The Demographics of Taxicab Riders

Thomas C. Proctor

March 3, 2016

Executive Summary

In this work I create an explanatory model of the location of New York City licensed taxicab (yellow cab) drop-offs based off demographic and location data. These findings may help to understand the demographics of yellow cab users. Also, it may be combined with similar analysis of other for-hire vehicle services to compare the demographics of these services. This information will be especially useful to inform the re-evaluation of the regulation of for-hire vehicles which is required due to the introduction of smartphone based hailing. Two data sources are used: taxicab trip data provided by the NYC taxicab and livery commission and demographic data from the US Census bureau. As drop-offs are count data, we expect that the drop-off data is drawing from a Poisson distribution and Poisson regression is used, along with feature selection to sort through the large number of available features. I find that drop-offs can be well predicted by the income of census tracts, and on average, $\textit{dropoffs per-capita} \lesssim (\textit{per-capita income})^2$. However, there is clearly significant influence from other factors, and the Poisson distribution based model for the randomness in taxicab drop-offs is not very accurate.

1 Introduction

In 1937, in response to traffic congestion, New York City began regulating for-hire cars by limiting the total number of vehicles (yellow cabs) that can legally pick up passengers hailed from the street. This has created a system which many say is unfair to those who do not live within the core business district, as accepted wisdom says that unmet demand is high enough within the core business district that there is no incentive for drivers to leave it. Recently, the advent of other for-hire vehicles in NYC has brought further attention to issues with yellow cabs service across geographic, income, and ethnic groups. The city has started a system of so called "boro cabs", which

are allowed to pick up street hails only outside of a defined central business district. Smartphone hailing apps manage to bypass the regulation scheme based upon street hails, providing significant competition to yellow cabs, along with claims that they better serve poorer neighborhoods. Along with the conventional wisdom about the locations where cabs can be hailed, there are also other common tropes about cab coverage, such as a bias against non-white passengers and that they are only used by the rich.

In order to answer questions about the users served by for-hire vehicles and approach questions of unmet demand, I evaluate the demographics of passengers by looking at the demographics of passenger drop-off locations. A map of this drop-off data can be seen in Figure 1.

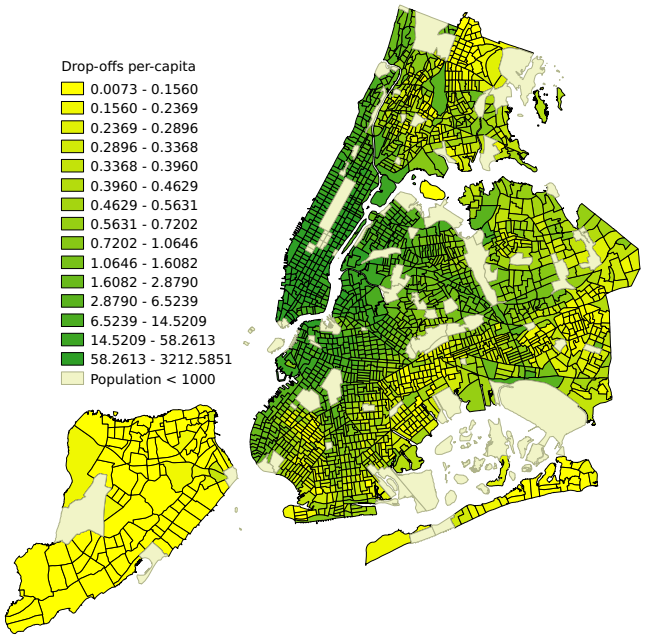

Figure 1: Map of New York City per-capita drop-offs. Staten Island is the island to the south-west with an overall low number of drop-offs, while Manhattan, where the chief central business district of the city is located, is the long, narrow island with a large number of drop-offs.

1.1 Data

I hope to be able to predict the number of drop-offs in an area based on the demographics of that area in order to understand the demographics of passengers. By comparing analysis for different types of for-hire vehicles, we can better understand the differences between the users of each service.

