UNDERSTANDING THE DETERMINANTS OF PUBLIC TRANSIT IN CANADA 1981-2016

by

Swiler Boyd (153966b)

Thesis submitted in partial fulfillment of the

Requirements for the Degree of

Bachelor of Science with

Honours in Economics and Computer Science

Acadia University

May, 2023

This thesis by Swiler Boyd

is accepted in its present form by the

Department of Economics

as satisfying the thesis requirements for the degree of

Bachelor of Science with Honours

Approved by the Thesis Supervisor

_____	_____
Dr. Justin Beaudoin	Date

Approved by the Head of the Department

_____	_____
Dr. Andrew Davis	Date

Approved by the Chair, Senate Honours Committee

_____	_____
Dr. Matthew McSweeney	Date

I, Swiler Boyd, grant permission to the University Librarian at Acadia University to reproduce, loan, or distribute copies of my thesis in microform, paper, or electronic formats on a non-profit basis. I, however, retain the copyright to my thesis.

_____

Signature of Author

_____

Date

Acknowledgments

I want to thank my supervisor, Dr. Justin Beaudoin, for his knowledge and guidance in this paper's data collection and writing process. Additionally, I want to thank Dr. Xiaoting Wang for her suggestions on improving this thesis in the editing process. Another important person to mention who significantly contributed to this study's data collection and cleaning process is my fellow student, Hunter Hache.

Most importantly, I want to acknowledge the entire Economics Department at Acadia University, who have positively impacted my academic growth and experience these past four years.

Table of Contents

ACKNOWLEDGMENTS…….. VII

LIST OF TABLES……..…… XI

ABSTRACT……..…… XIII

INTRODUCTION …… 1

LITERATURE REVIEW ……..……...…… 3

2.1 Determinants of Ridership Demand……..……. 3

2.2 Determinants of Ridership Supply ……..…….. 7

2.3 Simultaneous Relationship Between Demand and Supply of Public Transit and Instrumental Variables ……..……..……..……...…..…. 8

DATA SET INFORMATION ……..…. 13

3.1 Data…..……..……...…….....……...… 13

3.2 Data Limitations…….……..……..……. 16

METHODOLOGY …….……..…….…….…... 19

4.1 Model Development……...…….……. 19

4.2 Model I Specification…….…….. 20

4.3 Model II Specification…….……. 21

4.4 Model III Specification……...…….…… 22

RESULTS……..…….……. 23

5.1 Model I Results……..……..…… 23

5.2 Model II Results……..……..… 27

5.3 Model III Results …….……..……. 31

CONCLUSION AND POLICY IMPLICATIONS…….……..…….. 37

REFERENCES……...……..…….. 39

List of Tables

TABLE 1: SUMMARY OF DEPENDENT AND INDEPENDENT VARIABLES USED……...…… 13

TABLE 2: LEGEND FOR VARIABLES USED AND THEIR ABBREVIATIONS……..…… 14

TABLE 3: MODEL I RESULTS……..…. 24

TABLE 4: MODEL II RESULTS…… 28

TABLE 5: MODEL III RESULTS…….…….. 32

TABLE 6: OUTPUT FROM THE FIRST STAGE OF THE 2SLS REGRESSION……...…… 33

Abstract

In February 2021, Prime Minister Justin Trudeau announced a package that would allocate billions of dollars toward improving public transportation over the next eight years. This thesis aims to use regression techniques to determine what characterizes the demand for public transportation. The goal of studying fluctuations in different socioeconomic and demographic characteristics is to find which factors most significantly influence ridership demand. Identifying the variables that have the most substantial influence on ridership demand will allow policies to target different populations to maximize the impact of investment from Trudeau's administration.

This thesis explores 103 transit agencies between 1981-2016, with data coming from the Canadian Urban Transit Association and Statistics Canada. Different regression specifications and techniques are used to achieve results. Model (1) uses ordinary least squares regression analysis and examines the impacts of different socioeconomic variables on linked trips per capita. Model (2) uses the same specification as Model (1) but instead examines passenger revenue kilometers per capita as the dependent variable. Moreover, Model (3) investigates the difference between ordinary least squares and two-stage least squares regression analysis. Previous research has not contrasted the results of different measures of ridership demand, nor has it compared the quantitative differences from utilizing different regression specifications.

The conclusion drawn from analyzing these models was that linked trips per capita is the best measure of ridership demand, which is consistent with the results from previous studies. Additionally, across all specifications, transit supply (measured as revenue vehicle hours) is shown continually to be statistically significant in determining ridership demand. Furthermore,

the results showed that as the population and number of immigrants in a region increase, it can be expected that, on average, more people will take public transportation.

1. Introduction

On February 10th, 2021, Prime Minister Justin Trudeau pledged 14.9 billion dollars towards investing in public transit projects across Canada in the next eight years (Jones, 2021). Investing in public transit will enable Canada to expand its transit infrastructure to connect communities and reach environmental targets. In the announcement of this plan, Trudeau stated, "[w]hile these investments are good for the economy and crucial to our recovery from this global crisis, they are also helping us achieve our climate goals."

Canada has committed to achieving the net-zero emissions target by 2050, and an improved public transportation network will be essential to accomplish this goal. Even though the magnitude of the extent that public transit reduces pollution is unclear, prior research has shown that public transit reduces congestion and improves air quality (Beaudoin et al., 2015; Jiao et al., 2022).

To maximize the impact of this investment, it is essential to understand what determines a person's choice to take public transit. Whether it be the socioeconomic and demographic characteristics of specific populations or changes in a transit agency's operating features, understanding factors that influence a person's choice of mode of transit is essential. If certain populations are found to be more likely to take public transportation, then the government could choose to improve public transit in these regions to maximize the impact of this project.

Determining the amount of transit demanded in a region is usually done by regressing many explanatory variables on either linked or unlinked trips (Boisjoly et al., 2018; Collins & MacFarlane, 2018; Diab et al., 2020; Lee & Lee, 2013; Pasha et al., 2016; Taylor et al., 2009). As discussed in Section 2.1, an unlinked trips counts all transfers a person makes from the place of origin to their destination as multiple trips, whereas a linked trip only counts this as a singular

trip. Generally, it is accepted that linked trips are a more accurate (but less available) measure of ridership demand (Diab et al., 2020; Taylor et al., 2009); however, there is no literature comparing the results of using linked and unlinked trips as dependent variables in different regressions.

Additionally, few researchers have attempted to compensate for the simultaneity of ridership supply and demand and their cyclical impact on each other through two-stage least squares regression analysis. An in-depth analysis of these factors should provide more insight into the components that make up the demand for public transit in a city.

This thesis is laid out as follows. First, a literature review explains other researchers' work concerning transit supply and demand determinants. The literature review is followed by a discussion of the data set studied for this research and the limitations that it presents. Then, the rationale behind the methodology for the different specifications built in this study and an analysis of the results follows. Finally, this paper concludes by discussing the results in terms of policy implications and what it means for the allocation of funds from Trudeau’s transit investment project.

2. Literature Review

2.1 Determinants of Ridership Demand

Most of the available literature on public transit focuses on variables affecting ridership on a neighborhood or city level (Chakour & Eluru, 2016; Foth et al., 2014; Graehler et al., 2019; Kain & Zvi, 1999; Thompson et al., 2012). Focusing on a singular city and a smaller scale makes it difficult to conclude that the results can be applied to other regions with different demographic characteristics. This results in an inability to create generalized policy recommendations that can be readily used across cities and regions. Fewer studies focus on multi-city level studies due to data limitations and the need for uniformity in the data (Miller et al., 2018).

