Analysis of Taxi Data: Significant Predictors for Drivers Receiving a Tip

Members: Joseph Blubaugh, Anne Geraci, Rachael Glazner, Shannon Nitroy
Class: STAT 653 (Statistics in Research III) - Second Project Report

Date: 27 April (SP17 semester)

Background

Many workers in the service industry rely on tips as a necessary part of their income. One service industry that remains prevalent in heavily populated areas, particularly New York City, is the taxi. The Taxicab and Livery Passenger Enhancement Programs (TPEP/LPEP) approved the NYC Taxi and Limousine Commission (TLC) to collect data regarding taxi records in New York City starting in 2009. These records for yellow and green taxi trips include the date and times of pick-up and drop-off, trip distances, pick-up and drop-off locations, driver-reported passenger counts, payment types, rate types, as well as itemized fares. Additionally, For-Hire Vehicle (FHV) records were recorded. These records include taxi pick-up location, time, and date.

Project Objectives

Our objective is to determine which factors best predict the amount a passenger tips his or her driver. By better understanding these factors, we can determine the more profitable types of trips for drivers, which could potentially improve their salaries, as well as gain insight into what factors may affect the size of tip a passenger is willing to leave.

Data – Source and Variables

The full data set contains millions of observations for each month of the year 2014. Because of this large size, we selected a stratified random sample to represent our population. The month in which the trip was taken was used as the stratifying variable, in order to allow us to examine time of the year effects. 10,000 observations were randomly selected from each month for a total of 120,000 observations in our sample. The following table details each variable in the dataset:

Variable Name	**Data Type:**	**Description of Variable**
Month	Categorical	1 for January, 2 for February, etc.
Pickup_time	Discrete Numerical	Hour in which passenger was picked up (0 for midnight, 1 for 1am, etc.)
Dropoff_time	Discrete Numerical	Hour in which passenger was dropped off.
Passenger_count	Discrete Numerical	Number of passengers in the vehicle
Trip_distance	Continuous Numerical	Distance of trip (in miles)
Fare_amount	Continuous Numerical	Amount of fare (in US Dollars)
Tip_amount	Continuous Numerical	Amount of tip (in US Dollars)
Tip_pct	Computed Numerical	Percent of fare that was paid in tip (tip_amount/fare_amount)
toll_ind	Binary	Was there a toll charge (0 for no, 1 for yes)
pickup_location_id	Categorical	GPS for Pickup location was used to determine in which of 50 “sectors” the passenger was picked up. Prefixed with “P” - e.g P2. (See below)
dropoff_location_id	Categorical	GPS for Pickup location was used to determine in which of 50 “sectors” the passenger was dropped off. Prefixed with “D” - e.g. D13
Rate Code	Categorical	1 = Standard Rate; 2 = JFK; 5 = Negotiated Fare

Early in the analysis we discovered that cab rides paid with cash rarely reported tip data (10 of 49,695 observations). Cab rides paid by credit card reported a tip 95% of the time. The data dictionary accompanying this data set (see citation) confirmed this. As a result we felt we could not use observations for fares paid by cash because the tip data was not accurately reflected. Removing these observations effectively reduced our data set by 41%. We removed additional incomplete records such as observations with zero passengers, zero fares, and zero distance traveled. We also restricted latitudes and longitudes to the areas between Manhattan and JFK International Airport which removed a few trips that occurred outside of the New York area. Since we had an abundance of data we took a few additional data reductions to limit the effect of outliers in our analysis. We limited trip distance to 30 miles (36 observations removed) and fares to $100 (55 observations removed). We also restricted tip percent to 100% which removed 90 observations (.1% of remaining data). This was done primarily because we felt that these observations were very rare and not representative. The largest tip percent was 32 times the fare and would likely show up as an extreme outlier in our model. Finally, we decided to retain credit card records with no tip because, while infrequent, it does make up 5% of the observations. As a result our reduced data set was 67,193 observations (96% of the removed records were due to the incomplete tip data for cash rides).

Pickup and Drop off Locations

Pick up and Drop off location data were generated automatically when a cab driver started the fare timer. As a result, locations recorded by gps devices using latitude and longitude were unique. It was necessary to group the coordinates in order to use the location data in modeling. We chose KMeans clustering as a method for grouping the data into relevant locations. KMeans clustering is an iterative technique that creates "centers" that minimizes the distance between each observation and a particular cluster center. We arbitrarily chose 50 clusters to be generated for both pickup and drop off locations.

