EFFECTIVE GPS-BASED PANEL SURVEY SAMPLE SIZE FOR URBAN TRAVEL BEHAVIOR STUDIES

A Thesis
Submitted to
The Academic Faculty

By

Yanzhi Xu

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
May 2010

EFFECTIVE GPS-BASED PANEL SURVEY SAMPLE SIZE FOR URBAN TRAVEL BEHAVIOR STUDIES

Approved by:

Dr. Randall Guensler, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Laurie Garrow
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Michael Meyer
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Michael Rodgers
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Ming Yuan
School of Industrial and System Engineering
Georgia Institute of Technology

Date Approved: March 31, 2010

To my parents

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Randall Guensler, for his continuous guidance, encouragement, support and understanding. His dedication to quality research and devotion to students have inspired me to not only grow academically, but also grow as a person. I am also the most grateful for the invaluable advice and comments from the other members of my dissertation committee: Dr. Michael Rodgers, Dr. Laurie Garrow, Dr. Michael Meyer, and Dr. Ming Yuan. They have taught me the importance of attending to technical details while maintaining the perspectives of the "bigger picture". I would also like to thank Dr Jiawen Yang, who provided immense support and priceless suggestions at the dissertation defense.

My appreciation goes to Vetri Elango, without whose support on data processing this dissertation would have not been possible. My thanks also go to my office mates and fellow graduate students in the transportation program. Their friendship and company have filled my doctoral experience with joy. I also wish to thank the staff of transportation program: Lisa Baxter, J.J. Martino and Michael Anderson, for their invariably timely help.

I am profoundly thankful to my very soon-to-be husband, David, who has been my pillar of support. His love and patience are the sources of my strength, the very reason that I kept on going when I felt like giving up.

Finally, my gratitude towards my parents is beyond words. I am everything I am because of their selfless and unconditional love, the love that reaches around half the globe to always hold me up.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This research develops a framework to estimate the effective sample size of Global Positioning System (GPS) based panel surveys in urban travel behavior studies for a variety of planning purposes. Recent advances in GPS monitoring technologies have made it possible to implement panel surveys with lengths of weeks, months or even years. The many advantageous features of GPS-based panel surveys make such surveys attractive for travel behavior studies, but the higher cost of such surveys compared to conventional one-day or two-day paper diary surveys requires scrutiny at the sample size planning stage to ensure cost-effectiveness. Data collected from such surveys feature within-household correlations that arise from the panel design, and often non-normal distributions, both of which should be taken into account at the design and analysis stages.

The sample size analysis in this dissertation focuses on three major aspects in travel behavior studies: 1) to obtain reliable means for key travel behavior variables, 2) to conduct regression analysis on key travel behavior variables against explanatory variables such as demographic characteristics and seasonal factors, and 3) to examine impacts of a policy measure on travel behavior through before-and-after studies. The sample size analyses in this dissertation are based on the GPS data collected in the multi-year Commute Atlanta study. The sample size analysis with regard to obtaining reliable means for key travel behavior variables utilizes Monte Carlo re-sampling techniques to assess the trend of means against various sample size and survey length combinations.

The basis for the framework and methods of sample size estimation related to regression analysis and before-and-after studies are derived from various sample size procedures based on the generalized estimating equation (GEE) method. These sample size procedures have been proposed for longitudinal studies in biomedical research. This dissertation adapts these procedures to the design of panel surveys for urban travel behavior studies with the information made available from the Commute Atlanta study.

The findings from this research indicate that the required sample sizes should be much larger than the sample sizes in existing GPS-based panel surveys. This research recommends a desired range of sample sizes based on the objectives and survey lengths of urban travel behavior studies.

CHAPTER 1

INTRODUCTION

Estimating the minimum sample size is an important consideration in travel behavior studies. For conventional one-day or two-day travel surveys, sample size procedures are well known and widely applied; see, for example, the Travel Survey Manual by Cambridge Systematics (1996). The corresponding sample size procedures for Global Positioning System (GPS) based panel surveys, however, are less well developed. The many advantageous features of GPS-based panel surveys, as will be explored in this dissertation, make such surveys attractive for travel behavior studies. However, the higher cost of GPS surveys compared to conventional one-day or two-day paper diary surveys requires scrutiny at the sample size planning stage to ensure cost-effectiveness.

The essence of the problem lies in the cost of travel behavior data collection for travel behavior studies for modeling and/or policy evaluation purposes. The goal of sample size estimation is to collect data from sufficient numbers of households within well-controlled groups, where the travel behavior within each sampling group tends to be very similar to other households within the group and dissimilar to households in other control groups. The basic idea is to break up the overall population into rational subgroups which are more homogeneous, so that fairly precise estimates of parameters in the subgroups should be obtainable (Mace, 1964). However, in long-term panel surveys, the demographic characteristics that are often used to break up the overall population into subgroups tend to change over time, posing significant challenges at the design and analysis stages. In fact, the initial motivation for this work stemmed from the analysis of data collected during the Commute Atlanta Value Pricing Study where great demographic variability was observed (Xu*, et al.*, 2009b). The Commute Atlanta study was designed to assess the impact of mileage-based pricing incentives on travel behavior. The pricing experiment is implemented in different phases. In Phase I of the study from October 2004 to June 2005, the research team collected baseline travel data and travel diaries from participating households for nearly two years. In Phase II of the study from October 2005 to June 2006, the team implemented mileage-based in an effort to assess whether participating households would change their travel behavior in response to the pricing incentive. The households were given 5 cents/mile from October to December, 10 cents/mile from January to March, and 15 cents/mile from April to June. The change in household vehicle activity during the pricing element of the study was performed by comparing the before and after mileage levels in participating households for which valid baseline and pricing data were collected (95 households). The total vehicle miles of travel (VMT) for these households decreased by about 3% over the 9-month pricing period, compared to the baseline travel year. However, the noted reduction was not statistically significant due to the tremendous variability observed in household-to-household change in travel. Additionally, the travel patterns varied significantly day-to-day and month-to-month within each household. Furthermore, the demographics within these households changed considerably over time. Nearly 70% of the households experience one or more changes in their demographic characteristics between October 2004 and June 2006. Among these households, 35% experienced only vehicle ownership changes, which may or may not significantly impact travel behavior. The remainder experienced changes in household location, work location, employment status, income, household size, or other critical demographic variables, which most certainly impact travel behavior. The large variability observed in the Commute Atlanta study has raised the question about the adequate sample size for longitudinal surveys for travel response evaluations.

The primary goal of this research is develop a framework for sample size estimation that is specifically suitable for GPS-based urban panel travel behavior surveys. The key to success in this endeavor is to appropriately accommodate the longitudinal nature of the data, which will be explained and emphasized throughout this dissertation. Another challenge in sample size estimation for travel behavior studies is the violation of the normality assumption, as often seen in key travel behavior variables such as number of trips and daily or monthly VMT. This research aims to examine sample size requirements in situations where data are correlated and non-normally distributed.

The basis for the framework and methods developed in this dissertation are derived from the biomedical field. Liang and Zeger (1986) proposed the generalized estimating equation (GEE) method for the analysis of longitudinal data. Based on the GEE procedures, Liu and Liang (1997) developed algorithms to estimate sample size requirements for continuous and binary data. Later, Rochon (1998) prescribed a general methodology applicable in a wide variety of situations for continuous, binary and Poisson-distributed data.

To adapt the above mentioned procedures to travel behavior studies requires knowledge of the unique features of travel behavior data. The unprecedented multi-year continuous monitoring period of the Commute Atlanta study provides an excellent opportunity to gain such knowledge. Even though the analyses conducted in this dissertation will be limited to only the information available from a continuous small sample (95 household over two and a half years), the framework that can be developed using these data will prove useful in future panel studies.

1.1. Research Objectives

The sample size analysis in this dissertation focuses on three major uses of the data in travel behavior studies: 1) to obtain reliable means for key travel behavior variables, 2) to conduct regression analysis on key travel behavior variables against explanatory variables such as demographic characteristics and temporal factors, and 3) to examine the impacts of a policy measure and/or a change on travel behavior through before-and-after studies.

