NYPD and Crime Prevention in Transit: Analyzing the Effectiveness of Police Presence in U.S. Communities

By: Ray Portilla

**INTRODUCTION**

After a several fatal shootings that occurred in the NYC subway system in early 2024, the NYPD has added over 1,000 police officers into the subway tracks daily (Brachfeld, 2024), which has been the cause for discussion amongst New Yorkers. New Yorkers are known to be quite distrustful of the NYPD, with statements like, “defund the police,” taking charge after it was popularized by the Black Lives Matter Movement. That being said, it’s unsurprising that most commuters are displeased with the direction the MTA is taking, especially after recently raising fares from $2.75 to $2.90. Despite the growing allocation of resources to bolster law enforcement in train stations, questions remain about the effectiveness of these measures.

This analysis aims to evaluate the impact of heightened police presence in New York City’s transit system. By examining data from major US cities of police presence and other variables on crime rates, this research seeks to determine whether the increased deployment of officers is achieving its intended outcomes or simply represents an inefficient use of public funds. Using the results of our data, I will attempt to connect our findings specifically to NYC.

**LITERATURE REVIEW**

There have been many studies done to estimate the effect that police presence has on crime. In general, most studies tend to agree that police presence has a negative relationship with crime. For example, a systematic review of 49 studies measuring the effect of police presence found that police presence reduces most violent crimes (Dau, Vandeviver, Dewinter, et al., 2023).

Several theories attempt to explain this phenomenon. For example, the Deterrence Theory suggests that the presence of police makes committing crimes more impractical, which in turn reduces crime (Durlauf & Nagin, 2011). Other theories, such as the Broken Windows Theory, argue that visible signs of disorder, such as minor crimes or vandalism, encourage more crimes unless there are active police members (Wilson & Kelling, 1982).

However, some studies suggest that there are other more pressing variables to explain crime. A literature review conducted by Buonanno in 2003 asserted that decisions to commit crimes are rational decisions based on circumstances like inequality, poverty, level of education and any demographic background that can influence the aforementioned variables. Likewise, a study found that police presence can reduce crime in the short term but does not solve any of the underlying reasons why people are committing crime (Sherman, 1995).

Thus, studying the causes of crime and the impact of police is complex. One problem is that police presence is known to be an endogenous variable among scholars. Crime will increase police allocation and vice versa, making the correlation between the two difficult to interpret. As a result, many studies use instrumental variables as an attempt to view the real effect. A study by Weisburd in 2021 uses police calls as an instrumental variable for police allocation. Doing so they found that decreasing police presence increases crimes as well.

Keeping the past research in mind, this analysis will attempt to find the relationship between police presence and crime through a regression analysis that accounts for demographics and socio-economic factors that may bias our estimate. Additionally, this research will consider endogeneity through an instrumental variable model.

**DATA DESCRIPTION**

The dataset I will use for this analysis was obtained from Kaggle and contains demographic and crime-related data on 2,018 US communities. The dataset sources its’ data from the UCI Machine Learning Repo along with additional communities added in post. The dataset comprises of socio-economic data from the 1990 U.S. Census, law enforcement data from the 1990 U.S. LEMAS survey, and crime data from the 1995 FBI UCR.

It has 146 variables, which include demographics like population, race, age, household composition etc. and crime data like police allocation, budget and sworn officers. Although all demographic information is included, the police data has many missing values (over 75%).

**METHODOLOGY**

In this section, I would like to explore how predictive models can be used to better understand the relationship between police presence and crime. Below, I will outline the process of modeling a regression to analyze this effect, while addressing potential biases.

