The Demographics of Taxicab Riders

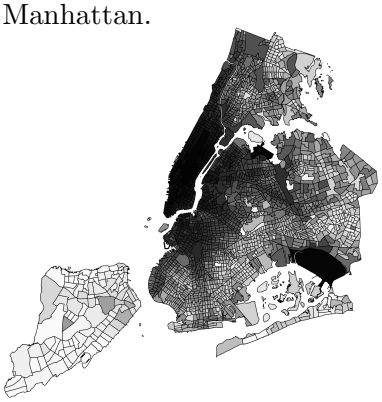
Thomas C. Proctor

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 any of these questions, we first have to understand the basic demographics of taxicab riders.

According to my analysis, the most important factor governing taxi-cab usage is income. More specifically, I find that the number of taxicab drop-offs in a given area per-capita is proportional to the square of the per-capita income in that area.

To understand taxicab usage in NYC, I took a look at the data released by the NYC Taxi and Livery Cab Commission detailing all the roughly 150 million metered taxi trip taken by licensed NYC cabs, commonly known as yellow cabs. This data includes longitude and latitude points for the drop-off and pickup locations of each trip. I compare the drop-offs to demographic data on residents from the U.S Census Bureau that is split up by census tract, which in NYC usually encompasses just a few blocks. I counted the number of drop-offs in each census tract so that I could compare the number of drop-offs to the demographic data from the census.

The map below shows New York City, with darker areas indicating a higher number of total drop-offs during 2013 and lighter lower. The darkest area of the map is the central business district of NYC, midtown Manhattan, with the two airports showing up as the black spots to the right of Manhattan.



Note that I removed Staten Island, the island on the lower left of the map, from my analysis. Staten Island is extremely demographically and economically distinct from the rest of New York City, so much so that Staten Island actually voted in a referendum to secede from the city in 1993.

By looking at the demographics of the locations where drop-offs occur, we can try to gain a bit of understanding of the demographic factors that effect ridership. It's important to note though that not every drop-off represents someone getting dropped off at their home, so the location of drop-offs will not directly tell us who is taking taxis - far from it. This is clear if you look the few areas, such as central park and major transit hubs such as Penn Station and Grand Central which have drop-off rates orders of magnitude above the rest of the pack. This large number of drop-offs does not come from residents, but from people from all over the city traveling to these popular destinations. The mantra that correlation does not imply causation is probably worth repeating here, and it may be that a model based only on demographic factors is only useful as a first step to a more complete model rather that one where we should be making decisions right off the bat.

To do this analysis, I first studied correlations between the number of drop-offs and the various demographic features of each census tract in order to select out the demographic features which did the best job at predicting the number of drop-offs. This indicated that the income of each census tract is highly correlated with the number of drop-offs, so I performed a regression to find how exactly income is related to drop-offs. The regression indicated

that

$$\textit{per-capita drop-offs} \approx \frac{(\textit{per-capita income})^2}{4.7 \times 10^8}, \tag{1}$$

When I say that the number of drop-offs per-capita is proportional to the square of the per-capita income, what this is really saying is something about the *average* number of drop-offs. If we have a bunch of census tracts with the exact same per-capita income, we would not expect them all to have the exact same number of drop-offs per-capita. Instead, the model says that they will all have different drop-offs per-capita, but their average will follow equation equation1. The model also gives a fairly specific description of how many tracts we'd expect to deviate from the average by how much, and this is where we start running into problems.

For the sake of comparison, lets imagine a New York where the probability of a taxicab dropping somebody off in a given census tract at any time depended on one thing, and one thing only - the per-capita income in that census tract. In this imaginary world, if we looked at two census tracts with the exact same income, we wouldn't expect them to have the exact same number of drop-offs within a year, but we would expect them to be pretty close. The thing is, if we look at two census tracts with similar per-capita incomes in the real data, the number of drop-offs is a whole lot less close than we'd expect if per-capita income was the only thing that mattered.

This means that there must be something other than per-capita income that matters here. Frankly, that's not surprising - if I told you that per-capita income really was the *only* thing that mattered, it would be much more shocking. The thing is that our data is *so* much farther apart than we'd expect if per-capita income was the only thing that mattered - about half a million times farther apart. This means that there must be a whole lot more important factors that come in to play.

However, using just income does do a pretty good job of explaining things. On a scale that goes from not explaining anything at all to explaining everything perfectly, using just income gets us about 65% of the way to perfect - or for those who are familiar with regression, $R^2 = 0.65$[1]. That's a whole lot of the way to perfect using just a single variable.

Which other factors are important is still an open question. A bunch of the possibilities are things that should be able to get data for such as the number of people who's place of work is in a given area, whether or not the

[1] Technically this isn't an R^2 which doesn't really exist for Poisson regression. Instead, it is a "pseudo-R^2", which behaves a bit like an R^2 from ordinary least squares in that it represents the improvement over a null model.

residents commute in cars, or how far the area is from the city center. But there also are plenty of factors that we can't hope to measure using a data driven approach, such as the presence of cultural institutions or a school dorm full of students who may be a whole lot wealthier than their income suggests.

Some of the factors that might be involved can't be seen by looking at the demographics of residents, like how many people's jobs are in a census tract. And some of the factors are things that we can't expect to see directly, like the presence of cultural institutions. However, many of the things that we may not be able to measure directly may have indirect effects on data that we can measure. It could be that this is the reason that income is particularly useful, as factors like race, local employment, and cultural institutions are all correlated with income.

It is also important to not take . I am looking at the demographics of the locations where drop-offs occur, not directly at the demographics of passengers, which could be dramatically different from data I'm measuring, especially if income is

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 **??**. 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.