Impact of Household Income on Crime Rate

W 203: Lab 2

Dominic Lim, Emerald Swei, Shalini Chawla

Mar 29, 2022

Contents

1 Introduction

We have been tasked by the Department of Justice (DOJ) to find correlations between community characteristics and crime rates. The results could help to improve policy around policing. By identifying communities that are likely to have higher crime rates as a product of these community characteristics, the DOJ may be able to prioritize high crime rate communities in allocating its $29 billion budget. [1]

Public policy decisions that affect law enforcement budgets, preventative programs, and funding for the criminal justice system are driven by divergent causal theories on criminology. For the purpose of this study, we hold a working hypothesis that a lack of familial resources and economic opportunities compel individuals to commit crimes. To operationalize our study, we are choosing to focus on predictor variables that relate to socioeconomic factors. Principally, we will be focusing on median household income as the primary predictive variable for crime rate. The covariates we have chosen are based on research that indicates these other factors may also have an impact on crime rate, such as level of education, levels of unemployment, and recent immigration into the community.

Our analysis is focused on addressing the following research question:

> *Do neighborhoods with lower median household incomes have higher crime rates?*

2 Data and Research Design

We will be conducting an observational study to understand the relationship between social-economic factors and crime rates in US communities. For this analysis, we have chosen to use an existing dataset available through the University of Irvine data archive at Communities and Crime Unnormalized Data Set

This dataset combines socio-economic data from the 1990 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 FBI UCR.

The response variable for our analysis is the per capita crime rate in a community. Our source has data on 8 different crimes that are identified by FBI as Index Crimes. Out of these 4 types - murder, rape, robbery, and assault are categorized as violent crimes and and the other 4 - burglaries, larcenies, auto thefts and arsons are categorized as non-violent crimes. The per capita crime rate for each type has already been calculated using the population for the community. For our analysis we are going to sum all 8 individual crime rates into a single response variable - Total Crime Rate per Capita, where capita is defined as per 100,000 of the population.

Our source dataset had 2215 samples with 147 attributes. We had 313 samples with values missing for at least one of the crime categories. We considered dropping these samples from our analysis but these entries included many of the prominent cities so we instead decided to proceed with filling the missing attributes for those communities with the median values of the attribute. Our final dataset has preserved all 2215 samples.

Table 1: Accounting Table

Data Source	Count	Description
Original Dataset	2215	Total number of records in original data
Missing Crime Data	313	Samples with missing data for one or more crime categories
Final Dataset	2215	Missing values populated with median value - no samples dropped

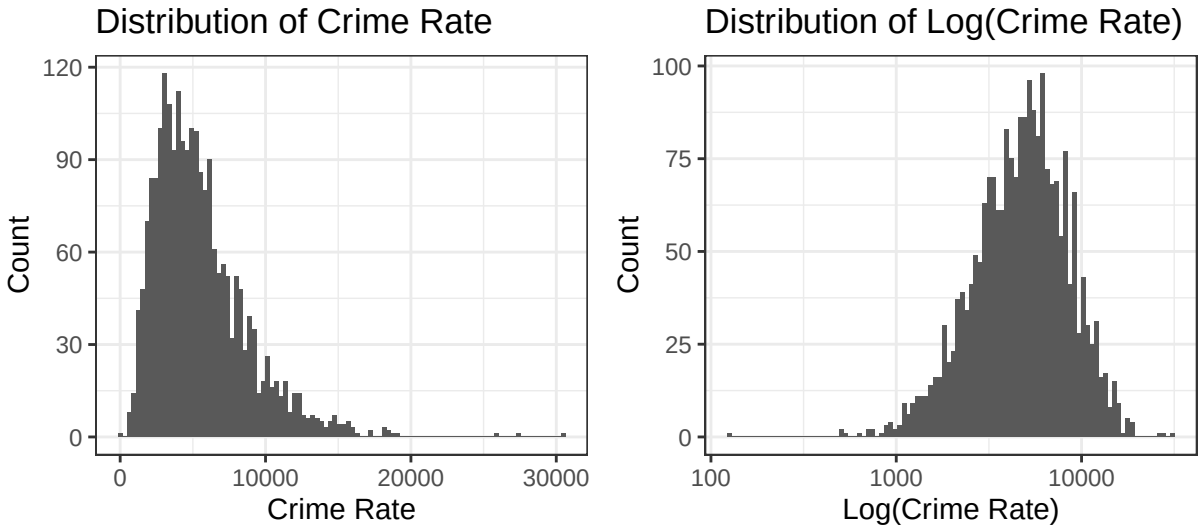
After filling in the missing values of crime rates in sub-categories with the median values for the corresponding categories, we have re-calculated the total crime rate as the sum of all 8 categories of crime rates and we will proceed with this set of 2215 samples for our analysis.