KMeans Clusters for Pickup Locations

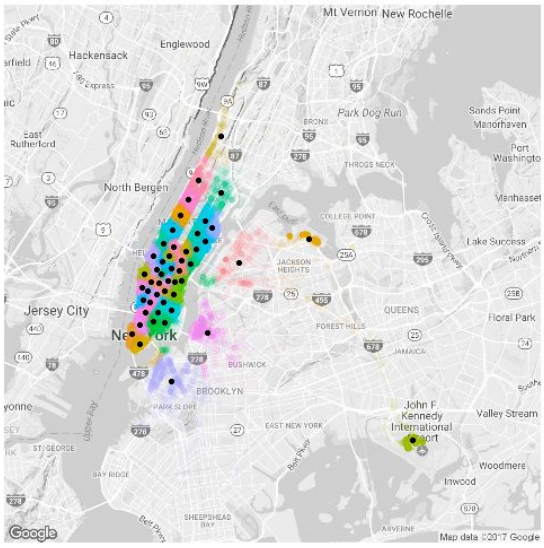

Figure 1: Pick-up location clusters for each drive.
Each color represents location ID.

KMeans Clusters for Dropoff Locations

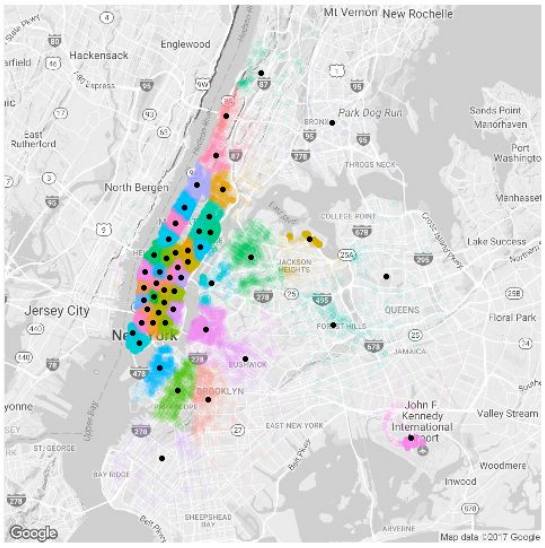

Figure 2: Drop-off location clusters for each drive.
Each color represents location ID.

Statistical Method

SAS was used to fit a generalized linear mixed model using the GLIMMIX procedure. Our response variable, tip amount, has a heavily right skewed distribution and we used the log transformation to improve our overall model fit. We also used the log transformation on our only covariate, distance. The size of the data set allowed us to test all fixed interactions in the full model.

Since all cab rides have a pickup and drop off location, it would have been interesting to test and analyze the interaction between locations. The number of KMeans clusters (50 x 50 = 2500) exceeded our computing power availability.

We used backward selection starting with the highest order interactions first as well as studentized residual plots for model selection.

Proposed Full Model

$$\begin{aligned} y_{ijklmnopq} =& \mu + \beta_1 \log(\text{distance}) + \alpha_i + \delta_j + \gamma_k + \theta_l + (\alpha\delta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\theta)_{il} + \\ & (\delta\gamma)_{jk} + (\delta\theta)_{jl} + (\gamma\theta)_{kl} + (\alpha\delta\gamma)_{ijk} + (\alpha\delta\theta)_{ijl} + (\alpha\gamma\theta)_{ikl} + (\delta\gamma\theta)_{jkl} + \\ & (\alpha\delta\gamma\theta)_{ijkl} + c_m + d_n + f_o + g_p + e_{ijklmnopq} \end{aligned}$$

Fixed Variables

$$\begin{aligned} \alpha_i &= \text{passenger\_count (6 levels)} \\ \delta_j &= \text{month (12 levels)} \\ \gamma_k &= \text{toll\_ind (2 levels)} \\ \theta_l &= \text{rate\_code (3 levels)} \end{aligned}$$

Random Variables

$$\begin{aligned} c_m &= \text{pickup\_location (50 levels)} \\ d_n &= \text{dropoff\_location (50 levels)} \\ f_o &= \text{pickup\_time (24 levels)} \\ g_p &= \text{dropoff\_time (24 levels)} \\ e_q &= \text{error} \end{aligned}$$

Full Model Results

We can see that the results of the full model indicate that we may have overparameterized our model. For certain interactions, we may not have a balanced or sufficient number of observations for each category, so we don't have enough degrees of freedom to provide useful F-statistics. So, we turn to backward selection to try to reduce the number of groupings. The four-way interaction was insignificant (p-value > 0.05) and since it is the largest interaction, was removed first. Then, one by one, the insignificant three-way interactions were removed. This left one three-way interaction that became significant in the model. The insignificant individual predictors were left in the reduced model because of their significant interaction effects. In addition, we can see that the standard error

of the pickup time is smaller than the variance estimate, so we removed that as well.

