For the percentage who did not graduate high school, we would expect a positive

relationship (+) with the number of violent crimes. This population group includes highschool dropouts who usually are more likely to get into trouble. We have decided to also include additional hypothesized parameters such as medium income (+), race distributions (percent of Asian/Black/White/Hispanic in the communities), population density (+), percent that does not speak English well (+) and unemployment rate (+) to assess other social-economic factors that can contribute to a high level of crime rate in a particular state.

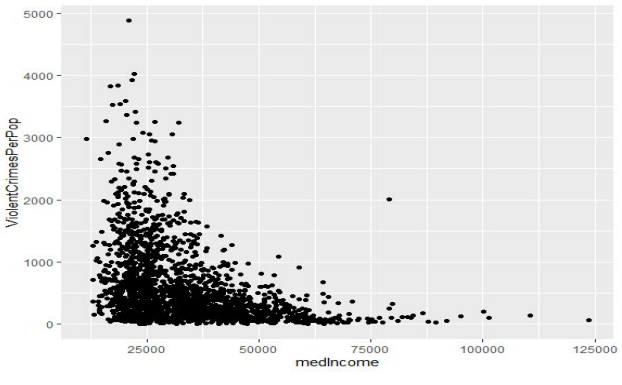
Data:

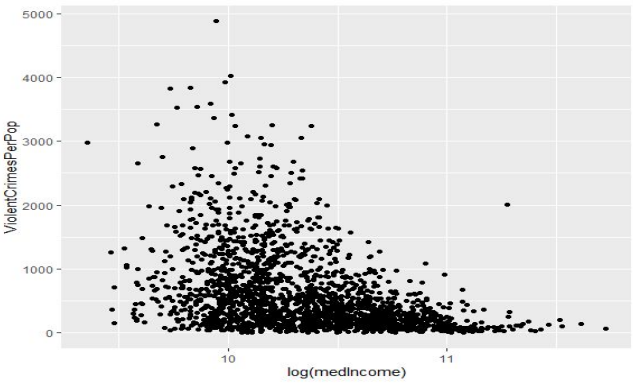
We have utilized the "Communities and Crime" dataset from the UC Irvine machine learning repository. This specific data is a combination of social-economic, law enforcement, and crime rate data collected by the US Census, US LEMAS survey, and FBI UCR in 1990. The data type is considered to be cross-sectional data which describes the activities of entities at a specific given point of time which in our case is 1990.

The original UCI dataset consists of 2215 rows, with each row pertaining to a specific community and 147 different column attributes. The columns consist of many socio-economic attributes varying from race, age, levels of education, to whether or not a specific community is urban, to even the number of police cars. This provided us with a wide plethora of factors to choose from.

Since our team decided to investigate the impacts and levels of influence a person's education level has on the number of crimes committed within communities, we decided to focus strictly on a smaller subset of the data. We first decided on the types of variables that we would include from the UCI dataset. We collected all the columns that strictly pertained to education and since there were only a total of 3, we chose to include all of them in our model. We also looked into other factors aside from education, to construct a more holistic data sample. Most of these additional factors were chosen if they had some potential correlation with education. This included columns pertaining to different races, population densities, rent, employment status, etc. Some of the columns we had also included "null" values, or more specifically the question mark strings in our data, indicating unknown values. To preserve the data, we wrote a function to determine what percent of the columns contained unknown values. We then set a threshold of 20% to remove any columns that exceeded that percentage of unknown question mark string values. After that, we plotted scatter plot distributions for each of our chosen features against our dependent variable ViolentCrimesPerPop to help visualize some of the data. This gave us a better understanding of the correlation of each of our chosen variables with our main dependent variable and determined which ones we would keep. This helped condense our dataset down to 20 columns and 1994 rows.

We also decided to apply a Log transformation to the medium income parameter to increase variability in independent variables. By doing so, we can reduce the standard error of the OLS coefficient and intercept term which makes the model more precise.





Correlation matrix:

Since our model includes multiple parameters, in order to better improve the model, we took into account that some of those variables might be related and even casual to each other. We decided that having 19 column attributes was still a lot, so we sought to reduce it down to just 8. In order to minimize the auto-correlation effects and the possible high collinearity effect between those variables, we calculated and created a correlation matrix of all the parameters that we integrated in our regression models. Our correlation matrix is a 19 * 19 matrix which includes correlation between every pair of the parameters excluding the only categorical variable: *state*. As shown in the reference section, we conducted our correlation matrix in the graphic form of a heat map. The number in the center of each small colored block is the correlation value between those two variables. As a result, the larger the absolute value of the number in a cell block indicates a stronger correlation between those two parameters. It is noteworthy that this correlation only reflects the possible collinearity between those two variables but not a casual relationship. The absolute value of each block in the heat map is identically reflected by the diagonal line but different in the sign of the correlation. Also, the beige color in the diagonal line across the graph indicates each variable is perfectly correlated to itself.

