Relations between Car Collisions and Population Variables in NYC

< Peng Jia, jackchina123, pj585 >

Abstract:

After analyzing the relations between total car collisions and different population variables by line regression, I have not found any acceptable relations because of the low R-squared value.

Introduction:

I want to analyze the relations between total car collisions and different population variables in different area of zip code in New York City. It is import to analyze the relations because if I can find the relations, I can do further analysis to find ways to reduce the car accidents. The populations variables are population density, mean household income, median earnings for workers, total units of buildings, average age, zip square Footage, residential Units, non-residential units, population 16 years and over, total population, and total households.

Data:

The data I will use is NYPD Motor Vehicle Collisions from NYC Open Data. The data is suitable and reliable because the data is provided by the New York Police Department (NYPD), which is professional dealing with vehicle collisions.

Methodology:

I use line regression to do data analysis. One important attribution is R-squared, or called coefficient of determination, which is a number that indicates how well data fit a statistical model. In general, the higher the R-squared, the better the model fits the data. I try to decide what R-squared Value is acceptable to interpret that model fits the data. It is hard to decide what the R-squared value is good enough. According to Jim Frost (2013), any field that attempts to predict human behavior, such as psychology, typically has R-squared values lower than 50%. It is because humans are simply harder to predict than physical processes. The car collision is caused by human for different reasons such as being drunk and distraction, so I can consider car collision as one kind of human behavior, and R-squared above 50% is a good prediction.

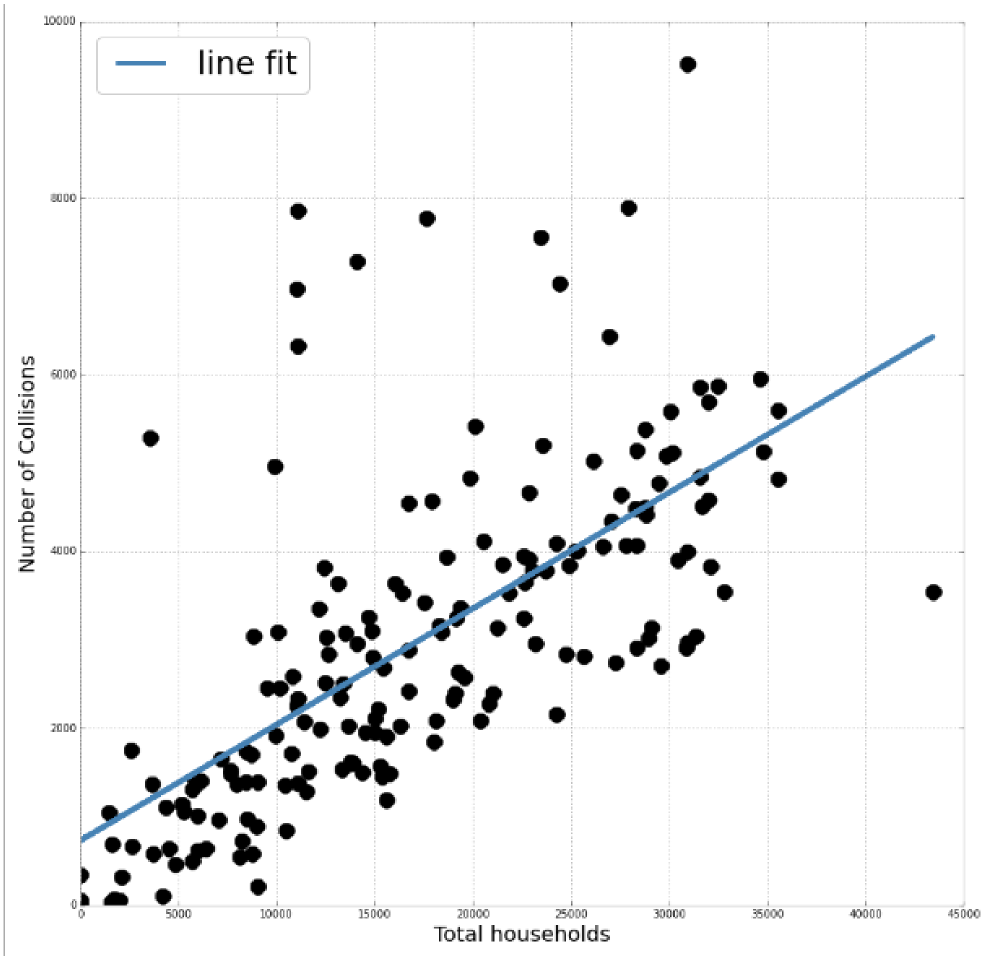
The R-squared value for the relations:

Number of collisions vs. Population Density: 0.007

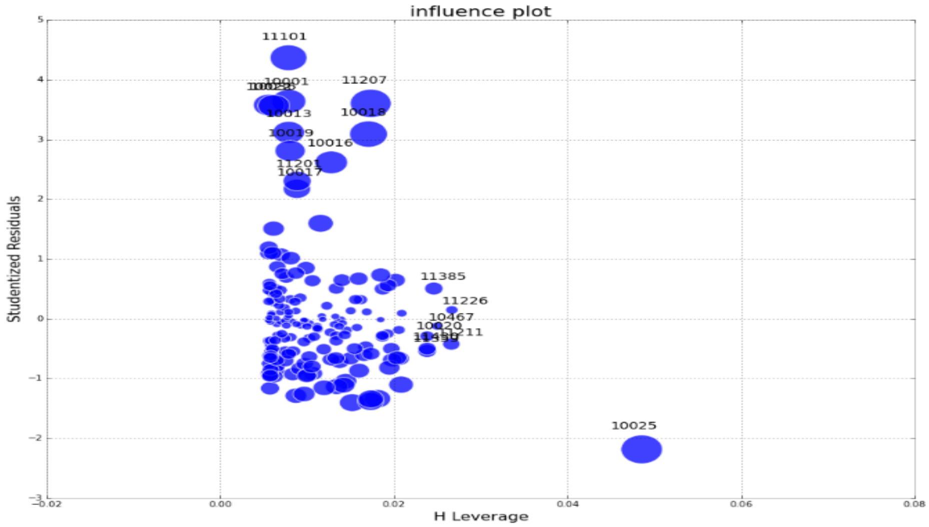
Number of collisions vs. Mean household income: 0.003
Number of collisions vs. Median earnings for workers: 0.000
Number of collisions vs. Average Age: 0.040
Number of collisions vs. Zip Square Footage: 0.004
Number of collisions vs. Residential Units: 0.025
Number of collisions vs. Non-residential units: 0.017
Number of collisions vs. Total Unit of Building: 0.029
Number of collisions vs. Total population: 0.383
Number of collisions vs. Population 16 years and over: 0.407
Number of collisions vs. Total households: 0.456

All of the R-squared values are below 50%, which are not good for analysis. I will try to remove outliers of (Number of collisions vs. Total households) because its R-squared value is the one that nearest 0.5.

The graph of Number of collisions vs. Total households:



Although an upward trend is visible, the data has significant scatter in the upper side. Moreover, the point between 40000 and 45000 has high leverage and may have large influence.



This is an influence plot, and the size of the points represent the influence it has on the fit. The point with high leverage and big influence is at zip code 10025, which is identified as the point on the bottom right of the previous plot graph as expected.

Then I decide the range of Total household is [0, 4000]. However, the R-squared values is 0.47, which is still lower than 50%, and the adjusted R-squared value is lower, only 0.467.

Conclusions:

After analyze line regression of the relations between number of car collision and population, I have not found any good relations between them because of the low value of R-squared. The relations of Number of collisions vs. Total households has the R-squared value near 50% without outliers, but it is still less than 50%. Even though if it would be more than 50%, the conclusion I will get is that the more households the area has, the more car collisions the area has, which is trivial and does not help for further analysis to reduce car collisions.

Future work:

In the future, I will try to find more variables about the population so that I may be able to find a good relations between number of car collisions and population variables.

Links:

Data of NYPD Motor Vehicle Collisions CSV version:

https://nycopendata.socrata.com/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

Data of Population in NYC:

https://github.com/jackchina123/PUI2015_peng/blob/master/pj585_EC/pupulation_nyc.csv

Bibliography

Jim Frost, “Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?”, May 30, 2013, http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Jonathan Welsh, “Which Types of Car Accidents Are Most Common?”, February 13, 2013, http://blogs.wsj.com/drivers-seat/2013/02/13/which-types-of-car-accidents-are-most-common/

Khosro Sadeghniiat-Haghighi, “Traffic crash accidents in Tehran, Iran: Its relation with circadian rhythm of sleepiness”, September 1, 2014, http://www.sciencedirect.com/science/article/pii/S1008127515000097

Michael Pines, “Top 25 Causes of Car Accidents”, https://seriousaccidents.com/legal-advice/top-causes-of-car-accidents/

Seidowsky Régine, “Processing Traffic and Road Accident Data in Two Case Studies of Road Operation Assessment”, March 8, 2015, http://www.sciencedirect.com/science/article/pii/S2352146515000368

Seyed Taghi Heydari, “Time analysis of fatal traffic accidents in Fars Province of Iran”, February 3, 2013, http://www.sciencedirect.com/science/article/pii/S1008127515301590