Full Model

The GLIMMIX Procedure

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
PICKUP_LOCATION_ID	0.001654	0.000407
DROPOFF_LOCATION_ID	0.004034	0.000877
PICKUP_TIME	0.000017	0.000053
DROPOFF_TIME	0.002620	0.000816
Residual	0.1155	0.000632

Fit Statistics	
-2 Res Log Likelihood	46577.59
AIC (smaller is better)	46587.59
AICC (smaller is better)	46587.60
BIC (smaller is better)	46597.15
CAIC (smaller is better)	46602.15
HQIC (smaller is better)	46591.24
Generalized Chi-Square	7723.42
Gener. Chi-Square / DF	0.12

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
LOG_DIST	1	54759	23268.2	<.0001
PASSENGER_COUNT	1	1	0.00	0.9998
MONTH	1	1	0.00	1.0000
TOLL_IND	1	66893	Infty	<.0001
RATE_CODE	2	66893	0.00	1.0000
MONTH*PASSENGER_COUN	1	1	0.00	1.0000
TOLL_IND*PASSENGER_C	5	1	Infty	<.0001
RATE_CODE*PASSENGER_	8	1	Infty	<.0001
MONTH*TOLL_IND	1	1	0.00	1.0000
MONTH*RATE_CODE	1	1	0.00	0.9999
TOLL_IND*RATE_CODE	2	66893	Infty	<.0001
MONTH*TOLL_I*PASSENG	1	1	0.00	1.0000
MONTH*RATE_C*PASSENG	1	1	0.00	1.0000
TOLL_I*RATE_C*PASSEN	1	2.912	0.00	0.9999
MONTH*TOLL_I*RATE_CO	1	1	0.00	1.0000
MONT*TOLL*RATE*PASSE	35	1	3.60	0.3985

Model Selection Criteria

It may be noted that the AICC for the full model is smaller than the AICC for the reduced model, a criterion that would typically lead the researcher to select the full model specification. However, we have a few reasons why we did not follow the selection criteria. One is the possible overparameterization/lack of degrees of freedom in the full model that provided unreliable estimates. Second, the AICC in the full model is only about 1% smaller than the reduced model, so the difference in fit is not radically different. Finally, the residual diagnostic plots in each model looked similar, so we do not believe there is a true difference in fit between the two models.

Fixed Effects Results

After the selection process, we are left with the results for the reduced model, found in table X. The degrees of freedom and resulting p-values and F-statistics are more useful in this model. Based on the tests for fixed effects, we see significant effects from trip distance, passenger count, month*passenger_count, rate_code*passenger_count, month*rate_code, toll_ind*rate_code, and the three way interaction of month*toll_ind*passenger_count in tip amounts.

Reduced Model

The GLIMMIX Procedure

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
PICKUP_LOCATION_ID	0.001657	0.000406
DROPOFF_LOCATION_ID	0.004015	0.000873
DROPOFF_TIME	0.002615	0.000801
Residual	0.1161	0.000635

Fit Statistics	
-2 Res Log Likelihood	47015.24
AIC (smaller is better)	47023.24
AICC (smaller is better)	47023.24
BIC (smaller is better)	47030.89
CAIC (smaller is better)	47034.89
HQIC (smaller is better)	47026.15
Generalized Chi-Square	7779.78
Gener. Chi-Square / DF	0.12

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
LOG_DIST	1	54891	23137.4	<.0001
PASSENGER_COUNT	5	66906	15.04	<.0001
MONTH	11	1	13.63	0.2085
TOLL_IND	1	66940	2.85	0.0913
RATE_CODE	2	1	79.28	0.0792
MONTH*PASSENGER_COUN	55	66896	3.43	<.0001
TOLL_IND*PASSENGER_C	5	66899	0.91	0.4720
RATE_CODE*PASSENGER_	8	66898	9.66	<.0001
MONTH*TOLL_IND	11	66894	3.48	<.0001
MONTH*RATE_CODE	22	1	8.56	0.2642
TOLL_IND*RATE_CODE	2	66728	30.59	<.0001
MONTH*TOLL_I*PASSENG	55	66897	3.87	<.0001

Fixed Effects: Main Effects

The main effects indicate that distance has a positive effect on tip amount, while the effect from passenger_count depends on the number of passengers. Trips with 2, 3, or 5 passengers resulted in significantly negative effects on tip amount relative to trips with 6 passengers. Trips with 1 or 4 riders were not significant relative to 6 passengers. There was not a significant month effect overall, but when we examine the solutions for each month coefficient, we find that trips taken in January have a significant positive effect on tip amounts relative to trips taken in December, and trips in the months of March, June, and August have a significant negative effect on tip amount relative to December. In the same way, the toll_ind effect was overall insignificant, but solutions show that trips without a toll have a significant positive effect on tip amounts relative to those with a toll.

