Umang Desai, Nelson Mei, Saarthi Jethi
Professor John Byers
CS 591: Final Paper
12/18/19

The Substitution Effect Of Blue Bikes on MBTA

Introduction

With the rise in the sharing economy, companies like Airbnb, Uber, and Lyft have been a popular choice for many in the hospitality and transportation industry. Another company that works under the same, 'fee-based' sharing model is Bluebikes.

Bluebikes (previously Hubway) is a publicly owned bike sharing service that allows users to rent bikes for 30-45 minute intervals for a small fee. It has become the main bike sharing system in the Boston metropolitan area since 2013. With over 200+ station across the Greater Boston region and over 2100+ available bikes, Bluebikes is prevalent in college campuses, downtown Boston, and in suburban locations around Boston. Their models generates a ridership of about 2 million rides each year with a total of over 8 million rides recorded since 2011.

There are two primary ways in which users can ride Bluebikes: pay for a single or day pass or become a subscribed member for a month or a year. These options allow Bluebikes to be a great choice for all classes and age groups, including businessmen and tourists. It has proven to be a great competitor to other modes of transport such as walking, driving, or taking the train, especially since it is accessible 24-hours a day.

The only other public mode of transport in a similar price range is the Boston MBTA system, more commonly known as "the T". Since the 1960s, the T has made all of Boston's areas very accessible. Today, the T consists of 149 subway stops over 5 main routes with a ridership of over 400 million each year. The T runs from 5am to around midnight.

It is worth mentioning that there is a Blue Bike stop near almost every MBTA subway stop which makes the two public transportation systems comparable in nature, and also motivates our research.

Motivation & Problem Statement

Knowing that we almost never have to wait to rent a Bluebike and that it is accessible at any point has been a blessing to college students like us. We all have used Bluebikes, whether it is out of desperation at 3am to get home or it is to hang out with friends in the middle of the day. Similarly, all of us have used the T extensively to travel around Boston, for the most part to escape the misery we would be in by standing in the unexpected rains or chilly wind.

As a team, we wondered on what factors play into this decision on taking the T vs. choosing to ride a bike. At what threshold do we have the switch between biking and riding the

T. Is it merely the fact the Bluebikes are more accessible or the weather is bad or there are other features at play?

More specifically, is Bluebikes a substitute or a compliment to the MBTA? If it is, what factors determine the substitution or complementary effect and what thresholds need to be met to make this switch?

Research Questions / Hypothesis

As the most basic question, Bluebikes is inconvenient because of the time it takes to travel the same distance is slightly higher.

The substitution effect over the weekdays might be higher because more people will be going to work and not many people use blue bikes for leisure..

We believe that Bluebikes are substitutes to the MBTA when the trains do not run and also when there are no easily accessible subway stop.

Considering other factors, we also believe that Bluebikes are used less in the morning when people are rushing to go to work but are used more in the evening when there is no rush.

Datasets, Collection, and Tools

For our research and analysis we have publically available datasets for Bluebikes and Boston weather, combined with data from Google maps.

Bluebikes Datasets

The Bluebikes public datasets included a list of bike stations consisting of 281 stations with their names, lattitudes, longitudes and capacity of bikes. We also used their trip history data for individual trips from 2011 to 2018 consisting of around 8 million trips. Each trip included the trip duration, start and stop times and stations and some user data.

Boston Weather Data

We found daily weather reports for Boston starting January 2013 through April 2018. Our main interest in this dataset were daily average temperature, precipitation, snowfall, and wind speed.