The objectives of this dissertation can be outlined as follows:

- Differentiate the between-household (cross-sectional) and within-household (longitudinal) information in household travel behavior
- Review the desirability of GPS-based panel studies for travel behavior studies including travel demand modeling and policy evaluation
- Characterize the variability of travel behavior associated with natural temporal rhythms and demographic characteristics
- Provide insight into the demographic instability within households and its implications on the design and analysis of travel behavior studies
- Examine the sampling distributions of the means of key travel behavior variables given various combinations of sample sizes and survey lengths

- Adapt GEE procedures to the analysis of GPS-based panel data for travel behavior studies
- Explore the distributional properties and correlation structures of key travel behavior variables
- Apply GEE procedures to the regression analyses of intra-regional and long-distance travel as the basis for sample size estimation for such analyses
- Develop a framework for sample size analysis in before-and-after studies in the transportation field
- Recommend a desired range of sample sizes based on objectives and lengths of longitudinal travel studies

1.2. Research Methodology

The objectives of this dissertation will be achieved by the following methodological approaches.

1.2.1. Data Collection and Processing

The analyses in this dissertation are based on data collected in the Commute Atlanta study. A detailed description of the Commute Atlanta study can be found in (Li, 2004) and (Ogle, 2005). The main objective of the Commute Atlanta study is to assess the effects of converting fixed automotive costs into variable driving costs. At the beginning of the study in 2003, the research team installed 487 GPS trip collectors in the vehicles of 268 participating households to collect second-by-second vehicle activity data. The Commute Atlanta study included the parallel collection of instrumented vehicle data, household socio-demographic surveys, and employer commute options surveys.

This dissertation utilizes the trip data collected from 95 households that stayed in the study from January 2004 to June 2006. In addition to the trip information collected by the in-vehicle GPS devices, this dissertation derives detailed household demographic information from the combination of household mail-in surveys and visualized household travel patterns such as trip-end frequencies. Households do not always report changes in their demographic information promptly, if at all, so the use of visualized travel patterns proves very useful in detecting changes within a household. The information about changes in demographic characteristics in turn benefits the analysis and understanding of household travel behavior. In travel demand forecasting, understanding the association between changes in household characteristics and changes in travel behavior can help planners predict long-term travel trends in the context of overall demographic trends such as population aging. In policy evaluation, household demographic changes need to be controlled for to establish a causal relationship between policy measures and changes in travel behavior. The roles of household demographic changes in travel demand forecasting and before-and-after policy studies will be discussed in detail in the subsequent chapters.

1.2.2. Descriptive Analysis

This research conducts extensive descriptive analysis of intra-regional and long-distance travel to provide insight into the longitudinal nature of panel data. The descriptive analysis will differentiate between-household and within-household information. The analysis is carried out in three aspects: 1) the general association between travel behavior and temporal and demographic factors, 2) the demographic changes that arise in a panel setting and their impacts on travel behavior, and 3) the trend of means as sample size and survey length vary. The bootstrap technique (Efron and Tibshirani, 1993; Davison and Hinkley, 1997) is applied widely in the descriptive analysis to accommodate the non-normal distribution of the data.

1.2.3. Generalized Estimating Equation (GEE) Procedures

The GEE approach is the formal statistical approach this dissertation adopts for relating sample size requirements to regression analysis and before-and-after studies. To implement the GEE procedures, the data distributions and correlation structures will be examined. The sample size estimation related to regression analysis is carried out by examining the changes in the significance of regression coefficients as sample size and survey length vary. The sample size estimation related to before-and-after studies is implemented by adapting existing algorithms that are applied in biomedical research to travel behavior studies.

1.3. Research Contributions

This research will fill a few gaps in the literature. First, with the growing popularity of longitudinal travel surveys, there is not much attention dedicated to sample size estimation. The researchers and practitioners tend to adopt sample size estimation methods suited for cross-sectional surveys and use these methods in panel surveys. This research will discuss and emphasize the significant difference in sample size requirements between traditional cross-sectional surveys and GPS-based panel surveys.

Second, the capabilities of GPS-based panel surveys to evaluate various transportation policies have not been fully explored. The transportation community has not characterized differences between the longitudinal and cross-sectional information available from panel data. This research will adapt GEE procedures which are widely applied in biomedical research to the analysis of longitudinal travel data.

Third, the magnitude of variability in the variables of interest (e.g. number of trips, VMT, etc) is also often overlooked in the literature. The temporal aspect of GPS-based panel surveys adds significant variability to what has been observed in cross-sectional data. The added variability mainly comes from the natural temporal rhythms of travel and the dynamic changes in households. This research will examine the patterns of variability, with special scrutiny to handle changes in confounding factors.

The sample size estimation methodology that the research aims to develop will take into account the policy objectives that a survey serves and the variability unique to longitudinal surveys resulting from temporal rhythms and dynamic changes. The estimated sample size requirement will be a function of study objectives and variability patterns.

To summarize, this research anticipates the following contributions:

- Assess the advantages of longitudinal travel behavior studies over traditional cross-sectional studies for travel behavior evaluation
- Better understand how intrinsic variability in the travel attributes of interest impacts transportation policy assessment
- Depict the sampling distributions of means for key travel behavior variables given a longitudinal design
- Develop a methodology to estimate sample size and the length of study combinations for longitudinal travel studies with regard to regression analysis

- Formulate guidelines for sample sizes pertaining to policy objectives of before-and-after studies

Even though the guidelines resulting from this research may have limited transferability to other metropolitan areas than Atlanta, the methodology will be applicable to a more generalized setting. The methodologies and results from this study should be further validated and updated with data from other regions as such data become available.

1.4. Dissertation Outline

Following this introductory chapter, Chapter 2 summarizes the literature that characterizes GPS-based panel surveys, reviews the desirability of such surveys in travel behavior studies, and states the need for comprehensive research on sample size requirements for such surveys. Chapter 3 describes the data collection and processing efforts of this dissertation. Chapter 4 conducts exploratory analysis that reveals the association between travel behavior and temporal and demographic factors. Chapter 5 elaborates on the impact of demographic characteristics on travel behavior with special attention to demographic instability. Chapter 6 examines the trends of means of key travel behavior variables with varying sample sizes and survey lengths. Together, Chapters 4 to 6 provide descriptive analysis on the general information of travel behavior made available by panel data. Chapter 7 begins to set the stage for the GEE procedures that will be applied in the rest of the dissertation by formally introducing the statistical approach. Chapter 8 explores the distributional properties and correlation structures of travel data as a preparation for implementing the GEE procedures. Chapter 9 relates sample size analysis to regression analysis. Chapter 10 develops a framework for before-and-after studies in the context of policy evaluation. Finally, Chapter 11 summarizes the research findings and suggests directions for future research.

CHAPTER 2

LITERATURE REVIEW

This section will first review the definitions, examples, and advantages of GPS-based panel surveys. Then the potential applications of such surveys in various aspects of transportation planning, including activity-based models and transportation policy studies, will be discussed at length. The final part of this section documents the need for further research on sample size requirements based upon the limited sample size studies that currently exist in the transportation field.

2.1. GPS-Based Panel Surveys

2.1.1. Definitions

Panel surveys can be defined as surveys of the same survey units at different times, measuring the same characteristics (Markus, 1979). In the literature, the terms "panel" and "longitudinal" are often used interchangeably. To clarify the exact definition of the surveys that this research is going to investigate, it is worthwhile to first compare panel data that arise from a panel survey to two other types of data: cross-sectional and time series.

The difference between a panel survey and a cross-sectional survey is that a panel survey has at least two repetitions on the same survey subjects - i.e. households, in most cases - whereas a cross-sectional survey conducts the survey only one time. It is also worthwhile to compare the data generated from a panel survey, namely, panel data, to a time series. With respect to a time series, observations are usually taken on a single subject (household, individual, vehicle, etc) at a relatively large number of time points. Panel data include observations on many subjects at a relatively small number of time points (Markus, 1979). Moreover, the unit of analysis is the time point in time series analysis (Ostrom, 1978), while the unit of analysis is the household, individual, or vehicle in panel analysis. Table 2.1 summarizes the differences and similarities among these three types of data.

Table 2.1 Comparison of Panel Data, Cross-Sectional Data, and Time Series

	Unit of Analysis	Number of Subjects	Number of Repetitions
Panel data	Household/individual/vehicle	Many	Relatively small
Cross-sectional data	Household/individual/vehicle	Many	One
Time series	Time point	One	Relatively large

Continuous GPS-based panel surveys do not fall under the above strict definition of panel surveys. Strictly speaking, panel surveys observe a few discrete points in time, as shown in Figure 2.1b. Each survey time point is referred to as a “wave” (Yee and Niemeier, 1996). However, GPS-based panel surveys often monitor travel behavior continuously, as shown in Figure 2.1c. In this case, it is useful to conceive of each wave as a continuous multi-day or multi-month survey. From the data analysis point of view, the entire panel data can be conceived of as a large number of time series - one for each household (Markus, 1979). In biomedical research, panel studies are often referred to as longitudinal studies (Diggle, *et al.*, 2002). In this dissertation, the terms “panel” and “longitudinal” will be used interchangeably.

