Predicting Uber and Lyft Fares Using Linear Regression

Simran Parajulee, directed by Dr. Giancarlo Schrementi
Department of Mathematics and Statistics
Hollins University, Roanoke, VA

OBJECTIVES

The goal of this research is to create a fare prediction model for ride hailing companies Uber and Lyft in the Greater Boston area. We create multiple linear regression models for the two companies and compare the difference in their pricing strategies.

INTRODUCTION

Uber and Lyft have made commute reliable and convenient, especially for individuals who do not own personal vehicles. Regular consumers of these services often experience unusual price fluctuations for a given source and destination. Finding a model that accurately predicts fares can help consumers decide the best choice for commute.

DATA

To build our models, we use a sample dataset available in Kaggle for Uber and Lyft price pings collected in Boston, MA. The dataset contains 110,190 data points for UberX and UberXL, and 102,470 data points for Lyft and LyftXL. We dropped rows with information regarding any other type of Uber/Lyft.

REGRESSION MODEL

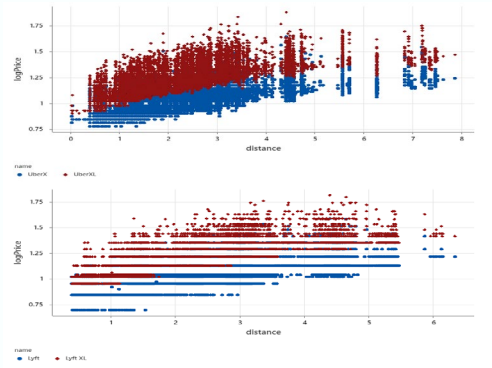
The following regression equations were used to build our models:

$$\text{Log(Fare)}_{\text{Lyft}} = 0.49 + 0.086\,D + 0.283\,S + 0 * \text{LyftX} + 0.198 * \text{LyftXL}$$

And $\text{Log(Fare)}_{\text{Uber}} = 0.82 + 0.073\,D + 0 * \text{UberX} + 0.202 * \text{UberXL}$

where Fare is the predicted fare, D is the distance between source and destination, S is the surge multiplier and LyftX, LyftXL, UberX and UberXL are the categorical encodings representing the type of Lyft or Uber respectively.

SCATTERPLOTS OF PRICE VS DISTANCE



RESULTS & DISCUSSION

Lyft
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0444113	91.31%	91.31%	91.31%

Uber
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0647167	80.68%	80.67%	80.67%

For both Uber and Lyft, the overall model, as well as all predictors used were statistically significant. For Lyft, our model explained 91.31% of the variability in Log(Fare) while for Uber, our model explained 80.67% of the variability in Log(Fare).

FUTURE RESEARCH

- We plan to get Uber and Lyft fare data from more areas in Boston and fit the data into out model.
- We also plan to study why we saw large standard deviation of fares for a given source and destination.

LIMITATIONS

One of the major limitations for our study was the data collection method. All data was collected during the month of November, in a particular part of Boston. Hence, the model might not accurately predict fares during other times of the year throughout the state.

Furthermore, for the Uber dataset, conditions for inference were not met as we can see a strong skew to the right in the histogram of residuals. This indicates that residuals are not normally distributed.

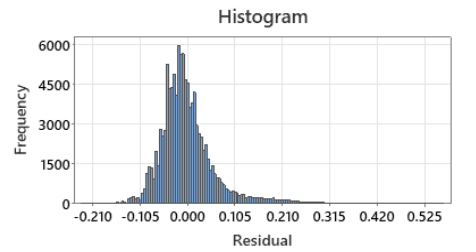


Figure 1 : Histogram of distribution of residuals for Uber

REFERENCES

1. Napitupulu, J. H. (2015, April 22). *Conditions and inference of linear regression*. Data Science, Python, Games. Retrieved April 20, 2022, from http://napitupulu-jon.appspot.com/posts/conditions-inference-linear-regression-coursera-statistics.html
2. *Normal probability plot of residuals*. PennState: Statistics Online Courses. (n.d.). Retrieved April 20, 2022, from https://online.stat.psu.edu/stat501/lesson/4/4.6#:~:text=Skewed%20residuals,terms)%20are%20not%20normally%20distributed.