Based on the information shown in the heat map, there are several cells with high correlation that are worth mentioning. Other than our dependent variable *ViolentCrimesPerPop*, we originally have 18 other variables in the dataset. After evaluating the correlation between those variables, we decided to drop 10 parameters that were highly correlated with each other and only kept 8 variables for our regression model to minimize the inaccuracy due to auto-correlation. We decided to drop *PctLess9th Grade* due to its high correlation with PctBSorMore (corr = 0.58). We also dropped racepctWhite and racePcgtAsian due to its high correlation with racePctBlack. The NumInShelters and MedRent parameters are also dropped due to their super high correlation with numStreet and Medincome variables.

Overall, after analyzing the auto-correlation in between our independent variables, the last eight variables we decided to keep are: *ViolentCrimesPerPop*, *PctBSorMore* (percentage of population with bachelor degree or higher), medIncome (median income), Racepctblack, *PctUnemployed*, *agePct12t29*, *PctNotSpeakEnglWell*, *NumStreet* (Number of homeless people) and *PctNotHSGrad* (percent of population who are not high school graduates). All of those variables are not only able to represent our dependent variables well but also have a low correlation with each other. As a result, by analyzing the parameter through the correlation matrix, we are able to return the variables that best represent the model without the auto-correlation bias.

Results:

Robust Standard Errors:

```
Coefficients:
                  Estimate Std. Error t value  Pr(>|t|)  CI Lower  CI Upper   DF
(Intercept)       3666.1773    598.094  6.1298 1.059e-09 2493.2185 4839.1362 1985
PctBSorMore         -2.6924      1.180 -2.2818 2.261e-02   -5.0065   -0.3783 1985
PctNotHSGrad        -5.1580      2.271 -2.2714 2.323e-02   -9.6116   -0.7044 1985
medIncome         -302.3256     52.287 -5.7820 8.553e-09 -404.8689 -199.7823 1985
racepctblack        25.3005      1.436 17.6160 1.191e-64   22.4838   28.1172 1985
PctUnemployed       12.8713      7.709  1.6696 9.515e-02   -2.2476   27.9902 1985
agePct12t29         -7.1077      1.726 -4.1177 3.984e-05  -10.4930   -3.7225 1985
PctNotSpeakEnglWell 42.3206      4.545  9.3106 3.257e-20   33.4063   51.2349 1985
NumStreet            0.1714      0.216  0.7935 4.276e-01   -0.2522    0.5950 1985

Multiple R-squared:  0.5254 ,   Adjusted R-squared:  0.5234
F-statistic: 113.2 on 8 and 1985 DF,  p-value: < 2.2e-16
```

Coefficient Interpretation:

Most of the parameters we found proved to be statistically significant based on the robust regression model we ran above. There were 2 specific parameters in our model that specifically corresponded to education level: *PctBSorMore* and *PctNotHSGrad*.

PctBSorMore corresponds to the percentage of people 25 years and over with a bachelor's degree or some other higher education. We expected higher education to be associated with lower committed crimes. Our robust model reported back a coefficient of -2.6924 for *PctBSorMore*. This is consistent with our expectations, as a 1% increase in the percentage of people 25 years and older with a bachelors' degree or higher, led to a 2.69 unit decrease in the number of violent crimes per 100,000 cases committed per community. At a 5% significance level, this parameter is statistically significant with a low p-value near 0.

PctNotHSGrad corresponds to the percentage of people 25 years and over that are not high school graduates. Similar to *PctBSorMore*, we expected higher education levels to be associated with a decreased number of committed crimes. Our robust model reported back a coefficient of -5.1580. This is actually inconsistent with our expectations, as it means that a 1% increase in the percentage of people 25 years and older that did not graduate high school, led to a 5.15 unit decrease in the number of violent crimes per 100,000 cases committed per community. At a 5% significance level, the parameter is statistically significant with a p-value of 0.02.

The other parameters we integrated into our model were included due to their potential correlation with education. They consist of race, income, employment, and other related social factors.

medIncome corresponds to the median household income. We took the natural log of this variable to help reduce any potential standard errors as mentioned earlier. This variable proved to be one of the more important ones, given that it has the largest coefficient in the robust model. We expected higher education levels to be associated with a decreased number of committed crimes. We hypothesized that higher education levels would lead to higher wages from higher paying jobs. Under that assumption, we would expect the medIncome variable and *ViolentCrimesPerPop* variable to have an inverse relation. The robust model results are consistent with our expectations. For every 1% increase in the median household income, a 3.023256 unit decrease was expected in the number of violent crimes per 100,000 cases committed per community. At a 5% significance level, the parameter is statistically significant.

racepctblack corresponds to the percentage of the population that is African American in the given community. As mentioned above, the other race parameters that correspond to White, Asian, and Hispanics, were dropped due to the high correlation between these four race variables. According to the NAACP, an African American is 5 times more likely than a white person to be stopped without cause. 65% of African American adults also expressed that they were often targeted because of their race, rather than for the nature of the situation or crime.

Similarly, approximately 35% of Latino and Asian adults also expressed similar sentiments. As such we expected a positive correlation between our race variables and the *ViolentCrimesPerPop* variable. In this specific case, we kept *racepctBlack*. Running the robust model, we see for every 1% increase in the percent of African Americans in a given community, a 25.003 unit increase was expected in the number of violent crimes per 100,000 cases committed per community. At a 5% significance level, the parameter is statistically significant. It also has the lowest p-value out of all our chosen variables, resting at 1.191e-64.