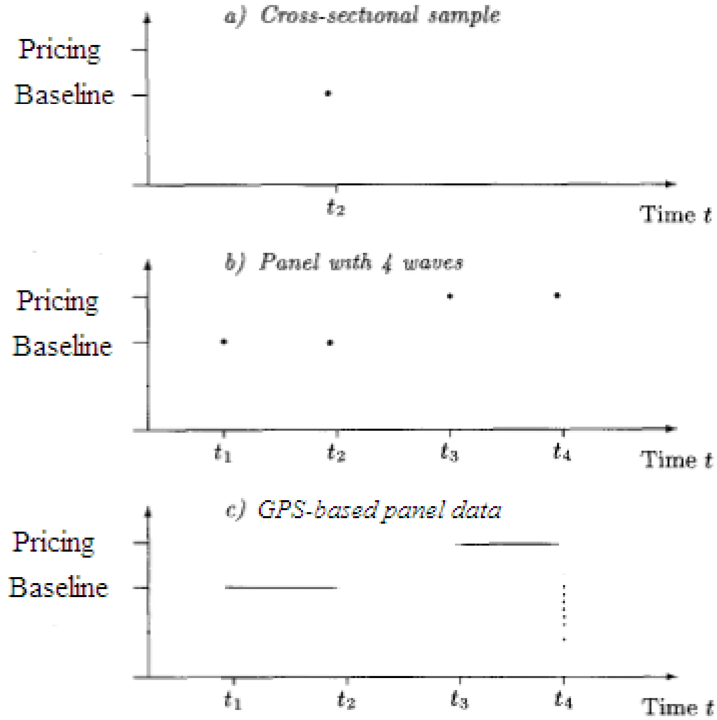

Figure 2.1 Comparison of Cross-Sectional Data, Traditional Panel Data, and GPS-Based Panel Data
Adapted from Blossfeld and Rohwer (2002)

In GPS survey efforts, researchers monitor trip origins, destinations, and real-time vehicle position using GPS tracking devices and collect supplemental travel surveys data. This research utilizes travel data collected through passive in-vehicle monitoring technology, where position data are collected by GPS systems without any information input by the drivers (Lee-Gosselin, 2002). GPS-based surveys are useful for exact time and destination recording and capture of trip underreporting (Wolf, *et al.*, 2003; Schönfelder, *et al.*, 2006).

2.1.2. Examples

This section focuses on the panel studies in the transportation field that are designed to evaluated mileage-based value pricing policies. These examples are panel surveys with multi-day observation periods using in-vehicle tracking technologies.

- Minnesota Mileage-Based User Fee Demonstration Project - This "Pay-As-You-Drive" (PAYD) pilot project simulated the conversion of vehicle lease and/or insurance pricing from traditional fixed payments to payments based on actual miles driven. In this study, personal vehicles of 130 participants had car chip technology installed for around one year. The car chip measures time, speed, and distance of travel, but is unable to track locations. The results from this project support the notion that some drivers will reduce mileage in response to price signals, although the range of responses, variability of the data, small sample size, short experiment period, and lack of negative consequences make it difficult to come to definitive conclusions (Buxbaum, 2008). To account for the small sample size, the study adopts a disaggregate method using a matching method that matched members of the treatment group to those of the control group based on the probability of participation in the experiment and their baseline mileage (Abou-Zeid, *et al.*, 2008). However, given the large variability in demographic characteristics across households, it is unlikely that the matching method could ensure meaningful comparisons of travel behavior across households.

- The Oregon Road User Fee Pilot Program - The pilot test is designed to demonstrate the technical and administrative feasibility of implementing an electronic collection system for mileage-based user fees and congestion tolls. In the spring of 2006, 260 trial participants in Portland, Oregon, had the on-board equipment added to their vehicles. For a period of one year, participants paid mileage-based charges rather than the fuels tax. When the participants filled up at gas stations, the fuel tax was deducted from the bill and the mileage charge was added. The charges were calculated according to VMT by zone and by time using in-vehicle GPS-based tracking device. The study of the results from this program reported reduction in VMT in response to price differential (Rufolo and Kimpel, 2008). However, the small sample size, especially the control group of only 27 vehicles, does not lead to the strong conclusions offered by this paper. The study did not investigate factors such as household demographic characteristics, seasonal effects and gas prices before drawing conclusions.

- Washington Traffic Choice Study - In this pilot study, in-vehicle GPS-based tolling devices with cellular communication capabilities were placed in the vehicles of voluntary participants. Different prices per mile were imposed depending upon the location and time of travel. More than 400 vehicles from 275 plus households participated in the study for up to 18 months. The preliminary results show a 7% reduction in total VMT of all participants (Puget Sound Regional Council, 2008). Similar to other projects mentioned above, this preliminary analysis did not factor in the demographic characteristics of the participants nor gas price fluctuations.

In reviewing the studies reported above, it becomes clear that it is difficult to separate the impact of the mileage-based pricing policy from the large variability in household travel behavior, when analyses are performed at the aggregate level. The impact of households' demographics, including their changes, and other exogenous factors that impact travel choices were seldom addressed in the literature. The aforementioned studies suffer from the large amount of variability and the arguably small sample sizes, therefore failing to achieve the objective of policy evaluation.

The magnitude of variability in household travel behavior revealed from the Commute Atlanta study (Xu*, et al.*, 2009a) further indicates that that much larger sample sizes and improved survey design will be required in longitudinal studies to ascertain how pricing affects travel behavior. There is both natural within-household variability as well as some potentially extreme variability resulting from demographic changes during panel studies. Significantly larger samples (much larger numbers of participating households within each demographic stratum, and much larger samples than those seen in other relevant studies reported in the literature) need to be coupled with more-intensive continuous surveys. Experimental controls over households that use their vehicles for business purposes also need to be introduced, possibly as a separate recruitment stratum (Elango*, et al.*, 2007). Accessibility to viable commuter transit is also an essential control variable in future studies, as households that have viable transit access maybe much more likely to respond to economic incentives.

2.1.3. Advantages of GPS-Based Surveys

2.1.3.1. Trip Under-Reporting and Non-Response

GPS-based surveys can recover unreported trips and activities, as the GPS devices automatically log all routes. Previous studies have documented the benefits of GPS-based surveys in revealing under-reported activities and trips; for example, (Wolf *et al.*, 2003; Zmud and Wolf, 2003; Bricka and Bhat, 2008). As Ogle, *et al.* (2006) pointed out, in paper-based travel diaries, participants under report trips and activities by as much as 30%, either because they forget some trips, or because they intentionally do so to shorten the time to fill out the diary or to answer the interview.

GPS-based surveys have also revealed the non-response issue in travel surveys. The Commute Atlanta study has revealed that households that have more than 16 trips per day or ones that have no trips at all tend to be more likely not to respond the survey (Ogle*, et al.*, 2005), indicating that the omitted trips are also not random in nature, i.e. there appears to be systematic under-reporting by trip making patterns.

2.1.3.2. Extended Monitoring Period

GPS-based surveys can provide continuous observations of over a week, a month, or even longer (Zumkeller*, et al.*, 2006). Several pilot GPS-based travel studies all have continuous monitoring periods ranging from one to two years. Table 2.2 summarizes the lengths of monitoring period of GPS-based surveys used in policy studies in the US. The extended monitoring period enabled by the GPS technology will benefit travel demand modeling in numerous ways as discussed below.

Table 2.2 Long-Term GPS-Based Surveys for Policy Analysis

Location	Start Year	Continuous Monitoring Period	Technology Type	Sample Size	Policy Areas
Atlanta, GA (Xu, *et al.*, 2009a)	2004	Two and a half years	In-vehicle	268 households recruited, 95 usable for pricing study	Mileage-based pricing, safety, emissions
Twin Cities, MN (Abou-Zeid, *et al.*, 2008)	2004	One year	In-vehicle	130 households	Mileage-based pricing
Oregon (Rufolo and Kimpel, 2008)	2006	One year	In-vehicle	201 households recruited, 168 households and 207 vehicles usable	Mileage-based pricing
Puget Sound, WA (Puget Sound Regional Council, 2008)	2002	Eighteen months	In-vehicle	275 households, 400 vehicles	Variable road tolling

Potential to Decrease Sample Size

Longitudinal surveys, in which each participant is monitored for an extended period of time, can provide very useful information with a smaller sample size than cross-sectional surveys, thanks to the greater statistical reliability that these surveys can achieve (Moser and Kalton, 1971). Even with an equivalent number of respondents, longitudinal surveys may prove cheaper than cross-sectional surveys in the long run, even though initially these surveys are a higher resource undertaking (Murakami, *et al.*, 2006). For example, Stopher, *et al.* (2008b) examined the sample size benefits of using a multi-day survey to measure distance traveled, number of trips and total travel time. However, such potential to reduce overall cost has to be examined rigorously, due to the much higher cost per correspondent, including equipment, data storage and processing, and statistical analysis. As well explained by Stopher (2008a), as the length of continuous monitoring period increases, the variability in observed travel patterns increases, requiring a larger sample size to allow good destination choice models to be developed. Because there is little experience to date with using multi-day data for model estimation, the sample size estimation is as yet an issue that is open for new research and exploration.