There is a consensus across transit literature that the amount of transit demanded, or the level of ridership, is measured by either linked or unlinked trips (Boisjoly et al., 2018; Collins & MacFarlane, 2018; Diab et al., 2020; Lee & Lee, 2013; Pasha et al., 2016; Taylor et al., 2009). Data for unlinked trips are widely available, and some transit agencies only report their ridership data regarding unlinked trips (Taylor et al., 2009). Unlinked trips do not account for all the transfers an individual makes from their place of origin to their destination. However, a linked trip counts as a singular trip from the source to the final stop, regardless of the transfers a trip takes. For example, if someone had to transfer three times between their first stop and last stop on a transit system, the measure of unlinked trips would count their journey as three separate trips, whereas a linked trip would count as one. Even though most studies use unlinked trips because of the availability of the data, unlinked trips often produce imprecise results by overestimating ridership (Diab et al., 2020; Lee & Lee, 2013; Taylor et al., 2009; Taylor & Fink, 2003). Beginning in 1996, the Canadian Urban Transit Association (CUTA) began reporting linked trips in their yearly report using the variable Total Regular Service Passenger Trips, which

is defined in the CUTA Transit Fact Book as "a linked trip, riding one way from origin to the final destination; passengers whose trips involve transferring from one vehicle to another are counted only once (i.e., transfers are not included)." Even though linked trips are considered a more accurate measure of public transit demanded, no literature has compared the results of unlinked and linked trips and analyzed the differing results. Researchers investigating the determinants of public transportation demand use either linked or unlinked trips because of the availability of the data to measure ridership demand.

Little work has been done using a different dependent variable (other than linked and unlinked trips) to capture ridership. Thompson et al. (2012) employed the variable "passenger kilometers per capita" in their research to explain the factors behind the success of Broward County, Florida's public transportation system. Broward County Transit has the highest ridership per capita and lowest cost per passenger kilometer of all busing systems in the United States metropolitan areas. They found passenger kilometers per capita to be a successful measure of ridership demand and effectively explained the success of this transit system in Florida. The primary advantage of this measure of ridership it that is adjusts for trip length, which gives more weight to longer trips.

Most transit literature examines the determinants of ridership by measuring factors such as vehicle revenue hours, fare price, and transit ridership while controlling for other variables. These variables are typically categorized as either "internal" or "external" variables (Boisjoly et al., 2018; Chakour and Eluru, 2013). Variables that a transit agency has control over are characterized as internal variables. These variables include but are not limited to vehicle revenue kilometers/hours, fare price, fleet size, net operating cost, direct operating expense, and total capital expenditure (CUTA, 2016).

External variables are more extensive, including socioeconomic and demographic characteristics, gas prices, car usage, parking availability, and urban form. Traditionally, these variables the transit operator has little to no control over are known as external variables, and they make up the demand for public transit ridership. Research from Taylor et al. (2009) has been widely accepted by those who have recently studied topics related to the determinants of transit demand region (Boisjoly et al., 2018; Chen et al., 2011; Diab et al., 2020; Pasha et al. 2016). Taylor et al. (2009) studied 265 urbanized regions in the United States. They found that a region's population characteristics, highway characteristics, metropolitan economy, and regional geography influence the number of transit trips taken in an area.

Factors found to be significant by Taylor et al. (2009) include population density, area of urbanization, household income, percentage of the population that are recent immigrants, percentage of college students, and percentage of households without a car. At a 5% significance level, they found that a 1% increase in population density results in a 0.487% increase in vehicle revenue kilometers. As expected, predicted vehicle revenue hours were the most significant indicator of unlinked trips, consistent with other findings. Additionally, they concluded that a 1% increase in population density would increase ridership by 0.42%, while a 1% increase in carless households would increase trips by 1.19%. Ticket prices were also statistically significant, with a 1% increase in fare prices translating to a 0.43% decrease in trips.

Standard socioeconomic variables included as controls in forecasting models are population, service area size, unemployment rate, labour force participation rate, GDP per capita, percent of households without a car, median household income, percent of post-secondary students, and percent of the population that recently immigrated (Boisjoly et al., 2018; Chakour and Eluru, 2013; Chen et al., 2013; Diab et al., 2020; Lee & Lee, 2013; Pasha et al., 2016; Taylor

et al., 2009; Thompson et al., 2012). Across sources, each variable's statistical significance and impact significantly differ between studies. For example, Boisjoly et al. (2018) found that a 1% increase in transit fares resulted in a 0.21% reduction in ridership, yet Taylor et al. (2009) found the same increase in fare price to result in a 0.43% reduction in trips taken. Diab et al. (2020) found that fare price was not statistically significant in their study. Studies of limited scope and size are often to blame for inconsistent results across transit research (Boisjoly et al., 2018; Diab et al., 2020; Durning & Townsend, 2015; Thompson & Brown, 2006).

This paper looks strictly at the impact of transit agencies' operating characteristics, the demographic traits of a given region, and their effects on transit ridership. Yet, it is important to note that lots of research has been done in other areas of North America about urban form and the impact it has on the demand for public transit (Chakour & Eluru, 2013; Foth et al., 2014; Lee & Lee, 2013; Pasha et al., 2016), which is not accounted for in this study due to time restrictions and limited data availability in Canada. An "urban form" that would be conducive to transit would reduce the demand for driving while making alternative forms of transportation more attractive through various urban planning policies that improve development density in areas near transit stops (Lee & Lee, 2013). Both Cao et al. (2009) and Ewing & Cevero (2001) found a link between carbon-efficient lifestyles and compact urban neighborhoods, including a decreased use of private vehicles. Even though most researchers find an association between these two variables, the magnitude of the urban form effect is unclear (Ewing & Cevero, 2001; Foth et al., 2014; Pasha et al., 2015). Critics of urban form policies state that appropriate pricing of transportation goods (such as cars, gas, and fare price) is more efficient and applicable in inducing consumers to change their mode of transit (Brueckner, 2007; Moore et al., 2010).

2.2 Determinants of Transit Supply

Across literature that studies transit supply, it is a common theme that researchers use either vehicle revenue kilometers or total vehicle revenue hours as their dependent variable. Taylor et al. (2009) and Storchmann (2001) utilize total vehicle revenue hours as their unit of measurement for transit supply, whereas Boisjoy et al. (2018), Lee and Lee (2013) and Diab et al. (2020) use revenue vehicle kilometers as their dependent variable in their respective regressions. Intuitively, it makes sense that vehicle revenue kilometers and vehicle hours are highly correlated. Diab et al. (2020) used Canadian transit data from 2002 to 2016 and found that the Pearson correlation coefficient between these two parameters is 0.995. Similarly, Taylor et al. (2009) studied cross-sectional transit data from 2000 and found that the correlation between these two variables was 0.95. Due to the high level of correlation, neither variable should be regressed on the other. When analyzing studies that use either total vehicle revenue hours or vehicle revenue kilometers, both dependent variables should be considered a near-equal measurement of transit supply. Including vehicle revenue hours (or a similar measure) in the regression is essential, as many researchers have found it to be the variable that can best explain the amount of ridership utilized within a transit agency (Boisjoly et al., 2018; Diab et al., 2020; Taylor et al., 2009).

In Canada, few studies have been conducted to analyze the determinants of public transit supply across all modes of provided transport. Many people have studied the determinants of ridership regarding bus travel or rapid light rail, but modes of public transit have yet to be looked at as a collective entity. Interestingly, prior research has shown that changes in bus vehicle kilometers significantly impact ridership (Boisjoly et al., 2018; Currie & Delbosc, 2011; Gong & Jin, 2014; Guerra & Cervero, 2011). Boisjoly et al. (2018) showed that the magnitude of the bus

vehicle revenue kilometers (RVK) coefficient is approximately five times greater than the RVK light rail coefficient. This indicates a closer association with buses and changes in ridership than light rail and ridership, which contradicts the belief that rail operations are sufficient to support growth in ridership. The reduction of bus vehicle kilometers in recent years is the most likely explanation for the slowed/reduced growth in ridership in major North American cities (Boisjoly et al., 2018).

