Essays on the Economic Analysis of Transportation Systems

Hyoungsuk Shim
Graduate Center, City University of New York

How does access to this work benefit you? Let us know!

Follow this and additional works at: https://academicworks.cuny.edu/gc_etds

Part of the Economics Commons

Recommended Citation
https://academicworks.cuny.edu/gc_etds/1129

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
Essays on the Economic Analysis of Transportation Systems

by

Hyoung Suk Shim

A dissertation submitted to the Graduate Faculty in Economics in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2015
©2015

Hyoung Suk Shim

All Rights Reserved
This manuscript has been read and accepted for the Graduate Faculty in Economics in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

__________________________________________

Dr. Jonathan R. Peters
Date: Chair of Examining Committee

__________________________________________

Dr. Wim Vijverberg
Date: Executive Officer

Dr. Jonathan R. Peters
Dr. David King
Dr. Michael E. Kress
Dr. Wim Vijverberg
Dr. Yochanan Shachmurove

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK
Abstract

Essays on the Economic Analysis of Transportation Systems

by

Hyoung Suk Shim

Adviser: Professor Jonathan R. Peters

This dissertation consists of four essays on the economic analysis of transportation systems. In the first chapter, the conventional disaggregate travel demand model, a probability model for the modeling of multiple modes, generally called random utility maximization (RUM), is expanded to a model of count of mode choice. The extended travel demand model is derived from general economic theory – maximizing instantaneous utility on the time horizon, subject to a budget constraint – and can capture the dynamic behavior of countable travel demand. Because the model is for countable dependent variables, it has a more realistic set of assumptions to explain travel demand then the RUM model. An empirical test of the theoretical model using a toll facility user survey in the New York City area was performed. The results show that the theoretical model explain more than 50 percent of the trip frequency behavior observed in the New York City toll facility users. Travel demand for facility users increase with respect to household employment, household vehicle count, and employer payment for tolls and decrease with travel time, road pricing, travel distance and mass transit access.

In the second chapter, we perform a statistical comparison of driving travel demand on toll facilities between Electronic Toll Collection (ETC) users, as a treatment group, and non users, as a control group, in order to examine the effect of ETC on travel demand that uses toll facilities. The data that is used for the comparison is a user survey of the ten toll bridges and tunnels in New York City, and the data contains individual user’s travel attributes and demographic characteristics, as well as the frequency of usage of the toll facilities so that the data thus allows us to examine
the difference in travel demand of E-ZPass, the Electronic Toll Collection System for Northeastern United States’ highway ETC system and compare tag holders and non tag holders. We find that the estimated difference of travel demand between E-ZPass users and non-users is biased due to model misspecification and sampling selection, and E-ZPass has no statistically significant effect on travel demand after controlling for possible sources of biases.

In the third chapter, we develop a parallel sparse matrix-transpose-matrix multiplication algorithm using the outer product of row vectors. The outer product algorithm works with the compressed sparse row (CSR) form matrix, and as such it does not require a transposition operation prior to perform multiplication. In addition, since the outer product algorithm in the parallel implementation decomposes a matrix by rows, it thus imposes no additional restrictions with respect to matrix size and shape. We particularly focus on implementation of this technique on rectangular matrices, which have a larger number of rows and smaller number or columns for performing statistical analysis on large scale data. We test the outer product algorithm for randomly generated matrices. We then apply it to compute descriptive statistics of the New York City taxicab data, which is originally given by a 140.56 Gbytes file. The performance measures of the test and application shows that the outer product algorithm is effective and performed well on large-scale matrix multiplication in a parallel computing environment.

In the last chapter, I develop a taxi market mechanism design model that demonstrates the role of a regulated taxi fare system on taxi drivers’ route choice behavior. In this model, a fare system is imposed by a taxi market authority with the recognition of asymmetric information, which in this case is about road network and traffic conditions, between passengers and drivers, and taxi trip demand is different and uncertain at its origin and destination. I derive a prediction from the model that shows the drivers have an incentive to make trip longer than optimal if they have passengers
whose trip origin has more taxi demand than the trip’s destination area, and the fare rule is a metered fare, which is non-negotiable. If the passengers’ trip destination has more demand than its origin and the fare rule is negotiable, on the other hand, then the drivers have an incentive to minimize the trip time and distance. I then perform large scale generalized methods of moment (GMM) estimations and find from the estimation that the theoretical prediction is consistent with empirical foundations from the GMM estimation results using 378,532,118 New York City taxicab trip records from 2008 to 2010.
I would like to quote Richard Feynman’s thought on quantum mechanics that “quantum mechanics has not been developed by one single scientist. It has been developed by many scientists.” to express my gratitude to all of my teachers, colleagues, friends and families who helped me to complete the Ph.D. program with this thesis.

Although my thesis is not about quantum mechanics, it could not have been completed without their help and support. First, I would like to thank Prof. Chung Sik Yoo, and Prof. Tae Jeong Lee at Yonsei University who motivated me and encouraged me to dream about becoming an economist. Their lectures on economics were life changing experience for me and it is still exciting to think about the ideas that they taught in the lectures. Without those lectures and their careful and thoughtful advisement, I would not have even dreamed about all the privileges that I have now.

Also, I would like to thank Prof. Michael Grossman, Prof. David Jaeger, Prof. Theodore Joyce, and Prof. Wim Vijverberg for their lectures on microeconomics and microeconometrics. I had expected to become a macroeconomist or a macroeconometrician before I joined the CUNY Graduate Center, but their lectures showed me that microeconomic analysis is intellectually sharp and very scientifically refined - this encouraged me to change my academic focus from macroeconomics to microeconomics without any hesitation.

Regardless of changing my academic interests, however, I have also learned a lot of interesting macroeconomic theories and arguments from Prof. Tom Thurston, Prof. Merih Uctum, and Prof. Yochanan Shachmurove. Especially important to me, the
applied economics seminar in spring 2011, organized by Prof. Shachmurove, provided me with a very exciting opportunity to catch up current topics on macroeconomics. I appreciate Prof. Shachmurove giving me a chance to see many notable and active macroeconomists’ present their work in this seminar.

I have to thank also Prof. Michael Kress, who gave me a tremendous opportunity that opened my eyes on scientific research in general. Because of his advice and support, I was given a chance to work on and get hand on experience with high performance computing systems. I was also was given the opportunity to collaborate with brilliant scientists from many different fields such as fluid mechanics, computational chemists and biologists, particle physicists, computer scientists and so on. Prof. Kress’ advice and guidance based on his broad academic interests and knowledge guided me to be more open minded with respect to scientific research and study on other subjects and fields besides of my own. I am still struggling not to be a narrow minded and isolated scholar, but prof. Kress’ model of a broad and creative scholar who contributes in many fields is definitely now my life time goal as a scholar that I would like to try to achieve.

I should mention how much I am indebted to my CUNY HPC center family. Paul Muzio, Nikoalos Trikoupis, Yauheni Dzedzits, Bimal Kulasekala, Michael Costantino, Nora Santiago, Bukurije Bajrami, and Richard Walsh were all contributing my study on high performance computing systems. Over the four years of my time in CUNY HPC center, I had a great help and support from them. I am sure no economics and social science Ph.D. students on earth have an access and permission to use such a prestigious supercomputer like me at the HPC center. It was absolutely great gift for me and all happened because of my HPC center families. I truly appreciate their support and help.

I am also indebted so much to my friends and classmates in the CUNY Graduate Center. Damir Cosic, Gabriel Movsessyan, Luz Salas, Kyle Bauser, Michael Male-
baum, Suleyman Taspinar, especially my beloved brother Jangwook Lee, and all my other friends who spent and shared this hard time with me. Sometimes they were good friends, sometimes they were my teachers, and sometimes they were my family who I could rely on. I cannot imagine me completing my Ph.D. courses without them. I truly appreciate their warm and thoughtful friendship.

My thanks to my committee members, Professors Peters, Vijverberg, Kress and Shachmurove of the CUNY Graduate School and a special thanks to Dr. David King of Columbia University, for his willingness to participate on my doctoral dissertation committee. Their review and comments regarding my work were invaluable in completing this task.

Finally, I should thank to my adviser Prof. Jonathan Peters who is not just an academic adviser to me, but who rather has become my intellectual father. Prof. Peters and I met and have worked together since I was in the first year of the Ph.D. program. I remember clearly how rough my communication skills were at that time, my difficulty with the English language, and my lack of interpersonal relationship building skills – besides my poor knowledge of economics as well. Because of these issues, advising me was a difficult and burdensome task – much like giving birth to a child and raising him up from infant to adult. But he never gave up on me or reduced his efforts to help and support my studies. I cannot describe in this short acknowledgement all things I got from Prof. Peters but I can say that his advisement was given with a strong sense of compassion and care. Prof. Peters showed me what real collegiality is and also how to cultivate a new generation of scholars. He is a great mentor and I wholeheartedly appreciate his help and support.
I dedicate this thesis to my mother Sueng Hee Lee, and my sister Jiyeon Shim for their sacrificing and support for me and my academic path.
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>iv</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiv</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xv</td>
</tr>
</tbody>
</table>

### 1 Disaggregate Multimodal Travel Demand Modeling Based on Road Pricing and Access to Transit

with Jonathan R. Peters, and Michael E. Kress

Reprinted with permission from

*Transportation Research Record: Journal of the Transportation Research Board, 2263, 57-65, (2011)*

1.1 Introduction .................................................. 2

1.1.1 Literature Review ..................................... 3

1.2 The Travel Demand Model .................................. 6

1.2.1 The Framework of Household Choice Behavior .......... 7

1.2.2 The Household Optimization under Uncertainty ......... 9

1.2.3 Mode Choice and Countable Travel Demand Model ....... 11

1.2.4 Characteristics of Travel Demand ......................... 12

1.3 Empirical Analysis ........................................ 14

1.3.1 Empirical Model Specification .......................... 15

1.3.2 Data Source ........................................... 16

1.3.3 Results ................................................ 19

1.4 Conclusion .................................................. 31

### 2 Electronic Toll Collection Systems and Travel Demand: A Field Experiment of Toll Facilities in New York City

with Jonathan R. Peters

Reprinted with permission from


2.1 Introduction .................................................. 34

2.2 Toll Facilities of New York City: A Field Experiment .... 38

2.2.1 Electronic Toll Collection System in the United States 39

2.2.2 E-ZPass in Detail ......................................... 40
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.3 Toll Prices of MTA Bridges and Tunnels</td>
<td>41</td>
</tr>
<tr>
<td>2.3 Theory of Travel Demand and Experimental Design</td>
<td>42</td>
</tr>
<tr>
<td>2.3.1 Travel Demand: Review</td>
<td>43</td>
</tr>
<tr>
<td>2.3.2 Average Treatment Effect on Travel Demand via Driving</td>
<td>45</td>
</tr>
<tr>
<td>2.3.3 Propensity Score Matching</td>
<td>46</td>
</tr>
<tr>
<td>2.4 Empirical Analysis</td>
<td>48</td>
</tr>
<tr>
<td>2.4.1 Data</td>
<td>49</td>
</tr>
<tr>
<td>2.4.2 Simple Difference in Difference Estimates</td>
<td>50</td>
</tr>
<tr>
<td>2.4.3 Propensity Score Matching</td>
<td>53</td>
</tr>
<tr>
<td>2.5 Conclusion</td>
<td>58</td>
</tr>
<tr>
<td>3 Parallel Sparse Algorithm for Matrix-Transpose-Matrix Multiplication with Michael E. Kress</td>
<td>60</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>61</td>
</tr>
<tr>
<td>3.2 Outer Product Algorithm</td>
<td>64</td>
</tr>
<tr>
<td>3.2.1 Parallel Algorithm for Dense Matrices</td>
<td>64</td>
</tr>
<tr>
<td>3.2.2 Parallel Algorithm for Sparse Matrices</td>
<td>66</td>
</tr>
<tr>
<td>3.2.3 Comparison with Conventional Algorithms</td>
<td>69</td>
</tr>
<tr>
<td>3.3 Experiment</td>
<td>72</td>
</tr>
<tr>
<td>3.3.1 Experimental Design</td>
<td>73</td>
</tr>
<tr>
<td>3.3.2 Random Matrix Results</td>
<td>75</td>
</tr>
<tr>
<td>3.4 Application: NYC Taxicab Data</td>
<td>79</td>
</tr>
<tr>
<td>3.4.1 Manipulation of Big Data</td>
<td>80</td>
</tr>
<tr>
<td>3.4.2 Parallel Execution of Taxi Data Analysis</td>
<td>82</td>
</tr>
<tr>
<td>3.4.3 Analysis: Taxicab data</td>
<td>84</td>
</tr>
<tr>
<td>3.5 Conclusions</td>
<td>90</td>
</tr>
<tr>
<td>4 Principal versus Agent: Market Operation Mechanism of the New York City Taxicab Industry</td>
<td>92</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>93</td>
</tr>
<tr>
<td>4.2 Mechanism Design of Taxi Industry: Model</td>
<td>97</td>
</tr>
<tr>
<td>4.2.1 Travel Demand for Taxi Service</td>
<td>97</td>
</tr>
<tr>
<td>4.2.2 Taxi Market Mechanism</td>
<td>101</td>
</tr>
<tr>
<td>4.3 Mechanism Design of Taxi Industry: Analysis</td>
<td>106</td>
</tr>
<tr>
<td>4.3.1 Equilibrium of the Mechanism</td>
<td>106</td>
</tr>
<tr>
<td>4.3.2 Analysis of Taxi Fare System</td>
<td>108</td>
</tr>
<tr>
<td>4.4 Issues in Empirical Analysis</td>
<td>116</td>
</tr>
<tr>
<td>4.4.1 An Observable Model for Taxi Market Mechanism</td>
<td>116</td>
</tr>
<tr>
<td>4.4.2 Data</td>
<td>118</td>
</tr>
<tr>
<td>4.4.3 Large Scale GMM Estimation</td>
<td>121</td>
</tr>
<tr>
<td>4.5 Empirical Analysis</td>
<td>125</td>
</tr>
<tr>
<td>4.5.1 Identifying (Natural) Experimental Fields</td>
<td>126</td>
</tr>
<tr>
<td>4.5.2 Additional Aspects of Heterogeneous Trip Characteristics- Time of Day</td>
<td>127</td>
</tr>
<tr>
<td>4.5.3 Model Estimation: The Main Result</td>
<td>129</td>
</tr>
<tr>
<td>4.5.4 Time and Spatial Variation of $\hat{\beta}_0$</td>
<td>133</td>
</tr>
<tr>
<td>4.6 Conclusion</td>
<td>139</td>
</tr>
<tr>
<td>4.7 Appendix</td>
<td>142</td>
</tr>
<tr>
<td>4.7.1 Mathematical Proofs</td>
<td>142</td>
</tr>
<tr>
<td>4.7.2 Dummy variable</td>
<td>144</td>
</tr>
</tbody>
</table>

Bibliography | 146

Bibliography | 150
# List of Figures

1.1 Toll Burdens per Capita ............................................. 17
1.2 Feasible Mode for Travel ........................................... 20

2.1 Locations of TBTA Toll Facilities ................................. 38
2.2 Diagram of Travel Modes ............................................ 43
2.3 GIS Characteristics of New York State Transportation System . 49
2.4 Histograms of Propensity Score: With Mass Transit ............ 54
2.5 Histograms of Propensity Score: Without Mass Transit .......... 55

3.1 Matrix decomposition for $A^T B$ .................................. 71
3.2 Spy plot of the 1,000 by 1,000 random matrix .................... 73
3.3 Gigaflop Performance Measures of $A^T B$ ......................... 76
3.4 Execution and communication times for $A^T B$ .................... 78
3.5 Spatial Distribution of NYC Taxicabs ............................... 82
3.6 Time Plot: Mean Taxi Fares and Travel Time ...................... 86
3.7 Spatial Distribution of Taxi Fares .................................. 88
3.8 Spatial Distribution of Trip Time by ZIP Code Polygons .......... 89

4.1 New York City Road Network and a Trip Route .................. 103
4.2 Spatial Distribution of NYC Taxicabs ............................... 110
4.3 Probability (Spatial) Distribution Function: Examples ............ 111
4.4 Comparison and Distance Determination ........................... 114
4.5 Traffic Analysis Zone (TAZ): John. F. Kennedy Airport .......... 119
4.6 Traffic Analysis Zone (TAZ): John. F. Kennedy Airport .......... 120
4.7 Taxi Trip Frequency Maps by TAZ Polygons ....................... 128
# List of Tables

<table>
<thead>
<tr>
<th>Section</th>
<th>Table Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Descriptive Statistics</td>
<td>18</td>
</tr>
<tr>
<td>1.2</td>
<td>Baseline Model Comparison 1</td>
<td>26</td>
</tr>
<tr>
<td>1.3</td>
<td>Baseline Model Comparison 2</td>
<td>27</td>
</tr>
<tr>
<td>1.4</td>
<td>Baseline Model Comparison 3</td>
<td>28</td>
</tr>
<tr>
<td>1.5</td>
<td>Baseline Model Comparison 4</td>
<td>29</td>
</tr>
<tr>
<td>1.6</td>
<td>Negative Binomial Estimation</td>
<td>30</td>
</tr>
<tr>
<td>2.1</td>
<td>Toll Price of MTA bridges and Tunnels</td>
<td>39</td>
</tr>
<tr>
<td>2.2</td>
<td>Differences in Differences Estimation for Average Treatment Effect</td>
<td>52</td>
</tr>
<tr>
<td>2.3</td>
<td>Propensity Score Matched Average Treatment Effect</td>
<td>56</td>
</tr>
<tr>
<td>2.4</td>
<td>Descriptive Statistics without Mass Transit Access</td>
<td>57</td>
</tr>
<tr>
<td>2.5</td>
<td>Descriptive Statistics without Mass Transit Access</td>
<td>58</td>
</tr>
<tr>
<td>3.1</td>
<td>Statistics of experimental matrices</td>
<td>75</td>
</tr>
<tr>
<td>3.2</td>
<td>Statistics of the Experimental Matrices</td>
<td>83</td>
</tr>
<tr>
<td>3.3</td>
<td>Means and Variance-Covariance Matrix of NYC Taxicab Data</td>
<td>85</td>
</tr>
<tr>
<td>4.1</td>
<td>Frequency of Taxi Trips in New York City Boroughs (Weekdays)</td>
<td>127</td>
</tr>
<tr>
<td>4.2</td>
<td>GMM Estimates of Taxi Trip Attributes</td>
<td>131</td>
</tr>
<tr>
<td>4.3</td>
<td>GMM Estimates of Mass Transit Terminal Indicators</td>
<td>135</td>
</tr>
<tr>
<td>4.4</td>
<td>GMM Estimates of Peak Time Indicators</td>
<td>136</td>
</tr>
<tr>
<td>4.5</td>
<td>GMM Estimates of Weekday Indicators</td>
<td>138</td>
</tr>
<tr>
<td>4.6</td>
<td>A matrix of dummy variables by category</td>
<td>145</td>
</tr>
</tbody>
</table>
1 Disaggregate Multimodal Travel Demand Modeling Based on Road Pricing and Access to Transit

with Jonathan R. Peters, and Michael E. Kress

Reprinted with permission from

Transportation Research Record: Journal of the Transportation Research Board,

2263, 57-65, (2011)
1. TRAVEL DEMAND MODEL

1.1 Introduction

In 2009, the National Academies released a document that identified nine critical transportation issues facing the United States in the first decade of the 21st century: congestion, energy, infrastructure maintenance, finances, equity, emergency preparedness, safety, inadequate institutions, and inadequate human intellectual capital.

Although the National Academies report focused on the transportation needs of the United States, countries around the world face these same critical transportation issues. To address two of these critical issues (congestion and finances), the Federal Highway Administration and many state and local governments have considered additional forms of taxation and road pricing. Pricing has the advantage in many cases of both lowering demand for a particular segment of highway or facility and providing revenue for road investment or other purposes. There are a variety of ways to apply pricing to road systems, and these include simple tolls, time of day pricing, congestion pricing zones, dynamic pricing systems and vehicle miles traveled charging systems. As such, there is a need to study the impact of road pricing on travel behavior and demand structure, as well as the interaction effect of pricing and mass transit services on mode choice through the use of advanced travel demand modeling.

To better understand why people travel more or less, the development of an appropriate theory of travel demand is a priority research task. Numerous papers on travel demand have been developed over the past 40 years to better understand empirical and theoretical concerns. Most of the travel demand literature has examined empirical and theoretical analysis simultaneously, rather than independently, because travel demand analysis has some potential pitfalls that are difficult to justify in practice—such as limited data sources, inappropriate model selection, and so forth—so the appropriateness of conduction empirical or theoretical analysis independently cannot be guaranteed.
1. TRAVEL DEMAND MODEL

1.1.1 Literature Review

The simultaneous consideration of empirical and theoretical analysis requires the development of a model specification from relevant theory and the empirical testing of that model. Choosing an appropriate model specification and combination of attributes is necessary to estimate a model consistently without bias. Quandt (1970) argued that an empirical travel demand study should be estimated by using a correctly chosen functional form, a relevant set of attributes, and an appropriate set of statistical inference methods; if at least one of these considerations is not upheld, then the analysis contains a serious bias problem.

In microeconomics, the correct functional form of demand can be chosen by solving the consumer’s utility maximization, subject to their budget constraints. Because the demand is derived from the utility maximization, which contains information about each individual’s preferences, the demand captures general consumption behavior. The use of utility maximization ensures that the explanatory variables are chosen from the consumer optimization. After the model has been developed under the appropriate theoretical constructs, the choice of statistical methods should be the next step.

The explanatory study of travel demand data is one of the two primary ways that the travel demand literature has developed; however, the literature is generally based on the idea that travel demand has an independent decision process, separated from the other usual economic decisions, such as consumption, and the labor-leisure choice. So, the functional form, and the set of attributes can be chosen in an empirical way. In other words, it is possible to set the functional form and attributes without theoretical verification of whether the model is correctly specified. The result of explanatory studies for travel demand, therefore, might have a serious bias because of absence of an appropriate theoretical model.
The most widely known model of transit demand is a mode choice type model derived from Random Utility Maximization (RUM). RUM is a theoretical framework of consumer choice used to analyze a particular type of consumer who has a discrete choice of modes of travel. The consumer in RUM has a known probability of choosing a particular mode that is associated with the mode’s attributes. The econometric estimation of RUM can be performed by multinomial logit (MNL), sometimes called the conditional logit model because the mechanics of RUM are to assign probabilities to each mode choice. The probability of each mode being chosen is treated as a function of the mode’s attributes. McFadden (1974) proposed a travel demand system for the mode choice of travel. The specification of individual choice is that, individuals are rational, and, therefore, users choose the optimal mode of travel to minimize the level of deprivation caused by the travel.

Since McFadden (1974) was published, much of the literature on travel demand has used RUM concepts of mode choice. McFadden (1978), Williams (1977), Daly and Zachary (1979) and Ben-Akiva and Lerman (1979) conducted RUM specification of travel demand and estimate using nested MultiNomial Logit (MNL). Ben-Akiva et al. (1987) use a combined model of two different specification: the first for revealed preference and, the second using stated preference data in a single model specification under the RUM concept. Morikawa (1989) used this model to analyze travel demand in Netherland. Because data for travel demand have collected, in many cases, in either a revealed preference or stated preference form, travel demand literature, especially empirical research, has utilized the available data. Morikawa (1989), Ben-Akiva and Morikawa (1990), Ben-Akiva and Yamada(1991), Hensher and Bradley (1993), Louviere (1993, 1999), and Hensher et al. (1989), Brownstone and Train (1999) all analysed and tested RUM specification models using stated, and revealed preference data. After 2000, almost all travel demand research that considered individual choice behavior was conducted using the RUM specification, whether the research interest
was empirical analysis or theoretical modeling. This recent trend has also affected the development of commercial software for analysis of transportation system, such as TransCAD.

In McFadden’s model, the RUM specification of consumer optimization for travel demand assumes that travel demand is not associated with general economic decision of the household but has independent decision-making process, which means that individuals or households decide to travel for travel itself, not to support general economic activities, such as consuming goods and services, working, and undertaking leisure activities. For this reason, it is hard to find theoretical consistency with the RUM model specification and general microeconomic literature. With the RUM specification, it is difficult to explain the functional purpose of travel, as travel itself might provide utility for some reason. The functional purpose of travel can be either work, leisure, or the buying or selling of goods and services. Utility is not directly created by travel in these cases; it is created by performing the other activities through travel. In addition, the mode-choice model is a probabilistic choice analysis that implies how much possibility an individual has for choosing a particular mode, not how many times the individual is willing to choose the particular mode. The probabilistic choice analysis is therefore hard to extend its physical demand for travel, such as the number of trips in a given mode (driving, for example) in a given period. In addition, not knowing the quantity of travel demanded in a given period creates an additional challenge, because it is difficult to draw policy implications from the results for transportation planning purposes.

This paper provides an extended travel demand model for trip frequency, based on the household lifetime utility maximization problem, subject to budget and time constraints. Travel events, in the extended model, are treated as supplements to support the general economic activities of consumption, labor, and leisure. As the household faces uncertainty in travel time for future travel, the household selects the
number of travel events for each period, and the process considers the uncertainty by updating travel time information. Thus, the model captures the dynamic behavior of countable travel demand associated with the attributes of that demand.

In the next section, the extended model for travel demand is outlined. The theoretical model is then tested through the use of a toll facility user survey from the New York City area. A description of the data used for the empirical analysis, and its characteristics, is then presented. Finally, the paper discusses the empirical analysis results and the study conclusions.

1.2 The Travel Demand Model

In this section, we consider a model of a household’s optimization behavior is presented in which the household chooses consumption and leisure to maximize its lifetime utility, subject to budget and time constraints; these choices are used to derive travel demand. The background concept of the household travel choice behavior is that the household decide on travel events for economic activities (consumption and leisure). In other words, travel is an instrument to support the household’s economic activities; it is not undertaken for the sake of travel itself. In addition, to analyze dynamic choice behavior, lifetime discounted utility maximization is considered in which the household chooses a quantity of consumption and leisure time to maximize the household’s utility over its lifetime. There are two distinguishing features of the model in this study, as compared with McFadden (1974). One is the role of travel; the other is a mechanism to assign levels of utility to choices. McFadden (1974) modeled travel demand under the disutility specification in which individual’s travel choice behavior depends on deprivation levels created by the travel. So the travel did not play a role as an instrument to help general economic activities, and utility were assigned directly to travel itself. This is somewhat paradoxical feature because, in his
1. TRAVEL DEMAND MODEL

paper, McFadden argued that travel is concomitant of other economic activities, such as work, shopping and recreation, not an objective itself. The model in McFadden, however, assigns utility directly to the deprivation level of a particular travel mode, so it implies that an individual chooses travel to minimize disutility from travel itself, not to promote economic activities. To treat travel as an instrument for the economic activities, therefore, travel is evaluated in the current study based on the household time and budget constraints, not on the utility function. Thus, the demand from the model will reflect travel as an instrument to support general economic activities.

1.2.1 The Framework of Household Choice Behavior

Consider $N$ identical households, each of which has a certain preference level over time, given by the utility function; each household is endowed with one unit of time, which can be allocated between work and leisure. The level of utility at each period depends on the amount of goods and services consumption and leisure time in that period. In other words, each household’s utility is a function of consumption and leisure, which can be maximized by the choice of consumption and leisure. Let $p_t \cdot d$ be the cost of travel, where $p_t$ is a unit price of the travel per a distance $d$ at time $t$, and let $h_t$ be the travel time. Because each household’s chooses the amount of consumption $c_t$ and leisure $l_t$ for a given budget and time, the household’s feasible amounts of consumption and leisure are $c_t - p_t \cdot d$, $l_t - h_t$ respectively. This implies that the households allocate time for labor and leisure within a particular period of time so that $l_t$ takes value between 0 and 1, and the allocation of labor is automatically determined as $1 - l_t$. Under these consideration, the instantaneous utility function for a representative household can be written as $u(c_t - p_t \cdot d, l_t - h_t)$. It is assumed that the function $u(\cdot, \cdot)$ is strictly increasing, strictly concave in each argument, and twice differentiable.$^1$

$^1$The discounted lifetime, consumption-leisure utility has been developed to analyze the business cycle of an economy. The Real Business Cycle model uses the utility function to keep their ag-
1. TRAVEL DEMAND MODEL

The representative household faces time and budget constraints when it chooses the amount of consumption and leisure. The household’s budget is determined by labor and capital income, and the labor income is determined by the household’s choice of work hours. Let $w_t$ be a rental price per unit labor, $r_t$ be a unit rental price of capital; $w_t$ and $r_t$ can be thought of as the household’s hourly wage and interest rate, which are evaluated at competitive market equilibrium. The household has limited hours for work, which can be treated as labor supply and calculated as the total unit of time minus the fraction of leisure. Because the travel time should be excluded from the amount of feasible leisure, the household income is $w^*_t(1 - l_t + h_t) + r_t k_t$, where $w^*_t$ is the wage earned by travel and $k_t$ is the amount of capital the household has at time $t$; $w^*_t$ is equal to $w_t$ when the user is working, otherwise it is zero.

To consider choice of travel under uncertainty, it is assumed that travel time is determined by travel distance $d$, and transit (in the general sense of time in motion) conditions\footnote{\textsuperscript{2}}\footnote{\textsuperscript{3}}, which are randomly given\footnote{\textsuperscript{4}}. This assumption comes from the notion that travelers face uncertainty in their travel time caused by, for example, traffic congestion or the failure of mass transit equipment. So the travelers’ choices depend on uncertain travel times, based on their previous experiences. This notion leads to a prediction rule for uncertain travel time based on information that is given by a traveler’s previous experience. Assume that the next period’s travel time is given only by the current period’s transit condition so that the previous transit condition does not affect the next period’s travel time. This can be described by the first-order aggregate macroeconomic analysis consistent with microeconomic foundations. Conducting the time discounted consumption leisure utility has advantages, the household optimization problem can consider two main economic activities of, consumption expenditure and labor supply, simultaneously, and solving the household optimization problem yields the dynamics of consumption, leisure, and labor choices. Hence, the travel demand model, derived from the optimization problem with time discounted utility is consistent with general economic activity because the solution will determine consumption, labor, and travel demand simultaneously, and reflect how the household update the given information to decide future demand.
autoregressive function as

\[ h_{t+1} = f(h_t, d, z_t), \]  

where \( z_t \) is stochastic shock on travel time, drawn according to a stationary transition function \( Q \). As \( h_{t+1} \) is a forecasted travel time based on information from previous period \( h_t \) and \( z_t \), not the expected travel time of the next period. Since \( h_t \) and \( d \) are exogenously given at time \( t \), and \( z_t \) is a stationary process, the evolution of travel time can be described by probability transition density \( Q \). We assume that the function \( f(\cdot, d, \cdot) \in \Gamma \) is strictly increasing in each argument, \( \Gamma \) is a non-empty set.

1.2.2 The Household Optimization under Uncertainty

If the household is rational and there is uncertainty about the future, the household chooses the optimal amount of consumption and leisure time to maximize the household’s expected utility over its lifetime. The household discounted expected lifetime utility maximization problem, therefore, can be written as

\[
\max U = E \left[ \sum_{t=0}^{\infty} \beta^t u(c_t - p_t d, l_t - h_t) \right]
\]

s.t. \( c_t - p_t d \leq w^*_t (1 - l_t + h_t) + r_t k_t \)

\[ h_{t+1} = f(h_t, d, z_t), \]

where \( \beta \) is a discount factor. Suppose the forecasting rule \( h_{t+1} = f(h_t, d, z_t) \) is additively separable, then \( h_t = g(h_{t+1}, d, z_t) \), where \( g(\cdot, d, z_t) \) is a unique inverse function of \( f(\cdot, d, z_t) \). Because the household is rational, it spends all of its budget

\[ ^3 \text{the term transit is generalized in this case to mean time in travel.} \]

\[ ^3 \text{The assumption, travel time is a function of travel distance with randomly given transit condition implies any transit for travel cannot arrive exactly on time to the final destination even though the traveler knows the distance of travel. It depends on transit conditions such as traffic jams for auto transits, unexpected delays due to the failure of mass transit equipment and other unknown conditions. This is the reason why a traveler chooses transit based on a calculated expected travel time for each mode of transit.} \]
and time so that inequality constraint of budget is binding. To reduce the equation to two constraints, $h_t$ is substituted into the budget equation, which gives

$$c_t - p_t d = w_t^* [1 - l_t + g(h_{t+1}, d, z_t)] + r_t k_t. \quad (1.2)$$

Substituting (1.2) into the objective function of the maximization problem yields a reduced form of the objective function. This equation is the household utility, a function of $h_t$, $h_{t+1}$, and $z_t$, because $w_t^*$, $r_t$, and $z_t$ do not change due to the choice of travel mode.

The objective function of the problem is denoted as $F(h_t, h_{t+1}, z_t)$. Under the above specification, the optimization problem can be solved by stochastic dynamic programming. Let $h' \in \Gamma(h, d, z)$ and let it be an endogenous state variable (i.e., a traveler’s forecasted travel time). Further, let $z, z' \in Q$, and let them be distinct stochastic processes that are identically distributed. The endogenous state variable $h'$ and the exogenous state variable $z'$ can be thought of as the future travel time and transit condition, respectively, which gives $h' = h_{t+1}$ and $z' = z_{t+1}$. By assuming that $z_t$ is the first-order Markov process, the time subscript can be dropped, and a dynamic programming solution can be applied for any time. By the principles of optimality, the solution of the stochastic dynamic programming is the same as the solution of a functional equation

$$v(h, z) = \max \left[ F(h, h', z) + \beta \int v(h', z') Q(z, dz') \right], \quad (1.3)$$

where $v$ is a function that represents the solution to the dynamic optimization problem.

Suppose the household determines its consumption and leisure from a different functional form as AR(1). That is $h_{t+1} = d + \rho h_t + z_t$. In this case, the unique inverse function of $h_{t+1} = f(h_t, d, z_t)$ is

$$h_t = g(h_{t+1}, d, z_t) = \frac{1}{\rho}(h_{t+1} - d - z_t).$$
optimization process. Let $v(h_t, z_t)$ be the solution of (1.3) for all $t$, then (1.3) can be rewritten as

$$v(h_t, z_t) = F(h_t, h_{t+1}, z_t) + \beta E_t [v(h_{t+1}, z_{t+1})], \quad \forall t \text{ and } z_t \in Q, \quad (1.4)$$

where $E_t[\cdot, \cdot]$ is the expectation operator. From (1.4), this can be interpreted as the expected benefit given by the future travel. Here, the household decides on travel based on this expected value of future travel so the expectation can be treated as a reference function of the household for the next period’s travel decision. In other words, the household chooses the mode and an amount of travel to maximize its utility, and the decision rule of the household can be summarized by the expectation $E_t[v(h_{t+1}, z_{t+1})]$.

