Addressing the reverse causation between the occurrence of crimes and income distributions with an instrumental variables approach

Alexander MAAS and Ryo INOUE

Abstract: The focus of this paper is how the income distribution of an area affects the occurrence of various types of crime in that and surrounding areas. This is done by utilising Instrumental Variables (IV) and Compositional Data with a Total (CoDa with a total) approaches. The results show that the occurrence of violent and property crimes are associated more with areas with a larger proportion of low or high income households respectively.

Keywords: Crime, Income, Instrumental Variables, Two-Stage Least Squares, Compositional Data

1. Introduction

Many researchers from different fields are interested in the study of crime for different reasons, from the economic impact of crime on the surrounding area to the demographic traits that are indicative of a higher crime area.

A problem that needs to be addressed is the endogeneity between crime and income and crime and population. This means that crime affects who lives in an area and who lives in an area affects the crime that occurs there and the larger the population in an area the more crime that can be expected to occur in that area but the more crime that occurs in an area the fewer residents who wish to live there. Failure to address this reverse causation will cause biased coefficient estimates and justifies using an instrumental variables approach.

This paper will investigate how the income distribution of an area is related to the frequency and types of crimes that occur in that area within NYC comparing the 2009-13 period with 2014-18. The originality of this research is in its focus on the income distribution rather than average income to better capture the interactions between the upper and lower class and the instrumental variables approach using land-use to instrument for income.

Alexander Maas

Tohoku University Graduate School of Information Sciences

maas.alexander.christian.p8@dc.tohoku.ac.jp

2. Literature Review

The focus of this paper on the income distribution and crime relationship means the relevant literature can broadly be broken into two categories: those that account for the endogeneity problem common in crime analyses and those that investigate the crime and demographics relationship. A brief overview of each is given below.

2.1 Crime and Endogeneity

The common theme among the papers with a focus on the crime and real estate prices link is their recognition that an endogeneity problem exists in their data which they need to address to obtain accurate estimates. Ihlanfeldt and Mayock (2010) focus on this problem and even offer a critique of existing literature's inadequate handling of the endogeneity problem. While the problem itself is known to authors, the difficulty in finding appropriate instruments causes many authors to ignore the problem entirely. If appropriate instruments are found, they are rarely tested to ensure their relevance.

2.2 Demographics and Crime

Most of the papers in this group investigate the effect that gentrification has on crime, for example

Papachristos et al. (2011) or that crime has on the gentrification process (Ellen et al., 2019). Despite the reverse causation between crime and the demographic variables that coincide with gentrification, very few papers try to account for endogeneity in their models. Another issue with these papers is the lack of clear definition of what gentrification actually refers to, how it is measured or if it is capable of being measured at all.

Despite these differences, the papers arrive at largely consistent results. An increasing income is associated with decreasing crime. Kreager et al. (2011) additionally found that initially there may be a slight increase in crime as friction between old and new residents increases while McIlhatton et al. (2016) found that higher income neighborhoods are associated with fewer violent crimes but more property crimes.

3. Methodology

This paper uses two distinct methodologies to prepare and model the data. The first deals with endogenous variables, and the second with compositional data.

3.1 Endogeneity Problem

By using variables which are correlated with the error term, the obtained coefficients will be biased upward or downward depending on the relationship between the endogenous and dependent variables. If such a situation exists, an instrumental variables approach can be used to produce unbiased estimates. For an instrument to be valid, it must satisfy several criteria: strong correlation must exist between the chosen instrument and the endogenous variable, the instrument cannot be endogenous, and the instrument cannot be an explanatory variable in the regression. If these conditions are met, an instrumental variables approach may output unbiased estimates.

This paper uses the two-stage least squares (2SLS) method. This method first takes the endogenous variable as the dependent variable and regresses it on all other exogenous variables and instruments. The fitted values from this first stage regression are then used in place of the endogenous variables in the second stage regression along with all other exogenous variables. The resulting coefficients are unbiased but possess a larger standard error than an ordinary least squares method would produce.