PART 1: Regression Analysis

Step 1: Limiting Regressor Variables

Step 2: Exploring Various Regression Models

PART 2: Instrumental Variable Analysis

Step 1: Validate my Instrumental Variable

Step 2: Imputing Missing Values

Step 3: Two-Stage Least Squares (2SLS) Analysis

PART 1: REGRESSION ANALYSIS

Step 1: Limiting Regressor Variables

To limit the variables provided by our dataset, I began by examining the F-statistics provided by f\_regression from scikit-learn and filtered out those with lower predictive power. I then conducted a simple OLS regression to identify and exclude more variables, guided by the findings from existing studies. Assuming that a simpler model is preferable, I filtered out variables based on their p-values (excluding police presence) until the adjusted R² began to decline. The final selection of variables, listed below, aligned with my research and intuition, so I will use this set for the remainder of this part.

1. PopDens: Population density of the community.
2. racepctblack: Percentage of the population that is Black.
3. racePctWhite: Percentage of the population that is White.
4. racePctHisp: Percentage of the population that is Hispanic.
5. agePct12t29: Percentage of the population aged 12 to 29.
6. medIncome: Median household income.
7. PctPopUnderPov: Percentage of the population living below the poverty line.
8. PctLess9thGrade: Percentage of the population with less than a 9th-grade education.
9. RentMedian: Median rent in the community.
10. NumInShelters: Number of individuals living in shelters.
11. PctUsePubTrans: Percentage of the population using public transportation.
12. PolicPerPop: Number of police officers per 100K population.
13. ViolentCrimesPerPop: Violent crimes per 100K population. (dependent variable)

Keep in mind, our variables of interest are PolicPerPop and ViolentCrimesPerPop which will be referred to as police presence and violent crimes respectively for the remainder of this analysis.

To test my models, I used both R²-adj for the OLS models and mean squared error calculated by how our predicted values compare to our test values.

To proceed with the analysis, the large subset of missing values must be addressed. After viewing the NA values, I could not find any indication that the values were missing not at random, so most of my analysis deals with the missing values by removing them. There are over 300 observations with non-missing values to use for this analysis.

Step 2: Trying Different Regression Models

To find the best model, I fitted my variables above to my dependent variable, violent crimes. Below are the models I used

1. OLS Regression
2. Polynomial Regression
3. K Neighbor Regression
4. Random Forest Regression

The results will be discussed in the next section.

PART 2: INSTRUMENTAL VARIABLE ANALYSIS

As mentioned before, police presence is known as an endogenous variable, which can lead to significant bias in our estimates. This dataset also has the problem of missing a significant number of values. To deal with both problems I decided to:

1. Impute the missing values for an instrumental variable.
2. Use those values to estimate PolicPerPop, effectively addressing endogeneity and our missing value problem.
3. Proceed with stage 2 of Two-Stage Least Squares (2SLS) analysis.

In doing so, we can ensure that there are less avenues of bias in our model. To have more points of reference, I will also conduct the 2SLS analysis without imputing any missing values.

Step 1: Validate my Instrumental Variable

To make the appropriate choice, an instrumental variable has to meet two criteria: it must be strongly correlated with the endogenous variable (PolicPerPop) and it also must be exogenous. I chose the variable PolicCars (police cars). To prove its relevance, I ran a regression of PolicCars on PolicPerPop. The results are shown below. With an F-stat of over 10 and a p-value under 0.01, I am confident in the police car variable’s ability to be a relevant instrument for police presence.

Regarding the second criteria, I believe police cars are exogenous because it is unlikely that it is related to the error term when estimating crime. More specifically, the number of police cars may only impact crime through the police allocation, making it a good instrument for police presence. Likewise, the car itself has no impact on crime if not for the policemen in it. As PolicCars meets both criteria, we will proceed to step 2.

Table 1: Relevance Test on IV: PolicCars

A screenshot of a computer

Description automatically generated

Step 2: Impute the missing values

To address the missing values for the PolicCars variable, I employed an Iterative Imputer from scikit-learn. There are several challenges that come up imputing missing values with one of them being that it assumes a relationship between the features which might not always be the case. Considering how we are imputing the instrumental variable, these errors can cause concern for the endogenous variable we will be estimating in the next step.