### 1.2.3 Mode Choice and Countable Travel Demand Model

Suppose that the individual has $i = 1, \ldots, J$ feasible transit modes. Let $v(h^i_t, z^i_t)$ be a maximized Bellman equation under the choice of the mode $i$; then the expected reference function for mode $i$ is

$$E_t[v(h^i_{t+1}, z^i_{t+1})] = \frac{1}{\beta} \left[ v(h^i_t, z^i_t) - F(h^i_t, z^i_t) \right], \quad (1.5)$$

where $F(h^i_t, z^i_t)$ is obtained by inserting $h^i_{t+1} = f(h^i_t, d, z^i_t)$ into the objective function.

It is assumed that the choice of mode satisfies the Independence from Irrelevant Alternatives (IIA)\textsuperscript{5}. The assumption implies that the relative probability of mode $i$, as it relates to the other modes, does not change. Therefore, the frequency of a specific mode $i$ is a function of its own attributes only. Let $y^i_t$ be a count of which the household chooses the mode $i$ at $t$, which is treated as a countable travel demand. Further, let $x_t$ be a vector of attributes. Assume that $y^i_t, \forall t$ is drawn according to a poisson distribution in which the mean is the expected benefit (1.4). Thus, $y^i_t$ has a
conditional pdf of:

\[
f(y^i_t | x_t) = \exp[-E\{v(h^i_{t+1}, z^i_{t+1})\} \cdot \frac{E\{v(h^i_{t+1}, z^i_{t+1})\} y^i_t}{y^i_t!}], \quad y^i_t = 0, 1, \ldots \quad (1.6)
\]

Therefore, the future demand for travel of a particular mode \(i\) measured as a countable random variable, is a function of the current period’s travel time \(h_t\) and experienced stochastic travel condition \(z_t\), which are given by the choice of mode \(i\) in the current period. The other attributes, \(w^*_t, r_t k_t, d\) and \(l_t\), affect the entire set of modes simultaneously because they are determined by the exogenous optimization process. In short, the household first determines its consumption and leisure to maximize its utility, then it decides on the mode and number of travel events that are necessary to achieve this optimum.

1.2.4 Characteristics of Travel Demand

In this section, the characteristics of trip frequency \(y^i_t\), which is a countable demand model of travel mode \(i\), are examined by using general properties of functional equations to show how \(y^i_t\) responds to each attribute of travel. Because the realization \(y^i_t\) is drawn according to a poisson distribution with mean \(E_t[v(h^i_{t+1}, z^i_{t+1})]\), the characteristic of \(E_t[v(h^i_{t+1}, z^i_{t+1})]\) will govern the stochastic behavior of \(y^i_t\) itself. Recall that the instantaneous utility function \(u(\cdot, \cdot)\), which is used as the objective function for a household maximization problem, is strictly increasing, strictly concave in each argument, and twice differentiable. From the given specification of the optimization problem and the general properties of functional equation (1.3) come the character-

\(^5\)IIA assumption is the relative probability of mode \(i\) as it relates to the other mode does not change if the additional modes are considered so that the probability affected only on attributes of mode \(i\), not the other mode’s attributes. This assumption support the idea that the count of a specific mode \(i\) is a function of it’s own attributes. See Wooldridge (2002)
1. TRAVEL DEMAND MODEL

The expectation $E_t[v(h^i_{t+1}, z^i_{t+1})]$ is increasing in wage $w^*_t$, financial return $r_t$, $k_t$, and decreasing in travel time of mode $i$ $h^i_t$, travel distance $d$, and the share of leisure $l_t$.

The proof of the proposition is straightforward. Substituting the two arguments of the utility function of the maximization problem into the constraints yields the reduced form objective function $u^{*}\{w^*_t[1 - l_t + g(h_{t+1}, d, z_t)] + r_t k_t\}, l_t - h_t\}$. This is the actual form of the objective function $F(h_t, h_{t+1}, z_t)$ because the other attributes $w^*_t, r_t, k_t, d$ are exogenously given. It is clear that the objective function $F$ is strictly increasing and concave in each argument of $u^{*}(\cdot, \cdot)$. Since $u$ is increasing in $w^*_t, r_t$, and $k_t$, $F$ is increasing in $w^*_t, r_t$, and $k_t$, and since $u$ is decreasing in $h^i_t, d$, and $l_t$ so is $F$. By the Contract Mapping Theorem, the Bellman equation $v(\cdot, \cdot)$ is also increasing in $w^*_t, r_t, k_t, d, l_t$.

In this next step, it is shown how travelers make decisions based on given information. The following proposition shows that the characteristic of $y^i_t$ can be established by examining the relationship between $y^i_t$ and $E_t[v(h^i_{t+1}, z^i_{t+1})]$.

**Proposition 1.2.4.2.** Consider the other attribute vector $\tilde{\mathbf{x}}_t = [\tilde{w}^*_t, \tilde{r}_t, \tilde{k}_t, \tilde{h}^i_t, \tilde{d}, \tilde{l}_t]$, and the associated probability distribution $f(y^i_{t+1}|\tilde{\mathbf{x}}_t)$. $f(y^i_{t+1}|\mathbf{x}_t)$ is second order stochastically dominates $f(y^i_{t+1}|\tilde{\mathbf{x}}_t)$ if one of the following inequalities holds:

(1) $\tilde{w}^*_t \geq w^*_t, \tilde{r}_t \geq r_t, \tilde{k}^*_t \geq k_t$

$f(y^i_t|\tilde{\mathbf{x}}_t)$ is second order stochastically dominated by $f(y^i_t|\mathbf{x}_t)$ if one of the following inequalities holds:

(2) $\tilde{h}^i_t \geq h^i_t, \tilde{d} \geq d, \tilde{l}_t \geq l_t$.

\(^5\)In general, utility function is strictly increasing, and strictly concave in each argument. Note that, concavity assumption implies diminishing marginal utility.
1. TRAVEL DEMAND MODEL

The inequalities (1) and (2) imply that the attribute vector \( \tilde{x}_t \) is greater than \( x_t \). From proposition 1.2.4.1, the greater attributes \( w^*_t, r_t, k_t \) yield the greater expected value \( E_t[v(h^i_{t+1}, z^i_{t+1})] \), and the greater attributes \( h^i_t, d, l_t \) yield the less expected value. In vector form, \( \tilde{x}_t \succeq x_t \) if one of inequalities (1) holds, than \( E_t[v(h^i_{t+1}, z^i_{t+1})|\tilde{x}_t] \geq E_t[v(h^i_{t+1}, z^i_{t+1})|x_t] \). By definition of second order stochastic dominance, hence, the probability distribution \( f(y^i_t|\tilde{x}_t) \) second order stochastically dominates \( f(y^i_t|x_t) \). Conversely, \( \tilde{x}_t \preceq x_t \) if one of inequalities (2) holds, that yields \( E_t[v(h^i_{t+1}, z^i_{t+1})|\tilde{x}_t] \leq E_t[v(h^i_{t+1}, z^i_{t+1})|x_t] \) so that \( f(y^i_t|\tilde{x}_t) \) is second order stochastically dominated by \( f(y^i_t|x_t) \). Therefore, proposition 1.2.4.2 is proven. This proposition can be interpreted as follows: the greater value of \( \tilde{y}^i_t \), which is drawn according to the probability distribution with attributes \( \tilde{x}_t \), occurs more frequently than \( y^i_t \) from a distribution with \( x_t \) when at least one of \( w^*_t, r_t, k_t \) in \( \tilde{x}_t \) is greater than those in \( x_t \). Similarly, the lesser value of \( \tilde{y}^i_t \) occurs more frequently than \( y^i_t \) when at least one of \( h^i_t, d, l_t \) in \( \tilde{x}_t \) is greater than those in \( x_t \). Together, the two propositions imply that, \( y^i_t \), the frequency of trip, is increasing in \( w^*_t, r_t, k_t \), and decreasing in \( h^i_t, d, l_t \).

1.3 Empirical Analysis

In this section, an empirical analysis for the count travel demand model is proposed to test whether the theoretical model is consistent with observed patterns of travel demand. Previous empirical studies on travel demand have focused largely on mode choice. From an econometric perspective, the mode-choice can estimates the probability of each mode as a function of the same set of attributes. The probability of choosing a particular mode only implies how much possibility of use of a given mode an individual has; it does not imply how many times the individual will choose the mode. This point makes the probabilistic choice model argument hard to extend to the physical quantity demanded for a travel mode, such as the number of driving
trips or mass transit rides per period. Furthermore, the absence of quantity demanded makes mode choice hard to use in policy decisions. To derive an appropriate empirical analysis, it is necessary to have an appropriate functional form of an empirical model for travel demand. As a theoretical model for a countable number of trips has been derived the demand for travel is better represented by frequency than by discrete choices; thus, possible empirical specifications that can be used to test this theory are examined, as opposed to mode-choice analysis. The number of occurrences of a particular event from binary choice can be modeled as a Poisson distribution so that the number of travel events in a given period can be estimated by Poisson regression analysis. Thus, three models are considered: the linear model, the Poisson model, and the negative binomial model. The results are compared to examined whether each of these models is theoretically well specified.

1.3.1 Empirical Model Specification

It is proposed that the model of travel demand be estimated by three different methods: ordinary least squares (OLS) under linear model specification, poisson, and the negative binomial model under a count data model specification. If the dependent variable of travel demand, trip frequency is given as real numbers, linear model specification is available to estimate the theoretical finding as outlined in section 2 without serious violations of statistical inference theory. The linear model for travel demand is given as

\[ y_i = \beta_0 + \beta_1 h_i + \beta_2 d + \beta_3 (r \cdot k) + \beta_4 w^* + \epsilon_i, \] (1.7)

where \( y_i \) is travel demand for mode \( i \) on \( (-\infty, \infty) \), \( h_i \) is travel time of mode \( i \), \( d \) is distance of travel, \( r \cdot k \) is disposable income, \( w^* \) is wage given by travel, and \( \epsilon_i \) is error term. The parameters directly imply the marginal effect of each attribute on travel demand and can be estimated through OLS estimation under the classical
linear regression model (CLRM) assumptions.

Poisson, and negative binomial model are the models most frequently used for a count data model. The count data model begins with the assumption that the dependent variable is drawn according to poisson distribution, given as

\[ Pr[y_i = k|x_i] = \frac{e^{-\mu} \mu^k}{y_i!} , \]  

(1.8)

where \( y_i \) is positive integer, \( k \) is number of event occurring. The count data model assumes \( \mu \) the expected value of \( y_i \) is a linear function of it’s attributes, as \( \mu = x_i \beta \). This model is estimated numerically through maximum likelihood estimation because the parameter vector \( \beta \) has no analytical solution. This is called the poisson regression model. One of the properties of a Poisson distribution is that the expected value and the variance of \( y_i \) are the same. That is \( E[y_i] = Var[y_i] = \mu \). Therefore, if the assumption is not satisfied, for example, in the overdispersion case \( \text{v (or heteroskedasticity)}\), \( E[y_i] \leq Var[y_i] \), it is hard to say that the model is correctly specified. In that case, negative binomial estimation can provide a solution because it can control the case \( E[y_i] \leq Var[y_i] \) and estimate the parameter vector. The negative binomial model represents the most suitable model estimation technique for the purpose of the current analysis.

1.3.2 Data Source

This study requires a source of data that would allow the frequency of travel demand for a given mode to be estimated. With the high level of interest in road pricing, the possibility of examining toll facility use is explored, specifically, an origin-destination survey conducted in New York City. The Triborough Bridge and Tunnel Authority (TBTA) – also know as Metropolitan Transportation Authority Bridges & Tunnels, conducted an origin-destination survey in October 2004. The authority traditionally conduct this type of survey roughly every 8 to 10 years. In 2004, the TBTA dis-
tributed 304,000 surveys at Cash toll lanes and mailed survey to 329,000 electronic
toll collection (ETC) system customers[6]. The survey contained 61,201 observations
of passenger car usage on the nine TBTA facilities in New York City. Further expla-
nation of the data source is provided in Peters and Kramer (2009). Table 1.1 provides
an overview of the descriptive statistics of the data.

In addition, the survey had a number of user characteristics reported that allows us
to estimate more fully our travel demand models using the survey data. In particular,
the survey reported information on frequency of use of the facility, household auto
ownership rates, self reported household income, travel time, number of employer in a

data. We acknowledge that these data are possibly subject to self-selection and, as a result, may
yield biased estimates as it is not a survey of randomly selected members of the population.
1. TRAVEL DEMAND MODEL

Table 1.1: Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Observations</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum value</th>
<th>Maximum value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trip Frequency</td>
<td>40765</td>
<td>2.1138</td>
<td>2.1401</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Toll price</td>
<td>41132</td>
<td>3.1602</td>
<td>1.5388</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Travel time</td>
<td>41132</td>
<td>97.7882</td>
<td>118.2886</td>
<td>1</td>
<td>1350</td>
</tr>
<tr>
<td>Distance (Miles)</td>
<td>41132</td>
<td>35.2099</td>
<td>60.4574</td>
<td>1.0538</td>
<td>2937.847</td>
</tr>
<tr>
<td>Household Income</td>
<td>37723</td>
<td>100644</td>
<td>59252.83</td>
<td>12500</td>
<td>2000000</td>
</tr>
<tr>
<td>Number of bus lines (Origin area)</td>
<td>20284</td>
<td>10.4912</td>
<td>13.1144</td>
<td>1</td>
<td>236</td>
</tr>
<tr>
<td>Number of bus lines (Destination area)</td>
<td>21373</td>
<td>10.4219</td>
<td>12.8826</td>
<td>1</td>
<td>236</td>
</tr>
<tr>
<td>Number of subway stops (Origin area)</td>
<td>13840</td>
<td>7.8066</td>
<td>5.4900</td>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td>Number of subway stops (Destination area)</td>
<td>14372</td>
<td>7.8336</td>
<td>5.5083</td>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td>Number of employed (Household)</td>
<td>40451</td>
<td>1.7397</td>
<td>1.0035</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Number of household vehicles</td>
<td>40639</td>
<td>1.9508</td>
<td>0.9502</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

household, and other demographic variables. A pricing algorithm was created in this study based on the direction of travel, payment method, pricing plan participation (resident plans) and facility used. This allowed the impact of price on demand for the facility to be estimated, which was not in the initial data survey. The distance of travel for each trip over the toll facilities was calculated based on the stated origin and destination ZIP Codes. In addition, knowing the facility used, two forms of travel distances are able to be predicted. The first was an estimate of the crow-flies distance, based on the shortest route from the origin zip to the facility used and then a second estimate from the facility used to the final destination zip. A linear estimate of these distances was measured based on longitude and latitude of the ZIP Codes and facilities.

To consider the effects of mass transit modes on the model estimate, a set of density of mass transit service variables was created utilizing data from the New York City Transit Authority. The variables were the number of local and express bus lines per ZIP Codes and the number of subway stations per ZIP Code for the ZIP

7Note that the household income is currently utilized in the model as a 10category metric. We are aware of the possibility that this categorized income variable may cause a measurement error in model estimation. We thus are consider utilizing a work dummy variable as a proxy variable for labor income in future work, and will estimate a model with the two income variables together.
Codes within the borders of New York City that have both facilities.

Each record in the sample had both an origin and a destination ZIP Code. As such, it was possible to study the impact of high or low transit density on both ends of a particular trip. It was expected that a trip that originated in a low transit density ZIP Code would have a higher frequency of trips on a toll facility compared with a trip that originates in a high transit density ZIP Code. A similar effect was posited for the destination ZIP Codes; that is, high mass transit density would predict a lower frequency of toll facility use.

In addition, the survey contained a number of user characteristics that allowed the travel demand models to be more fully estimated using the survey data. In particular, the survey reported information on the frequency of use of the facility, household auto ownership rates, self-reported household income, travel time, number of employed people in a household, and other demographic variables.

1.3.3 Results

This project represents an opportunity to examine the discrete choices of toll facility users understand their motivation for toll facility use. Figure 1.2 shows a tree diagram of the possible mode choices. In this study, the toll auto choice branch was examined. A number of model specifications were tested with regard to the general structural relationships. The robustness of the models was tested using the alternative model specifications outlined in the section 1.3.1.

The results of these model estimates are reported in Table 1.2. Model 1, 2 and 3 provide the results for the model using standard OLS estimation and with alternative specification that either included or excluded the availability of mass transit services. Models 4, 5 and 6 utilized the same explanatory variables, however, these models are estimated using a Poisson specification. in Models 7, 8 and 9, we estimate the model using the same explanatory variables, but the coefficients were estimated using
Figure 1.2: Feasible Mode for Travel
1. TRAVEL DEMAND MODEL

a Negative Binomial model specification.

The results were robust in terms of model specification, as well as the variables, which generally exhibited the correct signs and were for the most part, statistically significant. In addition, the models had a good level of explanatory power and consistently explained about 50% of the variation in the frequency of toll facility use. Given that this is a cross sectional data set, this represents a generally good level of exploratory power. In particular, the coefficients indicate that the model design is appropriate. The overall observations are discussed below.

Each user in this survey was asked about the frequency of trip that was examined by the survey. Users self reported data using frequency classes that ranged from “five or more trips per week” to “less than monthly”. These trip rates were converted into a weekly frequency of use based on the class marks of the individual class groups. The heavy users (five or more trips per week) tended to be located geographically closer to the facilities used than lower frequency users (one or less trips per week). Alternative groupings of frequency classes were examined to test the robustness of the model specification. We found that the model specification was robust across alternative groupings of frequency (three, four or five class groups). That is coefficients are minimally changed by alternative specifications of the trip frequency.

Alternative panels of explanatory variables were then examined to evaluate their impact on model explanatory power as well as to test the impact of individual coefficients. Poisson and Negative Binomial models were fit via maximum likelihood estimation using STATA. In some cases, the model did not converge so we were unable to obtain results on certain model specifications. The explanatory variables were selected based upon our prior discussion of their theoretical models related to transportation as a contributor to lifecycle utility maximization.

- Price of trip (toll paid). The observed coefficient is consistently negative and is highly statistically significant in most cases. With a range of 0.00 to 8.00 per trip in
terms of the toll price in this data series, there is considerable variation in pricing that is observed between users and facility. An additional improvement to the model would be to estimate the total trip cost based upon miles traveled and an expected cost per mile. Given the high level of tolls and the short average trip length in general, it is highly likely that the toll cost represents a significant portion of the variable costs of these trips. The results, related to travel time indicate a consistent, significant, and negative relationship as related to trip frequency. This is consistent with our theoretical expectations and the concept that users seek to minimize their time use in travel due to the negative value of travel time (minimizing the level of deprivation over their life cycle).

• Distance. This observed coefficient was found to be negative and statistically significant with longer trips tending to be less frequent than trips that are shorter in length. This is again consistent with our theoretical expectations, as consumers are expected to attempt in general to minimize their own time input into transportation services. As such, we expect longer trips to occur with less frequency, than shorter ones with the same fundamental characteristics.

• Work Purpose. Trips that were taken for work purposes were much more frequent than other trips (primarily recreational trips). The impact of work travel was found to be a significant and dominating factor in the decision to use a toll facility in New York City. This is in alignment with our expected model, as a consumer traveling for work purposes should be earning wage \( w_i \) per hour or be gaining some other compensation for their work effort. Given that the trip time, purpose and destination may have been dictated by the employer for these trips, it is not surprising that the decision to travel is largely determined by the work purposes. Coefficients had the expected positive sign and were generally very significant across specifications.

• Income. This variable is observed to exhibit a counterintuitive negative sign counter to our posited relationship. It is highly probable that there exists a high
level of multicollinearity between income and other explanatory variables. Income variables in a well specified model may have somewhat confusing signs, as it may mix in demand components as well as supply issues in the coefficient. As such, it has been reported on Table 4, and 5 that this variable may exhibit coefficients with counterintuitive signs. Further analysis of this relationship may be warranted.

- Household employment and household vehicle count. These variables both exhibited positive and statistically significant relationships as related to the frequency of toll facility use. Given that the number of workers in a household should increase the amount of travel for work purposes as well as increase household income, it is not surprising that additional employed members of the household would increase toll facility use. Correspondingly, increases in household vehicle ownership increased the frequency of toll facility usage.

- Transit density variables. These variables were mixed in terms of their explanatory power with the exception of having a subway stop in the destination ZIP Code. Clearly, access to the subway system at the destination ZIP Code is a key determinant in the decision to use toll facilities in New York City. Users with high quality (heavy rail metro mass transit) at their destination have a lower frequency of trip than users with other forms of mass transit. In some model specifications, the density of bus lines at the destination ZIP Code also decreased the frequency of toll facility use, however, the coefficient, while having a consistent sign, tended to be weaker in terms of explanatory power than the Subway density variable.

Table 1.6 presents the results of our estimation of the negative binomial models for a number of alternative sets of variables. The negative binomial models are considered to represent the best and most effective estimates of traveler frequency of toll facility use. We segmented the data to examine the impact of employer payment of toll facility use. The basic model of frequency of use in column 1 indicated a consistent and significant negative impact for toll (price of facility), travel time and distance of
The work trip variable had a strong and significant positive coefficient as did the number of people employed in the household. The number of vehicles in the household had a positive impact on frequency of trips, however the relationship was somewhat weaker statistically. In column 2, adding the variables that are related to payer type had minimal impact on the overall explanatory power of the model. Columns 3 and 4 splits the data by payer type (self pay versus employer pay/reimburse) and also adds the transit density (bus lines and subway stops) at trip origin. Columns 5 and 6 add the transit density variable at the destination ZIP Code as well as splits the data again by payer type.

Dividing the sample into self pay and employer pay provided some very interesting results. In columns 3 to 6, we explore the impact of transit density at the origin zip and destination zip as well as employer versus self pay. First and most significantly, when employers are paying the toll for a given user, the price effect becomes statistically insignificant as a predictor of use, while self payers retain a negative and very statistically significant price coefficient (p<0.01). Interestingly, travel time becomes statistically insignificant for users who have employers pay their tolls, but remains statistically strong and negative for self-payers. Household employment and vehicle counts are both positive and significant for self payers. In some part, the lower significance of the employer pay results may be related to the much smaller sample size of the employer pay data.

Finally, the availability of mass transit services at the destination ZIP Code—subway services in particular—caused a significant reduction in the frequency of toll facility use for self payers. Origin ZIP Code transit availability had no statistically significant impact on frequency of use for both employer pay and self pay users. Transit availability had no impact on users with employer paid tolls, but destination ZIP Code transit availability had an important negative impact on self payers use of toll facilities. The policy impacts of these results are intriguing. Toll facility users with
employer payment are unaffected by both the price of the toll, travel time and also
the availability of mass transit - they have a very high tendency to drive. Self payers
on the other hand are the majority of users on these facilities and they are sensitive to
the supply of mass transit at the destination ZIP Code, travel time, distance as well
as the price of the facility. Given that these user are responsive to these variables,
the opportunity exists to manage facility demand based upon these variables.
Table 1.2: Baseline Model Comparison 1

<table>
<thead>
<tr>
<th>Trip frequency# (5 categories)</th>
<th>OLS</th>
<th>Poisson</th>
<th>Negative Binomial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Toll</td>
<td>-0.0236***</td>
<td>-0.0438***</td>
<td>-0.0431***</td>
</tr>
<tr>
<td></td>
<td>[0.00570]</td>
<td>[0.00642]</td>
<td>[0.0106]</td>
</tr>
<tr>
<td>Travel time</td>
<td>-0.0558***</td>
<td>-0.0241</td>
<td>0.0242</td>
</tr>
<tr>
<td>(Log scale)</td>
<td>[0.0183]</td>
<td>[0.0167]</td>
<td>[0.0310]</td>
</tr>
<tr>
<td>Distance^2</td>
<td>-0.322***</td>
<td>-0.266***</td>
<td>-0.204***</td>
</tr>
<tr>
<td>(Log scale)</td>
<td>[0.0168]</td>
<td>[0.0152]</td>
<td>[0.0457]</td>
</tr>
<tr>
<td>Trip purpose</td>
<td>1.751***</td>
<td>1.715***</td>
<td>1.608***</td>
</tr>
<tr>
<td>(Working dummy)</td>
<td>[0.0220]</td>
<td>[0.0213]</td>
<td>[0.0372]</td>
</tr>
<tr>
<td>Income</td>
<td>-0.105***</td>
<td>-0.0847***</td>
<td>-0.121***</td>
</tr>
<tr>
<td>(Log scale)</td>
<td>[0.0155]</td>
<td>[0.0149]</td>
<td>[0.0272]</td>
</tr>
<tr>
<td>Bus access</td>
<td>-0.00247</td>
<td>-0.00509</td>
<td>-0.00153</td>
</tr>
<tr>
<td>(Origin area)</td>
<td>[0.00219]</td>
<td>[0.00385]</td>
<td>[0.00117]</td>
</tr>
<tr>
<td>Subway access</td>
<td>-0.00143</td>
<td>-0.000554</td>
<td>-0.00135</td>
</tr>
<tr>
<td>(Origin area)</td>
<td>[0.00199]</td>
<td>[0.00408]</td>
<td>[0.00107]</td>
</tr>
<tr>
<td>Bus access</td>
<td>-0.00428**</td>
<td>-0.00792**</td>
<td>-0.00271**</td>
</tr>
<tr>
<td>(Destination area)</td>
<td>[0.00210]</td>
<td>[0.00396]</td>
<td>[0.00114]</td>
</tr>
<tr>
<td>Subway access</td>
<td>-0.00890***</td>
<td>-0.0143***</td>
<td>-0.00488***</td>
</tr>
<tr>
<td>(Destination area)</td>
<td>[0.00199]</td>
<td>[0.00409]</td>
<td>[0.00106]</td>
</tr>
<tr>
<td>Constant</td>
<td>4.988***</td>
<td>4.602***</td>
<td>4.930***</td>
</tr>
<tr>
<td></td>
<td>[0.187]</td>
<td>[0.178]</td>
<td>[0.339]</td>
</tr>
<tr>
<td>Observations</td>
<td>10958</td>
<td>11593</td>
<td>3668</td>
</tr>
<tr>
<td>R^2 (Pseudo)**</td>
<td>0.450</td>
<td>0.451</td>
<td>0.409</td>
</tr>
</tbody>
</table>

# Heteroskedasticity Robust Standard Errors are reported in square brackets. The symbols, *, **, and *** indicate respectively that the estimated coefficient is statistically significant under 10%, 5%, and 1% significance levels.

## Because Maximum Likelihood Estimation does not provide sum of squared error and sum of squared regression, unlike to Ordinary Least Squares, Coefficient of Determinations, called R^2, plays no longer role as a measure of explanatory power. We calculate, thus, Pseudo R^2 which is calculated Corr(y, \hat{y})^2, where \hat{y} is fitted value of dependent variable.

1 The estimates have a dependent variable trip frequency with 5 categories, which are 0 if respondent drive a car once or less than once a week, 1 if twice a week, 2 if three, four, or five times per week, 3 if four times, 4 if five times or more per week.
2 The model is failed to estimate, caused by which the maximized log likelihood does not converge.
3 The model is failed to estimate, caused by which the maximized log likelihood does not converge.
4 The Poisson regressions and the negative binomial regression are all applied right censored regression model estimation.
Table 1.3: Baseline Model Comparison 2

<table>
<thead>
<tr>
<th>Trip frequency (6 categories)</th>
<th>OLS</th>
<th>Poisson</th>
<th>Negative Binomial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>TOLL</td>
<td>-0.0343 ***</td>
<td>-0.0655 ***</td>
<td>-0.0608 ***</td>
</tr>
<tr>
<td></td>
<td>[0.00796]</td>
<td>[0.00900]</td>
<td>[0.0152]</td>
</tr>
<tr>
<td>Travel time (Log scale)</td>
<td>-0.0663 ***</td>
<td>-0.0138</td>
<td>0.0513</td>
</tr>
<tr>
<td></td>
<td>[0.0261]</td>
<td>[0.0238]</td>
<td>[0.0449]</td>
</tr>
<tr>
<td>Distance (Log scale)</td>
<td>-0.455 ***</td>
<td>-0.408 ***</td>
<td>-0.292 ***</td>
</tr>
<tr>
<td></td>
<td>[0.0232]</td>
<td>[0.0210]</td>
<td>[0.0659]</td>
</tr>
<tr>
<td>Trip purpose (Working dummy)</td>
<td>2.696 ***</td>
<td>2.671 ***</td>
<td>2.540 ***</td>
</tr>
<tr>
<td></td>
<td>[0.0315]</td>
<td>[0.0306]</td>
<td>[0.0542]</td>
</tr>
<tr>
<td>Income (Log scale)</td>
<td>-0.166 ***</td>
<td>-0.147 ***</td>
<td>-0.191 ***</td>
</tr>
<tr>
<td></td>
<td>[0.0218]</td>
<td>[0.0209]</td>
<td>[0.0399]</td>
</tr>
<tr>
<td>Bus access (Origin area)</td>
<td>-0.00201</td>
<td>-0.00522</td>
<td>-0.00148</td>
</tr>
<tr>
<td></td>
<td>[0.00307]</td>
<td>[0.00566]</td>
<td>[0.01114]</td>
</tr>
<tr>
<td>Subway access (Origin area)</td>
<td>-0.00229</td>
<td>-0.00304</td>
<td>-0.00141</td>
</tr>
<tr>
<td></td>
<td>[0.00284]</td>
<td>[0.00602]</td>
<td>[0.00104]</td>
</tr>
<tr>
<td>Bus access (Destination area)</td>
<td>-0.00789 ***</td>
<td>-0.0104 *</td>
<td>-0.00386 ***</td>
</tr>
<tr>
<td></td>
<td>[0.00293]</td>
<td>[0.00577]</td>
<td>[0.00112]</td>
</tr>
<tr>
<td>Subway access (Destination area)</td>
<td>-0.0129 ***</td>
<td>-0.0231 **</td>
<td>-0.00580 ***</td>
</tr>
<tr>
<td></td>
<td>[0.00279]</td>
<td>[0.00599]</td>
<td>[0.00104]</td>
</tr>
<tr>
<td>Constant</td>
<td>5.918 ***</td>
<td>5.567 ***</td>
<td>5.851 ***</td>
</tr>
<tr>
<td></td>
<td>[0.266]</td>
<td>[0.252]</td>
<td>[0.500]</td>
</tr>
<tr>
<td>Observations</td>
<td>10539</td>
<td>11226</td>
<td>3411</td>
</tr>
<tr>
<td>$R^2$ (Pseudo)##</td>
<td>0.480</td>
<td>0.4845</td>
<td>0.4118</td>
</tr>
</tbody>
</table>

# Heteroskedasticity Robust Standard Errors are reported in square brackets. The symbols, *, **, and *** indicate respectively that the estimated coefficient is statistically significant under 10%, 5%, and 1% significance levels.

## Because Maximum Likelihood Estimation does not provide sum of squared error and sum of squared regression, unlikely to Ordinary Least Squares, Coefficient of Determinations, called $R^2$, plays no longer role as a measure of explanatory power. We calculate, thus, Pseudo $R^2$ which is calculated $Corr(y, \hat{y})^2$, where $\hat{y}$ is fitted value of dependent variable.

1 The estimates have a dependent variable trip frequency with 6 categories, which are 0 if respondent drive a car once or less then once a week, 1 if twice a week, 2 if three times a week, 3 if four times a week, 4 if five times per week, 5 if five times or more per week.

4 The Poisson regressions and the negative binomial regression are all applied right censored regression model estimation.