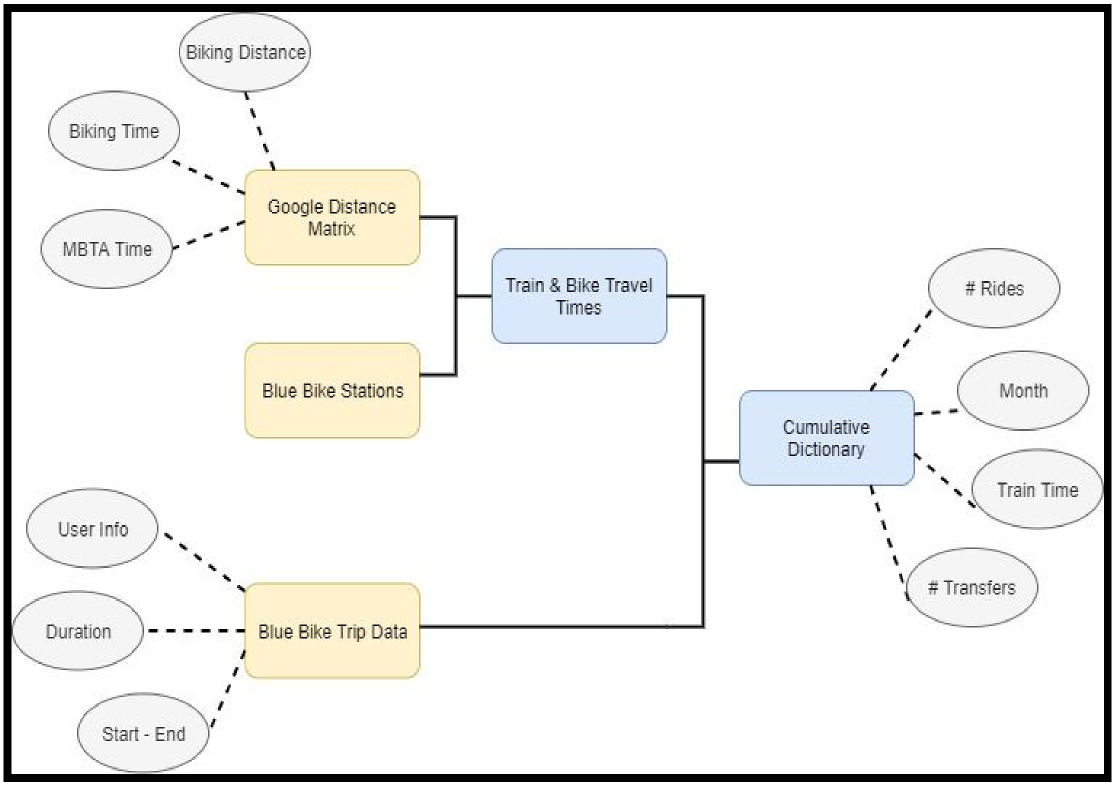
Google Distance Matrix

For every unique combination of Blue Bike station pairs, we used the Google Distance Matrix to find the biking distance, biking time, total time in the train, the total time to walk to and from the train stations, and the number of line transfers needed to travel between the two station pairs.

Data Cleaning Process:

Firstly, the Bluebikes dataset was used to gather the 281 bike stations across Boston. We used those stations to create 39,340 possible combinations of bike routes that can be taken. Each of these pairs of sources and destinations were passed through to the Google Directions API to find the distance between the source and destination, approximate time it would take to walk, bike, and take the train from a source to a destination. Also included was the number of transfers that would be involved during the trip if one were to take the train, and the approximate amount of walking time involved if the train option were to be taken. These 39,340 rows were stored into one csv file.

A main dictionary was then created for each source and destination pair. The keys in the dictionary are each of the (source, destination) pairs, and each value was another dictionary which included biking distance, and 12 sub-dictionaries for each month in a year. These sub-dictionaries include information such as total trips that month, total biking time, total morning trips, total evening trips, total weekday trips, and total weekend trips. The python datetime package was used to extract information out of the date, such as finding out what day of the week a certain trip occurred on.



Bluebikes trip data was then loaded into the dictionaries. There were 6,987,701 trips listed in the dataset from January 2011 to December 2018. After cleaning the data and filtering out trips that contained invalid values, we ended up with 4016190 trips.

The dictionary was then used to create a csv file that was ready for regression analysis. Each row represented a source and destination pair, and included metrics provided by Google Directions API as well as information from our main dictionary, such as average bike rides in January, the average number of daytime trips, and average number of nighttime trips.

On a separate table, we combined daily trip data for the past 8 years, which included the total number of rides that day for a given source destination pair, with the Boston weather dataset.