PctUnemployed corresponds to the percent of people 16 years and over who are unemployed. We hypothesized that higher education levels were correlated with a higher likelihood of being employed. Under this assumption, we expect a positive relationship between *PctUnemployed* and *ViolentCrimesPerPop*. Our robust model shows that for every 1% increase in the percent of people unemployed in a given community, a 12.8713 unit increase is expected in the number of violent crimes per 100,000 cases committed per community. At a 5% significance level, the parameter is statistically insignificant, as it has a p-value of 0.09.

agePct12t29 corresponds to the percent of the population that is between 12 to 29 years of age. As mentioned above, the other age variables were dropped due to high correlation. We kept this one, because it had the highest range. We also hypothesized that younger people were less likely to commit crimes than older people, given that they would most likely be in school during these age ranges. Under this assumption, we expected an inverse relationship between *agePct12t29* and *ViolentCrimesPerPop*. Our robust model shows that for every 1% increase in the percent of the population between 12 to 29 years of age, a 7.1077 unit decrease is expected in the number of violent crimes per 100,000 cases committed per community. At a 5% significance level, the parameter is statistically significant.

PctNotSpeakEngWell corresponds to the percent of people who do not speak English well in a given community. We hypothesized that being unable to speak English well in a given community in the United States may be tied to lower educational attainment. Under this assumption, we expected a positive correlation between *PctNotSpeakEngWell* and *ViolentCrimesPerPop*. Our robust model shows that for every 1% increase in the percent of the population that does not speak English well, a 42.302 unit increase is expected in the number of violent crimes per 100,000 cases committed per community. At a 5% significance level, the parameter is statistically significant.

NumStreet corresponds to the number of homeless people in the streets of a given community. We hypothesized that most likely higher education levels correspond with a smaller chance of ending up homeless. Under this assumption, we expect a positive correlation between *NumStreet* and *ViolentCrimesPerPop*. Our robust model shows that for every 1 unit increase in the number of homeless people in the streets, a 0.1714 unit increase is expected in the number of violent crimes per 100,000 cases committed per community. At a 5% significance level, the parameter is statistically insignificant, as its p-value is 0.4.

For the intercept term holding all other coefficients constant and 0, we expect a 3666.177 unit of increase in the number of violent crimes per 100,000 cases committed per community. However due to the complication and reality of the model, we should not look much into the intercept term in the model.

Fit: adjusted R^2 (coefficient of determination)

After building out a regression model and refining it with multiple analyzation, we came to the conclusion that the robust regression model fits best for the data we obtained. According to the summary of the robust regression model above, we have a coefficient of determination of 0.5254 and an adjusted coefficient of determination of 0.5234. Since we are doing multiple linear regression instead of singular linear regression, the adjusted R^2 value better explains the fit of the model as it takes into consideration the decreasing degree of freedom as the variables in the model increase. The adjusted R^2 of 0.5234 indicates that 52.34% of the variation in the change of our independent variable (ViolentCrimePop) can be explained by the variation in all the

dependent variables of our choice. This high value in the coefficient of determination further supports that our regression model is a good fit for our data.

MLR/SLR Assumptions:

We assume that the relationship between our X and Y variables to be linear in parameters. Before running the regression analysis, we wanted to assess whether the error term has a constant variance for all observations and satisfy the homoskedasticity MLR 4 assumption. A failure of this particular assumption will result in inefficient estimates and a biased test of hypothesis due to inaccurate standard errors. Therefore, we decided to test for heteroskedasticity by running a Breusch-Pagan Test. The null hypothesis is that the squared residuals should not depend on our regressors variables and we used a 5% alpha value to determine the rejection decision. The result of the test is shown below:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       1547005.93  572891.77   2.700  0.00699 **
PctBSorMore            59.72    1305.38   0.046  0.96352
PctNotHSGrad        -4290.63    2067.68  -2.075  0.03811 *
medIncome         -132033.38   50402.49  -2.620  0.00887 **
racepctblack        12915.46     810.77  15.930  < 2e-16 ***
PctUnemployed       16385.72    5946.73   2.755  0.00592 **
agePct12t29         -6049.43    1874.65  -3.227  0.00127 **
PctNotSpeakEnglWell 15714.59    3303.24   4.757  2.1e-06 ***
NumStreet              63.24      37.80   1.673  0.09445 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 428800 on 1985 degrees of freedom
Multiple R-squared:  0.197,     Adjusted R-squared:  0.1937
F-statistic: 60.86 on 8 and 1985 DF,  p-value: < 2.2e-16
```

We see that the p-value is significantly smaller than 5%. Therefore we reject the null hypothesis and conclude that there is statistically significant evidence that shows that the crime data is heteroskedastic. In order to fix this concern, we then proceeded to apply the weighted least square method in resolving the violation of heteroskedasticity. The steps we have taken is first we regress y on x and get the residual variables. Then we regress the log of squared residuals on x to get the predicted values. Lastly, we would weight the data by $\frac{1}{e^{predicted\ values}}$ and run an OLS regression with the new coefficients that adjust for heteroskedasticity. This is our result:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       2892.3975   339.1853   8.527  < 2e-16 ***
PctBSorMore         -1.0440     0.7097  -1.471 0.141414
PctNotHSGrad        -4.9425     1.4875  -3.323 0.000908 ***
medIncome         -243.5117    29.8534  -8.157 6.03e-16 ***
racepctblack        25.2604     1.1129  22.699  < 2e-16 ***
PctUnemployed       29.1989     4.5825   6.372 2.31e-10 ***
agePct12t29         -6.9746     1.0339  -6.746 1.99e-11 ***
PctNotSpeakEnglWell 48.9139     3.3208  14.730  < 2e-16 ***
NumStreet            0.6585     0.1173   5.613 2.27e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.193 on 1985 degrees of freedom
Multiple R-squared:  0.4435,    Adjusted R-squared:  0.4412
F-statistic: 197.7 on 8 and 1985 DF,  p-value: < 2.2e-16
```