1. TRAVEL DEMAND MODEL
Table 1.4: Baseline Model Comparison

<table>
<thead>
<tr>
<th></th>
<th>OLS (1)</th>
<th>OLS (2)</th>
<th>OLS (3)</th>
<th>Poisson (4)</th>
<th>Poisson (5)</th>
<th>Poisson (6)</th>
<th>Negative Binomial (7)</th>
<th>Negative Binomial (8)</th>
<th>Negative Binomial (9)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TOLL</strong></td>
<td>-0.0185***</td>
<td>-0.0322***</td>
<td>-0.0306***</td>
<td>-0.0102***</td>
<td>-0.0231***</td>
<td>-0.0210***</td>
<td></td>
<td>-0.0210***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.00426]</td>
<td>[0.00478]</td>
<td>[0.00803]</td>
<td>[0.00350]</td>
<td>[0.00382]</td>
<td>[0.00614]</td>
<td></td>
<td>[0.00614]</td>
<td></td>
</tr>
<tr>
<td><strong>Travel time</strong></td>
<td>-0.0417***</td>
<td>-0.0201</td>
<td>0.0284</td>
<td>-0.0190*</td>
<td>-0.00576</td>
<td>0.0286*</td>
<td></td>
<td>0.0286*</td>
<td></td>
</tr>
<tr>
<td>(Log scale)</td>
<td>[0.0140]</td>
<td>[0.0128]</td>
<td>[0.0237]</td>
<td>[0.0103]</td>
<td>[0.00977]</td>
<td>[0.0168]</td>
<td></td>
<td>[0.0168]</td>
<td></td>
</tr>
<tr>
<td><strong>Distance</strong></td>
<td>-0.247***</td>
<td>-0.222***</td>
<td>-0.158***</td>
<td>-0.139***</td>
<td>-0.124***</td>
<td>-0.0943***</td>
<td></td>
<td>-0.0943***</td>
<td></td>
</tr>
<tr>
<td>(Log scale)</td>
<td>[0.0127]</td>
<td>[0.0114]</td>
<td>[0.0346]</td>
<td>[0.00985]</td>
<td>[0.00928]</td>
<td>[0.0249]</td>
<td></td>
<td>[0.0250]</td>
<td></td>
</tr>
<tr>
<td><strong>Trip purpose</strong></td>
<td>1.312***</td>
<td>1.300***</td>
<td>1.197***</td>
<td>0.721***</td>
<td>0.711***</td>
<td>0.677***</td>
<td></td>
<td>0.677***</td>
<td></td>
</tr>
<tr>
<td>(Working dummy)</td>
<td>[0.0165]</td>
<td>[0.0159]</td>
<td>[0.0282]</td>
<td>[0.0126]</td>
<td>[0.0123]</td>
<td>[0.0217]</td>
<td></td>
<td>[0.0217]</td>
<td></td>
</tr>
<tr>
<td><strong>Income</strong></td>
<td>-0.0739***</td>
<td>-0.0671***</td>
<td>-0.0918***</td>
<td>-0.0431**</td>
<td>-0.0376**</td>
<td>-0.0537***</td>
<td></td>
<td>-0.0537***</td>
<td></td>
</tr>
<tr>
<td>(Log scale)</td>
<td>[0.0117]</td>
<td>[0.0112]</td>
<td>[0.0207]</td>
<td>[0.00909]</td>
<td>[0.00876]</td>
<td>[0.0159]</td>
<td></td>
<td>[0.0159]</td>
<td></td>
</tr>
<tr>
<td><strong>Bus access</strong></td>
<td>-0.00181</td>
<td>-0.00349</td>
<td>-0.00146</td>
<td>-0.00228</td>
<td>-0.00228</td>
<td>-0.00228</td>
<td></td>
<td>-0.00228</td>
<td></td>
</tr>
<tr>
<td>(Origin area)</td>
<td>[0.00164]</td>
<td>[0.00290]</td>
<td>[0.00128]</td>
<td>[0.00216]</td>
<td>[0.00216]</td>
<td>[0.00216]</td>
<td></td>
<td>[0.00216]</td>
<td></td>
</tr>
<tr>
<td><strong>Subway access</strong></td>
<td>0.000146</td>
<td>0.000846</td>
<td>0.000968</td>
<td>0.000962</td>
<td>0.000962</td>
<td>0.000962</td>
<td></td>
<td>0.000962</td>
<td></td>
</tr>
<tr>
<td>(Origin area)</td>
<td>[0.00149]</td>
<td>[0.00305]</td>
<td>[0.00117]</td>
<td>[0.00225]</td>
<td>[0.00225]</td>
<td>[0.00225]</td>
<td></td>
<td>[0.00225]</td>
<td></td>
</tr>
<tr>
<td><strong>Bus access</strong></td>
<td>-0.00271*</td>
<td>-0.00450</td>
<td>-0.00251**</td>
<td>-0.00341</td>
<td>-0.00341</td>
<td>-0.00341</td>
<td></td>
<td>-0.00341</td>
<td></td>
</tr>
<tr>
<td>(Destination area)</td>
<td>[0.00157]</td>
<td>[0.00300]</td>
<td>[0.00123]</td>
<td>[0.00220]</td>
<td>[0.00220]</td>
<td>[0.00220]</td>
<td></td>
<td>[0.00220]</td>
<td></td>
</tr>
<tr>
<td><strong>Subway access</strong></td>
<td>-0.00576***</td>
<td>-0.00875***</td>
<td>-0.00448***</td>
<td>-0.00652***</td>
<td>-0.00652***</td>
<td>-0.00652***</td>
<td></td>
<td>-0.00652***</td>
<td></td>
</tr>
<tr>
<td>(Destination area)</td>
<td>[0.00147]</td>
<td>[0.00307]</td>
<td>[0.0016]</td>
<td>[0.00233]</td>
<td>[0.00233]</td>
<td>[0.00233]</td>
<td></td>
<td>[0.00233]</td>
<td></td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>3.971***</td>
<td>3.822***</td>
<td>3.924***</td>
<td>1.822***</td>
<td>1.734***</td>
<td>1.806***</td>
<td></td>
<td>1.806***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.142]</td>
<td>[0.134]</td>
<td>[0.259]</td>
<td>[0.109]</td>
<td>[0.104]</td>
<td>[0.198]</td>
<td></td>
<td>[0.198]</td>
<td></td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>10539</td>
<td>11226</td>
<td>3411</td>
<td>10539</td>
<td>11226</td>
<td>3411</td>
<td></td>
<td>3411</td>
<td></td>
</tr>
<tr>
<td><strong>R² (Pseudo) #</strong></td>
<td>0.4426</td>
<td>0.4465</td>
<td>0.3665</td>
<td>0.4261</td>
<td>0.4340</td>
<td>0.3653</td>
<td></td>
<td>0.4261</td>
<td></td>
</tr>
</tbody>
</table>

# Heteroskedasticity Robust Standard Errors are reported in square brackets. The symbols, *, **, and *** indicate respectively that the estimated coefficient is statistically significant under 10%, 5%, and 1% significance levels.

## Because Maximum Likelihood Estimation does not provide sum of squared error and sum of squared regression, unlikely to Ordinary Least Squares, Coefficient of Determinations, called $R^2$, plays no longer role as a measure of explanatory power. We calculate, thus, Pseudo $R^2$ which is calculated $\text{Corr}(y, \hat{y})^2$, where $\hat{y}$ is fitted value of dependent variable.

1 The estimates have a dependent variable trip frequency with 4 categories, which are 0 if respondent drive a car once or less then once a week, 1 if twice a week, 2 if three times per week, 3 if four times or more per week.
2 The model is failed to estimate, caused by which the maximized log likelihood does not converge.
3 The model is failed to estimate, caused by which the maximized log likelihood does not converge.
4 The Poisson regressions and the negative binomial regression are all applied right censored regression model estimation.
<table>
<thead>
<tr>
<th>Trip frequency #</th>
<th>OLS</th>
<th>Poisson</th>
<th>Negative Binomial</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3 categories)</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>TOLL</td>
<td>-0.0118***</td>
<td>-0.0221***</td>
<td>-0.0225***</td>
</tr>
<tr>
<td></td>
<td>[0.00317]</td>
<td>[0.00358]</td>
<td>[0.00595]</td>
</tr>
<tr>
<td>Travel time</td>
<td>-0.0244**</td>
<td>-0.00437</td>
<td>0.0150</td>
</tr>
<tr>
<td>(Log scale)</td>
<td>[0.0101]</td>
<td>[0.00926]</td>
<td>[0.0171]</td>
</tr>
<tr>
<td>Distance</td>
<td>-0.180***</td>
<td>-0.147***</td>
<td>-0.122***</td>
</tr>
<tr>
<td>(Log scale)</td>
<td>[0.00918]</td>
<td>[0.00849]</td>
<td>[0.0254]</td>
</tr>
<tr>
<td>Trip purpose</td>
<td>0.937***</td>
<td>0.912***</td>
<td>0.849***</td>
</tr>
<tr>
<td>(Working dummy)</td>
<td>[0.0123]</td>
<td>[0.0119]</td>
<td>[0.0209]</td>
</tr>
<tr>
<td>Income</td>
<td>-0.0663***</td>
<td>-0.0542***</td>
<td>-0.0717***</td>
</tr>
<tr>
<td>(Log scale)</td>
<td>[0.00860]</td>
<td>[0.00830]</td>
<td>[0.0151]</td>
</tr>
<tr>
<td>Bus access</td>
<td>-0.00195</td>
<td>-0.00370*</td>
<td>-0.00188</td>
</tr>
<tr>
<td>(Origin area)</td>
<td>[0.00122]</td>
<td>[0.00219]</td>
<td>[0.00148]</td>
</tr>
<tr>
<td>Subway access</td>
<td>-0.000663</td>
<td>0.000396</td>
<td>-0.00154</td>
</tr>
<tr>
<td>(Origin area)</td>
<td>[0.00112]</td>
<td>[0.00227]</td>
<td>[0.00135]</td>
</tr>
<tr>
<td>Bus access</td>
<td>-0.00303**</td>
<td>-0.00581***</td>
<td>-0.00326**</td>
</tr>
<tr>
<td>(Destination area)</td>
<td>[0.00118]</td>
<td>[0.00223]</td>
<td>[0.00144]</td>
</tr>
<tr>
<td>Subway access</td>
<td>-0.00502***</td>
<td>-0.00752***</td>
<td>-0.00518***</td>
</tr>
<tr>
<td>(Destination area)</td>
<td>[0.00112]</td>
<td>[0.00229]</td>
<td>[0.00144]</td>
</tr>
<tr>
<td>Constant</td>
<td>3.071***</td>
<td>2.841***</td>
<td>3.060***</td>
</tr>
<tr>
<td></td>
<td>[0.103]</td>
<td>[0.0992]</td>
<td>[0.188]</td>
</tr>
<tr>
<td>Observations</td>
<td>10539</td>
<td>11226</td>
<td>3411</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.4232</td>
<td>0.4089</td>
<td>0.3479</td>
</tr>
</tbody>
</table>

# Heteroskedasticity Robust Standard Errors are reported in square brackets. The symbols, *, **, and *** indicate respectively that the estimated coefficient is statistically significant under 10%, 5%, and 1% significance levels.
# # Because Maximum Likelihood Estimation does not provide sum of squared error and sum of squared regression, unlikely to Ordinary Least Squares, Coefficient of Determinations, called $R^2$, plays no longer role as a measure of explanatory power. We calculate, thus, Pseudo $R^2$ which is calculated $Corr(y, \hat{y})^2$, where $\hat{y}$ is fitted value of dependent variable.

1 The estimates have a dependent variable trip frequency with 3 categories, which are 0 if respondent drive a car once or less then once a week, 1 if more than once a week and less than four times a week, 2 if more than four times per week.
2 The model is failed to estimate, caused by which the maximized log likelihood does not converge.
3 The model is failed to estimate, caused by which the maximized log likelihood does not converge.
4 The Poisson regressions and the negative binomial regression are all applied right censored regression model estimation.
Table 1.6: Negative Binomial Estimation

<table>
<thead>
<tr>
<th></th>
<th>Mass transit (Beginning)</th>
<th>Mass transit (Terminal)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Employer pay</td>
<td>Self pay</td>
</tr>
<tr>
<td>Toll</td>
<td>-0.0217 $$^*$$</td>
<td>-0.0212 $$^*$$</td>
</tr>
<tr>
<td></td>
<td>[0.00274]</td>
<td>[0.00274]</td>
</tr>
<tr>
<td>Travel time</td>
<td>-0.0718 $$^*$$</td>
<td>-0.0705 $$^*$$</td>
</tr>
<tr>
<td>(Log scale)</td>
<td>[0.00675]</td>
<td>[0.00674]</td>
</tr>
<tr>
<td>Distance mile</td>
<td>-0.207 $$^*$$</td>
<td>-0.206 $$^*$$</td>
</tr>
<tr>
<td>(Log scale)</td>
<td>[0.00575]</td>
<td>[0.00575]</td>
</tr>
<tr>
<td>Trip purpose</td>
<td>1.262 $$^*$$</td>
<td>1.264 $$^*$$</td>
</tr>
<tr>
<td>(Working dummy)</td>
<td>[0.00875]</td>
<td>[0.00879]</td>
</tr>
<tr>
<td>Income</td>
<td>-0.0896 $$^*$$</td>
<td>-0.0853 $$^*$$</td>
</tr>
<tr>
<td>(Log scale)</td>
<td>[0.00621]</td>
<td>[0.00622]</td>
</tr>
<tr>
<td>Employment</td>
<td>0.0398 $$^*$$</td>
<td>0.0394 $$^*$$</td>
</tr>
<tr>
<td>(Household)</td>
<td>[0.00467]</td>
<td>[0.00466]</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.0344 $$^*$$</td>
<td>0.0353 $$^*$$</td>
</tr>
<tr>
<td>(Household)</td>
<td>[0.00498]</td>
<td>[0.00497]</td>
</tr>
<tr>
<td>Payer (Employer)</td>
<td>-0.0601 $$^*$$</td>
<td></td>
</tr>
<tr>
<td>Payer (Self)</td>
<td>-0.144 $$^*$$</td>
<td></td>
</tr>
<tr>
<td>Bus access (Beginning area)</td>
<td>0.00152</td>
<td>-0.00157</td>
</tr>
<tr>
<td>Subway access (Beginning area)</td>
<td>0.00408</td>
<td>-0.000854</td>
</tr>
<tr>
<td>Bus access (Terminal area)</td>
<td>0.00249</td>
<td>0.00476 $$^*$$</td>
</tr>
<tr>
<td>Subway access (Terminal area)</td>
<td>0.01015</td>
<td>-0.00542 $$^*$$</td>
</tr>
<tr>
<td>Constant</td>
<td>1.189 $$^*$$</td>
<td>1.076 $$^*$$</td>
</tr>
<tr>
<td></td>
<td>[0.0507]</td>
<td>[0.0547]</td>
</tr>
<tr>
<td>Observations</td>
<td>36756</td>
<td>36703</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.5072</td>
<td>0.5088</td>
</tr>
</tbody>
</table>

# Heteroskedasticity Robust Standard Errors are reported in square brackets. The symbols, $^*$, $^{**}$, and $^{***}$ indicate respective coefficient is statistically significant under 10%, 5%, and 1% significance levels.

## Because Maximum Likelihood Estimation does not provide sum of squared error and sum of squared regression, unlikely to Ordinary Least Squares, Coefficient of Determinations, called $R^2$, plays no longer role as a measure of explanatory power. We calculate, thus, Pseudo $R^2 = Corr(y, \hat{y})^2$, where $\hat{y}$ is fitted value of dependent variable.

1. The estimates have a dependent variable trip frequency with 5 categories, which are 0 if respondent drive a car once or less than once a week, 1 if twice a week, 2 if three, four, or five times per week, 3 if four times, 4 if five times or more per week.

2. The Poisson regressions and the negative binomial regression are all applied right censored regression model estimation.
1.4 Conclusion

This study found that a model based on household choice behavior was consistent with the observed behavior of toll facility users in New York City in model estimation in which a model of count (frequency of use) of mode choices was used to analyze the behavior of toll facility users. By using the New York City toll facility user survey, it is found that the empirical estimation of lifecycle-based frequency models produced consistent results and had structural relationships that were stable across model specification. Although the behavior of toll facility users may be somewhat different than the users of other modes of travel, in that these users tend to be more automobile dependent as well as generally have significantly higher income as compared to the average resident of the communities that surround these facilities, the interest in utilizing pricing as a tool to manage demand on transportation facilities makes their behavior of significant interest. In addition, the models specified in this work can be logically extended to explore the behavior of other modes of travel (transit users, bicycle users and walkers).

A key component of our analysis was the ability to separate the behavior of self-paying users who had their road fees reimbursed. In fact, the structural relationships identified indicated a wide amount of variation in consumer demand depending on whether the user had to self-pay or had an employer who paid the toll.

Transit service availability was an important determinant of toll facility use in New York City. In particular, mass transit facility availability at the destination ZIP Code significantly reduced the frequency of toll facility use. The majority of the trips in this sample originate at the home ZIP Code of the user. As such, it is likely that the transit availability at the destination ZIP Code is more important than the origin ZIP Code as commercial locations may be more able to cluster around transit facilities as well as use a higher density land use pattern than residential areas.
such, destination ZIP Code transit availability is a key variable in the mode decisions.

The theoretical extension of the literature contained in this work, coupled with the fact that the statistical fit the theoretical models have high explanatory power, further analysis of additional regions and toll facilities is warranted. Models that include regional commuter rail service in addition to urban metro and local bus service would be an interesting extensions of this work.
2 Electronic Toll Collection Systems and Travel Demand: A Field Experiment of Toll Facilities in New York City

with Jonathan R. Peters

Reprinted with permission from

2. ETC SYSTEM AND TRAVEL DEMAND

2.1 Introduction

Installing a toll facility on road networks has been utilized for a long period time in the United States for funding transportation systems. Since locations of those toll facilities and toll prices vary over toll facilities, even for the same facilities within the same area, transportation authorities seem to have expanded the purpose of toll collection in that, it is not only for maintaining and operating road networks, but also for reducing congestion and some other environmental problems which might be caused by congestion and traffic overflow. In fact, the transportation authorities’ public press announcements and demand research indicates that their toll pricing and planning aim to address both issues. Congestion is rapidly becoming the most important concern for highway planning because of rapidly increasing social cost of congestion.

According to various transportation expert reports, total miles traveled (VMT) in the United States has seen a 111 percent increase in urban areas, and traffic congestion due to these same VMT increases wastes approximately 2.3 billion gallons of gasoline and causes 3.7 billion hours of delay each year. This tremendous amount of wasted resources has created a significant social cost of congestion that was valued at about $87.2 billion in 2007.\footnote{We quote these transportation statistics from Currie and Walker (2011), Vera and Preziosi (2011). Those two papers are well summarizing congestion facts and sources of the information in their introduction. See Currie and Walker (2011), and Vera and Preziosi (2011) for further information.} To reduce the congestion problem, toll facilities in some cases have charged different prices to users in different locations to use prices as an economic control on demand. From a user’s perspective, however, toll facilities are a different form of taxation in that takes their disposable income so that it may also reduce consumer welfare and creates deadweight loss, which is another social cost. Hence, toll payment price structures designed as price control for congestion becomes sensitive political issues because of the tradeoff between reducing congestion
2. ETC SYSTEM AND TRAVEL DEMAND

and losing social welfare.

To reduce the social costs of tolling, a number of researchers argue that an Electronic Toll Collection system is an effective alternative to deal with problems caused by congestion and delay in collecting tolls. The electronic toll collection system makes toll facilities more productive by reducing travel time and the time spent the toll payment process. The net effect is the same as reducing congestion without increasing toll price. For this reason, the effect of electronic toll collection system on travel demand has become an active research field and there are a large number of papers that investigate the effectiveness of electronic toll collection system on the various aspects of road prices. From the user perspective, Vera and Preziosi (2011) investigated what causes travelers to become electronic toll collection users. They perform empirical analysis of a E-ZPASS user survey, the name of electronic toll collection system used by U.S. Northeastern region, at the toll facilities operated by the Port Authority of New York and New Jersey to find a causal relationship between user satisfaction level and covariates that includes travel attributes and socio-economic status. Vera and Preziosi (2011) also found that reduced travel time by using E-ZPASS does not effect the probability of becoming a user, but the financial benefit, from E-ZPASS price has a significant effect on this same probability.

From environmental perspective, Currie and Walker (2011) performed a field experiment that investigates the relationship between traffic congestion and residents’ health condition around toll facilities. They perform a comparison study of the fraction of lower birth weight babies, and premature birth between the treatment group who are the residents within a 2 kilometer radius boundary of a toll plaza, and a control group who live close to highways but not in 2 km boundary, using Vital Statistics Natality records of Pennsylvania and New Jersey residents. They find a statistically significant positive effect of E-ZPASS for the residents in 2 km boundaries of toll plazas on infant health conditions and their analysis indicates that the the result is
due to reducing congestion at toll plazas and this makes the air pollution problems better around these same toll facilities.

On the taxation and social welfare perspective, Finkelstein (2009) argues that electronic toll collection system makes users less sensitive to the price of toll as compares to cash users. Their paper uses the toll facility usage data from 123 facilities, operated by 49 different authorities in 22 states for 20 years, from 1985 to 2005 to perform an empirical analysis that explore causal relationship between toll price and electronic toll collection use. Its empirical results show that toll facilities with electronic toll collection system tended to have greater price increases than the other facilities because electronic toll collection system makes toll price, as a salience of tax, less visible to users, and thus the users become less elastic to changes in toll price. Consequently the toll facilities can increase prices without losing demand.

Since the implementation of the electronic toll collection on transportation facilities broadly has been done mostly in the last two decades, those of studies depend highly on the empirical evidence they found because of absence of rigorous economic theory, and thus must consider the possibility to that their statistical inferences are biased and inconsistent. Most of them use linear model estimation to investigate differences of some variables of interest between the users and non-user or electronic tolling, called a difference in difference estimator, and it is well known fact that there are several sources of bias that can cause inconsistent estimation of difference-in-difference models. These include sample selection, failure to satisfy the exogeneity assumption, and model misspecification. The most frequently ignored source of bias that field experimental researchers miss out of three, is model misspecification. Even though it seems obvious that electronic toll collection systems would have an effect on is travel demand, it is hard to find empirical analysis that carefully examine travel demand in a baseline model that looks to estimate the effect consistently.

The absence of concern about theory may lead to the model being incorrectly
specified, thus, causing not only inconsistent estimation, but also yielding scientifically meaningless implications. As Deaton (2010) points out, criticisms of the strong dependency of development economics on Randomized Controlled Trials, an econometric method to estimate differences in differences parameter by collecting randomized samples, or manipulating data as randomized sample from non or quasi random sample, “experiment itself have no special ability to produce more credible knowledge than other methods”. In other words, Randomized Controlled Trials of types of fully empirical studies are subject to issues when we generalize out of the sample it uses because it depends too much on inductive reasoning that has to come up with a perfectly controlled situation that would need to be replicated again. Otherwise, the argument, given by the studies, will be isolated into their sample, or perfectly similar objects.

Thus, we examine all possible sources of bias in measuring the effect of electronic toll collection system on travel demand, then we discuss the consequence of the estimation under identification problems then of this bias as fully as we can. We use Triborough Bridges and Tunnel Authority’s user survey in 2004 to estimate the effect of E-ZPASS on travel demand of New York City metropolitan toll bridges and tunnel users, as explored in Chapter 1. We then design an experiment that captures differences in travel demand between EZ-PASS users and non-users based on a trip frequency travel demand model, As given by Chapter 1, it can be treated as a field experiment on the New York City toll facilities. From this experiment, we find serious upward bias given by misspecification, such as omitted variables, linear model specification problems. We find that the statistical significance of the estimate

\[2\] The papers, Deaton (2010) and Imbens (2010) argue about the controversial issue as to whether Randomized Controlled Trials can provide scientifically meaningful implication or not. Deaton (2010) criticizes the modern trend in development economic literature that highly depends on experimental result given by statistical inference without considering the economic theory that helps to find the essential mechanism that produces the outcome. Standing for the opposite side of Deaton (2010), Imbens (2010) argues that, because Randomized Controlled Trials concern a lot about statistical theory, it deserves reasonable consideration to provide to scientific implications, although economic theory does not play a role in that argument.
Figure 2.1: Locations of TBTA Toll Facilities

is thus eliminated by controlling for the sources of bias. In section 2, we introduce New York City toll facilities as a field test of our experiment, using the Metropolitan Transportation Authority’s (MTA) toll pricing policy and the consequence of those prices on user behavior. In section 3, we review a theoretical travel demand model to design an experiment that are representative of our methods.

2.2 Toll Facilities of New York City: A Field Experiment

In this section, we introduce the characteristics of electronic toll collection systems in the United States. In particular, E-ZPass, the name of the electronic toll collection
Table 2.1: Toll Price of MTA bridges and Tunnels)

<table>
<thead>
<tr>
<th>Facilities</th>
<th>General E-ZPass</th>
<th>General Cash</th>
<th>Residents E-ZPass</th>
<th>Residents Cash</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bronx-Whitestone</td>
<td>$4.80</td>
<td>$6.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brooklyn-Battery Park</td>
<td>$4.80</td>
<td>$6.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross Bay Veterans</td>
<td>$1.80</td>
<td>$3.25</td>
<td>$1.19</td>
<td>$1.62</td>
</tr>
<tr>
<td>Henry Hudson</td>
<td>$2.20</td>
<td>$4.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marine Parkway-Gil Hodges</td>
<td>$1.80</td>
<td>$3.25</td>
<td>$1.19</td>
<td>$1.62</td>
</tr>
<tr>
<td>Queens Midtown</td>
<td>$4.80</td>
<td>$6.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Throgs Neck</td>
<td>$4.80</td>
<td>$6.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triborough, Bronx</td>
<td>$4.80</td>
<td>$6.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triborough, Manhattan</td>
<td>$4.80</td>
<td>$6.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verrazano Narrow</td>
<td>$9.60</td>
<td>$13.00</td>
<td>$5.76</td>
<td>$7.72</td>
</tr>
</tbody>
</table>

Notes: The official name of Triborough Bridge is Robert F. Kennedy Bridge. The resident discount of Cross Bay, and Marine Parkway bridges are given to Rockaway, Queens’s resident. The Verrazano Narrow Bridge’s discount is for Staten Island resident. The toll prices are for 2-axle passenger vehicles under 7,000 lbs maximum gross weight, effective from December 30, 2010.

Electronic Toll Collection System in the United States

Electronic Toll Collection (ETC) systems are automatic toll payment systems that electronically debits via radio transponder a certain amount of toll costs from users’ account when the users are passing through an ETC toll plaza when the vehicle is equipped with payment equipment in their vehicles. ETC system are designed for two purposes, First, to reduce delay at toll plazas, and second to allow a toll authority
to price differently over users based on their individual characteristics. Because ETC system can identify user characteristic and charge a certain amount of toll based on these characteristics, transportation authorities are thus able to charge different amounts of toll for different users in terms of residence area, vehicle types and so on. Thus, transportation authorities all over the world have been installed ETC toll plazas not only for engineering purpose of improving travel flow in that eliminate time delay at toll plazas but also for the economic purpose that they can charge different toll price to different users. The authorities’ prices are normally aiming to reduce congestion in a particular area that can be managed by passing through their toll facilities. ETC system in the U.S. takes several different names in different group of states, and on different highway systems. EZ-Pass is deployed in the for Northeastern states including New York, New Jersey, Pennsylvania, and is also usable on other systems such as Fast Lane on Massachusetts, I-PASS for Illinois, i-Zoom in Indiana and FasTrak in California.

2.2.2 E-ZPass in Detail

E-ZPass is the name of the ETC system installed on toll facilities operated by the transportation agencies who are members of partiners of E-ZPass Interagency Group (IAG). E-ZPass was developed by the New York State Thruway Authority (NYSTA) and was first installed along the highways, operated by NYSTA. In 1991, IAG, consisted of seven independent toll agencies—The Port Authority of New York and New Jersey, The New Jersey Turnpike Authority, The New Jersey Highway Authority (operator of the Garden State Parkway at the time), the New York Metropolitan Transportation Authority, the New York State Thruway Authority, The Pennsylvania Turnpike Commission, and the South Jersey Transportation Authority (operator of the Atlantic City Expressway formed an operational consortium) Since the first E-ZPass was installed at the Spring Valley toll plaza in August 3, 1993, the system
has become the most well-known ETC system in the United States and it covers the widest ranges of highways over the United States. The toll facilities of MTA toll bridges and tunnels, that we are considering to perform an experiment, have also implemented the E-ZPass system and have used it for more than 10 years.

2.2.3 Toll Prices of MTA Bridges and Tunnels

There are 8 bridges and 2 tunnels are operated by Triborough Bridges and Tunnel Authority as toll facilities in New York Metropolitan area. Those are all located in the 5 boroughs in the New York City. Six bridges, the Bronx-Whitestone, Cross Bay, Henry Hudson, Marine Parkway, the two Triborough Bridges, and Queens Midtown tunnel all started to collect cash tolls by the end of 1930. The electronic tolling followed with the Verrazano Narrow Bridge implemented E-ZPass in 1995, and the rest of the toll plazas implementing ETC in 1996. The prices of tolls vary over facilities, and vary also by residence area, and vehicle types. Table 2.1 summarizes toll prices of MTA bridges and tunnels. As we seen, Verrazano Narrow Bridge charges the greatest amount of toll for all types of users, and there is also controversial that the toll price of Verrazano Narrow Bridge takes a significant amount of consumer surplus from residents of Staten Island, one of five boroughs in New York City, because the bridge has no alternative routes which are toll free, and mass transit access is relatively poor compare with the other New York City boroughs.

\[\text{Note that Verrazano Narrow Bridge is one way toll that charges the vehicles that across the bridge from Kings county (Brooklyn) to Richmond county (Staten Island).}\]
2.3 Theory of Travel Demand and Experimental Design

In this section, we discuss the econometric methods we look to use to demonstrate the effect of electronic toll collection system on the demand for driving. Since our experimental design is based on the question as to whether people tend to travel more when they are electronic toll collection system users, the method has to be an appropriate comparison study between treatment and control group. To find consistent estimation of an effect of a policy intervention, a number of empirical research methods in general have been developed not only in economics, but in social science and in other areas. And each of which has faced controversy about validity of the method from the perspective of statistical inference, for a number of reasons. First, the necessity of having a randomized sample, which is well collected without sampling bias, had turned out to be an issue since Lalonde (1986) shows that a well-controlled set of experimental data can yield accurate estimation even though it’s number of observations are relatively small as compared to than non-experimental data. Endogenously assigned treatment and control groups are also a problematic point that may cause inconsistent estimation of a comparison study. The other issue is the existence of implicit categories that a given sample can fall into but are also not observable\footnote{This is well known issue on comparison study by the name selection bias due to counterfactual outcomes. If the assignment of unobservable groups, to be “Always taker” or “Never taker”, has been done endogenously, then the assignment effects on estimation of average treatment effect. In other words, if individuals in an experiment can decide their status to be in unobservable groups, then the estimate of average treatment effect will be biased. We can consider two counterfactual groups in this study, a group of individuals who have electronic toll collection equipment but pay toll by cash, and individuals who have no equipment but pay by the equipment. We assume over this study that, however, no counterfactual groups since the second group is not possibly exist and the toll users have no incentive to be in the first group.} Endogenous assignment happens when individuals in an experiment can select their status whether to be in treatment group or control group, and implicit categories happen when the individual can decide their status to always be in treatment group,
called “always taker”, or to be in control group, called “never taker”. Those are the issues that we need to consider on our analysis as a comparison study between electronic toll collecting system users and non-users. We review in this section the travel demand model, proposed by Chapter 1, and design an experiment to perform an empirical comparison of travel demand between ETC users and non ETC users, then we look to develop an appropriate econometric method for the empirical comparison.

2.3.1 Travel Demand: Review

For the purpose of this study, when we are comparing travel demand between users and non-users, it is very important to use a correctly specified demand model for travel even though this study is more focused on empirical observation rather than theoretical concreteness. The most widely used demand model for travel is the mode choice model, proposed by McFadden (1974), which is based on the Random Utility Model. McFadden’s travel mode choice model assigns probabilities that are associated with the attributes to each travel mode such as walking, driving or mass transit. The assigned probabilities in the model implies how much an individual is willing to take a particular mode, given by a certain value of attributes. Because the mode choice

![Diagram of Travel Modes](image-url)

Figure 2.2: Diagram of Travel Modes
model takes probabilistic approach, it is better specified to analyze individual travelers decisions to choose a mode. But for the decision problem that choose number of travel events via a particular mode, these types of choices could not be captured by the mode choice model because it does not provide the number of travel events directly from the model.\footnote{Probability approach of demand analysis provides how much an individual has willingness to choose a mode, out of several alternatives because the demand for each mode is represented by probability that is associated with its covariates. But the question that how many times the individual will choose that mode again in a period of time such as per week, per year, is lying on a different dimension. The individual’s decision might be changed if she knows that she will face the same mode choice again on a certain time interval. In travel demand, for example, one shot decision of travel mode for sightseeing and repeated decision for work can never be treated as the same. This is the reason why we consider travel demand as frequency rather than probability.}

To design a field experiment that compares the demand for driving between electronic toll users and non-users, it is better to observe the number of travel demand events, rather than the probability of having a trip occur because driving demand events are actually countable numbers within a certain period of time, since being an electronic toll collection user is not a decision that is made for a one-shot event, thus we are looking at multiple consumption events. It is rate for travelers to purchase electronic toll collection equipment for a single trip. Travel demand modeled as a frequency of trip was proposed by Chapter 1. The model modifies the mode choice model to a frequency of trip on a particular mode by isolating multiple travel modes, given by Random Utility Maximization, into a single mode. Because the model is derived from a household’s lifetime consumption-leisure utility maximization, travel demand in the model implies the optimal number of trip via a particular travel mode that supports general economic activities such as consumption, leisure, and labor at a time period. For those of reasons, we expect that applying the trip frequency model to perform better for this comparison study as compared to the mode choice model.

The trip frequency model, given by Chapter 1, is poisson random variable that is drawn according to a conditional probability distribution function that is associated with a vector of attribute $x = [w^*, r, k, h_i, d, l]$, where the elements are wage given by trip, unit return on capital, capital, travel time of mode $i$, distance of travel, and
leisure time respectively. The conditional pdf is:

\[
f(y_i | x) = \exp \left[ -E \{ \nu (h_i', z_i') \} \right] \frac{E \{ \nu (h_i', z_i') \}^{y_i}}{y_i!}, y_i = 0, 1, \ldots
\]  

(2.1)

where \( y_i \) is trip frequency in the current period, \( h_i', z_i' \) are travel time and stochastic shock on travel time respectively when the household takes a travel mode \( i \). We can thus apply the travel demand model to analyze the effect of electronic toll collection system on driving demand by taking a mode \( i \) as the driving mode on a toll road.

2.3.2 Average Treatment Effect on Travel Demand via Driving

Suppose that individuals in the model are assigned into treatment and control group. Let \( y_i^1 \) be travel demand of driving events by treatment group which is for electronic toll collection users, and let \( y_i^0 \) be a travel demand for non-users. Define a category variable \( D_i \) that takes value 1 if an individual \( i \) is included treatment group, 0 if \( i \) is in control group. Now we have the average treatment effect on the demand for driving as:

\[
\delta = E \left[ y_i^1 | D_i = 1 \right] - E \left[ y_i^0 | D_i = 0 \right].
\]  

(2.2)

By considering travel model for trip frequency which we discussed in the previous section, we thus have the average treatment effect with correctly specified form of expectation:

\[
\delta_{ATE} = E \left[ y_i^1 | x_i, D_i = 1 \right] - E \left[ y_i^0 | x_i, D_i = 0 \right],
\]  

(2.3)

\( \nu(\cdot, \cdot) \) is the functional equation that maximize a stochastic dynamic programming. Note that the stochastic dynamic programming is given by household maximization problem. See chapter 1.

\( \nu \) The subscript \( i \), all over this section, implies an individual \( i \) in the experiment, not a particular travel mode.
where \( \mathbf{x} \) is a vector of attributes. Since \( y_i^1, y_i^0 \) are drawn according to the identical probability distribution (2.1) which is a conditional poisson distribution, we can have the estimates of \( E [y_i^1 | \mathbf{x}_i, D_i] \) via a count variable model such as a poisson regression or negative binomial regression, and the coefficient of category variable \( D_i \) is the estimate of the average treatment effect \( \delta_{ATE} \) itself.

Because \( y_i^1, y_i^0 \) are normally unobservable which thus implies that we cannot predict when a particular individual \( i \) is an electronic toll collection user and when she is a non ETC user, however, \( \delta_{ATE} \) cannot be estimated directly unless we observe each individual’s travel demand behavior twice once when they are non-users and once when they turn to be ETC users. In general, data for this type of experiment consists of a treatment and a control group which are completely non-overlapped along individuals. Let \( y_i \) be an observed trip frequency via driving. Then we redefine the average treatment effect with the frequency as:

\[
\delta_{ATE} = E [y_i | \mathbf{x}_i, D_i = 1] - E [y_i | \mathbf{x}_i, D_i = 0],
\]

and \( \delta_{ATE} \) is able to be estimated by performing a count variable model estimation of \( E [y_i | \mathbf{x}_i, D_i] \).

