

ABSTRACT

Title of dissertation: INTEGRATED DISCRETE-CONTINUOUS CHOICE MODELS: THEORY AND APPLICATIONS IN ESTIMATING HOUSEHOLD VEHICLE OWNERSHIP, TYPE AND USAGE

Yangwen Liu, Doctor of Philosophy, 2013

Dissertation directed by: Cinzia Cirillo, Associate Professor
Department of Civil and Environmental Engineering

In the United States, transportation uses approximately 26% of the nation energy consumption, and contributes 27% to the total greenhouse gas emissions. Given the important role of private transportation, it is urgent to develop effective and innovative quantitative methodologies to support public authority decision making. Although a substantial body of the literature investigates household vehicle ownership decisions - vehicle holding, type and usage; the majority of the existing studies focuses on only one of these three decisions, is often limited to specific geographic areas and is not calibrated with the most recent travel survey data available.

This dissertation proposes a modeling framework that is able to incorporate all the three decisions simultaneously, and takes into account the correlation across the discrete variable (vehicle holding) and the continuous variable (miles traveled). In this integrated discrete-continuous choice model, a multinomial probit model is used to estimate household vehicle holding decision, while a multinomial logit model is adopted to estimate the vehicle type decision. The vehicle usage decision variable

is integrated with the discrete variables by adopting an unrestricted correlation pattern between the discrete and the continuous variables. Since the problem has no closed-mathematical form, I use estimation techniques based on Monte-Carlo simulations and numerical computation of multivariate normal probabilities to derive the solutions.

Though a number of studies have demonstrated that unordered behavioral models outperform the ordered mechanisms for vehicle holding decisions, those comparative studies were only conducted for the discrete decisions concerning vehicle ownership. Therefore, an ordered discrete-continuous model structure is developed, in which an ordered probit replace the multinomial probit for the vehicle holding decisions. Both the unordered and ordered structures are estimated and validated on the 2009 National Household Travel Survey data. Ordered models are in general preferred to unordered models for the lower computational costs to derive the analytical solutions. However, results from operational data show that the unordered discrete-continuous models always outperform the ordered ones in terms of both statistical goodness of fit and predication capabilities.

The proposed modeling framework is then applied to the entire nation and a system of national vehicle ownership models is derived. The models are calibrated using the 2009 National Household Travel Survey data, each combining four regions (Northeast, Midwest, South and West) and three area types (urban, suburban and rural). In addition, the models are applied to the 2009 American Community Survey data for six randomly selected counties/areas. The prediction results for the six counties/areas demonstrate the prediction capability of the models calibrated. The

national models are valuable both for the national level (to evaluate federal policies) and small areas (that lack local household travel survey data). The results also demonstrate that the integrated discrete-continuous framework has good prediction capabilities in modeling household vehicle ownership decisions.

Lastly, the dissertation estimates a discrete-continuous model for the Washington D.C. Metropolitan Area and analyzes the impact of improved bus and metro services on household ownership and use decisions in that area. The 2009 National Household Travel Survey data and the General Transit Feed Specification data are integrated, and then both spatial and temporal measurements of transit services are created on the Census Tract level. The results show that improved transit is a significant factor in household vehicle ownership choices and that the proposed methods are able to effectively predict changes in vehicle ownership and usage with respect to the transit improvements.

In conclusion, the dissertation contributes to both the theoretical analysis and the practical applications of the household vehicle ownership problem. The results provide decision makers with advanced quantitative methods that are able to effectively analyze policies, aiming at promoting greener travel behavior and at mitigating energy consumption and emissions.

INTEGRATED DISCRETE CONTINUOUS CHOICE MODELS:
THEORY AND APPLICATIONS IN ESTIMATING HOUSEHOLD
VEHICLE OWNERSHIP, TYPE AND USAGE

by

Yangwen Liu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Dr. Cinzia Cirillo, Chair
Dr. Roberto Celi
Dr. David Chien
Dr. Hiroyuki Iseki
Dr. Don Pickrell
Dr. Paul Schonfeld

© Copyright by
Yangwen Liu
2013

Acknowledgments

I would like to express my deepest appreciation and gratitude to my advisor, Dr. Cinzia Cirillo, for her patience and guidance during my entire graduate study. I deeply appreciate all her contributions of time, ideas, and funding to make my Ph.D. study complete. Her joy and enthusiasm on research motivates me, especially during the toughest times in my Ph.D. pursuit. I am also thankful for the excellent example she has erected for me as a successful female engineer and professor.

My sincere thanks extend to all my academic committee members, Dr. Roberto Celi, Dr. David Chien, Dr. Hiroyuki Iseki, Dr. Don Pickrell and Dr. Paul Schonfeld, for their insightful comments on my thesis. Without their wisdom, advice, and insight, I could not make this far. Without the help from Dr. David Chien and Dr. Hiroyuki Iseki, I could not obtain the important research data, which are the source data to verify my models. Throughout the research work, Dr. David Chien, Dr. Hiroyuki Iseki and Dr. Don Pickrell offered numerous thought-provoking discussions, which definitely improved my dissertation in many aspects. I am also grateful to Dr. Paul Schonfeld for teaching and guiding me through his classes, as well as being a member in my qualifying exam and final defense.

I am thankful to my fellow officemates: Jean-Michel Tremblay, Michael Maness, Pratt Hetrakul, Renting Xu, Nayel Urea Serulle and Lynna Nguyen. I appreciate their time to discuss with me on my thesis, and I learned many things while studying and working together with them. I also appreciate the supports from my friends in Maryland, including but not limited to Yijing Lu, Xiang He, Mingyang Ji, Yuening

Hu, Xiaojie Cong, Zheng Zhang and Bo Sun, who are always there to cheer me up and stand by me throughout my up and down times.

Last but not the least, I would like to thank my parents, for giving birth to me and endlessly supporting me throughout my entire life. I also thank my husband for tolerating to live with an impatient lady, yet remain a gentleman with his unconditional love.

Table of Contents

| | |
|------------------------------------------------------------------------------------|-----|
| List of Tables | vi |
| List of Figures | vii |
| 1 Introduction | 1 |
| 1.1 Background and Motivation | 1 |
| 1.2 Current Research Status | 3 |
| 1.3 Research Objectives | 6 |
| 2 Literature Review of Vehicle Ownership Models | 10 |
| 2.1 Review of Vehicle Holding Models | 15 |
| 2.2 Review of Vehicle Type Models | 21 |
| 2.3 Review of Discrete-Continuous Models | 25 |
| 2.3.1 Model Derived from Conditional Indirect Utility Function | 26 |
| 2.3.2 Multiple Discrete-Continuous Extreme Value model | 29 |
| 2.3.3 Bayesian Multivariate Ordered Probit and Tobit model | 34 |
| 3 Methodology | 37 |
| 3.1 Unordered Discrete-Continuous Model | 37 |
| 3.1.1 The Discrete Choice Sub-model | 37 |
| 3.1.2 The Continuous Choice Sub-model | 40 |
| 3.1.3 The Integrated Discrete-Continuous Model | 41 |
| 3.1.4 Estimation with Simulation - 1st Attempt | 42 |
| 3.1.5 Estimation with Simulation - Modified Approach | 47 |
| 3.1.6 Estimation with Numerical Computation | 49 |
| 3.1.7 Normalization of Covariance Matrix | 52 |
| 3.2 Ordered Discrete-Continuous Model | 57 |
| 3.2.1 The Discrete and Continuous Sub-models and the Integrated Model | 57 |
| 3.2.2 Estimation with Numerical Computation | 58 |
| 3.3 Endogeneity | 60 |
| 3.4 Goodness of Fit Measures | 63 |
| 4 Data Sources | 65 |
| 4.1 National Household Travel Survey (NHTS) | 65 |
| 4.2 Vehicle Characteristics | 67 |
| 4.3 U.S. Census TIGER/Line shapefiles | 69 |
| 4.4 General Transit Feed Specification (GTFS) | 69 |
| 4.5 American Community Survey (ACS) | 71 |

| | | |
|-------|------------------------------------------------------------------|-----|
| 5 | Comparison of Unordered and Ordered Discrete-Continuous Models | 74 |
| 5.1 | Introduction | 74 |
| 5.2 | Data Statistics | 75 |
| 5.3 | Calibration of the Logsum | 80 |
| 5.4 | Estimation Results and Comparison | 84 |
| 5.5 | Application Results and Comparison | 88 |
| 5.6 | Chapter Summary | 92 |
| 6 | National Model of Vehicle Ownership and Usage | 95 |
| 6.1 | Introduction | 95 |
| 6.2 | Estimation Results with the NHTS Data | 97 |
| 6.3 | Application with ACS Data for Local Counties/Areas | 103 |
| 6.3.1 | County/Area Descriptions | 103 |
| 6.3.2 | Application Results | 114 |
| 6.4 | Chapter Summary | 114 |
| 7 | Measuring Transit Service Impacts on Vehicle Ownership and Usage | 123 |
| 7.1 | Introduction | 123 |
| 7.2 | Data Geo-Processing and Data Integration | 126 |
| 7.2.1 | Spatial Measurements of Transit Service | 126 |
| 7.2.2 | Temporal measurements of bus service | 127 |
| 7.2.3 | Transit service index (TSI) | 128 |
| 7.2.4 | Data integration and final database | 129 |
| 7.3 | Estimation Results | 129 |
| 7.4 | Policy Analysis | 134 |
| 7.5 | Chapter Summary | 136 |
| 8 | Conclusions and Future Research | 138 |
| 8.1 | Summary and Research Contributions | 138 |
| 8.2 | Future Research | 141 |
| | Bibliography | 144 |

List of Tables

| | | |
|-----|-------------------------------------------------------------------------|-----|
| 2.1 | Summary of vehicle ownership models | 12 |
| 2.2 | Summary of vehicle holding models | 19 |
| 2.3 | Summary of vehicle type models | 22 |
| 2.4 | Vehicle classification schemes | 24 |
| 2.5 | Selected empirical studies on purchase behavior (Chintagunta, 1993) . | 26 |
| 3.1 | Illustration example | 43 |
| 5.1 | Data Statistics | 77 |
| 5.2 | Estimation results of the vehicle type choice sub-model | 81 |
| 5.3 | Estimation results of the integrated discrete-continuous models | 86 |
| 5.4 | Application results from the unordered discrete continuous model . . . | 90 |
| 5.5 | Application results from the ordered discrete continuous model | 91 |
| 6.1 | Estimation Results of National Models | 100 |
| 7.1 | Sample calibration of TSI | 128 |
| 7.2 | Estimation results | 132 |
| 7.3 | Policy analysis based on different improvement of the transit service . | 136 |

List of Figures

| | | |
|------|----------------------------------------------------------------------------|-----|
| 3.1 | Distribution of conditional residuals | 46 |
| 4.1 | Data Structure of GTFS Data | 71 |
| 5.1 | Distribution of vehicle classes | 79 |
| 5.2 | Vehicle age profile | 79 |
| 5.3 | Application results from the unordered discrete continuous model | 93 |
| 5.4 | Application results from the ordered discrete continuous model | 94 |
| 6.1 | United States Regions (Census Bureau) | 97 |
| 6.2 | United States Urban Area (Census Bureau) | 98 |
| 6.3 | Maps of San Diego County, CA | 105 |
| 6.4 | Map of Queens, NY | 106 |
| 6.5 | Maps of Nassau County, NY | 108 |
| 6.6 | Maps of PUMA Area 1900, TX | 110 |
| 6.7 | Maps of Fairfax County, VA | 111 |
| 6.8 | Map of Henrico County, VA | 112 |
| 6.9 | Data Statistics from American Community Survey | 113 |
| 6.10 | Application results of San Diego County, CA | 115 |
| 6.11 | Application results of Queens, NY | 116 |
| 6.12 | Application results of Nassau County, NY | 117 |
| 6.13 | Application results of PUMA area 1900, TX | 118 |
| 6.14 | Application results of Fairfax County, VA | 119 |
| 6.15 | Application results of Henrico County, VA | 120 |
| 6.16 | Summary of applications for the six counties/areas | 121 |
| 7.1 | Data structure of the final database | 130 |

Chapter 1

Introduction

1.1 Background and Motivation

Increasing mobility demand, especially in urban areas has resulted in growing levels of motorization, congestion and pollution. Modern societies are still highly dependent on private vehicles to satisfy demand for activities; while fastest growing economies in the world are experiencing a rapid increase in motor vehicle ownership. It is clear that vehicle demand has to be optimally managed and regulated in order to reduce the adverse impacts of transportation.

In this context, the role of analysts and researchers is to expand the basic knowledge of the problem, develop better analytical tools and support decision makers in their strategic choices. Ultimately, the cost of information gathering and modeling development is covered by cost savings resulting from better decision making.

The importance of modeling household vehicle fleet choices has been recognized for several decades now, though the urgency in terms of GHG emission and fossil fuel energy dependence is definitively more recent [Vyas et al., 2012].

Car ownership models play an important role in transportation and land use planning and are a critical component of Transportation Modeling Systems. In the classical four-step forecasting model, the trip generation module uses the outputs

from car ownership models (i.e., [Golob and Vanwissen, 1989]; [Kitamura, 2009]) as its inputs. Furthermore, vehicle ownership greatly impacts mode choice (i.e., [Dis-sanayake and Morikawa, 2010]), frequency of trips (i.e., [Kitamura, 2009]; [Shay and Khattak, 2012]), destination choice, trip timing, activity duration and trip chaining properties (i.e., [Hatzopoulou et al., 2001]; [Roorda et al., 2009]; [Paleti et al., 2013]).

Models for car ownership are of interests to both public agencies and private organizations: a) The US Department of Energy, b) State Departments of Transportation, c) auto industry, d) Local transit Agencies and e) World Bank (Train, 1979).

A number of agencies have implemented vehicle ownership into their regional transportation models. The State of California has developed the Motor Vehicle Stock, Travel and Fuel Forecast (MVSTAFF) model that uses a macroeconomic approach to modeling statewide motor vehicle holdings, vehicle miles travelled (VMT) and total fuel consumption. Other model systems that include a car ownership component are: the Maryland Statewide Transportation Model, the Coordinated Travel-Regional Activity Based Modeling Platform (CT-RAMP) for the Atlanta Region from Atlanta Regional Commission (2009), and the activity based model from the Puget Sound Regional Council (2008), etc.

National governments use car ownership models to forecast tax revenues and the regulatory impact of changes in the level of taxation (i.e., [Hayashi et al., 2001]; [Brnnlund and Nordstrm, 2004]; [Giblin and McNabola, 2009]). This type of model systems examines the changes in the car market configuration, the life cycle CO_2 emission from automobile transport and the tax revenues due to different taxation

policies [Hayashi et al., 2001]. It specifically determines the effect of the varying weights of the tax components in the stages of: (a) car purchasing, (b) car owning, and (c) car using to the changes in the car class and age mix and the car users' driving pattern and behavior towards car class choice and decommissioning.

Vehicle ownership models are also used by policy makers to identify factors that affect VMT, and therefore address the problems related to traffic congestion, gas consumption and air pollution (i.e., [Dargay and Gately, 1997]; [Schipper, 2011]). Models for car ownership growth in developing countries are important for estimating the implications on energy demand and price and on the global CO_2 emissions [Dargay and Gately, 1997].

1.2 Current Research Status

There is a substantial body of literature that has investigated household vehicle ownership decisions of vehicle holding, type and usage. Unfortunately, the majority of these studies analyzed the three decisions separately, due to the fact that vehicle holding and type are discrete decisions while vehicle usage is continuous, and therefore it is hard to integrate them in one framework.

Most of the early studies focused on vehicle holding and type choices. Ordered discrete choice model (such as ordered logit and ordered probit) and unordered discrete choice model (such as multinomial logit, multinomial probit and nested logit) are the two major modeling structures that have been used for modeling these two choices. The ordered models assume that a vehicle ownership decision

is a latent variable, whereas the unordered model is based on the random utility maximization theory which assumes the households make decisions that provide the highest utility.

As policy makers have started to pay more attention to problems associated with car usage, a growing number of researchers in transportation are trying to unify both discrete decisions (how many cars and their type) and continuous decisions (amount of use) into one integrated modeling framework. Discrete-continuous models, which were firstly developed in economics, are capable of dealing with problems where both discrete and continuous choices are involved. To the best of my knowledge, three methods are available in the transportation literature to model simultaneously vehicle ownership and usage.

The first category of these models is derived from the conditional indirect utility function in the microeconomic theory. Roy's identity property is applied to estimate vehicle usage and the relationship between the discrete and continuous choices. The method is consistent with utility maximization theory. It has a elegant formulation and it is simple to implement. However, the interdependence between the discrete and continuous parts is only captured by means of observed variables and no correlations are accounted for the unobserved factors.

In 2005, Bhat [Bhat, 2005] developed a Multiple Discrete-Continuous Extreme Value (MDCEV) model which jointly estimates the holding of multiple vehicle types and miles for each vehicle type. The dependent variable in this model is the mileage for each vehicle type category. Utility for each household is maximized subject to a total mileage budget. Under the assumption that the error term is iid extreme

value distributed, the probability function simplifies to a closed form, and collapses to Multinomial Logit (MNL) model for one-car household. The MDCEV model is consistent with random utility maximization theory, and is able to capture trade-offs among the usage of different types of vehicles. However, this model requires finer classification of vehicles as one type of vehicle cannot be chosen twice by the household. This type of model is limited by the assumption of fixed total mileage budget for every household; this implies that it is not possible to predict changes in the total number of miles in response to policy changes. Moreover, there is only a single error term (represents the unobserved factors) that underlies both discrete and continuous choices.

The third method is the Bayesian Multivariate Ordered Probit and Tobit (BMOPT) model developed by Fang [Fang, 2008]. The BMOPT model is composed of a multivariate ordered probit model for the discrete choices and a multivariate Tobit model for the continuous choice. In the BMOPT model, household decisions on the number of vehicles in one of the two categories (cars and trucks) considered are estimated by means of ordered probit model. The multivariate Tobit model is applied to estimate the household decisions on miles driven with each vehicle type. The joint model is formulated with an unrestricted covariance matrix for the discrete and continuous parts. This method is easier to implement than the RUM based models, and can be applied to study policy implications. However, it cannot handle a large number of vehicle categories and the ordered mechanism may not perform as well as unordered mechanism in modeling car ownership models.

Two papers ([Bhat and Pulugurta, 1998] and [Potoglou and Kanaroglou,

2008]) have investigated the empirical performance of ordered and unordered mechanisms in modeling vehicle ownership. They provided strong evidence that the appropriate mechanism is the unordered response mechanism for the vehicle holding choice. However, no studies in the literature have compared the performance of discrete-continuous models under ordered and unordered mechanisms. There is little evidence to demonstrate the superiority of one model to the other.

1.3 Research Objectives

The dissertation develops a comprehensive modeling framework for both discrete and continuous decision variables in the context of household vehicle ownership; three main choices are considered: the number of vehicles, their type and vintage, the annual mileage traveled. The model system accounts for a large number of vehicle classes and vintages overcoming the limitations of previous models. Moreover, a flexible structure of the unobserved factors between the discrete and continuous parts offers an integrated and elegant form for household decisions that are naturally linked. In particular, the joint model allows the estimation of a full variance-covariance matrix that captures both correlation amongst the alternatives in the discrete models *and* between the number of cars owned and the correspondent mileage in the continuous equation.

The research also compares the ordered discrete-continuous structure and RUM-based unordered discrete-continuous structure in the context of joint models for vehicle holding and vehicle usage decisions. The ordered discrete-continuous

structure has a similar structure except that an ordered probit is used for the vehicle holding sub-model. This comparative analysis is motivated by the fact that ordered discrete-continuous models are relatively easier to estimate when compared to un-ordered model structures; however, the assumption that vehicle ownership decisions are measured by a single latent variable might affect the goodness of fit of the model and its practical performance. The analysis is performed on data extracted from the 2009 National Household Travel Survey (NHTS).

The study applies and validates the proposed modeling framework to both local and national geographical levels. More specifically, the model is tested and applied to a U.S. metropolitan area (Washington D.C. area) and to the entire nation. The Washington metropolitan area is one of the largest metropolitan areas in the U.S., has a diverse population and has recently adopted several Smart Growth planning strategies. The modeling framework is extended to the four Census Regions (Northeast, Midwest, South and West and three area types (urban, suburban and rural) and applied to calculate rates of vehicle ownership and mileage traveled. Several data are merged and used for model estimation and application, including the National Household Travel Survey, the General Transit Feed Specifications data and the American Community Survey.

The rest of the dissertation is organized as follows. Chapter 2 reviews the literature on models for households vehicle ownership choices. The review outlines significant factors that influence vehicle holding, type and use decisions. In particular, prior discrete-continuous frameworks that have been applied to vehicle ownership modeling are reviewed in this chapter.

Chapter 3 describes the methodological framework of the integrated discrete continuous model developed in this study. This chapter firstly presents the discrete choice sub-model, the continuous choice sub-model and the joint formulation. Then, it explains the two estimation approaches developed for model estimation; the first is based on simulation methods, while the second adopts numerical computation to approximate choice probabilities. The second section of this chapter also presents the ordered discrete continuous model which has a similar model structure except for the adoption of an ordered mechanism for the discrete choice sub-model. Several issues related to the normalization of the covariance matrix and endogeneity are treated in this chapter.

Chapter 4 describes the datasets that have been collected and used for model calibrations and applications, including the National Household Travel Survey data, data on vehicle characteristics, US Census TIGER data, the General Transit Feed Specifications data and the American Community Survey data.

Chapter 5 compares the unordered and ordered discrete-continuous models for the Washington D.C. Metropolitan area. A number of variables including household sociodemographic information, residential density and fuel cost are introduced in the model formulation and their relative coefficients estimated. Both estimation and application results are presented and general findings from the model comparisons outlined.

Chapter 6 presents a system of models representative of the entire United States; application results for randomly selected counties/areas are given and discussed. The national models are valuable both for the national planning level and

for small areas, especially those lacking local household travel survey data. The results further validate the proposed integrated discrete-continuous framework for modeling household vehicle ownership decisions.

Chapter 7 estimates household joint decisions on vehicle ownership and usage with transit service indicators for the Washington D.C. area. The analysis develops a method to integrate the household travel survey with geographic data, and generates spatial and temporal measurements of transit service for the model estimation. The results provide evidences on the impacts of improved bus and metro services on household ownership and use decisions in the Washington D.C. Metropolitan area.

Chapter 8 concludes with a summary of the major findings and research contributions. Future research directions for both methodological and applied aspects of the problem treated in this dissertation are identified.

Chapter 2

Literature Review of Vehicle Ownership Models

Models for predicting changes in the level of car ownership have been under development since the 30s (e.g. [Wolff, 1938]; [Rudd, 1951]; [Tanner, 1958] as they are essential to the transport planning process and are of interest to governments, vehicle manufactures, environmental protection groups, public transport authorities, and public transport operators.

Aggregate time series models have been widely used in very early modeling attempts. A sigmoid-shape function is usually used to explain the development of car ownership over time, and a growth function is related to income or gross domestic product (GDP). The function increases slowly in the beginning (at low GDP per capita), then rises steeply, and ends up approaching a saturation level. Examples along this line are the work done by Tanner (e.g. [Tanner, 1983]), [Button et al., 1993], [Ingram and Liu, 1999], the National Road Traffic Forecasts (NRTF) in the UK ([Whelan et al., 2000], [Whelan, 2001]), [Dargay and Gately, 1999], etc. These models have the lowest data requirements and can be used to predict the total number of cars in future years, which in turn is a potential starting point for more detailed analysis.

More recently, disaggregate car ownership models based on discrete choice models became the dominant method to deal with the number of cars owned by a

household. Examples are the work by [Gunn et al., 1978], which have been later implemented into the Dutch national model system (LMS) [HCG, 1989]. Similar models have been developed by [Bhat and Pulugurta, 1998] and by [Rich and Nielsen, 2001]; real applications of static discrete methods include the model developed for the city of Sydney ([HCG, 2000]), and the model for the National Roads Traffic Forecast (NRTF) in the UK ([Whelan, 2001]).

Disaggregate model systems are also used to explain households' choice of car type given car ownership. There are many publications on vehicle type choice models, such as [Berkovec, 1985], [Chandrasekharan et al., 1991], [Hensher et al., 1992], [Mannering and Winston, 1985], [Manski and Sherman, 1980] and [Train, 1986].

Table 2.1 provides a summary of modeling approaches for vehicle ownership existing in the literature.

Table 2.1: Summary of vehicle ownership models

| Reference | Data Source (Year) | Sample Size | Choices Examined | Model |
|----------------------------------|----------------------------------|-----------------------------------------------|-----------------------------------------------------------------------------------------------------|---------------------------------|
| [Lave and Train, 1979] | Seven US cities (1976) | 541 new car buyers | Vehicle type choice | MNL |
| [Manski and Sherman, 1980] | US (1976) | 1200 single-vehicle or two-vehicle households | Vehicle type choices in households holding one vehicle and two vehicles | MNL |
| [Beggs, 1980] | Baltimore (1977) | 326 households | Vehicle type choice | MNL |
| [Hensher et al., 1981] | Sydney (1980) | 151 households | Fleet-size choice, vehicle type choice | Nested Logit |
| [Hoeherman et al., 1983] | Haifa urban area, Israel, (1979) | 800 households | Transaction, Vehicle type | Nested Logit |
| [McCarthy, 1985] | San Francisco (1973-1975) | 269 households | choice between no transaction, replacing one auto, adding one auto, reducing one auto. | MNL |
| [Manning and Winston, 1985] | US (1978-1980) | 3842 households | quantity choice, type choice, utilization model | Nested Logit and OLS regression |
| [Berkovec and Rust, 1985] | US (1978) | 237 single-vehicle households | Vehicle type choice | Nested Logit |
| [Berkovec, 1985] | US (1978) | 1048 households | Vehicle quantity (0, 1, 2, 3), Vehicle type choice | Nested Logit |
| [Hensher and Le Plastrier, 1985] | Sydney (1980) | 400 households | Fleet-size choice (0, 1, 2, 3), vehicle type choice | Nested Logit |
| [Manning, 1986] | US (1978) | 272 households, 554 vehicles | vehicle usage | 3SLS-2 equations |
| [Train, 1986] | US (1978) | 1095 households | Vehicle quantity, class/vintage, usage | MNL and Regression |
| [McCarthy, 1989] | US (1985) | 726 households | choice of make/model for new vehicle purchases. Choice set is chosen plus 14 assigned alternatives. | MNL |

| | | | | |
|-----------------------------|-------------------------------------------|-------------------------------------|-----------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------|
| [Kitamura and Bunch, 1992] | Dutch National Mobility Panel Data set | Panel, 605 HH, (1984-1987) | vehicle quantity | Ordered Probit |
| [Golob, 1990] | Dutch (1985-1988) | 2119 households | choice between fleet size 0, 1, 2. | Ordered Probit |
| [Hensher et al., 1992] | Sydney (1981-1985) | 1444, 1295, 1251, 1197 | Static Vehicle choice and type-mix choice, Static vehicle use, Dynamic vehicle choice and use | Nested Logit and 3SLS regression |
| [Jong, 1996] | Dutch (Oct, 1992; Oct 1993) | Panel, 3241 respondents | Vehicle holding duration, Vehicle type choice, Annual kelometrage and fel efficiency | Hazard function, Nested logit, Regression |
| [Golob et al., 1997] | California (1993) | 4747 households | Vehicle use by type of vehicle | Structural equation model |
| [Bhat and Pulugurta, 1998] | US (1991, 1990, 1991), Dutch (1987) | 3665, 3500, 1822, 1807 | vehicle quantity (0, 1, 2, 3, 4) | MNL and Ordered logit |
| [Kitamura et al., 1999] | California, 1993 | Panel (First wave), 4747 households | 1 Vehicle holding model, and Num of Vehicle per household member and per driver, 2 Vehicle type choice, 3 Vehicle use | Ordered probit model, Tobit model; MNL; OLS regression |
| [Dargay and Gately, 1999] | UK, Family Expenditure Survey (1982-1993) | panel, cohort, 7200 hh | vehicle quantity | dynamic cohort (panel) |
| [Hanly et al., 2000] | UK (BHPS) (1993-1996) | Panel, about 4000-5000 households | vehicle quantity (0, 1, 2, 3+) | Ordered Probit model |
| [Mannering et al., 2002] | US (1995) | 654 households | vehicle acquisition type (cash, non-cash (lease, finance)); Vehicle type choice | Nested Logit model |
| [Choo and Mokhtarian, 2004] | San Francisco, 1998 | 1904 households | Vehicle type choice | MNL |

| | | | | |
|------------------------|-------------------------------------------|-------------------------|----------------------------------------------------|------------------------------------------------------------------|
| [Huang, 2005] | UK, Family Expenditure Survey (1957-2001) | Panel, 6,500 households | Number of cars owned or used by household (1+, 2+) | Dynamic Mixed Logit model with Saturation Level |
| [Bhat and Sen, 2006] | San Francisco (2000) | 3500 households | Vehicle type holding and usage | MDCEV (multiple discrete/continuous extreme value model) |
| [Whelan, 2007] | UK, (1971-1996) and NTS (1991) | unknown | vehicle quantity (0, 1, 2, 3+) | Hierarchical logit model with saturation level |
| [Cao et al., 2007] | Northern California, USA, 2003 | 1682 households | vehicle quantity (0, 1, 2, 3, 4, 5+) | Ordered probit and static-score model |
| [Fang, 2008] | NHTS (2001, CA) | 2299 households | Vehicle choice and usage (BMOPT and MDCEV) | BMOPT (Bayesian Multivariate Ordered Probit and Tobit) and MDCEV |
| [Bhat et al., 2009] | San Francisco (2000) | 15,000 households | Vehicle type/vintage and use, vehicle make/model | MDCEV-MNL |
| [Bhat and Eluru, 2009] | San Francisco (2000) | 15,000 households | residential neighborhood choice and daily VMT | copula-based approach |
| [Spissu et al., 2009] | San Francisco (2000) | 15,000 households | Vehicle type acquisition and VMT | copula-based approach |

2.1 Review of Vehicle Holding Models

In Table 2.2 I summarize a number of vehicle holding models in the literature, in particular I describe the data source, the sample size, model type and the dependent variables used for the analysis.

There are two types of discrete choice modeling structures that have been used in the household vehicle ownership studies: ordered-response mechanism and unordered-response mechanism. The ordered-response mechanism assumes that household vehicle ownership is represented as an ordinal variable and the choice is determined by a single latent variable which represents the propensity of the household vehicle ownership decisions. Examples of the application of ordered-response mechanism are [Kitamura, 1987], [Golob and Vanwissen, 1989], [Golob, 1990], [Kitamura and Bunch, 1992], [Bhat and Koppelman, 1993], [Kitamura et al., 1999], [Hanly et al., 2000], [Chu, 2002], [Kim and Kim, 2004a] and [Cao et al., 2007]. The unordered-response mechanism is based on the hypothesis that household vehicle ownership is represented as a nominal variable. It follows the random utility maximization (RUM) principle which assumes that the household makes the vehicle ownership decisions that provides the highest utility among all the possible choices. Examples of the studies with unordered-response mechanism are [Manering and Winston, 1985], [Train, 1986], [Bunch and Kitamura, 1990], [Hensher et al., 1992], [Purvis, 1994], [Ryan and Han, 1999], [Whelan, 2007], [Potoglou and Kanaroglou, 2008].

