MS&E 231 - Assignment 2

Samuel Hansen
sfhansen

Camelia Simoiu
csimoiu

Julie Zhu
lijingz

October 27, 2015

Background

Two competing labour supply theories may explain why it is more difficult to catch a cab in the rain. The neoclassical law of supply predicts that drivers will work longer hours on days which they can earn higher wages. The competing theory from behavioural economics predicts that cab drivers are "target earners" and will stop working when they have reached their personal income target. Assuming demand for taxis increases when it rains, the behavioural theory predicts that cab drivers meet their targets sooner and stop earlier in the day, which decreases the supply of taxis. Literature from the behavioural (Camerer et al., 1997) and neoclassical (Farber, 2014) schools of thought offer competing evidence concerning the estimated slope of the wage elasticity curve for New York City cab drivers. Camerer et al. argue the negative slopes of the wage elasticities imply drivers work fewer hours when wages are high, which indicates drivers are target earners. In contrast, Farber claims "drivers tend to respond positively to ... increases in earnings opportunities", meaning taxi wage elasticities for have positive slopes.

To test which model of labour supply is supported by the data, we replicated portions of the analyses conducted by Camerer et al. and Farber. In doing so, we used Amazon Web Services (AWS) to perform three MapReduce jobs that reduced a raw dataset of 700 million NYC taxi rides to a smaller set of aggregated driver statistics, including hourly earnings, number of drivers on duty, and miles driven. Overall, our results support the neoclassical model of labour supply that predicts "work hours should respond positively to transitory positive wage changes" (Camerer et al., 2014). Before interpreting these results, it is important to understand the methodological assumptions from which they were derived.

Data Processing

During the data cleaning phase, we only included observations that satisfied the following criteria:

Distance

We checked whether the distance recorded by the meter was greater than or equal to the haversine distance between the pickup and drop-off GPS coordinates. We used the haversine distance in order to determine the great-circle distance in miles between two points on a sphere (i.e. the Earth).

Because the spherical earth is a non-Euclidean surface, we could not use Euclidean formula to calculate the distance between two points.

GPS Coordinates

We checked that the recorded longitudes fell between Washington D.C. and the northernmost tip of Maine, and latitudes fell between the westernmost border of Pennsylvania and the easternmost seaboard of Maine. Although this range may seem overly generous, we sought to exclude clearly erroneous GPS coordinates (i.e. 0) while still including plausible long trips taken by taxis.

Passenger Number

We validated that the number of passengers was greater than 0 but less than 100. We set this high upper bound to account for possible cases in which a larger vehicle (e.g. charter bus) was commissioned to transport customers.

Dropoff and Pickup Times

We examined whether the drop-off time was after the pickup time and whether the total trip time was greater than 0 seconds. We did not assume a lower bound above zero to account for trips that were immensely short (i.e. a trip from one block to the end of the same block). To account for erroneous recordings of the trips time in seconds, we required that the absolute value of the difference between (dropoff - pickup) and (trip time in seconds) to be less than 10.

Speed

The average speed of the trip was less than 100 MPH. Although some NYC freeways have 70 MPH speed limits, it seemed unlikely for average trip speeds to exceed 100 MPH. We set this slightly higher bound to account for speeding drivers.

Rate Code

The rate code was in the set of valid options, strictly including 1,2,3,4,5, and 6. We obtained this list from the NYC Taxi and Limousine Commission description.

Aggregate Statistics

The goal of the data aggregation phase was to compute driver statistics grouped by date and hour. Most trips occurred within the same hour; however, to account for trips that crossed an hour mark, we introduced the following variables:

- **Rollover** is a boolean variable set to true if a trip crossed the hour mark and set to false otherwise. For instance, if a trip lasted from 9:50am-10:15am, we partitioned the trip into two periods: 9:50am-10:00am, for which rollover was false, and 10:00am-10:15am, for which rollover was true. To avoid double counting, we did not increment the number of trips and passengers when rollover was true.