I work with two data sources for this project. The first is yellow cab trip data provided by the New York City Taxicab and Livery Commission[1]. This data includes latitude and longitude coordinates for the start and end (pickup and drop-off) of every single yellow cab trip in NYC among lots of other trip information. The drop-off lat. lon. coordinates are used to generate the response variable, the yearly per-capita yellow cab drop-offs in a given census tract. When I first started exploring taxicab data, only the 2013 data had been released, so I used this data. The second source is demographic data from the US Census Bureau[2]. For 2010, the full US Census is a population count that the best possible attempt to record every single resident of the US, but is mostly just counting population, with only a small amount of other information. The American Community Survey, which runs constantly, includes many more demographic factors, including racial, income, and commute data. I used the 2009-2014 set of estimates, as the five-year estimates have reasonably low margins of error and are suggested for uses that aren't tracking demographic changes over time. It uses sampling and gets diverse demographic information, though some of this data can have very high errors.

The data from the Census Bureau is broken up by census tract, which divides all of the United States into relatively small geographic areas which ideally are largely demographically homogeneous. Thus, there can be a large range in total population and population density. For example, most parks and airports in NYC have their own census tract with little to no residents.

[1] Data is available from `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`

[2] Census data is published on many websites. I've used `https://nhgis.org/`, as it allows you to download only the columns you want and not waste hard drive space or time processing lots of data. The census supplies "shapefiles" of the different geographical areas they use at `https://www.census.gov/geo/maps-data/data/tiger-line.html`, including tracts. The city of New York supplies shapefiles that clip to the shoreline at `http://www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page`, as the census tracts are defined so that they include all river areas and much of the ocean. I expect that the census needs to be prepared for houseboats and changing shorelines.

2 Methods

I assume that the demographics of drop-off and pick up locations are representative of the demographics of users. As only a fraction of trips begin/end at the user's residence, this is a fairly significant assumption. It may be possible to better control for this by finding other data sources, such as measures of the number of people working in a given area or the commercial vs. residential zoning of an area.

2.1 Data Processing

There are over 150 million trips within NYC recorded for 2013. With the resources I have, it is not reasonable to hold all this data in RAM at once, so there is some wrangling required in order to study it. Furthermore, finding which census tract contains the lat. lon. point of a drop-off is a resource intensive process. Instead of assigning census tracts point-by-point, I separate the NYC area into a fairly grid, with one one-thousandth of a degree separation between each grid line, and find the census tracts of the points in that grid[3]. This reduces the points needed to match to tracts to only about 24 million. The actual work of matching the grid lat. lon. points to tracts is done using the PostGIS add-on to PostgreSQL. Then, lat. lon. points can be matched to tracts by simply rounding to the nearest one-thousandth of a degree and joining the table of drop-offs with the table of grid points[4]. Memory usage can be traded for time by only using a subset of the tracts in the table of grid points at once.

2.2 Data Analysis

New York City contains 2168 census tracts. Of these, I remove 83 that have populations under 1000 persons. These tracts are mostly parks, but also include the two airports, the Brooklyn Navy Yard, and an area of Midtown Manhattan that is mainly offices and transportation hubs. As the population of these areas is so low, we cannot expect drop-offs within them to be representative of their demographics.

[3] Python scripts for creating the grid to tract table and making it memory efficient can be found at my github page under `Generate tract lookup.py`, `Create Fipscodes dictionary.py`, and `Convert tract lookup low memory.py`.

[4] A Python script for computing statistics about the drop-offs in each NYC census tract can be found at `https://github.com/ThomasProctor/Slide-Rule-Data-Intensive/blob/master/DataStory/FinalFiles/ComputeTractStatsFromCSV.py`.

Also, my analysis does not cover Staten Island, which is geographically and demographically very distinct from the rest of New York City[5]. It is the only borough with no subway connection to Manhattan, where the New York City central business district is located, so public transit connections can be much more time consuming and expensive. Access to Manhattan by car requires crossing two bridges and driving through much of Brooklyn. It is the only borough where non-Hispanic whites make up a majority of the population[6], and its car ownership rate by household of 84% dwarfs that of the next closest, Queens, with 64%[7]. Unlike most of New York City, it is almost entirely suburban, and the separation between Staten Island is so pronounced that a referendum passed in 1993 to secede from the city, though the state prevented its implementation.