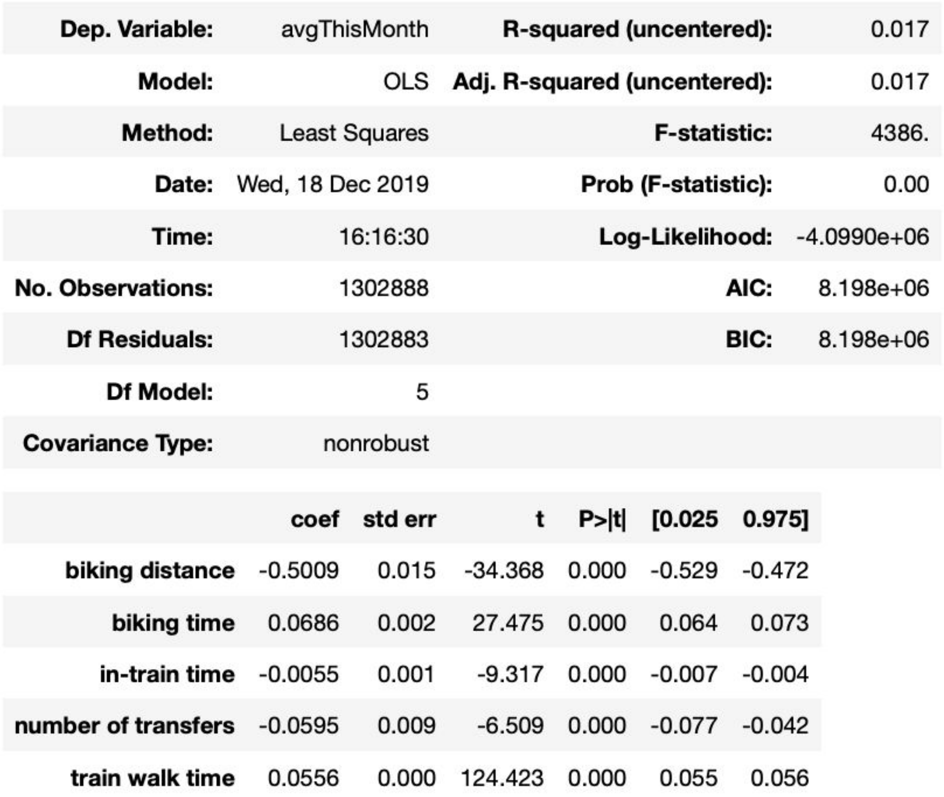
Our baseline regression included distance and biking time, and weather was one of the variables that we added to our baseline model.

Methodology and Discussion of Results

Baseline Model: MBTA vs. Bluebikes, substitution solely based on time

To quantify the impact of MBTA's inconvenience on Blue Bike ridership, we first start by modeling the ridership by inconvenience caused by time:

$$\#rides_{ijm} = \beta_1 \cdot bikingDistance_{ij} + \beta_2 \cdot bikingTime_{ij} + \beta_3 \cdot inTrainTime_{ij} + \beta_4 \cdot trainWalkTime_{ij} + \beta_5 \cdot \#transfers_{ij}$$



Dep. Variable:	avgThisMonth	**R-squared (uncentered):**	0.017
Model:	OLS	**Adj. R-squared (uncentered):**	0.017
Method:	Least Squares	**F-statistic:**	4386.
Date:	Wed, 18 Dec 2019	**Prob (F-statistic):**	0.00
Time:	16:16:30	**Log-Likelihood:**	-4.0990e+06
No. Observations:	1302888	**AIC:**	8.198e+06
Df Residuals:	1302883	**BIC:**	8.198e+06
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
biking distance	-0.5009	0.015	-34.368	0.000	-0.529	-0.472
biking time	0.0686	0.002	27.475	0.000	0.064	0.073
in-train time	-0.0055	0.001	-9.317	0.000	-0.007	-0.004
number of transfers	-0.0595	0.009	-6.509	0.000	-0.077	-0.042
train walk time	0.0556	0.000	124.423	0.000	0.055	0.056

Our baseline model shows all variables to be statistically significant **

In our model, *i* and *j* are the source and destination and we are predicting the number of rides per month:

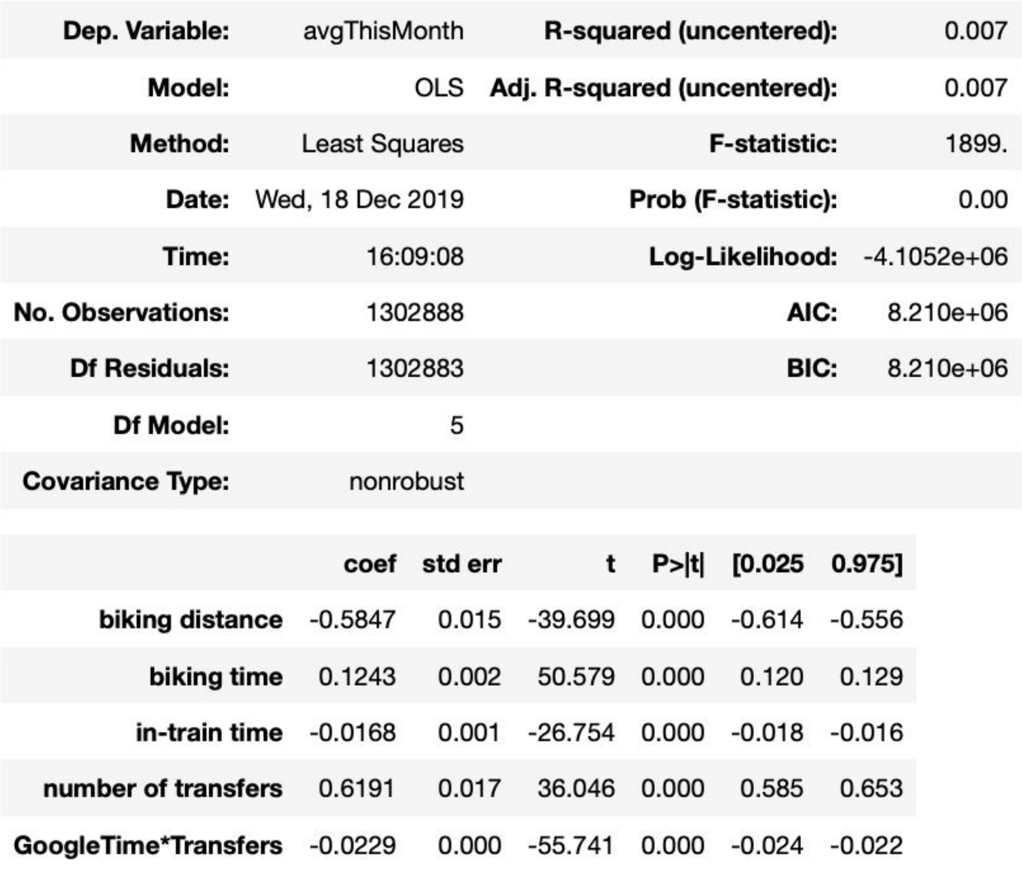
- Biking Distance: When the biking distance increases 1 mile, the average monthly ridership decrease by half a ride per station.
- Biking Time: When biking time increases by 14.5 minutes, average monthly ridership increases by 1 ride per station.
- In-train time: When in-train time increases by 181 minutes, average monthly ridership decreases by 1 ride per station.
- Number of transfers: When number of transfers increases by 1, average monthly ridership decreases by 0.0595 rides per station.
- Train Walk Time: When train walk time increases by 1, average monthly ridership increases by 0.0556 rides per station.

This model is not enough to decide whether Bluebikes is a substitute or a complement to the MBTA. It gives us preliminary data, and our model requires more granularity in the future.

Is there a substitution effect when the number of transfers in the subway increases?