Even though a transit agency's fleet size is an interesting parameter to include in a specification, it typically must be omitted due to the multicollinearity with vehicle revenue kilometers (Boisjoly et al., 2018). Vehicle revenue hours (or kilometers) is a more reliable statistic to use over long periods because it is the primary factor that characterizes the amount of service provided, and all transit agencies report it (CUTA, 2016). Sanauallah et al. (2021) did use fleet size as their primary variable characterizing transit supply. However, it is essential to note that their work was done on a case study in Belleville, ON over eight months. This study's short span and scope allowed Sanuallah et al. to carefully track fluctuations in Belleville's transit fleet and obtain results from their research. On a large scale, vehicle revenue kilometers are a more consistent and standardized measure of transit supply.

2.3 Simultaneous Relationship Between Demand and Supply of Public Transit and Instrumental Variables

It is essential to understand the many factors that influence public transit ridership correctly to inform policymakers with recommendations about ways to increase ridership. Most empirical studies looking at the determinants of transit ridership apply multiple linear regression analysis approaches to their data. These analyses do not account for the simultaneity between the

transit service provided and ridership (Diab et al., 2020; Taylor et al., 2009). Service providers respond to fluctuations in service consumption (ridership) levels by modifying the service supply, which will have a circular impact on transit use. If the model does not address this simultaneity, then coefficients could be inconsistent (Jung et al., 2016; Lee & Lee, 2013; Taylor et al., 2009).

The simultaneity in the data creates the problem of multicollinearity among the independent or explanatory variables. This usually occurs due to several relationships in the data – including transit operating characteristics, spatial variables, and demographic and spatial variables (Miller et al., 2018). Other studies have attempted to deal with this multicollinearity by reviewing correlation matrices and using logic and experience to eliminate problematic variables from their respective models. Currie and Delbosc (2011) contrasted the results from several regression specifications to analyze the impact multicollinearity played in their study. In one specification, they included many potentially relevant explanatory variables (even if they had a large correlation coefficient with each other), and in another, they only included explanatory variables with a correlation coefficient of less than 0.7. The results from these specifications varied, but they concluded that all models, a transit agencies' service level had a dominating impact on results.

To adjust for the endogeneity of transit demand and supply, instrumental variables (IVs) have been applied using the two-stage least square regression technique (Diab et al., 2020; Stochmann, 2001; Taylor et al., 2009). As the name implies, this regression is done in two stages. In the first, transit supply (vehicle revenue hours) is regressed on one or more instrumental variables. The instrumental variable(s) will be independent variables in the first stage of the regression and will be regressed on transit supply. It is important to note that the

instrumental variable should not be correlated with the error term from the regression in the second stage. The second stage of the regression will regress ridership on the predicted values of transit supply from the first stage and other independent variables associated with transit demand. Results from previous studies on transit demand using 2SLS have shown that transit supply is one of the main determinants of ridership, consistent with the findings of those using multiple linear regression analysis.

Those that have analyzed transit ridership using a two-stage least square regression have used different instrument variables to obtain their results. Diab et al. (2020) used transit supply (vehicle revenue hours) as the dependent variable in the first stage of the regression. They identified two factors directly impacting it – the operating budget and the population size. Diab et al. (2020) defended their choice of instruments by stating, "[t]he logic here is that greater vehicle revenue hours are expected to be delivered by transit agencies serving larger populations and enjoying higher operating budgets." Diab et al. (2020) estimated linked trips in the second stage of the regression by using the predicted value of vehicle revenue hours that they predicted in the first stage and other explanatory variables that impact ridership demand.

Taylor et al. (2009) noted that, on average, southern urbanized areas supply less transit than other parts of the United States and tend to be more conservative than their northern counterparts. They observed that regions such as upstate New York and Honolulu, Hawaii (both traditionally liberal regions) provided significantly more transit than the model predicted. In contrast, places like Montgomery, Alabama, offered drastically less transit service. Like Diab et al. (2020), Taylor et al. (2009) also created a predicted vehicle revenue hours variable. However, Taylor et al. (2009) used political orientation and population size as their instruments in the first stage of the regression. Both Taylor et al. (2009) and Diab et al. (2020) found the results from

the two-stage least squares regression to be still statistically significant, but the magnitude of the coefficients decreased compared to the ordinary least squares regression, after adjusting for the upward bias of the endogenous relationship.

3. Dataset Information

3.1 Data

The data used in this study come from two organizations: the Canadian Urban Transit Association (CUTA) and Statistics Canada. The data from CUTA includes forty years of data from 103 transit agencies from 1981 to 2016. In all models analyzed in this study**,** ridership will be the dependent variable as it is the best measure for the quantity of public transit demanded in each region. Ideally, this thesis hoped to compare the difference between unlinked and linked trips as a measure of ridership. However, CUTA only provides data concerning linked trips and does not provide the unlinked trips variable that most American transit agencies offer (Bosijoly et al., 2018; Taylor et al., 2009). Instead, Model I will use linked trips as the dependent variable, and Model II will utilize the variable "passenger kilometers per capita" as its respective dependent variable. The results of the two respective models will be compared to determine which factor is a more appropriate measure for ridership demand. Besides measures of ridership, all other variables taken from CUTA appear as independent variables in the regression. The exception would be Model III**,** which analyzes the results from a 2SLS regression (details will be described in subsequent sections). In this model, the variable vehicle revenue hours are used as the dependent variable in the first stage of the two-stage regression. The other variables analyzed from CUTA include net operating cost, fare, net capital cost, and service area size, which comprise the "supply side" information.

Information about factors describing socioeconomic and demographic regional differences comes from the census conducted by Statistics Canada. In this study, seven census years are included in the data set (1981, 1986, 1991, 1996, 2001, 2006, and 2016). The variables taken from the census will also be independent variables on the right-hand side of the regression.

The census variables included in the regression were chosen based on a combination of those found to be statistically significant in previous studies and those that seem logical to include. This methodology is consistent with most studies that utilize OLS or 2SLS to obtain their results (Diab et al., 2020; Durning & Townsend, 2015; Foth et al., 2014; Taylor et al., 2009).

The only independent variable included in some regression specifications not taken from census data or CUTA is that representing gas prices. Prior research has shown gas prices to have a mixed effect on public transit ridership (Boisjoly et al., 2018; Chakour & Eluru, 2013; Chen et al., 2011; Diab et al., 2020; Graehler et al., 2019; Jung et al., 2016; Lee & Lee, 2013). Data on Canadian gas prices was available through Statistics Canada for the years studied following 1991. Due to the unavailability of data from the first two years of the study, the variable for gas prices will not be included in any regression run on the entire data set. Statistics Canada only provided information about gas prices in a few cities within each province or territory. Because gas prices are near constant across a province or territory due to provincial regulation, average annual provincial gas prices were calculated and applied to each city within that province for a given year.

It is essential to mention that the transit agency service size and the census region surveyed do not have identical boundaries. Statistic Canada provides information on different regions in Canada and specifies an area as either a census subdivision, census division, census metropolitan area, or census agglomeration. The populations of these four classifications were cross-referenced with the transit agency service population size to ensure that the correct classification was chosen. For example, in 2016, Victoria, British Columbia, was listed as a census subdivision and a census metropolitan area. Statistics Canada (2016) listed the populations for both regions as 85,792 and 367,770, respectively. The transit agency responsible

for serving Victoria, BC, listed its service area population as 367,770: therefore, the census metropolitan area classification was deemed the proper measure for Victoria's census data.

Table 1 (below) summarizes the independent and dependent variables used in the different regression analyses included in this study. Table 2 provides a legend of the variables used in this study and their abbreviations.