3.2 Compositional Data

Compositional data is data where the components are parts of a whole and can be represented as the relative size of that component relative to some total or constraint. The problem with modeling composition data is the shared information in all the components, for example the total in the denominator, and the total constraint which prevents a single component from being unilaterally increased or decreased.

This paper uses the compositional data with a total (CoDa with a total) method (Coenders et al., 2017). While traditional compositional data analysis is only concerned with relative information contained in the components, CoDa with a total is able to also include the absolute information contained in the total in the analysis.

4. Data

The analysis is conducted on data which covers New York City census tracts spanning the years 2009-2018 excluding Staten Island. A census tract contains approximately 4,000 residents however this could be smaller or larger. The data used includes: information on the geographic boundaries of the census tracts and police precincts, complaint data from the NY Police Department (NYPD) which has information on all reported offences, income data from the US Census Bureau and data on the land-use of all properties in NYC.

Since the census tract is the spatial resolution of interest to this paper, the information on the census tract geographic boundaries is used to aggregate the crime

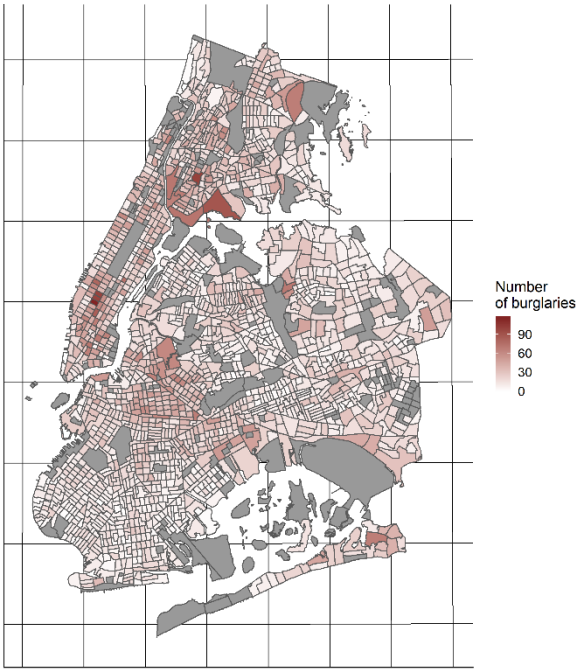

Figure 1. Number of burglaries by census tract in 2016-18

count data for later analysis. The police precinct boundaries are used to account for fixed effects around each census tract.

The crime data reported by the NYPD covers all reported offences, but this paper is only interested in six major felony offences: felony assaults, murders and manslaughters, grand larcenies, grand larcenies of motor vehicles, burglaries and robberies. Rape was excluded from the analysis due to the coordinates reported for the offence being the precinct headquarters not the actual location the offence occurred. Within the crime data, the analysis utilizes the information regarding the year the crime occurred and the coordinates of the occurrence. This information is aggregated at the census tract level to arrive at a crime count for each offence type for each census tract. To eliminate the presence of excess zeros and reduce the variability introduced into the data by short term crime counts, 2 three-year periods are used for this aggregation, 2011-2013 and 2016-2018. An example of this output is shown in **Figure 1** for burglaries for the 2016-18 period.

The income data used, from the US Census Bureau, reports the estimated number of households in each census tract as well as the proportion of households which belong to each income category. The income categories represent the proportion of the total number of households from that census tract whose yearly income fall within that category. For this paper, these categories are combined to reduce the total number of categories to two, the number of households that earn less than $35,000 a year and the number who earn over $35,000 a year. This $35,000 a year is approximately equal to the poverty threshold in NYC as determined by the Mayor's Office.

The final source of data is the information on land-usage within NYC from the Department of City Planning which will be used as the instruments for the 2SLS method. For each property lot in NYC the lot size, property size and land-use category are reported along with a large amount of additional information not relevant to this paper. There are 219 different land-use categories within NYC, many being unique to a certain area or representing only a small number of properties. This paper will only utilize the residential land-use categories of which there are 58. Since the income data is divided into high- and low-income, the same split is desired in the land-use data. To accomplish this the 58 categories are combined to make two new categories which roughly correspond to high density excluding condominiums, such as subsidized apartments for low-income households, and other residential. To better reflect the available residential space of each property, the larger of lot size or property size is selected to try and account for multistory and single-family building's different structures. For each census tract the total square footage of each of the two categories is calculated.