Potential to Identify Needed Increases in Sample Strata

Because GPS-based survey data provide more accurate and detailed information for an extended period, they enable researchers to gain insights into variability in household travel behavior and into changes in travel behavior over time that are not available in a conventional cross-sectional paper-based diary survey. The unique travel patterns of certain demographic groups, revealed through GPS-based survey data, can identify the need to add sample strata and/or include more variables during population synthesis in activity-based models. For example, based on the GPS data collected in the Commute Atlanta study, Elango, *et al.* (2007) noted that vehicles identified by participants as being used "always" or "occasionally" for business/commercial purposes undertake very different travel patterns than other vehicles. Households that include commercial vehicles also tend to have a much higher trip rates than those without, all other major demographic characteristics being equal. Therefore, without special weighting procedures, the presence of commercial-use vehicles in the sample can significantly bias analytical results (and most travel diary surveys do not currently explicitly control for the presence of such vehicles). The Georgia Tech researchers have previously argued that households with commercial vehicles must be treated as an independent sample strata in future travel diary data collection and travel demand analysis (Elango, *et al.*, 2007).

Benefits for Travel Demand Analysis

Extended monitoring periods can also improve travel demand analysis. Intuitively, population aging, demographic changes and temporal variability influence travel behavior, but these factors are often inadequately addressed in travel demand models due to insufficient data. To establish causality and estimating pricing elasticity, policy evaluation studies also need to include travel surveys with extended monitoring periods.

2.1.3.3. Improved Accuracy

In conventional paper-based travel diary surveys, respondents have to remember the locations and times of trip origins and destinations. In GPS-based surveys, however, trip origin, destination and route data are automatically collected without burdening the respondent (Kochan*, et al.*, 2008). In a recent demonstration of GPS-based travel diary surveys conducted by the University of Minnesota and Vehicle Monitoring Technologies, Inc., 43 university commuters completed travel diaries at a 94% completion rate, and even provided travel data for more than twice the number of days as requested, presumably because the surveys were interesting to complete (Elango and Guensler, 2010). Therefore, GPS-based surveys can be used to gather much more exact activity durations, and data for every segment of a vehicle tour. Such information is crucial for activity-based models. Furthermore, the GPS tracking technology has provided the possibility to analyze route choices and trip-chaining behaviors that have never been available before (Stopher, 2008b). For example, Li et al. (Li*, et al.*, 2005) examined morning commute route choice patterns using GPS data and found a strong relationship between the morning commute trip-chaining decision (single vs. multiple routes) and work schedule flexibility as well as commuters' socio-demographic characteristics and commute route attributes.

2.2. Potential Applications of GPS-Based Panel Surveys in Travel Demand Forecasting

2.2.1. Activity-Based Models

In recent years, MPOs in the US have demonstrated increasing interest in moving from conventional four-step models towards activity-based models. Currently, San Francisco, New York, Sacramento, and Lake Tahoe have developed activity-based models, while such models for Portland, Atlanta, and Denver are under development. Many more cities, such as Boston, are planning to convert their four-step models to activity-based models in the near future. This section examines existing urban activity-based models in the US, including: Portland (METRO), San Francisco (SFCTA), New York Best Practice Model, Columbus (MORPC), Atlanta (ARC), San Francisco Bay Area (MTC), Denver (DRCOG), and Lake Tahoe. First, the data sources for these models are compiled, revealing that the MPOs have not taken full advantage of GPS-based panel data. Next, the features of these activity-based models are summarized, indicating the need for better resolution in survey data for the purpose of establishing long-term causal relationships. Lastly, the aspects in which GPS-based panel surveys can contribute to improving activity-based models are summarized.

2.2.1.1. Data Sources

Household travel surveys serve as the basis for model estimation. Table 2.3 summarizes the travel survey data that support current activity-based models. Figure 2.2

zooms in on the months of the year that each survey covered. A few patterns emerge from the table and the chart.

Table 2.3 Household Travel Surveys that Support Activity-Based Models

	Portland	San Francisco	New York	Columbus	Atlanta	Sacramento	Bay Area	Denver	Lake Tahoe
Survey Year	1994-1995	1990	1997-1998	1999	2001-2002	2000	2000-2001	1997	2005
Number of Days Monitored	2	1, 3, 5	1	1	2	1	2	1	1
Number of Households Sampled	5,000	1,500*	11,264	5,433	8,609	3,492	15,000	4,196	1,220
Base Year Population (million)	1.4	0.75*	11[1]	1.5	4.7	2	6.8	2.2	0.63
Survey Instrument	Diary	Trip memory jogger	Diary	Diary	Diary	Diary	Diary	Diary	Diary
Weekend Travel	Yes, entire sample	No	275 house-holds	No	Yes, entire sample	No	Yes, entire sample	No	No
GPS Component	No	No	No	No	No	No	No	No	No

* SF residents only

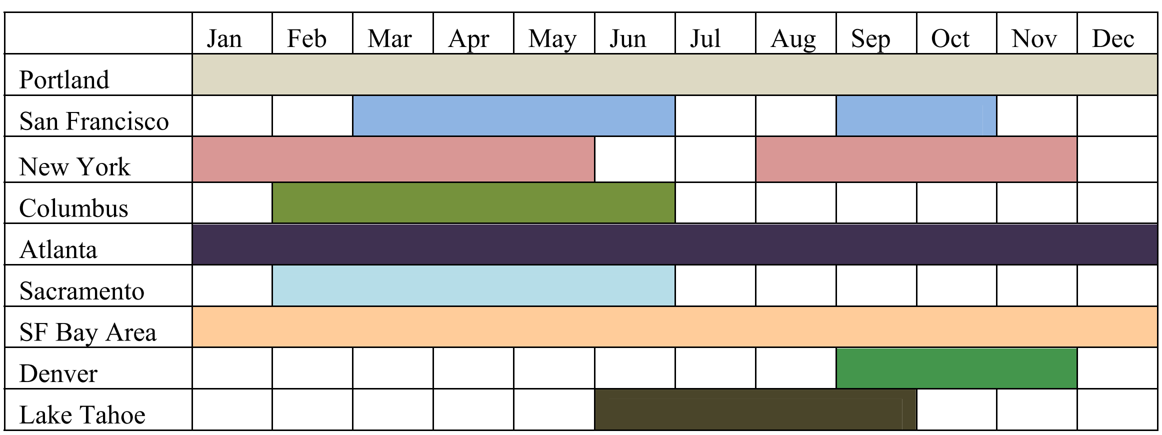

Figure 2.2 Months of a Year the Surveys Covered

[1] Not documented in survey reports. This number is an approximation by the population of the New York Metropolitan Transportation Council region published on http://www.nymtc.org/.

The surveys employed in the development of current activity-based models were primarily conducted in the 1990s and do not include a GPS component. The travel surveys have very limited continuous monitoring period. The longest is the 1990 San Francisco Bay Area travel behavior study that has three and five-day surveys. All other surveys only have one or two-day monitoring periods.

Five out of the nine surveys do not include weekend travel. Among the four surveys that did include weekend travel, the New York survey restricted the weekend sample to 275 households in the North Jersey Transportation Planning Authority (NJTPA) counties of northern New Jersey (Parsons Brinckerhoff Quade & Douglas, 2000).

Only the Portland, Atlanta, and San Francisco Bay Area surveys covered an entire year. Therefore, surveys in the other regions cannot reflect seasonal fluctuations of travel behavior. Furthermore, even though these three surveys lasted for a year or more, none of them tracked the same households continuously for that period. Hence, even these data cannot be used to isolate seasonal fluctuations in travel behavior as a function of demographic variability. Understanding this variability is important in assessing policies that affect general rather than one-day behavior, e.g. to assess the distribution of user charges for road pricing, or patterns of public transportation usage (Jones and Clarke, 1988).