Table 1: Summary of dependent and independent variables used

	(1)				
	mean	sd	count	max	min
tripspc	45.95976	78.24472	393	1318.747	.1365467
pkmpc	451.9164	545.0752	121	2687.568	.2449016
Pop	284955.6	571887.9	482	4098927	1387
Popimm	71667.97	181171	482	1266005	0
Particm	78.21746	43.2529	481	765	52.6
Unempm	11.56881	78.13158	481	1720	2
Particf	62.31331	41.67969	481	960	33
Unempf	15.9237	164.7024	481	3620	2
medincf2	76906.84	18343.92	482	214357	401.2506
avgmort2	1176.767	287.0308	482	2696	0
density	1271.126	982.4643	261	5657	.7522716
rvhpc	1.427473	1.968469	384	35.58158	.0563675
rentrate	.3416977	.1078057	340	.7057903	.0770197
foreign	.1581504	.1331634	482	1.90587	0
opertotald2	7.41e+07	1.98e+08	387	1.66e+09	43528.91
opercostnet2	3.99e+07	1.00e+08	393	7.99e+08	26580.41
fare2	2.548323	.7876187	297	6	1
gasprice2	93.06449	14.18272	347	128.8705	74.67135
Lowincfam	13.3049	13.79454	469	295	4
costpt	2.601867	2.775134	392	23.9968	.2451276
year== 1981.0000	.1428571	.3502608	525	1	0
year== 1986.0000	.1428571	.3502608	525	1	0
year== 1991.0000	.1428571	.3502608	525	1	0
year== 1996.0000	.1428571	.3502608	525	1	0
year== 2001.0000	.1428571	.3502608	525	1	0
year== 2006.0000	.1428571	.3502608	525	1	0
year== 2016.0000	.1428571	.3502608	525	1	0

Table 2: Legend for variables used and their abbreviations

Linked Trips per Capita	tripspc	Unlinked Trips per Capita
Passenger Revenue Kilometers per Capita	pkmpclog	Passenger Revenue Kilometers per Capita
Total Population by Sex and Age Groups	Pop	Total population of municipality
Total Population of Immigrants	Popimm	Total population of immigrants in municipality
Participation rate - Males 15 years and over	Particm	Participation rate - Males 15 years and over
Unemployment rate - Males 15 years and over	Unempm	Unemployment rate - Males 15 years and over
Participation rate - Females 15 years and over	Particf	Participation rate - Females 15 years and over
Unemployment rate - Females 15 years and over	Unempf	Unemployment rate - Females 15 years and over
Median income ($)	Medincf	Median family income
Average major payments for owners (monthly) ($)	Avgmort	Average mortgage payment, nominal $
Population Density	Density	Population Density
Revenue Vehicle Hours per Capita (transit supply parameter)	rvhpc	Revenue Vehicle Hours per Capita
Percentage of Homes Rented (rent/dwell)	Rentrate	% of Homes Rented
Percentage of Populations that are Immigrants	Foreign	% of the population that are immigrants
Total Direct Operating Expense	Opertotald	Direct operating costs in total, nominal $
Net Operating Cost	Opercostnet	Net operating cost, nominal $
Adult Cash Fare	Fare	Fare per trip, nominal $
Gas Price	Gasprice	Average annual provincial gas price, nominal $
Incidence of low income (%) - Number of Economic Families	Lowincfam	% of low-income families
Cost per trip (net operating cost/trips)	costpt	Cost per trip

3.2 Data Limitations

Unfortunately, this data set comes with limitations that restrict the extent of this study. This research aimed to acquire a more comprehensive data set that could include forty years of

transit and census data. Most previous studies have focused on public transit trends in either a specific city over a long period (Chakour & Eluru, 2013; Chiang et al., 2011; Collins & MacFarlane, 2018; Foth et al., 2014; Kain & Liu, 1999) or they have limited their focus to a small interval of time while examining ridership trends across one (or multiple) countries (Chen et al., 2011; Graehler et al., 2019; Pasha et al., 2016).

After observing the data, it is evident that measures from the first few years of study (1981, 1986, and 1991) lack information from relevant variables included in the regression specifications. Particularly, CUTA began its data collection in 1980 and only provided information about thirty-four measures of performance, whereas, in 2016, they offered a complete data set of ninety-one variables per transit agency. Variables such as fare price and regular service passenger trips (linked trips) did not appear in the CUTA data until 1996. If these variables are included in the regression specification that spans the forty years of data, they force the first three years of interest to be dropped from the model. Both fare and gas prices are relevant variables that would be ideally included in the model but cannot be if one runs a regression on the collected data set. Generally, the statistical software drops most of the observations (217) from the first few years due to missing information. Additionally, note that 2011 was omitted from the final data set because Statistics Canada canceled that polling year's long-form census, which resulted in a significant loss of information.

An advantage of this study is that it looks over a long period, which means there is a large selection of independent variables to consider for regression specifications. However, factors describing operating characteristics from CUTA and variables from Statistics Canada that illustrate demographic characteristics in a region are often correlated. If two variables depict a high level of correlation with each other, one of those variables will have to be omitted from the

specification. This reason is that a high degree of correlation between two independent variables will create a large variance in the parameter estimates, making coefficients from the regression challenging to interpret due to large confidence intervals.

The final noteworthy limitation of this study is that it takes observations at five-year intervals. Because the census is conducted every five years, data from CUTA is also taken at five-year intervals, even though CUTA provides operating statistics yearly. For example, suppose that a transit agency decides to put forth a large amount of capital investment into increasing its stock of vehicles in a non-census year. The impacts of that investment will show in subsequent years, and the next census year included in the data set could illustrate a significant increase in ridership. The regression would attribute this increase in ridership to other factors and not account for the capital investment missed between polling years. The inability to utilize CUTA data between polling years means a high likelihood of missing the correct quantity of investment of a transit agency across the forty years of study.

4. Methodology

4.1 Model Development

This thesis aims to analyze three specifications. The first model will attempt to explain transit demand by using linked trips as the dependent variable while including as many observations as possible to maximize the benefit of having a data set that contains observations over a forty-year period. The second model will drop the observations before 1996 to utilize passenger revenue kilometers per capita as the dependent variable and compare the results with the Model I specification. CUTA only began recording the variable "passenger revenue kilometers" in 1996, which resulted in the specification dropping all observations before this point in time. The first two models will use multiple linear regression analysis to explain changes in ridership. Finally, the third model will use a two-stage least squares regression to compensate for the simultaneous relationship between the supply and demand of public transit. Similar to the models created by Diab et al. (2020) and Storchmann (2001), the first stage of the regression will use population and operating budget as the instruments to predict vehicle revenue hours (the "supply" parameter) that will be used in the second stage of the regression. As discussed in previous sections, vehicle revenue hours and vehicle revenue kilometers are near equal measures for the transit supply in each region. The choice was made to use vehicle revenue hours because it had more observations in the data set.

Like research conducted by Diab et al. (2020), Kain and Liu (1999) and Taylor et al. (2009), this thesis uses natural logarithms to transform independent and dependent variables in the regression equation. Applying natural logarithms compensates for the skewness in the distributions of certain parameters in the model. Additionally, some variables, such as linked trips, passenger revenue kilometers, and vehicle revenue hours, were altered to create a new

variable measured per capita. In doing so, some multicollinearity issues in the model were eliminated.

Furthermore, all variables with a dollar measure (operating expenses, fares, and gas prices) were converted to be in terms of 2016 dollars using the Consumer Price Index for each year of study.

4.2 Model I Specification

This first model utilizes multiple regression analysis with linked trips per capita as the dependent variable. Linked trips are regressed on many socioeconomic variables, revenue vehicle hours, fare price, and gas prices. Three different specifications of this type will be run to show the limitations of this data set (particularly for the early years of the study). The general form for Model I is shown in Equation 1 below:

$$\log(linked\ trips\ per\ capita) = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_n X_{n,i} + \varepsilon_i \quad (1)$$

X_i denotes the explanatory variables in the specification where "n" independent variables are included. According to Equation 1, β_0 is the intercept, and β_1 represents the regression coefficient for the first explanatory variable. This structural form is used for all three specifications of Model I, with the only difference being the number of explanatory variables and dummy variables included. In the first regression, the primary goal is to have as many observations as possible to take advantage of the large data set. To accomplish this, several potentially relevant variables will be omitted as they do not have data containing observations from 1981 and 1986. In the regression software used for this analysis, if an observation is missing a data entry for any explanatory variable in the regression specification, the observation will be dropped entirely for a transit agency that year. The second regression of this type

includes other relevant variables but will contain a much smaller sample size due to missing observations. Furthermore, the third regression analyzes the difference in results when the dummy variables controlling for the year of study are omitted from the specification. Interestingly, many prior studies that have looked at the determinants of public transit do not use dummy variables to capture time-related effects (i.e., stock market crash or depression) that the model does not already account for (Diab et al., 2020; Storchmann, 2001; Taylor et al., 2009). The third regression gives insight into how the magnitude and significance of the regression results differ when dummy variables controlling for specific years are excluded from the regression.