- **Proportion Time** is a float representing the proportion of the total trip time that was spent in each hour. In the 9:50am-10:15am example, "proportion time" would equal (10 minutes)/(25 minutes) = 0.4 in the 9:50am-10:00am period, and (15 minutes)/(25 minutes) = 0.6 in the 10:00am-10:15am period. We then used this proportion to allocate the appropriate amount of earnings, miles, and other quantitative variables to their respective hourly periods.

Results

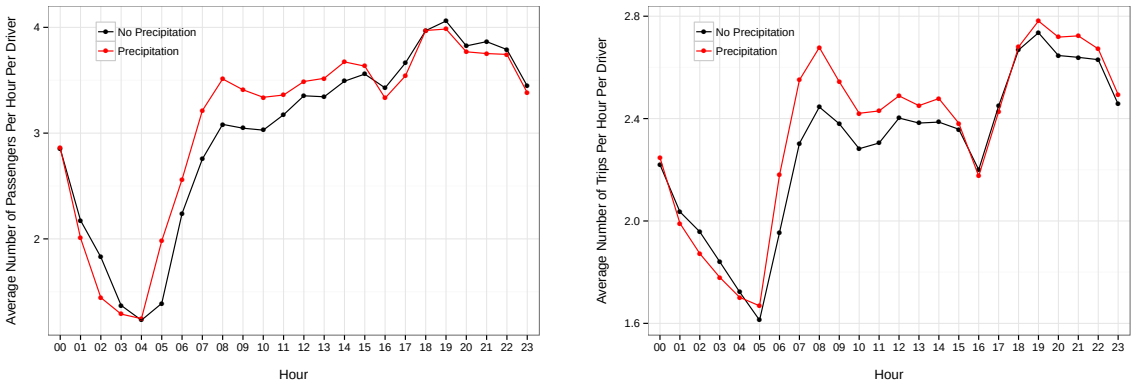
The resulting dataset after preprocessing has over 31,000 observations on approximately 1,300 days. Just under 8% of the original data were excluded due to failing validation checks. Computing some basic statistics on the dataset, we observe that the average trip distance when it rains is approximately equal to average trip distances when it does not rain (2.6 miles). The average number of drivers on duty is lower when it rains (8,395 drivers versus 9,891 drivers, or 15% less drivers).

Demand

After successfully obtaining aggregated driver statistics via AWS MapReduce, we first investigated whether demand for cabs actually increases when it rains. We examined two variables that can represent taxi demand: 1) hourly average number of trips per driver, and 2) hourly average number of passengers per driver. We argue that when the number of trips or passengers increases, demand for taxis is higher. On the other hand, when the number of trips or passengers decreases, demand for taxis is lower.

As depicted in Figures 1 (a) and (b), both measures of demand reveal a similar pattern: when precipitation is present, taxi demand is higher during the hours of 5:00am - 3:00pm. According to Figure 1(b), which depicts the number of trips, taxi demand is also slightly higher when it rains between 7:00pm-10:00pm; however, this pattern is not reflected in Figure 1(a), which quantifies demand by the number of passengers. Hours outside of these intervals have nearly overlapping demand curves in both the precipitation and no precipitation cases.

Based on these results, we can suggest that precipitation exhibits a non-trivial effect on taxi demand during morning-commute and mid-afternoon hours, which are traditionally peak times of intercity tourist traffic. Lastly, both demand curves are consistent with intuition because peak demand occurs during commuting hours.



(a) (b)

Figure 1: Average number of (a) passengers, and (b) trips per hour.

Supply

To examine the response of taxi supply to transitory changes in demand, we first quantified the supply of taxis as the number of drivers on duty in a given hour. Using this notion of supply, we inspected the average number of drivers on duty per hour (see Figure 2). The pattern in Figure 2 tells a coherent story with the demand metrics, albeit with some caveats. Roughly speaking, the supply curve follows a similar contour as both demand curves: supply is low when demand is low, and supply is high when demand is high. This phenomenon is consistent with neoclassical theory that posits supply and demand will adjust until market equilibrium is reached.

