Analyzing Income Levels by PUMAs

Afwan Kibria and Harrison Wang

How are income levels affected by education in a place like New York City? Oftentimes we arrive at these logical conclusions about economics and relationships through careful observations of micro and macro trends, but how do these thought processes line up within a more condensed timeframe? To investigate further we took aggregate summary statistical data from 55 Public Use Microdata Areas (PUMAs), which encompass and resemble the various Community Districts, albeit reshaping them to obtain data from population groupings in excess of 100,000 people. We then sourced averages for income, crime (number of major felonies committed), population 3+ years enrolled in school (from nursery to graduate school), population percentage with a Bachelor’s degree or higher, and unemployment rates for those of legal working age. Below were the summary statistics, indicating mean, median, maximums, minimums, standard deviations, and variance to depict the level of variability among data points across the board.

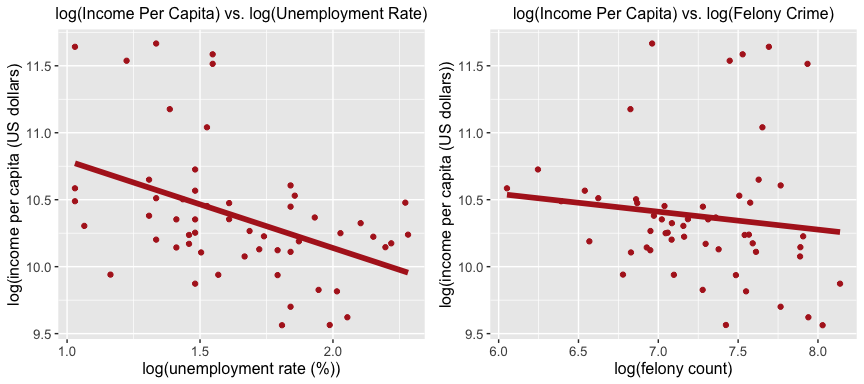
While summary statistics were determined and analyzed for all 6 variables used (Income per Capita, Unemployment Rate, Number of Major Felonies, Population age 3+ in schools, Population with a Bachelor’s degree or higher, and Total Population by PUMA), it appeared most logical to examine the summary statistics for the response variable, Income per capita, to examine its overall trend in the New York City area. For the following tibble, the summary statistics for Income per Capita for all PUMAs were determined as follows:

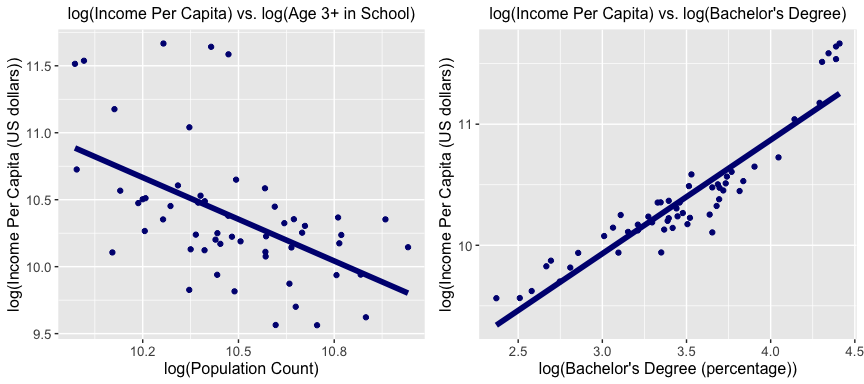
Summary Statistics for Income Per Capita (USD)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean Income Per Capita (USD) | Median Income Per Capita (USD) | Maximum Income Per Capita (USD) | Minimum Income Per Capita (USD) | Standard Deviation of Income Per Capita (USD) | Variance of Income Per Capita (USD) |
| 37,094.67 | 28,747 | 116,577 | 14,226 | 24,894.51 | 619,736,442.48 |

After evaluating the above values, it becomes clear that some degree of economic inequality exists, but as to whether or not there is some particular variable or influence that exists that yields such a disparity remains unclear. With a mean and median income of $37,094.67 and $28,747, respectively, compared alongside a staggering figure of $116,577 as a maximum, it no longer seems an unlikley prospect that some macroeconomic variable could potentially shed light on possible underlying influences on income levels.

## Scatterplots with Regression Line and Correlations





In the graphs above, correlations between pairs of variables were mapped together with a focus on income. When examining each of the charts, there was only one strong positive trend that came about, with an impressive correlation coefficient of -0.429 between percentage of population with bachelors degree and income per capita (USD). As for the graphs depicting the relationships between other variables and income, weaker correlations appeared, especially for crime with income levels, which had a correlation coefficient of 0.003. For unemployment and school enrollment, when compared alongside per capita income, there were noticeable trends that seemed to follow a more exponential decay-like trend. When assessing the underlying reasons for this, one might be able to note that with higher school enrollment, larger households exist, and thus with more dependents like students, total household income would likely become divided more frequently and therefore into smaller sums per capita. This explains the correlation coefficient of -0.429 between population over age three enrolled in school and income per capita (USD). As for unemployment, it would appear most logical for there to have been an exponential decay in income per capita with higher levels of unemployment as every 1% increase in unemployment extracts significant capital from households annually, and part-time workers are not counted. However, there was only a correlation of -0.369 between unemployment rate and income per capita (USD). When articulating the underlying reasons for why correlations may appear weaker, the main underlying sources of error here lie in the fact that many of the data sets used were summary statistics in a PUMA and as a result there was restricted access to data from individuals that would otherwise enable a macro analysis of these variables. Granted the macroeconomic nature of examining unemployment and crime rates, it also became difficult to discern a trend within one city given other variables at play (i.e., social welfare programs/public expenditures, real estate costs, other facets of inequality). For the sake of concisiveness, the correlations among the main variables highlighted at the beginning (Education and Income per Capita) were prioritized.

## Multiple Regression

In this part of the investigation, a multiple linear regression model, -760.324 (unemployment rate) + 0.07 (school enrollment) + 5.372 (felony count) + 1286.346 (bachelors degree attainment) +-1.6450423^{4}, was developed to effectively map the connectivity among the variables. Given the r-squared value, also known as the coefficient of determination, of 0.8895, the model seems to have an 88.95% match rate with the collected data points. The model could thereby be viewed as being a fairly accurate one. In addition, the coefficients for each of the subset linear equations, each depicting the rate of change in income per unit increase, presents helpful insights about the potential relationship between the four explanatory variables and response variable income. The only negative slope indicated by the multiple linear regression model was the unemployment rate, which had a slope of -760.324. The negative slope can be explained by the direct impact of lower employment on per capita income levels.

After having engaged in an analysis of city wide data, it would appear as though some of the basic correlations existed, though not enough to completely justify the more definitive relationships highlighted by preexisting economic principles and understandings. In order to have generated a more accurate visual and graphical depiction of the relationship between income per capita and education, individual data sets would needed to have been used instead of aggregated summary statistics by PUMA. In conducting the investigation as it currently stands, countless outliers could have affected the correlation depiction two-fold: (1) by being factored into PUMA averages and (2) by being folded into yet another set of summary statistics to generate regression lines of best fit. And while it can be said that none of the earlier hypotheses regarding potential relationships were proven in the case of this experiment, many questions can be posed relating to other external or systemic factors (i.e., redlining, socioeconomic profiling, etc.) that would otherwise impact the quality of the data and the subsequent measurement of correlation.