2.2.1.2. Features of Activity-Based Models

Current activity-based models consist of the following structure, with a hierarchy of levels from "top" to "bottom", where the lower-level choices predicted are conditional on higher-level choices (Bradley and Bowman, 2006; Davidson, *et al.*, 2007):

- Long-term level
  - Population synthesis
  - Auto ownership
  - Work and school locations
- Person/household-day level: number of tours and activities by purpose
- Tour-level
  - Main destination and mode
  - Begin/end times
  - Number of stops
- Trip-level
  - Intermediate stop location
  - The mode and departure time of each trip

Bradley and Bowman (2006) and Cambridge Systematics (2008) provide detailed summaries of the technical information of these models, including the model structure and variables included in each model component. PB Consult (2005) compared the different features between trip-based models and activity-based models. These reviews reveal that, compared to conventional trip-based models, activity-based models forecast travel behavior at a higher resolution for several aspects. Arguably, travel survey data need to provide high resolution observations with respect to the corresponding aspects. The areas in which activity-based models provide more details include:

- Improved Use of Household Demographic Characteristics

  Activity-based models predict basic travel decisions in a more disaggregate fashion than trip-based models. The demographic variables in aggregate trip-based models are limited to the predetermined dimensions that were used in the functional model segmentation (Davidson*, et al.*, 2007). With explicit micro-simulation of synthetic households and persons, activity-based models allow for the analyses of detailed population subgroups, such as ethnic groups or people with disabilities, and demographic changes, such as employment status and life-cycle stage changes (provided that the synthetic household population accurately reflects the regional demographic structures and inter-relationships of demographic variables).

- Spatial Structure

  In activity-based models, the micro-simulation framework is not tied as strongly to zone definitions, making it possible to specify variables related to land use, parking, and walk access (which do not need to be stored as matrices) at a finer level. According to Bradley and Bowman (2006), the Portland model uses such an approach for roughly 20,000 "blocks", while the Sacramento models employ more than 700,000 parcels.

- Time-of-Day Modeling

  Most conventional trip-based models cannot incorporate disaggregate time-of-day travel decisions unless they are calibrated for specific time periods. Most trip-based models predict daily trips, and the more refined models only use peak and off-peak periods. For example, the current trip-based model in Atlanta divides a day into four time periods. The activity-based models, however, have the capability of modeling travel behavior at a resolution of 30-minute intervals, e.g. the Sacramento, CA model (Kuppam*, et al.*, 2008). The advantages of the activity-based approach to modeling time-of-day choice decisions are analyzed in detail by Vovsha and Bradley (2004; 2006), and Vovsha, *et al.* (2005). The main constraint on how small the time periods can be is the sample size and adequacy of the self-reported times in the diary survey data, because there is evidence that people often round clock times to 10, 15 or 30 minute intervals (Bradley and Bowman, 2006).

- Trip-Chaining Behavior

  Trip-based models often do not model trip-chaining behavior (PB Consult Inc. and The Gallop Corporation, 2005). Because activity-based models conceptualize the tour as the analytical unit to construct daily activity patterns, trip-chain modeling is essential to these more advanced models. The accuracy and consistency in trip-chain modeling may improve overall model advancement (Ye*, et al.*, 2007).

- Household Joint Activity and Travel

  Some activity-based models (e.g. Columbus, Atlanta, and San Francisco Bay Area) explicitly treat the linkages between the predicted activities and travel of different household members. Such treatment takes full advantage of the micro-simulation approach in terms of its ability to reduce aggregation bias (Bradley and Bowman, 2006). A constraint on this modeling capability is the under-reporting of short trips and very young passengers (less than 5 years old), a phenomenon often seen in conventional household travel surveys (Davidson*, et al.*, 2007).

Current activity-based models also have room for enhancement, in that there are more features to be included in the models to improve model estimates. One of the major concerns is that current activity-based models do not adequately reflect the temporal variability of travel behavior. For example, the models do not include day of week as a variable. Very few of the models even model weekend travel. Similarly, seasonal variability of travel behavior is not reflected either. Additionally, none of the models documented the impact of special events on travel patterns. In a longer-term sense, household demographics change over time, but seldom do current activity-based models address longer term life-cycle stage variability - for example, gentrification of large neighborhoods, and population aging - in the future scenario population synthesis. Previous research has also suggested that the disaggregate modeling approach of activity-based models can model the impact of population aging on travel behavior change, but the implementation of this capability is not performed.

2.2.1.3. Using GPS-Based Panel Surveys to Improve Activity-Based Models

Accounting for Day-to-day Travel Variability

Previous studies have shown the significance of day-to-day variability in household travel behavior, and hence, the desirability of longitudinal survey data. For example, Jones and Clarke (1988) drew on studies in the UK and Australia and raised the policy and analytical rationale for using multi-day data. More recently, with the availability of GPS technologies, researchers have gained better understanding of day-to-day variability through the analysis of multi-day GPS data. For example, using GPS data collected for the Lexington pilot study, Pendyala (1999) reported that the intrapersonal variability in number of trips for a 3-day weekday sample was 49%. Stopher, *et al.* (2007) and Stopher, *et al.* (2008a) examined intrapersonal variability based on three waves of 28, 21, and 15 days of GPS data. They argued that a survey length of 15 days is optimal, based on evidence from the surveys in Australia.

The use of GPS devices extends greatly the potential duration of monitoring periods to multiple weeks, months or even years. The Commute Atlanta study spans an unprecedented length of two and a half years. The length of this continuous monitoring survey has provided an excellent opportunity to examine day-to-day variability in household travel behavior and its implications on household travel surveys and travel demand modeling. Li, *et al.* (2005) examined the morning commute route choice behavior of 182 drivers over a ten-day period and found significant correlation between the morning commute trip-chaining decision (single vs. multiple routes) and work schedule flexibility, commuters' socio-demographic characteristics and commute route attributes. Schönfelder , *et al.* (2006) provided evidence from the Commute Atlanta study that the monitoring period should last for about five to ten weeks of monitoring to gain some certainty about individual choice preference. Elango, Guensler et al. (2007) analyzed the variability using 12 months of data from this study and reported that households that have higher income, more members, more vehicles, children and students display more day-to-day variability, which arises from seasonal, temporal, and non-habitual activities. Long-term multi-day travel surveys are also able to capture the relatively rare events of multi-day journeys (Li*, et al.*, 2007), special events (e.g. ball games and construction) and even stay-at-home, zero-trip days.

One might argue that a cross-sectional survey spanning multiple years can reflect day-to-day variability of travel behavior. However, current cross-sectional surveys cannot separate day-to-day variability from demographic differences because these surveys employ different households at different points in time. Xu, *et al.* (2009a) analyzed the travel patterns of 95 households on a case-by-case basis in the Commute Atlanta Value Pricing Study, and found considerable variability across households even within the same demographic stratum defined by income, household structure and vehicle ownership. The variability with respect to time observed in multi-day data makes it possible to partition the total variability into between-household and within-household components.

Demographic Changes and Population Aging

The potential impact of changing demographics is at the heart of the uncertainty associated with assessment of whether a policy intervention in the transportation system played a role in changing a household's travel patterns. Potential impacts of transportation policies are overlaid upon the impacts of background changes (e.g. changes in gasoline price, changes in congestion levels, changes in parking pricing policies, changes in transit services, etc.). Furthermore, changes in household demographic characteristics are likely to influence travel behavior significantly more than the variety of policies being assessed; see, for example (Xu*, et al.*, 2009b). Furthermore, these changes may have very different impacts on different households as a function of their original and final demographic characteristics.

Explicit micro-simulation of synthetic households and persons, activity-based models can allow for testing the impacts of demographic changes (Davidson*, et al.*, 2007). However, the reliable realization of this capability depends on the support of accurate and detailed underlying survey data. To model causality and the magnitude of behavioral changes, conventional cross-sectional travel surveys fall short on the ability to single out variability across demographic groups from the changes within a household. It is also not practical to employ paper diaries to in longitudinal survey efforts, because of the increased respondent burden.

GPS tracking technologies make it possible to monitor household travel behavior for extended periods. The extended monitoring period can be used in measuring the effects of such changes. Xu, *et al.* (2009b) documented that over a two-and-a-half-year observing period, the impact of work status changes was most discernible among the demographic changes. Home location changes and household structure changes are also important sources of VMT change. This finding would have not been possible without the continuous monitoring period before and after these demographic changes.

2.2.2. Transportation Policy Studies

Many transportation policy studies can benefit from long-term GPS-based panel surveys to establish causal relationships. The discussion below focuses on pricing studies, but the implications apply to other policy studies that need to establish cause-and-effect relationships between household travel behavior and policy measures.

