**Predicting Public Transportation Demand Using Household and Neighborhood Characteristics**

Alexa Canaan

Colgate University

**Predicting Public Transportation Demand Using Household and Neighborhood Characteristics**

Abstract – *This paper aims to predict public transportation demand by comparing the household and neighborhood characteristics of those who do and do not use public transportation. Data from the 2013 American Housing Survey is analyzed to compare public transportation usage across a representative sample of households in the United States. Both OLS and logit regression models found that distance from public transportation is a statistically significant factor in determining public transportation usage. More coming soon…*

**Key Words:** Public Transportation, OLS, Logit Regression, LASSO Regression, Public Economics, *more coming soon*

**JEL Codes:** *coming soon…*

1. **Introduction and Motivation**

Public transportation has the ability to revolutionize a geographic area by increasing travel speeds, decreasing commute times, facilitating labor market matching, and providing more opportunities for leisure and consumption (Chin, Kahn, Moon, 2017, p.3). For individuals, the ability to reach necessary services and spatial opportunities is a key quality of life metric that can be improved with a range of transportation options, public and private (Karou et al., 2014, p.1). Not only that, but with traffic congestion being a source of economic cost for most industries and sectors, increasing the general use of public transportation can ease traffic flows and significantly increase productivity for individuals and benefit business (Weisbrod, Vary, Treyz, 2003, p.98). However, this is only true to the extent that public transportation is available to a population that will use it. For people who live far away from public transportation to the point that it is inaccessible, they cannot take advantage of its benefits even though they may choose to do so if it was accessible. My motivation for conducting this analysis is to predict public transportation usage across America using household and neighborhood characteristics. Variables of interest including distance to public transportation and metro status for households can be used to gauge accessibility to public transportation which is a key component of whether or not individuals actually opt to use it.

This paper will supplement existing research by examining the effect of accessibility on demand within the context of public transportation, as well as more generally. This paper specifically will aim to develop a machine learning model that can be improved upon and used in the future to predict demand of public transportation based on household and neighborhood characteristics of household observations. Ideally, this model will allow researchers to identify inefficiencies in access to public transportation. These inefficiencies occur when demand does not match accessibility, such as when a neighborhood has high demand for public transportation, but there are no stops nearby or when a neighborhood has little to no demand for public transportation, yet there are potentially multiple stops nearby. The predictive capability of the model will allow for policy makers and urban planners to better allocate infrastructures to communities across the United States based on whether or not their accessibility to these public goods correlates to them using said goods.

1. **Background Literature and Conceptual Framework**

Public transportation demand brings together several different avenues of economic thought, specifically discrete choice theory, estimating demand, and the transportation sector. Predicting demand has become increasingly important, especially within the public transportation sector, because it allows policy makers to optimize the transportation network holistically (Palacio, 2018, p.3).

Discrete choice models predict choices between discrete alternatives, like in the case of whether to use public transportation, or not use public transportation.

Recent research has focused on using machine learning to predict discrete choice and demand specifically for public transportation. Since most machine learning techniques are relatively new, many papers related to prediction have compared the accuracy of different techniques to justify their results (Omrani, 2015, p.841). This is also beneficial to economics and research as whole as it contributes to the goal of identifying what technique is best for making predictions (Kleinberg et al., 2015).

Access to a range of transportation options is a source of inequality. Over time, “the spatial growth of urban areas and the decentralization of employment and facilities have made it harder for people without access to a car to make the daily commute and to take advantage of distributed retail and leisure opportunities” (Karou et al., 2014, p.1).

Many papers focused on public transportation demand focus on attitudes towards public transportation. Borhan et al. (2014) studied travelers’ willingness to use public transportation by focusing on a wide range of variables from service quality, environmental impact, attitude, and behavior intention, with the specific intention of trying to find ways to better public transportation to increase its usage.

This paper adds to current literature by utilizing data from all population densities in the United States from urban centers to rural areas. Papers such as Omrani (2015), Palacio (2018), Chin et al. (2015) all focus solely on urban centers. While data on public transportation is usually only available for metro areas because this is where these systems are placed, researching these systems in a macro context may expose a need for public transportation in areas that are not considered typical areas for this kind of infrastructure.

By focusing on accessibility, this paper contributes to information on public transportation demand by identifying a baseline level of demand for public transportation. Palacio’s (2018) paper using smart card data, for example, aimed to predict daily and hourly demand of public transportation by looking at variables indicating peak travel hours, whether or not it was a workday, the season, and daily weather. This research allows the government to refine public transportation infrastructure that is already in place by increasing or decreasing the frequency of buses and trains on route at different time intervals. However, this paper will aim to predict whether or not certain areas require a large investment in infrastructure for public transportation using accessibility of public transportation as a determining factor of usage.