[1] U.S. Department of Justice FY 2020 Performance Budget https://www.justice.gov/jmd/page/file/1143926/download

3 Model Building Process

3.1 Distribution of Response Variable

The distribution of our response variable "Total Crimes per Capita" (totCrimesPerPop)in its original form is heavily right skewed. With the log transform, its distribution looks much closer to normal. We will proceed with using the log transform of the response variable for our models.



3.2 Build Models

The treatment variable of our study is the median household income of the community from the "medIncome" attribute in our data set. We have converted the dollar amount of median income to amount in 1000s to be used by our model(medIncomein000s).

- Our first model will study the relationship of our response variable "Total Crime Rate Per Capita" with a single treatment variable "Median Household Income" in the community.

$$log(totCrimesPerPop) = \beta_0 + \beta_1 medIncomein000s + \epsilon$$

Our base model shows that the Median Income is significant at the .001 alpha level which confirms our hypothesis. The coefficient for median income is -.024 which means that for every unit($1000) increase in median income, the crime rate goes down by 2.4%. The effect size may not be accurate here as there may be other factors in addition to income that impact the crime rate. We will include these addional variables as control in subsequent models.

- Besides income, there are many additional characteristics of a community that can significantly impact the crime rate. These additional variables will be our control variables in the second model. The covariates that we have identified to include in our second model are:
  - PctRecImmig5: percent of population who have immigrated within the last 5 years
  - PctNotHSGrad: percentage of people 25 and over that are not high school graduates
  - PctUnemployed: percentage of people 16 and over, in the labor force, and unemployed
  - PopDens: population density in persons per square mile

$$log(totCrimesPerPop) = \beta_0 + \beta_1 medIncomein000s + \beta_2 PctRecImmig5 + \beta_3 PctNotHSGrad + \beta_4 PctUnemployed + \beta_5 PopDens + \epsilon$$

- We have also identified additional variables that may potentially impact crime rate but may have linear relationship with our primary treatment variable and may explain the same variance. e.g.g more families with 2 parents may mean more income per household, more people under poverty level implies lower median. We will be building a third model with these variables in addition to the ones already included before:
  - PctFam2Par: percentage of families (with kids) that are headed by two parents
  - PctPopUnderPov: percentage of people under the poverty level
  - PctVacMore6Mos: percent of vacant housing that has been vacant more than 6 months
  - PctShelter: percent of population in homeless shelters
  - MedYrHousBuilt: median year housing units built

$$\begin{aligned} log(totCrimesPerPop) = \beta_0 &+ \beta_1 medIncomein000s + \beta_2 PctRecImmig5 + \beta_3 PctNotHSGrad \\ &+ \beta_4 PctUnemployed + \beta_5 PopDen + \beta_6 PctFam2Par + \beta_7 PctPopUnderPov \\ &+ \beta_8 PctVacMore6Mos + \beta_9 PctShelter + \beta_{10} MedYrHousBuilt + \epsilon \end{aligned}$$

3.3 Model Performance Evaluation

- The MSE(Mean Squared Residual) for model_1 is 0.2432341, for model_2, the value is 0.2257352 and for model_3, the value is 0.1569074, placing model_3 as our best performing model.
- Taking a look at the coefficient tests for the models, Median Income is significant at .001 alpha level in all 3 models with its coefficient going down for -0.024 to -0.021 to -0.005.
- We compare model_2 to model_1 using the anova() function to see if model_2 is a better representation of our population. Then we compare model_3 with model_2 to see if model_3 is a better fit. Both F-Tests return a p value lower than .05 indicating that the fuller model (the model with more variables) is better than the one with fewer variables again placing our third model as the best out of the three.