2.3.3 Propensity Score Matching

The next issue is how to estimate \( \delta_{ATE} \) consistently. As Lalonde (1986) shows, simple model estimation is able to achieve consistency if the data is collected under a randomize experiment. Non-experimental data can be treated as experimental data, even if the given data is collected under non-randomized, or quasi randomized experiment, only when the condition that the treatment and control group assignment is independent of the outcome, is satisfied. In other words, individuals in the experiment are assigned randomly into one of the two groups, treatment or control, that is \( \{y_i^1, y_i^0 \perp D_i\} \lor \mathbf{x} \). This assumption is hard to accept in our experiment, however, be-
cause of an obvious fact that individuals who are frequent traveler would have higher willingness to be electronic toll collection users and low demand users are likely to be cash payers. This assignment $D_i$ is thus not independent of frequency of travel, given travel attributes. In addition, the assignment $D_i$ is a function of attributes because individuals are able to determine their status whether to be users or non-users, based on $x$. Hence we can say that the estimate of average treatment effect, without considering selection bias due to the effect of counterfactual outcome, is a biased estimator.

Dehejia and Wahba (1999) developed the propensity score matching method that is able to estimate an average treatment effect consistently even though the assignment $D_i$ is not exogenous and independent of outcome variable $y_i$. The propensity score matching method is basically designed to compute an average difference of outcome variable $y_i$ between an individual in treatment group and outcome variable $y_i$ of the individual who has the closest propensity score, the estimated probability to be in the treatment, in the control group. The computed average difference of matched sample is therefore a consistent estimate of average treatment effect if the condition $\{y_i^1, y_i^0 \perp D_i\} \cup x_i$ holds. And if $p(x_i)$ is the correctly specified propensity score, then the distribution of attribute across treatment and control groups are identical. The two conditions are the reason why we need to match up the sample from treatment to control which share closed propensity score. By using matched samples that share similar propensity scores, the estimated probability distribution of attributes will satisfy the independent of assignment to treatment category so that its calculated average is the consistent estimator of $\delta_{ATE}$.

\footnote{Rosenbaum and Rubin (1983) shows that the independent assignment assumption, conditional on propensity score satisfies for all sample if the independent assignment assumption, conditional on attribute variables $x$. That is: $\{y_i^1, y_i^0 \perp D_i\} \cup x_i \Rightarrow \{y_i^1, y_i^0 \perp D_i\} \cup p(x_i), \forall i = 1, \ldots, N$. Dehejia and Wahba (1999) applies the above proposition to the average treatment effect that is associated with propensity score. The average treatment effect can consistently be estimated with observable outcome variable $y_i$ by conditioning propensity score, rather than the attribute vector itself. See Dehejia and Wahba (1999).}
The procedure to calculate the propensity matched average treatment effect is as follows. First, create strata for both the treatment and control groups based on the estimated propensity scores. The propensity score $p(x_i)$ is then obtained by estimating a discrete choice model: $Pr[D_i = 1 | x_i] = G(x_i'\beta)$, where $G(\cdot)$ is link function that can be either logistic or probit.\footnote{Notice that the propensity score is a measure of willingness to be in treatment group, so that it can be obtained the estimate of probabilistic choice model $Pr[D_i = 1 | x_i]$, where dependent variable is a dummy for treatment group. In this study, $D_i$ is dummy variable for electronic toll collection users, and the propensity score is the measure of willingness to be an electronic toll collection user.}

Second, eliminate observations in the control group, that are located outside treatment group’s the support range of propensity score\footnote{Since the main focus is to match the samples in the treatment group to the samples in the control groups that belong to the same strata of the propensity score, the samples in the control group that are out of the treatment group’s propensity score’s interval are useless. Stata provides the propensity score matching procedure as an external module with the name “psmatch2".}. Then we compute the difference in mean of treatment and control outcomes within each stratum, weighted by the number of treatment samples in the stratum.

### 2.4 Empirical Analysis

So far we have reviewed the econometric methods that can be used to estimate the average treatment effect of travel demand between electronic toll collecting system users and non-users. We present, in this section, results from various ways to estimate the average treatment effect of the electronic toll collection using The Triborough Bridge and Tunnel Authority user survey, the same data as chapter 1 used to test the theoretical travel demand model, they derived, empirically. The results show that the simple differences in differences estimations of the average treatment effect by adding a user dummy in the travel demand model, given by chapter 1, are statistically significant in overall, but the signs are dramatically changed when different combinations of attributes, and independent variables are considered. In addition, propensity score matching yield insignificant estimates of the average treatment effect for any combi-
2. ETC SYSTEM AND TRAVEL DEMAND

(a) ZIP Code Polygon
(b) ZIP Code Polygons and its centroids
(c) ZIP Code Polygon
(d) ZIP Code Polygons and its centroids

Figure 2.3: GIS Characteristics of New York State Transportation System

nations of attributes. Therefore, we can conclude that the toll payment method does not have an effect on travel demand but that users are affected solely by the difference in toll price between cash and ETC payment.

2.4.1 Data

The data that we use in this study is based on chapter 1 travel demand data. The paper used the Triborough Bridge and Tunnel Authority (TBTA) user survey. The TBTA operates 10 bridge and tunnel facilities around the New York City metropolitan area. The survey contains detailed addresses of origin and destination area, a number of self reported socio-economic status variable, origin-destination based travel time,
type of toll payment, frequency of trip per week, and so on. In the same way as
chapter 1, we use the number of bus lines and the subway stations within a zip code
area as measures of mass transit access in a particular area. We also use toll price for
each users by considering resident plans, direction of travel, payment method.

To obtain a more accurate measure of distance, we then calculate network distance
of each origin-destination area. The top-left panel of Figure 2.3 represents the ZIP
Code polygons of New York State, and top-right panel represent centroid points of
these same polygons. The distance in miles that we use in this study thus is the
measure of the distance between a centroid of the origin ZIP Code polygon and the
centroid of the destination ZIP Code polygon. The two panels at the bottom in
Figure 2.3 show road and highway network in New York State and New York City
respectively. We utilize the GIS maps of the U.S. roads and highway network, given
by National Transportation Atlas Database (NTAD) to compute network distance
along the U.S. road network between the origin and destination points of all zip code
centroids in the data.\footnote{Typical GIS software can trace this road network as a line layer and calculate the trajectory, which is drawn on the way between origin and destination zip code centroids, but it cannot take the shortest route when there are more than two routes on the way. So we use Transcad, a commercial software for analyzing transportation network, to take the optimal route by the shortest path finder module we then compute the optimal distance along the road and highway network.}

\subsection{2.4.2 Simple Difference in Difference Estimates}

To estimate the simple differences in differences estimator of toll payment, we then
perform regression analysis with a linear model, and negative binomial as the count
model specifications. The dependent variable of the regressions are trip frequency per
week. The first row of Table 2.2 presents estimates of the differences in differences
of toll payment type with different combinations of travel attributes. Notice that
regression coefficient of a categorical variable for the treatment group, without con-
sidering the other control variables yields a naïve estimator of the average treatment
effect $\delta_{ATE}$. The first five columns, in the left panel of the table, are from the linear model and remaining five columns in the right panel of the table, are from the negative binomial regression model. These model specifications were chosen by chapter 1’s theoretical travel demand model for a particular travel mode, and thus allow us to explore possible biases from various sources such as omitted variables and model misspecification. As the estimates in the first row show, there is an upward bias in the naïve estimator and it changes dramatically by controlling for the other attributes of travel demand. The naïve estimator from the negative binomial model specification in the first column of right panel is slightly less than that of the linear model’s one. This difference can be thought as model selection bias due to model misspecification because the trip frequency is drawn from poisson distribution, as chapter 1 showed.

The second column of each panel present estimates of $\delta_{ATE}$ with adjustment for travel attributes such time, distance and so on. The estimates are approximately a quarter of the unadjusted $\delta_{ATE}$ from the first column; even those are all statistically significant. It shows that the omitted variable bias due to the differences in travel attributes is about three times greater than the adjusted estimates. The estimates of $\delta_{ATE}$ with bus access in third column, are not significant and take negative values, and in the fourth columns, controlling for subway access, the coefficients are both significant and negative. The last column in both panels, presents a model where we controlling for bus and subway access in this case the models are not similar to each other, unlike to the other columns. In addition, negative Binomial regression with controls for mass transit yields a significant coefficient of $\delta_{ATE}$, while the linear regression specification does not.

From the differences in differences estimation for the average treatment effect of electronic toll collection usage on travel demand, we found the evidences that there are significant impacts of bias from omitted variables and model misspecification. As we have seen, the unadjusted estimates are upwardly biased and the test statistics
Table 2.2: Differences in Differences Estimation for Average Treatment Effect

<table>
<thead>
<tr>
<th></th>
<th>OLS Estimation</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E-ZPASS</td>
<td>0.175***</td>
<td>0.037***</td>
<td>-0.030</td>
<td>-0.078**</td>
<td>-0.064</td>
<td>0.122***</td>
<td>0.027***</td>
<td>-0.015</td>
<td>-0.042**</td>
</tr>
<tr>
<td></td>
<td>[0.015]</td>
<td>[0.012]</td>
<td>[0.024]</td>
<td>[0.037]</td>
<td>[0.040]</td>
<td>[0.011]</td>
<td>[0.009]</td>
<td>[0.012]</td>
<td>[0.019]</td>
</tr>
<tr>
<td>Toll Price</td>
<td>-0.027***</td>
<td>-0.045***</td>
<td>-0.029***</td>
<td>-0.032***</td>
<td>-0.015***</td>
<td>-0.041***</td>
<td>-0.023***</td>
<td>-0.015***</td>
<td>-0.016***</td>
</tr>
<tr>
<td></td>
<td>[0.003]</td>
<td>[0.005]</td>
<td>[0.008]</td>
<td>[0.008]</td>
<td>[0.006]</td>
<td>[0.002]</td>
<td>[0.002]</td>
<td>[0.004]</td>
<td>[0.004]</td>
</tr>
<tr>
<td>Travel Time</td>
<td>-0.082***</td>
<td>-0.001</td>
<td>0.007</td>
<td>0.022</td>
<td>-0.064</td>
<td>-0.057***</td>
<td>-0.000</td>
<td>0.005</td>
<td>0.012</td>
</tr>
<tr>
<td>(Log transformed)</td>
<td>[0.008]</td>
<td>[0.013]</td>
<td>[0.022]</td>
<td>[0.024]</td>
<td></td>
<td>[0.005]</td>
<td>[0.007]</td>
<td>[0.011]</td>
<td>[0.012]</td>
</tr>
<tr>
<td>Distance</td>
<td>-0.229***</td>
<td>-0.180***</td>
<td>-0.181***</td>
<td>-0.176***</td>
<td>-0.174***</td>
<td>-0.174***</td>
<td>-0.094***</td>
<td>-0.084***</td>
<td>-0.088***</td>
</tr>
<tr>
<td>(Log transformed)</td>
<td>[0.006]</td>
<td>[0.023]</td>
<td>[0.030]</td>
<td>[0.035]</td>
<td></td>
<td>[0.006]</td>
<td>[0.012]</td>
<td>[0.016]</td>
<td>[0.018]</td>
</tr>
<tr>
<td>Income</td>
<td>-0.093***</td>
<td>-0.124***</td>
<td>-0.117***</td>
<td>-0.106***</td>
<td>-0.053***</td>
<td>-0.063***</td>
<td>-0.059***</td>
<td>-0.053***</td>
<td></td>
</tr>
<tr>
<td>(Log transformed)</td>
<td>[0.007]</td>
<td>[0.013]</td>
<td>[0.020]</td>
<td>[0.022]</td>
<td></td>
<td>[0.004]</td>
<td>[0.007]</td>
<td>[0.010]</td>
<td>[0.011]</td>
</tr>
<tr>
<td>Working Travel</td>
<td>1.452***</td>
<td>1.219***</td>
<td>1.212***</td>
<td>1.181***</td>
<td>0.922***</td>
<td>0.642***</td>
<td>0.647***</td>
<td>0.617***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.010]</td>
<td>[0.018]</td>
<td>[0.027]</td>
<td>[0.029]</td>
<td></td>
<td>[0.007]</td>
<td>[0.014]</td>
<td>[0.017]</td>
<td>[0.018]</td>
</tr>
<tr>
<td>No. of Employments</td>
<td>0.036***</td>
<td>0.028***</td>
<td>0.038***</td>
<td>0.036***</td>
<td>0.022***</td>
<td>0.014***</td>
<td>0.019***</td>
<td>0.019***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.005]</td>
<td>[0.010]</td>
<td>[0.015]</td>
<td>[0.016]</td>
<td></td>
<td>[0.003]</td>
<td>[0.005]</td>
<td>[0.008]</td>
<td>[0.009]</td>
</tr>
<tr>
<td>No. of Vehicles</td>
<td>0.024***</td>
<td>0.092***</td>
<td>0.077***</td>
<td>0.086***</td>
<td>0.018***</td>
<td>0.046***</td>
<td>0.038***</td>
<td>0.042***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.005]</td>
<td>[0.011]</td>
<td>[0.017]</td>
<td>[0.018]</td>
<td></td>
<td>[0.003]</td>
<td>[0.006]</td>
<td>[0.008]</td>
<td>[0.009]</td>
</tr>
<tr>
<td>Bus Lines</td>
<td>-0.001</td>
<td>-0.003</td>
<td>-0.000</td>
<td>-0.001</td>
<td></td>
<td>-0.000</td>
<td>-0.000</td>
<td>-0.001</td>
<td></td>
</tr>
<tr>
<td>(Origin Area)</td>
<td>[0.001]</td>
<td>[0.003]</td>
<td></td>
<td></td>
<td></td>
<td>[0.000]</td>
<td>[0.000]</td>
<td>[0.001]</td>
<td>[0.001]</td>
</tr>
<tr>
<td>Bus Line</td>
<td>-0.003***</td>
<td>-0.004</td>
<td>-0.002</td>
<td>-0.002</td>
<td></td>
<td>-0.000</td>
<td>-0.000</td>
<td>-0.002</td>
<td></td>
</tr>
<tr>
<td>(Destination Area)</td>
<td>[0.001]</td>
<td>[0.003]</td>
<td></td>
<td></td>
<td></td>
<td>[0.001]</td>
<td>[0.001]</td>
<td>[0.002]</td>
<td>[0.002]</td>
</tr>
<tr>
<td>Subway station</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Origin Area)</td>
<td>[0.003]</td>
<td>[0.003]</td>
<td></td>
<td></td>
<td></td>
<td>[0.001]</td>
<td>[0.001]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subway station</td>
<td>-0.006**</td>
<td>-0.008**</td>
<td>-0.003**</td>
<td>-0.004**</td>
<td>-0.003**</td>
<td>-0.004**</td>
<td>-0.004**</td>
<td>-0.004**</td>
<td></td>
</tr>
<tr>
<td>(Destination Area)</td>
<td>[0.003]</td>
<td>[0.003]</td>
<td></td>
<td></td>
<td></td>
<td>[0.001]</td>
<td>[0.001]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>1.342***</td>
<td>3.081***</td>
<td>3.181***</td>
<td>3.059***</td>
<td>2.976***</td>
<td>0.294***</td>
<td>1.307***</td>
<td>1.236***</td>
<td>1.159***</td>
</tr>
<tr>
<td></td>
<td>[0.014]</td>
<td>[0.076]</td>
<td>[0.156]</td>
<td>[0.241]</td>
<td>[0.260]</td>
<td>[0.010]</td>
<td>[0.051]</td>
<td>[0.080]</td>
<td>[0.124]</td>
</tr>
<tr>
<td>No. of Obs</td>
<td>43,897</td>
<td>35,554</td>
<td>9,230</td>
<td>4,006</td>
<td>3,380</td>
<td>43,897</td>
<td>35,554</td>
<td>9,230</td>
<td>4,006</td>
</tr>
<tr>
<td>R2 (Pseudo)</td>
<td>0.003</td>
<td>0.525</td>
<td>0.388</td>
<td>0.378</td>
<td>0.379</td>
<td>(0.001)</td>
<td>(0.498)</td>
<td>(0.379)</td>
<td>(0.371)</td>
</tr>
</tbody>
</table>

1. Heteroskedasticity Robust Standard Errors are reported in square brackets. The symbols, *, **, and *** indicate respectively that the estimated coefficient is statistically significant under 10%, 5%, and 1% significance levels.
2. Because Maximum Likelihood Estimation does not provide sum of squared error and sum of squared regression, unlikely to Ordinary Least Squares, Coefficient of Determinations, called $R^2$, plays no longer role as a measure of explanatory power. We calculate, thus, Pseudo $R^2$ which is calculated $\text{Corr}(y, \hat{y})^2$, where $\hat{y}$ is fitted value of dependent variable.
3. The estimates have a dependent variable trip frequency with 6 categories, which are 0 if respondent drive a car once or less then once a week, 1 if twice a week, 2 if three times per week, 3 if four times per week, 4 if five times per week, 5 if six or more per week.
4. The model is failed to estimate, which is caused when the maximized log likelihood does not converge.
5. The model is failed to estimate, which is caused when the maximized log likelihood does not converge.
are large enough to conclude it is significant amount of bias. In addition, all of the average treatment effects $\delta_{ATE}$ estimations in left panel, linear regression model, report greater values than $\delta_{ATE}$ in right panel, negative Binomial regression. This difference can be defined as bias due to model misspecification because of two reasons. First, the dependent variable is a number of travel events per week so that it is precisely a positive integer not a real number to be treated as continuous random variable. Second, as chapter 1 pointed out, in theory, travel demand has to be treated as a positive integer because the number of travel events is definitely limited by a households time constraints. From the econometric perspective, a linear regression model would yield biased estimation of the parameters as dependent variable is a countable, positive integer; and thus it is censored at 0 because negative values are infeasible.

2.4.3 Propensity Score Matching

Since the data that we use is the collection of all MTA bridges and tunnel users, not only New York City residents or residents of New York state, but users from the other states as well, the sample for the second column covers entire sample of users, and thus is unlikely have the same sample characteristics as for the other columns. The samples for the parameter estimate while controlling for mass transit access thus eliminate users from outside of New York City because the bus and subway are available only inside the city borders. Other form of mass transit are available outside of the city border but their prices and service characteristics are much more complex than the services within the city borders. Hence, the differences between the model estimate with and without mass transit might come from sample selection bias. For this reason, in this section, we investigate the possible bias from sample section and try to perform a remedial measure of the average treatment effect using propensity score matching, and discuss about the consequence of the remedial measure technique.
2. ETC SYSTEM AND TRAVEL DEMAND

Figure 2.4: Histograms of Propensity Score: With Mass Transit

(a) Unmatched Sample: Cash User  (c) Unmatched Sample: EZ PASS User

(b) Matched Sample: Cash User  (d) Matched Sample: EZ PASS User
Figure 2.5: Histograms of Propensity Score: Without Mass Transit

To perform propensity score matching, we estimate a probit regression for the categorical variable of payment method variable, E-ZPASS user that is associated with travel attributes and mass transit access. Figure 2.4 is histograms of propensity score, and the predicted probability of probit regression without mass transit access. The top panel is for the entire sample and bottom panel is for the matched sample that is used for calculating the propensity score where we matched the average treatment effect that we introduced in section 2.3.3. Figure 2.5 is the same histograms of the estimated propensity score with mass transit access. Notice that the samples in figure 2.5 are restricted to New York City residents. Each histogram in figure 2.5 looks radically different within panel and between top and bottom. Table 2.3 reports the estimated averaged treatment effect. The second and third rows are calculated
### Table 2.3: Propensity Score Matched Average Treatment Effect

| Model                  | Sample   | Mean | ATE estimate | Test Statistics
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Treated</td>
<td>Controls</td>
<td>Obs</td>
</tr>
<tr>
<td>Unadjusted</td>
<td></td>
<td>1.342</td>
<td>1.517</td>
<td>43,897</td>
</tr>
<tr>
<td>No Mass Transit</td>
<td>Unmatched</td>
<td>1.576</td>
<td>1.459</td>
<td>5,780</td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>1.576</td>
<td>1.529</td>
<td>29,724</td>
</tr>
<tr>
<td>With Mass Transit</td>
<td>Unmatched</td>
<td>2.026</td>
<td>1.984</td>
<td>617</td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>2.022</td>
<td>2.053</td>
<td>2,738</td>
</tr>
</tbody>
</table>

$^1$ The ATE estimate of the first row is the mean difference between treatment and control group, and the rest of them are regression coefficients of the treatment group dummy variable from negative binomial model estimation. The symbols, *, **, *** indicate respectively that the estimated coefficient is statistically significant under 10%, 5%, and 1% significance levels.

$^2$ The test statistic for “Unadjusted” is t-statistics of group mean difference, and rest of them are z-statistics of MLE estimates from the negative binomial model estimation. Standard Errors are reported in parenthesis.

From the samples in which the propensity score is estimated without controlling for mass transit access, and are the same samples used in Figure 2.4, and the fourth and fifth rows are from the samples with controls for mass transit access. To identify how much selection bias is in the estimation, the naïve estimator is reported in the first row. The estimates in third, fourth, and fifth rows are around 0.04 and statistically insignificant, as compared to the first and second columns, that are around 1.2 to 1.7. This is clear evidence that the combined sources of bias for the effect of electronic toll collection system on travel demand exaggerates the estimated impact of ETC use and produce an upward bias such that the effect is identified as positive and statistically significant, but this is an incorrect conclusion. Even though much of the literature on electronic toll collection argue that it has significant effect on travel demand, the conventional wisdom of economic theory would suggest that consumer demand is all about price, and our results indicate that economic theory is correct.
Table 2.4: Descriptive Statistics without Mass Transit Access

<table>
<thead>
<tr>
<th></th>
<th>Unmatched Cash Users</th>
<th>Unmatched EZPASS Users</th>
<th>Matched Cash Users</th>
<th>Matched EZPASS Users</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs</td>
<td>Mean</td>
<td>Std. Dev</td>
<td>Obs</td>
</tr>
<tr>
<td>Toll price</td>
<td>7,253</td>
<td>3.775</td>
<td>1.762</td>
<td>37,036</td>
</tr>
<tr>
<td>Travel Time</td>
<td>7,253</td>
<td>4.401</td>
<td>0.859</td>
<td>37,036</td>
</tr>
<tr>
<td>Travel Distance</td>
<td>6,209</td>
<td>3.391</td>
<td>0.942</td>
<td>33,484</td>
</tr>
<tr>
<td>House Income</td>
<td>6,924</td>
<td>10.968</td>
<td>0.741</td>
<td>33,699</td>
</tr>
<tr>
<td>Working Travel</td>
<td>7,180</td>
<td>0.323</td>
<td>0.468</td>
<td>36,681</td>
</tr>
<tr>
<td>No. of Employments</td>
<td>7,134</td>
<td>1.821</td>
<td>1.052</td>
<td>36,421</td>
</tr>
<tr>
<td>No. of Vehicles</td>
<td>7,173</td>
<td>1.754</td>
<td>0.984</td>
<td>36,579</td>
</tr>
</tbody>
</table>

1. The samples of this table are given by the propensity score matching which are located in a common support of treatment and control group’s propensity score. 50 observations of control group are eliminated because those are located in out of common support.

2. The propensity score to get matched sample is given by predicted value of probit regression of E-ZPASS users on independent variables without mass transit access.

Table 2.4 and 2.5 report descriptive statistics of different samples, distinguished by propensity score matching and control variables also as to whether mass transit variables are included. All control variables, travel attribute and mass transit access, are not very different between E-ZPASS users and non-users, except for toll price. The mean and standard deviation of toll price between users and non-users of ETC are different and slightly greater than the others as is specified by the MTA pricing policy of discounting ETC prices relative to cash prices. This is an identical result to Table 2.2 that report highly significant toll price impact for all model specifications and combinations of control variables, while the significance and sign of E-ZPASS usage varies over models and control variables. This empirical evidence implies that people are concerned about price of travel when they make their travel decisions and payment method is apparently not that serious a matter for those same decisions.
Table 2.5: Descriptive Statistics without Mass Transit Access

<table>
<thead>
<tr>
<th>Unmatched Cash Users</th>
<th>EZPASS Users</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Obs</strong></td>
<td><strong>Mean</strong></td>
</tr>
<tr>
<td>Toll price</td>
<td>7,253</td>
</tr>
<tr>
<td>Travel Time</td>
<td>7,253</td>
</tr>
<tr>
<td>Travel Distance</td>
<td>6,209</td>
</tr>
<tr>
<td>Household Income</td>
<td>6,924</td>
</tr>
<tr>
<td>No. of Employments</td>
<td>7,134</td>
</tr>
<tr>
<td>No. of Vehicles</td>
<td>7,173</td>
</tr>
<tr>
<td>No. of Bus line (Origin)</td>
<td>3,488</td>
</tr>
<tr>
<td>No. of Bus line (Destination)</td>
<td>4,204</td>
</tr>
<tr>
<td>No. of Subway Station (Origin)</td>
<td>2,611</td>
</tr>
<tr>
<td>No. of Subway Station (Destination)</td>
<td>2,827</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Matched Cash Users</th>
<th>EZPASS Users</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Obs</strong></td>
<td><strong>Mean</strong></td>
</tr>
<tr>
<td>Toll price</td>
<td>456</td>
</tr>
<tr>
<td>Travel Time</td>
<td>456</td>
</tr>
<tr>
<td>Travel Distance</td>
<td>456</td>
</tr>
<tr>
<td>Household Income</td>
<td>456</td>
</tr>
<tr>
<td>Working Travel</td>
<td>456</td>
</tr>
<tr>
<td>No. of Employments</td>
<td>456</td>
</tr>
<tr>
<td>No. of Vehicles</td>
<td>456</td>
</tr>
<tr>
<td>No. of Bus line (Origin)</td>
<td>456</td>
</tr>
<tr>
<td>No. of Bus line (Destination)</td>
<td>456</td>
</tr>
<tr>
<td>No. of Subway Station (Origin)</td>
<td>456</td>
</tr>
<tr>
<td>No. of Subway Station (Destination)</td>
<td>456</td>
</tr>
</tbody>
</table>

1. The samples of this table are given by the propensity score matching which are located in a common support of treatment and control group’s propensity score. 50 observations of control group are eliminated because those are located in out of common support.

2. The propensity score to get matched sample is given by predicted value of probit regression of E-ZPASS users on independent variables without mass transit access.

2.5 Conclusion

We investigate, in this study, the effect of the E-ZPASS, electronic toll collection system on travel demand using the Triborough Bridges and Tunnel Authority’s user survey in 2004. The main focus of this work is to explore possible sources of bias in which the estimate of the average treatment effect, that is the difference of travel demand between E-ZPass users and non-users into. We observe clear evidence that simple difference-in-difference estimator of the average treatment effect, which is estimated by adding a categorical variable of treatment group, is upwardly bias due to model misspecification and omitted variables. By utilizing the propensity matching estimator of the average treatment effect, we are able to also detect a sample selection bias that possibly originated from the different transit condition that individuals have
in their living areas, such that individuals with alternative modes can choose when they decide to travel and by alternatives to traveling on a priced automobile route.

From a traditional economics perspective, this analysis is not that striking. Because economists expect, in principle, that individual respond to the prices of goods and services when they decide their quantity demanded. Since E-ZPASS is just a more convenient way to pay a toll than cash, therefore, one would expect that E-ZPASS itself cannot cause an effect on demand. In addition, we examine various econometric issues in the study and find that the absence of careful concern for model specification, omitted variables, and sample selection issues yield incorrect model estimations that are biased and not robust over sample selection. Much of the empirical literature on travel demand was written without consideration of the factors that can cause a biased estimate and many apparently are likely to be written with seriously biased result. A clear potential consequence of these problems are poor transportation planning decisions, based on these types of biased empirical estimates. These poor decision may in fact yield greater social cost than to not impose pricing on a road network.

From what we have explored in this paper, we can draw a prediction that Installation of ETC at toll plazas with the intent of reducing congestion may have no impact on congestion because travelers respond to price of the toll, not the ETC system itself. So travelers do not change their travel demand whether the toll entity installs an ETC system or not. The engineering literature argues that installing an ETC system can improve traffic flow and as such would stimulate demand based on a decrease in total travel time. Hence we would like to suggest to transportation researchers and planners that a careful examination of possible sources of bias, which includes model misspecification, omitted variables, and sample selection, matters have to be evaluated and controlled for when the empirical analysis of travel demand behavior is estimated.
3 Parallel Sparse Algorithm for Matrix-Transpose-Matrix Multiplication using Outer Product of Compressed Sparse Rows

with Michael E. Kress
3. PARALLEL SPARSE MATRIX MULTIPLICATION

3.1 Introduction

In practice, statistical data is often treated as a vector, or matrix, and because of that, there are many statistical computing algorithms that have been developed based on linear algebraic operations on numerical matrices. One of the most demanding linear algebraic operations in statistics is matrix-transposed-vector, or matrix-transposed-matrix multiplication. Variance-Covariance matrix, a basic multivariate statistical inference, is an example that can be computed by a transposed data matrix multiplied by the matrix, \( X^T X \), where \( X \) is the matrix containing data.\(^1\) Linear model estimation such as linear regression, and analysis of variance (ANOVA) is an example that requires calculating an inverse of the matrix \((X^T X)^{-1}\). It is thus a necessary condition for any linear model estimation, in order to solve for the normal equation for the parameter vector that turns out to be a least square estimator.\(^2\) Another example is the estimation of the generalized linear model with a matrix of instrumental variables, denoted by \( W \), that is used to control for endogenous exploratory variables. This estimation requires the inverse of the transposed instrumental variable matrix times the data matrix, \((W^T X)^{-1}\). These statistical analyses also thus begin with computing matrix-transpose-matrix multiplication \( X^T X \), or \( W^T X \).

Calculating the matrix-transposed-matrix multiplication quickly and accurately is therefore a very important technique in statistical computing. However, this calculation has not been supported well by modern numerical computation techniques,\(^1\) The exact formula of variance-covariance matrix is \( X^T MX \), where \( M = I - \frac{1}{N}(\iota \iota^T) \) with a column vector of ones \( \iota \).\(^2\) The normal equation comes from least square fitting of a linear model. Suppose that a researcher is interested in estimate of a parameter vector of interest \( \beta \), for the given linear model \( y = X\beta + \epsilon \), where \( y, X \) are a vector of dependent variable and a matrix of independent variables respectively, and \( \epsilon \) is a vector of unobservable error term. The least square estimate of \( \beta \) is given by solving the equation:

\[
X^T X \beta = X^T y
\]

This equation is called normal equation and the existence of the solution is guaranteed only by existence of the matrix inverse \((X^T X)^{-1}\).
especially parallel computing because statistical data matrices are merely square. In other words, the data matrices typically have the form in which the number of columns, which represents number of variables, is generally smaller than the number of rows that represents the number of observations. Previously developed parallel algorithms cannot be applied directly to the matrix-transpose-matrix multiplication for the statistical data matrices because they usually are designed for square matrices, or matrices where the number of rows is not much deviated from number of columns. Furthermore, the concept *big data* emphasizes that the size of data is large, so implementing parallel computing becomes more essential for contemporary large-scale data matrix computation.

Another critical computational issue for a large-scale statistical analysis is the absence of parallel algorithm for sparse matrix-matrix multiplication. Analyzing big data requires huge storage and sometimes the data size exceeds the maximum memory allocation for a single computing processor, which is 32 gigabytes (Gbytes) for the Intel processor. Reducing data size as small as possible is therefore another important matter as this serves to reduce the computational memory requirement. Since statistical data matrices in many cases have a large number of zero elements, manipulating the matrix in a sparse form provides a significant advantage in terms of size and efficiency of the computation. Unfortunately, to our knowledge there is no sparse matrix-transpose-matrix multiplication algorithm that has been developed for parallel computing system. Moreover, there are a few parallel algorithm for sparse matrix-vector multiplication but they use a fixed sparse matrix form that is inflexible in terms of the shape of a matrix for applying to statistical computing.

In order to develop parallel algorithms for sparse matrices, we use the outer product method for the matrix multiplication problem of $A^T B$, where $A_{n \times k}$, $B_{n \times l}$ have the same number of rows but the number of columns is not necessarily the same. The matrix multiplication is identical to the sum of all outer product of each rows
that is $A^T B = \sum_{i=1}^{n} a_i^T b_i$, where $a_i$, $b_i$ are $i^{th}$ row vector of each matrices. Since the terms in summation operator are completely independent of each other, and summation over the compute nodes is fully supported by a collective communication of parallel implementation standard such as MPI Reduce, the sum of vector operations are perfectly scalable to distribute its computation into multiple compute nodes.

This outer product algorithm also can be readily implementable in parallel computing by only utilizing compressed sparse row (CSR) form matrices, which are the most widely used sparse storage form in parallel sparse linear algebraic operations. In general, statistical computing deals with externally collected, or generated data from outside of the computation environment, and the data estimation generally involve more than two different computations to obtain a final output. If each computation can only handle certain types of sparse matrix storage forms, the data needs to be stored in several different sparse forms, and therefore, this causes significantly more memory consumption data storage and communication costs for parallel implementation. Since the outer product algorithm allows to store both left hand matrix $A$ and right hand side matrix $B$ in CSR form, it does not require to store the data in any particular sparse forms.

In this research, therefore, we present the outer products algorithm for both dense and sparse matrix multiplication in a parallel processing environment. Then we show the relative performances of its sparse matrix multiplication algorithm with randomly generated matrices as an experiment. The experimental matrix multiplication is coded in Fortran with MPI to implement them on a high performance computing system, provided by the CUNY High Performance Computing Center. Finally, we present an application of the outer product algorithm for a large-scale statistical computing using the New York City taxicab data, which has 378,532,118 taxi trip records for the period 2008-10. Because the taxicab data is originally stored in a 140.56 Gbytes text files, it provides a good opportunity to handle and analyze a big data
3. PARALLEL SPARSE MATRIX MULTIPLICATION

set. We describe how to manipulate the original taxicab data into a valid statistical data matrix, stored it into CSR form, and calculate basic descriptive statistics such as mean, variance, and covariance by applying the outer product algorithm.

The organization of this paper is as follows: In section 3.2, we present the outer product algorithm for both dense and sparse matrices. We then discuss about an experiment that we conducted to evaluate the performance of the algorithm in a parallel computing system in section 3.3. We then demonstrate an application of the outer product algorithm to perform statistical analysis of the New York City taxicab data in section 3.4.