5. Model

The first step in the analysis is the creation of the variables which measure the income distribution and

population of the census tracts using CoDa with a total. How these variables are calculated is shown in Eq.(1) and Eq.(2). The variable *y* is the measure of the income distribution of a census tract and is the logratio of the number of low-income households, $x_{under\ 35k}$, to the number of high-income households $x_{over\ 35k}$.

$$y = \sqrt{\frac{1}{2}} \ln \frac{x_{under\ 35k}}{x_{over\ 35k}} \quad (1)$$

The variable *t* is the measure of population and is calculated as the log of the number of low-income households, $x_{under\ 35k}$, plus the log of the number of high-income households, $x_{over\ 35k}$.

$$t = \frac{1}{\sqrt{2}} (\ln(x_{under\ 35k}) + \ln(x_{over\ 35k})) \quad (2)$$

The instruments used for the 2SLS method are also calculated using CoDa with a total and use the land-use area for each census tract. In Eq.(3) and Eq.(4), $x_{high\ density}$ is the square footage of high density residential properties excluding condominiums and x_{other} is the square footage of other residential properties. These instruments both pass the validity test proving their relevance to the *y* and *t* variables calculated in Eq.(1) and Eq.(2) above and justifying their use as instruments.

$$y^{IV} = \sqrt{\frac{1}{2}} \ln \frac{x_{high\ density}}{x_{other}} \quad (3)$$

$$t^{IV} = \frac{1}{\sqrt{2}} \left(\ln\left(x_{high\ density}\right) + \ln(x_{other})\right) \quad (4)$$

Once the *y* and *t* variables have been calculated their endogenous nature is addressed using 2SLS. In the first stage models, which are shown in Eq.(5) and Eq.(6), **β** are the coefficients and $\mathbf{X_{IV}}$ contains both instruments, the year and police precinct dummy variables and *y* and *t* variables calculated based on data from neighboring census tracts designed to capture spillover effects.

$$\hat{\boldsymbol{y}} = \boldsymbol{X_{IV}}\boldsymbol{\beta}_y \quad (5)$$

$$\hat{\boldsymbol{t}} = \boldsymbol{X_{IV}}\boldsymbol{\beta}_{\boldsymbol{t}} \quad (6)$$

The second stage then uses the fitted values from Eq.(5) and Eq.(6) as explanatory variables in the Poisson model shown in Eq.(7). Here, **C** is the crime counts of the census tracts, **β** are the coefficients and **X** contains the fitted values from the first stage, the year and police precinct dummy variables and the *y* and *t* variables calculated from the neighboring census tracts.

$$\mathbf{E}(\mathbf{C}) = \exp(\mathbf{X}\boldsymbol{\beta}) \quad (7)$$

6. Results

While demographics and crime are considered to be endogenous the construction of the variables used in this analysis is unique so this endogeneity should be confirmed. To do this the Hausman-Wu test is used. The results of this test show that for the property crimes the *y* and *t* variables are endogenous, interestingly however, the violent crimes show no endogeneity problem. This means the property crimes are better modeled using the 2SLS method while the violent crimes are better modeled with the standard Poisson method without any use of instruments. The results of the 2SLS regressions are reported in **Table 1** for property crimes and the results of the standard Poisson model are reported in **Table 2** for violent crimes. Upon inspection some patterns become immediately apparent.

The $\hat{t}$ and t_{nb} variables which measure the population of the target census tract and the surrounding census tracts respectively are all positive indicating that an increasing population, both within and without the target census tract, is associated with an increasing

number of crimes of all types.

The $\hat{y}/y$ and y_{nb} variables which are representing the income distribution of both the census tract and the surrounding census tracts respectively show mixed results for the effect of an increasing proportion of low-income households. $\hat{y}/y$ shows that a census tract with a higher proportion of high-income households is positively associated with the property crimes, theft, car theft and burglary and that a predominately lower income census tract is associated more with the violent crimes, assault, robbery and murder. However, the neighboring tracts all show the same effect, a higher proportion of low-income households in the surrounding areas is associated with an increase in all crime types in the target census tract.