We see that after adjusting for heteroskedasticity, all of our coefficients except for the variable percentage of population holding a bachelor's degree or more are all significant at the 0.1% level which is very interesting. We also observed that before adjusting for heteroskedasticity, the bachelor's degree or more was significant at the 5% level and now it is not. This is quite unexpected since we have always assumed that higher education levels will cause a lowering in the numbers of crime rate. Also by adjusting for the Heterskedascitiy, the variable percent unemployed and percent not high school graduates both became more significant at the 0.1% level.

Another MLR assumption we wanted to check was the autocorrelation test. We know that if our regressors correlate with the error term will induce omitted variable bias. So we ran the dwtest and here is the result:

```
        Durbin-Watson test

data:  reg3
DW = 2.0515, p-value = 0.8754
alternative hypothesis: true autocorrelation is greater than 0
```

Thus we see that the high p-value indicates that we failed to reject the null where the linear regression residuals are uncorrelated.

Another assumption we tried to test is the omitted variable bias. MLR is the most effective and efficient under the presumption that all variables that have influence in the dependent variable are included as parameters. In order to verify the hypothesis that there are no more omitted variables, we decided to test whether the square term of some variables should be included in the regression as well. We add in a new parameter the square of *ViolentCrimesPerPop*. Running the regression model with this new parameter allows us to test whether there is any square or power relationship between the independent and dependent variables. In order to test for the omitted variable bias, we perform the reset test to test for omitted squared term of variables.

Reset Test:

```
RESET = 24.153, df1 = 1, df2 = 1984, p-value = 9.629e-07
```

After running the regular regression model with the selected variables, we perform the reset-test to verify the MLR(8): omitted variable bias in nonlinear parameters. The reset-test tests for how all the independent variables affect the square residual. Our reset test is done by performing OLS with the additions of the squared and power term of *ViolentCrimePerPop* as a dependent variable. Corresponding to the MLR(8) assumption, the null hypothesis is that all the coefficients for the addition dependent variable should be jointly zero as they are supposed to have no additional impact to the original dependent variable. After we perform the reset test, we get the summary as shown above. The reset score is 24.153 and the p-value is 2.2e-16. This p-value is an extremely small number, which indicates that there is a high chance that our model rejects the null hypothesis. Even with a super large 99% significant interval, 2.2e-16 is smaller than our critical p-value of 0.01. As a result, we drew the conclusion that at 99% significant level, we reject the null hypothesis that all the coefficients for the squared and powered terms are jointly zero. Since all the coefficients are statistically significant, it breaks the MLR(8)

assumption that there is no omitted variable bias in our model. Our summary of the reset test indicates that at 99% significance interval, we reject the null hypothesis and admit that there is omit variable bias (with squared and power term of relationship with the original dependent model) that exists in our data.

Summary and Conclusion: (1-2 pages):

In this paper, we are conducting a regression analysis on the dependent variable number of crimes committed per 100,000 people. After analyzing the data and refining our model, we came to the conclusion that a robust multiple linear regression model was the best fit model for our regression. The robust regression model takes into consideration the heteroskedasticity of the data while still maintaining smaller standard errors for each coefficient. (compared to weighted-least squares regression). The robust regression model has a fitted coefficient of determination of 0.5234, which indicates that in our model, 52% percent of the change in our dependent variable (ViolentCrimePop) could be explained by the variation in the dependent parament of our choice.

The coefficients that we have hypothesized have interesting implications. After running the robust regression, we found that most of our selected independent variables were actually statistically significant. Most of our assumptions regarding the relations between each independent variable with our dependent variable, *ViolentCrimesPerPop*, were also for the most part accurate. We associated higher education levels with several other factors, such as higher incomes and employment status. *medIncome* was one of the more important variables given that it had one of the highest coefficients at -3.023256. Because we assumed higher education levels led to higher-paying jobs, we were able to support the hypothesis that education plays a significant role in bringing down the number of crimes committed in a given community. *PctBSorMore* was one of our main variables that specifically involved education- the percentage of people with a bachelor's degree or higher. The coefficient was -2.6924 and given our hypothesis, the robust regression model supported our findings.