In the context of the comparison of the ordered and unordered mechanisms,

there are two papers that explicitly investigate the empirical performance of the two structures in modeling vehicle ownership decisions, (see [Bhat and Pulugurta, 1998] and [Potoglou and Kanaroglou, 2008]). [Bhat and Pulugurta, 1998] compared the multinomial logit (MNL) models (represents unordered-response mechanism) and the ordered logit (ORL) models (represents ordered-response mechanism) with four datasets from Boston, Bay area, Puget Sound area and the Netherlands. The two mechanisms were evaluated by comparing elasticity effects, measure of fit and predictive performance. The results showed that the MNL model is able to capture elasticity patterns across alternatives, while the ORL is more rigid in elasticity effects. Meanwhile, the MNL model outperforms the ORL model in several measures of fit. The conclusion from this study is that the appropriate choice mechanism is the unordered-response structure for vehicle ownership modeling. [Potoglou and Kanaroglou, 2008] evaluated the multinomial logit (MNL) model, ordered logit (ORL) model and ordered probit (ORP) model for car ownership by using data from Baltimore, Dutch and Japan. The MNL, ORL and ORP models are compared with a number of data fit measures and the results clearly demonstrate the superiority of the MNL to the ordered ORL and ORP.

Those studies provided strong evidence that the appropriate mechanism is the unordered response mechanism for the vehicle ownership models. It is important to stress that the ordered and unordered models have been compared for vehicle holding models only.

In terms of the attributes adopted in existing vehicle holding, they can be classified into four categories: (1) information on the household, (2) information on

the household head or primary driver, (3) land-use factors and (4) other unclassified information).

Significant explanatory variables of the household includes: household's income, household structure, number of household members (household size), number of workers, number of adults, number of children, number of drivers (licensing holding) in the household. In terms of household income, usually the annual income is used in the model. In some studies, the logarithmic transformation of the income or the discretionary income (the amount of income left to the household after subtracting taxes and normal expenses) enter the model specification.

The estimation results showed that most of the household socio-economic characteristics have positive influence on car ownership. The positive coefficient of the income variable indicates that, for instance, a household is more likely to own more vehicles, with a higher household income. Same trends can be found in other attributes, such as the number of household members, number of workers, number of adults, number of children, and number of drivers in the household. All of the coefficients have considerable t-statistics. Few studies analyzed household structure variables, usually using the number of adults and the number of children in the household.

Significant explanatory variables about the household head or primary driver include: age, gender, education level and work status. The estimation results in the previous researches indicate that a household is likely to own fewer vehicles with older household head or female household head. With higher education level of the household head, a household is more likely to own more vehicles. Only few studies

included household head's work status in the utility function.

In terms of land use information, previous researches mainly use population density, and location variables (urban, suburban, and rural). Estimation results indicate that households in the area with high density or in urban area own fewer vehicles. A few studies included the accessibility to transit as this variable is difficult to obtain in many real cases.

Other variables, which do not belong to any of the three categories above, include for example dummy variables describing parking availability and the presence of company cars. These variables were mainly used in European studies; where parking space is limited and where the number of company cars can be significant.

Table 2.2: Summary of vehicle holding models

| Reference | Data Source (Year) | Sample Size | Model type | Independent Variables |
|-----------------------------|-------------------------------------------|-----------------------------------------------|--------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [Hensher et al., 1981] | Sydney (1980) | 151 households | Nested Logit | Choice between acquiring one vehicle given initial holdings or not |
| [Manning and Winston, 1985] | US (1978-1980) | 3842 single-vehicle or two-vehicle households | NL (choice between 1 and 2 vehicles for each period and combined period) | hh members, worker, income, urban indicator, log sum of type choice models, choice indicator |
| [Kitamura, 1987] | Dutch National Mobility Panel Data set | Panel, 605 HH, (1984-1987) | Ordered Probit model | Num of workers, Num of adults, num of children, HH size, num of drivers, HH education |
| [Golob, 1990] | Dutch (1985-1988) | 2119 households | Ordered Probit | HH income, Num of persons >18, Num of persons 12-17, Num of persons <12, Num of drivers, Num of workers, residence location |
| [Bhat and Pulugurta, 1998] | US (1991, 1990, 1991), Dutch (1987) | 3665, 3500, 1822, 1807 | ORL v.s. MNL | Num of non-working adults, Num of working adults, Annual HH income, Urban residential location, Suburban residential location, Single-family residential housing |
| [Kitamura et al., 1999] | California, 1993 | Panel (First wave), 4747 households | Ordered probit model, Tobit model | HH size, Num of drivers, num of workers, num of adults, dummy of couple, dummy of single person, dummy of income, owns home, dummy of parking space; accessibility, density |
| [Dargay and Gately, 1999] | UK, Family Expenditure Survey (1982-1993) | panel, cohort, 7200 hh | dynamic cohort (panel) | income, adults, children, % metropolitan, % rural, generation, car purchase cost car running cost, public transport fares |

| Reference | Data Source (Year) | Sample Size | Model type | Independent Variables |
|----------------------|---------------------------------------------------------------------------------------------|-----------------------------------|---------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [Hanly et al., 2000] | UK, British Household Panel Survey (BHPS) (1993-1996) | Panel, about 4000-5000 households | Ordered Probit model | household income, number of adults, number of children, number of worker, dummy of pensioner, regional dummy, population density |
| [Huang, 2005] | UK, Family Expenditure Survey (1957-2001) | Panel, 6,500 households | Dynamic Mixed Logit model with Saturation Level (GUASS) | Log of household disposable income, household size, number of workers, log of age of household head, log of index of real motoring costs, proportion of households living in Metropolitan area, proportion of households living in rural are, dummy of young household |
| [Whelan, 2007] | UK, family expenditure survey (FES) (1971-1996) and the national travel survey (NTS) (1991) | unknown | The hierarchical logit model with saturation level | household income, household structure, motoring costs, need/accessibility, company cars, time trend/license holding |
| [Cao et al., 2007] | Northern California, USA, 2003 | 1682 households | Ordered probit and static-score model (Limdep 8.0) | Female, HH income, HH size, Num. of adults, Num. of workers, Driving disability, Transit disability, Residential tenure, Outdoor spaciousness, num of business types, accessibility, car dependent, safety of car |

2.2 Review of Vehicle Type Models

Table 2.3 presents several vehicle type models, in particular it describes the data source, the sample size, model type, vehicle classification and the dependent variables.

The majority of the studies use multinomial logit model in the vehicle type estimation. MNL is chosen in the most cases because we can take advantage of one of the logit properties. An important property of the logit model is the independence from irrelevant alternatives (IIA) property. That is, the ratios of probabilities are necessarily the same no matter what other alternatives are in the choice set or what the characteristics of other alternatives are. This IIA property has several advantages, and one of them is that it is possible to estimate model parameters consistently on a subset of alternatives for each sampled decision maker. This fact is important because it saves computer time by estimating on a subset of alternatives when the total number of alternatives is large. In the vehicle type models, the combination of vehicle type choices increase exponentially with the number of vehicles in the household, hence it is computational impossible to take account all the alternatives. With the IIA property, multinomial logit model allows the modelers to get consistent coefficients with the estimation on a subset of the alternatives.

Table 2.3: Summary of vehicle type models

| Reference | Data (Year) | Source | Sample Size | Model Type | Vehicle Classification | Independent Variables |
|-----------------------------|------------------------|----------------------|----------------------------------------|--------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|
| [Lave and Train, 1979] | Seven US cities (1976) | US cities (1976) | 541 new car buyers | MNL | subcompact, sports, subcompact (A and B), compact (A and B), Intermediate, Standard (A and B) Luxury | purchase price/income, weight*age, HH member, vehicle |
| [Manski and Sherman, 1980] | US (1976) | US (1976) | 1200 single- or two-vehicle households | MNL | Chosen alternative plus 25 alternative makes/models/vintage (randomly selected from 600 vehicle type) | purchase price, seats, vehicle weight and age, acceleration time, luggage space, scrappage rate, transaction-search cost, operation cost |
| [Beggs, 1980] | Baltimore (1977) | Baltimore (1977) | 326 households | MNL | 5 classes (subcompact, compact, mid-size, full-size, luxury), 4 vintage (1942-1971, 1972-1974, 1975-1976, 1977) | purchase price, operating cost, wheelbase, "depreciated luxury", age of vehicle, income, hh members, distance to parking |
| [Hoeherman et al., 1983] | Haifa area, (1979) | urban Israel, (1979) | 800 households | Nested model | Chosen alternatives plus 19 alternative makes/models/vintages (randomly selected from 950 vehicle types) | purchase price, operating cost, engine size, vehicle age, income, brand loyalty, same make cars, horsepower to weight |
| [Manning and Winston, 1985] | US (1978-1980) | US (1978-1980) | 3842 single- or two-vehicle households | NL | Chosen alternative plus nine alternative makes/models/vintages (randomly selected from 2000 vehicles) | purchase price/income, operating cost/income, lagged utilization of same vehicle or same make |
| [Berkovec and Rust, 1985] | US (1978) | US (1978) | 237 single-vehicle households | Nested model | upper level: vehicle age groups (new, mid, old), lower level: 5 vehicle classes (subcompact, compact, intermediate, standard, luxury/sports) | purchase price, operating cost, seats, vehicle age, turning radius in urban, horsepower to weight, transaction |
| [Berkovec, 1985] | US (1978) | US (1978) | 1048 households | Nested model | 131 alternatives=10 years (1969-1978) * 13 vehicle classes (domestic subcompact, compact, sporty, intermediate, standard, luxury, pickup truck, van and SUV; foreign subcompact, larger, sports, and luxury) + all models before 1969 | purchase price, seats, proportion of makes/models in class to total makes/models |

| Reference | Data Source (Year) | Sample Size | Model Type | Vehicle Classification | Independent Variables |
|----------------------------------|-----------------------------|-------------------------------------|-------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [Hensher and Le Plastrier, 1985] | Sydney (1980) | 400 households | Nested Logit | Holdings: Choice of make/model/vintage given fleet size. Single model for all levels. Choice set is chosen plus 2 reported alternatives. Transaction: choice of make/model/vintage given fleet size adjustment. Choice set is chosen plus 1 or 2 alternatives randomly selected. | Registration charge, service and repair expense, sales tax on purchase price, seats, fuel efficiency, weight, luggage space, age of vehicle, age, passenger, dummy (≤ 600 miles per month, dummy (use for paid work) |
| [Jong, 1996] | Dutch (Oct, 1992; Oct 1993) | Panel, 3241 respondents | Nested logit model (diesel and non-diesel cars) | 133 make/model combinations; about 1000 make/model/age-of-car combinations (better); ALOGIT; 20 alternatives (the chosen one plus 19 random) | Log of remaining household income; fixed cost/income; fuel cost/income; dummy for brand loyalty, engine size, diesel, age |
| [Kitamura et al., 1999] | California, 1993 | Panel (First wave), 4747 households | MNL model | Four-door sedans, two-door coupes, Vans, wagons, sports car, SUVs. | dummy (same vehicle type), Age, male, education, employed, commuter, commute distance, other (same as the vehicle holding models) |
| [Manning et al., 2002] | US (1995) | 654 households buying new vehicles | Nested Logit model | Chosen alternative plus 9 alternative makes and models (randomly selected from 175 vehicle types) | purchase price/income, passenger side airbag, horsepower, vehicle residual value, consecutive purchases |
| [Choo and Mokhtarian, 2004] | San Francisco, 1998 | 1904 households | MNL model (LIMDEP) | small, compact, mid-size, large, luxury, sports, minivan/van, pickup, SUV | objective mobility, subjective mobility, travel liking, attitudes, personality, lifestyle, demographics |

The vehicle type classification methods in the literature mainly consists of five different categories: (1) models that only consider very general classes of vehicles, such as small car, compact car, large car, sporty car, etc; (2) models that consider general classes and vintages of vehicles, such as small old car, large new car, etc; (3) models that contain the chosen alternative and a number of randomly selected alternatives from the total number of combinations of makes and models (i.e. Toyota, Camry); (4) models that contain the chosen alternative and a number of randomly selected alternatives from the total number of combination of make, model and vintage (i.e. 2003 Honda Civic); (5) models that consider vehicle classes and vintages, such as 2005 mid-size car, 2007 SUV, etc.

Table 2.4 reports vehicle classification schemes in terms of vehicle size, vehicle function, or both. Most schemes for vehicle classification first group vehicle by size, and then special categories such sports, pickup and SUV are added.

| Source | Vehicle Classification | Basis |
|-------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|
| NHTS (FHWA, 2009) | Automobile (including wagon), van, SUV, pickup, other truck, RV, motorcycle, other | Function |
| NTS (BTS, 2009) | Subcompact car, compact car, intermediate car, full car, light pickup, large pickup, small van, large van, small utility, large utility | Size and function |
| EPA (2009) | Cars: two-seater, sedan(minicompact, subcompact, compact, mid-sized, large), station wagon (small, midsize, larg); Trucks: pickup (small and standard), van (cargo and passenger), minivans, SUV, special purpose vehicle | Size and function |
| Comsumer Reports (2009) | Convertible, small car, sedan, wagon, SUV, minivan, pickup, sporty car | Size and function |

Three types of variables are mainly used in existing vehicle type models: (1) vehicle characteristics, (2) household characteristics, and (3) other unclassified characteristics. Purchasing price, operating cost, space and engine related variables are usually found to be significant in vehicle type models.

2.3 Review of Discrete-Continuous Models

Discrete-continuous models have been investigated in marketing studies since 1980's. Marketing researchers developed discrete-continuous models to determine household purchase decisions for frequently purchased packaged goods by the impact of marketing mix and demographic variables. Previous studies have predicted one or more of the purchasing decisions by proposing relationships between the observed choices of households and variables such as product price, price cuts, feature advertisements, special displays and observed and unobserved household characteristics [Chintagunta, 1993]. Chintagunta summarized a partial list of previous studies dealing with household purchase behavior along with their important features (Table 2.5). Studies in marketing mainly focus on three different household purchase decisions: (1) the timing of a purchase or the category purchase decision, (2) the brand choice decision and (3) the purchase quantity decision. In transportation, discrete-continuous models have also attracted researchers' attention and recently have been investigated in studying household decisions on vehicle ownership (discrete choice) and vehicle use (continuous choice).

Table 2.5: Selected empirical studies on purchase behavior (Chintagunta, 1993)

| Preference | Decisions Studied |
|-------------------------------|-----------------------------------------------------|
| [Guadagni and Little, 2008] | Brand choice |
| [Neslin et al., 1985] | Purchasing timing, Purchase quantity |
| [Krishnamurthi and Raj, 1988] | Brand Choice, Purchase Quantity |
| [Tellis, 1987] | Brand Choice, Purchase Quantity |
| [Jones and Landwehr, 1988] | Brand Choice |
| [Gupta, 1988] | Purchase Timing, Brand Choice, Purchase Quantity |
| [Gupta, 1991] | Purchase Timing |
| [Bucklin and Lattin, 1991] | Purchase Incidence, Brand Choice |
| [Chiang, 1991] | Purchase Incidence, Brand Choice |
| [Jain and Vilcassim, 1991] | Purchase Timing |
| [Kamakura and Russell, 1989] | Brand Choice |
| [Schmittlein et al., 1988] | Purchase Timing |

2.3.1 Model Derived from Conditional Indirect Utility Function

The earliest generation of models that have investigated vehicle ownership choices with discrete-continuous models were derived from conditional indirect utility function (e.g., [Manning and Winston, 1985]; [Train, 1986]; [Hensher et al., 1992]; [de Jong, 1989b], [de Jong, 1989a] and [de Jong, 1991]), which is based on micro-economic theory. Originally developed by [Dubin and McFadden, 1984], and [Hannemann, 1984], the basic concept is that the households choose the combination of vehicle ownership and vehicle usage that gives the highest utility. Roy's identity is applied to estimate vehicle usage and the relationship between the two modeling stages. Although based on single discreteness, this series of studies based on the indirect utility function are able to capture the interdependence between the vehicle holding and the corresponding mileage by means of observed variables. This elegant formulation is consistent with economic theory and simple to implement.

Some terminologies of direct utility, indirect utility and Roy's identity:

- Direct utility function gives the utility that the consumer obtains at given quantities of each good ($U(x_1, x_2)$).
- Indirect utility gives the utility that the consumer obtains at given prices and income once he has chosen the quantities that maximize his (direct) utility subject to the budget constraint for the given prices and income ($Y(p_1, p_2, y)$).
- It can be shown that a consumer's preferences can be equivalently represented by either a direct utility function or an indirect utility function.
- If the consumer is a utility maximizer, then he will purchase the quantities of the two goods that solves the constrained maximization problem:

$$\max U(x_1, x_2) \text{ or } Y(p_1, p_2, y)$$

such that

$$y = p_1x_1 + p_2x_2$$

- Roy's identity states that the demand for a good is equal to (the negative of) the derivative of the indirect utility function with respect to the good's price divided by the derivative of the indirect utility function with respect to income. That is:

$$x_1 = -\frac{\partial Y/\partial p_1}{\partial Y/\partial y}$$

$$x_2 = -\frac{\partial Y/\partial p_2}{\partial Y/\partial y}$$

- For deriving demand functions, it is much easier to work with a consumer's indirect utility function rather than with his direct utility function.

Suppose the number of alternatives is J , the observed characteristics of each alternative i is z_i , the quantity of the good is x , the person's income is y , other observed characteristics of the person is s , and all unobserved factors is w_i . The price of the good is denoted as p_i , which is the price per unit of x given that alternative i is chosen.

The maximum utility that the person can obtain given that he has chosen alternative i :

$$Y_i = Y_i(p_i, y, z_i, s, w_i)$$

This is a conditional indirect utility function for alternative i . Conditional indirect utility functions can be constructed for each alternative in the set J . Each of these gives the maximum utility that the person can obtain if he chooses a particular alternative.

The person will choose alternative i if and only if the conditional indirect utility is higher for alternative i than for any other alternative:

$$Y_i(p_i, y, z_i, s, w_i) > Y_j(p_j, y, z_j, s, w_j)$$

for all j in J , $j \neq i$.

Consequently, the probability of alternative i being chosen is

$$P_i = \text{Prob}(Y_i(p_i, y, z_i, s, w_i) > Y_j(p_j, y, z_j, s, w_j))$$

for all j in J , $j \neq i$.

The indirect utility can be decomposed into observed and unobserved parts:

$$Y_i(p_i, y, z_i, s, w_i) = V_i(p_i, y, z_i, s) + e_i$$

Where e_i is a function of unobserved variables w_i and V_i is simply the difference between e_i and Y_i . The form of choice model is derived by specifying a distribution of e_i . For example, if each e_i is assumed to be distributed independently, identically extreme value, then the choice probabilities are in logit form.

The demand for good x is determined from the conditional indirect utility function using Roy's identity. That is, the demand for x , given that alternative i is chosen, is

$$x_i = \frac{\partial Y_i(p_i, y, z_i, s, w_i) / \partial p}{\partial Y_i(p_i, y, z_i, s, w_i) / \partial y} = g_i(p_i, y, z_i, s, w_i)$$

Under certain forms of the conditional indirect utility function, the conditional demand for x_i and the observed utility V_i can be derived as linear functions of income, price and other explanatory variables (Train, 1986).

2.3.2 Multiple Discrete-Continuous Extreme Value model

Multiple discrete-continuous extreme value (MDCEV) models, developed by [Bhat, 2005] and further applied in [Bhat and Sen, 2006] and [Bhat et al., 2009] are utility-based econometric models that jointly estimate the holding of multiple vehicle types and the miles for each vehicle type. The dependent variable in this model is the mileage for each vehicle type category. Utility for each household is maximized subject to a total mileage budget. Under the assumption that the error term is iid extreme value distributed, the probability function simplifies to an elegant and compact closed form, and collapses to Multinomial Logit (MNL) model for one car household.

In MDCEV model, the utility accrued to a household is specified as the sum of the utilities obtained from using each type of vehicle.

$$U = \sum_{j=1}^K \psi(x_j)(m_j + \gamma_j)^{\alpha_j} \quad (2.1)$$

Where there are K different vehicle types that a household can potentially own. m_j is the annual mileage of use for vehicle type j ($j = 1, 2, \dots, K$). $\psi(x_j)$ is the baseline utility for vehicle type j , and γ_j α_j are parameters. Ψ is a function of observed characteristics, x_j , associated with vehicle type j .

Eq. 2.1 is a valid utility function if $\psi(x_j) > 0$ and $0 < \alpha_j \leq 1$ for all j . the term γ_j determines if corner solutions are allowed (i.e., a household does not own one or more vehicle types) or if only interior solutions are allowed (i.e., a household is constrained by formulation to own all vehicle types).

the utility form is also able to accommodate a wide variety of situations characterizing vehicle type preferences based on the values of $\psi(x_j)$ and α_j ($j = 1, 2, \dots, J$). A high value of $\psi(x_j)$ for one vehicle type (relative to all other vehicle types), combined with a value of α_j close to 1, implies a high baseline preference and a very low rate of satiation for vehicle type j . This represents the situation when a household primarily uses only one vehicle type for all its travel needs (i.e., a "homogeneity-seeking" household). On the other hand, about equal values of $\psi(x_j)$ and small values of α_j across the various vehicle types j represents the situation where the household uses multiple vehicle types to satisfy its travel needs (i.e., a "variety seeking" household). More generally, the utility form allows a variety of situations characterizing a household's underlying behavioral preferences for different

vehicle types.

a multiplicative random element is introduced to the baseline utility as follows:

$$\psi(x_j, \epsilon_j) = \psi(x_j) \cdot e^{\epsilon_j} \quad (2.2)$$

Where ϵ_j captures the unobserved characteristics that impact the baseline utility for vehicle type j . The exponential form for the introduction of random utility guarantees the positivity of the baseline utility as long as $\psi(x_j) > 0$. To ensure this latter condition, $\psi(x_j)$ is parameterized further as $\exp(\beta'x_j)$, which then leads to the following form for the baseline random utility:

$$\psi(x_j, \epsilon_j) = \exp(\beta'x_j + \epsilon_j) \quad (2.3)$$

The overall random utility function then takes the following form:

$$U = \sum_{j=1}^K [\exp(\beta'x_j + \epsilon_j)] (m_j + \gamma_j)^{\alpha_j} \quad (2.4)$$

The satiation parameter, α_j , in the above equation needs to be bounded between 0 and 1, as discussed earlier. To enforce this condition, α_j is parameterized as $1/[1 + \exp(-\delta_j)]$. Further, to allow the satiation parameters to vary across households, δ_j is specified as $\delta_j = \theta_j' y_j$, where y_j is a vector of household characteristics impacting satiation for the j th alternative, and θ_j is a corresponding vector of parameters.

Eq. 2.4 subject to the constraint that $\sum_{j=1}^K m_j = M$, Where M is the total household motorized annual mileage.

Assuming that the ϵ_j terms are independently and identically distributed across alternatives, and are distributed standard Gumbel, the probability that the

household owns I of the K vehicle types ($I \leq 1$) is

$$P(m_i^* > 0 \text{ and } m_s^* = 0; i = 1, 2, \dots, I \text{ and } s = I + 1, \dots, K)$$

$$= \left[\prod_{i=1}^I c_i \right] \left[\prod_{i=1}^I \frac{1}{c_i} \right] \left[\frac{\prod_{i=1}^I e^{V_i}}{(\sum_{j=1}^K e^{V_j})^I} \right] (I - 1)! \quad (2.5)$$

Where $c_i = \left(\frac{1 - \alpha_i}{m_i^* + \gamma_i} \right)$ and $V_j = \beta' x_j + \ln \alpha_j + (\alpha_j - 1) \ln(m_i^* + \gamma_i)$. In the case when $I = 1$ for a particular household (i.e., only one vehicle type is chosen by the household), the model collapses to the standard MNL structure.

[Bhat and Sen, 2006] conducted an application of MDCEV that models jointly the decisions of holding multiple vehicle types (passenger car, SUV, pickup truck, minivan and van) and the mileage for each type in an integrated model system; data is extracted from the 2000 San Francisco Bay Area Travel Survey (BATS). In this study the authors analyze changes in vehicle type and usage due to changes in demographics, employment status, density and operating cost. Major conclusions can be summarized as follows: (1) there is a higher preference to own and use SUVs and minivans as the number of children in the household increases; (2) households with more members or with mobility-challenged household members have a higher preference for minivans; (3) households with more workers are less likely to prefer minivans; (4) households with more men or located in less dense area prefer pickup trucks; (5) vehicle operation cost has a negative effect on vehicle ownership and usage for all vehicle types except for passenger cars; (6) households are very likely to own passenger cars but put more miles on non-passenger cars (if available).

[Bhat et al., 2009] extended this study and formulated a nested model structure that includes a multiple discrete-continuous extreme value (MDCEV) compo-

ment to analyze the choice of vehicle class/vintage and usage in the upper level and a multinomial logit (MNL) component to analyze the choice of vehicle make/model in the lower level. The model accommodates heteroscedasticity and/or error correlation in both the multiple discrete-continuous component and the single discrete choice component of the joint model using a mixing distribution. The joint model also incorporates random coefficients in one or both components of the joint model. Again, using BATS data, the study derived several important findings: (1) household with higher income or more workers have higher preference towards newer vehicles and are less likely to use non-motorized transportation modes; (2) in terms of built environment characteristics, households located in urban areas are less likely to own/use large vehicles and more likely to use non-motorized transportation modes; (3) the preference of vehicle holding and use also depends on the age, gender and ethnicity of the household head; (4) households prefer vehicles with lower purchase price and operating cost, bigger luggage and seating capacity, higher engine performance and lower greenhouse gas emissions.

In conclusion, the MDCEV model recognizes multiple discreteness and is able to handle a large number of vehicle types. It well captures the interdependence between the vehicle type and the corresponding mileage and allows more complex specification forms as heteroscedasticity and correlation. However, this model requires finer classification of vehicles as one type of vehicle cannot be chosen twice by the household. This type of models is limited by the assumption of fixed total mileage budget for every household; this implies that it is not possible to predict changes in the total number of miles in response to policy changes. Moreover, there

is only a single error term underlies both discrete and continuous choices. Overall, the MDCEV model is consistent with random utility, it is able to capture trade-offs among the usage of different types of vehicles and can accommodate a large number of vehicle classifications.

2.3.3 Bayesian Multivariate Ordered Probit and Tobit model

[Fang, 2008] developed the BMOPT (Bayesian Multivariate Ordered Probit and Tobit) model, which is composed of a multivariate ordered probit model for the discrete choices and a multivariate Tobit model for the continuous choice. Household decisions on the number of vehicles in one of the two categories (cars and trucks) considered are estimated by means of ordered probit model. The multivariate Tobit model is applied to estimate the household decisions on miles driven with each vehicle type. The joint model is formulated with an unrestricted covariance matrix for the discrete and continuous parts.

Let two latent continuous variables y_1^* and y_2^* represent the preference levels for holding cars and trucks, let latent variables y_3^* and y_4^* represent uncensored average annual miles driven by cars and trucks. Indexing household by i , $i = 1, \dots, N$, the system for discrete-continuous choice of the vehicles can be written as:

$$y_{1i}^* = \mathbf{w}'_i \beta_{11} + \ln(d_i)' \beta_{12} + \epsilon_{1i} \quad (2.6)$$

$$y_{2i}^* = \mathbf{w}'_i \beta_{21} + \ln(d_i)' \beta_{22} + \epsilon_{2i} \quad (2.7)$$

$$y_{3i}^* = \mathbf{w}'_i \beta_{31} + \ln(d_i)' \beta_{32} + \epsilon_{3i} \quad (2.8)$$

$$y_{4i}^* = \mathbf{w}'_i \beta_{41} + \ln(d_i)' \beta_{42} + \epsilon_{4i} \quad (2.9)$$

where \mathbf{w}_i is a vector of characteristics for household i ; d_i is an indicator of residential density. The number of cars, y_{1i} , and trucks, y_{2i} , held by household i are determined by the value of the corresponding latent utility y_{1i}^* and y_{2i}^* ; specifically, $y_j = 0$, if $y_j^* \leq \alpha_1$, $y_j = 1$, if $\alpha_1 < y_j^* \leq \alpha_2$, $y_j = 2$ or more, if $y_j^* > \alpha_2$, for $j = 1, 2$. Average annual miles driven by cars y_3 is observed only when a household holds at least one car; that is,

$$y_3 = y_3^*, \text{ if } y_1 = 1 \text{ or } 2 \quad (2.10)$$

$$y_3 = 0, \text{ if } y_1 = 0 \quad (2.11)$$

The same logic applies to miles driven by trucks y_4 :

$$y_4 = y_4^*, \text{ if } y_2 = 1 \text{ or } 2 \quad (2.12)$$

$$y_4 = 0, \text{ if } y_2 = 0 \quad (2.13)$$

The whole system can then be written into a SUR (seemingly unrelated regression) form:

$$y_i^* = \mathbf{x}_i\beta + \epsilon_i \quad (2.14)$$

The error structure is a multivariate normal with zero means and unrestricted covariance matrix:

$$\epsilon_i \stackrel{i.i.d.}{\sim} N(\mathbf{0}, \Sigma) \quad (2.15)$$

The likelihood function is given as following:

$$L(\beta, \Sigma; y_1, y_2, y_3, y_4) \propto \prod_{i: y_{1i}=0, y_{2i}=0} f(y_{1i}^* < \alpha_1, y_{2i}^* < \alpha_1 | \beta, \Sigma)$$

$$\begin{aligned}
& \times \prod_{i_{1i}=0, y_{2i}=1} f(y_{1i}^* < \alpha_1, \alpha_1 < y_{2i}^* < \alpha_2, y_{4i} = y_{4i}^* | \beta, \Sigma) \\
& \times \prod_{i_{1i}=0, y_{2i}=2} f(y_{1i}^* < \alpha_1, y_{2i}^* > \alpha_2, y_{4i} = y_{4i}^* | \beta, \Sigma) \\
& \times \prod_{i_{1i}=1, y_{2i}=0} f(\alpha_1 < y_{1i}^* < \alpha_2, y_{2i}^* < \alpha_1, y_{3i} = y_{3i}^* | \beta, \Sigma) \\
& \times \prod_{i_{1i}=1, y_{2i}=1} f(\alpha_1 < y_{1i}^* < \alpha_2, \alpha_1 < y_{2i}^* < \alpha_2, y_{3i} = y_{3i}^*, y_{4i} = y_{4i}^* | \beta, \Sigma) \\
& \times \prod_{i_{1i}=1, y_{2i}=2} f(\alpha_1 < y_{1i}^* < \alpha_2, y_{2i}^* > \alpha_2, y_{3i} = y_{3i}^*, y_{4i} = y_{4i}^* | \beta, \Sigma) \\
& \times \prod_{i_{1i}=2, y_{2i}=0} f(y_{1i}^* > \alpha_2, y_{2i}^* < \alpha_1, y_{3i} = y_{3i}^* | \beta, \Sigma) \\
& \times \prod_{i_{1i}=2, y_{2i}=1} f(y_{1i}^* > \alpha_2, \alpha_1 < y_{2i}^* < \alpha_2, y_{3i} = y_{3i}^*, y_{4i} = y_{4i}^* | \beta, \Sigma) \\
& \times \prod_{i_{1i}=2, y_{2i}=2} f(y_{1i}^* > \alpha_2, y_{2i}^* > \alpha_2, y_{3i} = y_{3i}^*, y_{4i} = y_{4i}^* | \beta, \Sigma)
\end{aligned}$$

The BMOPT model is convenient to implement, and can be applied to study policy implications. It is able to handle a large number of vehicles, and captures the interdependence (correlation) between the number of vehicles and the total miles driven with each vehicle type considered, it also allows flexible specifications of error terms. There are a few limitations in this model structure. Firstly, the computation becomes intensive for a large number of vehicle categories, as the number of equations to be estimated increases proportionally with the number of vehicle types. Another concern is that the ordered mechanism may not perform as well as unordered mechanism in modeling car ownership models ([Bhat and Pulugurta, 1998]). Lastly, the same variables enter both discrete and continuous sub-model. Overall, the model is well suited for predicting the changes in the number of vehicles and miles traveled for each vehicle type.