Fixed Effects: Interaction Effects

Most of the interaction effects' interpretations are too specific to be useful. For example, trips with 3 passengers during the month of March give significantly smaller tips than trips with 6 passengers during March. However, there is one somewhat notable result from these interactions. Trips with rate codes 1 and 2 result in significantly different tips from rate code 5 for nearly all months relative to December (other than October and rate code 2 in November). Despite their impractical interpretations, it is still important to include these interaction effects in the model in order to obtain accurate estimates of the main effects.

Random Effects

Figures 3 and 4 depict the significant random effects due to pickup and dropoff location clusters. There are more significant dropoff locations than pickup locations, which could be related to the tip being an action performed at the end of the trip. High pickup effect locations are relatively close to the United Nations HQ and the Nomad District. In contrast, a low pickup effect location is relatively close to the Empire State Building. High drop off effect locations are near NYSE, N Bronx, Williamsburg Bridge Area, Hudson River Park, East Harlem, and JFK International Airport. Low dropoff effect locations are near NY University and Hudson Yards Railway Station.

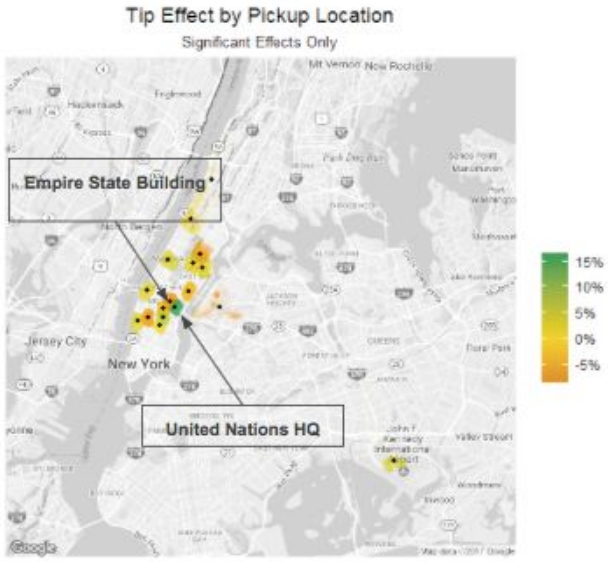

Figure 3: Significant predictors by pickup location. Percentage indicates percent of tip increase compared to insignificant location.

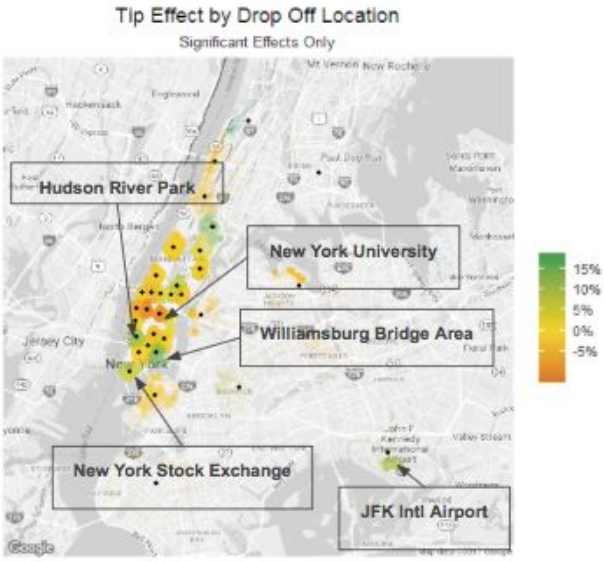

Figure 4: Significant predictors by dropoff location. Percentage Indicates percent of tip increase compared to insignificant location.