One of our main education variables *PctNotHSGrad* actually had the opposite effect of what we expected. The model reported back a coefficient of -5.1580. This means that a 1% increase in the percentage of people 25 years and older that did not graduate high school, led to a 5.15 unit decrease in the number of violent crimes per 100,000 cases committed per community. This is inconsistent with our expectations that higher education levels lead to reduced crime numbers. One possible solution that we can explore is to include entity fixed effect and see if the sign of the variable changes. The coefficients for the other additional variables pertaining to race, age, and homelessness also coincided with our hypotheses.

In order to improve the accuracy of our model, we also assessed various MLR assumptions and proposed potential ways of addressing the problems. The data did have some violations of MLR assumption (7) and (8): homoscedasticity and omitted-variable bias. The MLR model results in the most efficient and effective explanation but only under several assumptions. In order to validate the use of MLR models, we tested the accuracy for several MLR assumptions given the data we use. We find out that our data breaks MLR assumption (7) and (8) as the data is heteroscedastic and has omitted-variable bias. We also should include higher-order variables that better fit the data as we rejected the null hypothesis when conducting the reset test. Including a time-series, data may be useful as well, as it may provide more insight on how the number of crimes committed may change over time.

In conclusion, our robust model aligned with our hypothesis that higher educational levels were tied with lowering the number of crimes committed per community.

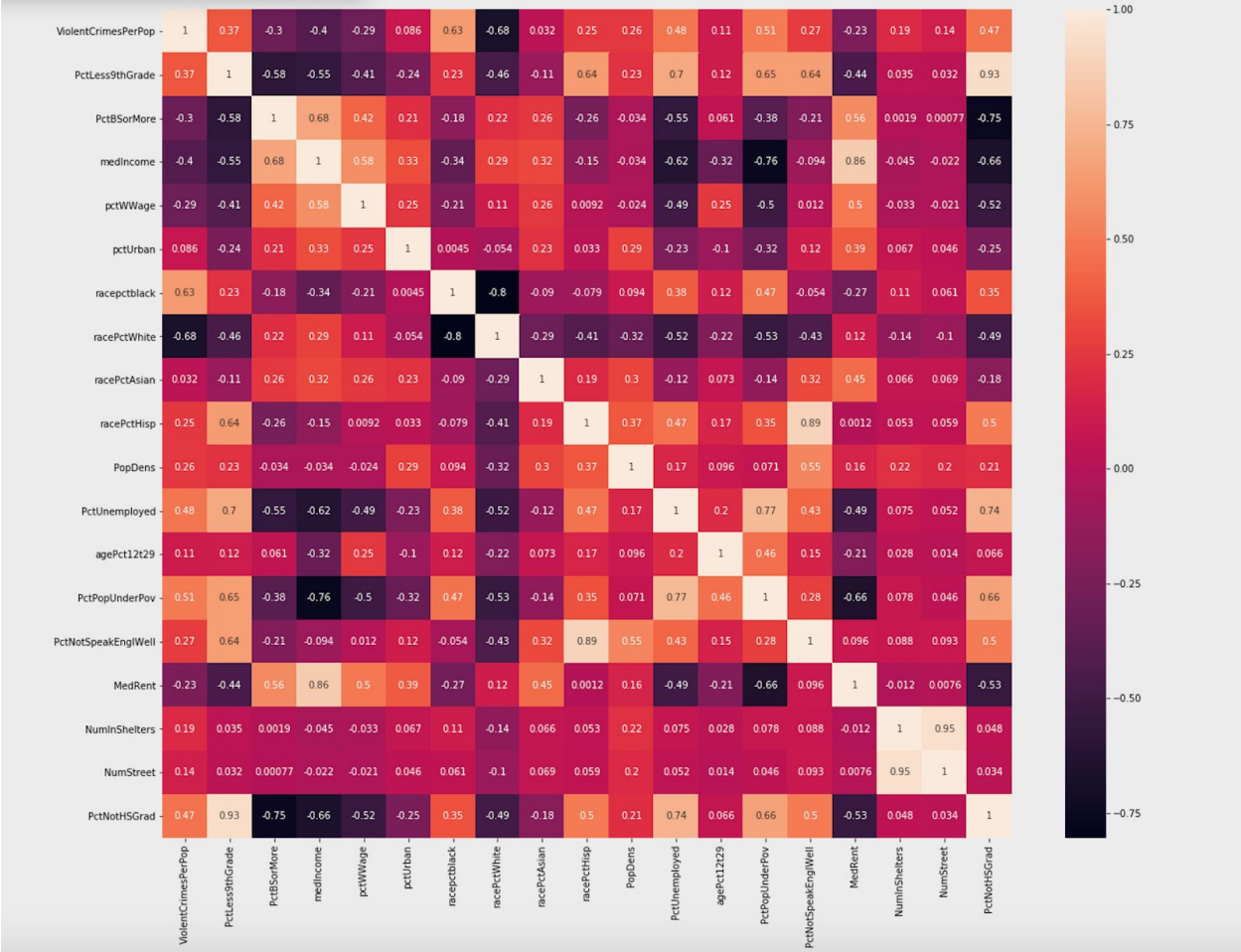
References:

Data source: https://archive.ics.uci.edu/ml/datasets/communities+and+crime

"Criminal Justice Fact Sheet." *NAACP*, 10 July 2020, www.naacp.org/criminal-justice-fact-sheet/

Lance Lochner, Enrico Moretti. The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reposts. March 2020.