Chapter 3

Methodology

3.1 Unordered Discrete-Continuous Model

3.1.1 The Discrete Choice Sub-model

Discrete choice models forecast the outcome of a categorical dependent variable Y_{disc} using some set of predictors. All k possible alternatives of Y_{disc} have a utility $(U_0, U_1, U_2, \dots, U_k)$ that consists of one observable part (systematic utility, \mathbf{V}) and one non-observable part (error term ϵ). In my modeling framework, the observed utility is decomposed into two parts (V_k and $V_{t_k|k}$):

$$U_0 = \epsilon_0$$

$$U_1 = V_1 + \lambda V_{t_1|1} + \epsilon_1$$

$$U_2 = V_2 + \lambda V_{t_2|2} + \epsilon_2$$

...

$$U_k = V_k + \lambda V_{t_k|k} + \epsilon_k$$

Where V_k is the utility of vehicle holding decision, which depends on factors that vary over k and $V_{t_k|k}$ is the utility of vehicle type choice conditional on k , which depends on factors that vary over $t_k|k$. $t_k|k$ is the choice set containing all the possible combinations of car types and vintages while λ is a parameter to be estimated.

I adopt a multinomial logit model for the vehicle type submodel; then the probability of choosing a certain type of vehicle is:

$$P_{t_k|k} = \frac{\exp(V_{t'_k|k})}{\sum_{t_k} \exp(V_{t_k|k})} \quad (3.1)$$

Where t'_k is the chosen alternative among total alternatives t_k . The utility that the household would obtain by its choice of vehicle type can be written as:

$$J_k = \ln \sum_{t_k} \exp(V_{t_k|k})$$

Therefore the utility of the discrete choice can be further written as:

$$\begin{aligned} U_0 &= \epsilon_0 \\ U_1 &= X_1^T \beta_1 + J_1 \lambda + \epsilon_1 \\ U_2 &= X_2^T \beta_2 + J_2 \lambda + \epsilon_2 \\ &\dots \\ U_k &= X_k^T \beta_k + J_k \lambda + \epsilon_k \end{aligned}$$

Where, X_1, X_2, \dots, X_k are the attributes in the utility functions; $\beta_1, \beta_2, \dots, \beta_k$ are the parameters to be estimated; $\epsilon_0, \epsilon_1, \dots, \epsilon_k$ are the error terms. The second part of utility J_k is also important because the choice of vehicle types affects the household's probability of choosing a certain number of vehicles.

The decision maker is assumed to be rational and to choose the alternative with the biggest utility. In my econometric setting I adopt a *probit* model for the discrete problem and therefore the error terms follow a multivariate normal distribution with full, unrestricted, covariance matrix.

For simplicity, let's assume that:

$$Y = Y_{disc}$$

$$X = (X_1, \dots, X_k)$$

$$J = (J_1, \dots, J_k)$$

$$\beta = (\beta_1, \dots, \beta_k)$$

$$\epsilon = (\epsilon_0, \epsilon_1, \dots, \epsilon_k)$$

$$\Sigma := \text{Covariance of the error term}$$

The likelihood of *one* observation can be expressed as follow:

$$P(Y = y|X, J, \beta, \lambda, \Sigma) = \int_{R^{k+1}} I(X_y^T \beta_y + J_y \lambda + \epsilon_y > X_j^T \beta_j + J_j \lambda + \epsilon_j \quad \forall j \neq y) \phi(\epsilon) d\epsilon \quad (3.2)$$

The functional indicator ($I()$) ensures that the observed choice is indeed the one with the biggest utility. The subscript y indicates the predictors and coefficients of the chosen alternative and the subscript j indicate the other alternatives.

Since only differences in utility matter, the choice probability can be equivalently expressed as (k) - dimensional integrals over the differences between the errors. Suppose we differentiate against alternative y , the alternative for which we are calculating the probability. Define:

$$\tilde{\epsilon}_{jy} = \epsilon_j - \epsilon_y \quad (3.3)$$

$$\tilde{V}_{jy} = (X_j^T \beta_j + J_j \lambda) - (X_y^T \beta_y + J_y \lambda) \quad (3.4)$$

$$\tilde{\epsilon}_y = \langle \tilde{\epsilon}_{1y}, \dots, \tilde{\epsilon}_{ky} \rangle \quad (3.5)$$

where the "...” is over all alternatives except y

Then :

$$P(Y = y) = \int_{R^k} I(\tilde{V}_{jy} + \tilde{\epsilon}_{jy} < 0 \quad \forall j \neq y) \phi(\tilde{\epsilon}_y) d\tilde{\epsilon}_y \quad (3.6)$$

which is a (k) -dimensional integral over all possible values of the error differences. The difference between two normals is normal and the covariance of $\tilde{\epsilon}_y$ can be easily transferred from the covariance of ϵ ([Train, 2009, p. 99]). Detailed explanation is given in section 3.1.7.

3.1.2 The Continuous Choice Sub-model

Regression is adopted to model the continuous part of the modeling framework or the decisions on the household vehicle mileage. In a regression, the dependent variable Y_{reg} is assumed to be a linear combination of a vector of predictors X_{reg} plus some error term (ϵ_{reg}):

$$Y_{reg} = X_{reg}^T \beta_{reg} + \epsilon_{reg} \quad \epsilon_{reg} \sim N(0, \sigma^2) \quad (3.7)$$

Usually, regression is solved by using the Ordinary Least Square (OLS) estimator [Weisberg, 2005], but the same problem can be expressed in the form of a likelihood function to be maximized [McCulloch et al., 2008, p. 117]. Indeed, given β_{reg} , X_{reg} and σ^2 , the likelihood of observing y_{reg} is given by the normal density function:

$$P(y_{reg} | \beta_{reg}, X_{reg}, \sigma^2) = \phi(y_{reg} | X_{reg}^T \beta_{reg}, \sigma^2) \quad (3.8)$$

The normal density is centered at $\hat{y} = X_{reg}^T \beta_{reg}$ and has variance σ^2 .

3.1.3 The Integrated Discrete-Continuous Model

In discrete-continuous choice models, I want to model Y and Y_{reg} jointly to capture the correlation between them. In this framework, I allow the error term of the regression to be correlated with the error terms of the utilities in the probit.

The specifications of the observable part of the utilities and of the regression's \hat{y} shall remain the same, while the error terms are assumed to follow an "incremented" normal distribution:

$$(\tilde{\epsilon}_{1y}, \tilde{\epsilon}_{2y}, \dots, \tilde{\epsilon}_{ky}, \epsilon_{reg}) \sim MN(0, \Sigma_{k+1}) \quad (3.9)$$

Therefore, the probability of observing Y and Y_{reg} is the product of the probability of observing Y and the probability of observing Y_{reg} conditional on observing Y

$$P(Y, Y_{reg}) = P(Y)P(Y_{reg}|Y) \quad (3.10)$$

or the product of the probability of observing Y_{reg} and the probability of observing Y conditional on observing Y_{reg}

$$P(Y, Y_{reg}) = P(Y_{reg})P(Y|Y_{reg}) \quad (3.11)$$

This is a general result about conditioning with random variables. [Rice, 2007, p. 88] The probability can be written as a form of density function:

$$f(Y, Y_{reg}) = f(Y)f(Y_{reg}|Y) \quad (3.12)$$

or

$$f(Y, Y_{reg}) = f(Y_{reg})f(Y|Y_{reg}) \quad (3.13)$$

3.1.4 Estimation with Simulation - 1st Attempt

I initially tried to calculate the probability function with the equation (3.10) as discussed in the last section, namely:

$$P(Y, Y_{reg}) = P(Y)P(Y_{reg}|Y) \quad (3.14)$$

This function consists of two parts: $P(Y)$ and $P(Y_{reg}|Y)$. The probability of probit ($P(Y)$) has integrals thus has no closed mathematical form. We could rely on simulation as described in [Train, 2009, p. 117]:

$$\hat{P}(Y = y|X, J, \beta, \lambda, \Sigma) = \frac{1}{B} \sum_{i=1}^B I(X_y^T \beta_y + J_y \lambda + \epsilon_y^{(i)} > X_j^T \beta_j + J_j \lambda + \epsilon_j^{(i)} \quad \forall j \neq y) \quad (3.15)$$

Where $\epsilon^{(i)}$ is a draw from a multivariate normal with mean 0 and variance Σ and B is the number of simulations.

However, it is not clear which functional form can be given to the conditional distribution $f(Y_{reg}|Y)$. Considering that we have a sample of points from the conditional distribution that can be used in order to estimate it, let's assume the following:

| | | |
|--------------------|---------------------|---------------------------------------------------------------------------------------------------|
| $\epsilon^{(i)}$ | $i = 1, \dots, B$ | Draws from the multivariate normal distribution |
| $\epsilon^{(i y)}$ | $i = 1, \dots, B^*$ | Subset of draws for which the biggest utility in the probit simulation was the observed Y |

In other words, when we simulate the probit probability, we keep the error terms that correspond to the regression whenever (conditional) the biggest utility is the

one of the observed choice. We rely on the sample $\{\epsilon^{(i|y)}\}_{i=1}^{B^*}$ to estimate $f(Y_{reg}|Y)$.

Consider the following illustrative example (3.1, where the chosen alternative among "Car", "Bus", "Bike" is "Car". We simulate the utilities B = 10 times:

| | Simulation # | | | | | | | | | |
|-----------|--------------|------------|------------|------------|-----------|------------|------------|------------|------------|----------|
| Utilities | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Car | 9.5 | 8.2 | 9.5 | 7.3 | 10 | 1.2 | 4.8 | 1.4 | 6.1 | 7.4 |
| Bus | 2.7 | 1.7 | 1.7 | 4.6 | 4.2 | 5.8 | 5.2 | 4.6 | 8.1 | 8 |
| Bike | 9.2 | 8.6 | 8.9 | 2.3 | 8.3 | 5.5 | 3.7 | 9.5 | 8.5 | 3.5 |

Table 3.1: Illustration example

In that case we would use the error terms of the regression corresponding to indexes 1,3,4 and 5 to estimate the density of the continuous variable.

To conclude, the problem of estimating the model likelihood reduces to collecting the regression error terms when we compute the probit. Those error terms are the product of the simulation and the problem reduces to a density estimation problem.

Interpretation of a density function

We know that the interpretation of a density function is that [Rice, 2007, p. 48]:

$$f(y_{reg})2\delta \approx P(y_{reg} - \delta < Y_{reg} < y_{reg} + \delta) \quad (3.16)$$

That is, the density of a random variable Y_{reg} evaluated at y_{reg} times the length of a small interval (δ) is approximately equal to the probability that Y_{reg} lies in this interval centered in y_{reg} . We can estimate the left hand side of this expression with

random draws, then we estimate $f(y_{reg})$ with:

$$\hat{p} = \hat{P}(y_{reg} - \delta < Y_{reg} < y_{reg} + \delta) = \frac{1}{B^*} \sum_{i=1}^{B^*} I(y_{reg} - \delta < X_{reg}^T \beta_{reg} + \epsilon_{reg}^i < y_{reg} + \delta) \quad (3.17)$$

$$f(y_{reg}) \approx \frac{\hat{p}}{2\delta} \quad (3.18)$$

To name only a few problems that can arise, it is possible that $\hat{p} = 0$ and we need to carefully select δ . However, this approximation is computable.

Kernel density estimation

Kernel density estimation uses a kernel, that is a density function whose purpose is to "weight" all the points in the sample in order to estimate the density [Parzen, 1962]:

$$\hat{f}(x) = \frac{1}{B^*} \sum_{i=1}^{B^*} K_h(x_i - x) \quad (3.19)$$

Where $K(\cdot)$ is a symmetric density function and

$$K_h(x) = \frac{1}{h} K(x/h) \quad (3.20)$$

Note that $K_h(\cdot)$ is also a symmetric density function that is only a scale transformation of $K(\cdot)$. We could use for instance a gaussian kernel (normal) in which case we would not have the problem of estimating the density by 0. However, this method is usually computationally expensive. A sample of 1,000 observations with 1,000 simulations each would require to compute the normal density one million time (!) to estimate the likelihood only once. The problem of finding h remains. The method described in the previous section happens to be the kernel density estimation using a uniform kernel where h was referred to as δ

- Large δ will give a biased estimate of the density
- Small δ will give a volatile estimate of the density

Possible simple approximation

There is little chance that we can derive the exact conditional distribution of Y_{reg} , but we may be able to find a known distribution that is a good approximation for it. If a good distribution that can estimate the conditional distribution of Y_{reg} given Y can be found, then it is possible to:

- Find the MLE estimator of the parameter (θ) of this distribution;
- Apply it to the conditional residuals from the probit simulation;
- Estimate $f(y_{reg}|y)$ with $f(y_{reg}|\hat{\theta}_{mle})$.

I tested this idea with the 2009 NHTS data used for the real case study proposed in this paper ([U.S. Department of Transportation, 2009]). The conditional residuals are those found at the maximum likelihood estimates:

We note that:

- Residuals appear to be normally distributed;
- The mean of the distribution of Y_{reg} is approximately $X_{reg}\beta_{reg} + \mu_{e(\cdot|y)}$
- The variance of the distribution of Y_{reg} is approximately $\sigma_{e(\cdot|y)}^2$

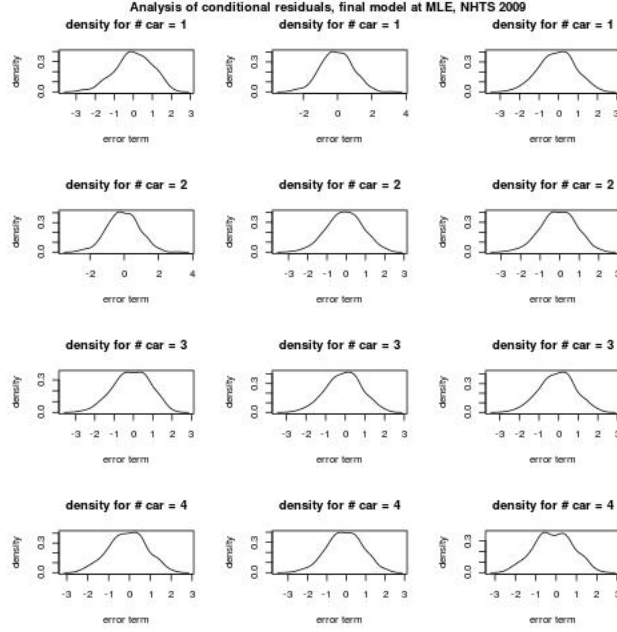


Figure 3.1: Distribution of conditional residuals

Where $\mu_{\epsilon(\cdot|y)}$ and $\sigma_{\epsilon(\cdot|y)}^2$ are the mean and variance of the conditional residuals. Therefore, the computation of the conditional density of y_{reg} is much more stable than the estimation obtained with a uniform kernel and much faster than the one issued by a Gaussian kernel.

In order to be able to estimate the conditional density, at least two conditional residuals are needed. Obviously, the precision of the density estimation depends on how well the discrete variables are predicted earlier in the simulation. However, given that the normal assumption seems to be a good approximation, there is no reason to believe that estimating it with few residuals will cause problems or will deteriorate estimates. The following table describes the amount of successes I observed for the probit, at convergence, for 1000 simulations:

| min | 1 st quartile | median | mean | 3 rd quartile | max |
|-----|--------------------------|--------|-------|--------------------------|-------|
| 4 | 282.8 | 415 | 403.4 | 540 | 954.0 |

Thus, at convergence, I always observe at least 2 successes; moreover, I do not face the problem of estimating a zero probability for the probit or the problem of lack of data for the regression density estimation.

The final Simulated Log Likelihood of the model is given by the following formula:

$$SLL(\beta, \beta_{reg}, \Sigma | Y, Y_{reg}, X, J, X_{reg}) = \sum_{i=1}^n \log \left(\frac{B_i^*}{B} \times \phi(y_{i,reg} | X_{reg}^T \beta_{reg} + \mu_{\epsilon(\cdot|y_i)}, \sigma_{\epsilon(\cdot|y_i)}^2) \right) \quad (3.21)$$

Where:

$$B_i^* := \text{number of success in } i^{th} \text{ probit simulation}$$

3.1.5 Estimation with Simulation - Modified Approach

The method that introduced in the previous section is intuitive and mathematically feasible, however, there are two drawbacks: (1) That by assuming the sup-sample of the error terms in regression is normally distributed may not reveal the true distribution (2) The computational error in the estimation accumulates with both the simulation in probit and re-sampling the error terms in regression. These problems might result in bias of the estimated coefficients thus worse goodness of fit.

Here I adopt the second form of the joint probability function (equation 3.11):

$$P(Y, Y_{reg}) = P(Y_{reg})P(Y|Y_{reg}) \quad (3.22)$$

In a multivariate normal distribution, if (\mathbf{A}, \mathbf{B}) follow a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad (3.23)$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \quad (3.24)$$

then $(\mathbf{A}|\mathbf{B} = \mathbf{B}_1)$ follows a multivariate normal distribution with mean and variance

$$\boldsymbol{\mu}_A = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{B}_1 - \boldsymbol{\mu}_2) \quad (3.25)$$

$$\boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \quad (3.26)$$

Back to the problem,

$$\begin{bmatrix} \tilde{\boldsymbol{\epsilon}}_y \\ \epsilon_{reg} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{disc} & \boldsymbol{\Sigma}_{disc,reg} \\ \boldsymbol{\Sigma}_{reg,disc} & \sigma^2 \end{bmatrix}\right) \quad (3.27)$$

$$P(Y_{reg}) = \phi(err|\mu = 0, \sigma^2 = \sigma_{reg}^2) \quad (3.28)$$

where $err = Y_{reg} - \hat{Y}_{reg}$ and

$$P(Y|Y_{reg}) = \int_{\mathbb{R}^{k-1}} I(\tilde{V}_{jy} + \tilde{\epsilon}_{jy} < 0 \quad \forall j \neq y) \varphi(\tilde{\boldsymbol{\epsilon}}_y) d\tilde{\boldsymbol{\epsilon}}_y \quad (3.29)$$

where, $\varphi(\boldsymbol{\epsilon})$ is the density function of a multivariate distribution and

$$\tilde{\boldsymbol{\epsilon}}_y \sim \mathcal{N}\left(0 + \frac{\boldsymbol{\Sigma}_{disc,reg}}{\boldsymbol{\Sigma}_{disc}}(err - 0), \sigma_{reg}^2 - \frac{\boldsymbol{\Sigma}_{reg,disc}\boldsymbol{\Sigma}_{disc,reg}}{\boldsymbol{\Sigma}_{disc}}\right) \quad (3.30)$$

$$\hat{P}(Y|Y_{reg}) = \frac{1}{B} \sum_{i=1}^B I(\tilde{V}_{jy} + \tilde{\epsilon}_{jy}^{(i)} < 0 \quad \forall j \neq y) \quad (3.31)$$

Where $\tilde{\epsilon}_{jy}^{(i)}$ is a draw from a multivariate normal with mean $(0 + \frac{\Sigma_{disc,reg}}{\Sigma_{disc}}(err - 0))$ and variance $\sigma_{reg}^2 - \frac{\Sigma_{reg,disc}\Sigma_{disc,reg}}{\Sigma_{disc}}$ and B is the number of simulations.

Then, the final Simulated Log Likelihood of the model is given by the following formula:

$$\begin{aligned} SLL(\beta, \beta_{reg}, \Sigma|Y, Y_{reg}, X, J, X_{reg}) = \\ \sum_{i=1}^n \log \left(\frac{B_i^*}{B} \times \phi(y_{i,reg}|X_{reg}^T \beta_{reg}, \sigma_{reg}^2) \right) \end{aligned} \quad (3.32)$$

Where:

$$B_i^* := \text{number of success in } i^{th} \text{ probit simulation}$$

3.1.6 Estimation with Numerical Computation

The modified estimation method with simulation greatly reduces the computational errors and bias because we know the exact distributions of $P(Y_{reg})$ and conditional probability $P(Y|Y_{reg})$, however, the accuracy of the estimated coefficients is highly depended on the draws from the simulation *and* it has very high computational cost when the number of draws increases.

In order to investigate the problems from the simulation, I also adopted a numerical method to compute the multivariate normal probabilities, which is developed by [Genz, 1992]. Genz (1992) proposed a transformation of multivariate normal probability function that simplifies the problem and places it into a form that allows

efficient calculation using standard numerical multiple integration algorithms. The method is explained in detail as following.

Numerical computation of multivariate normal probabilities ([Genz, 1992])

Given the multivariate normal distribution function

$$F(\mathbf{a}, \mathbf{b}) = \frac{1}{\sqrt{|\Sigma|(2\pi)^m}} \int_{a_1}^{b_1} \int_{a_2}^{b_2} \dots \int_{a_m}^{b_m} e^{-\frac{1}{2}\theta^t \Sigma^{-1} \theta} d\theta \quad (3.33)$$

where $\theta = (\theta_1, \theta_2, \dots, \theta_m)^t$ and Σ is an $m \times m$ symmetric positive definite covariance matrix.

A sequence of three three transformations are used to transform the original integral into an integral over a unit hyper-cube. This sequence begins with a Cholesky decomposition transformation $\theta = C\mathbf{y}$, where CC^t is the Cholesky decomposition of the covariance matrix Σ . Now $\theta^t \Sigma^{-1} \theta = \mathbf{y}^t C^t C^{-1} C^{-1} C \mathbf{y} = \mathbf{y}^t \mathbf{y}$, and $d\theta = |C| d\mathbf{y} = |\Sigma|^{\frac{1}{2}} d\mathbf{y}$. Since $\mathbf{a} \leq \theta = C\mathbf{y} \leq \mathbf{b}$ implies $(a_i - \sum_{j=1}^{i-1} c_{ij} y_j)/c_{ii} \leq y_i \leq (b_i - \sum_{j=1}^{i-1} c_{ij} y_j)/c_{ii}$ for $i = 1, 2, \dots, m$, we have

$$F(\mathbf{a}, \mathbf{b}) = \frac{1}{\sqrt{(2\pi)^m}} \int_{a'_1}^{b'_1} e^{-\frac{y_1^2}{2}} \int_{a'_2(y_1)}^{b'_2(y_1)} e^{-\frac{y_2^2}{2}} \dots \int_{a'_m(y_1, \dots, y_{m-1})}^{b'_m(y_1, \dots, y_{m-1})} e^{-\frac{y_m^2}{2}} d\mathbf{y} \quad (3.34)$$

with $a'_i(y_1, \dots, y_{i-1}) = (a_i - \sum_{j=1}^{i-1} c_{ij} y_j)/c_{ii}$ and $b'_i(y_1, \dots, y_{i-1}) = (b_i - \sum_{j=1}^{i-1} c_{ij} y_j)/c_{ii}$

Now each of the y_i 's can be transformed separately using $y_i = \Phi^{-1}(z_i)$, where

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-\frac{1}{2}\theta^2} d\theta \quad (3.35)$$

This is the standard univariate normal distribution. After these transformations, $F(\mathbf{a}, \mathbf{b})$ becomes

$$F(\mathbf{a}, \mathbf{b}) = \int_{d_1}^{e_1} \int_{d_2(z_1)}^{e_2(z_1)} \dots \int_{d_m(z_1, \dots, z_{m-1})}^{e_m(z_1, \dots, z_{m-1})} dz \quad (3.36)$$

with $d_i(z_1, \dots, z_{i-1}) = \Phi((a_i - \sum_{j=1}^{i-1} c_{ij} \Phi^{-1}(z_j)) / c_{ii})$ and $e_i(z_1, \dots, z_{i-1}) = \Phi((b_i - \sum_{j=1}^{i-1} c_{ij} \Phi^{-1}(z_j)) / c_{ii})$.

The integrand in this form is much simpler than the original integrand. The integration region is more complicated, however, and cannot be handled directly with standard numerical multiple integration algorithms. A solution to this problem is to put the integral into a constant limit form using $z_i = d_i + w_i(e_i - d_i)$. After this final set of transformations,

$$F(\mathbf{a}, \mathbf{b}) = (e_1 - d_1) \int_0^1 (e_2 - d_2) \dots \int_0^1 (e_m - d_m) \int_0^1 d(w) \quad (3.37)$$

with $d_i = \Phi((a_i - \sum_{j=1}^{i-1} c_{ij} \Phi^{-1}(d_j + w_j(e_j - d_j))) / c_{ii})$ and $e_i = \Phi((b_i - \sum_{j=1}^{i-1} c_{ij} \Phi^{-1}(d_j + w_j(e_j - d_j))) / c_{ii})$.

The innermost integral over w_m can be done explicitly because d_m and e_m have no dependence on w_m , so the complete sequence of transformations has reduced the number of integration variables by one.

Estimate the discrete continuous model with numerical computation

Recall that the likelihood of one observation is:

$$P(Y|Y_{reg}) = \int_{R^{k-1}} I(\tilde{V}_{jy} + \tilde{\epsilon}_{jy} < 0 \quad \forall j \neq y) \varphi(\tilde{\epsilon}_{\mathbf{y}}) d\tilde{\epsilon}_{\mathbf{y}} \quad (3.38)$$

After taking a few transformations:

$$\begin{aligned} P(Y|Y_{reg}) &= \int_{R^{k-1}} I(\tilde{\epsilon}_{jy} < -\tilde{V}_{jy} \quad \forall j \neq y) \varphi(\tilde{\epsilon}_{\mathbf{y}}) d\tilde{\epsilon}_{\mathbf{y}} \\ &= \int_{-\infty}^{-\tilde{V}_{1y}} \int_{-\infty}^{-\tilde{V}_{2y}} \dots \int_{-\infty}^{-\tilde{V}_{(k-1)y}} \varphi(\tilde{\epsilon}_{\mathbf{y}}) d\tilde{\epsilon}_{\mathbf{y}} \end{aligned}$$

This distribution function can be solved by Genz's method described above, thus no simulation is needed.

3.1.7 Normalization of Covariance Matrix

In the random utility maximization theory, the absolute level of utility is irrelevant to both the decision maker's behavior and the researcher's model ([Train, 2009]). If a constant is added to the utility of all alternatives, the alternative with the highest utility does not change. In other words, "Only differences in utility matter" and "The scale of the utility is arbitrary".

The decision maker chooses the same alternative with $U_{nj} \forall j$ as with $U_{nj} + k \forall j$ for any constant k . The level of utility does not matter from the researcher's perspective either. The choice probability is $P_{ni} = Prob(U_{ni} > U_{nj} \forall j \neq i) = Prob(U_{ni} - U_{nj} > 0 \forall j \neq i)$, which depends only on the difference in utility, not its absolute level. When utility is decomposed into the observed and unobserved parts, $P_{ni} = Prob(\epsilon_{nj} - \epsilon_{ni} > V_{ni} - V_{nj} \forall j \neq i)$, which also depends only on the difference.

The fact that only differences in utility matter has several implications for the identification and specification of discrete choice models. In general it means that

the only parameters that can be estimated (that is, are identified) are those that capture differences across alternatives.

Similarly, the scale of the utility does not matter because the utility of each alternative can be multiplied by a (positive) constant without changing which alternative has the highest utility.

In logit and nested logit models, the normalization for scale and level occurs automatically with the distributional assumptions that are placed on the error terms. As a result, normalization does not need to be considered explicitly for these models.

With probit models, however, normalization for scale and level does not occur automatically. The researcher must normalize the model directly.

Normalization of the model is related to parameter identification. A parameter is identified if it can be estimated, and is unidentified if it cannot be estimated.

An example of an unidentified parameter is k in the utility specification $U_{nj} = V_{nj} + k + \epsilon_{nj}$. While the researcher might write utility in this way, an might want to estimate k to obtain a measure of the overall level of utility, doing so is impossible. The behavior of the decision maker is unaffected by k , and so the researcher cannot infer its value from the choices that decision maker have made.

State directly, parameters that do not affect the behavior of decision makers cannot be estimated. In an unnormalized model, parameters can appear that are not identified; these parameters relate to the scale and level of utility, which do not affect behavior. Once the model is normalized, these parameters disappear. The difficulty arises because it is not always obvious which parameters relate to scale and level. In the preceding example, the fact that k is unidentified is fairly obvious.

In many cases, it is not at all obvious which parameters are identified.

In this study, the procedure proposed by [Train, 2009] has been applied to normalize the probit model and assure that all parameters are identified.

The probit model has five alternatives, and utility is expressed as $U_{nj} = V_{ni} + \epsilon_{nj}$, $j = 0, 1, \dots, 4$. The vector of errors is $\epsilon'_n = \langle \epsilon_{n1}, \epsilon_{n1}, \dots, \epsilon_{n4} \rangle$. It is normally distributed with zero mean and a covariance matrix that can be expressed explicitly as

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \cdot & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \cdot & \cdot & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \cdot & \cdot & \cdot & \sigma_{44} & \sigma_{45} \\ \cdot & \cdot & \cdot & \cdot & \sigma_{55} \end{pmatrix} \quad (3.39)$$

Where the dots refer to the corresponding elements on the upper part of the matrix. Note that there are 15 elements in this matrix, that is, 15 distinct σ 's representing the variance and covariance among the five errors. In general, a model with J alternatives has $J(J+1)/2$ distinct elements in the covariance matrix of the errors.

To take account of the fact that the level of utility is irrelevant, I take utility differences. Following the procedure from [Train, 2009], I take differences with respect to the first alternative. Define error differences as $\tilde{\epsilon}_{nj1} = \epsilon_{nj} - \epsilon_{n1}$ for $j = 2, 3, 4, 5$, and define the vector of error differences as $\tilde{\epsilon}_{n1} = \langle \tilde{\epsilon}_{n21}, \tilde{\epsilon}_{n31}, \tilde{\epsilon}_{n41}, \tilde{\epsilon}_{n51} \rangle$. Note that the subscript 1 in $\tilde{\epsilon}_{n1}$ means that the error differences are against the first alternative, rather than that the errors are for the first alternative.