Kuppam, *et al.* (2008) identified two issues to improve existing travel models for pricing studies: 1) inaccurate values of time due to aggregation biases, and 2) lack of temporal detail and behavioral choice for time-of-day models. These issues can be attributed to three major data deficiencies:

1. Lack of Disaggregate Detail - Sufficient detailed data are not available across individual travelers, vehicle types, trip purposes, travel modes, destination choices, trip distance, types of congestion, and road types. Traditional travel diaries provide limited information with respect to such segmentation.

    - Traveler and Vehicle Types - Traditional travel surveys are cross-sectional. That is, traditional surveys only provide a snapshot of household demographics and their observed travel behavior. However, previous research has provided evidence that household demographic characteristics and vehicle ownerships change rapidly over time. The Commute Atlanta study found out that more than 70% of the sampled households experienced changes between October 2004 and June 2006 in one or more of the major demographic characteristics of home location, employment status, income, household structure, schools attended, and vehicle ownership (Xu*, et al.*, 2009b). Without a longer monitoring period, the surveys cannot provide accurate information on the extent to which travel patterns correlate over time within a given household versus the impact of demographic changes.
    - Trip Purpose, Mode, and Destination - Values of time differ across trip purposes, modes, and destinations. Therefore, an accurate representation of such trip information is crucial to pricing studies. However, in traditional paper-based travel diaries, participants tend to under-report

trips, especially short, discretionary and non-home-based trips (Pierce*, et al.*, 2003; Ogle*, et al.*, 2005). This issue not only undermines the accuracy of model estimates, but also results in unreliable values of time, and hence inconsistent evaluation of pricing policies.

- Trip Distance, Congestion Level, and Road Type - Traditional paper-based diaries collect very limited information on route choice. Therefore, even though the surveys report origins and destinations, it is very difficult to calculate the actual trip distance. For the same reason, the road types the participants travel on are unknown because of the lacking route choice information. Furthermore, it is practically impossible for traditional paper-based travel diaries to reflect travel speed. Thus modelers do not know the severity of congestion the traveler was experiencing, or whether the congestion was recurrent or non-recurrent. Therefore, paper-based travel diary surveys do not collect the types and severity of congestion, whereas GPS-based surveys do collect these data

2. Lack of Temporal Activity Detail - Details of activity and trip scheduling and peak spreading behaviors are not available in cross-sectional, short-duration, traditional travel diary methods. The study of time-of-day policy measures require that survey data provide:

   - Complete Picture of Household Activity and Trip Scheduling Behavior - Due to trip under-reporting, the daily patterns that conventional household surveys portrait contain gaps where parts of a tour are frequently missing.

- High-Resolution of Departure/Arrival Times and Activity Duration - Participants often round up departure/arrival times to five-minute or even 15-minute increments. Activity duration is often a derived value from the arrival and departure times reported in the surveys, and is therefore often erroneous (Davidson*, et al.*, 2007). On the other hand, trip data collected by GPS devices provide accurate time stamps.

3. Lack of Before-and-after Studies to Validate the Effects of Policy Changes - This is arguably the most important drawback of conventional paper-based travel surveys. For the models to accurately evaluate pricing policies, there are two fundamental questions that survey data need to be able to answer: the causal link between travel behavior changes and pricing policies, and, if the causal link can be established, the magnitude of changes. It is well-understood in social sciences and biomedical research that longitudinal studies of the same sample is the only way to explicitly model causality (Diggle*, et al.*, 2002; Fitzmaurice*, et al.*, 2004), but as summarized in Table 2.3, the surveys supporting activity-based models are all cross-sectional in nature. Therefore, the causality and magnitude of the causal effects estimated from these survey data are at best presumptuous.

The lack of longitudinal design of the surveys can also explain why some intuitively important variables, such as trip distance, time of day, gender, age, etc., only have marginal significance on model output of values of time in pricing studies, as reported in some pricing studies using time-of-day choice models, e.g. (Kuppam*, et al.*, 2005). Longitudinal surveys are the only method to dynamically track such changes

(Kitamura, 1990). Because the temporal consequences of events are known, longitudinal data make it possible to separate the impacts of a background change, such as a change in the household demographic characteristics and an increase in gas prices, from the impacts of an intervention in the transportation system. Longitudinal surveys therefore allow for causal analyses at the household level.

2.3. Need for Research on Sample Size Requirement

Studies of the sample size requirements of longitudinal surveys are few. Among them are Stopher and Greaves (2007), which illustrated the issue using a panel to measure the change in vehicle kilometers traveled (VKT) before and after a simulated policy, and Stopher and Kockelman (2008b), which examined the sample size benefits of using a multi-day survey to measure VKT, number of trips and total travel time. These studies have several limitations. First, the survey lengths these studies are based on are relatively short (less than a month for each continuous monitoring period). The relatively short survey lengths do not allow these studies to reveal demographic instability and the corresponding travel behavior changes. Second, these studies often relied on the normality assumption of the data, which is not always the case in practice. Third, none of these studies controlled for confounding factors such as demographic characteristics, seasonal effects, and gasoline prices in their sample size calculation procedures. Finally, these studies did not take into account issues such as dropouts and missing data, both of which are common phenomena in GPS-based panel survey data collection.

Based on the few studies on sample size requirements for longitudinal travel surveys, the issues to be explored are many. First, the parameters to be measured are very limited in current studies, including only total travel distance, number of trips and total travel distance. As mentioned before, the objectives of longitudinal travel survey mainly lie in three aspects: modeling travel behavior, evaluating behavior change before and after transportation policies and investments, and understanding process decisions. Different transportation policies will require that the analysis of travel surveys focused on different parameters at various analytical levels. The rationale is that transportation policies affect people's decision-making process pertaining to travel, which in turn affects the results such as intra-regional number of trips and VMT, and long-distance travel frequency and VMT, etc. Therefore, travel response studies need to capture the variables involved in that process. On the one hand, more categories of parameters should be explored. For example, the study of long distance travel patterns will aid decision making concerning fuel consumption, whereas the study of intra-regional travel is critical to congestion management. Also, the GPS tracking technology has provided the possibility to analyze route choices that have never been available before (Stopher, 2008b). These travel attributes need to be analyzed separately because they result from very different decision-making processes. That is, the factors influencing the decision to take long-distance trips are not the same as those influencing the decision not to travel on a certain day. On the other hand, the general parameters should be analyzed at more detailed levels, such as by time of day, day of week, and month of year. By doing so, researchers gain insights into the temporal characteristics of travel behavior such as the rhythm of travel patterns. Such information will not only allow the evaluation of the impact of flexible work hours across travel days (Stopher*, et al.*, 2008b), for example, but also benefit the development of activity-based models.

Second, the impact of survey design considerations on sample size requirements are not well documented in the literature. The first design consideration that will affect sample size requirements is whether simple random sampling, stratified random sampling, or blocking designs are to be employed. Stratified random sampling is "the process of selecting a sample in such a way that identified subgroups in the population are represented in the sample in the same proportion as they exist in the population" (Fraenkel and Wallen, 2008). The basic concept of blocking design is to create homogeneous blocks in which the "nuisance factors", such as household income, household size, etc, are held constant, and the variable of interest, such as a mileage-based pricing incentive, is allowed to vary. Nuisance factors are those that may affect the measured result (e.g. number of trips per day), but not of primary interest (NIST/SEMATECH, 2003). In the case of household travel surveys, the nuisance factors are usually household demographic characteristics.

The sample size estimating methods in existing studies mentioned above assume simple random sampling. In practice, however, stratified random sampling is often employed, using sample size estimating methods adapted from simple random sampling. Such methods may not be applicable to the experimental designs for longitudinal surveys. First, the strict random sampling within subgroups of the population might increase the non-response rate because it is necessary for longitudinal surveys to have reliable participants (Zumkeller*, et al.*, 2006). Second, blocking designs are more appropriate than stratified random sampling for longitudinal surveys to minimize the effect of background conditions. Expectedly, blocking designs provide greater precision of estimates of the difference in travel response than that prepared with a completely randomized design if real block effects are present (Mace, 1964). These special design considerations for longitudinal surveys are likely to render the sample size estimating methods ineffective in the context of travel response studies, and hence the necessity of new methods tailored for longitudinal surveys.

The second and most obvious design consideration that influences sample size requirements is the duration of the continuous monitoring period. The long monitoring period may introduce significant temporal variability as described in Section 2.2.1.3. Such variability needs to be accounted for in the sample size planning procedures. Given the necessity and technical capability to continuously monitor travel behavior, there is a need to examine the impact of longer monitoring periods on sample size requirements. Even though the trade-offs between the duration of monitoring periods and the sample size is intuitive, it is important to quantify these trade-offs to ensure the cost-effectiveness of travel surveys.

