

Final Report

Formal report

Logan Miller | QMB 3200 | 12.4.2020

Table of Contents

[Introduction 2](#_Toc58015558)

[Methodology 2](#_Toc58015559)

[Results 4](#_Toc58015560)

[ANOVA 4](#_Toc58015561)

[Simple Linear Regression 4](#_Toc58015562)

[Conclusion 6](#_Toc58015563)

[Appendix 7](#_Toc58015564)

# Introduction

There are a lot of factors that play into the number of trips a household may take in a year including the age of the members of the household, the number of workers in the household, the average income of the members of the household, where the household is located, etc. This assignment is to determine the best model which can predict the number of trips this household will take based on a known set of variables and data from one thousand six hundred and thirty-one surveyed households. In this project, the simple linear regression model will be used to determine various possible r squared values and coefficients for different sets of variables. The goal of this project is to determine which model is the best model. The best model will be determined by the highest r squared value which will indicate which model is the best predictor of the expected value relative to the actual value as well as, show us some of the different relationships some variables have by analyzing the change in coefficients when the groups are made up more or less variables.

# Methodology

I used ANOVA to determine which variables were useless and which ones were helpful. According to the ANOVA FIG. 1, all the variables are good indicators of the likelihood of the next trip except for Income which has a Prob>F of 0.5796. Also, resloc3 seems to have no effect. I will continue to use it in future models unless the model chooses to omit it. Simple linear regression is a helpful model to use. Simple linear regression will make clear the coefficients of each variable and the r-squared value which tells us how spread out our data is from the predicted value, based on known data. With an r-squared value of 0.6252, I feel it has a high enough value to considered statistically relevant and accurate enough to predict the number of trips. I did three more linear regression tests, in hopes of seeing a noticeable change and maybe some more insight into how well certain variables stacked up against each other, each test with a decreasing number of variables. I noticed very little difference between the first and second linear regression tests. The difference between these two tests is that I removed the location variables. The r squared value and the coefficients made insignificant changes. The third model (Linear Regression Model Fig. 3) is reduced down to the number of household trips, the number of household residents under twelve years old, and the number of household residents over twelve years old. In this model, the r squared valued dropped from 0.6292 to 0.5926, thus indicating the actual values have a higher level of variance from the expected values. Also, a one unit change in the number of people in the household and in the number of people in the household under the age of twelve has a larger effect because their coefficients when further into the positive and negative, respectively whereas the change in the number of people in the household over the age of twelve has a diminished effect because it moved closer to zero. Finally, for Simple Linear Regression Fig. 4, only the variables for the number of workers in the household, the number of students in the household, and the number of cars owned by the household were included. This model was the least accurate in terms of r squared with a value 0.5177. This value is 5.8% smaller than the original model but the ten percent jump makes a big difference. In this model, the value of being a student or having a job doubles but the number of cars owned by the household triples, increasing the number of trips caused by one unit by three-fold.

# Results

## ANOVA



ANOVA Fig. 1

## Simple Linear Regression



Simple Linear Regression Fig. 1



Simple Linear Regression Fig. 2



Simple Linear Regression Fig. 3



Simple Linear Regression Fig. 4

# Conclusion

In conclusion, the first simple linear regression model is the best. This model has the highest coefficient of determination, or r squared value, which means it is the best predictor of the expected value relative to the actual value. This model also has all of the useful variables which shows how much of an effect a one unit change in each variable has on the total number of trips a household may take in a year.

# Appendix

cd "C:\Users\lmm56\Documents\School\Poly 20 - 21\QMB 3200\Final Project"

import excel "C:\Users\lmm56\Documents\School\Poly 20 - 21\QMB 3200\Final Project\Data\_Option2.xlsx", sheet("Sheet1") cellrange(A1:I1632) firstrow

summ hhsize nchlt12 nchgt12 nworker nstudent ncar income

tab resloc

help anova

anova ntrips hhsize nchlt12 nchgt12 nworker nstudent ncar income resloc

help gen

gen resloc1 - cond(resloc == 1, 1, 0)

gen resloc1 = cond(resloc == 1, 1, 0)

gen resloc2 = cond(resloc == 2, 1, 0)

gen resloc3 = cond(resloc == 3, 1, 0)

regress ntrips hhsize nchlt12 nchgt12 nworker nstudent ncar resloc1 resloc2 resloc3

regress ntrips hhsize nchlt12 nchgt12 nworker nstudent ncar resloc1 resloc2 resloc3

regress ntrips hhsize nchlt12 nchgt12

regress ntrips nworker nstudent ncar

log close

clear