The covariance matrix for the vector of error differences takes the form

$$\tilde{\Omega}_1 = \begin{pmatrix} \theta_{22} & \theta_{23} & \theta_{24} & \theta_{25} \\ \cdot & \theta_{33} & \theta_{34} & \theta_{35} \\ \cdot & \cdot & \theta_{44} & \theta_{45} \\ \cdot & \cdot & \cdot & \theta_{55} \end{pmatrix} \quad (3.40)$$

Where the θ 's relate to the original σ 's as follows:

$$\begin{aligned} \theta_{22} &= \sigma_{22} + \sigma_{11} - 2\sigma_{12} \\ \theta_{33} &= \sigma_{33} + \sigma_{11} - 2\sigma_{13} \\ \theta_{44} &= \sigma_{44} + \sigma_{11} - 2\sigma_{14} \\ \theta_{55} &= \sigma_{55} + \sigma_{11} - 2\sigma_{15} \\ \theta_{23} &= \sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13} \\ \theta_{24} &= \sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14} \\ \theta_{25} &= \sigma_{25} + \sigma_{11} - \sigma_{12} - \sigma_{15} \\ \theta_{34} &= \sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14} \\ \theta_{35} &= \sigma_{35} + \sigma_{11} - \sigma_{13} - \sigma_{15} \\ \theta_{45} &= \sigma_{45} + \sigma_{11} - \sigma_{14} - \sigma_{15} \end{aligned}$$

This matrix is obtained using the transformation matrix M_1 as $\tilde{\Omega}_1 = M_1 \Omega M_1'$.

$$\begin{aligned} M_1 &= \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{pmatrix} \\ \tilde{\Omega}_1 &= \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \cdot & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \cdot & \cdot & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \cdot & \cdot & \cdot & \sigma_{44} & \sigma_{45} \\ \cdot & \cdot & \cdot & \cdot & \sigma_{55} \end{pmatrix} \begin{pmatrix} -1 & -1 & -1 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{22} + \sigma_{11} - 2\sigma_{12} & \sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13} & \sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14} & \sigma_{25} + \sigma_{11} - \sigma_{12} - \sigma_{15} \\ \cdot & \sigma_{33} + \sigma_{11} - 2\sigma_{13} & \sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14} & \sigma_{35} + \sigma_{11} - \sigma_{13} - \sigma_{15} \\ \cdot & \cdot & \sigma_{44} + \sigma_{11} - 2\sigma_{14} & \sigma_{45} + \sigma_{11} - \sigma_{14} - \sigma_{15} \\ \cdot & \cdot & \cdot & \sigma_{55} + \sigma_{11} - 2\sigma_{15} \end{pmatrix} \end{aligned}$$

To set the scale of the utility, one of the diagonal elements is normalized. The top-left element of $\tilde{\Omega}_1$ is set to 1. This normalization for scale gives the following covariance matrix:

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta_{23}^* & \theta_{24}^* & \theta_{25}^* \\ \cdot & \theta_{33}^* & \theta_{34}^* & \theta_{35}^* \\ \cdot & \cdot & \theta_{44}^* & \theta_{45}^* \\ \cdot & \cdot & \cdot & \theta_{55}^* \end{pmatrix} \quad (3.41)$$

The θ^* 's relate to the original σ 's as follows:

$$\begin{aligned} \theta_{33}^* &= \frac{\sigma_{33} + \sigma_{11} - 2\sigma_{13}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\ \theta_{44}^* &= \frac{\sigma_{44} + \sigma_{11} - 2\sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\ \theta_{55}^* &= \frac{\sigma_{55} + \sigma_{11} - 2\sigma_{15}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\ \theta_{23}^* &= \frac{\sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\ \theta_{24}^* &= \frac{\sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\ \theta_{25}^* &= \frac{\sigma_{25} + \sigma_{11} - \sigma_{12} - \sigma_{15}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\ \theta_{34}^* &= \frac{\sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\ \theta_{35}^* &= \frac{\sigma_{35} + \sigma_{11} - \sigma_{13} - \sigma_{15}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\ \theta_{45}^* &= \frac{\sigma_{45} + \sigma_{11} - \sigma_{14} - \sigma_{15}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \end{aligned}$$

There are 9 elements in $\tilde{\Omega}_1^*$. These are the only identified parameters in the model. This number is less than the 15 elements that enter Ω . Each θ^* is a function of the σ 's. Since there are 9 θ^* s and 15 σ s, it is not possible to solve for all the σ 's from estimated values of the θ^* 's. It is therefore not possible to obtain estimates of all the σ 's.

In general, a model with J alternatives and an unrestricted covariance matrix will have $[(J - 1)J/2] - 1$ covariance parameters when normalized, compared to the $J(J+1)/2$ parameters when unnormalized. Only $[(J - 1)J/2] - 1$ are identified. This reduction in the number of parameters is *not* a restriction. The reduction in the number of parameters is a normalization that simply eliminates irrelevant aspects of the original covariance matrix, namely the scale and level of utility. The 15 elements in Ω allow for variance and covariance that is due simply to scale and level, which has no relevance for behavior. Only the 9 elements in $\tilde{\Omega}_1^*$ contain information about the variance and covariance of errors independent of scale and level. In this sense, only the 9 parameters can be estimated.

3.2 Ordered Discrete-Continuous Model

3.2.1 The Discrete and Continuous Sub-models and the Integrated Model

The ordered response structure uses latent variables to represent the vehicle ownership propensity of the household, thus it is not consistent with utility maximization theory. Suppose two latent variables y_d and y_r represent the preference levels for vehicle holding and vehicle usage (annual miles traveled). The ordered discrete-continuous model can be written as:

$$y_d = X_d^T \beta_d + \epsilon_d$$

$$y_r = X_r^T \beta_r + \epsilon_r$$

where, X_d and X_r are explanatory variables for the discrete choice and continuous choice, β_d and β_r are the coefficients to be estimated, ϵ_d and ϵ_r are the error terms, respectively. The number of vehicles holding by the household (Y) is determined by the value of latent variable y_d , specifically:

$$\begin{aligned}
 Y = 0 & \quad \text{if } y_d < \alpha_1 \\
 Y = 1 & \quad \text{if } \alpha_1 < y_d < \alpha_2 \\
 Y = 2 & \quad \text{if } \alpha_2 < y_d < \alpha_3 \\
 & \dots \\
 Y = k - 1 & \quad \text{if } \alpha_{k-1} < y_d < \alpha_k \\
 Y = k & \quad \text{if } \alpha_k < y_d
 \end{aligned}$$

Where $\alpha_1, \alpha_2, \dots, \alpha_{k-1}$ and α_k are the cut-points of the ordered probit equations. Similarly, in order to jointly to capture the correlation between the discrete and continuous parts, I allow the error terms to be correlated. Thus, the error terms follow a bivariate normal distribution:

$$(\epsilon_d, \epsilon_r) \sim BN(0, \Sigma)$$

Therefore, the model is composed of an ordered probit model and a regression with unrestricted correlation between the error terms.

3.2.2 Estimation with Numerical Computation

In the bivariate normal distribution, for example, if (X, Y) follow a bivariate normal distribution with mean (μ_x, μ_y) and covariance Σ , then $(Y|X = x)$ follows a normal distribution with mean $\mu_y + \rho(x - \mu_x)\frac{\sigma_x}{\sigma_y}$ and variance $\sigma_y^2(1 - \rho^2)$.

If Y_d follows an ordered probit and Y_r follows a regression with both error

terms correlated, we can write, *analytically*, the likelihood of one observation:

$$P(Y_d, Y_r) = P(Y_r)P(Y_d|Y_r)$$

or:

$$P(Y_r, Y_d) = P(Y_d)P(Y_r|Y_d)$$

Both approaches are feasible but the first one use *only* the property mentioned above, thus requires no numerical integration nor simulation. The second method requires numerical integration. if:

$$\rho = \frac{\sigma_{r,d}^2}{\sigma_r \sigma_d}$$

$$Z = X_d^t \beta_d$$

$$err = Y_r - \hat{Y}_r$$

Recall that α_k is the k-th cut-point used in the ordered probit to discretize the latent continuous variable, then :

$$P(Y_r) = \phi(err|\mu = 0, \sigma^2 = \sigma_r^2)$$

and:

$$P(Y_d|Y_r) = P(\alpha_{Y_d} < Z + \epsilon_d < \alpha_{Y_{d+1}}|Y_r)$$

$$= P((a = \alpha_{Y_d} - Z) < \epsilon_d < (b = \alpha_{Y_{d+1}} - Z)|Y_r)$$

Conditional on the regression, the only effect is that the error term of the ordered probit and its variance are:

$$\mu_{cond} = 0 + \rho(err - 0) \frac{\sigma_r}{\sigma_d}$$

$$\sigma_{cond}^2 = \sigma_d^2(1 - \rho^2)$$

Thus the conditional probability is simply:

$$P(Y_d|Y_r) = \Phi(b|\mu = \mu_{cond}, \sigma^2 = \sigma_{cond}^2) - \Phi(a|\mu = \mu_{cond}, \sigma^2 = \sigma_{cond}^2)$$

And the likelihood can be written like this :

$$L = \phi(err|\mu = 0, \sigma^2 = \sigma_r^2) \left(\Phi(b|\mu = \mu_{cond}, \sigma^2 = \sigma_{cond}^2) - \Phi(a|\mu = \mu_{cond}, \sigma^2 = \sigma_{cond}^2) \right)$$

3.3 Endogeneity

In a statistical model, a parameter or variable is said to be endogenous when there is a correlation between the parameter or variable and the error term. *Endogeneity* can arise as a result of measurement error, autoregression with autocorrelated errors, simultaneity, omitted variables, and sample selection errors. Broadly, a loop of causality between the independent and dependent variables of a model leads to *endogeneity*.

For example, in a simple supply and demand model, when predicting the quantity demanded in equilibrium, the price is endogenous because producers change their price in response to demand and consumers change their demand in response to price. In this case, the price variable is said to have total *endogeneity* once the demand and supply curves are known. In contrast, a change in consumer tastes or preferences would be an exogenous change on the demand curve.

The problem of *endogeneity* occurs when the independent variable is correlated with the error term in a regression model. This implies that the regression coefficient in an Ordinary Least Squares (OLS) regression is biased, however if the correlation is not contemporaneous, then it may still be consistent. There are many methods of overcoming this, including instrumental variable regression and Heckman selection correction.

In conclusion:

- An *endogenous* variable is one that is correlated with ϵ ;
- An *exogenous* variable is one that is uncorrelated with ϵ .

Generally, Instrumental Variables (IV) estimation is used when the model has endogenous variables. IV can thus be used to address the following important threats to internal validity:

- Omitted variable bias from a variable that is correlated with explanatory variables (\mathbf{X}) but is unobserved, so cannot be included in the regression;
- Simultaneous causality bias (endogenous explanatory variables; \mathbf{X} causes Y , Y causes \mathbf{X});
- Errors-in-variables bias (X is measured with error)

Instrumental variables regression can eliminate bias from these three sources.

An *instrumental variable*, Z is uncorrelated with the disturbance ϵ but is correlated with X . With this new variable, the IV estimator should capture only the effects on Y of shifts in X induced by Z whereas the OLS estimator captures

not only the direct effect of X on Y but also the effect of the included measurement error and/or endogeneity. IV is not as efficient as OLS (especially if Z only weakly correlated with X , i.e. when we have so-called "weak instruments") and only has large sample properties (consistency).

In order for a variable, z , to serve as a valid instrument for x , the following must be true

- The instrument must be exogenous (*instrument exogeneity*)

$$\text{Cov}(z, \epsilon) = 0$$

- The instrument must be correlated with the endogenous explanatory variable x (*instrument relevance*)

$$\text{Cov}(z, x) \neq 0$$

One computational method which can be used to calculate IV estimates is two-stage least-squares (2SLS or TSLS). In the first stage, each endogenous covariate in the equation of interest is regressed on all of the exogenous variables in the model, including both exogenous covariates in the equation of interest and the excluded instruments. The predicted values from these regressions are obtained. In the second stage, the regression of interest is estimated as usual, except that in this stage each endogenous covariate is replaced with the predicted values from its first stage model from the first stage.

- Stage 1: Regress X on all the exogenous regressors: regress X on Z_1, \dots, Z_m using OLS; compute predicted values \hat{X}

- Stage 2: Regress Y on \hat{X} and other explanatory variables using OLS.

In this study, it should be noted that the cost variable in the continuous part is estimated with an instrumental variable approach. This approach is required because when the household chooses which vehicle(s) it owns, it effectively chooses the operating cost of driving the selected vehicle(s) [Train, 1986]. The operating cost (endogenous variable) is regressed on the exogenous variables; those include household income, number of drivers, number of workers, owned or rental house, dummy of urban area, urban size, age of the household head and the education level of the household head. The predicted values from these regressions are obtained and used as exogenous variables to explain the vehicle miles traveled.

3.4 Goodness of Fit Measures

In statistics, the coefficient of determination ρ^2 is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. ρ^2 is most often seen as a number between 0 and 1.0, used to describe how well a regression line fits a set of data. An ρ^2 near 1.0 indicates that a regression line fits the data well, while an ρ^2 closer to 0 indicates a regression line does not fit the data very well. It is the proportion of variability in a data set that is accounted for by the statistical model. It provides a measure of how well future outcomes are likely to be predicted by the model.

In this study, the log-likelihood values from different models cannot be directly compared because of the different model structure, number of parameters

and number of observations. Therefore, we calculate the adjusted R^2 as follows:

$$\rho^2 = 1 - \frac{LL(\hat{\beta}) - n_{par}}{LL(0)}$$

Where $LL(\hat{\beta})$ is the log-likelihood value at convergence, $LL(0)$ is the log-likelihood value at zero, and n_{par} is the number of parameters estimated in the model.

A non-nested test has been also conducted for the ordered and unordered models. This test determines if the adjusted ρ^2 of two non-nested models are significantly different. I use the same method as in Bhat and Pulugurta (1998):

”If the difference in the adjusted ρ^2 is τ , then the probability that this difference could have occurred by chance is no larger than

$$\Phi \left\{ -[-2\tau LL(0) + (n_{par,2} - n_{par,1})]^{0.5} \right\}$$

in the asymptotic limit. A small value of the probability of chance occurrence indicates that the difference is statistically significant and that the model with the higher value of adjusted likelihood ratio index is to be preferred.”

Chapter 4

Data Sources

4.1 National Household Travel Survey (NHTS)

The main data sources used in this dissertation for car ownership modeling are extracted from the 2009 National Household Travel Surveys (NHTS). NHTS is conducted by the Federal Highway Administration (FHWA), the United States Department of Transportation (U.S.DOT) and serves as the nation's inventory of daily travel. It collected travel data from a national sample of the civilian, non-institutionalized population of the United States. NHTS is a microdata dataset, which contains a total of 150,147 households, 351,275 persons, 309,163 vehicles and 1,167,321trips in the final 2009 NHTS dataset (FHWA, 2011).

The NHTS is conducted as a telephone survey, using Computer-Assisted Telephone Interviewing (CATI) technology. The NHTS dataset includes all interviews from the national sample and the Add-on partners. The weighting factors have been adjusted to account for the oversampling in the Add-on areas.

States and MPOs have the unique opportunity to purchase samples of the household travel survey when it is conducted, approximately every five to seven years. These additional samples, along with random national samples collected in the add-on area, are compiled into a cleaned geocoded database for ready application to local planning and forecasting.

The 2009 NHTS dataset include information on:

- Household data. Relationship between household members, income, housing characteristics, and other socio-demographic information for each member of the household and for the head of the household;
- Information on each household vehicle, including year, make, model, and estimates of annual miles traveled;
- Data about drivers, including information on travel as part of work.
- Data about one-way trips taken during a designated 24-hour period (the household's travel day) including the time the trip began and ended, length of the trip, composition of the travel party, mode of transportation, purpose of the trip, and the specific vehicle used (if a household vehicle);
- Information to describe characteristics of the geographic areas in which the sampled household and its workplace are located;
- Data on telecommuting;
- Public perceptions about the transportation system;
- Data on Internet usage; and
- The typical number of transit, walk and bike trips made over a period longer than the 24-hour travel day.

The 2009 NHTS Data is organized into four different data files:

- Household Record;
- Vehicle Record;
- Person Record;
- Travel Day Trip Record.

4.2 Vehicle Characteristics

The NHTS data does not contain the detailed vehicle information needed for the estimation of the car type model. Vehicle characteristics are computed from the *Consumer Reports*. *Consumer Reports* contains vehicle specification data for models tested within the past 10 years; up to four model years are available and classified by performance, crash protection, fuel economy, and specifications; market value or price of each new or used car are also part of the dataset.

I collected all the vehicle specifications and price for each make, model and year, including:

- Tested Model (i.e. 2003 SR5 4-door SUV 4WD, 4.0-liter V6, 4-speed automatic (Toyota 4Runner))
- Price
- Seating (front, rear, third)
- Engine size
- Transmission (manual or automatic)
- Acceleration
- 0 to 30 mph, sec.
- 0 to 60 mph, sec.

- 45 to 65 mph, sec.
- Quarter-mile, sec
- Quarter-mile, mph
- Emergency handling
- Braking
- Braking from 60 mph dry, ft.
- Braking from 60 mph wet, ft.
- Comfort/convenience
- Ride
- Noise
- Driving position
- Seat comfort
- Shoulder room, in
- Leg room, in
- Head room, in
- Controls and display
- Interior fit and finish
- Trunk/Cargo Area
- Luggage/cargo capacity, cu. ft.
- Climate System
- Fuel Economy (MPG)
- Cruising range, mi.
- Fuel capacity, gal.
- Fuel type
- Safety (Crash and rollover tests)
- Specifications
- Length, in.

- Width, in.
- Height, in.
- Turning circle, ft.
- Curb weight, lb.
- Max. load, lb.
- Typical Towing capacity, lb.

4.3 U.S. Census TIGER/Line shapefiles

The U.S. Census TIGER/Line shapefiles contain the geographic extent and boundaries of both legal and statistical entities. The 2009 data on Census Tract level is obtained for the State of Maryland, Virginia and District of Columbia because the main data source (NHTS data) was conducted in 2009 and was geo-referenced on Census Tract level.

4.4 General Transit Feed Specification (GTFS)

The General Transit Feed Specification (GTFS) , which was originally developed by Google and Portland TriMet defines a common data format for public transportation schedules and the associated geographic information. The GTFS is an open format and it is composed of a series of text files; each file contains a particular aspect of the transit service: stops, routes, trips and other schedule data. The GTFS data for the Washington D.C. Metropolitan area is obtained from the Washington Metropolitan Area Transit Authority (WMATA). The database consists of the following files:

- Agency: contains the transit agency id, name and website.
- Stops: individual locations where vehicles pick up or drop off passengers. The data contains information on stop id, stop name, latitude and longitude and stop location.
- Transit Routes: a route is a group of trips that are displayed to riders as a single service. The data contains information of route id, route name, route type (i.e., subway, rail and bus), etc.
- Trips for each route: a trip is a sequence of two or more stops that occurs at a specific time. The data contain information on the trip id, trip name, trip head sign, and the corresponding route id and service id.
- Stop times: times that a vehicle arrives at and departs from individual stops for each trip.
- Calendar dates: specify when service starts and ends, as well as days of the week when the service is available. The data contains information on the service id and service dates.
- Shapes: rules for drawing lines on a map to represent a transit organizations routes.

The data structure of GTFS is presented in Figure 4.1.

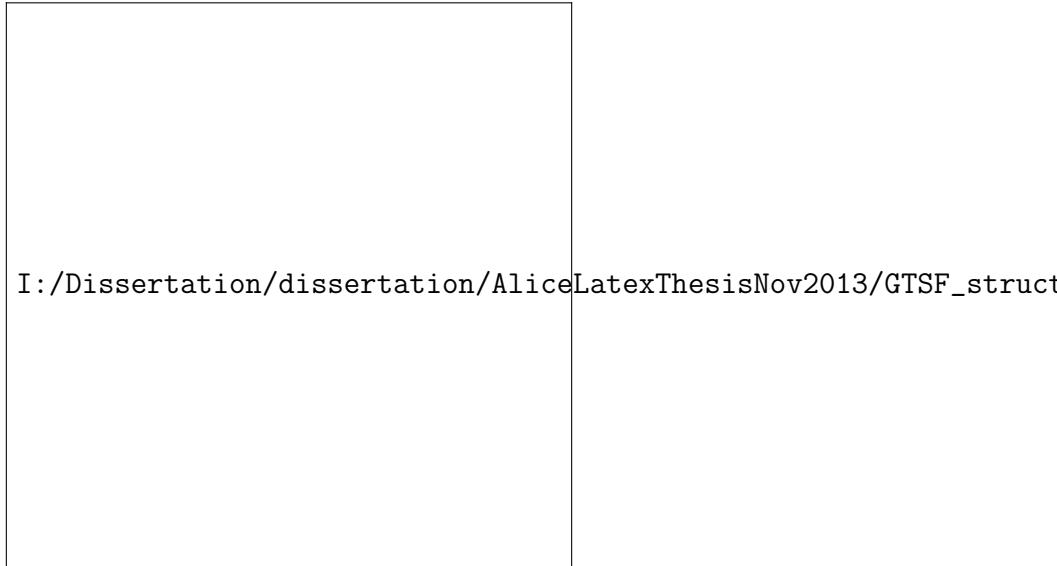


Figure 4.1: Data Structure of GTFS Data

4.5 American Community Survey (ACS)

The American Community Survey (ACS) is an ongoing statistical survey by the U.S. Census Bureau, sent to approximately 250,000 addresses monthly (or 3 million per year). It regularly gathers information previously contained only in the long form of the decennial census. It is the largest survey other than the decennial census that the Census Bureau administers.

Every 10 years since 1790, Congress has authorized the government to conduct a national census of the U.S. population, as required by the U.S. Constitution. In the twentieth century, the questions were divided between a short and long form. Only a subset of the population was required to answer the long-form questions. The most recent census consisted of a short form, which included basic questions about age, sex, race, Hispanic origin, household relationship, and owner/renter status. After the 2000 Census, the long form became the ACS and will continue to collect

long-form-type information throughout the decade. The ACS includes not only the basic short-form questions, but also detailed questions about population and housing characteristics. It is a nationwide, continuous survey designed to provide communities with reliable and timely demographic, housing, social, and economic data every year [U.S. Census Bureau, 2013].

The primary benefit of ACS is that the data are being collected and will be disseminated more frequently than the once-in-10-years decennial census Long Form data. Data users will no longer need to rely on aging snapshot estimates of population and housing characteristics. Instead, they will be able to use more recently collected data whose accuracy and relevance will not depend on how closely the analysis year conforms to the decennial census year. In addition, the increased frequency of data releases will enable data users to analyze trends over shorter time periods [Transportation Research Board, 2007].

Particularly, The American Community Survey (ACS) Public Use Microdata Sample (PUMS) files show the full range of population and housing unit responses collected on individual ACS questionnaires. The PUMS files contain records for a subsample of ACS housing units and group quarters persons, with information on the characteristics of these housing units and group quarters persons plus the people in the selected housing units.

In terms of the geo-reference information, Region, Division, State, and Public Use Microdata Areas (PUMAs) are the only geographic areas identified in the ACS PUMS. Public Use Microdata Areas (PUMAs) are non-overlapping areas that partition each state into areas containing about 100,000 residents and are the most

detailed geographic areas available in the ACS PUMS files [U.S. Census,].

Chapter 5

Comparison of Unordered and Ordered Discrete-Continuous Models

5.1 Introduction

This comparison is motivated by the fact that ordered discrete-continuous models (Fang, 2008) are relatively easier to estimate when compared to unordered model structure; however the assumption that vehicle ownership decisions are measured by a single latent variable might affect the goodness of fit of the model and its performance in model application and policy analysis.

In this chapter, I apply the unordered and ordered discrete-continuous models for the Washington D.C. Metropolitan area with the 2009 NHTS and vehicle characteristics data. I assume that the choice set of the vehicle holding sub-model includes zero, one, two, three and four or more vehicles. The types of vehicle owned by each household are categorized by classes and vintages. This classification is based on the classes proposed in the 2009 National Household Travel Survey (NHTS) and in the 2009 National Transportation Statistics (NTS); it is mainly based on vehicle size, function, and brand loyalty (domestic or imported). Therefore, each household is assumed to have a choice among 12 classes and 10 vintages; 120 alternatives are in the final choice set for vehicle type and vintage sub-model. The twelve vehicle classes are: (1) small domestic car; (2) compact domestic car; (3) mid-size domestic car; (4) large domestic car; (5) luxury domestic car; (6) small import car; (7) mid-

size import car; (8) large import car; (9) sporty car; (10) minivan/van; (11) pickup trucks; (12) SUVs. The 10 vintages are pre-1999 and 2000 through 2008.

5.2 Data Statistics

The primary data source used in this case study is the 2009 National Household Travel Survey (NHTS). The comparison analysis is restricted to the Washington D.C. Metropolitan area, for which 1,420 observations in the dataset. Household characteristics, land-use variables and information on each household vehicle, are the main variables extracted from the original dataset. Table 5.1 lists the basic statistics relative to the household sample. For the Washington D.C. Metropolitan area, the average vehicle ownership per household is 1.87 in 2009. The percentage of the households without a car is 7.28%, 26.72% own one vehicle, 43.49% own two vehicles, 17.03% own three vehicles and 5.48% own four or more vehicles. The average household income increases for household having up to two cars, but remains stable for household with 3 or 4+ cars. The number of cars in the household is highly associated with the number of adults and number of drivers in the family. About half of the households who do not have a car do not own a house. The land use variables, such as dummy of urban area, urban size, population density and housing density, greatly influence the household car ownership decisions. The households with more cars are generally located in less dense or more rural area. In the Washington D.C. metropolitan area, the average age of the household head is around 55 years old in 2009, which is somehow an indication of the aging society happening in

western countries. Households with zero or one car have older household head. The average education level in this area is college/bachelors degree; however, households without a car have much lower education level. The average annual mileage traveled by a household is around 20,000 miles per year. The mileage traveled increases accordingly with the household car ownership.

Table 5.1: Data Statistics

| <i>variables</i> | by Number of cars | | | | | Statistics for all car holding cases | | | | |
|--------------------------|-------------------|--------|--------|--------|-------|--------------------------------------|------------|---------------|-------------|-------------|
| | 0 | 1 | 2 | 3 | 4 | <i>min</i> | <i>max</i> | <i>median</i> | <i>mean</i> | <i>s.d.</i> |
| 2009 NHHS | | | | | | | | | | |
| Vehicle ownership | 7.28% | 26.72% | 43.49% | 17.03% | 5.48% | 0 | 4 | 2 | 1.87 | 0.96 |
| Hhld. Income level | 6.47 | 10.98 | 14.55 | 15.29 | 15.99 | 1 | 18 | 16 | 13.21 | 5.30 |
| Num. of adult | 1.31 | 1.40 | 1.99 | 2.21 | 2.76 | 1 | 5 | 2 | 1.86 | 0.66 |
| Num. of workers | 0.53 | 0.69 | 1.16 | 1.44 | 1.63 | 0 | 4 | 1 | 1.06 | 0.83 |
| Num. of drivers | 0.75 | 1.26 | 1.98 | 2.32 | 2.83 | 0 | 5 | 2 | 1.80 | 0.77 |
| Owned house | 0.45 | 0.75 | 0.91 | 0.97 | 0.98 | 0 | 1 | 1 | 0.85 | 0.36 |
| Urban area | 0.94 | 0.84 | 0.75 | 0.60 | 0.56 | 0 | 1 | 1 | 0.75 | 0.43 |
| Urban size | 4.94 | 4.25 | 3.60 | 2.95 | 2.55 | 1 | 6 | 5 | 3.70 | 2.29 |
| Use of PT | 0.26 | 0.08 | 0.06 | 0.05 | 0.06 | 0 | 1 | 0 | 0.08 | 0.27 |
| Age of hhld head | 59.13 | 60.51 | 53.12 | 52.16 | 53.00 | 18 | 95 | 54 | 55.36 | 14.91 |
| Female hhld head | 0.78 | 0.64 | 0.52 | 0.55 | 0.44 | 0 | 1 | 1 | 0.57 | 0.49 |
| Educ. of hhld head | 2.78 | 3.40 | 3.67 | 3.52 | 3.33 | 1 | 5 | 4 | 3.49 | 1.21 |
| Housing unit per sq mile | 7233 | 3637 | 1341 | 797 | 626 | 50 | 30000 | 750 | 2252 | 4220 |
| Percent renter-occupied | 50 | 32 | 22 | 18 | 17 | 0 | 95 | 20 | 26 | 21 |
| Population per sq mile | 12153 | 6931 | 3408 | 2363 | 1888 | 50 | 30000 | 3000 | 4725 | 6333 |
| Workers per sq mile | 2870 | 1899 | 1042 | 640 | 453 | 25 | 5000 | 350 | 1303 | 1660 |
| AMT | 0 | 10361 | 23890 | 35781 | 48728 | 0 | 91329 | 19097 | 21554 | 16381 |

In terms of vehicle class and vintage (Figure 5.1 and Figure 5.2), for the households with only one car, about half of the vehicles are imported cars. Households with more than one car own more vans. There are much more pickup trucks for the households with three or more cars. The average age of the cars in the study area is 8.6 years old, the majority of the cars are between 4 and 10 years old. The households with more cars tend to hold older cars in average, since the average vehicle age in 1-car and 2-car households is around 8 to 8.5 and more than 10 years old for the 4+ car households.

The 2009 NHTS data does not include information on vehicle price, fuel efficiency, seating, engine, and other vehicle characteristics by vehicle make and model, which are important attributes for the analysis of factors associated to vehicle type decisions. The vehicle characteristic data were obtained from the *ConsumerReports*. *ConsumerReports* provide the vehicle specification data on models tested within the past 10 years, having up to four model years by performance, crash protection, fuel economy, and specifications. *ConsumerReports* also indicates the sale price or the price of each new or used car. Then we aggregated all the information we collected by 12 vehicle classes and 10 vintages. Therefore, there are totally 120 alternatives (12 classes * 10 vintages), with detailed and aggregated vehicle specification and price information.