3.2 Outer Product Algorithm

3.2.1 Parallel Algorithm for Dense Matrices

Let $A_{n \times k}$, and $B_{n \times l}$ be matrices, not necessarily square, but with the same number of rows. We can express the matrices $A$ and $B$ as column vectors of row vectors, which are:

$$A = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, \quad B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

where $a_i = [a_{i1} \ a_{i2} \ldots \ a_{in}], i = 1, \ldots, k$ is a row vector of $k$ elements, and $b_i, i = 1, \ldots, n$ is a row vector with $l$ elements. Consider a matrix multiplication of the transposed $A$ by $B$, denoted $A^T B$. This $k \times l$ matrix can be obtained by applying
inner (dot) product of the column vectors as:

\[
A^T B = \left[ \begin{array}{c}
\mathbf{a}_1 \\
\mathbf{a}_2 \\
\vdots \\
\mathbf{a}_n
\end{array} \right]'
\cdot
\left[ \begin{array}{c}
\mathbf{b}_1 \\
\mathbf{b}_2 \\
\vdots \\
\mathbf{b}_n
\end{array} \right] = \sum_{i=1}^{n} \mathbf{a}_i^T \mathbf{b}_i.
\]

Since \( \mathbf{a}_i \) is originally defined as a row vector, \( \mathbf{a}_i^T \mathbf{b}_i \) is the outer product of \( \mathbf{a}_i \) and \( \mathbf{b}_i \). This shows that \( A^T B \) can be expressed as the sum of outer products of each row vectors in \( A \) and \( B \), and \( A^T B \) operation can be parallelized by dividing the sum amongst multiple number of processors. Let \( p \) be a number of processors that are available for \( A^T B \) operation. The sum of outer products can be divided into \( p \) pieces as:

\[
A^T B = \sum_{i=1}^{n^*} \mathbf{a}_i^T \mathbf{b}_i + \sum_{i=1}^{2n^*} \mathbf{a}_i^T \mathbf{b}_i + \ldots + \sum_{i=1}^{pn^*} \mathbf{a}_i^T \mathbf{b}_i
\]

\[
= \left[ \begin{array}{c}
\mathbf{a}_1^T \\
\mathbf{a}_2^T \\
\vdots \\
\mathbf{a}_{n^*}^T
\end{array} \right] 
\cdot
\left[ \begin{array}{c}
\mathbf{b}_1 \\
\mathbf{b}_2 \\
\vdots \\
\mathbf{b}_{n^*}
\end{array} \right] + \ldots

+ \left[ \begin{array}{c}
\mathbf{a}_{(p-1)n^*+1}^T \\
\mathbf{a}_{(p-1)n^*+2}^T \\
\vdots \\
\mathbf{a}_{pn^*}^T
\end{array} \right] 
\cdot
\left[ \begin{array}{c}
\mathbf{b}_{(p-1)n^*+1} \\
\mathbf{b}_{(p-1)n^*+2} \\
\vdots \\
\mathbf{b}_{pn^*}
\end{array} \right]
\]
3. PARALLEL SPARSE MATRIX MULTIPLICATION

The $A^T B$ operation turns out to be a sum of row-wise partitioned matrix multiplications and the partitioning strategy can be consistent with the (row) size of the matrices. In other words, $A$ and $B$ can be equally divided and sent to each processors because $A$ and $B$ have the same row size $n$.

The number of rows of a local matrix $A[i]$ is given by the floor of $n/p$. If $\text{mod}(n/p) \neq 0$, then we can attach $\text{mod}(n/p)$ number of remaining rows to any processors’ partitioned matrix to deal with remainders. The $A^T B$ operation can thus be summarized by the following algorithm:

### Algorithm 1 Outer Product Algorithm for Dense Matrices

1: **procedure** $\text{MAT}(A, B, p)$
2: \hspace{1em} $n \leftarrow \text{size}(A, 1)$
3: \hspace{1em} $n^* \leftarrow \text{int}(n/p)$
4: \hspace{1em} for $i \leftarrow 1, p$ do \hspace{1em} # Iterate through processors
5: \hspace{2em} $A[i] \leftarrow A[(n^*(i - 1) + 1) : (n^* \cdot i)]$ \hspace{1em} # Send data to $i$th processor
6: \hspace{2em} $B[i] \leftarrow B[(n^*(i - 1) + 1) : (n^* \cdot i)]$ \hspace{1em} # Send data to $i$th processor
7: \hspace{1em} $A'[i] \leftarrow 0$ \hspace{3em} # Define local target matrix as zero
8: \hspace{1em} end for
9: \hspace{1em} for $j \leftarrow 1, n^*$ do \hspace{1em} # Iterate within processors
10: \hspace{2em} $A'[i](j) \leftarrow A'[i](j - 1) + A[i](j)B[i], (j)$ \hspace{1em} # Compounding outer products
11: \hspace{1em} end for
12: **Reduce** $A' B$ \hspace{1em} # Collective communication of sum through processors
13: **end procedure**

3.2.2 Parallel Algorithm for Sparse Matrices

Most parallel sparse matrix algorithms perform for various mathematical operations, not only for multiplication but other techniques such as transposition and matrix decompositions which require choosing a particular storage form prior to performing the computation due to a designated inter-process data exchange. In other words, the matrix multiplication in a parallel computation environment divides the given matrices into several block matrices and distributes them in an inflexible way because a block matrix must be matched with the exact target block matrix. This inflexible

---

3. Floor is an integer function that rounds down a division, and mod is the modulo operation that finds the remainder of a division.
partitioning demands a pre-determined storage form when the algorithm considers a matrix operation with sparse matrices. Here in our example, we choose \textit{compressed sparse rows} (CSR) form to develop the algorithm. The CSR storage form is widely accepted as a current standard because the least number of indices, number of rows plus number of non-zero elements, are needed.

Since many of parallel sparse matrix operation algorithms are designed based on CSR, the multiplication algorithm with the CSR form would yield flexibility that can be incorporated with the other computations, once the matrices are stored in CSR form. For example, suppose that a matrix, which is used in a multiplication, need to be used in another matrix operation. If the multiplication algorithm does not support CSR, while the other operation has only an algorithm for CSR form, the matrix has to be stored into two different forms. The simple and practical example of the above case is to perform matrix-vector multiplication and then matrix-transposed-vector multiplication. The matrix-vector multiplication can be done with the CSR form, but matrix-transposed-vector multiplication cannot use the same form because of the absence of an appropriate computation algorithm, so the matrix has to be stored in both CSR and CSC at the same time, in order to perform the two matrix-vector multiplications simultaneously thus is extremely costly in terms of data storage for large data sets.

As a starting point, the outer product algorithm can be implemented in CSR form.\footnote{Buluc and Gilbert [2008] points out that the CSR form cannot be applied to matrix-transpose-vector multiplication, even though the form is the most efficient sparse storage form in terms of required number of indices; therefore, that the matrix should be stored in \textit{compressed sparse columns} (CSC) form, or \textit{compressed sparse blocks} (CSB) form that they propose. Therefore, by using the CSR form for the matrix-transposed-vector, we propose that the efficiency and accuracy of the algorithm is guaranteed.}
Consider a matrix $A$ as:

$$
A = \begin{bmatrix}
1 & 0 & 0 & 2 & 0 \\
3 & 4 & 0 & 5 & 0 \\
6 & 0 & 7 & 8 & 9 \\
0 & 0 & 10 & 11 & 0 \\
0 & 0 & 0 & 0 & 12
\end{bmatrix}.
$$

Any matrix can be represented by a CSR form that consists of three vectors, a vector of nonzero elements, a vector of column index, and a vector of an index for the nonzero element vector that represents the first nonzero element in each row. The CSR form of the matrix $A$ is therefore given as:

$$
AA = [1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10 \ 11 \ 12],
$$

$$
JA = [1 \ 4 \ 1 \ 2 \ 4 \ 1 \ 3 \ 4 \ 5 \ 3 \ 4 \ 5],
$$

$$
IA = [1 \ 3 \ 6 \ 10 \ 12 \ 13],
$$

where $AA$ is the nonzero element vector, $JA$ is the column index vector, and $IA$ is the index vector for $AA$ that represents the first element of every row. Note that all nonzero elements in a row can be indexed by $IA$. For example, the elements of the first row of the matrix $A$ is from $IA(1)$ to $IA(2) - 1$ of the $AA$'s elements.

Using this property, the matrix $A$ in CSR form can be partitioned into rows in the same way as the dense form matrix is partitioned for the outer product multiplication. Suppose that there are two available compute nodes for the operation that is $p = 2$, and since $A$ has five rows that is $n = 5$, the number of partitioned matrices would be three and two, or two and three that are $n^* = 3$ or $2$. In this way, we can divide $A$ in CSR form into two local matrices as:
A[1] = \[
\begin{align*}
AA[1] &= [1 \ 2 \ 3 \ 4 \ 5], \\
JA[1] &= [1 \ 4 \ 1 \ 2 \ 4], \\
IA[1] &= [1 \ 3 \ 6]
\end{align*}
\]

\begin{align*}
AA[2] &= [6 \ 7 \ 8 \ 9 \ 10 \ 11 \ 12], \\
JA[2] &= [1 \ 3 \ 4 \ 5 \ 3 \ 4 \ 5], \\
IA[2] &= [1 \ 5 \ 7 \ 8]
\end{align*}
\]

The above example shows how a CSR matrix can be partitioned by row \(^5\) as the same way as the dense matrix algorithm parallelizes its matrices, so that the outer product algorithm can be implemented in CSR form matrix-transposed-matrix multiplication. Consider again the multiplication problem \(A^T \times B\). Let \(p\) be the number of compute nodes that incorporates with the parallel matrix multiplication. Consider CSR form of \(A\) and \(B\) are given by \(AA, JA, IA\) and \(BB, JB, IB\) respectively. The parallel sparse algorithm of matrix-transpose-matrix multiplication is then given by following:

3.2.3 Comparison with Conventional Algorithms

There are three computational efficiencies and advantages that the outer product algorithm provides. First the matrix-transpose-matrix multiplication can be done with the CSR form even though it involves transposition of a matrix. A tuple form representation, which stores each nonzero of a matrix as a triple consisting of its row index, its column index, and the nonzero elements, is able to be transposed by

\(^5\)Note that \(IA[2]\), the row index of the second local matrix \(A[2]\), needs to be adjusted as it begins with one, in order to use it as a new CSR form matrix in the second compute node that is assigned \(A[2]\).
### Algorithm 2: Outer product algorithm for CSR form matrix

1. **procedure** `smat(AA, JA, IA, BB, JB, IB, p)`
2. \( n \leftarrow \text{size}(IA) - 1 \)
3. \( n^* \leftarrow \text{int}(n/p) \)
4. \( \text{for } i \leftarrow 1, p \text{ do} \)
5. \( IA[i] \leftarrow IA[(n^*(i - 1) + 1) : ((n^* \cdot i) + 1)] \)
6. \( AA[i] \leftarrow AA(IA[i])((n^* + 1) - 1) \)
7. \( JA[i] \leftarrow JA(IA[i])((n^* + 1) - 1) \)
8. \( IB[i] \leftarrow IB(IA[i])(n^* + 1) - 1 \)
9. \( BB[i] \leftarrow BB(IB[i])((n^* + 1) - 1) \)
10. \( JB[i] \leftarrow JB(IB[i])(n^* + 1) - 1 \)
11. \( A'B[i] \leftarrow 0 \)
12. \( \text{end for} \)
13. \( \text{for } j \leftarrow 1, n^* \text{ do} \)
14. \( k_1 \leftarrow IA(j) \)
15. \( k_2 \leftarrow IA(j + 1) - 1 \)
16. \( k_3 \leftarrow IB(j) \)
17. \( k_4 \leftarrow IB(j + 1) - 1 \)
18. \( \text{for } l \leftarrow k_1, k_2 \text{ do} \)
19. \( \text{for } m \leftarrow k_3, k_4 \text{ do} \)
20. \( A'B(j)(JB(l), JA(m)) \leftarrow A'B(j - 1)(JB(l), JA(m)) + AA(j)(l) \cdot BB(j)(m) \)
21. \( \text{end for} \)
22. \( \text{end for} \)
23. \( \text{end for} \)
24. \( \text{Reduce A'B} \) \( \triangleright \) Collective communication of sum through processors
25. \( \text{end procedure} \)

This algorithm iterates through processors to distribute and collect data for matrix multiplication. It involves handling sparse matrices in CSR form and ensuring efficient computation based on parallel processing.

---

6For example, a nonzero element of a matrix \( A \)'s \( i \)th row and \( j \)th column can be represented in tuple form as \((i, j, A_{i,j})\), where \( A_{i,j} \) is the nonzero element. The tuple representation allow to obtain \( A^T \) by switching row and column indices as \((j, i, A_{i,j})\).
3. PARALLEL SPARSE MATRIX MULTIPLICATION

(a) One dimensional decomposition

(b) Two dimensional decomposition

Figure 3.1: Matrix decomposition for \(A^T B\)

multiplications as itself and its transpose, both CSR and CSC indices are required simultaneously, and therefore, twice the number of integers have to be stored, which is \(2 \times (n + \text{nnz}(A))\). Since the outer product stores \(n + \text{nnz}(A)\), it becomes more efficient as the matrix size increases.

The second advantage is that the outer product algorithm has no inter-processor communication and thus it does not need to exchange data during its execution, except for distributing data to compute nodes at the beginning, and collecting outputs from the nodes at the end of the execution. This distinguishes this method from conventional parallel matrix-matrix multiplication algorithms with more than two dimensional (2D) block decomposition of its matrices. In conventional parallel matrix multiplications, during the computation, a 2D or more than 2D block decomposed matrix has to be sent from a compute node that stores a submatrix to a number of compute nodes that store submatrices that needs to be matched with the sent matrix.

Figure 3.1 shows that a block matrix in 2D decomposition has to be distributed to all block matrices in the same row. Even though the entire communication volume and number of floating point operations are the same, longer 2D block decomposition would have longer computation time than the outer product algorithm like 1D decomposition that isolates multiplications of submatrices at each compute nodes. Because 1D algorithms that have no communication between processors as compute nodes can avoid communication time as a part of its computation time, 2D algorithms are thus
The third advantage is that the allocation of data is relatively flexible. As Buluç and Gilbert (2008) points out, parallel algorithm for matrix-matrix multiplication requires a somewhat formally designated way of decomposition so that the scalability of matrices that are involved in the multiplication play a crucial role in determining performance of the algorithms, especially for sparse matrices. In 2D decomposition, for example, the number of rows and number of columns of each submatrices $A^T[p], B[p]$ have to be matched to compute a target matrix $A^T B[p]$. In other words, the applicability of the algorithm is subject to the shape of matrices and therefore, not every matrix-matrix multiplications can be done with that algorithm. In statistical computing, moreover, size of data matrix, especially number of rows that is number of observations, is not predictable before the data is given. Since the outer product algorithm decomposes the given matrices with the same number of rows, it is suitable for matrix-matrix multiplication with an unpredictable number of rows such as is commonly found in the statistical data matrices.

### 3.3 Experiment

In this section, we describe an experiment that we conducted to evaluate the outer product algorithm for sparse matrix-transpose-matrix multiplication. The experiment focuses particularly on how the algorithm performs when the matrix given is similar to statistical data, which is rectangular with a large number of rows, and randomly scattered zero elements on the matrix. In addition, the experiment demonstrates the role of parallel execution for the algorithm of statistical operations like matrix multiplication.
3. PARALLEL SPARSE MATRIX MULTIPLICATION

3.3.1 Experimental Design

We design and conduct an experiment to evaluate the performance of parallel execution of the outer product algorithm based on three specific characteristics. First, zero and nonzero elements are randomly scattered over the matrix. Second, the matrix has a relatively large number of rows and small number of columns so it is a rectangular matrix. Third, the size of the matrix is large scale, especially for number of rows. The first and second are general characteristics for statistical data as a matrix, and the third one is required to test the algorithm for a large-scale statistical computing, or big data analysis.

Figure 3.2: Spy plot of the 1,000 by 1,000 random matrix
We generate uniform random numbers that takes a value from minus one to one, then substitute the value, which are smaller than zero, to zero. This provides a matrix in which approximately half of its elements have zero and those are randomly located in the matrix. By converting the matrix, which is dense form, into CSR form, we eventually obtain a random sparse matrix in CSR form. This procedure of generating random sparse matrix is done via R, a language and environment for statistical computing and graphics, which is installed in a Dell shared memory system of the CUNY High Performance Computing Center.

Figure 3.2 is a spy plot of a 1,000 by 1,000 random matrix that is generated by the above procedure. The dots in the matrix represent nonzero elements, the blanks represent zero elements that also represents sparsity pattern of the matrix. As we can see, the nonzero elements seem to be scattered over the matrix without making any particular patterns such as clustering into some particular area.

Hence, we generate three different matrices with the same number of rows but different number of columns using the random matrix generating procedure. In order to keep the three characteristics of matrices in this experiment, which are rectangular shape with large number of rows and randomly scattered zero elements, we generate one million by ten, one million by 100, and one million by 1,000 as the left hand side matrix $A$. Then we perform matrix-transpose-matrix multiplication of $A$ with a one million by 1,000 matrix as the right hand side matrix $B$. The basic information about the experimental matrices and $A^T B$ operations is summarized in Table 3.1.

The second and fourth rows in the table report sparcity, ratio of nonzero elements to the entire number of elements in a matrix. All three matrices are approximately 50% non zero elements as we planned. We calculate number of flops for $A^T B$ in a way

---

Table 3.1: Statistics of experimental matrices

<table>
<thead>
<tr>
<th>Size of Target Matrix</th>
<th># of non-zero of ( A )</th>
<th>Sparsity of ( A )</th>
<th># of non-zero of ( B )</th>
<th>Sparsity of ( B )</th>
<th># of flops for ( A^T B )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 10 \times 1,000 )</td>
<td>( 4.9990 \times 10^6 )</td>
<td>0.4999</td>
<td>( 5.0003 \times 10^6 )</td>
<td>0.500025</td>
<td>( 19.999 \times 10^9 )</td>
</tr>
<tr>
<td>( 100 \times 1,000 )</td>
<td>( 4.9995 \times 10^6 )</td>
<td>0.4999</td>
<td>( 5.0003 \times 10^6 )</td>
<td>0.500025</td>
<td>( 19.999 \times 10^9 )</td>
</tr>
<tr>
<td>( 1,000 \times 1,000 )</td>
<td>( 5.0002 \times 10^6 )</td>
<td>0.5000</td>
<td>( 5.0003 \times 10^6 )</td>
<td>0.500025</td>
<td>( 19.999 \times 10^9 )</td>
</tr>
</tbody>
</table>

that \( A \) and \( B \) are assumed to be dense. The number contains both useful and not useful flops, which are number of floating point operations of zero elements. The flops itself is therefore not accurate in terms of actual number of floating point operation in sparse matrix multiplication but it is still useful to measure performance of parallel execution of the operation and compare it with different number of CPUs.

3.3.2 Random Matrix Results

We execute matrix multiplication of the three different target matrices on a SGI cluster system with 744 processor cores and 96 NVIDIA Fermi processor accelerators located at the CUNY High Performance Computing Center. This particular SGI cluster system, named Andy, is separated into two parts. One is an SGI ICE system with 45 dual-socket, compute nodes each with Intel 2.93 GHz quad-core Intel Core 7 (Nehalem) processors providing a total of 360 compute cores. Each computing node has 24 Gbytes of memory or 3 Gbytes of memory per core. The first part of Andy’s interconnect network is a dual rail, DDR Infiniband (20 Gbit/second) network in which one rail is used to access Andy’s Lustre storage system and the other is used for

\^[8]Number of flops for \( A^T B \) stay constant as long as the size of the matrices \( A \) and \( B \) stay the same, although number of useful flops for \( A^T B \) vary over the sparsity pattern of the two matrices. Therefore, conventional performance measure such as gigaflops, or megaflops still work to show how the performance parallel execution of the algorithm is differed when a greater the more number of CPUs get involved in the execution.
inter-processor communication. The other part of the Andy is a cluster of 48 SGI x340 1U compute nodes, also connected to 24 1U quad-GPU Fermi s2050 accelerator nodes. The multiplication of the three matrices are coded in Fortran 95 with OpenMPI 1.6.3, which is an open source MPI-2 implementation for Intel compiler and compiled using Intel Compiler Suite 13.0.1.117[^9].

The execution results are reported in Figure 3.3 and 3.4. Figure 3.3 is line plots of gigaflops per second (Gflops), from the three different matrix multiplications. We measure CPU clock time in seconds between the MPI_receive is done and when the MPI_reduce is done, then use it as the denominator of Gflop/sec. We obtain Gflops with 19 different combinations of CPUs, which range from 1 to 128, for each matrix multiplication. The straight line in Figure 3.3 is Gflops for the matrix multiplication with 10 columns on the left hand side that thus becomes a 10 by 1,000 target matrix. The trend goes up similarly with the dotted-dashed line, which is for 100 by 1,000.

[^9]: Note that because our programming execution do not involve GPU accelerations, we use the first part of Andy in which the number of available CPUs is 360.
matrix multiplication, from one to 16 CPUs, then the upward trend gets stronger. 100 by 1,000 multiplication has ten times more floating point operations (flops) than a 10 by 1,000 matrix but those two calculations take almost the same until the 16 CPU cores. This result can be interpreted as follow; the outer product algorithm with small number of CPUs well performs for a large scale matrix multiplication, but it poorly performs for a small scale multiplication.

By comparing the dotted-dashed line with dashed line, Gflops for 1,000 by 1,000 in Figure 3.3, however, the same interpretation cannot be applied. The gap between the dashed line and dotted-dashed line is getting wider until 48 CPUs, and it gets closer at 64 CPUs. Then the two lines go up by roughly the same slope. The comparison of Gflops between 100 by 1,000 and 1,000 by 1,000 seems to be reversed with the pattern of 10 by 1,000 and 100 by 1,000 case. Although these comparison create some ambiguity on performance, we can observe two consistent patterns over the number of CPUs. First, the smaller the matrices’ size, the higher the Gflops. Second, Gflops are slightly increased by adding more CPUs.

Because one of the outer product algorithm’s distinguished feature is to remove the need for communication time between the computing nodes and thus reduce computing time, it is important to show communication time pattern over the number of CPUs and flops. We report the total execution time and the associated communication time in Figure 3.4. The solid line is total execution time, which is CPU clock in seconds from beginning of the \texttt{MPI\_send} to end of the \texttt{MPI\_reduce}. The shaded region under the straight line is communication time, which is CPU clock time from beginning of \texttt{MPI\_send} to end of \texttt{MPI\_receive}. Panel (a) shows that total execution time goes up by increased communication time, when number of flops is not large enough to take advantages from parallel execution. Unlike the 10 by 1,000 case, matrix multiplication with a larger number of flops has reduced total execution time with slightly increased communication time. Panel (b) shows that the communication
Figure 3.4: Execution and communication times for $A^T B$
time takes a relatively smaller portion of total execution time with less than 24 CPUs, but it dominates most of the execution time. In panel (c), the communication takes very small portion of the entire execution time and it keeps declining over CPUs. Figure 3.4 shows a consistent pattern that the portion of communication time, out of the entire execution time, gets smaller and the advantage of parallel execution gets larger, as the size of matrices is increased.

### 3.4 Application: NYC Taxicab Data

In this section, we describe an application of the outer product algorithm on statistical computing for a big data set. The data for this test is the New York City yellow medallion taxicab data, and the statistical computing challenge is calculation of the variance-covariance matrix, and an estimations of conditional expectation of travel demand. The NYC taxicab data is a geographic positioning system (GPS) tracking record for each of the entire totallity of trips made by NYC medallion taxis during 2008-10. The data was collected by New York City Taxicab Limousine Commission (NYCTLC). The NYCTLS mandated their medallion taxicabs to equip each vehicle with an electronic device that also allowes to passengers to use credit/debit card payment and enhanced driver and passenger information monitoring under the name *Taxicab Passenger Enhancement Program* (T-PEP), which was legislated as TLC policy on March 30, 2004. The T-PEP equipment was originally promoted to enhance customer service by providing credit card payment, and automatic trip reporting in order to detect and intercept drivers’ fraud. The equipment records detailed information about a taxi trip such as time, location, number of passengers, payment method, and itemized taxi fares. It then reports the recorded information to NYC TLC server.\(^{10}\)

The taxicab data, provided by NYC TLC, contains 378 million taxi trip records

\(^{10}\)Note that trip time and location are recorded as origin and destination pairs.
from January 1, 2008 to December 31, 2010. The original data was separated into two comma-separated-value (csv) formatted text files and the size of these files was 52.11Gbytes and 88.45Gbytes. The headers in every column of the data file are variables, which are vehicle id, medallion id, driver’s name, shift number, trip id, trip pick up date time, trip drop off date time, trip pick up location, trip drop off location, passenger count, payment type, fare amount, toll amount, tip amount, surcharge, total amount, trip time, trip distance, start longitude, start latitude, end longitude, end latitude, vendor name, start zipcode, end zipcode, distance between service, and time between service.

3.4.1 Manipulation of Big Data

In order to perform any statistical analysis, values in data must be numeric that are either integers or real numbers but some of variables in the taxicab data such as payment method, O-D time are recorded as character variables. We read the two csv data files into SAS, a statistical computing package developed by SAS Institute Inc. that was installed on a Dell shared memory system with 24 processor cores at the CUNY High Performance Computing Center. This system, named Karle has a total of 96 Gbytes of memory or 4 Gbytes per core so that the maximum memory for a serial job execution is up to 90 Gbytes. The main advantage to use SAS for big data manipulation is that SAS performs its operations line-by-line for the DATA step, which is one of four major part of SAS programs. Executing the SAS data-step program for big data takes a significantly longer time because of its line-by-line looping procedure but it allows it also to avoid unexpected termination of procedures due to insufficient memory or memory spillover\(^{11}\).

Once the data is imported into SAS, we create time and data variables using the

\(^{11}\)Reading a text file formatted data into SAS, however, requires available memory space at least as big as the data size. Shared memory system in which maximum memory allocation can exceed 32 Gbytes, the maximum physical memory allocation for a single Intel CPU, such as Karle server in CUNY High Performance Computing center that we use for this study.
recorded beginning and end time for each trip. The time record in the data takes the format like “01JAN08:23:59:59” for 23:59:59 of January 1, 2008 and SAS has implicit functions to recognize and convert these same values of specific time frame such as hour, minute, and second. It also provides functions to convert these time records formats into specific date such as year, month, and date. We create all possible time and date variables for both the beginning and end of trips that can be obtained from the original records, in order to use them as time identifiers. We also create spatial indices for origin and destination (O-D) of each trips based on the recorded location of beginning and end of each trips.

The data has geographic coordinates, reported as longitude in numeric form as real number, and latitude, of where individual taxi trips begins and ends. This geographic coordinates can be used as spatial identifier to analyze the distribution of taxi trips over the New York City metropolitan region, but we use five-digit ZIP Code polygons and assign the matched ZIP Code number to each origin and destination coordinates. The reasons to use ZIP Code as spatial identifier are following: First, a pair of real numbers can only indicate an individual location, if the geographic coordinates are used as the identifier. In other words, two variables in real numbers are needed to provide a spatial identifier. By assigning ZIP Codes numbers for each coordinates, however, a single location can be indicated by one variable in integer form. Second, there are up to 300 million non-missing observations have geographic coordinates. The distribution pattern of this large-scale data is difficult to analyze, either visually or analytically, because there are too many neighboring points, not thus a single observation is anguish in a sea of other points.

Figure 3.5 presents identification of spatial distribution of large scale data. We take a random sample of 300k observation that have LaGuardia airport as its origin

\[[-75.1994, -72.0103] \text{ and } [39.5401, 42.0339], \text{ and the destinations are } [-79.4417, -71.9881] \text{ and } [39.6439, 42.0558] \text{ respectively.} \]
and Manhattan below at 59th street as its destination. Then we plot the data of its destination location. Panel (a) is 3D histogram with geographic coordinate grids, and panel (b) is 2D scatterplot of the geographic coordinates that use latitude and longitude as vertical and horizontal axes respectively. From panel (b), we can see how accurate the trip locations were recorded with six place precision of roughly 10 centimeter or 4 inches. The destination points generally lie along the New York City road network. However, it is hard to visually figure out distributional pattern in detail such as what is the location of the most, or the least frequent destination of the taxi. But panel (a) clearly emphasizes the distribution pattern over the somewhere Manhattan region is very dense. ZIP Code polygon are used as spatial identifiers in the geographic coordinate grid portrayed in panel (a).

3.4.2 Parallel Execution of Taxi Data Analysis

From the taxicab data, we thus obtain two matrices. One is the independent variable matrix, denoted $X$, and the other is the instrumental variable matrix, denoted $W$. These are used in a linear model estimation of taxi drivers’ trip route choice behavior as explored in Shim (2013). These two matrices have 293,693,470 rows but number of
columns is different. $X$ contains 62 variables and a vector of ones at the first column, so it has 63 columns. $W$ has 151 variables, and the vector of ones at the first column, so 152 is the total number of columns. We compute three matrix multiplications, $X^TX$, $W^TW$, and $X^TW$ using the outer product algorithm.

The multiplications are coded in Fortran 95 with message passing interface (MPI), and compiled using the Intel Fortran Compiler (version 3.1.61\textsuperscript{13}, and executed on a two cabinet Cray XE6m system in CUNY High Performance Computing Center, named Salk. Salk consists of 176 dual-socket compute nodes each containing two 8-core AMD Magny-Cours processors running at 2.3 GHz for a total of 16 cores per node. This gives the system a total of 2816 cores for the production processing of CUNY’s HPC applications. Each node has a total of 32 Gbytes of memory or 2 Gbytes of memory per core.

\textsuperscript{13}The reason to use Intel compiler is that the intel compiler is able to specify its memory model based on the size of data, which is able to be read into a head node or compute nodes. The matrix $X$ consists of 6.2 Gbytes nonzero elements, a 5.7 Gbytes row index, and a 2.9 Gbytes column index as its data files. The matrix $W$ consists of a 4.2 Gbytes nonzero elements, a 5.6 Gbytes row index, and a 2.6 Gbytes column index. Since the Intel compiler’s memory model allows to read bigger than a 2 Gbytes as a file input, therefore, large-scale data, or big data analysis demands the use of the Intel compiler’s memory model option.
3. PARALLEL SPARSE MATRIX MULTIPLICATION

3.4.3 Analysis: Taxicab data

From the three target matrices $X^T X$, $W^T W$, and $X^T W$, we obtain variance-covariance matrix of $X$ and $W$, and moment estimators of conditional expectation of a variable given time, date, and location indicator variables. Variance-covariance matrix, which also refers to as sample variance-covariance matrix, has variances at its diagonal, and covariance of the variables at its off-diagonal elements. Since the first column of $X$ and $W$ are vectors of ones, we can obtain sample mean of all variables. Hence we can compute most of descriptive statistics from the three target matrices from the first row, or the first column of the target matrices. Basic information about the target matrices $X^T X$, $W^T W$, and $X^T W$, and the associated performance measures are reported in Table 3.2. Because $W$ has 148 indicator variables, which take value one or zero, out of 152, whereas $X$ has 40 indicator variables, out of 63, the number of zero elements in $W^T W$ is larger than $X^T X$. This fact is represented by lower sparsity and higher gigaflops of $W^T W$ for all number of CPUs. The bottom panel of Table 3.2 shows that each computation’s gigaflops is almost doubled when the number of CPUs is doubled. Thus, indicating a positive unit relationship between processor frequency and calculating performance.

In addition, we also can obtain sample moment estimator of conditional expectation, if one of two variables is an indicator variable that takes either one or zero. For example, the average taxi fare of trip from JFK airport is an estimation of conditional expectation of taxi fare given the origin of trip is JFK airport, denoted $E[Y|X = 1]$, where $Y$ is taxi fare and $X$ is an indicator variable that takes one if trip origin is JFK airport, and zero otherwise. The data matrix $X$ have taxi fare and JFK airport

---

$^{14}$For example, the very first element of $X^T X$ is the number of observations 293,693,470 because it is $\sum_{i=1}^{N} 1 \cdot 1$, where $N$ is the number of observations, or number of rows in $X$. The first column in $X^T X$ has sum of columns of $X$ over its rows as $\sum_{i=1}^{N} X_{i,1}, \sum_{i=1}^{N} X_{i,2}, \ldots$. By dividing each element of the first column in $X$ with the very first element of $X$, these turn out to be sample means as $N^{-1} \sum_{i=1}^{N} X_{i,1}, N^{-1} \sum_{i=1}^{N} X_{i,2}, \ldots$. Note that $X^T X$ is a symmetric matrix so the sample means can be obtained from the first row of $X^T X$. 

Table 3.3: Means and Variance-Covariance Matrix of NYC Taxicab Data

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>Variance-Covariance Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Passenger count</td>
<td>Total fare</td>
</tr>
<tr>
<td>Passenger count</td>
<td>1.6612</td>
<td>1.4836</td>
</tr>
<tr>
<td>Total fare</td>
<td>9.4705</td>
<td>0.0911</td>
</tr>
<tr>
<td>Tolls</td>
<td>0.1208</td>
<td>0.0059</td>
</tr>
<tr>
<td>Tip</td>
<td>0.5870</td>
<td>-0.0149</td>
</tr>
<tr>
<td>Surcharge</td>
<td>0.2060</td>
<td>0.0276</td>
</tr>
<tr>
<td>Trip time</td>
<td>11.4684</td>
<td>0.2178</td>
</tr>
<tr>
<td>Vacant cruise</td>
<td>10.1295</td>
<td>0.0368</td>
</tr>
<tr>
<td>Credit card</td>
<td>0.2859</td>
<td>-0.0130</td>
</tr>
</tbody>
</table>

origin indicator variable as two of its variables, so $N^{-1} \sum_{i=1}^{N} X^T X_{fare,JFK}$, where $X^T X_{fare,JFK}$ is an element at which the row is taxi fare and the column is the JFK airport indicator.

Table 3.3 reports the mean, variance, and covariance of the independent variable matrix $X$. The first column in the table 3.3 contains the mean of variables, and the rest of columns contains the variance-covariance matrix. In the variance-covariance matrix, the diagonal elements are variances and the off-diagonal elements are covariance. We can thus obtain and calculate every independent and instrumental variables’ descriptive statistics by the following way: The mean total fare, as reported in the first column and the second row, is $9.47$, and its variance, reported at the third column and the second row, and the standard deviation is $46.26$, and $6.80$ respectively. The covariance between total fare and trip time, reported at the third column and sixth row.