Correlation Matrix (Heatmap)



```r
library("readxl")
library("tidyverse")
library("broom")
library("lmtest")
library("stargazer")

library("knitr")
library("plm")
library("plyr")
library("ggplot2")
library("tidyr")
library("Hmisc")
install.packages("Hmisc")
install.packages("corrplot")
library(estimatr)


# File name with path
library(readr)
crime_data <- read_csv("C:/Users/henry bi/Downloads/crimedata_try2.csv")
df <- crime_data

reg1 <- lm(ViolentCrimesPerPop~PctLess9thGrade+PctBSorMore+medIncome  +pctWWage +pctUrban +racepctblack
           +racePctWhite  +racePctAsian +racePctHisp  +PopDens  +PctUnemployed
           +agePct12t29 +PctPopUnderPov +PctNotSpeakEnglWell  +MedRent
           +NumInShelters +NumStreet +PctNotHSGrad, data = df)

summary(reg1)

df$ViolentCrimesPerPopPercent <- (crime_data$ViolentCrimesPerPop / 100000)*100


reg2 <- lm(ViolentCrimesPerPopPercent~PctLess9thGrade+PctBSorMore +medIncome  +pctWWage +pctUrban +racepctblack
           +racePctWhite  +racePctAsian +racePctHisp  +PopDens  +PctUnemployed
           +agePct12t29 +PctPopUnderPov +PctNotSpeakEnglWell  +MedRent
           +NumInShelters +NumStreet , data = df)

summary(reg2)

stargazer(reg2,
          type= "html",
          out= "termcolumns.doc",
          title= "Term Project Columns",
          align = TRUE)


## the matrix of scatter plot
df %>%
  gather(-ViolentCrimesPerPop, key = "var", value = "value") %>%
  ggplot(aes(x = value, y = ViolentCrimesPerPopPercent)) +
  geom_point() +
  facet_wrap(~ var, scales = "free") +
  theme_bw()
```

```r
# Henry BI Transformation
df$medIncome <- log(df$medIncome)

# robust linear regression
fit_robust <- lm_robust(ViolentCrimesPerPop ~ PctBSorMore + PctNotHSGrad + medIncome+racepctblack +PctUnemployed
                        +agePct12t29  +PctNotSpeakEnglWell +NumStreet, data = df)

ggplot(df, aes(x= log(medIncome), y= ViolentCrimesPerPopPercent)) + geom_point()
ggplot(df, aes(x= medIncome, y= ViolentCrimesPerPopPercent)) + geom_point()
ggplot(df, aes(x= log(medIncome) , y= ViolentCrimesPerPop)) + geom_point()
ggplot(df, aes(x= medIncome , y= ViolentCrimesPerPop)) + geom_point()




#  selected data and easy MLR
reg3 <- lm(ViolentCrimesPerPop ~ PctBSorMore + PctNotHSGrad + medIncome+racepctblack +PctUnemployed
           +agePct12t29 +PctNotSpeakEnglWell +NumStreet, data = df)
summary(reg3)

summary(fit_robust)


# residual, heteroskedastic test Henry BI
reg4 <- lm(reg3$residuals^2 ~ PctBSorMore + PctNotHSGrad + medIncome+racepctblack +PctUnemployed
           +agePct12t29 +PctNotSpeakEnglWell +NumStreet, data = df)
summary(reg4)


# fixes using wls Sven & henry BI
lnres3 <- log(reg3$residuals^2)
weights_reg3 <-lm(lnres3 ~ PctBSorMore + PctNotHSGrad + medIncome+racepctblack +PctUnemployed
                  +agePct12t29  +PctNotSpeakEnglWell +NumStreet, data = df)
hhat <- exp(weights_reg3$fitted.values)

wls.reg3 <- lm(ViolentCrimesPerPop ~ PctBSorMore + PctNotHSGrad + medIncome+racepctblack +PctUnemployed
               +agePct12t29 +PctNotSpeakEnglWell +NumStreet, weights = 1/hhat, data =df)
summary(wls.reg3)
summary(fit_robust)
summary(reg3)


# reset test Jennifer
resettest(ViolentCrimesPerPop~ PctBSorMore + PctNotHSGrad + medIncome+racepctblack +PctUnemployed
          +agePct12t29  +PctNotSpeakEnglWell +NumStreet, power=2:3,type=c('fitted'),data = df)


# auto correlation test Henry BI OVB
dwtest(fit_robust)
dwtest(reg3)
```

Econ 140 Term Project

By: Sven Wu, Henry Bi, Haojuan He

- Research Topic: Communities and Crime Data Set

```
import pandas as pd
import numpy as np
```

```
from google.colab import drive
drive.mount('drive')
```

```
Mounted at drive
```