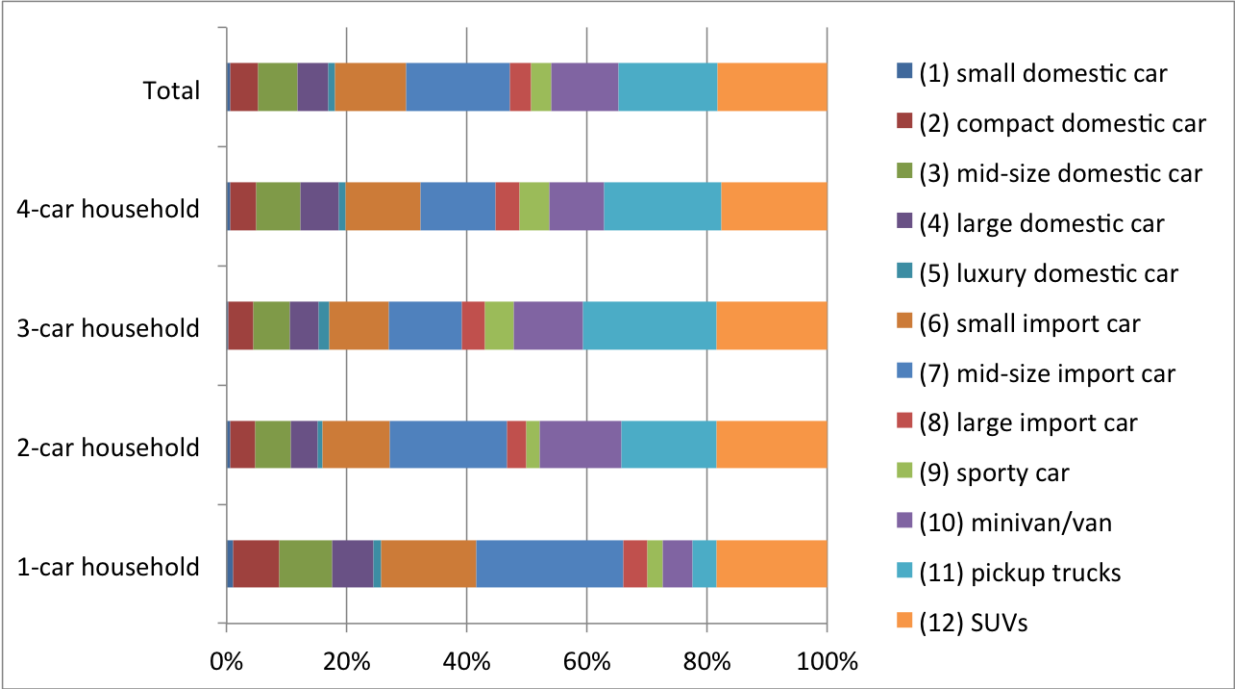


Figure 5.1: Distribution of vehicle classes

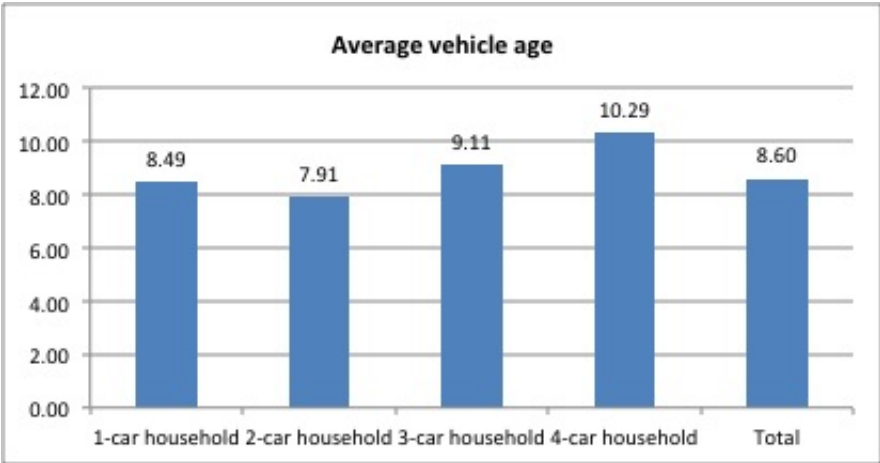


Figure 5.2: Vehicle age profile

5.3 Calibration of the Logsum

We first calibrate a multinomial logit model for the vehicle type submodel. The number of alternatives for the vehicle type choice increases exponentially with the number of vehicles in the household, for example, a family which has three cars would have 120^3 choices in total. Because of the large number of alternatives, estimation of this model on the full set of alternatives is considered infeasible. We take the advantage of IIA property of multinomial logit model. The vehicle type sub-model is then estimated on a subset of alternatives which includes the households chosen alternative and 20 alternatives randomly selected from the 120 alternatives. Tests ([Train, 1986]) indicate that, once the number of alternatives exceeds a minimal threshold, the estimated parameters are not sensitive to the number of alternatives included in the choice set. Results from the class/vintage sub-model are reported in Table 5.2.

The vehicle holding and vehicle type sub-models are then linked using a logsum variable derived from the calibration of the multinomial logit for the vehicle type/vintage decisions.

Table 5.2: Estimation results of the vehicle type choice sub-model

| Variable | One-Car Household | | Two-Car Household | | Three-Car Household | | Four-Car Household | |
|----------------------------------------|-------------------|--------|-------------------|--------|---------------------|--------|--------------------|--------|
| | Estimate | t-stat | Estimate | t-stat | Estimate | t-stat | Estimate | t-stat |
| Sum of shoulder room | 0.0329 | 5.6 | 0.04837 | 19.9 | 0.0471 | 18.4 | 0.0678 | 9.5 |
| Sum of luggage space | 0.2162 | 7.3 | 0.1563 | 5.8 | - | - | 0.1032 | 4.70 |
| Log(n. of make/model in the class) | 1.382 | 28.4 | 0.8015 | 40.1 | 0.8241 | 38.5 | 1.182 | 16.3 |
| Auto's MPG | 0.0026 | 0.3 | - | - | - | - | - | - |
| Difference of max MPG and min-MPG | - | - | 0.0158 | 3.0 | 0.02465 | 3.1 | 0.02202 | 1.1 |
| D. at least one foreign car | -0.4959 | -7.6 | - | - | - | - | -0.3361 | -2.4 |
| D. both foreign cars | - | - | 1.031 | 14.0 | - | - | - | - |
| D. one old vehicle | 0.4515 | 8.3 | 0.4133 | 10.2 | 1.094 | 10.9 | 1.144 | 3.5 |
| D. at least one SUV (Hhld 3 memb) | 2.087 | 8.2 | 0.2794 | 3.0 | - | - | -0.7771 | -3.6 |
| D. at least one pickup (Hhld 3 memb) | -1.332 | -1.3 | 0.293 | 3.0 | 0.6034 | 5.1 | 0.5366 | 2.6 |
| D. at least one van (Hhld 3 memb) | - | - | 1.152 | 12.7 | 0.3519 | 3.0 | -0.3945 | -1.9 |
| D. auto (Hhld 3 memb) | 1.1 | 4.7 | - | - | - | - | - | - |
| D. of both autos | - | - | -1.333 | -4.7 | - | - | - | - |
| D. of at least one sporty car | - | - | - | - | - | - | 0.4704 | 2.1 |
| Purchase price (Hhld income <20k) | -0.2037 | -15.2 | - | - | - | - | - | - |
| Purchase price (Hhld income = 20k-40k) | -0.0683 | -8.5 | - | - | - | - | - | - |
| Purchase price (Hhld income > 40k) | -0.0069 | -1.2 | - | - | - | - | - | - |

| Variable | One-Car Household | | Two-Car Household | | Three-Car Household | | Four-Car Household | |
|-----------------------------------------|-------------------|----------|-------------------|----------|---------------------|----------|--------------------|----------|
| | Estimate | t-stat | Estimate | t-stat | Estimate | t-stat | Estimate | t-stat |
| Purchase price (Hhld income < 45k) | - | - | -0.157 | -28.2 | - | - | - | - |
| Purchase price (Hhld income = 45k-80k) | - | - | -0.0730 | -17.8 | - | - | - | - |
| Purchase price (Hhld income > 80k) | - | - | -0.0457 | -7.7 | - | - | - | - |
| Purchase price (Hhld income < 55k) | - | - | - | - | -0.1688 | -22.8 | - | - |
| Purchase price (Hhld income = 55k-100k) | - | - | - | - | -0.0953 | -18.1 | - | - |
| Purchase price (Hhld income > 100k) | - | - | - | - | -0.0289 | -6.1 | - | - |
| Purchase price (Hhld income < 60k) | - | - | - | - | - | - | -0.1841 | -12.5 |
| Purchase price (Hhld income = 60k-100k) | - | - | - | - | - | - | -0.1175 | -11.2 |
| Purchase price (Hhld income > 100k) | - | - | - | - | - | - | -0.0678 | -8.1 |
| Number of observations | | 2995 | | 4525 | | 2015 | | 545 |
| Initial Likelihood | | -9118.34 | | - | | -6134.71 | | -1659.26 |
| Final Likelihood | | -7859.95 | | 13776.46 | | -3755.26 | | -788.66 |
| Rho-Squared | | 0.138 | | 0.2327 | | 0.3879 | | 0.5247 |

Generally households prefer vehicles with more shoulder room and bigger luggage space; moreover, they are more likely to own a car type for which more choices (make and model combinations) are available. The variable of difference of MPGs is a proxy to test whether the households prefer cars with similar engine size or not. The positive coefficient indicates that households with multiple cars have higher tendency to own cars with different horsepower. However, when it comes to the households with four or more cars, this factor becomes less significant. Households with only one car do not prefer foreign cars, while two-car households prefer both domestic cars. Households are in general holding older vehicles. Households are more likely to own only one car if there are less than three members in the family, whereas households with more than three members prefer to own SUVs. For the two-car households, the ones with three or more household members prefer to own one SUV/pickup/van rather than two autos. Similarly, the households with three cars are more likely to own a pickup or a van. However, households with four or more cars tend to own at least one pickup, but not SUVs or vans; they also have higher tendency of owning a sporty car. The coefficients related to vehicle purchase price are negative and significant; their magnitude is decreasing with the increase of household income. The lower income group is more sensitive to the vehicle purchase price, while higher income group are found to be less sensitive to vehicle price (as expected). The logsum of the class/vintage submodel is then calculated and included into the discrete continuous model.

5.4 Estimation Results and Comparison

The unordered discrete-continuous (UDC) model with both simulation and numerical computation and the ordered discrete-continuous (ODC) model have been estimated with the 2009 NHTS data. Estimation results of the three models are presented in Table 5.3.

All the estimated coefficients are significant and have the expected sign, with only a few exceptions. Positive coefficients of household income indicate that households with higher income have higher tendency to own more vehicles and drive more. The negative coefficient in the one-car household alternative means high-income group are less likely to own only one car. In the unordered models, the magnitude increases as the number of vehicles in the household increases. The coefficients of number of drivers in the household are very significant, indicating that this factor has high effects on how many cars a household owns. This coefficient is positive in the ordered structure, and also positive in the unordered structure with an exception for the one-car households. The negative coefficient for one-car household alternative indicates that, the more drivers in the household, the less likely they own only one car. Similarly, households with female household head are less likely to own more cars.

Urban size is the size of urban area in which home address is located, in which the lower value represents urban areas and the higher value represents rural areas. Residential density is an indication of the built environment around the household location. The coefficients of these two variables are significantly negative (with a

few exceptions in the one-car household alternative) and have higher magnitude as the households own more cars in the unordered structure. Both of the coefficients in the ordered and unordered structures infer that the households located in highly residential areas are more likely to own fewer cars and drive less while the households located in a more rural area have higher probability of having more cars and drive more.

The driving cost is measured by dollars per mile. As expected, the coefficient of driving cost is significant and negative, indicating that higher driving cost induces the households to drive less.

Table 5.3: Estimation results of the integrated discrete-continuous models

| Variable | Model 1 | | Model 2 | | Model 3 | | Model 4 | |
|------------------------------------|----------------|--------|----------------|--------|---------------|--------|----------------|--------|
| | Coefficient | t-stat | Coefficient | t-stat | Coefficient | t-stat | Coefficient | t-stat |
| Dependent variable: Number of cars | | | | | | | | |
| logsum | 0.388 | 0.012 | 0.514 | 0.018 | 0.363 | 0.111 | | |
| constant | | | | | | | | |
| 1 car | 2.863 | 0.237 | -2.948 | 0.217 | | | 1.368 | 0.096 |
| 2 cars | -8.700 | 0.098 | -21.889 | 0.284 | | | -4.432 | 1.039 |
| 3 cars | -14.404 | 0.188 | -28.931 | 0.257 | | | -4.918 | 0.113 |
| 4+ cars | -21.385 | 0.201 | -35.658 | 0.245 | | | -11.288 | 0.114 |
| income | | | | | | | | |
| 1 car | -0.051 | 0.011 | -0.101 | 0.020 | 0.086 | 0.006 | 0.151 | 0.010 |
| 2 cars | 0.056 | 0.006 | 0.692 | 0.072 | | | 0.395 | 0.029 |
| 3 cars | 0.105 | 0.010 | 0.745 | 0.077 | | | 0.429 | 0.028 |
| 4+ cars | 0.111 | 0.012 | 0.693 | 0.074 | | | 0.327 | 0.026 |
| num. of drivers | | | | | | | | |
| 1 car | -0.010 | 0.007 | -0.304 | 0.236 | 0.608 | 0.057 | -0.048 | 0.027 |
| 2 cars | 3.223 | 0.079 | 9.236 | 0.214 | | | 2.111 | 0.215 |
| 3 cars | 4.041 | 0.102 | 10.167 | 0.190 | | | 1.742 | 0.086 |
| 4+ cars | 4.432 | 0.092 | 10.120 | 0.165 | | | 3.314 | 0.142 |
| gender (female) | | | | | | | | |
| 1 car | -0.129 | 0.551 | -0.063 | 0.249 | -0.235 | 0.063 | -0.054 | 0.068 |
| 2 cars | -0.874 | 0.054 | -3.434 | 0.213 | | | -0.732 | 0.245 |
| 3 cars | -0.928 | 0.073 | -3.605 | 0.211 | | | -0.854 | 0.281 |
| 4+ cars | -0.885 | 0.059 | -3.667 | 0.194 | | | -2.208 | 0.360 |
| urban size | | | | | | | | |
| 1 car | 0.077 | 0.035 | -0.109 | 0.058 | -0.032 | 0.013 | -0.013 | 0.028 |
| 2 cars | -0.120 | 0.074 | -0.270 | 0.178 | | | 0.103 | 0.277 |
| 3 cars | -0.199 | 0.093 | -0.354 | 0.186 | | | -0.038 | 0.018 |
| 4+ cars | -0.201 | 0.084 | -0.406 | 0.183 | | | -0.368 | 0.063 |

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---------------------------------------------|---------------|---------------------------|-----------------------------|---------------|
| Variable | Coefficient | Coefficient | Coefficient | Coefficient |
| | std. err | std. err | std. err | std. err |
| res. density | | | | |
| 1 car | 0.041 | 0.101 | -0.103 | -0.168 |
| | 0.005 | 0.015 | 0.010 | 0.017 |
| 2 cars | -0.223 | -1.112 | | -0.472 |
| | 0.034 | 0.159 | | 0.036 |
| 3 cars | -0.442 | -1.298 | | -0.740 |
| | 0.054 | 0.181 | | 0.070 |
| 4+ cars | -0.484 | -1.262 | | -0.599 |
| | 0.064 | 0.170 | | 0.183 |
| α_1 | | | 1.580 | |
| α_2 | | | 3.149 | |
| α_3 | | | 4.201 | |
| Dependent variable: AMT (10k) | | | | |
| constant | 1.130 | 1.385 | 1.473 | 1.456 |
| | 0.102 | 0.116 | 0.105 | 0.068 |
| income | 0.129 | 0.128 | 0.132 | 0.127 |
| | 0.005 | 0.007 | 0.006 | 0.006 |
| own home | 0.671 | 0.328 | 0.258 | 0.296 |
| | 0.277 | 0.098 | 0.072 | 0.060 |
| gender (female) | -0.056 | -0.095 | -0.080 | -0.035 |
| | 0.034 | 0.061 | 0.059 | 0.013 |
| res. density | -0.113 | -0.118 | -0.120 | -0.117 |
| | 0.008 | 0.009 | 0.011 | 0.006 |
| driving cost (\$ per mile) | -5.103 | -4.670 | -5.133 | -4.967 |
| | 0.283 | 0.285 | 0.238 | 0.098 |
| Log-likelihood at zero | -9583.87 | -9583.87 | -9583.87 | -9583.87 |
| Log-likelihood at convergence | -3349.812 | -3288.934 | -3607.747 | -3472.514 |
| Number of parameters | 25 | 25 | 10 | 24 |
| Number of observations | 1420 | 1420 | 1420 | 1420 |
| Adjusted ρ^2 | 0.648 | 0.654 | 0.623 | 0.635 |
| $-2(LL(\hat{\beta}^4) - LL(\hat{\beta}^2))$ | | $367.16 > \chi_{25,0.01}$ | | |
| Non-nested test result | | | $\Phi(15.98) = 1.34e^{-56}$ | |

**Note: Model 1 is the unordered discrete-continuous model with simulation; Model 2 is the unordered discrete-continuous model with numerical computation; Model 3 is the ordered discrete-continuous model; Model 4 is the same as Model 2 except excluding the "logsum" variable, which make it comparable to Model 3. *Note: Variables that are significant at 95% level or above are bold.*

The covariance matrices of the models are reported below. In the unordered discrete-continuous models, the bottom line of the matrix explains the correlation between the mileage traveled and the utilities of the vehicle holding alternatives. The positive values mean higher demand on mileage usage increase the utility of owning two or more cars in the household. In the ordered discrete-continuous models, the correlation between the number of vehicles and mileage traveled is 0.5, which means that the demand of vehicle usage increase the propensity of owning more cars.

$$\hat{\Sigma}_1 = \begin{pmatrix} 2.00 & -1.14 & -1.31 & -1.30 & -0.27 \\ -1.14 & 1.63 & 0.37 & 0.76 & 0.10 \\ -1.31 & 0.37 & 2.37 & 1.68 & 0.67 \\ -1.30 & 0.76 & 1.68 & 1.36 & 0.46 \\ \mathbf{-0.27} & \mathbf{0.10} & \mathbf{0.67} & \mathbf{0.46} & \mathbf{1.23} \end{pmatrix}$$

$$\hat{\Sigma}_2 = \begin{pmatrix} 2.00 & -10.34 & -10.24 & -10.57 & -0.73 \\ -10.34 & 58.26 & 61.44 & 61.57 & 4.46 \\ -10.24 & 61.44 & 68.64 & 67.11 & 5.21 \\ -10.57 & 61.57 & 67.11 & 66.34 & 5.00 \\ \mathbf{-0.73} & \mathbf{4.46} & \mathbf{5.21} & \mathbf{5.00} & 1.25 \end{pmatrix}$$

$$\hat{\Sigma}_3 = \begin{pmatrix} 1.00 & 0.50 \\ \mathbf{0.50} & 1.56 \end{pmatrix}$$

$$\hat{\Sigma}_4 = \begin{pmatrix} 2.00 & 3.31 & 3.95 & 3.43 & 1.48 \\ 3.31 & 12.89 & 5.69 & 4.64 & 2.38 \\ 3.95 & 5.69 & 11.67 & 12.19 & 3.43 \\ 3.43 & 4.64 & 12.19 & 36.93 & 4.56 \\ \mathbf{1.48} & \mathbf{2.38} & \mathbf{3.43} & \mathbf{4.56} & \mathbf{1.24} \end{pmatrix}$$

5.5 Application Results and Comparison

The models estimated have been applied to test policy scenarios; the variables of interest are density and driving cost. The following three scenarios have been

tested:

- Household income: 10% decrease, 5% decrease, 5% increase and 10% increase
- Residential density: 50% decrease, 25% decrease, 25% increase and 50% increase
- Driving cost: 50% decrease, 25% decrease, 25% increase and 50% increase

Results in Table 5.4 and Table 5.5 show the effects of those variables on both vehicle holding and mileage traveled. It appears that results are consistent between ordered and unordered structures except for the "household income" scenarios. There are slight effects on vehicle holding changes with respect to all the scenarios, except that a 10% change of household income level in ordered discrete continuous model will result in up to 4.23% change in vehicle holding stock. Changes in fuel cost have great effects in increasing/reducing vehicle usage. For example, vehicle usage will be reduced by around 17% - 20% when the driving cost is increased by 50%. However, even when the density increased by 50% people only cut less than 6% in their car use. The observations from the "density" scenarios are consistent with the findings in the previous studies (i.e., [Fang, 2008]).

Table 5.4: Application results from the unordered discrete continuous model

| | 0-car hh | 1-car hh | 2-car hh | 3-car hh | 4-car hh | average vehicle ownership | miles |
|----------------|----------|----------|----------|----------|----------|---------------------------|-------------------|
| Actual | 7.22% | 22.59% | 46.82% | 17.91% | 5.45% | 1.92 | 22,490.65 |
| Income -10% | 7.20% | 24.52% | 45.93% | 16.92% | 5.44% | 1.89 | 20,829.10 -7.39% |
| Income -5% | 7.22% | 23.49% | 46.39% | 17.46% | 5.45% | 1.90 | 21,666.00 -3.67% |
| Income +5% | 7.22% | 21.75% | 47.18% | 18.39% | 5.46% | 1.93 | 23,310.70 3.65% |
| Income +10% | 7.25% | 20.70% | 47.71% | 18.88% | 5.47% | 1.95 | 24,151.20 7.38% |
| Density -50% | 7.24% | 20.10% | 47.99% | 19.20% | 5.46% | 1.96 | 23,730.66 5.51% |
| Density -25% | 7.25% | 21.38% | 47.42% | 18.49% | 5.46% | 1.94 | 23,080.49 2.62% |
| Density +25% | 7.12% | 23.79% | 46.21% | 17.43% | 5.45% | 1.90 | 21,850.73 -2.85% |
| Density +50% | 6.97% | 24.91% | 45.82% | 16.86% | 5.45% | 1.89 | 21,231.07 -5.60% |
| Fuel cost -50% | 7.22% | 22.58% | 46.86% | 17.89% | 5.45% | 1.92 | 26,479.62 17.74% |
| Fuel cost -25% | 7.23% | 22.54% | 46.81% | 17.95% | 5.47% | 1.92 | 24,501.11 8.94% |
| Fuel cost +25% | 7.23% | 22.56% | 46.74% | 18.02% | 5.45% | 1.92 | 20,488.95 -8.90% |
| Fuel cost +50% | 7.22% | 22.54% | 46.84% | 17.94% | 5.46% | 1.92 | 18,501.26 -17.74% |

Table 5.5: Application results from the ordered discrete continuous model

| | 0-car hh | 1-car hh | 2-car hh | 3-car hh | 4-car hh | average vehicle ownership | miles |
|----------------|----------|----------|----------|----------|----------|---------------------------|-------------------|
| Actual | 7.05% | 25.17% | 43.79% | 17.97% | 6.03% | 1.91 | 22,552.90 |
| Income -10% | 7.66% | 27.50% | 44.10% | 15.98% | 4.76% | 1.83 | 20,802.00 -7.76% |
| Income -5% | 7.36% | 26.23% | 44.04% | 17.01% | 5.36% | 1.87 | 21,709.60 -3.74% |
| Income +5% | 6.77% | 24.09% | 43.58% | 18.83% | 6.73% | 1.95 | 23,426.30 3.87% |
| Income +10% | 6.50% | 23.21% | 43.02% | 19.78% | 7.50% | 1.99 | 24,333.50 7.90% |
| Density -50% | 5.03% | 24.53% | 45.05% | 18.88% | 6.51% | 1.97 | 23,857.50 5.78% |
| Density -25% | 6.00% | 24.92% | 44.39% | 18.40% | 6.29% | 1.94 | 23,214.90 2.94% |
| Density +25% | 8.11% | 25.37% | 43.28% | 17.43% | 5.80% | 1.87 | 21,904.60 -2.87% |
| Density +50% | 9.07% | 25.49% | 42.78% | 17.02% | 5.64% | 1.85 | 21,278.60 -5.65% |
| Fuel cost -50% | 7.02% | 25.22% | 43.85% | 17.89% | 6.02% | 1.91 | 27,030.40 19.85% |
| Fuel cost -25% | 7.01% | 25.23% | 43.85% | 17.87% | 6.04% | 1.91 | 24,801.60 9.97% |
| Fuel cost +25% | 7.03% | 25.29% | 43.78% | 17.87% | 6.03% | 1.91 | 20,322.00 -9.89% |
| Fuel cost +50% | 7.05% | 25.23% | 43.83% | 17.94% | 5.96% | 1.91 | 18,087.80 -19.80% |

5.6 Chapter Summary

In this chapter, both ordered and unordered models are applied to estimate a joint model of household vehicle holding and mileage traveled on data extracted from the 2009 NHTS and representative of the Washington metropolitan area. Variables related to household characteristics, land use and driving cost have been estimated for both datasets and result to significantly affect decisions regarding the number of cars in the household and annual mileage driven. Although coefficients are not directly comparable, the results from the model application to policy testing show that both density and driving costs do not affect much the vehicle holding under analysis. Changes in driving costs only marginally affect the number of cars but greatly affect the AMT in households residing in the Washington metropolitan area.

In terms of the methodological comparison, the advantage of the ordered structure over the unordered is that it offers a closed mathematical form for the choice probabilities and does not require simulations for the estimation, that are proven to be quite difficult in probit model calibration. However, the unordered discrete-continuous models always performs better in terms of goodness of fit statistics when compared to ordered discrete-continuous models, which is consistent to previous results obtained in the literature and related to vehicle holding decisions. Finally, although the superiority of discrete-continuous unordered probit over ordered probit might be case specific, this analysis confirms once again that the unordered structure is better suited for vehicle holding and use decisions even in the context of joint discrete-continuous decisions.

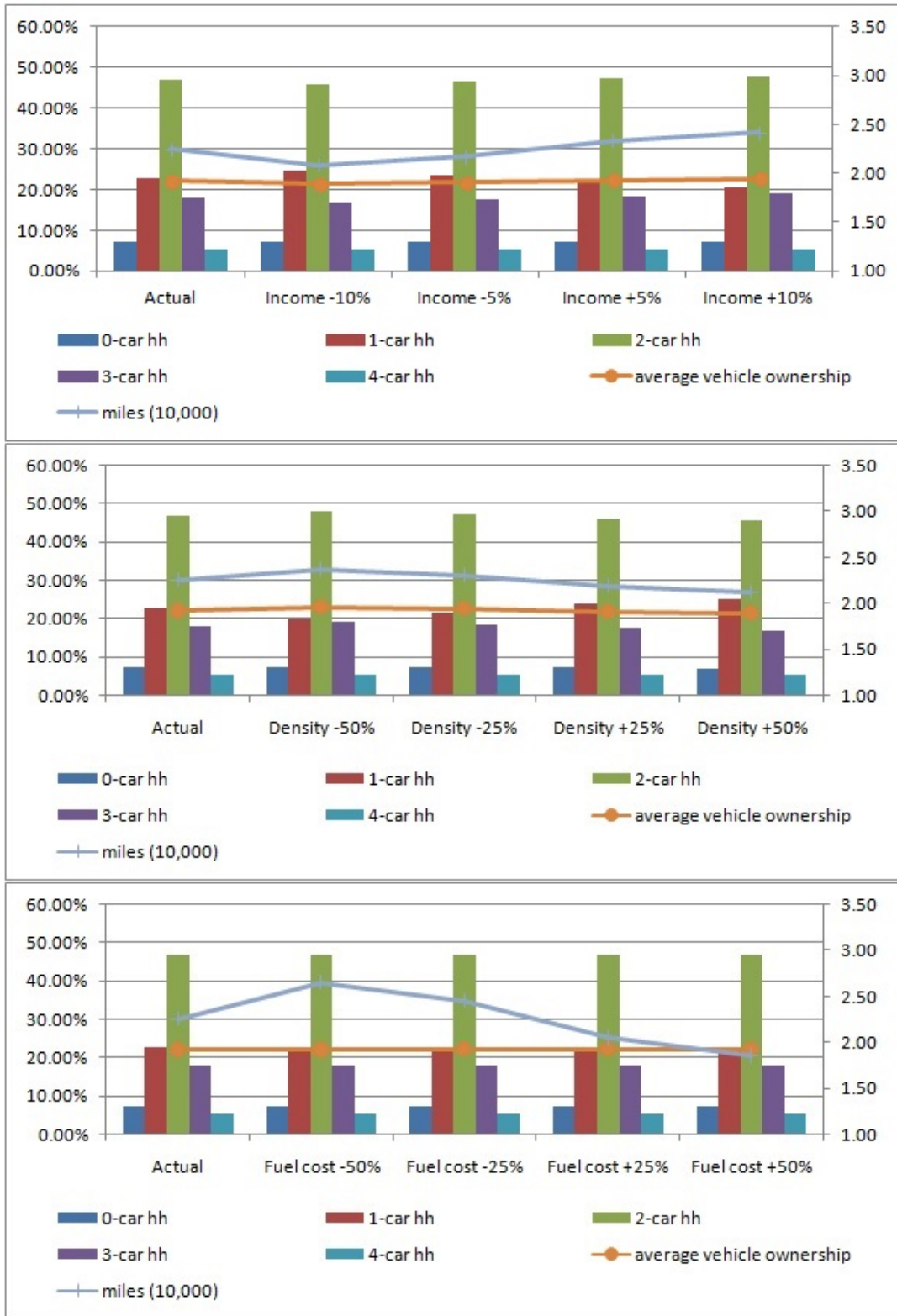


Figure 5.3: Application results from the unordered discrete continuous model

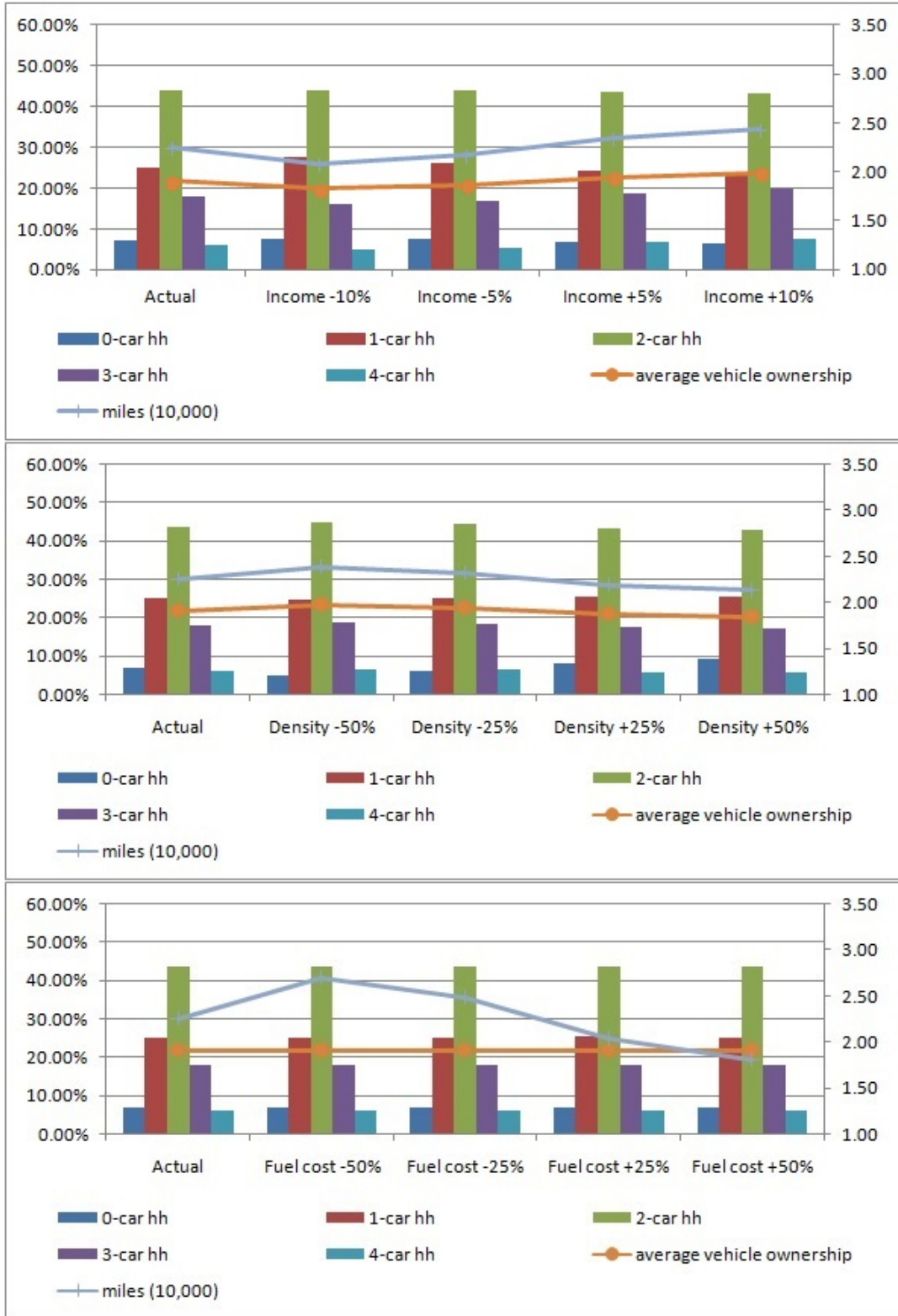


Figure 5.4: Application results from the ordered discrete continuous model

Chapter 6

National Model of Vehicle Ownership and Usage

6.1 Introduction

In the literature, there are a large number of studies that have developed vehicle ownership models for large cities and metropolitan areas (See Chapter 2). The majority of them are based on household travel survey data. However, very few studies conducted such research for the entire U.S., especially in more recent years. The barriers include the difficulties to capture demand levels for different population segments across the nation, and the poor data sources for small cities/areas.