Table 3.3 only reports on a part of the entire variance-covariance matrix for $X$ because the size of the matrix is actually $63 \times 63$. We report the rest of the variance-covariance matrices for $X$ and $W$ on various plots and maps to present its distributional characteristics clearly, using moment estimation of conditional expectations. The instrumental variable matrix $W$ contains indicator variables for time,
date, and O-D locations’ ZIP Codes. These conditional expectation estimates allow us to demonstrate the mean distributions of variables in both $X$ and $W$. Figure 3.6 reports (mean) distributions of taxi fare and trip time over hours in the day, and days in the week. This figure helps to emphasize the importance of controlling for the hours in the day, when performing any statistical analyses using the taxicab data because it does not seem to be distributed randomly. The lines in panel (a) fluctuate over time along certain pattern and it is not independent over time. Thus, the independent and identical distribution (i.i.d) assumption for a sample, a necessary condition to be accepted as a random sample and that provides the argument for statistically valid analysis, is violated. Thus, if a statistical analysis is conducted without controlling for hours in the day, one expects to estimate biased results especially for taxi fare and trip time.

The taxicab data thus has dependency over not only time variation but also spatial distribution. As described above, we assign five-digit ZIP Codes and use the ZIP Code as spatial index. There are 866 ZIP Codes that are assigned for both origin and destination locations. We exclude ZIP Codes that have the number of observations less than 0.1% of the entire observations in the taxi data. Figure 3.7 reports
the (mean) distributions of taxi fare and trip time over ZIP Code polygons around New York City. The map shows that taxi fare and trip time are getting higher in some particular clusters such as Manhattan, or JFK airport. This gives us the same implication that spatial distribution has to be also controlled as well as hourly distribution, in order to obtain unbiased statistical results when we analyze taxi fares and trip times.
Figure 3.7: Spatial Distribution of Taxi Fares
3. PARALLEL SPARSE MATRIX MULTIPLICATION

Figure 3.8: Spatial Distribution of Trip Time by ZIP Code Polygons

(a) Mean Trip Time at Origin

(b) Mean Trip Time at Destination
3.5 Conclusions

In this study, we introduce a parallel algorithm sparse matrix-transpose-matrix multiplication using the outer product of row vectors in matrices. The outer product algorithm works with compressed sparse row (CSR) form matrices, and it does not require transposition operations prior to perform multiplication. Since the parallel implementation of the outer product algorithm decomposes a matrix by row, it imposes no additional restrictions with respect to the size and shape of matrices. We focus on the value of this technique on rectangular shaped matrices, which have a larger number of rows and smaller number or columns, for performing statistical analysis on big data sets. We test the general performance for the outer product algorithm for randomly generated matrices. The performance measures of the test show that the outer product algorithm is well implemented and performs well on large-scale matrix multiplication in a parallel computing environment.

As we discussed above, this method can produce three computational efficiencies and advantages. First, the matrix-transpose-matrix multiplication can be done with the CSR form even though it involves transposition of a matrix. Second, the outer product algorithm has no need for inter-processor communication to exchange data during its execution, except to distributed the data to compute nodes at the beginning, and collecting outputs from the nodes at the end of the execution. The third advantage is relatively flexible allocation of the data. These three advantages make the outer product algorithm more applicable to and efficiently with large-scale statistical computing for big data. Further, we conduct an experiment of the application on New York City taxicab data. We show that the advantages of parallel execution of the outer product algorithm increases, as the size of data and the associated data matrices gets larger.

Since the concept of big data has become popular in statistics and the other sci-
ences, demand for support from high performance computing systemshas increased because of technical difficulties required to manipulate and analyze big data. The outer product algorithm provides the very beginning step in developing enhanced methods for statistical analyses for big data by giving a linkage between high performance computing and statistical computing. We thus expect the outer product algorithm to help develop applied statistical computing algorithms that will enhance high performance computing systems so that they can be used more efficiently for big data problem.
4 Principal versus Agent: Market Operation Mechanism of the New York City Taxicab Industry
4. TAXI MARKET MECHANISM

4.1 Introduction

In this paper, I apply mechanism design, a game theory that demonstrates a game situation with incomplete information between a principal who sets up the rules and the finite number of agents who play the game under the rules, in order to establish a model for taxi market resource allocation and then analyze it with asymmetric information that exists between passengers and drivers, along with heterogeneous trip characteristics, where the origin and the destination points of a trip have different general levels of quantity demanded for taxi trips at each location. The main advantage of applying mechanism design to the taxi market is that it provides equilibrium conditions by which the agents in the game situation accept an allocation mechanism, that is designated by the principal, and announce their true market behavior characteristics. In other words, we can characterize the equilibrium conditions that a taxi driver, as the agent, chooses an optimal route under a regulated fare system imposed by a taxi market authority, as the principal. The equilibrium yields a socially optimal allocation if the optimal route is a shortest route in terms of trip time and distance. We can define the resulting allocation, as a Pareto optimal allocation in the taxi market, and hence the model is able to provide implications that show as under what conditions the taxi market achieves the Pareto optimal allocation.

The New York City yellow medallion taxicab market is an example of the model. The New York City Taxicab & Limousine Commission (TLC) issues a limited number of medallion, and set up the fare rates system for NYC medallion taxicab, the yellow cab, and the other taxi services. The fare system is not a single fare rule; there are a number of different fare rules that are applied for different types of trip. Furthermore, the passengers obviously prefer to make their taxi trip shorter and lower fare, but they are subject to limited information about trip routes because the New York City’s road network is one of the most complicated road networks and its traffic condition
are dynamically changing continuously. The passengers are thus more likely to rely on information from their taxi drivers for the choice of exact trip route.

The economic literature on taxi industry tend to analyze the market under the assumption that the price of trips and the market entry vehicles into a taxi industry are regulated by a government authority. This seems to be a reasonable and realistic assumption in most cases and cities because the authority may be concerned about passenger protection from crime and risky driving behavior. Regardless of the reason for entry control of the taxi market, also known as a medallion system, moreover, it also tends to cause monopoly pricing behavior of the licensed taxi drivers, and this makes the price control, also known as a regulated fare system, a justifiable policy. The price control is therefore a necessary condition to achieve the second best efficient allocation in the taxi market, provided there exists an entry control, and thus finding an optimal taxi fare system becomes a key general approach that the taxi market literature constructs and uses to analyze taxi industry.

Leading the way to analyze taxi market, Douglas (1972), in a seminal paper that applies supply and demand analysis to the taxicab industry, models the taxi market as a monopoly and shows how the regulated price leads to the allocation of resources that become socially optimal whereas an unregulated monopoly pricing system yields an inefficient allocation. De Vany (1975) extends the supply and demand analysis for the taxi markets to show the impact of entry control on economic welfare. If the taxi market entry is regulated through a medallion system, De Vany finds it turns out to be a monopoly market so that the market allocation does not achieve its socially optimum unless the price regulation is imposed. Beesley and Glaister (1983) consider information structure on taxi market analysis when the supply and demand model has waiting time as a stochastic component. Cairns and Liston-Heyes (1994) argues that the medallion system can be implemented as a de facto bond to provide an evaluation tool of appropriate market behaviors of the drivers. Arnott (1996) argues that one
4. TAXI MARKET MECHANISM

needs to provides subsidies to taxi drivers to reduce waiting time for passengers, using
the supply and demand model.

But the optimal regulated price in taxi industry may not always lead to a socially
optimal allocation as a second best market solution because of demand uncertainty,
and also information asymmetry between drivers and passengers. For a given road
network in a particular area, taxi drivers normally have high levels of information
about the network, whereas passengers may have limited information.\footnote{The
information about road network for a particular trip can contain both time and spatially
invariant components, and time and spatially variant components. The time and spatially invariant
components of road network for an individual trip are such as route, and its toll price and distance.
And the time and spatially variant components can be congestion, travel time, and some accidental
event due to whether condition, or car accidents.} Thus, taxi
drivers have an incentive to choose a longer trip route other than the optimal, which
is the shortest distance route that takes the minimum travel time, to increase the
total fare of the trip, in the case where the passenger generally wants to go from an
origin with greater taxi demand to a destination with less demand for taxi, and the
passenger has insufficient information about the road network to identify whether the
driver goes along the optimal route. If this inference is true, any regulated fare system
fail to achieve a socially optimal allocation, where drivers transport their passengers
with as shortest trip time and distance as possible so that the passengers and the
drivers both spend the least amount of money and time on a given trip.

For the purpose of applying the theory of mechanism design to a taxi industry, in
this study, I define the utility of a taxi market authority, as a principal, and drivers,
as players in a mechanism, which is a function of trip distance that is drivers’ hidden
characteristic. Then I analyze how the equilibrium distance, and the corresponding
utilities are determined with respect to the taxi fare system, designed and imple-
mented by the authority.\footnote{In the taxi industry, the compensation mechanism is the regulated fare system and the agents are
the taxi drivers. Since the objective of TLC is to provide taxi services to the passengers as much as
they demand, and since the passengers’ objective is to minimize travel time, the fare system should
let the taxi drivers seek and take the passengers and drive along the shortest route. So the socially
optimal equilibrium is where the taxi drivers takes the shortest distance routes. This assumption

\[\text{utility} = f(\text{trip distance})\]
fare system that yields a Pareto efficient allocation, if taxi trips have heterogeneous characteristics due to the spatial distribution of passengers that is there exists different probabilities of finding a new passenger between origin and destination for the next trip. This is a key market point - as passenger may well be finished with their transport activity at the end of a given trip - but the driver is looking forward to the next trip in a series of trips during a working day. So the probability of finding a customer at the destination point may then conduct an alternative trip route behavior on the current trip.

An empirical analysis is then conducted to examine the theoretical prediction. The data for the empirical analysis is NYC taxicab trip records with individual trips’ geographic coordinates of origin and destination, number of passengers, payment method, itemized fares, trip time, and distance. There are 378,532,118 trips have been collected from January 2008 to November 2010. The empirical analysis aims to estimate a parameter that implies a difference between the equilibrium distance and the optimal distance of a taxi trip for a given set of attributes, and I use recorded distances on the taxi meters for the equilibrium distances and shortest route distances for the optimal distances of the trips. So, the difference between the recorded distance and the shortest route distance becomes an observed difference between the equilibrium and the optimal distance. I perform GMM estimation to obtain consistent estimates of the average difference of the distances in several different groups of trips that have different fare rules or clearly have different taxi demand at their origins and destinations. GMM estimation is employed to control for possible endogenous factors over time and the spatial dimension of the data. The results show that the theoretical prediction is consistent with our empirical foundation that a metered fare rule can make drivers choose the shortest routes when a given trip has more taxi demand at its destination than origin so that the fare rule does not work to prevent an inefficient will be discussed in detail later.
taxi trip, where the drivers make trip longer when they drive back from the origin to the destination, and the opposite pattern is observed for the trips in negotiated fare zones. The drivers are thus more likely to make trip shorter when their trip origin has less taxi demand than its destination.

In the next Section, the travel demand model, proposed by chapter 1, is reviewed and modified for taxi service. Then a mechanism design model that is associated with the travel demand model is established and discussed. In Section 4.3, I introduce the equilibrium conditions of the mechanism design model for taxi market and analyze how drivers’ behave differently under what circumstances. In Section 4.4, I discuss econometric issues on the empirical analysis of the taxi market mechanism using the New York City taxicab GPS tracking data. The results of the empirical analysis is discussed in Section 4.5. Then I summarize the arguments of this study in Section 4.6 as a concluding remark. Finally, mathematical proofs for a lemma, theorem, and proposition will be provided in Section 4.7.

4.2 Mechanism Design of Taxi Industry: Model

In this section, a theory of mechanism design for a taxi market is proposed and discussed. I review the travel demand model, proposed by chapter 1 and offer some reasons to justify the assumptions, which are necessary to impose for the taxi market mechanism. Then I establish the mechanism of taxicab industry based on the given assumptions.

4.2.1 Travel Demand for Taxi Service

Because taxi trip demand is the most crucial determinant not only for drivers’ route choice but also of a given taxi market authority’s fare rule set which could be selected
to maximize social welfare in the market\textsuperscript{3} we can simply think that the social welfare is a function of the authority’s utility and passengers’ utility. The authority’s utility can be maximized by providing sufficient number of taxi services to the passengers subject to the drivers’ willingness of making trips for the passengers as the way in which both drivers and passengers can maximize their utilities, and thus social welfare.

The conventional taxi travel demand model has waiting time and fare amount as its attributes\textsuperscript{4} but this does not seem to be a realistic and applicable model design for this study because i) distance, and travel time are the attributes of demand for taxi travel as important as fare, ii) waiting time is a part of the entire travel time and the entire travel time is a function of the travel distance, so the demand model with fare and waiting time can only explain a special case of the taxi trip demand, and iii) having an absence of minimum trip distance means the model cannot provide any implications on drivers’ route choice which determines the distance of a trip based on asymmetric information about road network between passengers and drivers.

The idea that the entire trip time and distance are the attributes, which determine taxi trip demand, rather than waiting time that is a part of the entire trip time, is provided by chapter 1 through their travel demand model. The baseline assumption of the model is that individuals choose a number of trips to maximize their utility, subject to their budget and travel time forecast. The utility is a function of consumption and leisure, and the consumption and leisure time are discounted by cost and time of the travel, respectively. The design model is thus a disutility of travel model in which individuals can achieve maximum utility level by minimizing travel time and cost, in order to allocate additional money and time from travel to consumption and leisure.

\textsuperscript{3}Social welfare in economics terms means the sum of consumer and producer surplus. Note that taxi market authorities’ goal may not to maximize the social welfare by designing the market operating system in an efficient way. The goal can be differed by the authorities’ own mandates such as let the drivers transport passengers safety.

\textsuperscript{4}See Douglas (1972), and its subsequent researches on taxi demand.
4. TAXI MARKET MECHANISM

The assumptions of the model are: the travel time is randomly given so that the individuals have to use their own travel time forecast, based on their previous experience that provides perfect information to the individual on prior travel events. The travel time forecast can be summarized by a time evolving function of previous travel time, a random component of travel time that is thus a stochastic process, and distance. Under these assumptions, the individual’s utility maximization problem can be described by:

$$\max U = E \left[ \sum_{t=0}^{\infty} \beta^t u(c_t - p_t \cdot d, l_t - h_t) \right]$$

s.t. $$c_t - p_t \cdot d \leq w_t^* (1 - l_t + h_t) + r_t k_t$$

$$h_{t+1} = g(h_t, d, z_t),$$

where $$u(\cdot, \cdot)$$ is an intertemporal consumption-leisure utility, $$c_t, k_t, r_t, p_t$$ is a consumption, a household capital holding, a unit price of capital, and a unit price of travel events respectively, and $$h_t, d, z_t$$ are the traveler attributes that are travel time, distance, and the stochastic component of the travel time respectively[^5]. The utility maximization problem is a stochastic dynamic programming problem, and the solution path for the travel demand for a mode $$i$$ can be derived as:

$$f(y_i|x) = \exp \left[ -E \{v(h'_t, z'_t)\} \right] \cdot \frac{E \{v(h'_t, z'_t)\}}{y_i}, \quad y_i = 0, 1, \ldots, \quad (4.1)$$

where $$y_i$$ is a number of travel by a travel mode $$i$$, $$x$$ is a vector of decision variables that

[^5]: Note that the second constraint $$h_{t+1} = g(h_t, d, z_t)$$ is a time evolving function that describes a forecasting rule of future travel time for a given information about travel time. The main idea to obtain the travel demand $$y_i$$ is to treat all traveler attributes as exogenous so the Bellman equation $$v(h_t, z_t)$$, a solution of the stochastic dynamic programming problem is given by:

$$v(h_t, z_t) = F(h_t, h_{t+1}, z_t) + \beta E_t [v(h_{t+1}, z_{t+1})], \quad \forall t$$ and $$z_t \in Q.$$  

This yields the expected return of future travel $$E_t [v(h_{t+1}, z_{t+1})]$$ and the demand $$y_i$$ is assumed to be Poisson random variable with mean $$E_t [v(h_{t+1}, z_{t+1})]$$. The solution is generally derived to solve out maximization problem of the Bellman equation, but it turns out to be a given equation by assuming exogenous attributes. The details of derivation is described in chapter 1.
have been determined exogenously, and \( \nu(h'_i, z'_i) \) is the Bellman equation, a solution of the stochastic dynamic programming problem.

The travel demand \( y_i \) is a poisson random variable with mean \( E \{ \nu(h'_i, z'_i) \} \) that denotes the expected benefit that the individual traveler expects to achieve from a future travel event. The baseline assumptions behind the count travel demand model (4.1) is that travel is time consuming and costly which therefore takes away the opportunity to consume goods and services, and also potentially wastes leisure time. The traveler is thus able to achieve maximum utility by minimizing travel distance and time. The demand model assumes Independence of Irrelevant Alternatives (IIA) that the traveler chooses a particular travel mode to minimizing expected time and distance of the travel event and this decision does not depend on any other possible alternative travel modes.

Chapter 1 assumes further that the traveler’s travel time forecast \( h_{t+1} \) and its random component on the travel time \( z_{t+1} \) are fixed over travel node, which means that travelers do not think their travel time is differed by choice of trip mode. In other words, the traveler expects to spend same travel time regardless of any transit modes such as driving, mass transit like bus and subway. This assumption is too strong to apply directly to the taxi market model because each travel mode has own trip route and the associated distance, and the traveler would recognize and consider it for her travel time forecasting. The main focus of this study, in addition, is to analyze how the social welfare which is the taxi market authority’s objective, is changed with respect to the distance of a taxi trip, given by the drivers’ route choice, and thus taxi trip demand has to be varied over the distance. For this reason, I consider a modified travel demand model that can have different distance at different travel modes. The

\[ E \{ \nu(h'_i, z'_i) \} \]

for a travel mode \( i \) is the future realization of \( h_i, z_i \) respectively. These can be thought as \( h_{t+1} \) and \( z_{t+1} \) so the expectation implies the future benefit. I use \( h'_i, z'_i \) to avoid to use messy notation with super and subscript at the same time.

Note that the travel in the model is a particular trip from an origin to a destination, so that the distance is the physical length between the origin and the destination along road network.
travel time forecast $h_{i,t+1}$ is then a function of the distance, denoted $h_{t+1}(d_i)$. By assuming the other attributes are being held constant, the taxi trip demand function can be expressed as $y(d_i)$, where $d_i$ refers to trip distance by taxi.

To construct a relevant model for the taxi market mechanism, it then needs to focus only on trip distance between origin and destination which implies that the distance represents route choice and determines taxi market equilibrium. In fact, the distance can be the single determinant for the following reasons: First, the taxi authority’s policy instrument is taxi fare rule and the rule normally calculate a trip fare based on the distance and time of a trip; Second, the absolute distance of a trip is determined immediately once the driver choose a route of the trip, given that the driver does not change the route during the trip, but the absolute travel time is uncertain due to unexpected traffic conditions. Furthermore, the other travel attributes are more likely to be given for a single trip because the other travel attributes such as wage rate, financial income and leisure time are given exogenously from the decision of the route choice for the trip. For this reason, I use the taxi trip demand as a function of distance for the following taxi market allocation mechanism.

4.2.2 Taxi Market Mechanism

Consider a taxi market for a metropolitan area that is assumed to have a finite number of Origin-Destination (OD) trips. Suppose that there are taxi drivers who participate the market as players in the region, and the types of the drivers are homogeneous, which means that all drivers have the same utility for a trip. Suppose also that there is a taxi market authority that plays a role as a principal.

---

8The list of travel attribute is given by chapter 1. They assume that the wage, financial income, and leisure time in their model are given by having a trip through a particular travel mode $i$. They also show how the number of travel through the mode $i$ are determined by the travel attributes in two propositions. In the travel demand model, travel time forecast varies over the travel modes so the travel mode choice is done by the trip time forecasting and the other attribute do not affect on the mode choice. For the demand for taxi, therefore, we can argue that travel time does not effect on the number of taxi trips because passengers consider the travel time forecast when they decide to catch a taxi, not a bus or a subway.
Assume that the authority has no private information, whereas the taxi drivers have \( \theta = (\theta_1, \ldots, \theta_I) \), a vector of distances for all possible OD trip \( i = 1, \ldots, I \). An element \( \theta_i \) of the vector \( \theta \) is the distance of a trip \( i \) that is determined when the route of the trip is chosen. The distance \( \theta_i \) is drawn from a set \( \Theta \) so \( \theta_i \) is a random variable and the set \( \Theta \) is a set in a \( \sigma \)-algebra that represents a road network of the metropolitan area, denoted \( \mathcal{R} \). Figure 4.1 shows the New York City road network and a trip route from 42nd street and Broadway, to John F. Kennedy (JFK) Airport that represents an example of the \( \sigma \)-algebra for a trip distance. The line layer and dots represent the road network, so it is \( \mathcal{R} \) for the NYC Taxi cabs. The red line is one out of a set of finite but many alternative routes for the trip on the road network. Hence, the set \( \Theta \) can be thought of as a set of all possible routes for the trip, and the distance \( \theta_i \in \Theta \), which refers to the distance of red line, can be drawn from the infinitely many different routes. Moreover, the route can be changed during the trip due to unexpected traffic conditions, so it is defined as a random variable.

Let \( y : \Theta \mapsto \mathbb{N}_0 \) be the taxi trip demand function that maps from the distance to the non-negative integer, including zero. Note that the demand function \( y(\theta) \) is the travel demand that I reviewed in section 4.2.1. Let \( t : \Theta \mapsto \mathbb{R}_+ \) be a monetary transfer from passengers to a driver so \( t(\theta_i) \) implies fare amount for taxi trip \( i \) with respect to the distance \( \theta_i \). Now we define an allocation in the taxi market, denoted

---

8 The distance vector \( \theta \) represents the driver’s characteristic, which implies individual players’ hidden behavior or action that is unknown for the principal. As will be seen below, in addition, it determines demand and the associated market allocation. Note that the number “I” implies the number of OD trips in the region.

9 As I assumed about \( \theta_i \), the distance, for a trip \( i \), and the set \( \Theta \) contains all possible distances for the trip \( i \) that are determined by the choice of route. In order to establish a model for the taxi market mechanism with respect to trip distance, the distance has to cover all possible trips in the metropolitan area, so that the road network \( \mathcal{R} \) is assumed to be a \( \sigma \)-algebra, which is a set of sets \( \Theta \). This set up allows the distance \( \theta_i \) is well-defined random variable. This is a type space in the taxi market mechanism.

10 Note that the taxi trip demand \( y(\theta) \) is a composite function that contains fare, as well as the distance so it is \( y(t(\theta), \theta) \). Since the taxi fare \( t(\theta) \) is also a function of the distance, here I use \( y(\theta) \) to make notation simpler. Because the demand for taxi is the composite function of fare and distance, in addition, the characteristic of the function with respect to the distance is not consistent. In other words, it may or may not increase as the distance gets longer.
4. TAXI MARKET MECHANISM

$x(\theta) = (y(\theta), t(\theta))$, that consists of taxi trip demand and the associated fare amount. The allocation is a major component of a mechanism that defines incentive structure of the players in the mechanism, with respect to the players’ action. In the taxi market, therefore, the allocation implies that, for all trip $i$, there exists an exchange of taxi riding services and its fare between passengers and drivers. And the total amount of the exchange is determined by distance $\theta_i$. The conventional theory of mechanism design assumes that the monetary transfer is made by a principal in the mechanism, but here I treat the transfer as being made by passengers under the fare rule, given by the authority, so that the authority can choose $t(\cdot)$ as a rule but it cannot choose the exact amount of fare for every single trip $i$.

The market allocation $x : \Theta \mapsto \mathbb{N}_0 \times \mathbb{R}_+$ allows us to argue whether a regulated taxi fare system allocates resources between passengers and drivers efficiently. In order to make an efficient allocation argument, here I consider a representative driver’s utility $u(x, \theta)$ and the principal’s utility $u_0(x, \theta)$, where $u$ is strictly increasing and $u_0$ is strictly decreasing in the monetary transfer $t(\cdot)$. The principal’s utility $u_0(x, \theta)$, as its objective in the taxi market operation, can be thought of as the social welfare.
that is given by the market allocation $x$. The authority needs to design a fare system $t(\cdot)$ that satisfies both passengers and drivers at maximum, to achieve the maximum social welfare. The utility function for the driver $u(x, \theta)$ implies that the driver prefers to earn a higher fare in a particular trip so he or she can extract a sort of profit, while the principal does not because a higher fare yields less demand for taxi, which reflects utility of passengers. Since the monetary transfer $t(\theta_i)$ is primarily a function of distance, the major determinant of utility of both principal and driver is the distance. Without loss of generality, any fare systems presented to the drivers is assumed to be a fare system that maximizes the authority’s utility, and that is the maximum social welfare of the taxi market, if the authority designs the fare system under perfect information about distances $\theta$ for all trips.

The other component of a mechanism is a message of the players that conveys the players’ hidden type to a principal. In general, a mechanism defines a message $\mu \in \mathcal{M}$, where $\mu$ is a function that represents how players announce their hidden characteristic $\theta$. Suppose that, for instance, a taxi driver makes a trip with 10 miles and the fare increases by a dollar per mile without a baseline fare. The driver can charge more than 10 dollars by cheating and adding travel distance about 10 miles in a circuitous route to his passengers. If he normally put one additional mile on the fare, his message, the announcement of the distance, is $\mu(\theta_i) = \theta_i + 1$. The message is only for a specific trip, and he might have different scheme for a different trip, so that the message can be treated as a random variable, denoted $\mu_i(\theta)$, drawn along a road network as a $\sigma$-algebra $\mathcal{M}$. The primary task in the mechanism design is thus to find a monetary transfer $t^*(\theta)$ that leads players to announce their true characteristics through the message space $\mathcal{M}$, and the existence of the monetary transfer is guaranteed by the revelation principle. But I ignore the role of message in the taxi market mechanism with the assumption that the announced distance at the end of taxi trip is to be true distance in which the driver has been driven through a
4. TAXI MARKET MECHANISM

The next is to define an objective of the principal and the players in the mechanism. Let $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_I)$ be a vector of all possible alternative trips’ distance for a trip $i$. Since the utility level is conditional on utilities of all possible alternative trips in which these utility levels depend upon the probability of finding passengers, and since the authority has an uncertainty about how drivers choose the distance of trip $i$, the objective function is an expected utility, rather than a deterministic, we can think of this as applying a Bayesian game with incomplete information. Here, I consider a representative driver’s expected utility, for a given allocation $\{x(\theta)\}_{\theta \in \Theta}$, as $E_{\theta_{-i}} [u(x(\theta_i, \theta_{-i}), \theta_i, \theta_{-i})]$, and the principal’s expected utility is $E_{\theta} [u_0(x(\theta_i, \theta_{-i}), \theta)]$. The driver’s expected utility can be interpreted as the driver’s belief about benefit from a trip $i$ with the distance $\theta_i$, when the distances of the other trips are given by $\theta_{-i}$. In the same way, the expected utility can be interpreted as the authority’s belief about benefit of the entire taxi market from the trip $i$ with the distance $\theta_i$, when distances of all trips $\theta = (\theta_1, \ldots, \theta_I)$ are given. Note that the utilities of both the authority and the driver are determined by the market allocation $\{x(\theta)\}_{\theta \in \Theta}$ that consists of demand for taxi trip $y(\theta_i)$ and the associated fare $t(\theta_i)$. In other words, the benefit of a trip for drivers and the authority in a taxi market is given by the demand and the fare system.

---

11 This type of mechanism is called a direct revelation mechanism in which the type space $\mathcal{R}$ is the message space $\mathcal{M}$ in the taxi market. In other words, the driver always charge the fare based on true distance. The meter based taxi fare system is able to achieve this goal because it calculate the trip distance automatically so the driver has no way to cheat on his passengers about the trip distance. Since the most of metropolitan taxi authorities have conducted the meter based system, the assumption that the message space in the taxi market mechanism is the type space itself is justifiable.

12 I use the notation of Fudenberg and Tirole (1991). They interpret it as “the expectation is taken over agent $i$’s beliefs about $\theta_{-i}$ conditional on his type $\theta_i$.” See p.270 of Fudenberg and Tirole (1991).
4.3 Mechanism Design of Taxi Industry: Analysis

In this section, I introduce equilibrium conditions in a general mechanism design model and discuss how it can be applied to the taxi market analysis. With the equilibrium conditions, I analyze a taxi market under the notion that the spatial distribution of passengers is non-uniform over a region so the probability of finding the next passengers after a given taxi trip is different at different locations and times. The spatial distribution plays a role to consider as heterogeneous trip characteristics at various locations, where origin and destination have different quantity demanded for taxi trip so that the drivers can discrimination against certain trips not only by the distance but also the probability to fine the next passenger at the destination of a trip. In order to demonstrate drivers’ route choice and the associated utility for a given fare system, I define a concept of a well-defined market, and provide a mathematical logical evaluation as to how the trip distance that is given by the trip route choice determines market allocation, which are summarized by Lemma 4.3.2.1 and Theorem 4.3.2.1. Finally I show a unique and socially efficient single taxi fare system cannot exist for the heterogeneous trip characteristics, which is presented in Proposition 4.3.2.1.

4.3.1 Equilibrium of the Mechanism

Because the demand cannot be directly controlled by the authority, the authority can achieve its maximum utility through the fare system, and the Pareto efficiency of the fare and the associated allocation \( \{x(\theta)\}_{\theta \in \Theta} \) can be argued. The theory of mechanism design provides a condition in which players in a mechanism are like to accept a monetary transfer, announce their true characteristics to the public, and eventually implement a decision function that is proposed by the principal. This can
be summarized by the following inequality:

\[ E_{\theta_i} \left[ u(x(\theta_i, \theta_{-i}), \theta_i, \theta_{-i}) \right] \geq E_{\theta_i} \left[ u(x(\hat{\theta}_i, \theta_{-i}), \theta_i, \theta_{-i}) \right], \quad \forall i = 1, \ldots, I \]  

(4.2)

where \( \hat{\theta}_i \in \Theta_i \) is an announced type of the player \( i \), drawn from \( \Theta_i \), a set of all possible types of the player. This inequality is called incentive compatibility (IC) condition, and since the IC is defined on the expected utility, it is a Bayesian IC.

In order for an equilibrium to exist, in a game situation, the utility function as a payoff of players has to be bounded. The Bayesian IC (4.2) plays a role to specify an upper bound of the players’ utility. And the lower bound is given by a feasibility of the payoff with respect to strategy profile, called individual rationality (IR). Here, I state an IR condition for taxi drivers as:

\[ E_{\theta_i} \left[ u(x(\theta_i, \theta_{-i}), \theta_i, \theta_{-i}) \right] \geq u, \]  

(4.3)

where \( u \) is a reservation utility of drivers. It can be thought as a fixed cost of a trip, if the utility is monetary profit of the trip. By the two boundary conditions IC and IR, the utility can be defined as a bounded function \( u : \Theta \mapsto [u, u(\hat{\theta}_i)] \), so the expected utility \( E_{\theta_i} \left[ u(x(\theta_i, \theta_{-i}), \theta_i, \theta_{-i}) \right] \) does.

Let \( \theta_i \) be the optimal distance along the optimal route for a trip \( i \), and let \( \hat{\theta}_i \) be a distance that maximizes driver’s utility. The implication of the Bayesian IC in the taxi market mechanism is that the drivers are willing to take passengers, given by the demand \( y(\theta_i) \), under a fare system \( t(\theta_i) \) from the authority. The fare system, in addition, eliminates any incentives to make distance \( \hat{\theta}_i \) differ from the optimal distance \( \theta_i \) so the driver always drives along the optimal route. In other words, the fare system \( t(\cdot) \) yields higher utility for the drivers when they make a trip through a route with \( \theta_i \) than any other trip routes with \( \hat{\theta}_i \). The taxi demand function \( y(\theta_i) \) is called implementable if the fare system satisfies the Bayesian IC condition. The taxi
market allocation \( \{x(\theta)\}_{\theta \in \Theta} \) is called *truthfully implementable* if the demand function is implementable and the message space of the mechanism that is associated with the allocation is the type space.\(^{13}\)

The theory also provides a criterion for the principal that is able to evaluate whether the market allocation \( \{x(\theta)\}_{\theta \in \Theta} \) yields a socially optimal allocation. Holmström and Myerson (1983) proposes a definition that an incentive compatible allocation is *ex-ante efficient* if and only if there does not exist any incentive compatible allocation that yield better outcomes for both players and the principal.\(^{14}\) The term *ex-ante efficiency* allows an analysis of the model to examine whether an allocation in a taxi market mechanism achieves the socially optimum, by implementing proper license issuing and fare charging system. If the taxi market authority designs an optimal fare system and equips it to its drivers properly, for example, the system leads the drivers always taking the optimal route so the passengers are able to create taxi demand based on a certain belief that taxi trips are always made along the optimal route.

### 4.3.2 Analysis of Taxi Fare System

In this section, I introduce the concept of a well-defined taxi market, and demonstrate taxi drivers’ attitude towards the risk of vacancy cruising and the effect of their attitude towards this on the efficiency of allocation. Then I examine under what condition a unique and ex-ante efficient allocation can exist. The underlying reasoning is that the drivers have a preference to be in locations when they are looking for

---

\(^{13}\)The property is called *revelation principle* for Bayesian Nash Equilibrium in which the allocation \( x(\theta_i, \theta_{-i}) \) satisfies the Bayesian IC \(^{12}\), and the revelation principle that provides enough incentive to the players announcing their true characteristics. See p.256 of Fudenberg and Tirole (1991)

\(^{14}\)Holmström and Myerson (1983)’s definition covers *interim* and *ex-post* efficiency allocation. The term *ex-ante* is more appropriate for taxi market because a taxi trip is made by passengers who decides to take taxi, out of the other travel mode such as driving their own car, or various mass transit modes. The *interim* and *ex-post* efficiencies are infeasible, or ineffective for the passengers’ travel mode decision, and thus for the taxi market mechanism it would be better to argue about *ex-ante* efficiency.
passengers based on prior beliefs about the probability of finding their next potential passengers. These prior beliefs can be modeled as a spatial distribution of potential taxi passengers over the region. Furthermore, there exists a distinction between more or less preferable locations, as long as the distribution is not uniform. The concept of a well-defined taxi market is thus to distinguish the taxi drivers’ preference over the spatial distribution into the most preferable and the least preferable locations, and model how the drivers behave differently across the different locations.\footnote{The distinction, made by introducing the well-defined taxi market, can also be interpreted as a consideration of taxi drivers’ heterogeneous behavior on location choice. Treating all taxi trips within a region as a single homogeneous trip is thus unable to model the taxi drivers’ clustering behavior toward several trip dense locations.}

A well-defined taxi market is a place where drivers can find passengers to fill up empty trips with the least risk. Let the spatial distribution of potential passengers be the most dense in the well-defined market, and dispersed over the other locations. This spatial distribution implies that the probability of finding the next passengers near the place of destination of a trip gets smaller as the trip distance from the well-defined market gets further. This set-up provides an identical model of taxi trip to the conventional theory of choice under uncertainty. Let $f(\cdot)$ be a probability (spatial) distribution function of finding next passenger. Now I construct the mathematical definition of well-defined taxi market as follows:

**Definition 4.3.2.1 (Well-defined Taxi Market).** An origin (a destination) of a taxi trip $i$ is called a “well-defined taxi market” if the probability distribution function $f(\theta_i)$ is monotonic decreasing (increasing) in $\theta_i$ for a destination (origin) of the trip $i$.