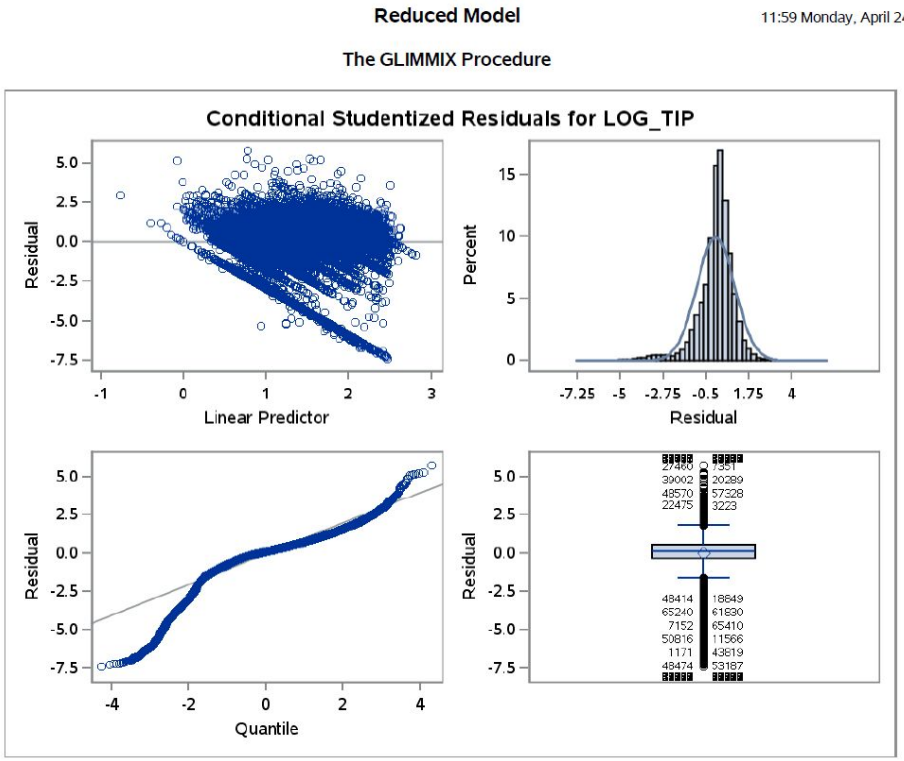

Figure 5: Model diagnostics for the reduced model.
Upper left. Linear predictor by residual. **Upper Right.** Distribution of residuals.
Lower Left. QQ plot of residuals. **Lower Right.** Boxplot Residual error.

Residual Diagnostics

The top left figure of the residual diagnostics shows a generally random pattern. The series of points that appear to be a straight line are likely due to observations where passengers did not tip, but the model predicted a non-zero tip amount, given the trip characteristics. The top right figure shows a somewhat normal distribution of residuals with some slight left skewing. The bottom left figure illustrates a relatively normal QQ Plot considering the large size of the data. The lower left tail exhibits some departure from normality, which again is likely because of the prediction of tips for observations where there were no tips given. The bottom right figure shows fairly symmetrical errors with outliers, common for large datasets.

Summary

Our objective was to determine which factors best predict the amount a passenger tips his/her taxi driver. From a stratified random sample taken from the year 2014, we first processed the data by removing outliers, questionable data, cash transactions. Using K-means clustering we interpreted the GPS location for the pickup and drop-off and coded the location according to one of 50 "sectors" so that we could observe trends for various locations. We used SAS to generate a generalized linear mixed model to predict the trip amount (in US dollars) based on the Month, Toll-indicator, Rate-code, Passenger Count, Trip Distance, Pickup Time, Drop Off Time, Pickup Location and Drop off Location. Log transformations were used on the continuous variables.

In our full model, we found that the Standard Error for the Pickup Time was larger than the effect estimate, so this variable was not included in the reduced model. In addition, we found that many of our four-way and three-way

interactions were insignificant, so these were also dropped. Our reduced model indicates that significant Fixed effects on Tip Amount are: Trip Distance, Passenger Count, Month*Passenger Count, Rate-code*Passenger Count, Month*Toll-indicator, Toll-indicator*Rate_code, and Month*Toll-Indicator*Passenger Count. Significant Covariates include the Pickup Location, Drop off Location, and Drop off Time.

We found that taxi trips of longer distances result in higher tips. The tip amount seems to be somewhat related to the number of passengers on the trip, particularly for trips with 2, 3, or 5 passengers. If no toll is paid, the tip amount seems to be significantly higher than if a toll is paid. There seems to be no significant differences in tips based on the month in which the trip is made. The visual information on pickup and drop-off location indicates that tips are higher when dropped off at locations where riders are likely to be more affluent such as Wall Street, JFK Airport, or the United Nations. Analysis of the effect of the Hour in which a rider is dropped off indicates that passengers tend to tip more in the morning hours between 4am and 1pm and in the evening between 7pm and 11pm.

Citations

TLC Trip Record Data. (2016). *NYC Taxi & Limousine Commission*. Retrieved from http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

Latex equation generated in R. See accompanying code.