This chapter develops a series of vehicle ownership and usage models for the entire United States, which is motivated by the lack of national vehicle ownership models in the literature, and the needs to determine vehicle/driving demand in small areas with limited data availability.

The models are estimated for four Census Regions (Northeast, Midwest, South and West; Figure 6.1) and 3 area types (urbanized area, urban clusters and rural). The categories are selected according to the U.S. Census definitions. Then the models are applied to small areas using ACS PUMS data (2009 1-year estimate), which is a new data source from the U.S. Census Bureau and was firstly implemented in 2005. The idea developed in this chapter is inspired by the most recent studies on model/data transferability and by the need to integrate different sources of data for

transportation analysis. For example, Hu et al. (2007) combined the 2001 NHTS data and 2000 census data to provide estimates of regional or local travel, including vehicle trips (VT), vehicle miles of travel (VMT), person trips (PT), and person miles of travel (PMT) by trip purpose and a number of demographics ([Hu et al., 2007]).

The NHTS data contains a wealth of nation's daily travel information, however, it is not as rich as ACS data in terms of the sample size. The NHTS is only conducted every 5-7 years whereas ACS is collected continuously. In fact, the NHTS data were not recommended for analysis of categories smaller than the combination of Census division, MSA size, and the availability of rail. In addition, some metropolitan areas conduct their own household travel surveys, but many lack the necessary resources to collect local data.

In this analysis, the entire NHTS data set and model estimations are performed for 12 groups, composed by 4 Census Regions (Northeast, Midwest, South and West; Figure 6.1) and 3 area types (urbanized area, urban clusters and rural). As defined by the U.S. Census Bureau, urban areas are contiguous census block groups with a population density of at least 1,000 /sq mi with any census block groups around this core having a density of at least 500 /sq mi. Urban areas are delineated without regard to political boundaries. The census has two distinct categories of urban areas. Urbanized Areas have populations greater than 50,000, while Urban Clusters have populations of less than 50,000 but more than 2,500. An urbanized area may serve as the core of a metropolitan statistical area, while an urban cluster may be the core of a micropolitan statistical area ([U.S. Census,]).

In the NHTS data, Region, Division, State and the area type indicators are derived from the household's home address (confidential) and the U.S. Census boundary files. On the other side, Region, Division, State, and Public Use Microdata Areas (PUMAs) are the geographic areas identified in the ACS PUMS files.



Figure 6.1: United States Regions (Census Bureau)

6.2 Estimation Results with the NHTS Data

With the 2009 NHTS data, twelve discrete-continuous models for household vehicle ownership and usage are estimated (combination of 4 Regions and 3 area types), and the results are shown in Table 6.1. The explanatory variables in the final model specification are: household annual income level, household size, number



I:/Dissertation/dissertation/AliceLatexThesisNov2013/us_urban.jpg

Figure 6.2: United States Urban Area (Census Bureau)

of workers in the household, dummy of having child(ren), dummy of owned home, residential density, and driving cost (\$/mile), which are common variables (except density and cost) between NHTS and ACS data. Almost all of the coefficients are significant at 95% level and have the expected signs, with only a few exceptions. Although the magnitudes cannot be compared directly, it still can be seen that there are diversities among different Regions and area types. The models are then applied and validated; results are presented in the next section.

Table 6.1: Estimation Results of National Models

| | Northeast | | Midwest | | South | | West | |
|------------------------------------|-----------|---------|---------|---------|---------|----------|---------|---------|
| | Urban | Rural | Urban | Rural | Urban | Rural | Urban | Rural |
| Dependent variable: Number of cars | | | | | | | | |
| constant | | | | | | | | |
| 1 car | -0.103 | 0.309 | -0.003 | 0.886 | -0.284 | -0.172 | 0.313 | 1.014 |
| 2 cars | -21.336 | -19.78 | -15.918 | -5.157 | -23.602 | -37.925 | -16.584 | -3.28 |
| 3 cars | -49.132 | -23.278 | -26.848 | -26.475 | -51.843 | -56.695 | -18.601 | -11.177 |
| 4+ cars | -46.506 | -76.713 | -38.732 | -20.739 | -53.254 | -129.845 | -36.641 | -15.186 |
| income | | | | | | | | |
| 1 car | 0.086 | 0.08 | 0.074 | 0.031 | 0.103 | 0.113 | 0.065 | 0.089 |
| 2 cars | 0.904 | 0.684 | 0.811 | 0.274 | 1.097 | 1.608 | 0.729 | 0.279 |
| 3 cars | 1.429 | 0.747 | 0.829 | 0.516 | 1.619 | 1.906 | 0.749 | 0.401 |
| 4+ cars | 1.03 | 1.125 | 0.801 | 0.434 | 1.668 | 2.737 | 0.885 | 0.37 |
| num. of hh members | | | | | | | | |
| 1 car | 0.315 | 0.202 | 0.286 | -0.105 | 0.257 | 0.366 | 0.154 | 0.121 |
| 2 cars | 3.376 | 4.528 | 3.289 | 1.387 | 3.88 | 7.211 | 3.31 | 0.897 |
| 3 cars | 5.351 | 4.618 | 4.152 | 2.245 | 5.081 | 7.786 | 3.528 | 1.363 |
| 4+ cars | 4.512 | 7.659 | 3.181 | 2.173 | 5.888 | 14.037 | 4.251 | 1.446 |
| num. of workers | | | | | | | | |
| 1 car | 0.504 | 0.615 | 0.592 | 1.027 | 0.665 | 0.526 | 0.094 | 0.789 |
| 2 cars | 2.367 | 4.457 | 3.287 | 2.464 | 3.758 | 5.908 | 2.372 | 1.823 |
| 3 cars | 5.745 | 5.008 | 4.7 | 6.643 | 7.132 | 8.617 | 2.595 | 3.037 |
| 4+ cars | 4.025 | 9.631 | 9.813 | 5.982 | 7.622 | 11.447 | 4.748 | 3.339 |
| own home | | | | | | | | |
| 1 car | 0.398 | 0.64 | 0.543 | 1.117 | 0.888 | 0.603 | 0.406 | 0.579 |

| | | | | | | | | | | | | |
|-------------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 2 cars | 6.946 | 5.597 | 7.456 | 5.189 | 3.956 | 3.956 | 7.071 | 12.591 | 5.136 | 5.146 | 2.929 | 2.357 |
| 3 cars | 6.599 | 11.192 | 8.067 | 7.763 | 3.614 | 10.074 | 10.235 | 17.372 | 5.926 | 5.54 | 5.116 | 4.222 |
| 4+ cars | 11.289 | 3.433 | 18.069 | 0.575 | 3.188 | 7.983 | 8.214 | 30.763 | 5.357 | 8.587 | 10.486 | 6.382 |
| res. Density (1,000) | | | | | | | | | | | | |
| 1 car | -0.063 | -0.16 | -0.197 | -0.059 | -0.089 | 0.38 | -0.084 | -0.179 | 0.275 | -0.049 | -0.239 | -0.322 |
| 2 cars | -0.62 | -0.721 | -1.424 | -1.268 | -0.626 | -2.718 | -0.594 | -2.574 | -0.991 | -0.589 | -0.969 | -0.798 |
| 3 cars | -1.171 | -2.948 | -1.477 | -1.305 | -0.843 | -7.747 | -1.194 | -4.358 | -3.758 | -0.607 | -2.62 | -1.99 |
| 4+ cars | -1.334 | -0.189 | -1.974 | -2.203 | -1.036 | -5.415 | -1.382 | -4.055 | -4.981 | -0.858 | -4.991 | -2.567 |
| Dependent variable: AMT (10k) | | | | | | | | | | | | |
| constant | 0.077 | 0.136 | 0.759 | 0.342 | 0.702 | 1.319 | 0.162 | 0.354 | 0.212 | 0.216 | 0.625 | 1.328 |
| income | 0.062 | 0.072 | 0.055 | 0.06 | 0.099 | 0.051 | 0.071 | 0.087 | 0.071 | 0.069 | 0.07 | 0.054 |
| num. of hh members | 0.252 | 0.255 | 0.22 | 0.166 | 0.221 | 0.154 | 0.271 | 0.23 | 0.239 | 0.387 | 0.301 | 0.183 |
| num. of workers | 0.346 | 0.427 | 0.507 | 0.488 | 0.452 | 0.627 | 0.503 | 0.501 | 0.645 | 0.425 | 0.451 | 0.533 |
| own home | 0.218 | 0.209 | 0.428 | 0.452 | 0.468 | 0.691 | 0.566 | 0.397 | 0.588 | 0.467 | 0.409 | 0.384 |
| has child(ren) | -0.004 | 0.165 | 0.207 | 0.168 | 0.179 | 0.328 | 0.128 | 0.406 | 0.144 | -0.056 | -0.129 | 0.086 |
| res. Density (1,000) | -0.041 | -0.107 | -0.144 | -0.039 | -0.017 | -0.875 | -0.048 | -0.139 | -0.102 | -0.045 | -0.186 | -0.227 |
| driving cost (\$ per mile) | -1.137 | -1.175 | -4.351 | -2.732 | -6.706 | -7.184 | -3.661 | -2.996 | -2.051 | -5.105 | -4.418 | -6.048 |
| Log-likelihood at zero | -9153 | -7697 | -8789 | -8771 | -8871 | -7567 | -9385 | -9224 | -7840 | -8547 | -6576 | -7125 |
| Log-likelihood at convergence | -3114 | -2476 | -3488 | -3396 | -3202 | -3838 | -3457 | -3580 | -3927 | -3517 | -3205 | -3836 |
| Number of parameters | 32 | 32 | 32 | 32 | 32 | 32 | 32 | 32 | 32 | 32 | 32 | 32 |
| Number of observations | 1500 | 1155 | 1500 | 1500 | 1333 | 1500 | 1500 | 1500 | 1500 | 1500 | 1297 | 1500 |
| Adjusted R2 | 0.656 | 0.674 | 0.600 | 0.609 | 0.635 | 0.489 | 0.628 | 0.608 | 0.495 | 0.585 | 0.508 | 0.457 |

$$\hat{\Sigma}_{NE,urban} = \begin{pmatrix} 2.00 & 6.25 & 9.31 & 8.98 & 1.46 \\ 6.25 & 67.98 & 129.25 & 36.71 & 4.91 \\ 9.31 & 129.25 & 332.54 & 61.85 & 7.63 \\ 8.98 & 36.71 & 61.85 & 89.06 & 6.73 \\ 1.46 & 4.91 & 7.63 & 6.73 & 1.06 \end{pmatrix}$$

$$\hat{\Sigma}_{NE,suburban} = \begin{pmatrix} 2.00 & 8.59 & 12.58 & 11.78 & 1.48 \\ 8.59 & 72.33 & 128.74 & 89.07 & 6.09 \\ 12.58 & 128.74 & 377.38 & 60.82 & 8.79 \\ 11.78 & 89.07 & 60.82 & 204.98 & 8.20 \\ 1.48 & 6.09 & 8.79 & 8.20 & 1.10 \end{pmatrix}$$

$$\hat{\Sigma}_{NE,rural} = \begin{pmatrix} 2.00 & 5.05 & 5.69 & 10.59 & 1.43 \\ 5.05 & 82.50 & 84.65 & 104.52 & 6.84 \\ 5.69 & 84.65 & 89.30 & 105.84 & 7.35 \\ 10.59 & 104.52 & 105.84 & 332.62 & 11.36 \\ 1.43 & 6.84 & 7.35 & 11.36 & 1.18 \end{pmatrix}$$

$$\hat{\Sigma}_{MW,urban} = \begin{pmatrix} 2.00 & 7.37 & 8.53 & 9.69 & 1.51 \\ 7.37 & 78.10 & 89.20 & 68.67 & 6.54 \\ 8.53 & 89.20 & 117.14 & 53.37 & 7.55 \\ 9.69 & 68.67 & 53.37 & 161.23 & 7.96 \\ 1.51 & 6.54 & 7.55 & 7.96 & 1.16 \end{pmatrix}$$

$$\hat{\Sigma}_{MW,suburban} = \begin{pmatrix} 2.00 & 6.55 & 5.77 & 6.44 & 1.06 \\ 6.55 & 22.20 & 22.37 & 24.84 & 3.57 \\ 5.77 & 22.37 & 39.18 & 38.60 & 4.62 \\ 6.44 & 24.84 & 38.60 & 43.48 & 5.01 \\ 1.06 & 3.57 & 4.62 & 5.01 & 1.22 \end{pmatrix}$$

$$\hat{\Sigma}_{MW,rural} = \begin{pmatrix} 2.00 & 3.24 & 3.43 & 4.74 & 1.48 \\ 3.24 & 13.36 & 29.38 & 30.17 & 3.33 \\ 3.43 & 29.38 & 196.83 & 131.85 & 8.23 \\ 4.74 & 30.17 & 131.85 & 101.17 & 7.51 \\ 1.48 & 3.33 & 8.23 & 7.51 & 1.31 \end{pmatrix}$$

$$\hat{\Sigma}_{S,urban} = \begin{pmatrix} 2.00 & 5.93 & 16.64 & 12.20 & 1.51 \\ 5.93 & 97.61 & 30.17 & 118.66 & 6.13 \\ 16.64 & 30.17 & 190.82 & 125.52 & 10.84 \\ 12.20 & 118.66 & 125.52 & 199.50 & 9.69 \\ 1.51 & 6.13 & 10.84 & 9.69 & 1.22 \end{pmatrix}$$

$$\hat{\Sigma}_{S,suburban} = \begin{pmatrix} 2.00 & 8.87 & 12.80 & 19.90 & 1.59 \\ 8.87 & 290.51 & 215.82 & 359.41 & 9.16 \\ 12.80 & 215.82 & 262.21 & 292.96 & 11.42 \\ 19.90 & 359.41 & 292.96 & 881.16 & 18.09 \\ 1.59 & 9.16 & 11.42 & 18.09 & 1.28 \end{pmatrix}$$

$$\hat{\Sigma}_{S,rural} = \begin{pmatrix} 2.00 & -8.82 & -1.84 & -9.54 & -0.35 \\ -8.82 & 47.61 & 14.77 & 57.01 & 4.02 \\ -1.84 & 14.77 & 45.22 & 7.22 & 5.11 \\ -9.54 & 57.01 & 7.22 & 81.25 & 5.28 \\ -0.35 & 4.02 & 5.11 & 5.28 & 1.32 \end{pmatrix}$$

$$\hat{\Sigma}_{W,urban} = \begin{pmatrix} 2.00 & -0.60 & -0.57 & 0.68 & 1.18 \\ -0.60 & 24.59 & 25.61 & 24.81 & 3.01 \\ -0.57 & 25.61 & 27.17 & 29.50 & 3.22 \\ 0.68 & 24.81 & 29.50 & 53.94 & 4.33 \\ 1.18 & 3.01 & 3.22 & 4.33 & 1.18 \end{pmatrix}$$

$$\hat{\Sigma}_{W,suburban} = \begin{pmatrix} 2.00 & 3.45 & 5.37 & 5.49 & 1.56 \\ 3.45 & 17.76 & 27.14 & 40.06 & 3.24 \\ 5.37 & 27.14 & 176.57 & 143.66 & 6.49 \\ 5.49 & 40.06 & 143.66 & 156.55 & 6.76 \\ 1.56 & 3.24 & 6.49 & 6.76 & 1.27 \end{pmatrix}$$

$$\hat{\Sigma}_{W,rural} = \begin{pmatrix} 2.00 & 4.98 & 7.11 & 8.01 & 1.29 \\ 4.98 & 13.83 & 21.37 & 24.18 & 3.07 \\ 7.11 & 21.37 & 40.02 & 41.96 & 4.48 \\ 8.01 & 24.18 & 41.96 & 48.89 & 5.03 \\ 1.29 & 3.07 & 4.48 & 5.03 & 1.29 \end{pmatrix}$$

6.3 Application with ACS Data for Local Counties/Areas

6.3.1 County/Area Descriptions

The estimated models are then applied for small areas with ACS PUMS data. Six counties/areas are random selected, which are San Diego County in California (Figure 6.3), Queens in New York (Figure 6.4), Nassau County in New York (Figure 6.5), PUMA 1900 area (5 counties) in Texas (Figure 6.6), Fairfax County in Virginia (Figure 6.7) and Henrico County in Virginia (Figure 6.8):

- San Diego County, CA - West, Urban

Although California has a large dataset from the statewide household travel survey, it is still worth to examine the performance of the national models on a metropolitan area with big cities; here we have selected San Diego County to perform this analysis. The basic demographic information and sample size from NHTS and ACS data are:

Total area: 4,525.52 *mile*²

Total population: 3,095,313 (2010 Census)

Population density: 680/*mile*²

ACS : 11653 obs.

NHTS: 3712 obs.

- Queens, NY - Northeast, Urban

Queens Borough is a highly populated area in New York City, which is the most dense city in the U.S. People in this area may have different travel behavior than those residing in other regions. Meanwhile, this area has many immigrants from all around the world which may affect their travel choices as well. Again, although the New York City has good household travel surveys, it is still good to test the national models for this extremely dense area. The basic demographic information and sample size from NHTS and ACS data are:

Total area: 178.28 *mile*²

Total population: 2,272,771 (2010 Census)



Figure 6.3: Maps of San Diego County, CA

Population density: 21,116/*mile*²

ACS : 6985 obs.

NHTS: 251 obs.



Figure 6.4: Map of Queens, NY

- Nassau County, NY - Northeast, Urban

Nassau County is located next to the east bounder of Queens in New York and many households in this county have jobs in New York City. This county is

still within the New York metropolitan area and this application is to validate the different travel styles within the same metropolitan area but different counties, thus validate the effectiveness of the national models. The basic demographic information and sample size from NHTS and ACS data are:

Total area: 453 *mile*²

Total population: 1,339,532 (2010 Census)

Population density: 4,669/*mile*²

ACS : 4875 obs.

NHTS: 265 obs.

- PUMA 1900, TX - South, Rural

This area includes Hill County, Navarro County, Limestone County, Freestone County and Navarro County in Texas. This area is very scattered and it is located at roughly the middle point between Austin and Dallas - two big metropolitan areas in Texas. The 2009 NHTS only has less than 100 observations in this area, however ACS has around 900 observations. This is a good example that local household travel survey is not available and the national data sample has very limited observations. The basic demographic information and sample size from NHTS and ACS data are:

Hill County, TX Total area: 986 *mile*² Total population: 35,089 (2010 Census) Population density: 34/*mile*²



Figure 6.5: Maps of Nassau County, NY

Navarro County, TX Total area: 1,086 *mile*² Total population: 47,735
(2010 Census) Population density: 18/*mile*²

Limestone County, TX Total area: 933 *mile*² Total population: 23,384
(2010 Census) Population density: 23/*mile*²

Freestone County, TX Total area: 892 *mile*² Total population: 19,816
(2010 Census) Population density: 21/*mile*²

Navarro County, TX Total area: 779 *mile*² Total population: 17,866 (2010
Census) Population density: 57/*mile*²

ACS : 894 obs.

NHTS: 93 obs.

- Fairfax, VA - South, Urban

Fairfax County is located in the Washington DC metropolitan area and west to the District of Columbia. It is one of the counties that have the highest household income in the country. Many people live in the Fairfax County commute to DC. The basic demographic information and sample size from NHTS and ACS data are:

Total area: 407 *mile*²

Total population: 1,118,602 (2010 Census)

Population density: 2,738.5/*mile*²

ACS : 4033 obs.



Figure 6.6: Maps of PUMA Area 1900, TX

NHTS: 205 obs.



Figure 6.7: Maps of Fairfax County, VA

- Henrico, VA - South, Urban

Henrico County is a portion of the Richmond Metropolitan area, surrounding the City of Richmond. Henrico is one of the oldest counties in the United States. The basic demographic information and sample size from NHTS and ACS data are:

Total area: 245 *mile*²

Total population: 314,881 (2010 Census)

Population density: 1,323/*mile*²

ACS : 1274 obs.

NHTS: 379 obs.



Figure 6.8: Map of Henrico County, VA

Figure 6.9 present some basic statistics from the ACS PUMS files. Fairfax County has the highest average household vehicle ownership, whereas the average number of vehicles per household in Queens is less than 1. Generally, the households in Fairfax County and Nassau County have bigger household size, more workers and

children and much higher income. On average, about three-quarters household own their home, while half of the households in San Diego County and Queens rent their home.



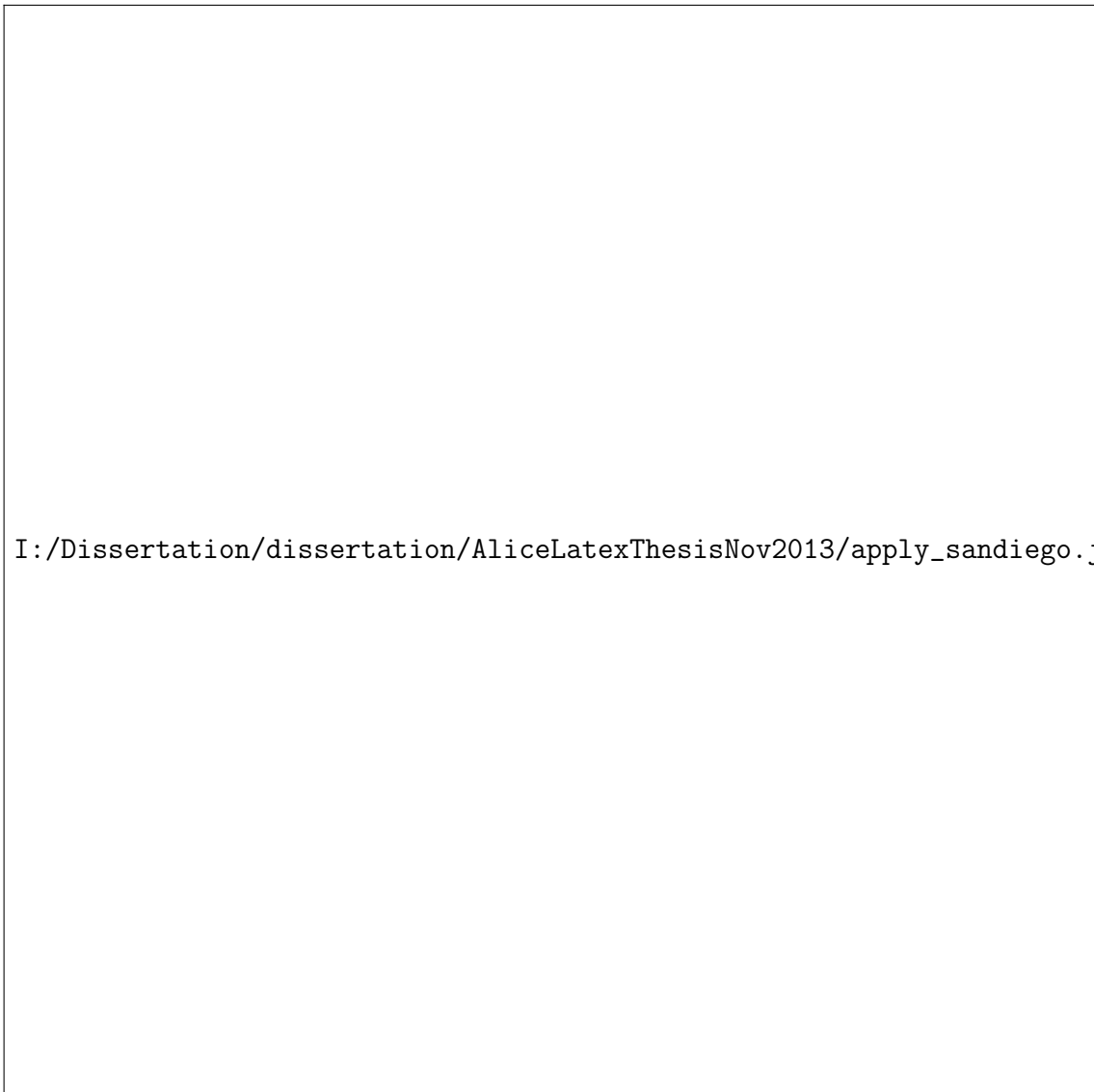
Figure 6.9: Data Statistics from American Community Survey

6.3.2 Application Results

Figure 6.10 to Figure 6.15 presents the application results of the national models on vehicle ownership and usage. Generally, the models are able to replicate the actual values in each county/area. The model slightly underestimates the average vehicle ownership and mileage in San Diego County. For Queens, NY, the model overestimates the portion of 0-car households thus it overestimates the average number of vehicle per household. Nevertheless, the prediction of mileage is close to the actual value. The model slight underestimates the average household vehicle ownership but overestimates the average annual mileage per household. The estimates for the PUMA 1900 area in Texas are very close to the actual numbers, with the exception of small shifts in the share of the alternatives. The application results for the Fairfax County shows that the model underestimates the share of 1-car households but overestimates the share of 2-car households, and it overestimates the average mileage for this county. The predictions for the Henrico County are fairly close to the real values, both for the vehicle ownership and the annual mileage. Finally, Figure 6.16 summarizes the application results for the six counties/areas.