4.3 Model II Specification

The primary difference between the first and second models is that the second specification uses passenger vehicle kilometers per capita as the dependent model in the regression equation. Model 2 still utilizes ordinary least squares as the regression technique to achieve results. As noted previously, linked trips are considered to be a reasonably reliable estimate of ridership demand (Diab et al., 2020; Taylor et al., 2009). Even though the consensus is that linked trips provide a more reliable measure than unlinked trips, few studies have been conducted that utilize a different variable to measure ridership demand. Model II will follow the work of Thompson et al. (2012) and use passenger kilometers per capita as the dependent variable and contrast the results from this specification with results from Model I. The interpretation of the coefficients for this model will be identical to Model I because the dependent variable is the only difference between the two specifications. Below is the regression equation for the second model:

$$\log(passenger\ revenue\ kilometers\ per\ capita) = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_n X_{n,i} + \varepsilon_i \quad (2)$$

CUTA began collecting data on linked trips in 1996, so 1996 will be this model's first year of study. A motivating factor for creating this specification is to investigate the impact that explanatory variables have on different dependent variables. Similar results would indicate that conclusions that are drawn from respective regressions are reliable.

4.4 Model III Specification

The third model specification employs the two-stage least squares (2SLS) regression technique to achieve its results. Like research conducted by Diab et al. (2020), this model will create a predicted revenue vehicle hours variable using net operating cost and population in the first stage of the 2SLS regression. The equations for both stages are included below:

First stage:

$$\log(rvh) = \beta_0 + \beta_1 \log(pop) + \beta_2 \log(nominal\ operating\ cost) + \varepsilon_i \quad (3)$$

Second stage:

$$\log(linked\ trips\ per\ capita) = \beta_0 + \beta_1 \widehat{revenue\ vehicle\ hours} + .. + \beta_n X_{n,i} + \varepsilon_i \quad (4)$$

The rationale behind using population and operating cost as instruments is that transit agencies with higher revenue vehicle kilometers are most likely to serve larger populations and have more expenses (hence loftier operating costs). In the second stage of the 2SLS regression, fluctuations in the population and operating budget are incorporated through the predicted revenue vehicle hours and are not explicitly included as explanatory variables in the second stage regression.

5. Results

5.1 Model I Results

This portion of the thesis presents and analyzes the results obtained from running an ordinary least squares regression on the collected data. As described in equation (1), Model 1 utilizes the linked trips per capita variable as the dependent variable. A logarithmic function form accounts for the distributional skewness in the data. Additionally, using logarithms on both sides of the regression allows the results to be interpreted in elasticities. The regression coefficients on the right-hand side of the equation explain the impact of an explanatory variable's one percent increase on the percent of unlinked trips per capita.

The results from the three specifications in equation (1) are displayed below. The asterisk next to the coefficients indicates the statistical significance of each variable included in the respective regression. One asterisk means a coefficient is significant at a 5% significance level; two asterisks denote that a coefficient is significant at a 1% significance level, and three asterisks mean that a coefficient is significant at a 0.1% level of significance.

Table 3: Model I Results

	(1) tripspclog	(2) tripspclog	(3) tripspclog
rvhpclog	**1.070*** **	**1.191*** **	**1.197*** **
	(25.98)	**(17.70)**	**(18.13)**
poplog	**0.167*** **		
	(7.67)		
particm	**−0.000187**	**0.000222***	**0.000233****
	(−1.79)	**(2.45)**	**(2.74)**
unempm	**−0.00185**	**0.00139**	**0.00467**
	(−0.20)	**(0.09)**	**(0.33)**
particf	**0.0112****	**0.00438**	**0.00438**
	(2.63)	**(0.98)**	**(0.98)**
unempf	**0.0106**	**−0.00958**	**−0.0119**
	(0.85)	**(−0.44)**	**(−0.56)**
avgmort2log	**−0.267***	**−0.342**	**−0.324**
	(−2.12)	**(−1.75)**	**(−1.80)**
lowincfam	**0.0126***	**−0.0130***	**−0.0108**
	(2.54)	**(−2.07)**	**(−1.71)**
costpt	**−0.0783*** **	**−0.0794*** **	**−0.0786*** **
	(−6.58)	**(−7.46)**	**(−7.42)**
1981.year	**0**		
	(.)		
1986.year	**−0.0835**		
	(−1.45)		
1991.year	**−0.156***		
	(−2.38)		
1996.year	**−0.286*** **		
	(−4.67)		
2001.year	**−0.251*** **	**0**	
	(−4.26)	**(.)**	
2006.year	**−0.215*** **	**−0.0672**	
	(−3.81)	**(−0.69)**	
2016.year	**−0.235*** **	**−0.0128**	
	(−3.54)	**(−0.17)**	
popimmlog		**0.107*** **	**0.104*** **
		(6.09)	**(6.11)**
rentrate		**1.502*** **	**1.473*** **
		(3.66)	**(3.48)**
densitylog		**−0.0467***	**−0.0465***
		(−2.16)	**(−2.12)**
fare2log		**0.0927**	**0.102**
		(0.67)	**(0.75)**
gasprice2log		**0.320**	**0.138**
		(1.06)	**(0.82)**
_cons	**2.639*** **	**2.978**	**3.641****
	(3.66)	**(1.53)**	**(2.68)**
N	**365**	**148**	**148**
R-sq	**0.939**	**0.963**	**0.963**

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

The primary motivation of the first specification is to include as many observations as possible. Out of the 482 entries in the data set, the first regression captures 365 observations. This sample size is sufficient for reliable results. To achieve this large sample size, potentially relevant variables such as density, percent of homes rented, fare price, and gas price were omitted because they needed to contain data from the early years of the study. Even though these variables were dropped, the r-squared value shows that 93.9 percent of the variation in linked trips per capita can be explained using the explanatory variables included in the model.

Unsurprisingly, the explanatory variable for revenue vehicle hours per capita was statistically significant at a 0.1% significance level. A 10% increase in revenue vehicle hours per capita would be associated with a 10.7% increase in linked trips per capita, with everything else remaining constant. This is consistent with Boisjoly et al. (2018), but the magnitude of the revenue vehicle hours coefficient is larger than what Taylor et al. (2009) and Diab et al. (2020) found. Another logical result from the regression came from the explanatory variable denoting population. A 1% increase in the population is estimated to have a 0.167% increase in trips per capita, which was found to be significant at a 0.1% significance level.

More interestingly, in the first specification, the female labor force participation rate is statistically significant at a 1% significance level. Tyrinopoulos & Antoniou (2013) concluded from their research that working females preferred taking transit compared to their male counterparts. The first specification confirms this finding by illustrating that a 1% increase in the participation rate for females has an estimated 1.12% increase in linked trips.

Average mortgage and incidence of low-income families were also statistically significant at a 5% significance level. A 1% increase in the incidence of low-income families

would increase linked trips per capita by 1.26%, consistent with research conducted by Durning and Townsend (2015).

Variables illustrating the male labor force participation rate, male unemployment rate, and female unemployment rate were insignificant at a 5% significance level. The results from the first specification indicate some similarity with previous studies; however, the magnitude and significance of some variables contradict what other prior research has shown (Boisjoly et al., 2018; Diab et al., 2020; Taylor and Fink, 2013; Taylor et al., 2009; Tyrinopoulos and Antoniou, 2013). This could be attributed to the decision to omit potentially relevant variables to capture a large sample size.