Figure 4.2 presents the distribution of NYC yellow cab trip origins in Manhattan below 59th street. The most dense location in the Panel (a) plot is at 42nd street and Broadway, the Times Square area, and it is an example of the most well-defined taxi market in the region because no other location has higher frequency of passengers in the region.\footnote{The data of the figure 1 is given by a random sample with 300,000 number of observations,}
4. TAXI MARKET MECHANISM

Assume that a fare system for a taxi market with a fixed initial fare. It is clear that a fare system should be increasing in trip distance. For instance, the unit fare system, the most general regulated taxi fare system, is a uniformly increasing function of trip distance, whereas the drivers’ expected utility is monotonically increasing, in general, by the definition of risk aversion. In the taxi market mechanism, as I established in Section 4.2.2, the taxi fare for a trip $i$ is characterized by $t(\theta_i)$, the monetary transfer rule that is given by the taxi authority, and the taxi driver of trip $i$ has a utility $u_i(x(\theta_i, \theta_{-i}), \theta_i, \theta_{-i})$, which is strictly increasing in the given fare $t(\theta_i)$. From now, let denote $t(\theta_i)$ be the unit fare, and $u(\theta_i)$ be the driver’s utility for a trip $i$. Recall $\hat{\theta}_i$, the announced type of player $i$ for a given mechanism. Let $\theta_i$ and $\hat{\theta}_i$ be the optimal and actual distance from the origin to the destination of a trip $i$. The optimal distance $\theta_i$ is the distance that yields the maximum utility to the passengers for a given fare $t(\theta_i)$, and it determines demand for taxi trip $y(\theta_i)$. The actual distance $\hat{\theta}_i$ is the distance that the driver selects for the trip $i$ by her route choice, in order to achieve the maximum utility. Assume that the optimal distance $\theta_i$ is common knowledge that drawn from New York City Taxicab Limousine Commission’s GPS tracking data during 2008 to 2010. Note that I use the standard random sampling without any stratification for the entire region of NYC yellowcabs, and draw the histogram and the scatter plot only for midtown and downtown Manhattan from 300,000 random sample.
In order to build a model with heterogeneous characteristics of taxi trip, here I define probability distribution functions for the trip from a well-defined market to a non well-defined market, denoted $f(\theta)$, and the trip from a non well-defined market to the well-defined market, denoted $g(\theta)$. By the definition 2.3.1, $f(\theta)$ is monotonic decreasing, and $g(\theta)$ is monotonic increasing in distance $\theta$. As we can see in Figure 4.3, the probability of finding the next passengers gets smaller as the distance gets further from the well-defined market, presented in panel (a), and the probability gets larger as the distance gets further from non well-defined market.\[17\] Let $F(\cdot)$ and $G(\cdot)$ be cumulative distribution functions of $f(\cdot)$ and $g(\cdot)$ respectively. These spatial distributions allow the drivers to define expected return finding a frequency of fast fare of taxi trips, under a given fare system, as $\int t(\theta)dF(\theta)$ for the trip from a well-defined market to a non well-defined market, and $\int t(\theta)dG(\theta)$ for the trip from a non well-defined market to a well-defined market.

The assumption behind the expected return is the following: the taxi drivers expect monetary income from a trip based on the given fare and the future uncertainty.

---

17 Note that longer distance in $g(\theta)$ implies that the trip gets closer to the well-defined market so the probability goes up as the distance gets far from the non well-defined market.
about having vacant cruise for an extended period of time. So, the drivers’ trip
decision is made by comparison between trip fare, the expected return fare, and their
own utility. Thus, a driver extracts a fare rent from his passenger in case where he
feels the trip length is sub-optimal given his utility. Assume that the drivers have no
right to refuse passengers for any trip $i$, so that they are only able to fit their utility
by adjusting the trip distance $\hat{\theta}_i$. Let $B(X)$ be a space of the utility function, which is
bounded. The necessary conditions for the existence of $\hat{\theta}_i$ in taxi market mechanism
are summarized by following lemma:

**Lemma 4.3.2.1 (Existence of the trip distance $\hat{\theta}_i$).** If the fare system $t(\theta)$ is defined
on the space $B(X)$ and increasing in trip distance $\theta$, then:

(a) there exists a unique $\hat{\theta}_i$ such that

$$u(\hat{\theta}_i) = \int t(\theta_i)dF(\theta_i),$$

(4.4)

for all trip $i$ with the probability distribution function $f(\theta_i)$.

(b) In the same way, there exists a unique $\hat{\theta}_j$ such that

$$u(\hat{\theta}_j) = \int t(\theta_j)dG(\theta_j),$$

(4.5)

for all trip $j$ with the probability distribution function $g(\theta_j)$.

Lemma 4.3.2.1 provides a necessary condition to build the model for the taxi drivers’
trip distance decision based on the comparison between $\theta_i$ and $\hat{\theta}_i$. Recall the Bayesian
IC (4.2). By examining the inequality of (4.2), whether a given fare system leads the
drivers to choose the optimal route, which yields the optimal distance, can be argued.
Since the actual distance $\hat{\theta}_i$ for a trip $i$ is determined by the condition (4.4) or (4.5),
we can demonstrate under what conditions \( \hat{\theta}_i \) satisfies the Bayesian IC. This analysis is summarized by the following theorem:

**Theorem 4.3.2.1** (Truthful Implementation of Taxi Demand). The taxi demand function \( y(\theta) \), and the associated taxi market allocation \( x(\theta) = (y(\theta), t(\theta)) \) is truthfully implementable, if the fare system \( t(\theta) \) yields

(a) for all trip \( i \) with \( f(\cdot) \),

\[
\hat{\theta}_i \geq \theta_i \text{ such that } u(\hat{\theta}_i) = \int t(\theta_i) dF;
\]

(b) for all trip \( j \) with \( g(\cdot) \),

\[
\hat{\theta}_j \geq \theta_j \text{ such that } u(\hat{\theta}_j) = \int t(\theta_j) dG.
\]

The implication of Theorem 4.3.2.1 can be interpreted with the graphs in Figure 4.4. As shown in Panel (a), the drivers can earn more than they expect if the driver’s utility \( u(\hat{\theta}_i) \) is defined below at the fare at \( \theta \), which is \( t(\theta_i) \). In this case, the drivers have no incentive to make a longer trip than the optimal, and it satisfies the Bayesian IC (4.2). On the other hand, the case where the utility \( u(\hat{\theta}_i) \) is defined above at the fare at \( \theta \) that is \( t(\theta_i) \), as shown in Panel (b) of Figure 4.3, leads the drivers to make longer trip than the optimal so it does not satisfy the Bayesian IC. Non truthfully implementable taxi demand implies that the drivers are likely to make longer distance trips than the optimal, and passengers who know about this tendency do not take a taxi for their trip, while passengers who do not recognize this tendency might still use a taxi service. This reflects the adjustment of the demand \( y(\theta) \), and it is obviously decreasing in this case.

In order to argue about efficiency of a taxi fare system, here I consider a round
4. TAXI MARKET MECHANISM

Figure 4.4: Comparison and Distance Determination

trip that goes from an origin to a destination, then goes back to the origin. Note that the actual passengers in the trip do not necessarily have to be the same for both legs of the trip, so that the drivers face the prospect of finding passengers on the way back to the origin when they arrive the destination. Suppose that there exists a fare system $t^*(\theta_i)$ that yields the expected return, which is the same as the drivers’ utility $u(\theta_i)$, for all trip $i$ with $f(\theta_i)$. Without loss of generality, the fare system $t^*(\cdot)$, as a monetary transfer in a taxi market mechanism, yields ex-ante efficient allocation for trip $i$. But it is not for the other trip $j$ with $g(\theta_j)$, as long as the trip $j$ has a different probability distribution. This analysis is summarized by the following proposition:

**Proposition 4.3.2.1.** If the cumulative probability distribution $F(\cdot)$ is the first order stochastic dominance of the cumulative distribution function $G(\cdot)$, then a unique fare system $t^*(\cdot)$, which yields ex-ante allocation, does not exist, for any round trips with $F(\cdot)$ and $G(\cdot)$.

The proposition provides an analysis of taxi market that an inflexible fare system

---

18Round trip helps to build up a model for taxi market mechanism when the spatial distribution of passengers matter. In other words, the characteristics of trip is heterogeneous due to non-uniform distribution of passengers over the region, and this yields different incentive structure to the drivers for each trip. So the efficiency of a fare system can be compared between trips with different probabilities of return fares.
cannot fit a market with heterogeneous taxi trips due to the spatial distribution of passengers.\footnote{Note that this does not necessary to be a support argument of negotiated fare system because it is affected by information structure in which I do not impose on the taxi market mechanism model. This analysis cannot expand to further details about how a negotiated fare work, without careful discussion and consideration of the information structure between drivers and passengers.} Now I consider a round trip between well-defined taxi markets, as an implication that the spatial distribution is uniform. This analysis requires an additional assumption that the trip has an origin and a destination, which have the same probability, and the probability does not change over the distance. This uniform distribution is identical to both ways for any round trips with the above characteristic so it is $F(\theta) = G(\theta)$ for all $\theta$. Even though the round trip case between well-defined markets does not imply there always exists a unique $t^*(\cdot)$ with ex-ante efficient allocation, it shows the $t^*(\cdot)$ can possibly exist on the round trip with $F(\theta) = G(\theta)$.

Proposition 4.3.2.1 and the analysis of the round trip with $F(\theta) = G(\theta)$, for all $\theta$, shows that the necessary condition for the existence of an ex-ante efficient market mechanism through a particular fare system $t^*(\cdot)$ is to equalize the probabilities of finding the next passengers at both origin and destination. In other words, providing clustered locations for taxi passengers can make the socially efficient market allocation through $t(\cdot)$. This analysis is thus providing the implication that turning a non well-defined market into a well-defined market should be the authority’s policy of highest priority, because this allows the authority finds a fare system that makes passengers and drivers better off by reducing inefficiency of their taxi market allocation.\footnote{Flexible fare system might not be a feasible alternative, in order to achieve the ex-ante efficient allocation because it charges different trip fares between passengers who make the trip from a well-defined taxi market to a non well-defined market, and the passengers from a non well-defined market to a well-defined market, even if their trips have the same trip length.}
4.4 Issues in Empirical Analysis

In this section, I discuss issues in an empirical analysis to examine the theoretical predictions, which is addressed by the mechanism design model for taxi market in Section 4.4.3. First I derive an observable econometric model that is associated with the Bayesian incentive compatibility condition (4.2). Then I discuss characteristics of the New York City TLC’s taxicab trip data and variables that I use in the econometric model estimation. Finally, given the scale of NYC’s yellow taxi markets, I introduce a computational method, which is implemented in a parallel computing system, for large-scale generalized methods of moment estimations of the econometric model.

4.4.1 An Observable Model for Taxi Market Mechanism

As I discussed in the section 4.2.1, taxi trip demand \( y(\theta_i) \) is implementable if it satisfies the Bayesian incentive compatibility condition (4.2). The implementability of a taxi trip demand means that a driver is willing to make a trip for passengers in the region that a given fare system satisfies the inequality (4.2). Let \( \delta \) be a parameter that is the difference between the left and the right hand side of the inequality (4.2), defined as:

\[
E_{\theta_{-i}} [u(x(\theta_i, \theta_{-i}), \theta_i, \theta_{-i})] - E_{\hat{\theta}_{-i}} [u(x(\hat{\theta}_i, \theta_{-i}), \theta_i, \theta_{-i})] = \delta, \quad \forall i = 1, \ldots, I, \tag{4.6}
\]

where \( \hat{\theta}_i \) is the actual distance and \( \theta_i \) is the optimal distance of a trip \( i \). The parameter \( \delta \) offers us a way to examine whether a particular taxi fare system yields \textit{ex-ante} efficient allocation and implements taxi demand by leading drivers to choose an optimal route, and \( \delta \) becomes positive. On the other hand, a negative \( \delta \) implies the fare system is \textit{ex-ante} inefficient and the fare system cannot implement the taxi trip demand truthfully. By estimating \( \delta \) and evaluating its sign, therefore, a partic-
ular taxi fare system can be examined empirically whether it leads drivers to choose optimal routes.

In order to identify the parameter $\delta$, the driver’s expected utility function $E_{\theta_{-i}} [u(\cdot, \theta_{-i}), \theta_i, \theta_{-i}]$, which is a composite function of the expectation $E_{\theta_{-i}} [\cdot, \theta_i, \theta_{-i}]$, the utility $u_i(\cdot, \theta_{-i})$, and the allocation $x(\cdot, \theta_{-i})$, has to be identified, and the identification of the composition function using taxi trip data is irrelevant. For this reason, I consider another way that specifies $\beta_0 = \theta_i - \hat{\theta}_i$, the difference between the optimal distance $\theta_i$ and the actual distance $\hat{\theta}_i$ of a trip $i$. The actual distance $\theta_i$ is longer than the optimal distance $\hat{\theta}_i$ if a given fare system is ex-ante inefficient, and therefore, the sign of $\beta_0$ is consistent with the $\delta$ in (4.6). The conditional expectation equation that identifies $\beta_0$ given a fare rule, information over time and space is then:

$$E \left[ \theta_i - \hat{\theta}_i \bigg| t(\theta_i), \Omega_s, \Omega_t \right] = \beta_0, \quad \forall i = 1, \ldots, I, \quad (4.7)$$

where $t(\theta_i)$ is total fare, for a trip $i$ with $\theta_i$, $\Omega_s$ is an information set for the spatial distribution of passengers, and $\Omega_t$ is information set of time. Note that the spatial information set $\Omega_s$ can be thought as a road network because it has to be observable and known to all drivers. By Theorem 4.3.2.1, $\hat{\theta}_i$ is the actual distance of the taxi trip $i$ that gives equivalent utility to the expected fare of $\theta_i$, the optimal distance, conditional on the spatial distribution of passengers. So $\hat{\theta}_i$ can be treated as a distance from the driver’s utility $u(\cdot, \theta_{-i}), \theta_i, \theta_{-i}$, and since $\theta_{-i}$ contains the optimal distances of all other trips, the model needs to consider the road network $\Omega_s$ to control for $\theta_{-i}$, the optimal distances of all other trips.\footnote{The spatial distribution of taxi passengers is not directly observable because it is based on individual drivers’ experiences, but it can be assumed that is indirectly observable by the combination of time and road network. For instance, taxi drivers have their way to obtain information about where and when they can most likely to find passengers. So the place where a driver is at can be chosen based on time and the location. In general, theory of trip is selected by the driver based on their expectation of cruising for a fare and precise destination is selected by the passenger.} With data of the road network for $\Omega_s$, time for $\Omega_t$, and the fare system $t(\cdot)$, the parameter $\beta_0$ can be correctly identified and
4. TAXI MARKET MECHANISM

estimated via the model (4.7).

4.4.2 Data

In order to estimate \( \beta_0 \) to compare the effects of various fare rules on trip distance, I use New York City Taxicab trip record data. The New York City Taxi & Limousine Commission (TLC) has selected an electronic taximeter system, under a program named the Taxicab Passengers Enhancement Project (T-PEP) and these meters are now deployed in all its medallion taxicabs. The main purpose of installing the new meters was to provide credit card payment as an option for the taxicab passengers. Monitoring drivers’ individual operation is another desired goal of the project so the T-PEP system records every single trip made by all New York City medallion taxicab drivers, and reports the record to the TLC’s server immediately. The record contains geographic coordinates of both origin and destination of individual trip with the date and time of when the trip began and ended. It also includes travel time, distance, and the associated itemized taxi fares. This study utilizes 378,532,118 trips records that have been collected from January 2008 to November 2010.

From the taxi trip record data, I can have \( \hat{\theta}_i \), the actual trip distance in the region for individual trip \( i \) from the data because taxi fare is calculated based on the recorded distance and travel time. But the data has no appropriate measure of \( \theta_i \), the optimal distance in the taxi market. So I use the shortest distance between an origin and a destination (O-D) on the road network around the New York City metropolitan area, under the assumption that the optimal distance, as common knowledge for everyone who is involved in taxi market, is the shortest distance because travel time is uncertain due to the unexpected traffic condition. The shortest distance along the road network for an individual trip can be obtained on geographic polygons and the associated road network layer. The reason for using geographic polygons, rather than to use geographic coordinate themselves, is to reduce the number of all possible
4. TAXI MARKET MECHANISM

(a) TAZ Polygons and its centroids.
(b) Centroid connectors

Figure 4.5: Traffic Analysis Zone (TAZ): John. F. Kennedy Airport

O-D combinations. Note that the number of combinations is almost infinite if the geographic polygons are not used, so that the $3.78^2 \times 10^{16}$ pairs of distances would have to be calculated to cover the New York City metropolitan region.\(^{22}\)

The geographic polygons that I use to calculate the optimal distance $\theta_i$ is Traffic Analysis Zone (TAZ) polygons from the New York Metropolitan Transportation Council (NYMTC)’s Best Practice Model, a Geographic Information System (GIS) model for analyses of New York City metropolitan area’s transportation systems. Figure 4.5 shows the TAZ system that I use for the distance calculation. There are two more alternative standards of geographic polygon, which are zipcode and Census tract, but TAZ polygon is the most detailed geographic region in terms of the size because a TAZ polygon is bounded by road and street network. In other words, a single TAZ polygon does not contain any streets and roads inside, so that it can be used to analyze transportation systems without eliminating any road network information and they are bounded by the road network. Applying TAZ polygons plays a role of subdividing an interval of the geographic coordinates, into a manageable number of

\(^{22}\)The origin and the destination location of an individual taxi trip have been recorded as geographic coordinates, longitude and latitude. Each record takes six decimal places in the original dataset and the interval of the longitude and latitude of the origins are $[-75.1994, -72.0103]$ and $[39.5401, 42.0339]$, and the the destinations are $[-79.4417, -71.9881]$ and $[39.6439, 42.0558]$ respectively.
4. TAXI MARKET MECHANISM

(a) TAZ Polygons and its centroids.  
(b) Centroid connectors

Figure 4.6: Traffic Analysis Zone (TAZ): John. F. Kennedy Airport

strata, which indicate street and road blocks. The trip within a TAZ polygon is assumed to have the same optimal distance so that the number of the optimal distances trip pairs (origin and destination) that cover the entire region becomes manageable.

The way to calculate a road network O-D distance is following: First, compute geographic centroid of every polygon. Then calculate linear distance from the centroid to the closest point on the road network. This linear distance is called centroid connector, and it plays a role of imposing a penalty on the optimal distance calculations. Finally, calculate a distance along the road network between the points where centroid connector of the origin polygon touches the road network, and the point in which the destination’s centroid connector touches the road network. Eventually we get the shortest distance by adding the road network distance with two centroid connectors’ distances.

Panel (a) of Figure 4.6 shows an example of centroid of a polygon, and the asso-

\footnote{The NYC taxicab data requires 3,430 and 3,445 TAZ polygons for origin and destination respectively to compute the optimal distances that corresponds to each observations. The number of distance pairs is thus 11,816,350. The manageable number means that 11,816,350 number of distances are more easy to calculate than 378 million squared, the number of O-D pairs in the data.}

\footnote{The centroid connector can be more than two within a polygon. It depends on how many roads and streets the polygon has so the number of centroid connectors and the number of roads and streets are matched to eliminate the irrelavent distance calculation case in which there is a road closer to the centroid but the connector links to a distant road.}
ciated centroid connector is presented in Panel (b). Panel (b) also shows an example of the optimal distance calculation scheme from Grand Central Train Station to JFK Airport in New York City. The meaning of the optimal distance can be varied by considering many different traffic condition factors such as congestion in certain areas, rush hour travel, and so on. But these factors are all randomly and exogenously given for most taxi passengers, and it is difficult to present prior to when a passengers plan to make a taxi trip. This ambiguity of traffic condition is the reason to use the shortest-physical distance as a measure of the optimal distance. The measurement error, caused by eliminating traffic conditions, is assumed to be random across the road network.

4.4.3 Large Scale GMM Estimation

In this section, I introduce a large scale linear Generalized Method of Moments (GMM) estimation procedure for the New York City taxicab data. Recall that the observable model for the Bayesian IC condition (4.6). That is:

$$E \left[ \theta_i - \hat{\theta}_i \mid t(\hat{\theta}_i), \Omega_s, \Omega_t \right] = \beta_0.$$  (4.8)

The ideal outcome of $\beta_0$ is zero which implies that the fare system $t(\cdot)$ is well designed and implemented for drivers, so they do not need to make longer trips than optimal. By assuming that the parameter $\beta_0$ is constant in the presence of the proper control variables for fare system $t(\cdot)$, information set over time $\Omega_t$, and spatial distribution $\Omega_s$, the model (4.8) becomes an econometric model for $\theta_i - \hat{\theta}_i$ with constant parameter $\beta_0$.

In order to estimate $\beta_0$ consistently, therefore, here I consider a linear model that is associated with the observable model (4.8) as:

$$y_i = \beta_0 + x_i' \beta + z_i' \gamma + \epsilon_i,$$  (4.9)
4. TAXI MARKET MECHANISM

where \( y_i = \theta_i - \hat{\theta}_i \) is the difference between the optimal distance and the recorded distance, \( x_i \) is a vector of control variables for fare, and \( z_i \) is a vector of control variables for time and date of a trip \( i \), so that \( z_i \in \Omega_t \). I obtain \( y_i \) by subtracting the recorded distance in the taxi data from the calculated O-D distance that corresponds to the trip. \( x_i \) consists of all itemized fares and its related variables such as payment method, specialized fare zone (dummies), number of passengers, vacancy time from the end of previous trip. \( z_i \) consists of hours, weekday, and month dummy variables, which are recorded when a trip begins based on origin time and date.\(^{26} \)

These \( y_i, x_i, \) and \( z_i \) can be obtained from the data but \( x_i \) obviously depends on the location and route of the trip, which is represented by the information set \( \Omega_s \), drawn from spatial distribution of passengers. To control for endogeneity of \( x_i \), caused by the absence of the information set \( \Omega_s \), I consider a vector of instrumental variables \( w_i \) that belong to the information set and are exogenous to the error term, such that \( w_i \in \Omega_s \) and \( E[\epsilon_i|w_i] = 0 \). The NYC taxicab data has geographic coordinates of origin and destination locations, and since the origin and destination of a trip are determined solely by passengers, it is exogenous to the drivers. Furthermore, the location of origins and destinations is a part of the road network, so the vector of instrumental variables with location indicators can be \( E[\epsilon_i|w_i] = 0 \) and \( w_i \in \Omega_s \). The location indicators, which are used as instrumental variables, are dummy variables for ZIP Code polygons of origin and destination locations.\(^{27} \)

\(^{25}\)The simplest estimator of \( \beta_0 \) is the sample mean but this estimator may be inconsistent due to model misspecification that comes from the following way. Suppose that \( \beta_0 \) is estimated by sample mean as \( \hat{\beta}_0 = N^{-1} \sum_i (\theta_i - \hat{\theta}_i) \). This simple estimation requires the structural model \( \theta_i - \hat{\theta} = \beta_0 + \epsilon_i \) with the assumption that the error term \( \epsilon \) has mean zero, so \( \epsilon \) is eliminated asymptotically when the sample mean operator is taken. The justification of the sample mean as a consistent estimator of \( \beta_0 \) is then proved by assuming that the error term \( \epsilon_i \) has mean zero, but this assumption does not make sense as the way how the parameter \( \beta_0 \) is derived theoretically.

\(^{26}\)Since the data also contains time and date, recorded at the destination of a trip, \( z_i \) can either be dummies for origin, or destination time and date, or both. But most trips end up within an hour so that origin time and destination time are virtually equivalent. And since the model is to analyze \textit{ex-ante} efficiency, I choose the origin time and date as control variables for the information set \( \Omega_t \).

\(^{27}\)Note that \( w_i, \) the vector of instrumental variables, contains \( z_i, \) the vector of time and date indicators, under the assumption that the time and date of a trip are also exogenously given to the drivers.
The data only have geographic coordinates of origin and destination points of individual trip. There are several ways to use these coordinates to control for the location and time simultaneously. I use ZIP Code polygons of corresponding origin and destination coordinates under the assumption that drivers are able to figure out the location and the associated trip route immediately after they take passengers into the taxi, and that procedures are assumed to be based on ZIP Code-wise level spatial distribution, which is able to contain buildings and streets, unlike TAZ polygons.

The reason to use ZIP Code polygons, rather than TAZ, is that the number of observations within a TAZ polygon is quite small and becomes too small to compute a matrix inverse of $\sum_i w_i' w_i$, especially for low frequency origin and destination area. By expanding the geographic polygon a bit, we solve this computation issue. Since ZIP Code polygons are wider than TAZ, each polygons contain more observations so it provide safer indicator variables against the computational failure of the matrix inversion. By using ZIP Code polygons, in addition, the number of indicator variables as instruments can be reduced. There are 865 5-digit ZIP Code polygons that corresponds to 3,430 and 3,445 TAZ polygons for origin and destination respectively, in the region where I map the taxi data and analyze it. So there are 3,430 times 3,445 of origin and destination dummy variables for TAZ polygons becomes 865 times 865 ZIP Code polygon dummies. The matrix computation for the estimation is thus to be much easier when the instrumental variables for ZIP Code polygons are used.

The most notable issue on estimation of the parameters $\beta_0$, $\beta$, and $\gamma$ in (4.9) is size of the data. The entire sample has 293.7 million observations, and this number of observations takes 1.51 gigabytes for a variable. As I used in the estimation, 152 instrumental variables are being used and a matrix of the instrumental variables in dense form takes 229.52 gigabytes. There is no programming routine that can handle

\footnote{Note that the estimation of the parameters $\beta_0$ and $\beta$ in (4.9) requires to compute inverse of $\sum_i w_i' w_i$ as an instrumental variable estimation. The numerical computation of the inverse matrix might fail if some of indicator vectors in $w_i$ have too few observation of ones, out of the entire observations because an algorithm for the matrix inversion may recognize the matrix as a singular.}
this kind of large scale data matrix on a serial computing machine that has a single processor. So I develop a sparse matrix multiplication programming subroutine and apply it to the GMM estimation. The subroutine allows to implement large scale matrix multiplication on a parallel computing system via a language-independent parallel programming and communication protocol such as Message Passing Interface (MPI). Let $\theta$ be a vector of the parameters in (4.9), $x_i$ and $w_i$ be a vector of independent variables, and a vector of instrumental variables respectively. The GMM estimator of $\theta$ is then given by:

$$
\hat{\theta} = \left( \sum_{i=1}^{N} x_i'w_i \right) \hat{\Sigma} \left( \sum_{i=1}^{N} w_i'x_i \right)^{-1} \left( \sum_{i=1}^{N} x_i'w_i \right) \hat{\Sigma} \left( \sum_{i=1}^{N} x_i'y_i \right),
$$

(4.10)

where $\Sigma$ is a weighting matrix that controls for heteroskedasticity of the error term.\(^{29}\)

The estimation requires to compute a transposed matrix-matrix multiplication $\sum_{i=1}^{N} w_i'x_i$, a vector-matrix multiplication $\sum_{i=1}^{N} w_i'y_i$, and a matrix-diagonal-matrix multiplication $\sum_{i=1}^{N} u_i^2 \cdot w_i'w_i$. In the entire sample case, the multiplication $\sum_{i=1}^{N} w_i'x_i$ becomes a transposed of 293,693,496 by 62 matrix multiplies a 293,693,496 by 152 matrix.

The subroutine and the associated parallel computing program work for large scale linear algebraic operation. Since $x_i$ and $w_i$ contain many indicator variables, the matrices contain many zero elements, so the subroutine bring computational

\(^{29}\)The ways to obtain the optimal weighting matrix for linear GMM estimation begins with computing residuals from two-stage least square (2SLS) estimator $\hat{\theta}$ as:

$$
\hat{\theta} = \left( \sum_{i=1}^{N} x_i'w_i \right) \left( \sum_{i=1}^{N} w_i'w_i \right)^{-1} \left( \sum_{i=1}^{N} x_i'w_i \right) \left( \sum_{i=1}^{N} w_i'y_i \right).
$$

Then the (2SLS) residuals can be computed as $\hat{u}_i = y_i - x_i\hat{\theta}$. With the residual, the optimal weighting matrix $\hat{\Sigma}$ is given by:

$$
\hat{\Sigma} = \left( N^{-1} \sum_{i=1}^{N} u_i^2 \cdot w_i'w_i \right)^{-1}.
$$

Note that a matrix representation of $\hat{\Sigma}$ is $W'\Lambda W$, where $\Lambda$ is a diagonal matrix with $\Lambda_{ii} = u_i^2$, for all $i = 1, \ldots, N$. Computation of $\hat{\Sigma}$ is thus done by applying the parallel sparse matrix multiplication subroutine with a diagonal matrix that is in between given two matrices.
efficiency by treating the matrices as a sparse form.\footnote{30} The other advantage is parallel implementation of the computation. Even if the matrices are stored in a sparse form, they are still too large to handle by a serial computing system or programming, which deals with a computation problem in a single computing processor, because of memory limitations.\footnote{31} Note that Compressed Sparse Row (CSR) form, a sparse matrix storage form that I use in the subroutine, of $x_i$ requires 15.95 gigabytes, and 13.44 gigabytes memory space for $w_i$. The parallel computing divides the matrix computation into several pieces such that the output matrix of each pieces does not depend on the others, the machine then performs the computation independently at the assigned processors, and then collects the output matrix to the assigned processors. Parallel computing thus allows that each computing processor is assigned a manageable size of matrix so that the parallel computing algorithm can handle a large scale matrix computation problem.

### 4.5 Empirical Analysis

From the observable model \[ (4.8) \] and the Theorem 4.3.2.1, two theoretical predictions about fare system can be addressed using the parameter $\beta_0$. A negative $\beta_0$ estimate implies that a given fare system does not satisfy the Bayesian IC so the drivers are likely to make longer trip than optimal. In the same way, a positive $\beta_0$ implies that a fare system provides enough incentive to the drivers to make the optimal length

\footnote{30}{Scalable Linear Algebra Package (ScALAPACK), the most widely used library of high-performance linear algebra routines for distributed-memory message-passing MIMD, does not provide sparse matrix computation. There are some scientific computation programming libraries that support sparse matrix computation in parallel but no libraries support a transposed matrix-matrix multiplication because once a matrix is stored in sparse form, then it is hard to transpose.}

\footnote{31}{The memory size that requires to deal with a real number, as known as floating point number, is 4 bytes on a 32-bit computing system, and 8 bytes on a 64-bit system. Since the data matrix of the independent variable vector $x_i$, for the entire sample, has $293,693,496 \times 62$ numbers of elements, it requires approximately 67.83 gigabytes in 32-bit system, and 135.66 gigabytes in 64-bit system. The maximum memory allocation per processor within a CPU is known to be 32 gigabytes, and therefore, no serial computing system are capable of holding the data matrices of $w_i$, $x_i$, and the vector of dependent variable $y_i$, and performing computation at the same time.}
of trip. Under the predictions around \( \beta_0 \), I test the truthful implementability of the several different taxi fare systems under different trip characteristics using the New York City medallion taxicab data, introduced in Section 4.3.2.

### 4.5.1 Identifying (Natural) Experimental Fields

During 2008 to 2010, when the data had been collected, the New York City medallion taxicab had unit fare rule for a trip within New York City, and negotiated fare rule for a trip beyond the City. The New York City TLC specified also some special fare rules for some particular trips such as flat fare between New York County (Manhattan) and JFK airport. The variety of NYC medallion taxicab’s fare rule provides several testable hypotheses about how the fare rules effect drivers’ behavior differently. First, since the NYC taxicab had both unit and negotiated fare zones, the estimates of \( \beta_0 \) from both zones allow to compare which fare system best implements the demand for taxi service. Second, it also has flat fare zone so the comparison between unit and negotiated fare can be extended and examined the ex-ante superiority of a system. Finally, the data is capable of identifying trips i) from a (relatively) well-defined market to a (relatively) non well-defined market, ii) from a non well-defined market to a well-defined market, and iii) between well-defined markets.

Furthermore, this identification of the different trip characteristics can be applied on the different fare rules. For example, finding passengers inside New York City is easier than outside of the City, and the trip beyond the City has negotiated fare rule, so that the trip from the City to the outside of the city can be examined as a trip for which the origin is a well-defined market and the destination is a non well-defined market, under negotiated fare system. The opposite way trip can be examined as a trip for which the origin is a non well-defined market and the destination is a well-defined market.

---

\[ \text{32 These particular fare rules that have been affected during 2008 to 2010 are written in chapter 54 of the NYCTLC rule book. The rules were effective for the trip between Newark airport and the City, Westchester County and other locations.} \]
Table 4.1: Frequency of Taxi Trips in New York City Boroughs (Weekdays)

<table>
<thead>
<tr>
<th>Borough</th>
<th>Origin</th>
<th>Destination</th>
<th>Origin</th>
<th>Destination</th>
<th>Origin</th>
<th>Destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Richmond (Staten Island)</td>
<td>2,358</td>
<td>3,971</td>
<td>2,857</td>
<td>9,403</td>
<td>12,582</td>
<td>40,130</td>
</tr>
<tr>
<td>Kings (Brooklyn)</td>
<td>239,129</td>
<td>367,607</td>
<td>528,837</td>
<td>1,674,867</td>
<td>2,493,394</td>
<td>7,126,259</td>
</tr>
<tr>
<td>Bronx</td>
<td>29,340</td>
<td>98,659</td>
<td>35,628</td>
<td>163,444</td>
<td>156,802</td>
<td>784,234</td>
</tr>
<tr>
<td>Queens</td>
<td>990,575</td>
<td>1,362,050</td>
<td>1,584,558</td>
<td>1,471,592</td>
<td>6,777,447</td>
<td>8,003,375</td>
</tr>
<tr>
<td>New York (Manhattan)</td>
<td>35,589,615</td>
<td>34,970,168</td>
<td>50,095,673</td>
<td>48,894,388</td>
<td>201,330,894</td>
<td>194,533,903</td>
</tr>
<tr>
<td>New York City</td>
<td>36,851,017</td>
<td>36,802,455</td>
<td>52,247,553</td>
<td>52,213,694</td>
<td>210,771,119</td>
<td>210,487,901</td>
</tr>
<tr>
<td>Others</td>
<td>119,180</td>
<td>167,742</td>
<td>223,382</td>
<td>257,241</td>
<td>854,216</td>
<td>1,137,434</td>
</tr>
</tbody>
</table>

Notes: The proportions of trips in each New York City boroughs are reported in parentheses, and the proportions in square brackets are of the entire trips that are observed at each time durations. The peak time in morning is between 6 am and 9 am, and the peak time in evening is between 5 pm and 8 pm.

well-defined market, under the negotiated system.