However, there is a notable difference: after 4:00pm on Figure 2, taxi supply is strictly higher when it does not rain compared to when it does. Figures 1(a) and (b) tell conflicting stories as to whether taxi demand is higher during this period when it rains; however, the demand curve quantified by average hourly passengers (Figure 1(a)) more closely resembles the supply curve than does the demand curve quantified by average hourly trips (Figure 1(b)), suggesting the former demand metric may be a better proxy than the latter.

Lastly, the main take-away from Figure 2 is that neither the red supply curve (precipitation) nor the black supply curve (no precipitation) consistently dominate each other. Taxi supply is higher when it rains between approximately 8:00am-11:00am, whereas taxi supply is higher when it does not rain between 6:00pm-11:00pm and 1:00am-3:00am. This indicates an interaction effect between time of day (i.e. morning, afternoon, or night) and precipitation level, which in turn, implies taxi supply is not solely affected by wages, but rather a multitude of interacting variables.

Average Hourly Wages

Given the interdependent nature of demand, supply, and wages, we then investigated average hourly wages between precipitation periods and non-precipitation periods. As depicted in Figure 3, average hourly wages are consistently higher when it does not rain compared to when it does.

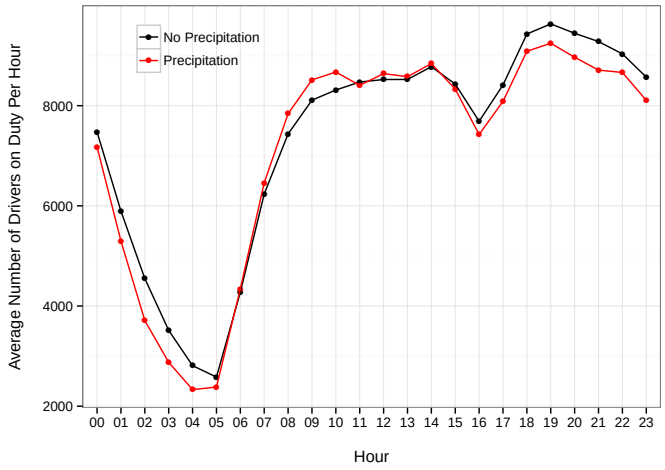


Figure 2: Average hourly supply of drivers on duty.

When interpreting this plot, there are two main caveats to bear in mind: first, average hourly wages are markedly higher than those reported by Camerer et al. (1997) and Farber (2014). Secondly, there is a large peak in hourly wages around 5:00am, which is either due to an anomaly in the data or a real increase in taxi demand in early morning hours. Indeed, it is possible that early morning trips are often those that span large distances, such as intracity commutes or airport rides. Overall, this plot confirms the intuition that wages are higher during commuting hours (i.e. early morning and later evening). However, the results may be affected by various sources of error, which are explained in the discussion section.

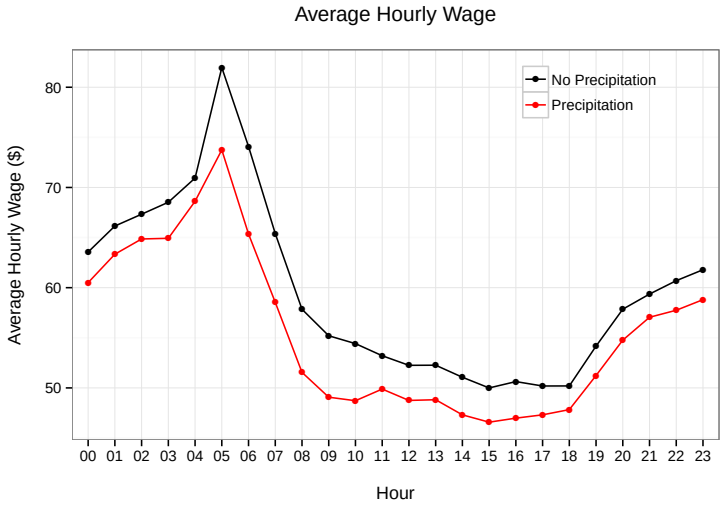


Figure 3: Average hourly wage.