4 Results

4.1 Model Estimates

4.2 Statistical Significance

We can see with each increase in variables for each model that there is an increase in fit to the data. With only one variable, median income, we achieve an R^2 of 0.301 and as we increase to ten variables with our "kitchen sink model," R^2 increases to 0.547. At each stage of increasing the number of variables in the model, the R^2 increases. In the first model, median income has a coefficient of -0.024, with a very low p-value. It has a small and negative relationship with crime rate. In the second model, median income has a very similar coefficient of -0.021. The other variables, PctRecImmig5, PctNotHSGrad, PctUnemployed, and PopDens have coefficients of 0.035, -0.00082, 0.030, and 0.0000077 respectively. Of these variables, only PctRecImmig5 and PctUnemployed had p-values that suggested significance.

Median income remains significant through all three models, and percentage of immigrants (PctRecImmig5) and population density (PopDens) are significant through the two models they are included in. In the last

Table 2: Impact of median income on crime rate

	Dependent variable:		
	Crime Rate Per Capita		
	(1)	(2)	(3)
Median Income in 000s	-0.024^{***}	-0.021^{***}	-0.005^{***}
	(0.001)	(0.001)	(0.001)
Percent Immigrants		0.035^{***}	0.027^{***}
		(0.006)	(0.005)
Percent Non HS Graduates		-0.001	0.001
		(0.002)	(0.001)
Percent Unemployed		0.030^{***}	-0.008
		(0.005)	(0.005)
Population Density		0.00001	-0.00001^{**}
		(0.00001)	(0.00000)
Percent 2 Parent Families			-0.038^{***}
			(0.002)
Percent Under Poverty			-0.00002
			(0.002)
Percent Vacant Homes			-0.004^{***}
			(0.001)
Percent Living in Shelters			0.225^{***}
			(0.070)
Median Yr House Built			0.006^{***}
			(0.001)
Constant	9.270^{***}	8.916^{***}	-0.106
	(0.034)	(0.080)	(1.841)
Observations	2,215	2,215	2,215
R^2	0.302	0.352	0.550
Adjusted R^2	0.301	0.350	0.547
Residual Std. Error	0.493 (df = 2213)	0.476 (df = 2209)	0.397 (df = 2204)
Note:			*p<0.1; **p<0.05; ***p<0.01

model, the significance of the percentage unemployed (PctUnemployed) variable drops and the coefficient actually flips to a negative one, from 0.030 to -0.008. Also in the third model, the percentage of families with two parents (PctFam2Par), percentage of vacant homes (PctVacMore6Mos), percentage living in shelters (PctShelter), and the median year of houses built (MedYrHousBuilt) are all significant. The only variable added to the third model that is not significant is the percentage under poverty (PctPopUnderPov).

4.3 Practical Significance

We demonstrated from the increasing fit of our models that more factors than median income contribute to the crime rate of a community. This may be something we intuitively knew already, but have shown through statistical modeling is true. As we increased the number of variables considered, we saw an increase in the fit of the model through the R^2. We also noticed that when we included variables representing the percentage unemployed (PctUnemployed) and percentage under poverty (PctPopUnderPov) in the third model, these variables did not seem to be significant even though our intuition was that they would be. While reasoning why this may have occured, we may guess that some of the contribution that these variables would have provided have actually been captured in the median income variable. Residents who are unemployed or under the poverty line would by definition likely have low or no income, and this impact could have already been measured in the median income. Since these variables have some collinearity, we could expect that their significance would be dampened by the existence of a collinear variable in the model.

Lastly, we found that almost all of the coefficients for our variables were small numbers, with absolute values between 0 and 1. The variables with the strongest coefficients were percentage of recent immigrants (PctRecImmig5) and percentage of two parent families (PctFam2Par), which had coefficients of 0.027 and -0.038 respectively in the third model. It would be interesting to examine the specific cities or communities with a high percentage of recent immigrants to try an infer whether there is a reverse causal relationship or some omitted variable that might be feeding the strength of this predictor. We know that there is some collinearity between median income and the percentage of two parent families; in future research, we might want to take a look at the percentage of two parent families variable alone to see whether or measure how much median income is absorbing its significance.