We hypothesized that if the number of transfers needed in the T are high, people would rather bike than take the T. To model that, we used an interactive variable model in which the bikiting time is multiplied by the number of transfers. The number of transfers being a discrete variable which can take on values [0, 1, 2, 3, 4] would show us how much do we lose ridership if the number of transfers increases based on how long the riders need to bike.

$$\#\text{rides}_{ijm} = \beta_1 \cdot bikingDistance_{ij} + \beta_2 \cdot bikingTime_{ij} + \beta_3 \cdot bikingTime_{ij} \cdot transfers_{ij}$$



Dep. Variable:	avgThisMonth	**R-squared (uncentered):**	0.007
Model:	OLS	**Adj. R-squared (uncentered):**	0.007
Method:	Least Squares	**F-statistic:**	1899.
Date:	Wed, 18 Dec 2019	**Prob (F-statistic):**	0.00
Time:	16:09:08	**Log-Likelihood:**	-4.1052e+06
No. Observations:	1302888	**AIC:**	8.210e+06
Df Residuals:	1302883	**BIC:**	8.210e+06
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
biking distance	-0.5847	0.015	-39.699	0.000	-0.614	-0.556
biking time	0.1243	0.002	50.579	0.000	0.120	0.129
in-train time	-0.0168	0.001	-26.754	0.000	-0.018	-0.016
number of transfers	0.6191	0.017	36.046	0.000	0.585	0.653
GoogleTime*Transfers	-0.0229	0.000	-55.741	0.000	-0.024	-0.022

Note: biking time is the same as google time

When we added this new factor, one thing to notice is that the coefficient on the biking time doubled. This means that with no transfers, the ridership increases by 0.12 per station pairs per month. But surprisingly, against our hypothesis, with an increase in the number of transfers, has an inverse effect on the biking time. This means that if there are numerous transfers, then an increase in biking time by 1 minute increases the ridership by 0.1 per station pair per month.

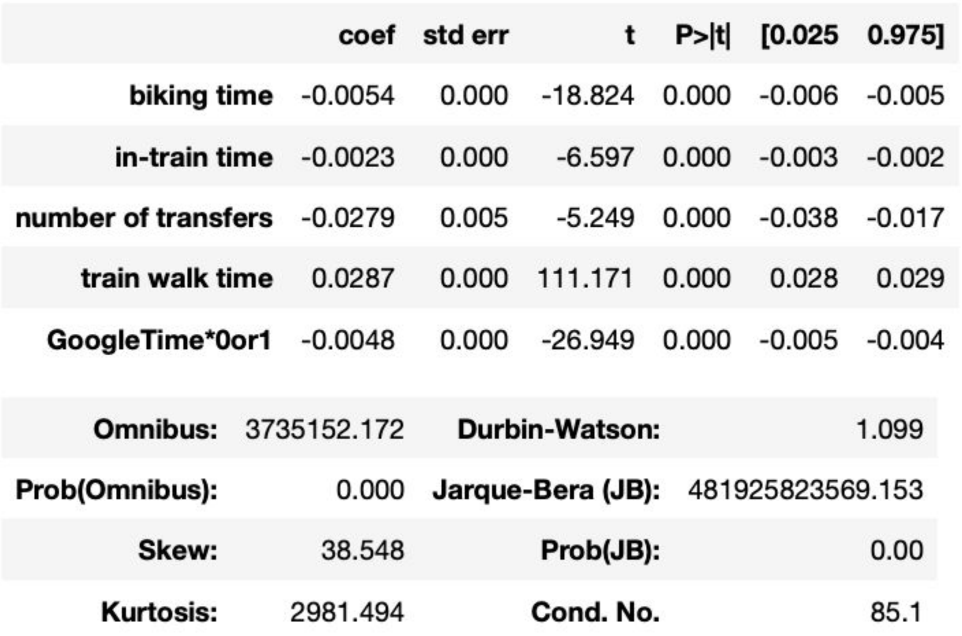
This could potentially be because the number of transfers could be an indicator of distance and the higher the distance, the longer the biking time. And the longer the biking time, the less likely is the substitution rate from Bluebike ridership to the MBTA

Modeling weekend trips vs. weekday trips

Similar to our previous model, we wanted to see if people would choose Bluebikes over MBTA more over the weekdays than weekends.