The second specification for Model 1 includes more explanatory variables to explain more of the variation in the dependent variable. The r-squared value increases from 93.9 to 96.3 between the first and second specifications. However, the impact of adding more variables to a data set that is not complete is evident – the sample size decreases to 148. The second specification utilizes population density instead of the population as a measure of the compactness of a city; Lee and Lee (2013) conducted their research using this parameter as well. The second specification includes fare price, immigrant population, home rental rate, and the average provincial gas price. The immigrant population explanatory variable was excluded from the first specification due to the high correlation (0.9471) with population size. Interestingly, the correlation between density and immigrant population is much smaller (0.4883), which is small enough not to violate econometric assumptions.

Like the first specification, revenue vehicle hours per capita and the incidence of low-income families were statistically significant. The magnitude of the coefficient associated with revenue vehicle hours per capita (rvhpc) increased and signified that a 10% increase in rvhpc

would increase ridership by 11.91%, on average. Even though the coefficient for the incidence of low-income families stayed statistically significant, the sign changed from positive to negative, indicating that an increase in the percentage of low-income families will cause a reduction in linked trips. The reason for this is unclear, yet logically it makes sense that the coefficient for low-income families should be positive. The first specification allows for a better interpretation of the coefficient for low-income families and shows that a 1% increase in the incidence of low-income would increase ridership by 1.26%, on average.

Of the explanatory variables added to the second specification, immigrant population, percentage of homes rented, and population density were all significant at a 5% significance level. The most exciting result coming out of the second specification was that both fare price and average provincial gas price were not found to be statistically significant, contradicting the work of Boisjoly et al., 2018; Diab et al., 2020; Guerra and Cervero, 2011; and Lee and Lee, 2013. Further research should be done to ensure the validity of these results, but if they hold, this finding could have significant policy implications. According to the second specification, we cannot conclude that gas and fare price fluctuations have a statistically significant impact on ridership.

5.2 Model II Results

The second model uses passenger revenue kilometers per capita and runs similar specifications as the first model, with the only difference being the dependent variable. For easy comparison, the second specification in Model I will be included next to the first specification in Model 2 specification. Because Model II will only be using data after 1996, there will not be a specification that aims to have as many observations as possible since the first three years of

collected data are already dropped from the regression. Additionally, to ensure that both models utilize the same information, the statistical software dropped all observations before 1996 so the first and second specifications cover the same time. The first specification in Model II will consist of the same explanatory variables as the second specification.

Table 4: Model II Results

	(1) pkmpclog	(2) tripspclog
rvhpclog	0.627**	1.126***
	(3.44)	(19.78)
particm	0.0914	0.000160
	(1.69)	(1.92)
unempm	0.0130	-0.0135
	(0.19)	(-0.89)
particf	-0.122	0.00366
	(-1.90)	(0.88)
unempf	-0.0733	0.0165
	(-0.77)	(0.77)
avgmort2log	0.247	-0.640***
	(0.29)	(-3.99)
medinc2log	4.035***	1.200***
	(3.79)	(4.05)
popimmlog	0.227***	0.111***
	(4.57)	(7.12)
lowincfam	0.0593	0.00982
	(1.71)	(1.17)
rentrate	0.167	1.535***
	(0.16)	(4.84)
densitylog	-0.157	-0.0473*
	(-1.80)	(-2.36)
fare2log	-0.739*	0.141
	(-2.41)	(1.43)
gasprice2log	4.961**	0.746**
	(3.08)	(2.63)
costpt	-0.145***	-0.0881***
	(-5.47)	(-9.49)
2001.year	0	0
	(.)	(.)
2006.year	-1.127**	-0.184*
	(-2.95)	(-1.98)
2016.year	-0.981**	-0.236**
	(-2.82)	(-2.92)
_cons	-61.29***	-9.598**
	(-4.16)	(-2.69)
N	71	148
R-sq	0.870	0.969

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Table 4 contains the results from two regressions that explains the effects of using different dependent variables as a measure of ridership. Preliminary observation of the table leads to the conclusion that all statistically significant explanatory variables have the same sign for their respective coefficients, which is a positive indicator that both linked trips and passenger revenue vehicle kilometers are a similar measure of ridership demand. Using the 5% significance level, the conclusion can be drawn that regression (1) contains seven statistically significant explanatory variables, and regression (2) contains eight significant independent variables. Additionally, the r-squared value explains that the explanatory variables that are shown in the first specification explain 87.0% of the variation in passenger revenue kilometers per capita. The second specification that uses linked trips per capita as the dependent variable shows that the independent variables explain 96.9% of the variation.

At a 0.1% significance level, revenue vehicle hours per capita is statistically significant in both specifications, with the magnitude of the coefficient doubling in the second specification. The first specification illustrates that a 1% increase in revenue vehicle hours would increase passenger vehicle kilometers per capita by 0.627%, whereas the second specification shows that a 1% increase in vehicle revenue hours per capita would increase linked trips by 1.126%. The result of the magnitude of specification (1) coefficients being smaller than specification (2) coefficients does not hold true for all statistically significant explanatory variables. For example, both regressions (1) and (2) reveal that at a 0.01% level of significance, median family income is significant in determining ridership demanded. However, in the first specification, a 1% increase in median family income is associated with a 4.035% increase in passenger revenue kilometers per capita, but specification (2) shows that a 1% increase in median family income is associated with a 1.2% increase in linked trips per capita. The intuition for this result would be that, on

average, people in larger cities earn more money and are also more likely to utilize public transportation.

Like the conclusions drawn from Model I, the percentage of immigrants in a population was found to be statistically significant at a 0.01% significance level. Regression (1) that utilizes passenger revenue kilometers per capita as the dependent variable displayed a coefficient double the magnitude of regression (2), 0.227 compared to 0.111, respectively.

Interestingly, Model II, specification (1), was the only regression analyzed in this thesis that showed fare price to be statistically significant at a 5% significance level. This specification disclosed that a 1% increase in fare price would decrease passenger revenue kilometers per capita by 0.739%, on average, with everything else remaining the same. Logically, this result makes sense that as ticket fare price increases, some people will forgo traveling by public transportation and instead turn to a private vehicle. This result is consistent with the findings of (Boisjoly et al., 2018; Chen et al., 2011; Kain & Liu, 1999; Taylor et al., 2009).

Even though specification (1) provides results that make more sense regarding fare price, it is important to note that the magnitude of some of the coefficients seems to largely surpass what other studies have found and are frankly unrealistic. This indicates that the parameter "passenger revenue kilometers per capita" may not be the best measure for ridership demand. For example, regression (1) displays the result that a 1% increase in gas prices would result in a 4.961% increase in the dependent variable explaining ridership. Whereas regression (2) illustrates a more realistic conclusion that a 1% increase in gas prices would result in a 0.746% increase in linked trips per capita. A similar result arises from the median family income explanatory variable where the coefficients for specification (1) and (2), though both statistically significant, are 4.035 and 1.200, respectively. The conclusion can be drawn from this analysis of

Model II that linked trips per capita is a better indicator of transit demand ridership than passenger revenue kilometers per capita.

5.3 Model III Results

The results contained in this section include four specifications: the output from the first and third specifications are obtained through a two-stage least squares regression, while the second and fourth specifications are OLS regressions and are included for comparison. Like Model 1, the first regression aims to have as many observations as possible (n = 266). The third specification adds more explanatory variables, such as gas and fare prices, creating a more useful regression equation. Additionally, the output from the first stage regression is included in how much variation in revenue vehicle hours can be explained by population and net operating cost, which partially indicates the validity of the instruments used.