5 Limitations

5.1 Statistical Limitations

We have 2215 samples in our dataset making it a candidate for a large sample model. The two assumptions that must be met for the large sample model are as follows:

1. I.I.D Data
2. Unique BLP Exists

Independent and Identically Distributed (I.I.D)

There are a number of concerns with I.I.D as it relates to the dataset. The dataset, which merges socio-economic data from the 1990 Census and crime data from the 1995 FBI Uniform Crime Reporting dataset, excludes observations that are missing Census or Crime Reporting data. As a result, our observations include communities with populations of at least 10,000 residents and police departments that are complying with the Uniform Crime Reporting standards. This may introduce a bias to the dataset in which less populous communities and/or smaller police departments are being excluded. This is made clear in comparing the sum of populations in our entire dataset compared to the actual population tabulated in the 1990 Census. (117.7M vs 248.7M) [2]

[2]Stat AB of U.S2 - Census.gov. https://www2.census.gov/prod2/statcomp/documents/1991-02.pdf.

In terms of independence, geographic data can exhibit clustering effects as the socio-economic factors and crime rates of a given community may have an impact on neighboring communities. As we can see below, our plot shows that our residuals suggest homoscedasticity (even if we formally fail based the Breusch–Pagan test). Out of caution, we have chosen to use Robust Standard Errors.

Unique BLP Exists

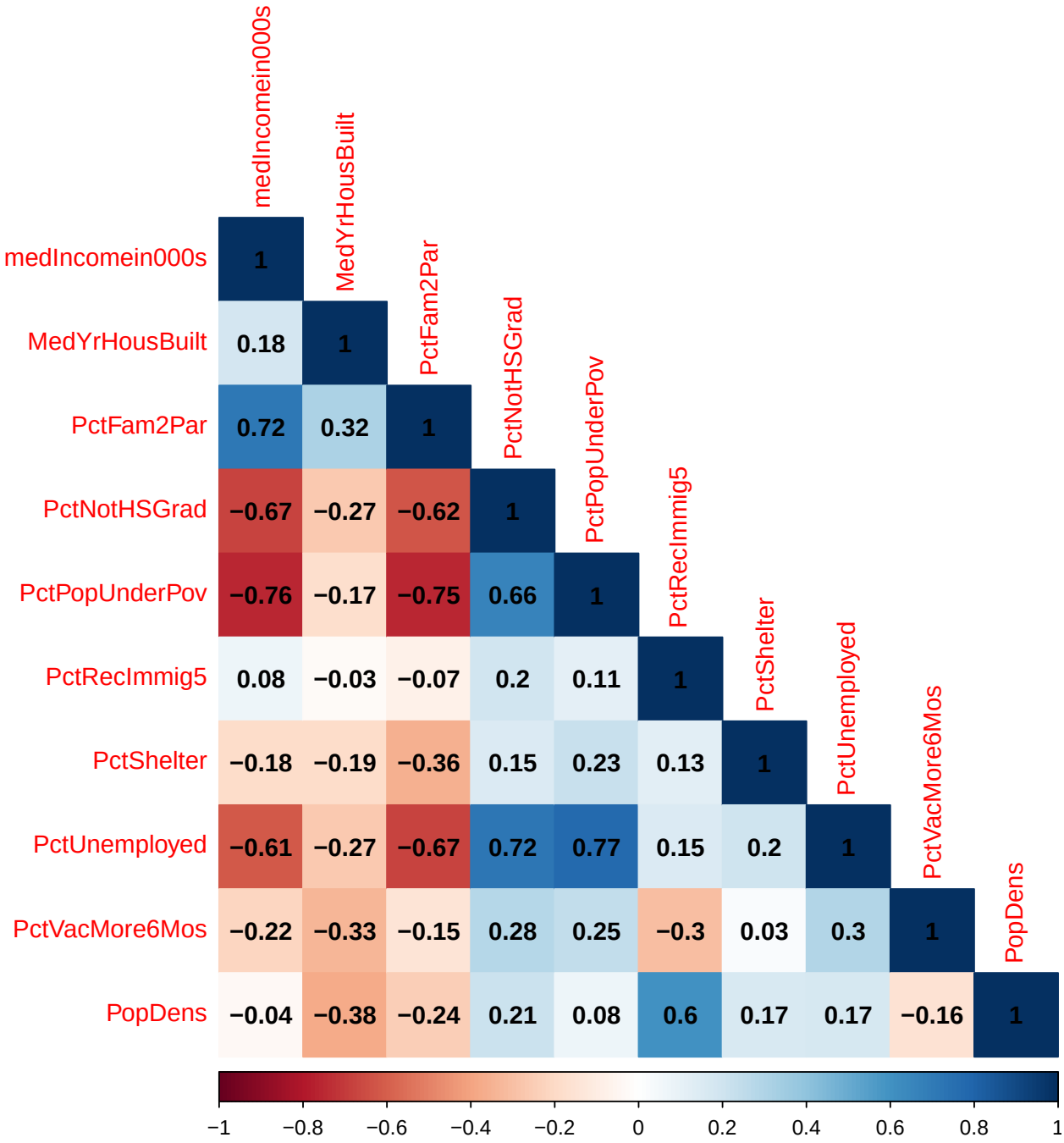
The existence of a unique BLP implies that there is no perfect collinearity with the variables of choice. For this assumption to be met, no variable can be written as an exact linear combination of any other variable(s). We can conclude that a unique BLP does exist after interrogating these assumption by looking at the following:

1. The VIF (Variance Inflation Factor)
2. The Correlation Matrix showing no perfect collinearity

The VIF for most of our covariates are reasonable (below 4) except PctPopUnderPov with a value of 4.35. We expect this variable to have some relationship to the median income, since a higher median income of a community would indicate that there are less people under the poverty line and more people living under the poverty line would likely bring down the median income of a neighborhood. As such, we can understand that there would be some collinearity between the two variables.

```
## medIncomein000s    PctRecImmig5    PctNotHSGrad   PctUnemployed         PopDens
##        3.366165        2.159610        2.871300        3.284123        2.161771
##      PctFam2Par  PctPopUnderPov  PctVacMore6Mos      PctShelter  MedYrHousBuilt
##        3.456131        4.384764        1.523340        1.191277        1.561858
```

When we examined the correlation plot below, we found that the correlation between the socioeconomic variables tended to be higher, around 0.7 on average. For example, the correlation between medIncome and PctNotHSGrad, PctFam2Par, and PctPopUnderPov are all around 0.7. Furthermore, the highest correlation value of 0.77 is between PctPopUnderPov and PctUnemployed. Practically, these variables likely have collinearity as factors such as income, employment, and higher education are generally related in our society. The percentage of residents under the poverty line directly has a relationship with median income, as by definition those who are under the poverty line are below some federally established income level. The relationship between median income and the percentage of families with two parents is also not surprising because a household with two parents is typically more likely to have a higher combined household income than a household with only one parent. Seeing these variables in the context of our society, we can see how they could either influence or otherwise have correlation with one another.