Loading in the data

```
from google.colab import files
uploaded = files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving crimedata.csv to crimedata.csv
```

```
import io
crime_data = pd.read_csv(io.BytesIO(uploaded['crimedata.csv']), encoding = "ISO-8859-1")
# Dataset is now stored in a Pandas Dataframe
```

```
crime_data
```

	Êcommunityname	state	countyCode	communityCode	fold	population	household
0	BerkeleyHeightstownship	NJ	39	5320	1	11980	
1	Marpletownship	PA	45	47616	1	23123	
2	Tigardcity	OR	?	?	1	29344	

```
for i in crime_data.columns:
  print(i)
```

```
Êcommunityname
state
countyCode
communityCode
fold
population
householdsize
racepctblack
racePctWhite
racePctAsian
racePctHisp
agePct12t21
agePct12t29
agePct16t24
agePct65up
numbUrban
pctUrban
medIncome
pctWWage
pctWFarmSelf
pctWInvInc
pctWSocSec
pctWPubAsst
pctWRetire
medFamInc
perCapInc
whitePerCap
blackPerCap
indianPerCap
AsianPerCap
OtherPerCap
HispPerCap
NumUnderPov
PctPopUnderPov
PctLess9thGrade
PctNotHSGrad
PctBSorMore
PctUnemployed
PctEmploy
PctEmplManu
PctEmplProfServ
PctOccupManu
PctOccupMgmtProf
MalePctDivorce
MalePctNevMarr
FemalePctDiv
```

```
TotalPctDiv
PersPerFam
PctFam2Par
PctKids2Par
PctYoungKids2Par
PctTeen2Par
PctWorkMomYoungKids
PctWorkMom
NumKidsBornNeverMar
PctKidsBornNeverMar
NumImmig
PctImmigRecent
PctImmigRec5
```

Medincome & wage Education

UCI Communities and Crime Unnormalized Data Set

- Some Potential Questions to Answer

1) Analyze if number of vacant and occupied houses and the period of time the houses were vacant had contributed to any significant change in violent and non-violent crime rates in communities

2) How has unemployment changed crime rate(violent and non-violent) in the communities?

3) Were people from a particular age group more vulnerable to crime?

4) Does ethnicity play a role in crime rate?

5) Has education played a role in bringing down the crime rate?*

EDA

```
#Looking into the columns
crime_data.columns
```

```
Index(['Êcommunityname', 'state', 'countyCode', 'communityCode', 'fold',
       'population', 'householdsize', 'racepctblack', 'racePctWhite',
       'racePctAsian',
       ...
       'burglaries', 'burglPerPop', 'larcenies', 'larcPerPop', 'autoTheft',
       'autoTheftPerPop', 'arsons', 'arsonsPerPop', 'ViolentCrimesPerPop',
       'nonViolPerPop'],
      dtype='object', length=147)
```

```
#The median income
crime_data[['medIncome']]
```

	medIncome
0	75122
1	47917
2	35669
3	20580
4	17390
...	...
2210	24727
2211	20321
2212	27182
2213	19899
2214	23287

```
#Our dependent variable
crime_data[['ViolentCrimesPerPop']]
```

	ViolentCrimesPerPop
0	41.02
1	127.56
2	218.59
3	306.64
4	?
...	...
2210	545.75
2211	124.1
2212	353.83
2213	691.17
2214	918.89

2215 rows × 1 columns

```
crime_data
```

	Êcommunityname	state	countyCode	communityCode	fold	population	household
0	BerkeleyHeightstownship	NJ	39	5320	1	11980	
1	Marpletownship	PA	45	47616	1	23123	
2	Tigardcity	OR	?	?	1	29344	
3	Gloversvillecity	NY	35	29443	1	16656	
4	Bemidjicity	MN	7	5068	1	11245	
...	...	...	...	...	...	...	
2210	Mercedcity	CA	?	?	10	56216	
2211	Pinevillecity	LA	?	?	10	12251	
2212	Yucaipacity	CA	?	?	10	32824	
2213	Beevillecity	TX	?	?	10	13547	
2214	WestSacramentocity	CA	?	?	10	28808	

Variables Kept:

1. PctLess9thGrade
2. PctBSorMore
3. medIncome
4. pctWWage
5. pctUrban
6. racepactblack
7. racePctWhite
8. racePctAsian
9. racePctHisp
10. popDens
11. PctUnemployed
12. agePct12t29
13. PctPopUnderPov
14. state
15. PctNotSpeakEnglWell
16. MedRent
17. NumInShelters
18. NumInShelters
19. NumStreet
20. PolicCars
21. PolicOperBudg

```
variable_kept = ['ViolentCrimesPerPop', 'PctLess9thGrade', 'PctBSorMore', 'medIncome', 'pctWWage'
```

```
variable_kept = ['ViolentCrimesPerPop','PctLess9thGrade','PctBSorMore','medIncome','pctWWage',
                 'pctUrban','racepctblack','racePctWhite','racePctAsian',
                 'racePctHisp','PopDens','PctUnemployed','agePct12t29',
                 'PctPopUnderPov','state','PctNotSpeakEnglWell','MedRent','NumInShelters',
                 'NumStreet','PolicCars','PolicOperBudg','PolicAveOTWorked','PctNotHSGrad']
```

```
crime_data[variable_kept]
copy1 = crime_data[variable_kept]
copy1.head()
```