Table 5: Model III Results

	(1) tripspclog	(2) tripspclog	(3) tripspclog	(4) tripspclog
rvhlog	0.451***	1.016***	0.900***	1.130***
	(12.66)	(20.98)	(10.82)	(24.98)
particm	−0.000419*	−0.000110	−0.000544***	−0.0000975
	(−2.17)	(−1.16)	(−5.00)	(−1.25)
unempm	−0.0236	0.00263	−0.0761***	−0.0166
	(−1.04)	(0.25)	(−4.00)	(−1.35)
particf	0.0195*	0.00970*	−0.00137	0.00202
	(2.35)	(2.42)	(−0.21)	(0.48)
unempf	0.0159	0.0115	0.0464	0.00945
	(0.58)	(0.87)	(1.82)	(0.61)
avgmort2log	−0.947***	−0.299**	−0.0255	−0.0272
	(−4.39)	(−2.66)	(−0.10)	(−0.20)
lowincfam	−0.0161	−0.0125*	−0.00262	−0.00851
	(−1.19)	(−2.03)	(−0.30)	(−1.58)
rentrate	3.168***	1.472***	2.216***	1.366***
	(4.85)	(5.14)	(4.23)	(5.10)
1981.year	0	0	0	
	(.)	(.)	(.)	
1986.year	0		0	
	(.)		(.)	
1991.year	0		0	
	(.)		(.)	
1996.year	−0.256	−0.196**	−0.147	0
	(−1.65)	(−2.84)	(−1.37)	(.)
2001.year	−0.289*	−0.173**	−0.166	−0.0148
	(−2.10)	(−2.68)	(−1.69)	(−0.27)
2006.year	−0.331*	−0.125*	0.133	0.0592
	(−2.48)	(−2.10)	(1.32)	(0.58)
2016.year	−0.292	−0.0891	0	0.0571
	(−1.71)	(−1.30)	(.)	(0.73)
poplog		−0.833***		−0.792***
		(−13.30)		(−12.25)
costpt		−0.0849***		−0.0744***
		(−7.74)		(−7.69)
popimmlog			−0.462***	−0.142***
			(−6.12)	(−3.48)
fare2log			−0.0130	0.00942
			(−0.10)	(0.13)
gasprice2log			−1.198**	−0.216
			(−2.85)	(−0.75)
_cons	2.870*	2.484***	2.346	1.492
	(2.19)	(3.56)	(1.13)	(1.03)
N	266	266	216	216
R-sq	0.772	0.952	0.892	0.959

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Table 6: Output from the First Stage of the 2SLS regression

Source	SS	df	MS	Number of obs	=	377
				F(2, 374)	=	3040.68
Model	1018.15154	2	509.07577	Prob > F	=	0.0000
Residual	62.6156735	374	.167421587	R-squared	=	0.9421
				Adj R-squared	=	0.9418
Total	1080.76721	376	2.87438089	Root MSE	=	.40917

rvhlog	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
poplog	.3223468	.0414816	7.77	0.000	.2407804	.4039131
opercostnet2log	.677608	.0287976	23.53	0.000	.6209825	.7342335
_cons	-2.624838	.1935572	-13.56	0.000	-3.005435	-2.244241

.

Table 6 displays the output from the first stage of the two-stage least squares regression. In this stage, the variable revenue vehicle hours is regressed on population and net operating cost. Logarithms have been applied to all three variables. As seen from the table, both population and net operating costs are statistically significant and have large t-statistics of 7.77 and 23.53, respectively. The r-squared value from this regression is 0.9421, which means that the two explanatory variables can explain 94.21% of the variation in revenue vehicle hours. Additionally, the F-statistic for the first stage is 3040.68 which far surpasses the “rule of thumb” that an F-statistic greater than ten indicates a good instrument.

As seen from the first and third specifications, the 2SLS regression decreases the r-squared value from 0.952 to 0.772 and from 0.959 to 0.892, respectively. This is an indicator that, as hypothesized, there is endogeneity present in the model and that a 2SLS regression is needed to compensate for this issue.

In both the first and third specifications, revenue vehicle hours remain statistically significant, but the magnitude of the coefficient greatly decreases when compared to their OLS

counterparts (2) and (4). Taylor et al. (2009) also found this when they used political orientation as an instrument to predict revenue vehicle hours. The first specification displays that a 10% increase in transit supply would increase ridership by 4.5%, whereas (3) shows that ridership would increase by 9%; both specifications showed that revenue vehicle hours were statistically significant at a 0.1% level of significance. The specifications that used 2SLS both illustrate that the male labor force participation rate was statistically significant and indicate that an increase in male participation in the workforce would decrease linked trips per capita. Specification (3), which includes more explanatory variables, also shows that the male unemployment rate is statistically significant at a significance level of 0.1. Additionally, the first specification showed that the average mortgage was significant at a 0.1 significance level. The regression showed that a 1% increase in the average mortgage would decrease transit ridership by 0.947% on average, with everything else remaining the same.

The most significant difference between the OLS and 2SLS results is that gas price becomes statistically significant, which makes logical sense. Yet, the ordinary least squares regression suggested that it was irrelevant. Intuitively, an increase in gas price would translate to an increase in transit ridership because the opportunity cost of driving a car increase with the gas price hike. However, the regression results indicate that a 1% increase in gas price would decrease linked trips per capita by 1.198%. Knowing this is not a logical conclusion indicates that the instrumental variables used for this model do not suffice to compensate for the endogeneity present in this data. Two hypothesized reasons for these confusing results could be that gas prices are correlated with net operating cost, which would invalidate the use of net operating cost as an instrument. The other reason could be that this data set annualizes and takes a provincial average of gas prices and does not capture price fluctuations monthly. Studies

conducted in the United States, which had better data availability, illustrated that gas prices were a primary factor in determining transit demand (Boisjoly et al., 2018; Diab et al., 2020; Lee & Lee, 2013; Graehler et al., 2019).

6. Conclusion and Policy Implications

This thesis examines different regression techniques that best capture what factors characterize ridership demand. Model (1) follows the standard from other transportation studies by regressing linked trips per capita on many explanatory variables that make up the demographic traits of a population. Model (2) continues by analyzing a less studied dependent variable, passenger revenue kilometers per capita. The results from Model (1) and Model (2) were examined, and the conclusion was drawn that linked trips per capita is a better variable to measure ridership demand. For this specific study, using linked trips per capita allows for an extensive data set and does not overestimate the magnitude of coefficients, as seen in Model (2). Finally, Model (3) attempts to compensate for the simultaneity of transit supply and demand using two-stage least squares regression analysis. The results from this specification indicate that a 2SLS regression is the appropriate technique when exploring the determinants of the demand for public transportation.

Regarding policy implications drawn from this study, revenue vehicle hours per capita, population, and percentage of immigrants in a population were statistically significant across all models. This suggests that investment will be maximized by targeting areas with a large population and a large share of immigrants, which is a reasonable conclusion. Regarding Trudeau's plan to invest fifteen billion into public transportation to meet environmental goals, the findings of this paper suggest that he should focus his investment on metropolitan areas in Canada. Moreover, the prevalent theme across all specifications was that revenue vehicle hours remained statistically significant at a 0.01% significance level, illustrating that the supply of public transportation provided to a region is the greatest factor in determining the level of ridership. Further research could be conducted to analyze the impact and differences of

expanding revenue vehicle hours in urban regions compared to rural regions. To accomplish this, interaction terms should be used in the regression specification to determine the demand elasticity with respect to transit supply (measured in revenue vehicle hours) between rural and urban regions.

The data studied in this thesis presented its limitations, and there is room for more growth and research in this field. Specifically, different instrumental variables should be examined to amplify the use of 2SLS regression analysis. The argument could be made that some of the instrumental variables used in (3) may correlate with the error term from the second stage of the 2SLS regression. As suggested by Taylor et al. (2009) and Beaudoin et al. (2015), voting data could be a valid instrument for two-stage least squares regression analysis and should be investigated further. Additionally, contrasting results from a pooled time-series cross-sectional analysis with OLS and 2SLS regressions could help compensate for the endogeneity in transportation data.

