**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*

**Key Words:** Public Transportation, OLS, Logit Regression, LASSO Regression, Infrastructure Economics,

**JEL Codes:**

1. **Introduction and Motivation**
2. **Background Literature and Conceptual Framework**
3. **Data Description and Analysis**

I use data from the American Housing Survey, a longitudinal housing unit survey, to develop my predictive model. 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.

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. 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”. 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.

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. 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. Further, 10,000 units had no response after repeat visits or refused to be interviewed. This led to an overall response rate of 86%.

The data does come with its own set of limitations in terms of incomplete data, wrong answers, and sampling variability. For example, incomplete data are adjusted by assuming that the respondents are similar to those not answering and the size of these errors is estimated. The data is also not adjusted for wrong answers and does not estimate the size of these errors. The same applies to sampling variability.

In order to learn more about public transportation usage in America, I utilized several categorical variables to create my simplified model.

1. **Regression Models and Estimation Methods**
2. **Regression Results and Analysis**
3. **Summary and Conclusions**
4. **References**