6.4 Chapter Summary

This chapter develops a system of national vehicle ownership models - twelve discrete-continuous models for the United States. The models are estimated using 2009 NHTS data for each combination of four regions (Northeast, Midwest, South and West) and three area types (urban, suburban and rural). In addition, the



`I:/Dissertation/dissertation/AliceLatexThesisNov2013/apply_sandiego.jpg`

Figure 6.10: Application results of San Diego County, CA

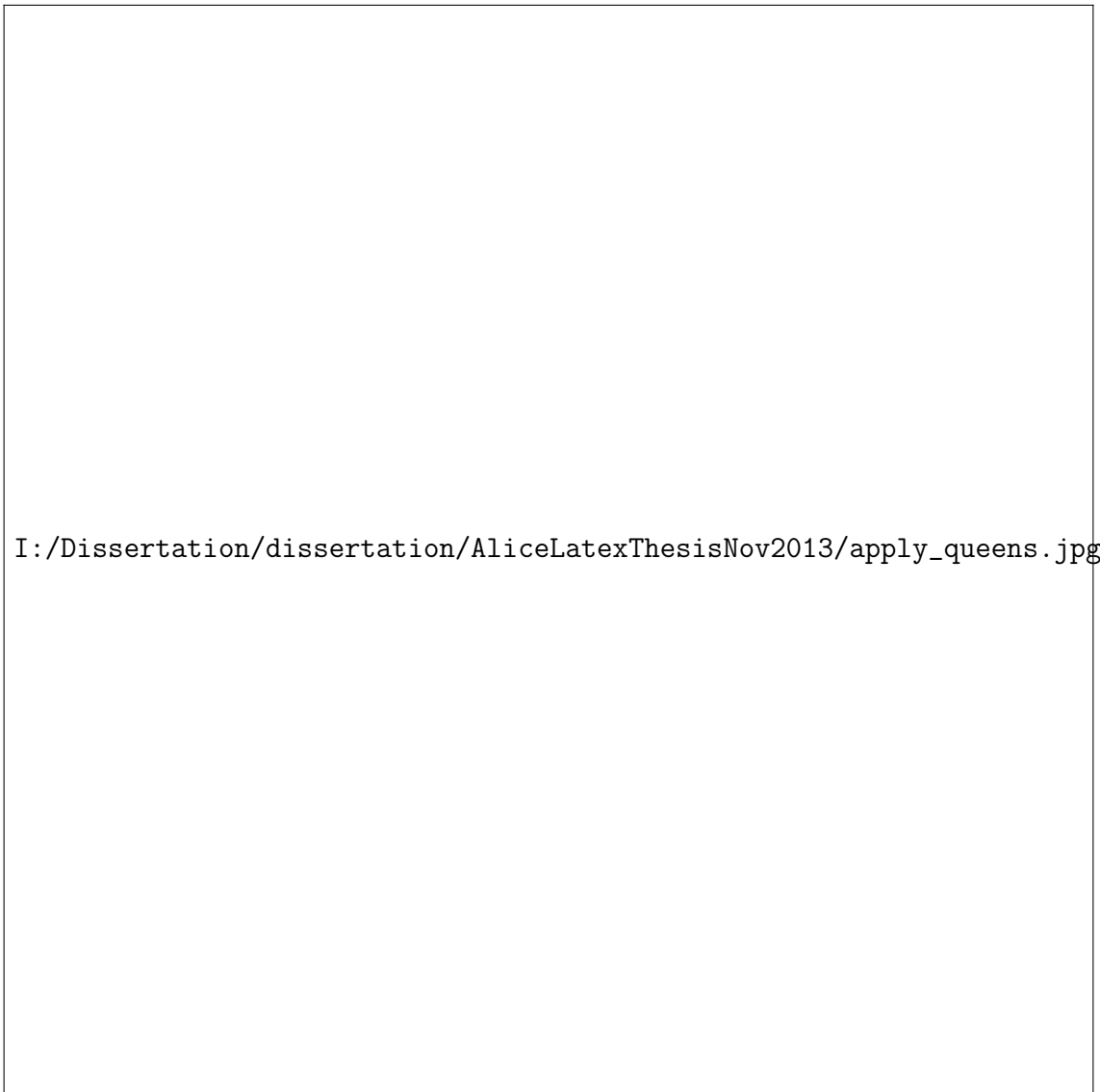


Figure 6.11: Application results of Queens, NY



Figure 6.12: Application results of Nassau County, NY

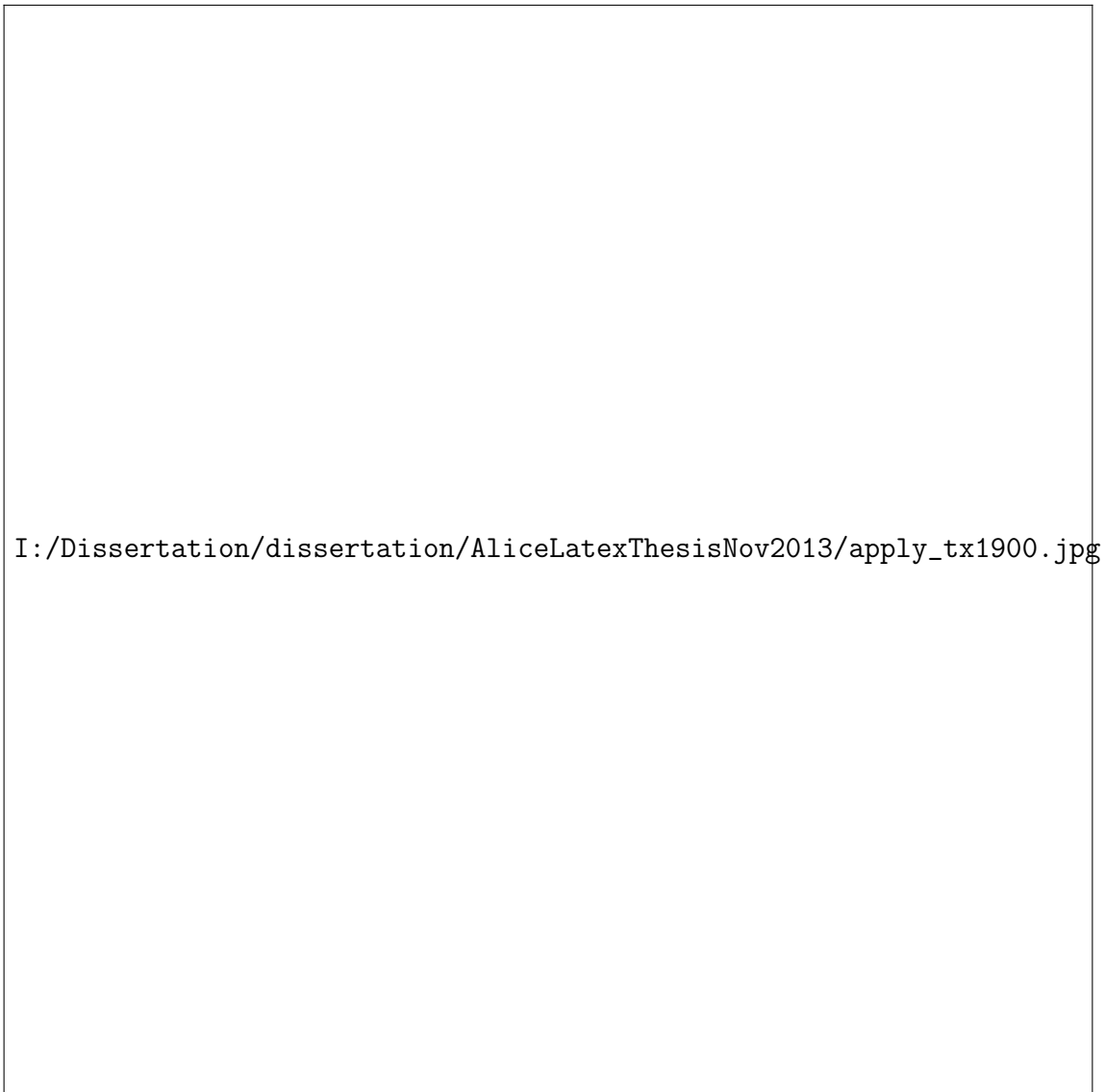


Figure 6.13: Application results of PUMA area 1900, TX

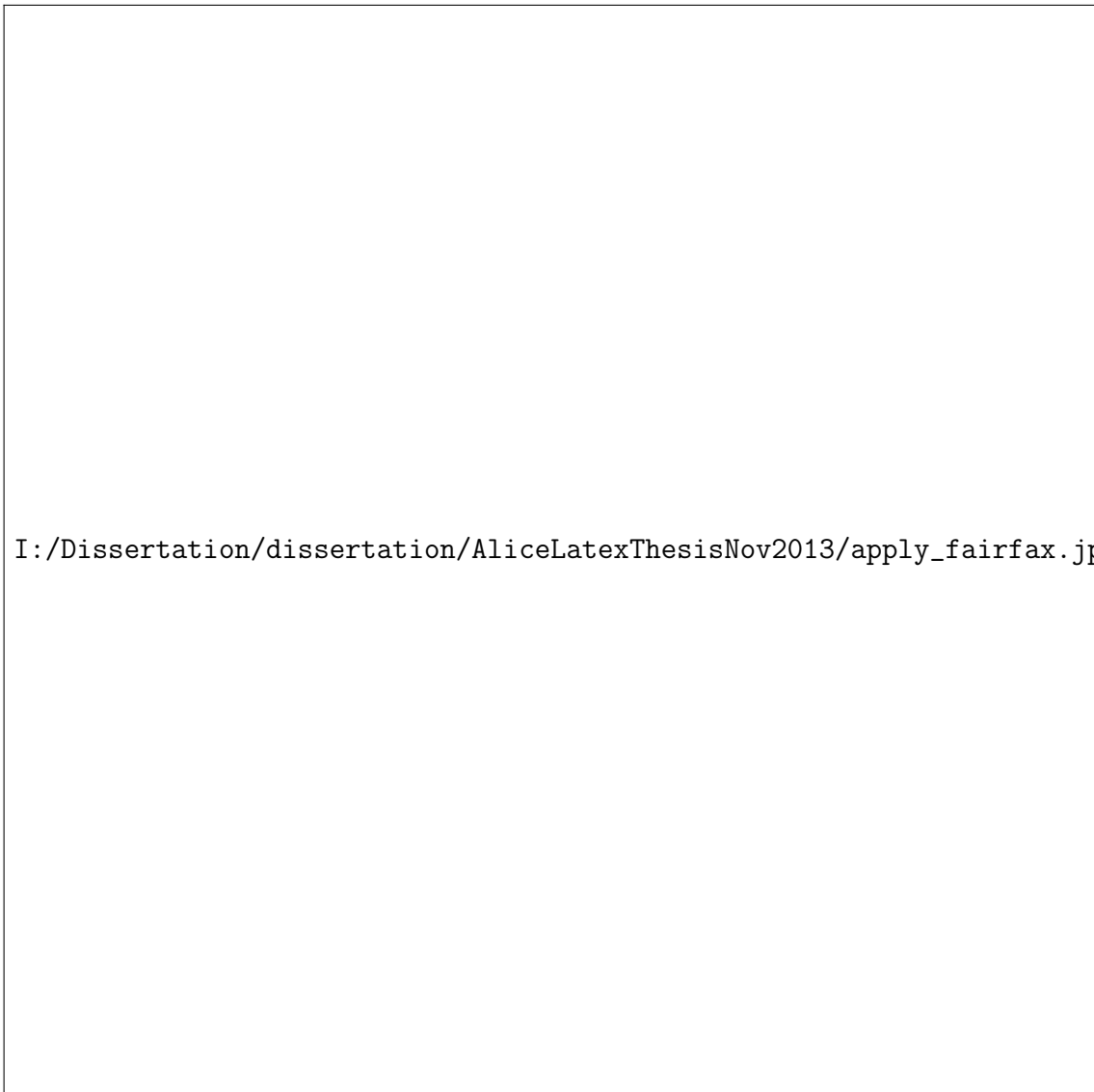
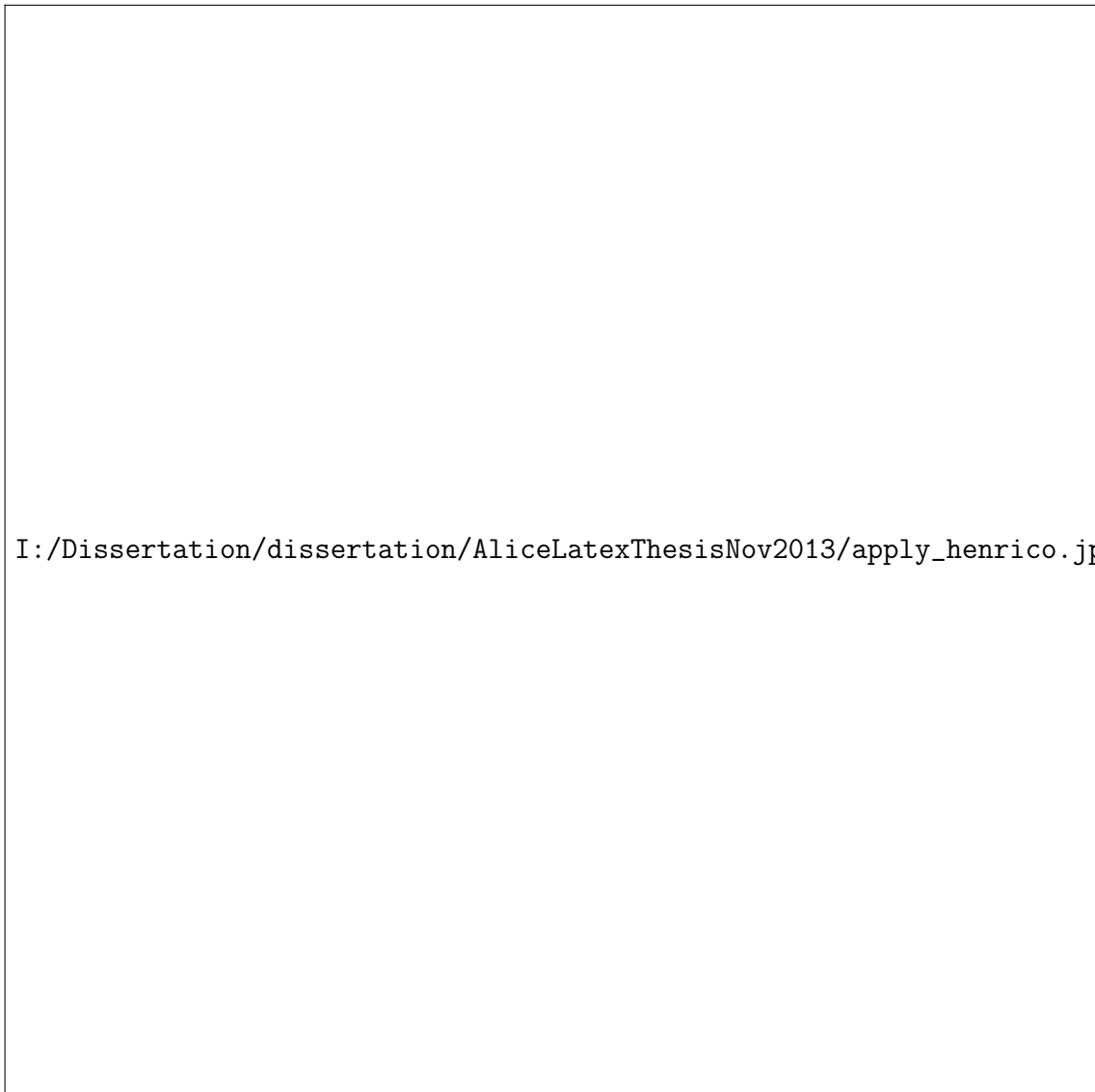


Figure 6.14: Application results of Fairfax County, VA



I:/Dissertation/dissertation/AliceLatexThesisNov2013/apply_henrico.jpg

Figure 6.15: Application results of Henrico County, VA



`I:/Dissertation/dissertation/AliceLatexThesisNov2013/apply_acs_summary.jpg`

Figure 6.16: Summary of applications for the six counties/areas

model is applied to six counties/areas using the 2009 ACS PUMS data. Although some deviations from the real values are observed, the integration of NHTS and ACS data is valid, and the results from the six applications demonstrate the ability of the national models in providing accurate estimates for various city/area types. The national models are valuable planning tools both at the national level and for small areas, especially those lacking local household travel survey data. The results further validate the proposed discrete-continuous framework for modeling household vehicle ownership decisions.

Chapter 7

Measuring Transit Service Impacts on Vehicle Ownership and Usage

7.1 Introduction

A recent article published in the press by Addison [Addison, 2010] shows that Americans scrapped 14 million cars in 2009, while they only bought 10.5 million new ones. The 2009 drop was the first large decline in vehicle ownership registered in the past 50 years. Although the recession probably played a major role, this decline might be also due to the introduction of smart growth policies and the consequent increase in urban density, the adoption of employer commute and flex-work programs, the expansion of car sharing, the introduction of the Car Allowance Rebate System (CARS), colloquially known as "Cash for Clunkers", and improved rail connectivity and inter-modality. Addison also reported that the increase in the use of public transit is one of the top ten reasons for the drop in car ownership especially in large metropolitan areas . In February 2013, President Barack Obama fleshed out plans to invest in public transportation and repair the nation's aging infrastructure. In fact, the administration has invested in more than 350 miles of new rail and bus rapid transit, 45,621 buses, and 5,545 railcars [American Public Transportation Association, 2013].

The effects of transit service level on car ownership has been examined in a number of national studies, the U.S. ([Deka, 2002], [Kim and Kim, 2004b], UK

([Cullinane, 1992], [Goodwin, 1993]), Australia ([Hensher, 1998]), Canada ([P. and P., 1998]), The Netherlands ([Kitamura, 1989]), Germany ([Bratzel, 1999]), and China ([Cullinane, 2002], [Li et al., 2010]). More specifically, Kitamura ([Kitamura, 1989]) investigated the causal relation between car ownership and transit use on data obtained from the 1984 Dutch National Mobility Panel survey. The results show that car use determines transit use, and that transit use does not determine car use. Nevertheless, the current situation is very different from the 80s, when the "car boom" was taking place, the number of household with access to one or more cars was limited, and fuel price was relatively low. Bunt and Joyce [Hensher, 1998] conducted a household survey to test the effectiveness of Vancouver's SkyTrain and its effect on car ownership patterns near the rapid transit stations. Statistics from the survey show that the average car ownership is much lower for households located near SkyTrain stations. Cullinane [Cullinane, 2002] found that good public transport can deter car ownership based on an attitudinal survey in Hong Kong, where public transport is plentiful and cheap and car use is low. Deka [Deka, 2002] applied regression models to examine the relationship between transit availability and auto ownership with travel survey data from Los Angeles. The conclusion is that significant improvements will be needed in transit services to bring a slight decrease in auto ownership among the general population. Kim and Kim [Kim and Kim, 2004b] developed econometric models to predict the effect of accessibility to public transit on automobile ownership and miles driven. Important findings in their analysis are: (i) the number of licensed drivers is the primary determinant of the number of automobiles owned, (ii) the presence of children is not a significant

factor in automobile ownership and VMT, and (iii) VMT is affected more by transit in multi-vehicle households than in one-vehicle households.

Recent studies provide evidence that good public transportation might encourage people to reduce vehicle ownership and use. However, very few studies use advanced quantitative methods to investigate the relationship between public transit service and vehicle ownership and use. Other difficulties include collecting geographic data and quantifying the transit service level. Moreover, many metropolitan areas are interested in improving public transportation in order to reduce traffic congestion and in providing more efficient transportation systems ([Washington Metropolitan Area Transit Authority, 2012a], [Maryland Transit Administration,]). Therefore, it is crucial to explore the impact of public transportation on vehicle ownership and use with advanced methods and accurate data based on geographic information systems.

This chapter aims to investigate the effects of improved public transportation services on household vehicle ownership and use with the unordered discrete-continuous models that proposed in Chapter 3. The analysis is conducted for the Washington D.C. Metropolitan Area, which is a mix of urban and suburban areas with a relatively good public transportation system for which further improvements are foreseen. The information used for model estimation was obtained from different sources. The 2009 National Household Travel Survey (NHTS) data with geographic reference (U.S. Census Tract level) was kindly provided by the Federal Highway Administration (FHWA), U.S. DOT, while the General Transit Feed Specification (GTFS) data was obtained from the Washington Metropolitan Area Transit Au-

thority (WMATA).

Different measurements for transit level of service are found in the literature. The Local Index of Transit Availability (LITA) [Rood, 1998] measures the transit service intensity of an area with transit data and census data (demographic information). Depending on the data availability, LITA scores can be computed for any area unit. Transit Capacity and Quality as defined in the Service Manual (TCQSM) [Transportation Research Board, 2003] also uses transit data and census data but incorporates a service coverage measure to assess transit accessibility. TCQSM offers a comprehensive guide for infrastructure enhancements specific to public transportation systems. The Time-of-Day Tool (Polzin02) provides the relative value of transit service accessibility for each time period and requires data on temporal distribution of travel demand in addition to transit and census data.

The method to measure transit service in this analysis is similar to the one proposed by Keller [Keller, 2012]. The method mainly follows the TCQSM manual and takes into account both spatial and temporal characteristics of the transit system. Data on the temporal distribution of travel demand is not available so that the transit service measurements are calculated on a yearly average level only.

7.2 Data Geo-Processing and Data Integration

7.2.1 Spatial Measurements of Transit Service

This section follows the TCQSM manual recommendations to calibrate the coverage of public transportation services. In particular, a service buffer is created

for each area surrounding a station to derive the area of usage for potential transit users. The TCQSM ([Transportation Research Board, 2003]) suggests a 0.25-mile buffer around bus stops and a 0.5-mile buffer around rail stations. These buffers are based on willingness to travel studies; buffers based on these distance ranges tend to represent between 75 and 80 percent of all walking trips to a transit stop.

The GTFS data is firstly converted from .txt files to shapefiles for both transit stations and routes, and then projected in ArcGIS along with Census TIGER files. The buffer zones for the bus stops with radius 0.25-mile and metro routes with radius 0.5-mile are then created. The overlapped buffers are dissolved to eliminate double counting. The coverage area is joined to the census tract zone and the percentage of coverage is computed. The process is repeated for each stop/route and for each census tract zone. The final variables that are produced in this process include (1) percentage of bus stops coverage, (2) percentage of metro routes coverage, (3) total length of bus routes, (4) total length of metro routes, and (5) total number of bus stops. All the variables above are calibrated for each census tract in the Washington D.C. Metropolitan area.

7.2.2 Temporal measurements of bus service

The data related to transit timetable in the GTFS files is utilized to calibrate the temporal measurements of bus services. Firstly the GTFS files are merged with the key IDs (see Figure 4.1). The merged data have information on the bus arrived time for each route and stop for an entire day (24 hours). Then for each stop and

for each route, the bus service duration and average headway is computed with data mining techniques. Finally the average duration and headway are aggregated for each census tract zone.

7.2.3 Transit service index (TSI)

The transit service index takes into account both spatial and temporal measurements and it is calculated with the percent service coverage area, the average service headway and the service duration. For each census tract zone, the TSI is calculated as:

$$TSI = \frac{\text{percent service coverage area}}{\text{average service headway}} \times \text{service duration}$$

Table 7.1 presents some examples of TSI calibration from real data. In this analysis, TSI is calculated for the bus service only. The reason of not including metro service is because the time schedule of metro subways in the DC area is comparatively rigid and does not create variation among different census tract zones. Instead, the percent service coverage area is created as the measurement of transit service.

Table 7.1: Sample calibration of TSI

| Zone ID | % coverage | Service headway | Service duration | TSI |
|-------------|------------|-----------------|------------------|-------|
| 11001009204 | 100% | 0.44 hr | 7.83 hr | 17.96 |
| 24033801309 | 6.68% | 0.34 hr | 4.03 hr | 0.80 |
| 51177020104 | 0 | 0 hr | 0 hr | 0 |
| 11001001402 | 49.55% | 0.38 hr | 14.69 hr | 19.19 |

7.2.4 Data integration and final database

The final database consists of three components: NHTS data, GIS output and vehicle characteristics. As shown in Figure 7.1, the data sets are linked with key IDs. Specifically, the 2009 NHTS data includes household socio-economic information, such as household income, household size, number of drivers, number of workers, and land use characteristics around the household location, such as the residential density of the census tract zone, the urbanization level, etc. The GIS output includes data on bus stop coverage percentage, metro route coverage percentage, total length of bus routes, total length of metro routes, total number of bus stops, transit service index, average bus headway, and average bus service duration for each census tract zone. The vehicle characteristic data includes purchase price, operating cost, fuel economy, seating, performance, and other specifications for each vehicle type.

7.3 Estimation Results

Table 4 presents the parameter estimates of the joint vehicle ownership and usage model; it should be noted that the model includes a logsum variable derived from the vehicle type and vintage model in Table 5.2. The variables TSI, created to represent bus and metro coverage percentage, are significant and have a negative impact on household vehicle ownership and miles traveled. The variable "TSI of bus" is selected instead of other measures because it gives a more comprehensive representation with respect to both spatial and temporal bus service information. In terms of metro subways, the time schedule is comparatively rigid so that the metro



I:/Dissertation/dissertation/AliceLatexThesisNov2013/GTSF_database.jpg

Figure 7.1: Data structure of the final database

service is measured using the percentage coverage only. With good accessibility of bus and metro service, households tend to own fewer cars. The magnitudes of the coefficients increase with the number of cars owned by the household, indicating that the transit service level has greater impacts on multi-vehicle households. In particular, the coefficient of metro service coverage for 4-car household is significantly greater than the one obtained for other alternatives. Coefficients of household income are positive and significant; the value of the coefficients is larger for households owning more cars. Households with higher income tend to own multiple cars and drive more, and the higher their income, the more likely that they will own more cars. Households with owned house are more likely to have higher mileage on their vehicles. Households with more drivers own more vehicles and drive more. The coefficients related to the number of drivers are significant except in the one-car alternative. In terms of the characteristics of the household head, the dummy variable "female household head" is significant except for the one-car household; the negative sign meaning that households with a female head tend to own fewer cars and to drive less. The coefficients of residential density are significant and negative (except for the one-car household), inferring that the households located in a more dense area have lower probability of owning more cars and of driving less. The parameter of driving cost is negative and significant, indicating that higher operational cost induces households to drive less.

In addition to the coefficients of the variables, the covariance matrix between the discrete and the continuous independent variables is estimated. In particular, the bottom line of the matrix explains the correlation between the mileage traveled

and the utility differences (with respect to the zero-car alternative) of the vehicle ownership alternatives. The positive numbers mean that higher mileage usage increases the utility of owning more cars; the magnitude of the correlation factors increases with the number of household vehicles. The negative value found for the correlation across mileage and zero-car alternative can be explained by the fact that zero miles of very low mileages further decrease the difference in utility of owning a car or not owning a car.

Table 7.2: Estimation results

| Variable | Coefficient | Std. Err |
|----------------------------------------------------------------------------------|----------------|----------|
| <i>Dependent variable: Number of cars</i> | | |
| logsum (expected utility from vehicle type choice alternative specific constant) | 0.430 | 0.011 |
| 1 car | -3.161 | 0.193 |
| 2 cars | -17.050 | 0.267 |
| 3 cars | -22.913 | 0.219 |
| 4+ cars | -27.934 | 0.178 |
| household income level | | |
| 1 car | -0.090 | 0.021 |
| 2 cars | 0.446 | 0.053 |
| 3 cars | 0.490 | 0.054 |
| 4+ cars | 0.440 | 0.052 |
| number of drivers | | |
| 1 car | -0.038 | 0.197 |
| 2 cars | 7.185 | 0.193 |
| 3 cars | 7.982 | 0.193 |
| 4+ cars | 7.791 | 0.183 |
| gender of household head (female) | | |
| 1 car | 0.089 | 0.199 |
| 2 cars | -2.350 | 0.189 |
| 3 cars | -2.495 | 0.194 |
| 4+ cars | -2.637 | 0.159 |
| urban size | | |
| 1 car | -0.049 | 0.048 |
| 2 cars | -0.071 | 0.115 |
| 3 cars | -0.153 | 0.113 |
| 4+ cars | -0.204 | 0.112 |
| residential density (census tract level) | | |
| 1 car | 0.051 | 0.014 |

| | | |
|------------------------------------------|---------------|-------|
| 2 cars | -0.524 | 0.135 |
| 3 cars | -0.704 | 0.150 |
| 4+ cars | -0.549 | 0.151 |
| TSI of bus | | |
| 1 car | 0.018 | 0.008 |
| 2 cars | -0.103 | 0.038 |
| 3 cars | -0.105 | 0.036 |
| 4+ cars | -0.116 | 0.039 |
| percentage coverage of metro routes | | |
| 1 car | 0.280 | 0.164 |
| 2 cars | -2.212 | 0.267 |
| 3 cars | -1.756 | 0.296 |
| 4+ cars | -9.442 | 0.185 |
| <i>Dependent variable: Miles (10k)</i> | | |
| constant | 1.470 | 0.121 |
| household income level | 0.124 | 0.006 |
| own home | 0.372 | 0.107 |
| gender of household head (female) | -0.080 | 0.076 |
| residential density (census tract level) | -0.055 | 0.012 |
| driving cost (\$ per mile) | -4.823 | 0.294 |
| TSI of bus | -0.025 | 0.004 |
| percentage coverage of metro routes | -0.324 | 0.159 |
| Log-likelihood at zero | -5880.231 | |
| Log-likelihood at convergence | -3260.811 | |
| Number of parameters | 41 | |
| Number of observations | 1420 | |
| Adjusted ρ^2 | 0.44 | |

Note: Variables that are significant at 95% level or above are bolded.

$$\hat{\Sigma} = \begin{pmatrix} 2.00 & -7.51 & -7.43 & -7.54 & -0.68 \\ -7.51 & 29.22 & 30.68 & 30.41 & 2.92 \\ -7.43 & 30.68 & 35.04 & 33.67 & 3.55 \\ -7.54 & 30.41 & 33.67 & 32.72 & 3.34 \\ \mathbf{-0.68} & \mathbf{2.92} & \mathbf{3.55} & \mathbf{3.34} & 1.23 \end{pmatrix}$$

In order to investigate the significant role of the transit service attributes, the car ownership model is re-estimated without transit related variables. A log-likelihood ratio test is conducted to test the significance of transit service variables in the vehicle ownership model:

H_0 : Coefficients of transit service variables are not zero (full model)

H_1 : Coefficients of transit variables are zero (reduced model)

degree of freedom (DOF) = 10

$$\begin{aligned} & -2[LL(\hat{\beta}^1) - LL(\hat{\beta}^0)] \\ & = -2[(-3349.812) - (-3260.811)] \\ & = 178.002 > \chi_{10,0.05}^2 = 25.188 \end{aligned}$$

The test statistic is much larger than the Chi-square with 10 degrees of freedom at the 95 percent confidence level. Therefore we reject the hypothesis that the coefficients of transit service variables are zero and I conclude that the model could not be reduced. The testing result confirms again the significant role of transit service variables in vehicle ownership models.

7.4 Policy Analysis

The Washington Metropolitan area is developing a 30-year transit plan [Washington Metropolitan Area Transit Authority, 2012b], which aims to provide a long term vision for future growth and to improve and expand transit service. The goal of the regional plan is to seek solutions such as making pedestrian and rail connections between lines to bypass bottlenecks, adding new rail lines through the downtown core and improving surface transit. A recent announcement [Washington Metropolitan Area Transit Authority, 2012a] from WMATA says that in 2013, \$5 million will be invested to provide customers with better bus service. One of the biggest efforts is a new limited-stop MetroExtra route, which improves the transit system with more frequent service, additional capacity, and expanded hours of operation. On

the other hand, the metrorail ridership is expected to top 1 million daily rides by 2040 and the system's core will be severely crowded [Johnson, 2011]. WMATA has been looking at long-term strategies for expanding transit. The Purple Line [Maryland Transit Administration,], which is a 16-mile transit line that will connect the Red, Green and Orange lines of the metro system in the suburban area of Maryland, can be seen as part of the long term plan. Meanwhile, a Beltway Metro Line is under consideration.

Given the numerous investments foreseen for the public transportation system in the Washington DC Metropolitan area, it is worth to examine the impacts of improved transit services on household vehicle ownership and usage. In this chapter, the model estimated in the previous section is applied to evaluate different policy scenarios. I first analyze the effects of improved bus services; in this hypothetical scenario every census tract zone has at least 50% bus stop coverage, 15-minute average headway and 6 peak hours duration (6:30AM - 9:30AM and 3:30PM - 6:30PM). In the improved metrorail service scenario, the core area of Washington Metropolitan area (urban size greater than 1 million) has at least 50 percent metro route coverage.

The application results are presented in Table 7.3. The short-run impacts of improved transit service generally reduce both vehicle ownership and miles traveled. The average vehicle ownership is reduced by 2 percent in the improved bus service scenario and 1.5 percent in the improved metro service scenario. The annual mileage traveled decreases by about 8 percent with improved bus service and 1.6 percent with improved metro service. Comparatively, the improved bus service has greater

impacts on reducing both the vehicle ownership and the mileage traveled.

It should be noted here that the NHTS data has limited number of households in the DC and Maryland area due to the fact that neither of these regions are in the NHTS add-on program. The predictions provided could be more accurate with an increased number of observations available for model calibration.

Table 7.3: Policy analysis based on different improvement of the transit service

| | current | Improved bus service | | Improved metro service | |
|---------------------------|----------|----------------------|---------|------------------------|---------|
| | | predicted | %change | predicted | %change |
| 0-car household | 7.16% | 7.17% | 0.01% | 7.17% | 0.01% |
| 1-car household | 23.06% | 26.05% | 2.99% | 24.60% | 1.55% |
| 2-car household | 46.56% | 44.32% | -2.25% | 44.84% | -1.72% |
| 3-car household | 17.82% | 17.19% | -0.64% | 19.44% | 1.61% |
| 4-car household | 5.40% | 5.28% | -0.12% | 3.95% | -1.45% |
| Average vehicle ownership | 1.91 | 1.87 | -2.03% | 1.88 | -1.49% |
| Mileage | 22231.70 | 20410.40 | -8.19% | 21879.50 | -1.58% |

7.5 Chapter Summary

The Washington Metropolitan area is a diverse region with both dense urban areas and suburban areas. This region is also served by a good public transportation system that will undergo several improvement plans in the short and long term. Given the raising interests on transit investments from both federal and state governments, as well as, the traffic concerns on the Beltway, it is important to understand and quantify the relation between public transportation service and household vehicle ownership and usage. In particular, this chapter has analyzed the impact of

improved bus and metro services on household ownership and use decisions in the Washington Metropolitan area.