Wage Elasticities

Building on our previous findings concerning demand, supply, and average hourly earnings, we extended our analysis to examine aggregated wage elasticities, which measure how long drivers work at different wage levels. A wage elasticity curve with a positive slope implies drivers work more when higher wages are available, whereas a negative slope indicates drivers work less. In accordance with the methods proposed by Camerer et al., we plotted the log of average daily hours worked versus the log of average daily earnings (see Figure 4) (Camerer et al., p. 417). When interpreting this plot, bear in mind that we normalized these averages by the number of drivers on a given day, and each point represents one day in sampled period between 2010-2013.

There are two main take-aways from Figure 4. First, the wage elasticity curve has a clearly positive slope, implying higher wages are associated with more hours worked. This trend holds true for days with precipitation and days without precipitation. Second, the red points (days with precipitation) are clustered in the bottom left quadrant, whereas the black points (days without precipitation) are clustered in the top right quadrant. This clustering is merely a reiteration of the finding from Figure 4, which showed wages for days without precipitation are consistently higher than those for days with precipitation.

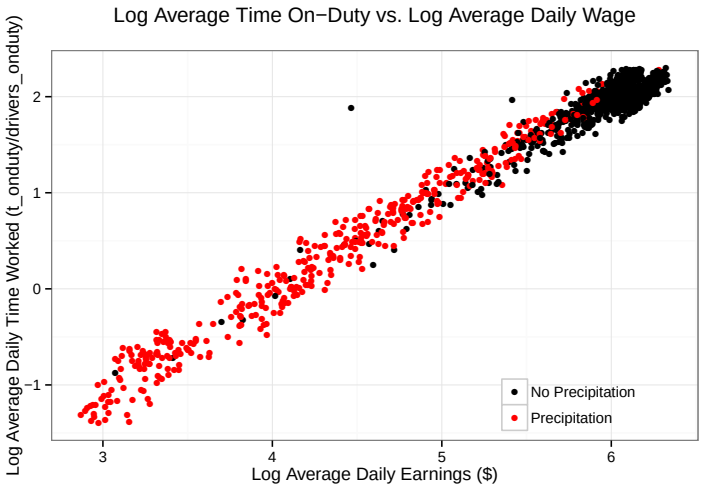


Figure 4: Wages of drivers versus hours worked.

Discussion

Given the positive slope of the wage elasticity curve in Figure 4, our results corroborate Farber's analysis, but challenge Camerer et al.'s claim that the slopes of wage elasticities for NYC taxi drivers are negative. In turn, our results support the neoclassical law of supply, which predicts a positive relationship between wages and hours worked. To understand why our findings differed from Camerer et al.'s, we investigated their assumption that "wages tend to be correlated within days and uncorrelated across days" (p. 408), which, they claim, "is necessary to explain strongly

negative wage elasticities" (Ibid. 409).

Upon further investigation, we believe this theory may not be supported by the data. As an example, consider Figure 5, which plots hourly average wages during the randomly-chosen period of March 2012. The cyclical pattern of peaks and valleys corresponds to weekends and weekdays, suggesting wages are correlated across days. To corroborate this finding, we examined random months throughout all years and found similar cyclical patterns during weekend periods. We argue that the cyclical nature of wages across days challenges Camerer et al.s assumption of the "one-day time horizon for labour supply decisions." (Ibid.). Indeed, cab drivers may work less when wages are lower during weekdays, then compensate by working more when wages are high on weekends.