References

Beaudoin, Justin, Farzin, Yeganeh, & Lawell, Cynthia. (2015). Public transit investment and sustainable transportation: A review of studies on transit's impact on traffic congestion and air quality. *Research in Transportation Economics, 52,* 15-22.

Boisjoly, Geneviève, Grisé, Emily, Maguire, Meadhbh, Veillette, Marie-Pier, Deboosere, Robbin, Berrebi, Emma, & El-Geneidy, Ahmed. (2018). Invest in the ride: A 14-year longitudinal analysis of the determinants of public transport ridership in 25 North American cities. *Transportation Research. Part A, Policy and Practice, 116,* 434- 445.

Brueckner, Jan. (2007). Urban growth boundaries: An effective second-best remedy for unpriced traffic congestion? *Journal of Housing Economics, 16*(3–4), 263–273.

Cao, X. Y., Mokhtarian, P. L., & Handy, S. L. (2009). Examining the impacts of residential self -selection on travel behaviour: A focus on empirical findings. *Transport Reviews, 29(3),* 359–395.

Cervero, R. (2003). Growing smart by linking transportation and land use: Perspectives from California. *Built Environment, 39(1),* 66–78.

Chakour, Vincent, & Eluru, Naveen. (2013). Examining the Influence of Urban form and Land Use on Bus Ridership in Montreal. *Procedia – Social and Behavioral Sciences, 104,* 875-884.

Chen, Cynthia, Don Varley, & Jason Chen. (2011). What Affects Transit Ridership? A Dynamic Analysis involving Multiple Factors, Lags and Asymmetric Behaviour. *Urban Studies (Edinburgh, Scotland), 48(9),* 1893-1908.

Chiang, Wen-Chyuan, Russel, Robert, & Urban, Timothy. (2011). Forecasting ridership for a metropolitan transit authority. *Transportation Research Part A: Policy and Practice, 45(7),* 696-705.

Collins, Patricia & MacFarlane, Robert. (2018). Evaluating the determinants of switching to public transit in an automobile-oriented mid-sized Canadian city: A longitudinal analysis. *Transportation Research Part A: Policy and Practice, 118,* 682-695.

Cui, Boer, Boisjoly, Geneviève, Miranda-Moreno, Luis, & El-Geneidy, Amhed. (2020). Accessibility matters: Exploring the determinants of public transport mode share across income groups in Canadian cities

CUTA. (2016). Canadian Transit Fact Book.

Currie, G., & Delbosc, A. (2011). Understanding bus rapid transit route ridership drivers: An empirical study of Australian BRT systems. *Transport Policy, 18(5),* 755-764.

Diab, Ehab, Kasraian, Dena, Miller, Eric J, & Shalaby, Amer. (2020). The rise and fall of transit ridership across Canada: Understanding the determinants. *Transport Policy, 96*, 101-112.

Durning, Matthew, & Townsend, Craig. (2015). Direct Ridership Model of Rail Rapid Transit Systems in Canada. *Transportation Research Record, 2537(1)*, 96-102.

Foth, Nicole, Manaugh, Kevin, & El-Geneidy, Ahmen. (2014). Determinants of Mode Share over Time: How Changing Transport System Affect Transit Use in Toronto, Ontario, Canada. *Transportation Research Record, 2417(1),* 67-77.

Gong, Huawei & Jin, Wenzhou. (2014). "Analysis of urban public transit pricing adjustment program evaluation based on trilateral game." *Procedia - Social and Behavioral Sciences, Vol. 138*, 332- 339.

Graehler, Michael, Mucci, Alex & Erhardt, Gregory. (2019). Understanding the Recent Transit Ridership Decline in Major US Cities: Service Cuts or Emerging Modes?

Guerra, Erick, & Cervero, Robert. (2011). Cost of a Ride. *Journal of the American Planning Association, 77(3)*, 267-290.

Kain, John F, & Liu, Zvi. (1999). Secrets of success: Assessing the large increases in transit ridership achieved by Houston and San Diego transit providers. *Transportation Research. Part A, Policy and Practice*, 33(7), 601-624.

Lee, Bumsoo & Lee, Youngsung. (2013). Complementary Pricing and Land Use Policies: Does It Lead to Higher Transit Use? *Journal of the American Planning Association, 79(4),* 314-328.

Jiao, Liudan, Luo, Fenglian, Wu, Fengyan, Zhanng, Yu, Huo, Xiaosen & Wu, Ya. (2022). Can the opening of urban rail transit improve urban air quality? Evidence from 94 lines in China. *Environmental Impact Assessment Review, 97*.

Jones, R. P. (2021, February 10). Trudeau pledges billions in permanent funding for public transit. CBC News. Retrieved from https://www.cbc.ca/news/politics/trudeau-transit fund-1.5908346

Jung, Hojin, Yu, Gun, & Kwon, Kyoung-Min. (2016). Investigating the Effect of Gasoline Prices on Transit Ridership and Unobserved Heterogeneity. *Journal of Public Transportation, 19(4),* 56-74.

Miller, Eric, Shalaby, Amer, Diab, Ehab, & Kasraian, Dena. (2018). Canadian Transit Ridership Trends Study. *Canadian Urban Transit Association (CUTA)*.

Moore, Adrian, Staley, Samuel, & Poole, Robert. (2010). The role of VMT reduction in meeting climate change policy goals. *Transportation Research Part A: Policy and Practice, 44(8),* 565–574.

Pasha, Mosabbir, Rifaat, Shakil, Tay, Richard, & DeBarros, Alex. (2016). Effects of Street Pattern, Traffic, Road Infrastructure, Socioeconomic and Demographic Characteristics on Public Transit Ridership. *KSEC Journal of Civil Engineering, 20,* 1017-2022.

Tyrinopoulos, Yannis & Antoniou, Constantinos. (2013). Factors affecting modal choice in urban mobility. *European Transport Research Review, 5,* 27–39.

Salon, D., Boarnet, M. G., Handy, S., Spears, S., & Tal, G. (2012). How do local actions affect VMT? A critical review of the empirical evidence. *Transportation Research Part D: Transport and Environment, 17(7),* 495–508.

Sanaullah, Irum, Alsaleh, Nael, Djavadian, Shadi, & Farooq, Biblal. Spatio-temporal analysis of on-demand transit: A case study of Belleville, Canada. *Transportation Research Part A: Policy and Practice, 145,* 284-301.

Storchmann, Karl. (2001). The impact of fuel taxes on public transport – an empirical assessment for Germany. *Transport Policy, 8(1),* 19-28.

Statistics Canada. (2016). Geographic Classifications. Retrieved from https://www.statcan.gc.ca/eng/subjects/standard/pcrac/2016/introduction

Syed, Sharfuddin & Khan, Ata. (2000). Factor Analysis for the Study of Determinants of Public Transit Ridership. *Journal of Public Transportation, 3(3),* 1-17.

Taylor, Brian, Miller, Douglas, Iseki, Hiroyuki, & Fink, Camille. (2009). Nature and/or nurture? Analyzing the determinants of transit ridership across US urbanized areas. *Transportation Research. Part A, Policy and Practice*, 43(1), 60-77.

Taylor, Brain, & Fink, Camille (2003). The Factors Influencing Transit Ridership: A Review and Analysis of the Ridership Literature. *UC Berkeley: University of California Transportation Center.*

Taylor, Brain, & Fink, Camille. (2013). Explaining transit ridership: What has the evidence shown? *Transportation Letters, 5(1),* 15-26.

Thompson, Gregory, & Brown, Jeffrey. (2006). Explaining Variation in Transit Ridership in U.S. Metropolitan Areas Between 1990 and 2000: Multivariate Analysis. *Transportation Research Record*, 1986, 172-181.

Thompson, Gregory, Brown, Jeffery, & Bhattacharya Torsha. (2012). What Really Matters for Increasing Transit Ridership: Understanding the Determinants of Transit Ridership Demand in Broward County, Florida. *Urban Studies, 49(15),* 3327-3345.