4.5.2 Additional Aspects of Heterogeneous Trip Characteristics—Time of Day

In Table 4.1, the observed frequency of the New York City taxicab trips by the city boroughs and the City metropolitan areas at different times are reported. The frequency in New York County (Manhattan) at all time, reported in the fifth column, are about 95% of the entire trips that occur within the New York City. This shows that no other New York City borough have better taxi market than Manhattan in terms of probability to find taxi services so it is the most well-defined market in the City. As the same way, the frequency of the whole New York City takes 99% of the entire trips around the city metropolitan area at all time so it is the well-defined market, from
Another remarkable observation is the difference of frequency between origin and destination in the region. Manhattan and the rest of New York City, which I refer to as the well-defined market, have significantly less frequency of origin than destination, while other regions have higher frequency of destination than origin. This is an indirect evidence of excess supply for Manhattan and the New York City as the well-defined markets, and excess demand for the other regions, which are implying the taxi market allocation between drivers and passengers, is socially inefficient.

Figure 4.7 presents unconditional distributions of taxi trips around New York City metropolitan area. The polygons in the maps are TAZ blocks, and high density TAZ block shows up being red. The remarkable pattern that is observed in Table 4.1, most of trips were reported in Manhattan, seems to be consistent with the pattern in Figure 4.7. The distribution of trip destination in Panel (b) of Figure 4.7 is a little more dispersed over the outside of Manhattan than the origin distribution, which is

\[ \chi^2 \] statistics for 2-way contingency table for morning peak time, from the first to the second columns in Table 4.1, of the New York City boroughs, from the first row to the fifth row in Table 4.1, is 129,218.356. The table for evening peak time, and weekday are 700,292.996, and 2,882,135.957 respectively. The \( \chi^2 \) statistics of the table between New York City and the other regions, the sixth and the seventh rows in Table 4.1, for morning, evening peak time, and weekday, are 8251.21, 2396.28, and 40464.77 respectively. The p-values of all \( \chi^2 \) statistics are approximately zeros, regardless of degrees of freedom.
4. TAXI MARKET MECHANISM

the pattern that is observed in Table 4.1 too. The distinct pattern of Figure 4.7 shows up at the two airports in New York City. The two TAZ blocks at the outside of Manhattan with highest density, red colored, are LaGuardia and JFK airports. It is obvious that large number of passengers are willing to go, or find taxi services at an airport but those two airports are the unique TAZ blocks, however, which have the highest density of taxi trips outside of Manhattan. The two unique TAZ blocks provide chances to demonstrate taxi drivers’ behavior on a particular trip characteristic such as a trip between a well-defined and another well-defined market, or a well-defined and a non well-defined market. Moreover, since the trip between JFK airport and Manhattan has fixed fare rule, the trip characteristic under fixed and unit fare system can be compared with a distinction of the trip between well-defined and well-defined, and well-defined and non well-defined market.

4.5.3 Model Estimation: The Main Result

The GMM estimates of the model \( (4.9) \) with seven different subsamples are reported in Table 4.2. The sample selection strategy is based on the theoretical prediction, discussed in section 4.3.2, that taxi drivers are likely to behave differently at the different locations where fare systems and spatial distribution of passengers differ. Recall that the fare system of within New York City is unit fare, between the city and beyond the city is negotiated fare, and there are two special fare zones that the fixed fare between Manhattan and JFK airport, and from the city to Westchester or Nassau county. Note that the instrumental variables are the five-digit ZIP Code dummies of both origin and destination that have greater than 0.1% of observations, in order to avoid computational failure. As the second to the last column reported,

\[ \text{The observed sample of taxi trip between well-defined markets is the trip records of Manhattan and JFK airport, or LaGuardia airport. From the econometric perspective, this particular trip can hardly be identified within Manhattan because the TAZ blocks are all neighboring each other so treating the trip within Manhattan as the trip between well-defined market and another well-defined market. In other words, those two TAZ blocks might be in a single well-defined market.} \]
4. TAXI MARKET MECHANISM

the set of instrumental variables for every sample satisfies exclusion restriction that implies the instrumental variables were appropriately chosen.

The first row of Table 4.2 reports GMM estimates of the target parameter $\beta_0$. The $\beta_0$ estimate of the entire sample, in the first column, is negative and it is statistically significant. To take the theoretical prediction, given by Theorem 4.3.2.1, into account, negative $\beta_0$ implies the inequality $\hat{\theta}_i > \theta_i$, so the taxi fare system fails to satisfy the Bayesian IC (4.2) in the sample region. The New York City TLC’s fare system is therefore not able to control its drivers’ choice of longer route than optimal. The $\beta_0$ estimate of the sample within the city in the second column of Table 4.2 is also negative with strong statistical significance. Unlike the other regions, the Manhattan trip sample yields positive and significant $\beta_0$ estimate that implies the satisfaction of the Bayesian IC (4.2). In other words, the drivers of the New York City taxicabs can obtain sufficient utility level from a trip with an optimal route only for Manhattan related trips, under the given New York City TLC’s designated fare system. In overall metropolitan region, on the other hand, the drivers need to choose a longer route than optimal, in order to achieve sufficient utility level from the trip.

The frequency pattern, presented in Table 4.1 and Figure 4.7 that 95% of trips occurred in Manhattan, and 99% in the city, this can thus be explained in part by the above analysis. In the overall New York City metropolitan region, taxi drivers are likely to choose longer route that increases fare amount and travel time. If the passengers know this fact, they are likely to choose the other travel modes that provides mobility with shorter travel distance, so the overall demand for taxi in the region may decrease. The consequence of this scenario is therefore clustering into the region where the drivers can achieve sufficient utility level with an optimal distance route, and Manhattan is the region, out of any other New York City metropolitan area.

However, this interpretation of $\beta_0$ needs to consider also the other explanatory
### Table 4.2: GMM Estimates of Taxi Trip Attributes

<table>
<thead>
<tr>
<th></th>
<th>Entire Sample</th>
<th>New York City Related Trip</th>
<th>Manhattan Related Trip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Within</td>
<td>From</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NYC</td>
<td>NYC</td>
</tr>
<tr>
<td>( \hat{\beta}_0 )</td>
<td>-0.5009</td>
<td>-0.1696</td>
<td>0.8144</td>
</tr>
<tr>
<td></td>
<td>(0.044)</td>
<td>(0.003)</td>
<td>(0.190)</td>
</tr>
<tr>
<td>Number of Passengers</td>
<td>-0.0890</td>
<td>-0.0570</td>
<td>0.1521</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>Total Fare</td>
<td>-0.1187</td>
<td>-0.1008</td>
<td>-0.0657</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>Itemized Fare (Toll)</td>
<td>0.1117</td>
<td>-0.3006</td>
<td>-1.2376</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.002)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>Itemized Fare (Tip)</td>
<td>-0.0311</td>
<td>0.0066</td>
<td>-0.0959</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.000)</td>
<td>(0.004)</td>
</tr>
<tr>
<td>Itemized Fare (Surcharge)</td>
<td>4.6212</td>
<td>2.8913</td>
<td>-8.3353</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td>(0.015)</td>
<td>(3.966)</td>
</tr>
<tr>
<td>Trip Time</td>
<td>0.0273</td>
<td>0.0124</td>
<td>0.1585</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>Time Gap</td>
<td>0.0076</td>
<td>0.0015</td>
<td>0.0349</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Payment Method (Credit)</td>
<td>-0.0890</td>
<td>-0.0842</td>
<td>2.2890</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.001)</td>
<td>(0.073)</td>
</tr>
<tr>
<td>Fixed Fare Zone (From JFK Airport)</td>
<td>0.9717</td>
<td>0.3573</td>
<td>-2.2239</td>
</tr>
<tr>
<td></td>
<td>(0.039)</td>
<td>(0.012)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>Fixed Fare Zone (To JFK Airport)</td>
<td>0.4800</td>
<td>1.2023</td>
<td>-2.3145</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.008)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>Special Fare Zone (From Newark Airport)</td>
<td>-1122.6</td>
<td>8.0926</td>
<td>-3.0177</td>
</tr>
<tr>
<td></td>
<td>(23.33)</td>
<td>(0.874)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Special Fare Zone (To Newark Airport)</td>
<td>60.089</td>
<td>3.6383</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.992)</td>
<td>(0.021)</td>
<td></td>
</tr>
<tr>
<td>Special Fare Zone (To Westchester)</td>
<td>21.285</td>
<td>-3.9340</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.206)</td>
<td>(0.059)</td>
<td></td>
</tr>
<tr>
<td>Special Fare Zone (To Nassau)</td>
<td>-80.985</td>
<td>-5.3379</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.713)</td>
<td>(0.072)</td>
<td></td>
</tr>
<tr>
<td>Hour Fixed Effect</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Weekday Fixed Effect</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Month Fixed Effect</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Hansen’s J</td>
<td>0.0584</td>
<td>0.0745</td>
<td>0.0446</td>
</tr>
</tbody>
</table>

Note: Parameter estimates are GMM estimators with 5-digit ZIP Codes origin, and destination indicator dummy variables each of which possess more than 0.1% number of observations as instrumental variables. The asymptotic standard error estimates of the optimal GMM estimators are reported in parentheses. Degrees of freedoms for overidentification test are reported in square brackets. The Chi-squared statistics for overidentification restriction are reported in the last row. The degrees of freedom, number of instrumental variables minus number of independent variables are 90, 92, 110, 115, 71, 107, and 65 respectively.

Base seasonality categories for month, day of week, and hour of day which are control dummy variables were chosen in a given sample by selecting the category with the lowest frequency of trips for that period. Table 4.6 provides a matrix of dummy variables by category.
variables in the model because the $\beta_0$ is a constant parameter estimate in the regression models that this study examines. Trip extension occurs when the overall effect is negative and so we should examine this composite effect. For example, in the first column of Table 4.2, the coefficient of the variable “number of passengers” is -0.089, so more passengers in the vehicle increases the propensity to extend the trip by about 1/10th of a mile for each additional passenger. This extension in trip is capped by the capacity of the cab and would top out at around 4/10ths of a mile with 4 passengers in the cab (generally the maximum capacity). Similarly, if we examine the payment of fare by credit card, we find that by paying by credit card, we see an extension of trip length by a little less than 1/10th of a mile. Since fare payment is assigned a 1 for credit card trips, payment by card increases expected trip extension.

In the same way, the coefficient of the total fare is -0.1187, and this implies that a driver’s willingness to extend a trip increases as a total fare for the trip increases. Higher total fare trips may hide from users this trip extension activity as there is perhaps greater potential for minor route deviations that adds travel distance to a trip. This may be less observable over a long journey as opposed to a short trip with a clear direct route. Drivers may exploit this situation and extend trips.

The third and fourth columns of Table 4.2 are the samples for the trips beyond the New York City. Clearly, the trip from the city is the trip between a well-defined market to a non well-defined market, and the trip to the city is an opposite case. The $\beta_0$ estimates of those two sample regions are both significantly positive but the magnitudes differ. The trips to the city have about ten times greater $\beta_0$ than the trips from the city, which means the drivers in the trip moving to the city center tend to choose as short as possible route. In the seventh and the eighth columns, I examine how the utilities that are given by the $\theta_i$ and $\hat{\theta}_i$ are differed under unit fare system. The $\beta_0$ estimate in the seventh column of which the sample represents trips from a well-defined market is about four times greater than the $\beta_0$ of eighth column, for the
trip to the well-defined market, so that it is Pareto superior. The comparison of $\beta_0$ estimates provide the evidence that unit fare system is more suitable for the trip to a well-defined market and the negotiated fare system is more suitable for the trip from a well-defined market.

This analysis is consistent with the theoretical prediction, summarized by Proposition 4.3.2.1, that there is no unique fare system that yields a Pareto optimal allocation for any round trip where the market has different probability of finding new passengers at origin and destination at various locations. Under negotiated fare rule, drivers would be able to offer higher fare for the trip to a non well-defined market, when they find the passengers desire these trip. On the way back from the trip, however, the driver are willing to make a trip back to the well-defined market regardless of fare amount so the negotiated fare would not work well on the way back trip. On the other hand, under a unit fare system, drivers cannot offer a fare for a trip to a non well-defined market as high as they desire to make the trip so the drivers do not prefer these trips, unlike passengers who prefer to make taxi trips with known predicted fare amount. But the trip from the destination to the origin, both drivers and passengers are better off because the drivers little care about fare of the current trip than the previous one.

4.5.4 Time and Spatial Variation of $\hat{\beta}_0$

The GMM estimates of mass transit terminal dummies are reported in Table 4.3. Each of the six estimates of the JFK airport trip dummies at the second and third row of the table is statistically significant and has opposite sign to the corresponding $\beta_0$ estimate. Moreover, the magnitudes are all greater or equal to the $\beta_0$ estimate. Recall that trip between JFK airport and Manhattan has fixed taxi fare rate, so

---

35The mass transit terminal indicators are TAZ id indicators that contain the terminals in their polygons. All of each terminal take place in a whole TAZ polygon, so the observed trip within that TAZ polygons can be identified as the taxi trip that has originated from or is destined to the terminal.
the JFK related trips leads the drivers choosing an optimal distance route in both unit fare, represented in the second column, and fixed fare zones. LaGuardia airport related trip with Manhattan has also opposite and statistically significant coefficients to the corresponding $\beta_0$ estimate. But in the entire region and within NYC, the coefficients are either have the same sign or are too small or too big in scale to deflate the magnitude of $\beta_0$. Newark airport, which is located outside of New York City, has the same pattern in its coefficient estimation that is either too big or small to deflate $\beta_0$. The other bus and train terminals in Manhattan such as Pennsylvania Station, Grand Central, and Port Authority bus terminal, have the same sign of each $\beta_0$ estimates.

The three airports around the New York City are a representative example of well-defined taxi market, so the coefficients of the airport indicators in different region provides an implication on the characteristic of trip between well-defined markets. That is, fixed fare rule would work well on the trip between well-defined markets, in order to deflate longer trip distance than optimal and the associated market inefficiency. This is an supporting evidence of the theoretical prediction, addressed by Proposition 4.3.2.1 because both argue that high enough probabilities of finding new passengers at origin and destination of a trip rules out the role of the fare system on route choice. In other words, drivers will go along the optimal route regardless of what fare rule is applied for the trip if they are quite certain to find the next passengers at the destination of the trip who are coming back to an optimal origin area, without significant vacant cruising time during the round trip. Unlike the fixed fare rule, however, the unit fare rule between well-defined taxi markets does not work to deflate the gap parameter $\beta_0$. As shown by sign and magnitude of Newark and LaGuardia airports’ coefficients, these trips have either opposite sign, or too big magnitude to adjust the gap parameter $\beta_0$ as close as zero, the ideal value.

Table 4.4 represents hourly variation of $\beta_0$ estimates in different regions. Note
Table 4.3: GMM Estimates of Mass Transit Terminal Indicators

<table>
<thead>
<tr>
<th></th>
<th>Entire Sample</th>
<th>New York City Related Trip</th>
<th>Manhattan Related Trip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Within</td>
<td>From</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NYC</td>
<td>NYC</td>
</tr>
<tr>
<td>$\hat{\beta}_0$</td>
<td>-0.5009</td>
<td>-0.1696</td>
<td>0.8144</td>
</tr>
<tr>
<td></td>
<td>(0.044)</td>
<td>(0.003)</td>
<td>(0.190)</td>
</tr>
<tr>
<td>From JFK Airport</td>
<td>0.9717</td>
<td>0.3573</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.039)</td>
<td>(0.012)</td>
<td></td>
</tr>
<tr>
<td>To JFK Airport</td>
<td>0.4800</td>
<td>1.2023</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.008)</td>
<td></td>
</tr>
<tr>
<td>From LaGuardia Airport</td>
<td>-0.8254</td>
<td>0.0709</td>
<td>1.0268</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.005)</td>
<td>(0.106)</td>
</tr>
<tr>
<td>To LaGuardia Airport</td>
<td>-0.9922</td>
<td>0.0064</td>
<td>-48.4769</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.005)</td>
<td>(0.840)</td>
</tr>
<tr>
<td>From Newark Airport</td>
<td>-1122.6</td>
<td></td>
<td>8.0926</td>
</tr>
<tr>
<td></td>
<td>(23.33)</td>
<td></td>
<td>(0.874)</td>
</tr>
<tr>
<td>To Newark Airport</td>
<td>60.089</td>
<td></td>
<td>9.7205</td>
</tr>
<tr>
<td></td>
<td>(0.992)</td>
<td></td>
<td>(1.801)</td>
</tr>
<tr>
<td>From Grand Central</td>
<td>-0.0571</td>
<td>-0.0545</td>
<td>-0.8953</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.205)</td>
</tr>
<tr>
<td>To Grand Central</td>
<td>-0.0309</td>
<td>-0.0228</td>
<td>3.5470</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.000)</td>
<td>(1.483)</td>
</tr>
<tr>
<td>From Port Authority</td>
<td>-0.5686</td>
<td>-0.4414</td>
<td>0.4463</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.003)</td>
<td>(0.535)</td>
</tr>
<tr>
<td>To Port Authority</td>
<td>-0.4544</td>
<td>-0.2327</td>
<td>-7.9999</td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td>(0.003)</td>
<td>(5.253)</td>
</tr>
<tr>
<td>From Penn Station</td>
<td>-0.0384</td>
<td>-0.0360</td>
<td>0.0541</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.000)</td>
<td>(0.143)</td>
</tr>
<tr>
<td>To Penn Station</td>
<td>-0.0169</td>
<td>0.0080</td>
<td>3.2632</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.000)</td>
<td>(0.995)</td>
</tr>
</tbody>
</table>

Note: Parameter estimates are GMM estimators with 5-digit ZIP Codes origin, and destination indicator dummy variables each of which possess more than 0.1% number of observations as instrumental variables. The asymptotic standard error estimates of the optimal GMM estimators are reported in parentheses.

Base seasonality categories for month, day of week, and hour of day which are control dummy variables were chosen in a given sample by selecting the category with the lowest frequency of trips for that period. Table 4.6 provides a matrix of dummy variables by category.
### Table 4.4: GMM Estimates of Peak Time Indicators

<table>
<thead>
<tr>
<th>Time</th>
<th>Entire Sample</th>
<th>New York City Related Trip</th>
<th>Manhattan Related Trip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Within</td>
<td>From</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NYC</td>
<td>NYC</td>
</tr>
<tr>
<td>( \hat{\beta}_0 )</td>
<td>-0.5009</td>
<td>-0.1696</td>
<td>0.8144</td>
</tr>
<tr>
<td></td>
<td>(0.044)</td>
<td>(0.003)</td>
<td>(0.190)</td>
</tr>
<tr>
<td>7 am</td>
<td>1.4335</td>
<td>0.9190</td>
<td>-1.3219</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.004)</td>
<td>(0.168)</td>
</tr>
<tr>
<td>8 am</td>
<td>1.4537</td>
<td>0.9420</td>
<td>-1.7212</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.004)</td>
<td>(0.178)</td>
</tr>
<tr>
<td>9 am</td>
<td>1.4284</td>
<td>0.9282</td>
<td>-2.0391</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.004)</td>
<td>(0.175)</td>
</tr>
<tr>
<td>6 pm</td>
<td>-0.9539</td>
<td>-0.5624</td>
<td>1.6226</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.004)</td>
<td>(0.206)</td>
</tr>
<tr>
<td>7 pm</td>
<td>-0.9433</td>
<td>-0.5640</td>
<td>2.3440</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.004)</td>
<td>(0.211)</td>
</tr>
<tr>
<td>8 pm</td>
<td>0.1183</td>
<td>0.0951</td>
<td>0.9142</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.001)</td>
<td>(0.181)</td>
</tr>
</tbody>
</table>

Note: Parameter estimates are GMM estimators with 5-digit ZIP Codes origin, and destination indicator dummy variables each of which possess more than 0.1% number of observations as instrumental variables. The asymptotic standard error estimates of the optimal GMM estimators are reported in parentheses.

Base seasonality categories for month, day of week, and hour of day which are control dummy variables were chosen in a given sample by selecting the category with the lowest frequency of trips for that period. Table 4.6 provides a matrix of dummy variables by category.
that the coefficients of omitted hours in the table have consistent patterns that it is normally negative from midnight to early morning, and positive at evening before midnight. The overall pattern of the estimates is that it has statistically significant opposite sign in morning, and it changes to the same sign of $\beta_0$ in evening, except within Manhattan and to NYC trips. This represents dynamic changes of demand for taxi service over time within a day. The trip with a coefficient that turns $\beta_0$ out to be positive has increasing demand, and vice versa. By applying this logic, the entire region, within NYC, to NYC, and within Manhattan trips are identified as the region where taxi trip demand goes up in morning peak hours, then goes down in evening peak hours. It seems to reflect the difference of residence characteristic between urban and suburban areas. New York City, in general and Manhattan in particular are an urbanized region where people who live at outside of the region come for work; and this pattern is captured as the time variation in Table 4.4.

Table 4.5 represents weekday variation of $\beta_0$. The magnitudes of weekday indicators from Monday to Friday are too small to conclude that there is significant dynamics along weekdays but weekend, in the entire region, within Manhattan, from Manhattan trips. The other regions have either opposite way of the magnitude, or similar to each other. Because the signs are not consistent, there is no justifiable pattern but it is clear that the demand for taxi services are different between weekday and weekend.
### Table 4.5: GMM Estimates of Weekday Indicators

<table>
<thead>
<tr>
<th></th>
<th>Entire Sample</th>
<th>New York City Related Trip</th>
<th>Manhattan Related Trip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within NYC</td>
<td>From NYC</td>
<td>To NYC</td>
</tr>
<tr>
<td>$\hat{\beta}_0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.5009</td>
<td>-0.1696</td>
<td>0.8144</td>
</tr>
<tr>
<td></td>
<td>(0.044)</td>
<td>(0.003)</td>
<td>(0.190)</td>
</tr>
<tr>
<td>Sunday</td>
<td>0.7320</td>
<td>0.4653</td>
<td>-0.0631</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.002)</td>
<td>(0.085)</td>
</tr>
<tr>
<td>Monday</td>
<td></td>
<td>1.1211</td>
<td>-37.1793</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.102)</td>
<td>(1.010)</td>
</tr>
<tr>
<td>Tuesday</td>
<td>-0.0063</td>
<td>0.0095</td>
<td>1.3105</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.000)</td>
<td>(0.106)</td>
</tr>
<tr>
<td>Wednesday</td>
<td>-0.0041</td>
<td>0.0130</td>
<td>1.1570</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.000)</td>
<td>(0.104)</td>
</tr>
<tr>
<td>Thursday</td>
<td>0.0238</td>
<td>0.0334</td>
<td>1.2636</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.000)</td>
<td>(0.102)</td>
</tr>
<tr>
<td>Friday</td>
<td>0.0134</td>
<td>0.0237</td>
<td>0.8717</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.000)</td>
<td>(0.104)</td>
</tr>
<tr>
<td>Saturday</td>
<td>0.7052</td>
<td>0.4492</td>
<td>4.4011</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.002)</td>
<td>(0.499)</td>
</tr>
</tbody>
</table>

Note: Parameter estimates are GMM estimators with 5-digit ZIP Codes origin, and destination indicator dummy variables each of which possess more than 0.1% number of observations as instrumental variables. The asymptotic standard error estimates of the optimal GMM estimators are reported in parentheses.

Base seasonality categories for month, day of week, and hour of day which are control dummy variables were chosen in a given sample by selecting the category with the lowest frequency of trips for that period. Table 4.6 provides a matrix of dummy variables by category.
4.6 Conclusion

In this paper, I apply mechanism design, a game theory that demonstrates a game situation with incomplete information between a principal who sets up the rules and the finite number of agents who play the game under the rules, in order to establish a model for taxi market resource allocation and then analyze it with asymmetric information that exists between passengers and drivers, along with heterogeneous trip characteristics, where the origin and the destination points of a trip have different general levels of quantity demanded for taxi trips at each location. The main advantage of applying mechanism design to the taxi market is that it provides equilibrium conditions by which the agents in the game situation accept an allocation mechanism, that is designated by the principal, and announce their true market behavior characteristics. In other words, we can characterize the equilibrium conditions that a taxi driver, as the agent, chooses an optimal route under a regulated fare system imposed by a taxi market authority, as the principal. The equilibrium yields a socially optimal allocation if the optimal route is a shortest route in terms of trip time and distance. We can define the resulting allocation, as a Pareto optimal allocation in the taxi market, and hence the model is able to provide implications that show as under what conditions the taxi market achieves the Pareto optimal allocation.

For the purpose of applying the theory of mechanism design to a taxi industry, in this study, I define the utility of a taxi market authority, as a principal, and drivers, as players in a mechanism, which is a function of trip distance that is drivers’ hidden characteristic. Then I analyze how the equilibrium distance, and the corresponding utilities are determined with respect to the taxi fare system, designed and implemented by the authority.\footnote{In the taxi industry, the compensation mechanism is the regulated fare system and the agents are the taxi drivers. Since the objective of TLC is to provide taxi services to the passengers as much as they demand, and since the passengers’ objective is to minimize travel time, the fare system should let the taxi drivers seek and take the passengers and drive along the shortest route. So the socially} Using this model I show that there does not exist a
unique fare system that yields a Pareto efficient allocation, if taxi trips have heterogeneous characteristics due to the spatial distribution of passengers that is there exists different probabilities of finding a new passenger between origin and destination for the next trip. This is a key market point - as passenger may well be finished with their transport activity at the end of a given trip - but the driver is looking forward to the next trip in a series of trips during a working day. So the probability of finding a customer at the destination point may then conduct an alternative trip route behavior on the current trip.

An empirical analysis is then conducted to examine the theoretical prediction. The data for the empirical analysis is NYC taxicab trip records with individual trips’ geographic coordinates of origin and destination, number of passengers, payment method, itemized fares, trip time, and distance. There are 378,532,118 trips have been collected from January 2008 to November 2010. The empirical analysis aims to estimate a parameter that implies a difference between the equilibrium distance and the optimal distance of a taxi trip for a given set of attributes, and I use recorded distances on the taxi meters for the equilibrium distances and shortest route distances for the optimal distances of the trips. So, the difference between the recorded distance and the shortest route distance becomes an observed difference between the equilibrium and the optimal distance. I perform GMM estimation to obtain consistent estimates of the average difference of the distances in several different groups of trips that have different fare rules or clearly have different taxi demand at their origins and destinations. GMM estimation is employed to control for possible endogenous factors over time and the spatial dimension of the data. The results show that the theoretical prediction is consistent with our empirical foundation that a metered fare rule can make drivers choose the shortest routes when a given trip has more taxi demand at its destination than origin so that the fare rule does not work to prevent an inefficient

optimal equilibrium is where the taxi drivers takes the shortest distance routes. This assumption will be discussed in detail later.
4. **TAXI MARKET MECHANISM**

taxi trip, where the drivers make trip longer when they drive back from the origin to the destination, and the opposite pattern is observed for the trips in negotiated fare zones. The drivers are thus more likely to make trip shorter when their trip origin has less taxi demand than its destination.

This paper explores computation and market structure aspect of taxi market with particular focus on large scale data analysis. My finding indicates that there are significant structural patterns in the market and that these pattern seemed to be driven by market force. This work provides interesting and new avenues of research in the area of shared transport and services. Further development of this work can be extended to many research topics.
4.7 Appendix

4.7.1 Mathematical Proofs

Lemma 4.3.2.1. Suppose to the contrary that there does not exist a unique \( \hat{\theta}_i \) such that

\[
u(\hat{\theta}) = \int t(\theta)dF.
\]

By the incentive compatibility (IC) and the individual rationality (IR) condition, the utility function \( u(x(\cdot, \theta_{-i}), \cdot) \) is bounded, so it is \( u : \Theta \mapsto [u, u(\theta_i)] \). Without loss of generality, the utility function \( u \) is assumed to be continuous on \( \mathbb{R}_+ \), so the codomain of \( u \) is a subset of \( \mathbb{R}_+ \). Let \( B(X) \) be a space of the bounded function \( u \) and \( X \) be the codomain set \( [u, u(\theta_i)] \subset \mathbb{R}_+ \). Since the set \( X \) is a subset of positive real numbers, \( B(X) \) is a complete metric space with the metric \( \rho(x, y) = |x - y| \).

Define an operator \( T : B(X) \mapsto B(X) \) such that

\[
Tu = \int t(\theta_i)dF.
\]

The operator \( T : B(X) \mapsto B(X) \) satisfies the following conditions:

(a) (monotonicity) For all \( \tilde{u}, u \in B(X) \) with \( \tilde{u} \geq u \),

\[
T\tilde{u} = Tu = \int t(\theta_i)dF.
\]

(b) (discounting) For any \( \alpha \geq 0 \) and for some \( \beta \in (0, 1) \),

\[
[T(u + \alpha)](\theta) = \int t(\theta_i)dF = \leq (Tu)(\theta) + \beta \cdot \alpha,
\]

where \( (u + \alpha)(\theta) = u(\theta) + \alpha \).
Hence the operator $T$ satisfies the Blackwell’s sufficient conditions for a contraction, so it is a contraction mapping.

By contraction mapping theorem, there exists a unique fixed point $u^*$ in $B(X)$ that is $u^* = \int t(\theta_i)dF$ so a unique $\hat{\theta}_i$ such that $u(\hat{\theta}_i) = \int t(\theta_i)dF$ has to exist. Therefore, the contrary cannot hold.

**Theorem 4.3.2.1.** Suppose to the contrary that a taxi market allocation $x(\theta) = (y(\theta), t(\theta))$ satisfies the Bayesian IC with $\hat{\theta}_i > \theta_i$ such that $u(\hat{\theta}_i) = \int t(\theta_i)dF$, for all $i$. By the property of $t(\cdot)$ that is increasing in its argument, $t(\hat{\theta}_i) \geq t(\theta_i)$, for all $\hat{\theta}_i > \theta_i$. This implies $u(y(\theta_i), t(\hat{\theta}_i), \theta) > u(y(\theta_i), t(\theta_i), \theta)$, and hence

$$E_{\theta}[u(y(\theta_i), t(\hat{\theta}_i), \theta_i, \theta_{-i})] > E_{\theta}[u(y(\theta_i), t(\theta_i), \theta_i, \theta_{-i})].$$

This is contradiction of the Bayesian IC.

**Proposition 4.3.2.1.** Let $t^*(\cdot)$ be a solution of the principal’s utility maximization problem for both trip with the spatial distribution $f(\cdot)$ and $g(\cdot)$ as:

$$\max_{t(\cdot)} E_{\theta}[u_0(y(\theta), t(\theta), \theta)],$$

s.t. $u(y(\theta_i), t(\theta_i), \theta_i, \theta_{-i}) \geq u(y(\theta_i), t(\hat{\theta}_i), \theta_i, \theta_{-i}),$

$$u(y(\theta_i), t(\theta_i), \theta_i, \theta_{-i}) \geq u.$$

Assume that $t^*(\cdot)$ is a unique maximum for both trip with $f(\cdot)$ and $g(\cdot)$. Then it is Pareto (ex-ante) superior than any other $t(\cdot)$, so it is an ex-ante efficient allocation.

By the definition of well-defined market, $F(\cdot)$, cumulative distribution function of the probability distribution function $f(\cdot)$, first-order stochastically dominates $G(\cdot)$, cumulative distribution function of $g(\cdot)$. By the definition of stochastic dominance, the distributions $F(\cdot), G(\cdot)$ yield the following inequality:

$$\int u_0(y(\theta), t(\theta), \theta)dF(\theta) > \int u_0(y(\theta), t(\theta), \theta)dG(\theta).$$

Since the operators $\int dF$ and $\int dG$ are contraction mapping, which is $T : B(X) \mapsto$
$B(X)$ in the proof of lemma 2.3.1, the left and right hand sides have unique fixed point individually. Each constrained maximization problems have thus a unique fixed point, as their solutions, and therefore $t^*(\cdot)$ cannot be a solution for both trips with $f(\cdot)$ and $g(\cdot)$.

\[\square\]

4.7.2 Dummy variable
Table 4.6: A matrix of dummy variables by category

<table>
<thead>
<tr>
<th>Sample</th>
<th>Trips</th>
<th>Month Dummy</th>
<th>Day Dummy</th>
<th>Hour Dummy</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>Entire Sample</td>
<td>November</td>
<td>Monday</td>
<td>5:00 AM</td>
</tr>
<tr>
<td>NYC</td>
<td>Within NYC</td>
<td>November</td>
<td>Monday</td>
<td>5:00 AM</td>
</tr>
<tr>
<td>NYC</td>
<td>From NYC</td>
<td>November</td>
<td>Saturday</td>
<td>4:00 AM</td>
</tr>
<tr>
<td>NYC</td>
<td>To NYC</td>
<td>October</td>
<td>Sunday</td>
<td>5:00 AM</td>
</tr>
<tr>
<td>Manhattan</td>
<td>Within Manhattan</td>
<td>October</td>
<td>Sunday</td>
<td>3:00 AM</td>
</tr>
<tr>
<td>Manhattan</td>
<td>From Manhattan</td>
<td>November</td>
<td>Monday</td>
<td>3:00 AM</td>
</tr>
<tr>
<td>Manhattan</td>
<td>To Manhattan</td>
<td>November</td>
<td>Thursday</td>
<td>4:00 AM</td>
</tr>
</tbody>
</table>


*Developing Statistical Software in Fortran 95* (n.d.).


Pacheco, P. S. (1997), *Parallel programming with MPI*, Morgan Kaufmann.