Step 3: Two-Stage Least Squares (2SLS) Model

Here, I will be doing a 2SLS model on the data without missing values and with the imputed values. The missing values 2SLS model was calculated through statsmodel in python and was fairly straight forward. I decided to limit the variables once more to include the those that give the highest R²-adj. The variables are shown in the regression summary in Table 4.

To find the 2SLS model from the imputed values I manually performed both stages using PolicCars’ imputed values making sure to estimate PolicPerPop to compensate for its previous missing values.

**RESULTS AND FINDINGS**

The results of each part’s regression models will be discussed and analyzed in this section. Six models were used in this analysis with the measures of fit listed below.

Table 2: Part 1 and 2 Models

|  |  |  |
| --- | --- | --- |
| MODEL | MEAN SQUARED ERROR | R²-adjusted |
| OLS Regression | 204041 | 0.575 |
| Polynomial Regression | 419081 | - |
| K Neighbor Regression | 212682 | - |
| Random Forest Regression | **161484** | - |
| 2SLS (NA removed) | - | 0.2620 |
| 2SLS (imputed values) | - | **0.544** |

The two best fitting models will be discussed from each part.

PART 1: REGRESSION ANALYSIS

OLS Regression

Table 3: OLS Regression

A screenshot of a computer

Description automatically generated

Table 3 shows the OLS regression of our set of independent variables (listed above) on our dependent variable, ViolentCrimesPerPop. Like much of the research done before, it indicates that demographics such as race, educational background and population density have a significant impact on crime. Interestingly enough, police per population is nowhere near a significant variable, likely caused by some omitted variable reflected in our error term. This will be addressed in our instrumental variable analysis. Regardless, here there is a negative correlation between police and crime, even though it is not strong.

The Random Forest Regression was our best fitting model with the lowest MSE. Although there are challenges in interpreting the effects of my variables, below is the feature importance of all the variables.

A graph with blue and white bars

Description automatically generated

Figure 1

As one can see, race tends to be the most important feature from our variables with percent under poverty following in after. As seen in the OLS regression, police presence is not as relevant as other variables, but still contributes almost 10% to the predicted value. As this had the lowest MSE, it may be safer to say that police presence has more of an impact on crime than the OLS regression may lead us to believe.

PART 2: INSTRUMENTAL VARIABLE ANALYSIS

My first 2SLS model was constructed by only using non-null values. The results are below.

Table 4

A screenshot of a computer

Description automatically generated

This model suggests that only 26.2% of the variance in violent crimes is explained by our model after using police cars to estimate police presence. This number has decreased from our last OLS model that did not have any instrumental variables. However, our variable, PolicPerPop is considerably more significant in this model than the previous one. This suggests that our instrument is addressing the endogeneity problem, making our estimate less biased.

The coefficient in the regression changed drastically, shifting from negative to positive. This suggests that police presence is positively correlated with crime, which contradicts findings from many other studies. One possible explanation for this discrepancy could be the limited number of observations used in the original analysis. To address this issue, I performed another 2SLS regression using imputed values to ensure a more robust analysis. The findings are shown below.

Table 5

A screenshot of a computer

Description automatically generated

In Table 5, the estimate of the endogenous variable, PolicPerPop, is denoted as “0”. After using most of the observations, the R²-adj increases to suggest that this model explains 54.4% of the variation in violent crimes. It also makes our variable of interest, police presence, significant at 10%, which is a huge step from the previous two OLS models. Out of our OLS models, this likely has the strongest potential to explain violent crimes through police presence even if other variables have more predictive power than police presence.

The estimate of the coefficient, however, is still positive which is a questioning discovery when interpreted literally. There are several explanations I would like to offer as to why the coefficient is positive.

1. Inaccurate Imputed Values – As mentioned before, a significant portion of my data was compromised due to missing values. These missing values may have made it more difficult for the imputer to estimate a proper relationship between the variables, thus providing biased estimates.
2. Insufficient Data – Our data was from several communities across the US. Differences in state culture or attitudes towards crime will not be reflected in the dataset but may be in the error term. Perhaps a panel dataset over time per state may have been better to estimate the effects that police have holding entity effects constant.
3. Police Deployment – Areas with higher crime are likely to have more police deployed. Although an instrumental variable model was used to curb this effect, a causal effect may still not be the correct interpretation of the coefficient.