Finally, the inherent characteristics including variability and correlation of the attributes of interest are not well understood for longitudinal travel data. The longitudinal design gives rise to potential association for repeated observations within a household, which must be taken into consideration both in the design and the analysis stages (Liu and Liang, 1997). Sample size calculation methods for studies with repeated or correlated data have been proposed since the 1990s, and there is generally no explicit formula that can be used to handle correlated data. Such methods have mainly been applied in biomedical studies, but not yet in travel survey designs. Specifically, the following topics need to be discussed in detail for sample size planning in the transportation field:

- The sources and magnitude of variability; particularly,
    - The influence of changes in background environment on the variability of travel variables
    - The influence of changes within households on the variability of travel variables
- The distributional properties of each travel variable
- The correlation within a household
- The influence of missing data on sample size requirement

The variability in travel behavior variables arises from both between-household and within-household sources, as will be discussed in Chapter 4. The understanding of the magnitude of variability is the first prerequisite for sample size estimation. The greater the variability, the larger the sample size is required. The knowledge of distributional properties, which will be explored in Chapter 8, is required for the adoption of the correct sample size calculation formula or the correct model form for numerical methods. The within-household correlation is characteristic of longitudinal data. The degree of positive correlations affects the required sample size in different directions depending on the main objective of a study. Positive correlation increases the required sample size when estimating cross-sectional averages or differences between averages for more than one group, but decreases the required sample size when estimating a change over time (Diggle*, et al.*, 2002). The correlation structures of key travel behavior variables will be explored in Chapter 8. The design of a longitudinal study also differs from that of a cross-sectional study in that the required sample size should take attrition into account. In a GPS-based panel travel survey, missing data occur due to various reasons such as dropouts and equipment issue. At the design stage, proper consideration of missing data can help ensure the availability of valid data for analysis, as will be discussed in Chapter 10 for before-and-after studies.

2.4. Summary

GPS-based panel travel surveys can extend the monitoring period while decreasing respondent burden. The GPS technology can continuously monitor household travel behavior for multiple days, weeks, months, and even years. This capability can improve the accuracy of activity-based models, as well as policy evaluation, such as congestion pricing, because a long monitoring period allows for causal analysis and the realistic measurement of policy effects. This capability also enables the models to include long-term temporal elements that reflect day-to-day variability of travel behavior and demographic changes.

The desirability of GPS-based panel data warrants research to examine the sample size requirements for collecting such data. The few existing studies have a series of limitations with regard to the assumptions and working data they relied on. More comprehensive procedures to estimate sample size requirements for longitudinal studies have been applied in biomedical studies. Such procedures often require numerical methods, and need to be adapted to suit transportation studies.

This dissertation will adapt the existing procedures in biomedical research to the design of GPS-based panel travel surveys. The variability, distributions, and correlation structures of GPS-based panel travel data will be studied to estimate required sample sizes for studies of various objectives - to obtain reliable statistical references, to conduct regression analysis, and to evaluate transportation policies through before-and-after studies. The data upon which the analyses are based come from the Commute Atlanta study. The detailed data description will follow in Chapter 3.

CHAPTER 3

DATA DESCRIPTION

This chapter describes the data on which the analyses in this dissertation are based. The first part provides an overview of the data collection effort of the Commute Atlanta study, based on the information provided by Li (2004) and Ogle (2005). The second part describes the data processing procedures that extracted certain aspects of travel information for later analyses from the Commute Atlanta data.

3.1. Data Collection

The Commute Atlanta study was designed to assess the effects of converting fixed automotive operating costs into mileage-based and congestion-based operating costs. The monitoring program originally started in 2003, but the pricing study did not formally start until October 2004 due to a delay in funding schedule (Ogle, 2005). For approximately three years between the commencement of monitoring in 2003 and the end of the pricing study in 2006, the Commute Atlanta study has collected detailed information for more than 1.8 million vehicle trips. The pricing experiment of the Commute Atlanta study was implemented in two phases. In Phase I of the study from October 2004 to June 2005, the research team collected baseline travel data and travel diaries from participating households for almost two years. Employer commute options surveys and parallel travel diary data were also collected. In Phase II of the study from October 2005 to June 2006, the team implemented mileage-based incentives at 5 cents/mile from October to December, 10 cents/mile from January to March, and 15 cents/mile from April to June, in an effort to assess whether participating households would change their travel behavior in response to the pricing incentive. The final quarter of pricing at 15 cents per mile is roughly the equivalent of a household paying their insurance, vehicle registration fees, and gasoline taxes on a cent per mile basis. If households respond to cent/mile incentives by reducing vehicle miles of travel, strategies such as pay-as-you drive insurance or replacement of gasoline taxes with cent/mile fees could reduce congestion, fuel consumption, and vehicle emissions.

Data are classified into three primary groups including trip data, demographic data, and support data.

3.1.1. Trip Data

The proposed research effort will use the trip data collected in the Commute Atlanta Value Pricing Study from 2004 to 2006. The Commute Atlanta study recruited 273 households in the 13-county Atlanta metropolitan area. The project installed GT Trip Data Collectors in more than 475 vehicles in the participating households. These devices collected second-by-second vehicle activity data that include position, speed and heading. A trip is defined as the vehicle activities between an engine-on event and an engine-off event.

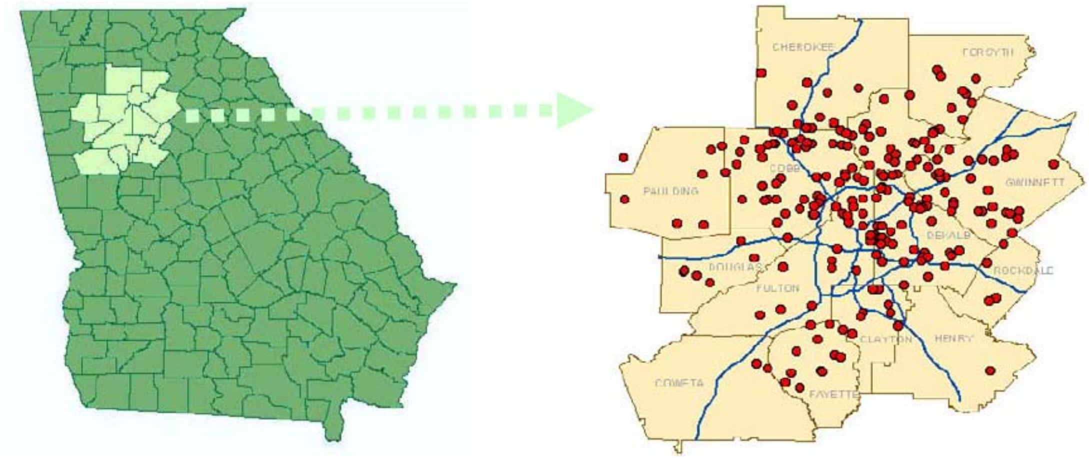

Figure 3.1 Commute Atlanta Study Area and Household Locations
(Li, 2004)

3.1.2. Demographic Data

The Commute Atlanta study also sent out monthly household surveys to update household demographic information so that every trip in the Commute Atlanta database is associated with specific vehicle information, specific household information, and specific primary driver information based on the survey information received. The survey data are supplemented by geo-spatial analysis of vehicle activity data (observation of trip frequencies by time of day) to verify the home, work, and school locations. This methodology is described in detail in Section 3.2.1.

3.1.3. Support Data

Other data that support the data analysis include economic and spatial information. For example, the gas price information is crucial in the analysis of VMT change over time. Spatial information includes various maps such as business locations, street network, transit routes and stops, etc. The Commute Atlanta research team obtained access to land use database and business address database. A variety of mapping routines were employed to incorporate and visually display these data.

3.2. Data Processing

Not all data collected in the period from 2004 to 2006 in the 273 households are usable due to various reasons such as equipment failures and panel attrition. Equipment failures include instrument component failures, disconnecting of power during vehicle maintenance, etc. Instrument component failures, including random memory card failures, GPS, antenna, and other board-level failures, occur approximately 3-5% per year. Panel attrition is another major reason why there are periods for which travel information could not be obtained for every vehicle in the fleet. Part of the attrition is due to a delay in funding in 2003. More than 30% of the households opted out of the study during the study period between the initial deployment and the experimental implementation. The other part of the sample group attrition comes from the instability of household demographics. Changes in participant vehicle ownership, household structure, and residence locations occur at a rate of approximately 8% per year. When households did not report the household changes promptly, the research team would lose contact with those households.