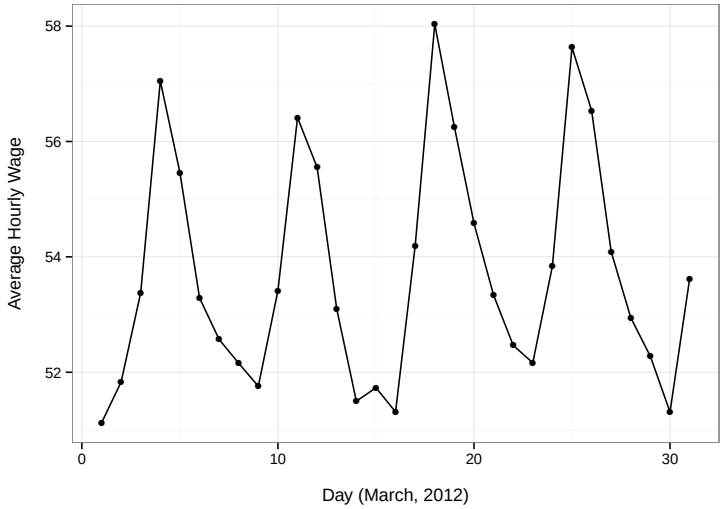


Figure 5: Weekly trend in earnings.

Furthermore, the three data sets used by Camerer et al. had small sample sizes and spanned small intervals of time (e.g. November 1 - 3, 1988, October 29 - November 5, 1990). By contrast, our data set included 700 million rides and spanned 5 years, which is similar to the data used by Farber. The small sample size of Camerer et al.'s data likely failed to account for confounding factors such as surge demand during holidays, summer tourist season, and changing wage laws over the 8 year time horizon from which their data came. Given that our data more closely resemble those used by Farber, it is unsurprising that both of our conclusions support the neoclassical theory of wage elasticity. However, we offer this conclusion bearing in mind that several sources of error are yet to be resolved.

Sources of Error

First, note that the classification of precipitation by the Central Park weather station is not limited to just rainfall. Precipitation includes both snowfall and rainfall; however, each of these weather conditions may lead to different patterns in taxi supply and demand. One plausible hypothesis may be that driving in the snow is typically harder than driving in the rain, causing taxi drivers to

supply fewer working hours when it snows compared to when it rains. In a similar vein, customers may demand fewer taxi rides when it snows due to concerns of heavy traffic and opt to take subways instead. A source of error was that our analysis treated precipitation as a binary variable rather than a factor with multiple levels (i.e. fog, rain, snow, sleet). A deeper investigation concerning wage elasticities and precipitation types would provide a more nuanced answer to the question of why it is hard to catch a cab in the rain.

Second, the "missingness" of a variable can have predictive value. If the data are missing not at random (MNAR), the missingness of one variable can be correlated with another variable. We did not account for patterns in missing data because we excluded observations that failed our data validation checks. Because we assumed data are missing completely at random (MCAR), it is possible we are not capturing the effect of a certain subset of our variables on wages. This set of variables may influence our results by highlighting an underlying set of confounding predictors that will lead our results to possess more bias.

Lastly, we again note the limitations of the raw data. Wage data is specified only as fares with tips paid by credit card. We are unable to record any tips paid with cash. This would influence the daily hourly wage of a driver by a non-trivial amount. Another observation is that fleet and independent medallions cannot be discerned. This may be because owner-drivers "face different incentives regarding labour supply than do drivers who lease their cabs, whether from a fleet owner or an individual" (Farber, p.9). We cannot see the effects of medallion type on supply and demand, which may influence our results as well.

References

[1] Camerer, Colin et al. "Labor Supply of New York City Cabdrivers: One Day at a Time." The Quarterly Journal of Economics 112.2 (1997): 407441.

[2] Farber, Henry S. "Why You Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers." Cambridge: N. p., 2014. Print.