	ViolentCrimesPerPop	PctLess9thGrade	PctBSorMore	medIncome	pctWWage	pctUrban	ra
0	41.02	5.81	48.18	75122	89.24	100.0	
1	127.56	5.61	29.89	47917	78.99	100.0	
2	218.59	2.80	30.13	35669	82.00	100.0	
3	306.64	11.05	10.81	20580	68.15	0.0	
4	?	12.15	25.28	17390	69.33	0.0	

```
cars = copy1['PolicCars']
print(len(cars))
sum(cars == '?')
```

```
2215
1872
```

```
1872/2215
```

```
0.8451467268623025
```

```
holder = []
for i in copy1.columns:
  column = copy1[i]
  invalid = sum(column == "?")
  percentage = invalid / 2215
  if percentage >= 0.20:
    holder.append(i)
```

```
holder
```

```
['PolicCars', 'PolicOperBudg', 'PolicAveOTWorked']
```

```
copy1 = copy1.drop(columns = holder)
```

copy1

	ViolentCrimesPerPop	PctLess9thGrade	PctBSorMore	medIncome	pctWWage	pctUrban
0	41.02	5.81	48.18	75122	89.24	100.00
1	127.56	5.61	29.89	47917	78.99	100.00
2	218.59	2.80	30.13	35669	82.00	100.00
3	306.64	11.05	10.81	20580	68.15	0.00
4	?	12.15	25.28	17390	69.33	0.00
...	...	...	...	...	...	...
2210	545.75	17.12	15.79	24727	75.05	100.00
2211	124.1	12.51	19.28	20321	75.06	100.00
2212	353.83	7.82	12.42	27182	59.79	100.00
2213	691.17	24.37	12.40	19899	71.67	0.00
2214	918.89	13.93	8.86	23287	68.89	99.19

2215 rows × 20 columns

copy1

	ViolentCrimesPerPop	PctLess9thGrade	PctBSorMore	medIncome	pctWWage	pctUrban
0	41.02	5.81	48.18	75122	89.24	100.00
1	127.56	5.61	29.89	47917	78.99	100.00
2	218.59	2.80	30.13	35669	82.00	100.00
3	306.64	11.05	10.81	20580	68.15	0.00
4	?	12.15	25.28	17390	69.33	0.00
...	...	...	...	...	...	...
2210	545.75	17.12	15.79	24727	75.05	100.00
2211	124.1	12.51	19.28	20321	75.06	100.00
2212	353.83	7.82	12.42	27182	59.79	100.00
2213	691.17	24.37	12.40	19899	71.67	0.00
2214	918.89	13.93	8.86	23287	68.89	99.19

2215 rows × 20 columns

```
df = copy1[~copy1[copy1.columns].isin(['?'])]
```

```
df
```

	ViolentCrimesPerPop	PctLess9thGrade	PctBSorMore	medIncome	pctWWage	pctUrban
0	41.02	5.81	48.18	75122	89.24	100.00
1	127.56	5.61	29.89	47917	78.99	100.00
2	218.59	2.80	30.13	35669	82.00	100.00
3	306.64	11.05	10.81	20580	68.15	0.00
4	NaN	12.15	25.28	17390	69.33	0.00
...	...	...	...	...	...	...
2210	545.75	17.12	15.79	24727	75.05	100.00
2211	124.1	12.51	19.28	20321	75.06	100.00
2212	353.83	7.82	12.42	27182	59.79	100.00
2213	691.17	24.37	12.40	19899	71.67	0.00
2214	918.89	13.93	8.86	23287	68.89	99.19

2215 rows × 20 columns

```
df = df.dropna()
```

```
df
```

	ViolentCrimesPerPop	PctLess9thGrade	PctBSorMore	medIncome	pctWWage	pctUrban
0	41.02	5.81	48.18	75122	89.24	100.00

```
df.to_csv('/content/drive/My Drive/crimedata_try2.csv', index = False)
```

2	218.59	2.80	30.13	35669	82.00	100.00

```
df
```

	ViolentCrimesPerPop	PctLess9thGrade	PctBSorMore	medIncome	pctWWage	pctUrban
0	41.02	5.81	48.18	75122	89.24	100.00
1	127.56	5.61	29.89	47917	78.99	100.00
2	218.59	2.80	30.13	35669	82.00	100.00
3	306.64	11.05	10.81	20580	68.15	0.00
5	442.95	8.76	20.66	21577	75.78	100.00
...	...	...	...	...	...	...
2210	545.75	17.12	15.79	24727	75.05	100.00
2211	124.1	12.51	19.28	20321	75.06	100.00
2212	353.83	7.82	12.42	27182	59.79	100.00
2213	691.17	24.37	12.40	19899	71.67	0.00
2214	918.89	13.93	8.86	23287	68.89	99.19

1994 rows × 20 columns