1. **Data Description and Analysis**

I use data from the American Housing Survey, a longitudinal housing unit survey, to develop my predictive model with all computations and graphics performed in the R programming language(www.r-project.org). The American Housing Survey is a comprehensive survey sponsored by the Department of Housing and Urban Development and conducted by the U.S. Census Bureau biennially in odd-number years, spanning from 1973 to 2017 between May and September. The same housing units from all 50 states and the District of Columbia are surveyed every other year until new samples are drawn, allowing for analysis of households over time. The goal of the data set is to provide timely information on the quality and cost of housing in the United States and American metropolitan areas using the participating housing units, chosen to represent all housing units in the United States. It is used by policymakers to make decisions about housing for all demographics in America (US Census Bureau, 2020).

Each observation in the data set is defined as a “housing unit” or any house, townhouse, apartment, mobile home or trailer, single room, group of rooms, or other location that is occupied as separate living quarters, or if vacant, is intended for occupancy as separate living quarters (US Census Bureau, 2020). The survey is conducted using computer-assisted personal interviewing using laptops. Data is collected from two types of respondents: occupied housing units and unoccupied housing units. Data collected from an occupied housing unit is defined by the US Census Bureau as “A household respondent, who must be a knowledgeable household member 16 years of age or over, provides information on the unit, the household composition, and income” (US Census Bureau, 2020). Data collected from a vacant housing unit is defined by the US Census Bureau as coming from a respondent such as “a landlord, owner, real estate agent, or knowledgeable neighbor” who can provide data on the unit (US Census Bureau, 2020).

This particular data set is applicable to the research question because it has been used by policy makers to plan community development such as infrastructure. I will specifically focus on data from 2013 due to its large range of questions covering public transportation. There were approximately 84,400 sampled housing units with a supplemental sample of 15,533 housing units in the Chicago, Detroit, New York City, Northern New Jersey, and Philadelphia metropolitan areas (US Census Bureau, 2013). Out of the 84,400 sampled housing units, 2,715 were ineligible because the unit no longer existed or because it did not meet the definition of a housing unit (US Census Bureau, 2013). Further, 10,000 units had no response after repeat visits or refused to be interviewed which led to an overall response rate of 86% (US Census Bureau, 2013).

The data does come with its own set of limitations in terms of incomplete data, wrong answers, and sampling variability, much of which is unavoidable. For example, incomplete and missing data are adjusted by assuming that the respondents are similar to its peer households that did report data, assigning them an average value instead of ignoring the missing data (US Census Bureau, 2013). The dataset is not adjusted for wrong answers and does not estimate the size of these errors (US Census Bureau, 2013). Sampling errors are accounted for by the US Census Bureau by proving standard error tables and allowing users to calculate their own sampling errors, although these calculations may vary by user (US Census Bureau, 2013). Recognizing the limitations of the data is important because while they are unavoidable, our ability to take the findings out of context is not. Recognizing that these observations can be flawed is critical to our treatment of the findings as estimations.

In order to learn more about public transportation usage in America, several categorical variables were utilized to create my simplified model. The main variable of interest is PTPUBTRN, recoded to be an indicator variable that is 1 if the respondent uses public transportation and 0 if the respondent does not. METRO3 is a variable that indicates whether or not a person lives in a city center (1), metropolitan area (2), or rural area (3). CARS and TRUCKS indicate the number of cars and trucks per household, respectively. ZINC2 is a numeric variable that represents total household income. The variables PTDISPUB, PTDISBUS, PTDISRAIL, PTDISSHUT, and PTDISSUB represent how far each household is from a general public transportation stop, bus stop, rail stop, shuttle stop, and subway stop, respectively. The distances from public transportation stops for each household are measured on a gradient, with 1 representing less than ¼ mile, 2 representing ¼ to less than ½ mile, 3 representing ½ to less than 1 mile, and 4 representing 1 mile or more. Additional variables that are included in the model accounting for neighborhood effects are WNTRAN, WNAMEN, and WNJOB, which indicate with 1 if the respondent moved to their neighborhood to be closer to public transportation, amenities, or their job, respectively and 0 if they did not.

Before performing a regression analysis, we can look at the summary statistics to get an initial understanding of our dataset. After removing NA also known as missing variables from the dataset, there were 60,096 total observations.