Regardless, we can interpret this finding as, in areas with higher crime, there are more police.

**CONCLUSION**

The analysis of the relationship between police presence and crime rates has yielded complex results that can be interpreted to challenge conventional beliefs. While much of the existing literature suggests that increased police presence correlates with a reduction in crime, my findings show a positive correlation between the two. My findings are summarized below:

1. There is a positive correlation between police presence and crime.
2. Demographics and socio-economic indicators are more significant predictors of crime as shown by our Random Forest and IV OLS Regression.

For New York City, this finding is particularly important. The figure below illustrates crimes specifically on transit for this year, showing an upward trend despite the addition of over 1,000 police officers to the subway system in early 2024. Although this is not conclusive evidence, it raises questions about the traditional belief that increased police presence effectively reduces crime. Even if we were to assume a negative relationship between police presence and crime, the results might suggest that the impact of policing on crime rates may be overstated. Moreover, these findings point to the possibility that there are other, more significant factors contributing to the rise in crime, such as socio-economic conditions, inequality, or the effectiveness of law enforcement strategies.

A graph with a line

Description automatically generated

Figure 2

Considering these complexities, this analysis encourages a more nuanced approach to crime prevention in New York City, where the focus should not be on increasing police presence but also on addressing underlying socio-economic issues that drive criminal behavior.

**NEXT STEPS**

Further research and more detailed data analysis are needed to better understand these dynamics and to ensure that resources are allocated effectively for the long term. Below are a few suggestions for additional projects.

1. Panel Data with Longitudinal Analysis – Analysis with a new dataset that keeps time and entity effects fixed may provide us with a better estimate of the effect of police presence on crime.
2. More Variables – In this analysis I was limited to the variables given by the dataset. Perhaps including more socio-economic indicators may have explained more of our variance.
3. More Research on the Root Causes of Crime – This would allow us to better fit variables to our models to make more accurate predictions.

Through these next steps, future research can refine our understanding of the effectiveness of policing strategies and explore more holistic approaches to crime prevention.

**REFERENCES**

Buonanno, P. (2003). *The socioeconomic determinants of crime: A review of the literature* [Department working paper]. Department of Economics, University of Milan - Bicocca.

Dau, P. M., Vandeviver, C., Dewinter, M., et al. (2023). Policing directions: A systematic review on the effectiveness of police presence. European Journal of Criminology, 29(2), 191–225. <https://doi.org/10.1007/s10610-021-09500-8>

Durlauf, S. N., & Nagin, D. S. (2011). Imprisonment and crime: Can both be reduced?. *Criminology & Public Policy*, *10*(1), 13-54.

Kelling, G. L., & Wilson, J. Q. (1982). Broken windows. *Atlantic monthly*, *249*(3), 29-38.

Lin, M.-J. (2009). More police, less crime: Evidence from U.S. state data. *International Review of Law and Economics, 29*(2), 73–80. <https://doi.org/10.1016/j.irle.2008.12.003>

Mb. (2022, September 12). *Crimes in US communities dataset*. Kaggle. https://www.kaggle.com/datasets/michaelbryantds/crimedata

Sherman, L. W. (1995). Hot spots of crime and criminal careers of places. *Crime and place*, *4*(enero).

*Transit/Bus Crime Reports*. Reports - Transit-Bus Crime - NYPD. (n.d.). https://www.nyc.gov/site/nypd/stats/reports-analysis/transit-bus.page

Wang, Y. (2013). *Estimate the effect of police on crime using electoral data and updated data* (Master's thesis, University Name). All Theses. <https://digitalcommons.universityname.edu/theses/1677>

Weisburd, S. (2021). Police Presence, Rapid Response Rates, and Crime Prevention. *The Review of Economics and Statistics, 103*(2), 280–293. <https://doi.org/10.1162/rest_a_00889>