This chapter proposes a methodology to integrate the household travel survey with geographic data. Specifically, the main data sources are the 2009 National Household Travel Survey (NHTS) and the General Transit Feed Specification (GTFS). Secondary data includes the 2009 Census TIGER shapefiles and vehicle characteristics from the Consumer Reports. Both spatial and temporal measurements of transit service are created based on the GTFS data and geographic information data using data mining techniques. The transit service index is calculated with these measurements and then integrated with the NHTS data, the GIS output data and the vehicle characteristics into one database referenced at the census tract level.

This chapter jointly estimates the household decisions on vehicle ownership and usage with the integrated database; estimates are obtained for household social-demographic attributes, land-use characteristics, vehicle characteristics and transit service variables. The model is then applied to policy scenarios that accounts for transit investments. The results obtained show that transit service generally reduces both vehicle ownership and miles traveled. The average vehicle ownership is reduced by 1.5 - 2.0 percent and the mileage decreases by about 1.6 - 8.0 percent respectively with improved bus service and with improved metro service.

Chapter 8

Conclusions and Future Research

8.1 Summary and Research Contributions

Vehicle ownership plays an important role in the overall transportation planning process, due to its impacts on the environment, energy consumption, economic system and public health. In this dissertation, an integrated discrete-continuous model is proposed to simultaneously estimate the household decisions on vehicle holding, type and use. The model uses a multinomial probit model to estimate household vehicle holding decisions and a multinomial logit model to estimate the vehicle type. The vehicle usage decisions have been integrated into these discrete models with an unrestricted correlation pattern between the discrete and the continuous parts. The dissertation also compares the outcomes of the ordered and unordered discrete-continuous structures. Results obtained from the 2009 National Household Travel Survey show that significant correlation exists between the vehicle holding and use decisions. Therefore, significant estimation bias is expected when ignoring correlations among these decisions and when assuming that they are independent. The comparison results also indicate that unordered discrete-continuous model outperforms the ordered structure in terms of the goodness of fit.

The second half of the dissertation focuses on the applications of the proposed modeling framework and on the related policy analysis. The 2009 National

Household Travel Survey data is the main data source to estimate household vehicle ownership decisions across the United States. Twelve models are calibrated for the four Census Regions of the United States (Northeast, Midwest, South and West) and three area types (urban, suburban, rural). Due to the different demographic profiles and area types (e.g., urban, rural, etc.), a number of sites are selected to account for heterogeneity in regional locations and residential density levels. Then the estimated models are applied to six randomly selected counties/areas, using the 2009 American Community Survey Public Use Microdata Sample. Results from the six applications demonstrate the capability of the national models in providing accurate estimates for the various city/area types selected, although small prediction errors are found when comparing real data and estimates.

The proposed modeling framework is also applied with additional transit service variables to analyze the impact of improved bus and metro services on household vehicle ownership and use decisions. In order to derive the transit variables, the household travel survey data is integrated with transit data, namely the General Transit Feed Specification (GTFS) data. In the analysis, spatial measurement, temporal measurement and the combination of the two measurements of transit service are computed in GIS. Results show that transit service variables are significant factors in household vehicle ownership choices and that the proposed methods are able to effectively predict changes in vehicle ownership and usage due to transit service improvements.

In conclusion, this dissertation contributes to both theoretical analysis and practical applications of the household vehicle ownership problem:

- An integrated discrete continuous choice model is developed to simultaneously estimate the household choices on vehicle ownership (discrete), the types (discrete) and annual mileage traveled (continuous).
 - The model is able to include a large number of alternatives in both the vehicle holding and the vehicle type choices.
 - The model allows unrestricted correlations of the unobserved factors between the discrete and continuous parts.
 - The model accommodates flexible specifications.
 - The model can be applied for policy analysis.
 - The model can generate reasonable estimates of the coefficients.
 - The covariance matrix explains well the correlations between the unobserved factors from the utilities of the discrete choices and the demand function of the continuous choice.
 - The non-simulation approach provides a better model fit.
 - The performance of the model would be improved if the information about vehicle type choice is included.
- A comparison of unordered and ordered structures in discrete-continuous framework is conducted with operational data. The results show that the unordered discrete continuous model is more appropriate than the ordered discrete continuous model in estimating household vehicle ownership and usage decisions.
- A system of national models on household vehicle ownership choices is devel-

oped with National Household Travel Survey data and American Community Survey data. Applications for six randomly selected areas demonstrate that the models are able to produce accurate estimates.

- The model is further applied using geographic data to study the impacts of improved transit service on household vehicle ownership choices in the Washington D.C. metropolitan area.

8.2 Future Research

There are several future directions in this research that are worth further investigation. The general ideas for improving the current research are summarized as follows.

- First and foremost, it would be valuable to analyze the direct correlation between vehicle type and vehicle usage, and estimate mileage for each vehicle in the household in the future. Because the vehicle type could affect how many miles the household travel with the vehicle, meanwhile, the demand for vehicle mileage traveled also could be a key factor on the vehicle type choice. For example, a family with both a compact car and a pick-up truck may travel with the more fuel-efficient compact car on a daily basis and may only use the pick-up truck when it must. A family member who drives 20,000 miles a year for commuting may choose a vehicle with high MPG.
- Another limitation in the dissertation is that all of the coefficients in the models are assumed to be constant and they do not vary over different groups

of households. Therefore, random parameter approach could be integrated into the framework to capture the taste variation among the population.

- The proposed framework is a static model and only provides short-run forecasting results. It could be further extended into a dynamic discrete-continuous model with a module to capture the household's dynamic choices on vehicle holding. For example, [Xu, 2011] developed a dynamic vehicle ownership choice model which allows the estimation of the probability of buying a new vehicle or postponing this decision; if the decision to buy is made, the model further investigates the vehicle type choices. Dynamic models explicitly account for consumers' expectations of future vehicle quality or market evolution, arising endogenously from their purchase decisions. By incorporating this component into the discrete-continuous framework, the modeling results would be able to provide the policy makers a reference for medium to long term planning.
- In the dissertation, the error terms between the discrete and continuous parts are assumed to be multivariate normal distributed. Although the correlations are estimated with an unrestricted covariance matrix, this part can be improved with a more flexible correlation pattern. For example, the copula models permit the combination of any univariate marginal distributions that need not come from the same distributional family. They are very general, encompassing a number of existing multivariate models and providing a framework for generating many more [Danaher and Smith, 2011].

- Alternative fuel vehicles have drawn increasing attention, because of their potential to reduce greenhouse-gas emissions and utilize renewable energy sources. However, alternative fuel vehicles face barriers to adoption such as lack of knowledge by potential adopters, low consumer risk tolerance, and high initial purchase costs. A number of consumer incentives for purchasing alternative fuel vehicles have been put in place to address the market barriers. In the future, it would be necessary to include the choice of the new-technology vehicle types in the vehicle ownership modeling framework to investigate the effectiveness of the policy incentives and address the solutions to overcome the market barriers.
- In regional travel modeling and simulation, the combination of the number of vehicles owned by a household, the type choice of the vehicles, and the usage of the vehicles are important travel determinants of greenhouse gas (GHG) emissions, fuel consumption, and pollutant emissions. The proposed discrete-continuous model in this dissertation provides a good basis to forecast vehicle fleet and the usage in response to changes in fuel prices, socio-economic shifts and policy decisions. Therefore, another interesting future direction is to integrate this modeling framework into the emissions/energy models (such as MOVES [EPA, MOVES,] and MOBILE6 [EPA, MOBILE6,]) in order to calculate greenhouse gas emission calculations.

Bibliography

- [Addison, 2010] Addison, J. (2010). Ten reasons for drop in car ownership. <http://www.cleanfleetreport.com/clean-fleet-articles/car-ownership-declines/>.
- [American Public Transportation Association, 2013] American Public Transportation Association (2013). Obama proposes investments for public transportation. <http://newsmanager.commpartners.com/aptaapt/issues/2013-02-22/index.html>.
- [Beggs, 1980] Beggs, S. (1980). Choice of smallest car by multi-vehicle households and the demand for electric vehicles. *Transportation Research Part A: General*, 14(5-6):389–404.
- [Berkovec, 1985] Berkovec, J. (1985). Forecasting automobile demand using disaggregate choice models. *Transportation Research Part B: Methodological*, 19(4):315–329.
- [Berkovec and Rust, 1985] Berkovec, J. and Rust, J. (1985). A nested logit model of automobile holdings for one vehicle households. *Transportation Research Part B: Methodological*, 19(4):275–285.
- [Bhat, 2005] Bhat, C. R. (2005). A multiple discrete-continuous extreme value model: Formulation and application to discretionary time-use decisions. *Transportation Research Part B*, 39(8):679–707.
- [Bhat and Eluru, 2009] Bhat, C. R. and Eluru, N. (2009). A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B-Methodological*, 43(7):749–765.
- [Bhat and Koppelman, 1993] Bhat, C. R. and Koppelman, F. S. (1993). An endogenous switching simultaneous equation system of employment, income, and car ownership. *Transportation Research Part a-Policy and Practice*, 27(6):447–459.
- [Bhat and Pulugurta, 1998] Bhat, C. R. and Pulugurta, V. (1998). A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. *Transportation Research Part B*, 32(1):61–75.
- [Bhat and Sen, 2006] Bhat, C. R. and Sen, S. (2006). Household vehicle type holdings and usage: an application of the multiple discrete-continuous extreme value (mdcev) model. *Transportation Research Part B*, 40:35–53.
- [Bhat et al., 2009] Bhat, C. R., Sen, S., and Eluru, N. (2009). The impact of demographics, built environment attributes, vehicle characteristics, and gasoline prices on household vehicle holdings and use. *Transportation Research Part B*, 43(1):1–18.

- [Bratzel, 1999] Bratzel, S. (1999). Conditions of success in sustainable urban transport policy change in relatively successful european cities. *Transport Reviews*, 19:177-190.
- [Brnnlund and Nordstrm, 2004] Brnnlund, R. and Nordstrm, J. (2004). Carbon tax simulations using a household demand model. *European Economic Review*, 48(1):211-233.
- [Bucklin and Lattin, 1991] Bucklin, R. E. and Lattin, J. M. (1991). A two-state model of purchase incidence and brand choice. *Marketing Science Marketing Science*, 10(1):24-39.
- [Bunch and Kitamura, 1990] Bunch, D. S. and Kitamura, R. (1990). Multinomial probit model estimation revisited: testing estimable model specifications, maximum likelihood algorithms and probit integral approximations for trinomial models of car ownership. Technical report, Institute of Transportation Studies Technical Report, University of California, Davis.
- [Button et al., 1993] Button, K., Ngoe, N., and Hine, J. (1993). Modelling vehicle ownership and use in low income countries. *Journal of Transport Economics and Policy*, 27(1):51-67.
- [Cao et al., 2007] Cao, X. Y., Mokhtarian, P. L., and Handy, S. L. (2007). Cross-sectional and quasi-panel explorations of the connection between the built environment and auto ownership. *Environment and Planning A*, 39(4):830-847.
- [Chandrasekharan et al., 1991] Chandrasekharan, R., McCarthy, P., and Wright, G. (1991). Models of brand loyalty in the automobile market. *Paper presented at 6th IATBR Conference, Quebec*.
- [Chiang, 1991] Chiang, J. (1991). A simultaneous approach to the whether, what and how much to buy questions. *Marketing Science Marketing Science*, 10(4):297-315.
- [Chintagunta, 1993] Chintagunta, P. K. (1993). Investigating purchase incidence, brand choice and purchase quantity decisions of households. *MARKETING SCIENCE*, 12(2):184.
- [Choo and Mokhtarian, 2004] Choo, S. and Mokhtarian, P. L. (2004). What type of vehicle do people drive? the role of attitude and lifestyle in influencing vehicle type choice. *Transportation Research Part A: Policy and Practice Transportation Research Part A: Policy and Practice*, 38(3):201-222.
- [Chu, 2002] Chu, Y. B. (2002). Automobile ownership analysis using ordered probit models. *Travel Demand and Land Use 2002: Planning and Administration*, (1805):60-67.

- [Cullinane, 1992] Cullinane, S. (1992). Attitudes towards the car in the uk: some implications for policies on congestion and the environment. *Transportation Research Part A*, 26:291301.
- [Cullinane, 2002] Cullinane, S. (2002). The relationship between car ownership and public transport provision: A case study of hong kong. *Transport Policy*, 9:2939.
- [Danaher and Smith, 2011] Danaher, P. J. and Smith, M. S. (2011). Modeling multivariate distributions using copulas: Applications in marketing. *Marketing Science*, 30(1):4–21.
- [Dargay and Gately, 1997] Dargay, J. and Gately, D. (1997). Vehicle ownership to 2015: Implications for energy use and emissions. *Energy Policy*, 25(14-15):1121–1127. doi: 10.1016/S0301-4215(97)00104-3.
- [Dargay and Gately, 1999] Dargay, J. and Gately, D. (1999). Income’s effect on car and vehicle ownership, worldwide: 1960-2015. *Sage Urban Studies Abstracts*, 27(4).
- [de Jong, 1989a] de Jong, G. C. (1989a). *Simulating car cost changes using an indirect utility model of car ownership and car use*. PTRC SAM, Brighton.
- [de Jong, 1989b] de Jong, G. C. (1989b). *Some joint models of car ownership and car use; Ph.D. thesis*. Faculty of Economic Science and Econometrics, University of Amsterdam.
- [de Jong, 1991] de Jong, G. C. (1991). An indirect utility model of car ownership and car use. *European Economic Review*, 34(5):971–985.
- [Deka, 2002] Deka, D. (2002). Transit availability and automobile ownership: Some policy implications. *Journal of Planning Education and Research*, 21:285–300.
- [Dissanayake and Morikawa, 2010] Dissanayake, D. and Morikawa, T. (2010). Investigating household vehicle ownership, mode choice and trip sharing decisions using a combined revealed preference/stated preference nested logit model: case study in bangkok metropolitan region. *Journal of Transport Geography*, 18(3):402–410.
- [Dubin and McFadden, 1984] Dubin, J. A. and McFadden, D. L. (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica*, 52(2):345–362.
- [EPA, MOBILE6,] EPA, MOBILE6. <http://www.epa.gov/otaq/m6.htm>.
- [EPA, MOVES,] EPA, MOVES. <http://www.epa.gov/otaq/models/moves/>.
- [Fang, 2008] Fang, H. A. (2008). A discretecontinuous model of households vehicle choice and usage, with an application to the effects of residential density. *Transportation Research Part B*, 42:736–758.

- [Genz, 1992] Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics Journal of Computational and Graphical Statistics*, 1(2):141–149.
- [Giblin and McNabola, 2009] Giblin, S. and McNabola, A. (2009). Modelling the impacts of a carbon emission-differentiated vehicle tax system on CO_2 emissions intensity from new vehicle purchases in Ireland. *Energy Policy*, 37(4):1404–1411.
- [Golob, 1990] Golob, T. F. (1990). The dynamics of household travel time expenditures and car ownership decisions. *Transportation Research Part A-Policy and Practice*, 24(6):443–463.
- [Golob et al., 1997] Golob, T. F., Bunch, D. S., and Brownstone, D. (1997). A vehicle use forecasting model based on revealed and stated vehicle type choice and utilisation data. *Journal of transport economics and policy.*, 31(1):69.
- [Golob and Vanwissen, 1989] Golob, T. F. and Vanwissen, L. (1989). A joint household travel distance generation and car ownership model. *Transportation Research Part B-Methodological*, 23(6):471–491. ISI Document Delivery No.: CD269 Times Cited: 13 Cited Reference Count: 47 GOLOB, TF VANWISSEN, L PERGAMON-ELSEVIER SCIENCE LTD OXFORD.
- [Goodwin, 1993] Goodwin, P. (1993). Car ownership and public transport use: Revisiting the interaction. *Transportation*, 20:21–33.
- [Guadagni and Little, 2008] Guadagni, P. M. and Little, J. D. C. (2008). A logit model of brand choice calibrated on scanner data. *Marketing Science Marketing Science*, 27(1):29–48.
- [Gunn et al., 1978] Gunn, H., Bates, J., and Roberts, M. (1978). A model of household car ownership. *Traffic Engineering and Control*.
- [Gupta, 1988] Gupta, S. (1988). Impact of sales promotions on when, what, and how much to buy. *Journal of Marketing Research*, 25:342–355.
- [Gupta, 1991] Gupta, S. (1991). Stochastic models of inter-purchase time with time dependent covariates. *Journal of Marketing Research*, 28:1–15.
- [Hanly et al., 2000] Hanly, M., Dargay, J. M., and Trb (2000). Car ownership in Great Britain - panel data analysis. *Activity Pattern Analysis and Exploration: Travel Behavior Analysis and Modeling: Planning and Administration*, (1718):83–89.
- [Hannemann, 1984] Hannemann, M. W. (1984). The discrete/continuous model of consumer demand. *Econometrica*, 52(3):542–561.
- [Hatzopoulou et al., 2001] Hatzopoulou, M., Miller, E., and Santos, B. (2001). Integrating vehicle emission modeling with activity-based travel demand modeling: Case study of the greater Toronto, Canada, area. *Transportation Research Record: Journal of the Transportation Research Board*, 2011(-1):29–39.

- [Hayashi et al., 2001] Hayashi, Y., Kato, H., and Teodoro, R. V. R. (2001). A model system for the assessment of the effects of car and fuel green taxes on co2 emission. *Transportation Research Part D: Transport and Environment*, 6(2):123–139.
- [HCG, 1989] HCG (1989). Resource papers for landelijk model, volume 2. *Hague Consulting Group*.
- [HCG, 2000] HCG (2000). 9009-3b, chapter 3: Sydney car ownership models. *Hague Consulting Group*.
- [Hensher, 1998] Hensher, D. A. (1998). The imbalance between car and public transport use in urban australia: why does it exist? *Transport Policy*, 5(4):193–204.
- [Hensher et al., 1992] Hensher, D. A., Barnard, P., Smith, N., and Milthorpe, F. (1992). *Dimensions of automobile demand; a longitudinal study of automobile ownership and use*. North-Holland, Amsterdam.
- [Hensher and Le Plastrier, 1985] Hensher, D. A. and Le Plastrier, V. (1985). Towards a dynamic discrete-choice model of household automobile fleet size and composition. *Transportation Research Part B: Methodological Transportation Research Part B: Methodological*, 19(6):481–495.
- [Hensher et al., 1981] Hensher, D. A., Manefield, T., of Economic, M. U. S., and Financial, S. (1981). *A structured-logit model of automobile acquisition and type choice*. School of Economic and Financial Studies, Macquarie University, North Ryde, N.S.W.
- [Hocherman et al., 1983] Hocherman, I., Prahsker, J., and Ben-Akiva, M. (1983). Estimation and use of dynamic transaction models of automobile ownership. *Transportation Research Record*, (944):134–141.
- [Hu et al., 2007] Hu, P. S., Reuscher, T., Schmoyer, R. L., and Chin, S.-M. (2007). Transferring 2001 national household travel survey. <http://nhts.ornl.gov/tx/TransferabilityReport.pdf>.
- [Huang, 2005] Huang, B. (2005). Car demand forecasting using dynamic pseudo panel model. *Model MVA and Department of Economics, Birkbeck College*.
- [Ingram and Liu, 1999] Ingram, G. K. and Liu, Z. (1999). Determinants of motorization and road provision. *POLICY RESEARCH WORKING PAPERS-WORLD BANK WPS*, (2042):ALL.
- [Jain and Vilcassim, 1991] Jain, D. C. and Vilcassim, N. J. (1991). Investigating household purchase timing decisions: A conditional hazard function approach. *Marketing Science Marketing Science*, 10(1):1–23.

- [Johnson, 2011] Johnson, M. (2011). Metro planners contemplate system’s second generation. <http://greatergreaterwashington.org/post/10965/metro-planners-contemplate-systems-second-generation/>.
- [Jones and Landwehr, 1988] Jones, J. M. and Landwehr, J. T. (1988). Removing heterogeneity bias from logit model estimation. *Marketing Science Marketing Science*, 7(1):41–59.
- [Jong, 1996] Jong, G. D. (1996). A disaggregate model system of vehicle holding duration, type choice and use. *Transportation Research Part B: Methodological Transportation Research Part B: Methodological*, 30(4):263–276.
- [Kamakura and Russell, 1989] Kamakura, W. A. and Russell, G. J. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, 26(4):379–390.
- [Keller, 2012] Keller, A. (2012). Creating a transit supply index. *the 12nd Transport Chicago Conference*.
- [Kim and Kim, 2004a] Kim, H. S. and Kim, E. (2004a). Effects of public transit on automobile ownership and use in households of the usa. In *RURDSThe Applied Regional Science Conference (ARSC)*, volume 16, pages 245–262.
- [Kim and Kim, 2004b] Kim, H. S. and Kim, E. (2004b). Effects of public transit on automobile ownership and use in households of the usa. *Review of Urban Regional Development Studies*, 16:245262.
- [Kitamura, 1987] Kitamura, R. (1987). A panel analysis of household car ownership and mobility, infrastructure planning and management. In *Proceedings of the Japan Society of Civil Engineers*, volume 383/IV-7, pages 13–27.
- [Kitamura, 1989] Kitamura, R. (1989). A causal analysis of car ownership and transit use. *Transportation*, 16(2):155–173.
- [Kitamura, 2009] Kitamura, R. (2009). A dynamic model system of household car ownership, trip generation, and modal split: model development and simulation experiment. *Transportation*, 36(6):711–732.
- [Kitamura and Bunch, 1992] Kitamura, R. and Bunch, D. (1992). *Heterogeneity and state dependence in household car ownership: A panel analysis using ordered-response probit models with error components*. Elsevier, Amsterdam.
- [Kitamura et al., 1999] Kitamura, R., Golob, T. F., Yamamoto, T., and Wu, G. (1999). Accessibility and auto use in a motorized metropolis.
- [Krishnamurthi and Raj, 1988] Krishnamurthi, L. and Raj, S. P. (1988). A model of brand choice and purchase quantity price sensitivities. *Marketing Science*, 7(1):1–20.

- [Lave and Train, 1979] Lave, C. A. and Train, K. (1979). A disaggregate model of auto-type choice. *TRAGj/cja:jidj Transportation Research Part A: General*, 13(1):1–9.
- [Li et al., 2010] Li, J., Walker, J. L., Srinivasan, S., and Anderson, W. P. (2010). Modeling private car ownership in china. *Transportation Research Record: Journal of the Transportation Research Board*, 2193:76–84.
- [Manning, 1986] Manning, F. (1986). Equilibrium in automobile markets using dynamic disaggregate models of vehicle demand. *Unpublished manuscript, Pennsylvania State University, University Park, Pennsylvania.*
- [Manning and Winston, 1985] Manning, F. and Winston, C. (1985). A dynamic empirical-analysis of household vehicle ownership and utilization. *Rand Journal of Economics*, 16(2):215–236. ISI Document Delivery No.: A6380 Times Cited: 117 Cited Reference Count: 27 MANNERING, F WINSTON, C RAND CORP LAWRENCE.
- [Manning et al., 2002] Manning, F., Winston, C., and Starkey, W. (2002). An exploratory analysis of automobile leasing by us households. *Journal of Urban Economics*, 52(1):154–176.
- [Manski and Sherman, 1980] Manski, C. and Sherman, L. (1980). An empirical analysis of household motor vehicle holdings. *Transportation Research A (Policy)*, 14:349–366.
- [Maryland Transit Administration,] Maryland Transit Administration. <http://www.purplelinemd.com/en/>.
- [McCarthy, 1989] McCarthy, P. (1989). Consumer valuation of new car attributes: An econometric analysis of the demand for domestic and japanese/western european imports. *Transportation Research Part A: General Transportation Research Part A: General*, 23(5):367–375.
- [McCarthy, 1985] McCarthy, P. S. (1985). An econometric analysis of automobile transactions. *Rivista internazionale di economia dei trasporti.*, 12(1).
- [McCulloch et al., 2008] McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, Hoboken, New-Jersey, second edition.
- [Neslin et al., 1985] Neslin, S. A., Henderson, C., and Quelch, J. (1985). Consumer promotions and the acceleration of product purchases. *Marketing Science Marketing Science*, 4(2):147–165.
- [P. and P., 1998] P., B. and P., J. (1998). Car ownership patterns near rapid transit stations. *Bunt Associates Engineering Ltd.*

- [Paleti et al., 2013] Paleti, R., Bhat, C. R., and Pendyala, R. M. (2013). An integrated model of residential location, work location, vehicle ownership, and commute tour characteristics. *Transportation Research Record*.
- [Parzen, 1962] Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076.
- [Potoglou and Kanaroglou, 2008] Potoglou, D. and Kanaroglou, P. S. (2008). Modelling car ownership in urban areas: a case study of hamilton, canada. *Journal of Transport Geography*, 16(1):42–54.
- [Purvis, 1994] Purvis, L. (1994). Using census public use micro data sample to estimate demographic and automobile ownership models. *Transportation Research Record*, 1443:21–30.
- [Rice, 2007] Rice, J. A. (2007). *Mathematical statistics and data analysis*. Duxbury, Belmont, California, third edition.
- [Rich and Nielsen, 2001] Rich, J. and Nielsen, O. (2001). A microeconomic model for car ownership, residence and work location. *Paper presented at European Transport Conference 2001, PTRC, Cambridge*.
- [Rood, 1998] Rood, T. (1998). The local index of transit availability: An implementation manual. *Local Government Commission, Sacramento, California*. <http://www.lgc.org/>.
- [Roorda et al., 2009] Roorda, M. J., Carrasco, J. A., and Miller, E. J. (2009). An integrated model of vehicle transactions, activity scheduling and mode choice. *Transportation Research Part B-Methodological*, 43(2):217–229. Roorda, Matthew J. Carrasco, Juan A. Miller, Eric J.
- [Rudd, 1951] Rudd, E. (1951). The relationship between the national income and vehicles registrations. *Research Note RN/1631, Road Research Laboratory, Harmondsworth*.
- [Ryan and Han, 1999] Ryan, J. and Han, G. (1999). Vehicle-ownership model using family structure and accessibility application to honolulu, hawaii. *Transportation Research Record: Journal of the Transportation Research Board*, 1676:1–10.
- [Schipper, 2011] Schipper, L. (2011). Automobile use, fuel economy and co2 emissions in industrialized countries: Encouraging trends through 2008? *Transp. Policy Transport Policy*, 18(2):358–372.
- [Schmittlein et al., 1988] Schmittlein, D. C., Helsen, K., and Wharton School. Marketing, D. (1988). *Analyzing duration times in marketing research*. Wharton School, University of Pennsylvania, Marketing Dept., Philadelphia, Pa.

- [Shay and Khattak, 2012] Shay, E. and Khattak, A. J. (2012). Household travel decision chains: Residential environment, automobile ownership, trips and mode choice. *International Journal of Sustainable Transportation*, 6(2):88–110.
- [Spissu et al., 2009] Spissu, E., Pinjari, A. R., Pendyala, R. M., and Bhat, C. R. (2009). A copula-based joint multinomial discrete-continuous model of vehicle type choice and miles of travel. *Transportation*, 36(4):403–422.
- [Tanner, 1958] Tanner, J. (1958). The an analysis of increases in motor vehicles in great britain. *Research Note RN/1631, Road Research Laboratory, Harmondsworth*.
- [Tanner, 1983] Tanner, J. (1983). International comparisons of car ownership and car usage. *Transport and Road Research Laboratory Report 1070; Department of the Environment and of Transport, Crowthorne, Berkshire*.
- [Tellis, 1987] Tellis, G. J. (1987). *Advertising exposure, loyalty, and brand purchase : a two-stage model of choice*. Marketing Science Institute, Cambridge, MA.
- [Train, 1986] Train, K. (1986). *Qualitative choice analysis : theory, econometrics, and an application to automobile demand*. MIT Press, Cambridge, Mass.
- [Train, 2009] Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, England, second edition.
- [Transportation Research Board, 2003] Transportation Research Board (2003). Transit capacity and quality of service manual, 2nd ed. tcrp project 100. *Washington, D.C.: TRB, National Research Council*. <http://onlinepubs.trb.org/onlinepubs/tcrp/tcrp100/part>
- [Transportation Research Board, 2007] Transportation Research Board (2007). *A guidebook for using american community survey data for transportation planning*.
- [U.S. Census,] U.S. Census. <http://www.census.gov/acs/www/>.
- [U.S. Census Bureau, 2013] U.S. Census Bureau (2013). *American community survey information survey*. https://www.census.gov/acs/www/Downloads/ACS_Information_Guide.pdf.
- [U.S. Department of Transportation, 2009] U.S. Department of Transportation, F. H. A. (2009). *National household travel survey*. <http://nhts.ornl.gov>.
- [Vyas et al., 2012] Vyas, G., Paleti, R., Bhat, C. R., Goulias, K. G., Pendyala, R. M., Hu, H.-H., Adler, T. J., and Bahreinian, A. (2012). Joint vehicle holdings, by type and vintage, and primary driver assignment model with application for california. *Journal of the Transportation Research Board*, 2302:74–78.

- [Washington Metropolitan Area Transit Authority, 2012a] Washington Metropolitan Area Transit Authority (2012a). Metro invests in 'better bus' service in dc, maryland, virginia. http://www.wmata.com/about_metro/news/PressReleaseDetail.cfm?ReleaseID=5233.
- [Washington Metropolitan Area Transit Authority, 2012b] Washington Metropolitan Area Transit Authority (2012b). Web page launched to gather public input on 30-year transit plan. http://www.wmata.com/about_metro/news/PressReleaseDetail.cfm?ReleaseID=4753.
- [Weisberg, 2005] Weisberg, S. (2005). Applied linear regression. John Wiley & Sons, Hoboken, New-Jersey, third edition.
- [Whelan, 2001] Whelan, G. (2001). Methodological advances in modelling and forecasting car ownership in great britain. Paper for European Transport Conference 2001, PTRC, Cambridge.
- [Whelan, 2007] Whelan, G. (2007). Modelling car ownership in great britain. Transportation Research Part a-Policy and Practice, 41(3):205–219.
- [Whelan et al., 2000] Whelan, G., Wardman, M., and Daly, A. (2000). Is there a limit to car ownership growth? an exploration of household saturation levels using two novel approaches. Transportation planning methods : proceedings of Seminar C held at the PTRC Transport and Planning Summer Annual Meeting, University of Sussex, England, from 11-15 September 1989., (445):255–264.
- [Wolff, 1938] Wolff, P. d. (1938). The demand for passenger cars in the united states. Econometrica, 6(2):113–129.
- [Xu, 2011] Xu, R. (2011). Dynamic discrete choice models for car ownership modeling. Ph.D. Thesis, University of Maryland.