Feature selection of a large body of demographic data, including commute patters, racial make-up, and economic data of residents indicates that per-capita income is the most promising predictor for drop-offs in a given census tract. This is not surprising. Taxicabs are a luxury service. In NYC, taxi cabs compete with public transit that costs about as much per trip as the minimum charge for a taxi, and just a little bit over 1% of New York commuters commute mainly via taxicab. Thus, I would expect that those with more income are more likely to take cabs, and this is the case.

Clean up feature selection notebook, add github link to endnotes

2.2.1 Poisson Regression

The log-log plot of per-capita drop-offs vs per-capita income shown in Figure 2 indicates a roughly linear relationship of the logarithms of the two variables. The deviance from this linear behavior in the lower left of this plot can be explained by the fact that the data is drawn from a Poisson distribution, and data will not be symmetrically distributed around the mean. Using

$$\log d - \log p = \theta_c + \theta_j \log j \tag{1}$$

as my model, where d is the drop-offs in a given census tract, p is the population (included in the model as an offset), and θ_c and θ_j are unknown parameters of the model, I do a Poisson regression on this data and find that

$$y \approx \frac{j^2}{4.7 \times 10^8}, \tag{2}$$

[5] `https://en.wikipedia.org/wiki/Staten_Island`. Yes, I cited Wikipedia. The geographic position of Staten Island can be seen in Figure 1.

[6] `https://en.wikipedia.org/wiki/Demographics_of_Manhattan`

[7] `http://www.nycedc.com/blog-entry/new-yorkers-and-cars`

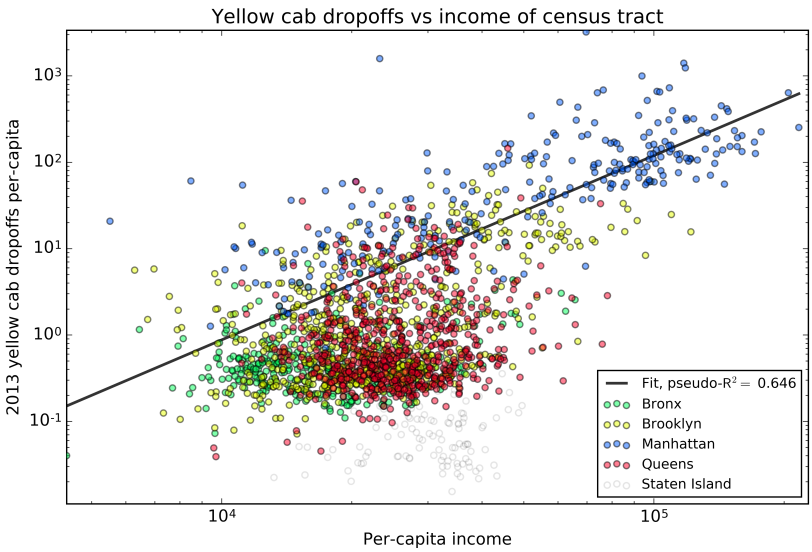

Figure 2: Per-capita drop-offs vs per-capita income with a Poisson model. The fit is given in equation 2.

where y is the drop-offs per-capita in a given census tract, and j is the per-capita income of that tract. This fit is illustrated in Figure 2.

Clean up Poisson analysis notebook and add footnote

One of the more useful methods for judging the quality of Poisson regression is by looking at the deviance, which is the factor that is minimized in the regression algorithm. I find a very high deviance of 2×10^8, which gives a p-value of 1 when compared as a hypothesis with the saturated model as the null hypothesis. This isn't all that surprising. A Poisson distribution has a variance equal to it's mean. Because there are obviously more factors at play than the predictors that we have used, we would expect that there will be a lot of extra variance than that expected by the pure Poisson distribution. This *overdispersion* is generally the rule rather than the exception with real world data.

The normal way to account for this overdispersion is by relaxing the requirement that the variance is equal to the mean, and instead have the variance equal to $\tau\lambda$, where λ is the mean, a function of the predictors, and τ is a fitted constant called the *dispersion parameter*. The resulting model, called a *quasi-Poisson* model, will always have the exact same prediction for the mean as the corresponding Poisson model, as the only thing that has changed is the random part of the model.