The year dummy reported is taking 2013 as the base year. Therefore, a negative year dummy coefficient represents a decrease in crime because of time passing. This decrease is found for all crime types except assaults which experienced an increase in frequency. This could be due to several reasons such as a general improving economic condition or improving security measures. So, while the results show the majority of crimes decrease over time the reasons behind this are uncertain.

The precinct fixed effects are predominately significant and can successfully capture some of the spatial variation and clustering that the crime data displays.

Table 1. 2SLS Results for property crimes

	Theft		Car Theft		Burglary	
Term	Est.	Sig.	Est.	Sig.	Est.	Sig.
Inter.	-1.43	***	-3.31	***	-1.60	***
$\hat{\boldsymbol{y}}$	-0.53	***	-0.85	***	-0.58	***
$\hat{\boldsymbol{t}}$	0.33	***	0.45	***	0.36	***
$\boldsymbol{y}_{\mathbf{nb}}$	0.29	***	0.49	***	0.46	***
$\boldsymbol{t}_{\mathbf{nb}}$	0.20	***	0.12	***	0.13	***
Year	-0.03	***	-0.41	***	-0.43	***
*** p < 0.001, ** p < 0.01, * p < 0.05, . p < 0.1, p > 0.1						

Table 2. Poisson Results for violent crimes

	Murder		Assault		Robbery	
Term	Est.	Sig.	Est.	Sig.	Est.	Sig.
Inter.	-6.65	***	-2.03	***	-2.58	***
y	0.53	***	0.40	***	0.18	***
t	0.43	***	0.35	***	0.30	***
$\boldsymbol{y}_{\mathbf{nb}}$	0.45	***	0.52	***	0.49	***
$\boldsymbol{t}_{\mathbf{nb}}$	0.18	***	0.20	***	0.29	***
Year	-0.24	***	0.15	***	-0.25	***
*** p < 0.001, ** p < 0.01, * p < 0.05, . p < 0.1, p > 0.1						

7. Conclusion

To find the relationship between the income distribution of an area and the crimes which occur in that area an instrumental variables approach was utilized to address the endogeneity which exists between income and crime and correct bias which would result from using more traditional methods.

The results obtained from this analysis are largely in line with the previous literature. However, previous literature often focuses on average income levels which fails to capture the complete picture of what the income situation in the area is. To better address this structure of the income distribution within census tracts, this paper has utilized an approach which not only accounts for the reverse causation between crime and income but successfully separates the effects of population size and income distribution to allow for the individual effects to be examined.

Future research could additionally consider the effect of the racial composition of an area on the occurrence of crime and incorporate average income to better explain and isolate the income distributions effect.

References

Coenders, G., Martín-Fernández, J.A. and Ferrer-Rosell, B., 2017. *When relative and absolute information matter: compositional predictor with a total in generalized linear models. Statistical Modelling*, **17**, 494-512.

Ellen, I.G., Horn, K.M. and Reed, D., 2019. *Has falling crime invited gentrification?. Journal of Housing Economics*, **46**.

Ihlanfeldt, K. and Mayock, T., 2010. *Panel data estimates of the*

effects of different types of crime on housing prices. *Regional Science and Urban Economics*, **40**, 161-172.

Kreager, D.A., Lyons, C.J. and Hays, Z.R., 2011. *Urban revitalization and Seattle crime, 1982-2000*. *Social Problems*, **58**, 615-639.

McIlhatton, D., McGreal, W. de la Paz, P.T., and Adair, A., 2016. *Impact of crime on spatial analysis of house prices: evidence from a UK city*. *International Journal of Housing Markets and Analysis*, **9**, 627-647.

Papachristos, A.V., Smith, C.M., Scherer, M.L. and Fugiero, M.A., 2011. *More coffee, less crime? The relationship between gentrification and neighborhood crime rates in Chicage, 1991 to 2005*. *City and Community*, **10**, 215-240.