Note: In this case our response variable is the number of rides on either a weekday/weekend in a given month from stop *i* to *j*. To do this, we created a 0-1 binary variable to mark trips to be either a weekend trip or a weekday trip.

$$\#\text{rides}_{ijm} = \beta_1 \cdot bikingDistance_{ij} + \beta_2 \cdot bikingTime_{ij} + \beta_3 \cdot bikingTime_{ij} \cdot weekdays/weekend$$



	coef	std err	t	P>|t|	[0.025	0.975]
biking time	-0.0054	0.000	-18.824	0.000	-0.006	-0.005
in-train time	-0.0023	0.000	-6.597	0.000	-0.003	-0.002
number of transfers	-0.0279	0.005	-5.249	0.000	-0.038	-0.017
train walk time	0.0287	0.000	111.171	0.000	0.028	0.029
GoogleTime*0or1	-0.0048	0.000	-26.949	0.000	-0.005	-0.004

Omnibus:	3735152.172	**Durbin-Watson:**	1.099
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	481925823569.153
Skew:	38.548	**Prob(JB):**	0.00
Kurtosis:	2981.494	**Cond. No.**	85.1

The new multiplicative variable is called googletime*0or1

Our conclusion was that the coefficient 0.0048, is too small to show any real difference and our hypothesis could not be tested with the format of our data with more granularity.

Related Work

ScienceDirect's "Sustainable bike-sharing systems: characteristics and commonalities across cases in urban China" is a paper that investigates the varying levels of success of different bike sharing apps in different major cities, such as Shanghai, Beijing, and Wuhan. It first goes into describing the various advantages of using bike sharing, which are the main reasons people use bike sharing. The article notes that the traffic congestion problem in China's major cities is part of what led to higher demand for this sustainable commuting option. The article concludes by describing the causes for the various success levels of the different competing companies. These reasons include lack of instructions for users, not enough bike stations, bikes stations being placed in inefficient locations, and poor bike design. This article's focus is different than that of our paper, as our paper involves weather and accessibility to public transit and how they affect the amount of Blue Bike rides.

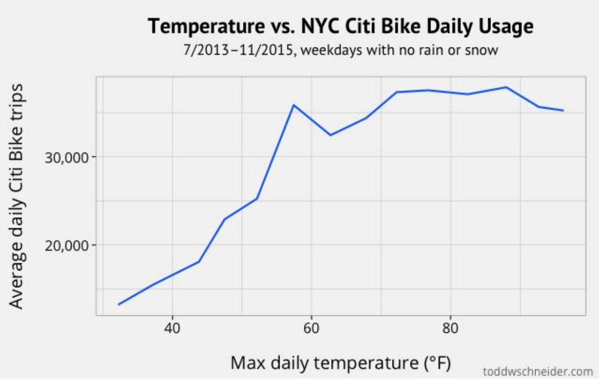
Adopted from Todd Schneider's Analysis

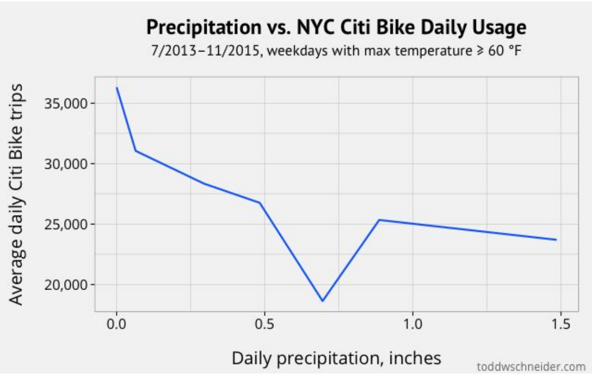
Linear Regression Model:

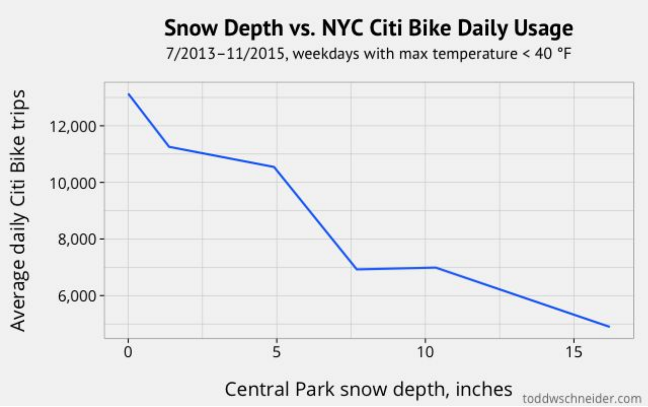
The author ran a Regression Model that calculates an estimate of how long each CitiBike ride took based on the Google Map route they used. The various factors that were used in this were their gender, age, and distance in miles. Another factor that we could predict was, how many bikes were manually moved to a different station after being dropped off. By taking every bike drop off and correlating it with the percentage of rides where the bike's next trip started at a different station from where the previous bike was dropped off. In March 2015, around 13% of bikes were transported from their dropoff stations to a different station.