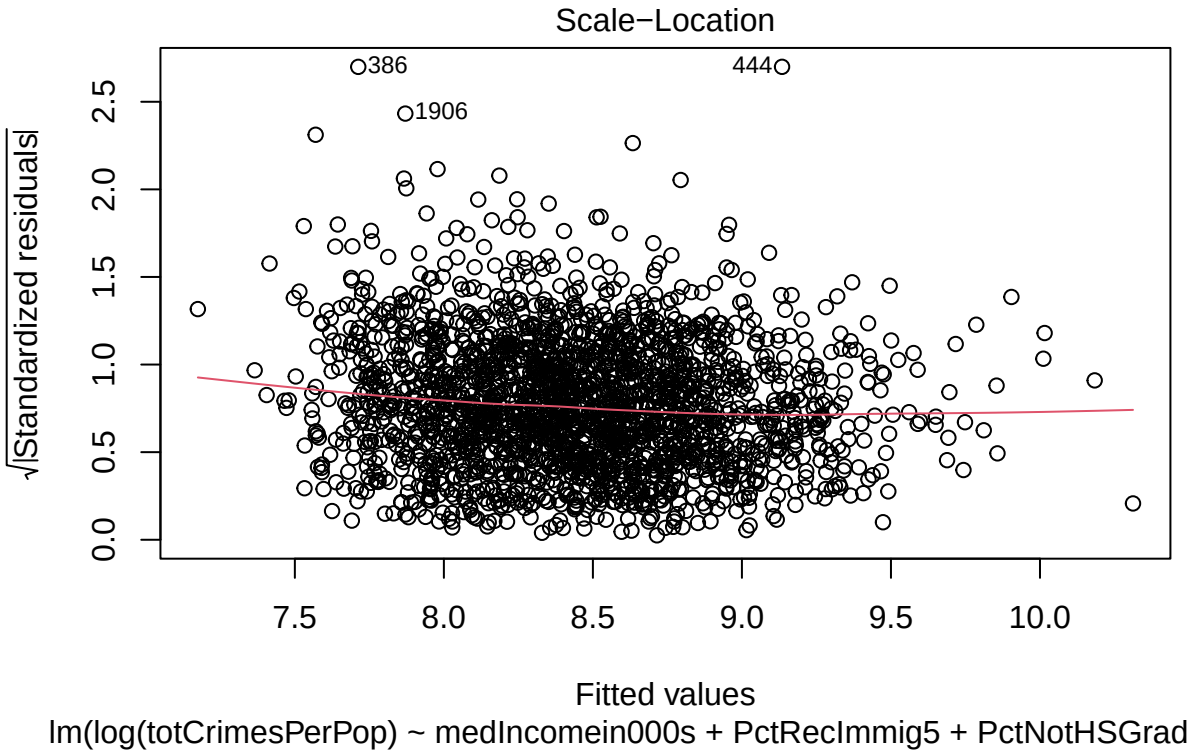
5.2 Additional Assumptions:

In addition to the two assumptions of our large-sample model, we wanted to consider additional assumptions that may strengthen our trust in our estimators.

Homoscedasticity of Residuals

To test for homoscedasticity, we conducted the Breusch–Pagan test in which we rejected the NULL hypothesis of homoscedasticity. However, from the below Scale-Location plot, we can see that the residuals are spread equally along the ranges of predictors. There is a general linear quality to the plot suggesting Homoscedascitiy.

```
##
##  studentized Breusch-Pagan test
##
## data:  model_3
## BP = 38.622, df = 10, p-value = 0.00002956
```