For the machine learning model, it is necessary to remove factor variables that have only one level, because this means that there is no difference in condition and all observations share the same value for said variable. When doing this in R, the number of variables decreased from 4114 to 3717.

1. **Regression Models and Estimation Methods**
2. **Regression Results and Analysis**

It is important to recognize that due to the nature of the dataset, some groups were better represented than others. The American Housing Survey is supposed to provide a representative sample of the United States based on housing characteristics, but this means that certain populations are given less weight than others because they make up less of the United States population.

There is possible measurement error in regard to measuring distance from public transportation stops to households in our OLS and logit models. Since the distance to public transportation stops are grouped in quarter mile increments and then in a group of greater than 1 mile away from public transportation, we are potentially both underestimating and overestimating how far away households are from public transportation stops. These distance gradients could also mean that there is attenuation bias in our coefficients, meaning the estimate coefficient is closer to zero than it should be.

1. **Summary and Conclusions**

Other topics that beg further questioning from this paper include other household characteristics and public transportation use, such as whether those who rent are more likely to use public transportation than those who own property. Not only that, but investigating topics related to accessibility such as neighborhood safety to build upon the model and applying the model to accessibility of other resources more generally like food security could provide

1. **References**

**(APA Style)**

Borhan, M. N., Syamsunur, D., Mohd Akhir, N., Yazid, M., Razuhanafi, M., Ismail, A., & Rahmat, R. A. (2014). Predicting the use of public transportation: a case study from Putrajaya, Malaysia. *The Scientific World Journal*, *2014*.

Chin, S., Kahn, M. E., & Moon, H. R. (2017). Estimating the Gains from New Rail Transit Investment: A Machine Learning Tree Approach. *NBER Working Paper Series*, 1–49. doi: 10.3386/w23326

Karou, S., & Hull, A. (2014). Accessibility modelling: predicting the impact of planned transport infrastructure on accessibility patterns in Edinburgh, UK. *Journal of Transport Geography*, *35*, 1-11.

Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z., 2015. Prediction Policy Problems. American Economics Review. Vol. 105, No. 5, pp. 491-95.

Omrani, H. (2015). Predicting travel mode of individuals by machine learning. *Transportation Research Procedia*, *10*, 840-849.

Palacio, S. M. (2018). Machine Learning Forecasts of Public Transport Demand: A comparative analysis of supervised algorithms using smart card data. *XREAP WP*, *3*.

Team, R., et al., 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna Austria .

**US Census Bureau**. (2013). Appendix B. Sample Design and Weighting: 2013. Retrieved February 1, 2020, from[[https://www2.census.gov/programs-surveys/ahs/2013/2013%20AHS%20National%20Sample%20Design%20and%20Weighting.pdf?#](https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/30941/summary)](https://www2.census.gov/programs-surveys/ahs/2013/2013%20AHS%20National%20Sample%20Design%20and%20Weighting.pdf?)

**US Census Bureau**. (2013). Appendix D. Nonsampling and Sampling Errors. Retrieved February 1, 2020, from [https://www2.census.gov/programs-surveys/ahs/2013/2013%20AHS%20National%20Errors.pdf?#](https://www2.census.gov/programs-surveys/ahs/2013/2013%20AHS%20National%20Errors.pdf?)

**US Census Bureau**. (2019, June 26). AHS 2013 National Public Use File (PUF). Retrieved February 1, 2020, from[[https://www.census.gov/programs-surveys/ahs/data/2013/ahs-2013-public-use-file--puf-/ahs-2013-national-public-use-file--puf-.html](https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/30941/summary)](https://www.census.gov/programs-surveys/ahs/data/2013/ahs-2013-public-use-file--puf-/ahs-2013-national-public-use-file--puf-.html)

**US Census Bureau**. (2020, January 6). American Housing Survey (AHS) About. Retrieved February 1, 2020, from<https://www.census.gov/programs-surveys/ahs/about.html>

**US Census Bureau**. (2020, January 6). American Housing Survey (AHS) Methodology. Retrieved February 1, 2020, from[[https://www.census.gov/programs-surveys/ahs/about/methodology.html](https://www.census.gov/programs-surveys/ahs/about.html)](https://www.census.gov/programs-surveys/ahs/about/methodology.html)

**US Census Bureau**. (2020). AHS Codebook. Retrieved February 1, 2020, from <https://www.census.gov/data-tools/demo/codebook/ahs/ahsdict.html>

Weisbrod, G., Vary, D., & Treyz, G. (2003). Measuring Economic Costs of Urban Traffic Congestion to Business. Transportation Research Record, 1839(1), 98–106. https://doi.org/10.3141/1839-10