Using a quasi-Poisson model, the deviance is lowered significantly to 386, and the p-value is reduced to zero. However, this should be taken with a grain of salt, as the dispersion parameter is an incredibly high 5×10^5, indicating that there is a huge amount of extra variance above and beyond the Poisson model expectation.

The quality of the random component of the model can also be judged by looking at the *Anscombe residuals*, which transform the residuals so that they should have an approximately normal distribution if the model is correct. However, the quantile-quantile plot shown in Figure 3 shows that distribution of the Anscombe residuals is far from normal. The distribution is narrow around the mean, with relatively longer tails than that of a normal distribution.

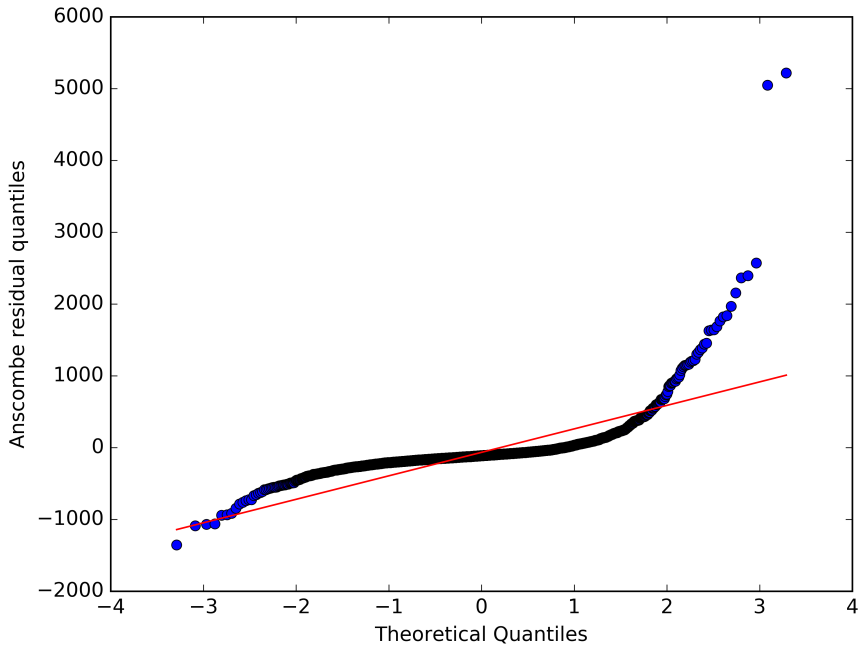

Figure 3: Q-Q plot for a quasi-Poisson fit. If the residuals followed a normal distribution, they would lie on the red line.

Using the dispersion, we can find a parameter analogous to the R^2 used in ordinary least squares which measures the improvement over the null model, called *Mcfadden's pseudo-*R^2. This pseudo-R^2 is 0.65 for our quasi-

Poisson model, indicating that a dramatic improvement over the null model, although the behavior of the residuals runs counter to the expectations of the model, as indicated by both the very high value for τ and the poor fit to a normal distribution.

3 Conclusions

Based on this analysis, we can see that per-capita income is an impressive predictor for taxicab drop-offs, with a pseudo-$R^2$of 0.65. However, the high dispersion parameter, along with the poor fit of the deviance residuals to a normal distribution indicates that the model may still have much room for improvement. A possible avenue may be adding more demographic predictors, such as the average commute time, car usage, spatial position, or the racial make-up of census tracts. Borough, which may be a rough proxy for any of these possible predictors can be seen to correlate will with drop-offs in Figure 2. However, while there may be interesting information to be gleaned from exploring which predictors are the most successful, the room for improvement in this model largely lies in the distribution of the predicted error, and thus would probably be relatively marginal.

It is also possible that no improvement can be achieved from a demographic data based approach. It may be that the effect from factors like transit hubs and cultural institutions that are particular to each individual census tract cannot be fully expressed through demographic data, and may not be able to be reasonably tracked by any data.

I, for one, find the simplicity of this model appealing: per-capita dropoffs are proportional to the square of per-capita income. While there are clearly more details, the simple take away here is very satisfying.

Write a lay-person readable blog post.

Todo list