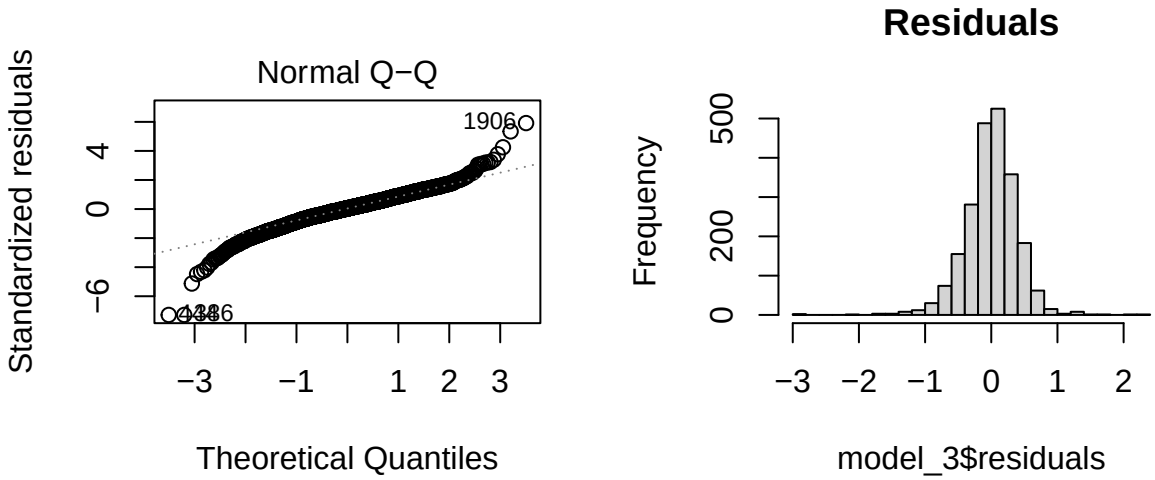


Normality of Errors

To test for Normality of Errors, we conducted the Shapiro–Wilk test and Jarque–Bera test in which we rejected the NULL hypothesis of Normal Distribution of Residuals. However, from the below Normal QQ plot and Histogram of Residuals, we can see that the residuals are appear normally distributed. One thing to note is that Normal QQ plot shows some curvilinearity in the tails, suggests that there are some extreme values at the tails.

```
##
##  Shapiro-Wilk normality test
##
## data:  sample(model_3$residuals, size = 2215, replace = TRUE)
## W = 0.95757, p-value < 0.00000000000000022
```

```
##
##  Jarque-Bera Normality Test
##
## data:  model_3$residuals
## JB = 2026, p-value < 0.00000000000000022
## alternative hypothesis: greater
```



5.3 Structural Limitations

1. We intended to include the variable LemasSwornFT(number of sworn full time police officers) in our analysis, but our source dataset is missing this data for 1872 out of the 2215 samples. Since the number of police officers can not be estimated, we did not find it reasonable to estimate this value and had to exclude it form our models. Since the number of police officers in a community can have an impact on managing the crime rate, we believe that this missing information has introduced Omitted Variable Bias in our models. Since the strength of police force has a negative relationship to crime rate and higher income areas can have more budget for police hence more police officers, the overall bias is -ve(+ve * -ve), pushing our coefficient away from zero (the coefficient for median income is negative). Since the description of our dataset states that all communities with large police departments(with 100+ officers) have the data included and only a random sample of the smaller police departments have the data available. This implies that the communities with the missing data all have small police departments(< 100 officers). We can use this information to create an additional binary variable for police department size as small vs large and include it in our model to mitigate some of the OVB.

2. Another omitted variable is the number of visitors. Communities with large numbers of visitors will have higher per capita crime than communities with fewer visitors. And more visitors also mean more business for local residents bringing the median income up. The overall bias in this case is +ve (+ve * +ve), pushing our coefficient towards zero.

3. Another omitted variable is drug and alcohol usage which may impact multiple coefficients. Communities with greater drug and alcohol usage will likely have higher per capita crime, especially drug related crime. Higher alcohol and drug usage may also adversely affect employment and income. Finally, drug usage might impact law enforcement presence as additional narcotics officers are tasked to combat the spread of illicit substances. The overall bias in this case of substance usage pushes our income coefficient away from zero.

6 Conclusion

The goal of our study was to understand the impact of Median Household Income in a community on the Crime Rate. To understand this relationship, we analyzed the data set curated from a combination of three different data sources - socio-economic data from the 1990 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 FBI UCR. With median income being our primary treatment variable, we used additional covariates to control for additional factors impacting the crime rate.

As shown by all three models, median income does have a significant impact on the crime rate in a community, even though in reality the impact is not as large as it is in our first model. Median income by itself does not completely explain the full variance in crime rate; instead, there are many different factors correlated to the income and wealth distribution in a community including population under poverty, 2 parent families and unemployment rate that together have an impact on overall crime rate. Our study confirms our hypothesis and shows evidence that supports the idea that lack of familial resources and economic opportunities may compel individuals to commit crimes.

According to our selected Model 3:

> *A $1,000 USD increase in median household income is associated with a 0.5% percent reduction in total crime rate, ceteris paribus.*

This information is practically important for the DOJ and policy makers to direct additional budget and create policies toward bringing in more jobs and skill training opportunities to lower income communities.

We would also encourage additional investigation in which the impact of household income on crime rates is studied through direct intervention. This may come in the form of experimentation with tax breaks and incentives for employers to provide jobs in lower income communities, or direct cash assistance akin to a universal basic income. With an estimated annual aggregate cost of $2.8 trillion[3] to the U.S. economy from violent and non-violent crimes, we believe that such preventative measure may help to reduce crime and its impact on our communities.

[3]Anderson, David A., and Centre College. "The Aggregate Cost of Crime in the United States: The Journal of Law and Economics: Vol 64, No 4." The Journal of Law and Economics, 1 Nov. 2021, https://www.journals.uchicago.edu/doi/abs/10.1086/715713.