Non Linear Regression Model:

Biking is probably less popular on rainy and snowy days and so we have to look at the various factors that show a relationship between ridership and the weather. Some factor as stated above daily max temperature, daily precipitation, and daily snow depth. This is a non-linear regression model because the total number of trips would have to be positive, but a standard regression model can have a negative number. The other reason is the relationship between bike rides and weather is nonlinear because the change in temperature can have different magnitudes.







In this model, we can see that the difference between no rain and 0.1 inches of rain is more significant than the difference between 0.5 and 0.6 inches of rain.

$$d_{trips} = Baseline(d) + Weather(d) \tag{1}$$

$$Baseline(d) = e^{\beta_{const} + \beta_{wday} \cdot d_{weekday} + \beta_{expansion} \cdot d_{expansion}} \tag{2}$$

$$Weather(d) = \beta_{weather} \cdot \frac{1}{1 + e^{\frac{-(WeatherFactor(d) - \beta_{center})}{\beta_{width}}}} \tag{3}$$

$$WeatherFactor(d) = d_{maxtemp} + \beta_{precip} \cdot d_{precip} + \beta_{snowdepth} \cdot d_{snowdepth} \tag{4}$$

Dtrips is the number of Citi Bike trips on a specific date. The Baseline variable is an exponent that will give a positive output. It is a factor for bikes before there are more stations that are added. Then, Weather is a combination of the maximum temperature, precipitation, and snow depth. The residuals on this graph are normally distributed. The residuals have a lower variance on these dates with fewer trips. These models are able to see how weather and ridership are correlated with each other.

Conclusion

It is evident that BlueBike has a substitution effect as a mode of transportation in Boston. It is visible from our first regression that the ridership is dependent on the time MBTA takes to run and the distance from the rider's current location to the nearest MBTA stop.

It was hypothesized that an increase in the number of transfers or an increase of train time would cause an increase in the amount of BlueBike ridership. Running the regression showed that longer train times and more transfers, factors that are inconvenient for train passengers, do not increase the ridership of the nearby Blue Bike stations. Therefore, this analysis has clearly disproved our initial hypothesis of having train time and transfer as a direct correlator with how many people ride BlueBikes.

Regarding the substitution effect over the weekend, the effect was too small to be visible because of the granularity of the data. Because Boston is so well-connected, we did not find enough locations where BlueBikes was the only mode of transportation.

Future Works:

Modeling Ridership and Substitution to the MBTA based on weather

It is apparent that snow and rain would increase ridership of the MBTA and decrease the ridership of Bluebikes. Future research could be conducted based on interactive linear models like the ones proposed below.

Ridership based on all conditions

$$\#rides_{ijm} = \beta_1 \cdot bikingtime_{ij} + \beta_2 \cdot inTrainTime + \beta_3 \cdot snow + \beta_4 \cdot rain + \beta_5 \cdot temp + \beta_6 \cdot wind$$

Isolating the effect on snow and rain -- high substitution against ridership

$$\#rides_{ijm} = \beta_1 \cdot bikingtime_{ij} + \beta_2 \cdot inTrainTime + \beta_3 \cdot snow \cdot inTrainTime + \beta_4 \cdot rain \cdot inTrainTime + \\ \beta_5 \cdot snow \cdot bikingDistance + \beta_6 \cdot rainc \cdot bikingDistace$$

Citations

Schneider, T. (2016, January 13). A Tale of Twenty-Two Million Citi Bike Rides: Analyzing the NYC Bike Share System. Retrieved December 2, 2019, from.

Zhang, Lihong, et al. “Sustainable Bike-Sharing Systems: Characteristics and Commonalities across Cases in Urban China.” Journal of Cleaner Production, vol. 97, 15 June 2015, pp. 124–133., doi:10.1016/j.jclepro.2014.04.006.