Extensive data processing effort is thus required considering all the aforementioned factors. If households are the sampling units, households lacking proper demographic information need to be removed from the study. If the equipment in one or more vehicles of one household did not function properly for a period, that household will also have to be removed for at least the period where data are omitted (because complete household data are not available for that period). For vehicle- based analysis, for example, in commute behavior studies, only the vehicle in which the equipment failed will need to be removed for that period. For individual-based analysis, the same criterion as for vehicle-based analysis will apply, but valid individual demographic information must also be available.

3.2.1. Demographic Information

The data sources for household demographic characteristics and changes come mainly from monthly household surveys. The survey data are supplemented by geo-spatial analysis of vehicle activity data (observation of trip frequencies by time of day) to verify the home, work, and school locations.

The analytical procedure for each case study begins with a review of the household information relational database to identify basic information such as household structure, work and school status, income and vehicle ownership. In terms of household structure, the analyst reviews number of household members, age group of each member, and the relationships between members. The database also provides information on the work and school status of each participant, including the specific job type, work location, education level, and school location, over time.

The number of household members, income and the number of vehicles owned determine to which GT sampling strata a household belongs. However, one or more aspects of the household demographic characteristics often change during the study period. Some households report these changes in the surveys, but not necessarily in a timely fashion. Other households do not report the changes at all. Analysts are generally able to identify work and school location changes through trip end frequencies, as described below.

Trip ends provide the most insight into household travel activities. The time-of-day analysis of trip end frequencies provides evidence about where household members go and usually infers what they do. Home and work locations are among the most frequently visited places. Also, the trip end with the highest frequency during the morning peak usually signals the work location. Using this information, the analysts rank the trip end frequencies and match the first ten to home and work locations obtained from household surveys. This automated process only matches a limited number of locations. The reasons can be that the household did not report these locations, or the locations changed but they did not report the changes, and/or that the geo-coding process did not identify the right locations. Therefore, a manual process matches the remaining household locations. Household results are integrated into a spatial graphics package to produce a map showing its home, work and school locations, as illustrated in Figure 3.2.

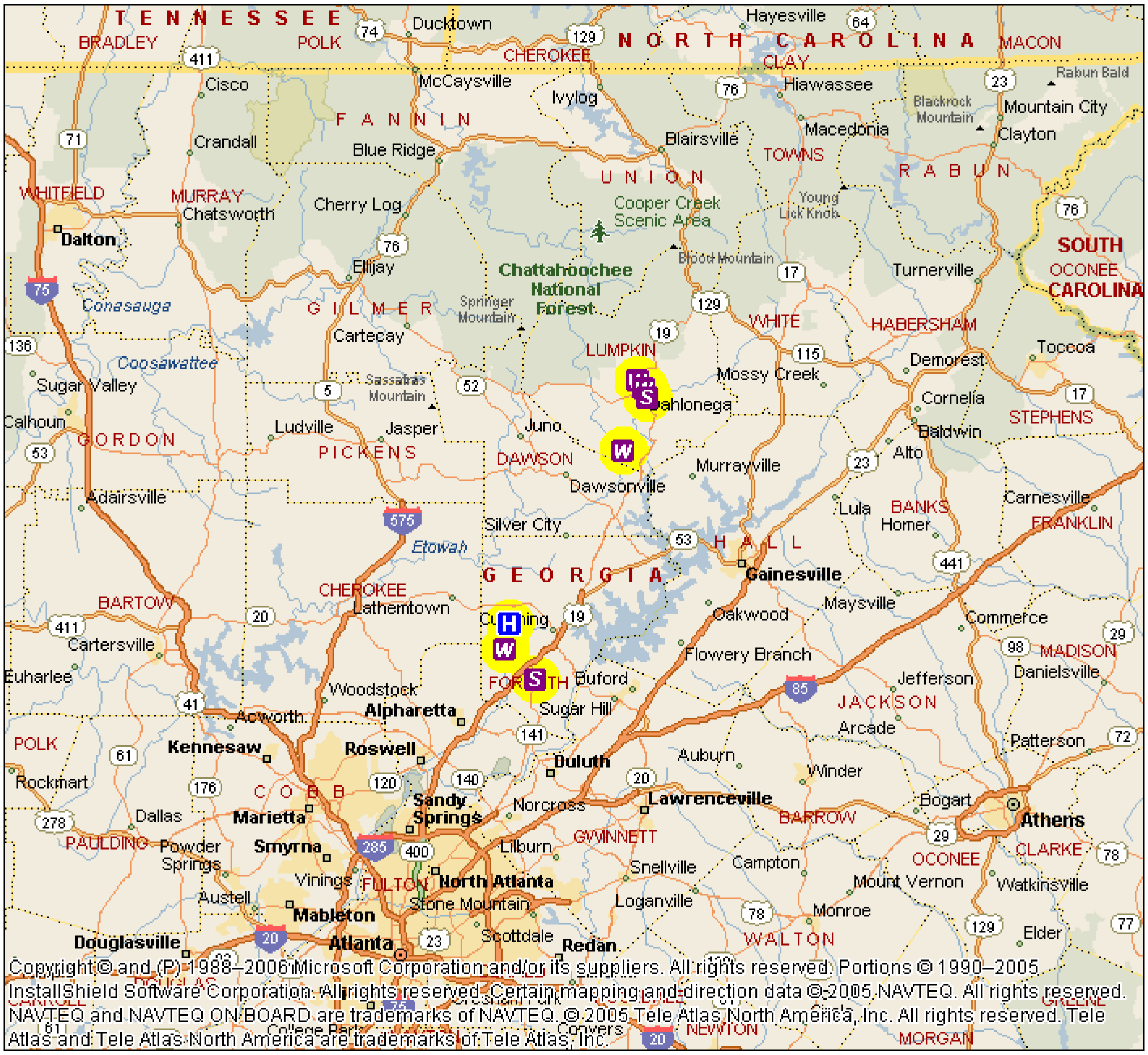

Figure 3.2 Sample Household Home, Work and School Location Map
(Xu, *et al.*, 2009a)

3.2.2. Intra-Regional Trips

Each trip recorded by the GPS devices in the Commute Atlanta study has an associated travel region status, indicating whether a trip: 1) stays inside of the study region from start to end, 2) starts in the region and ends outside of the region, 3) stays outside of the study region from start to end, and 4) starts outside of the region and ends inside the region. The Commute Atlanta intra-regional dataset used in this dissertation includes only the trips that stay inside of the study region from start to end.

The intra-regional dataset in this dissertation is extracted from the pricing study element of the Commute Atlanta study, spanning from October 2004 to June 2005 as the baseline period, and from October 2005 to June 2006 as the pricing period, forming the basis for sample size analysis in Chapter 9, since intra-regional travel is normally the main concern of metropolitan planning organizations (MPOs) for travel demand modeling and policy studies, the intra-regional data. For consistency, all analyses concerning intra-regional travel in this dissertation use this same dataset.

3.2.3. Long-Distance Tours

The algorithms for identifying long-distance tours in this dissertation are adapted from Li *et al.* (2007). A long-distance tour is defined as a series of trips that start inside of the study area, go to a destination outside of the study area and then come back inside the region again, with a linear distance between home and destination of more than 50 miles. The 50-mile criterion is chosen based on both the definition of long-distance travel in the 2001 National Household Travel Survey (NHTS) as reported by Hu and Reuscher, (2004) and recommendations by Li, *et al.* (2007). Travel distance and duration are aggregated from the trips in that tour.

The Commute Atlanta long-distance dataset in this dissertation includes long-distance travel information for the same 95 households as the intra-regional dataset, except that one of the 95 households was excluded. The excluded household has one member who is a college student living outside of the study area but stays with the parents inside the study area extensively. Given the algorithms for identifying long-distance tours, all the intra-regional trips this person makes while staying with the parents would be coded as parts of long-distance tours, rendering the numbers and VMT of long-distance tours observed in this household extremely high. The long-distance dataset provides data spanning from January 2004 to June 2006.

3.3. Data Summary

This dissertation utilizes the GPS-based panel data collected in the Commute Atlanta study. The panel data include trip data collected by the in-vehicle GPS devices, demographic information of the households in the study, and support data such as background economic information and spatial information made available by various mapping routines.

Two separate datasets were processed for the subsequent analyses - the intra-regional travel dataset and the long-distance travel dataset. The intra-regional travel datasets include travel information of the 95 households for which updated demographic information were available. The intra-regional travel dataset covers household travel from October 2004 to June 2005, and from October 2005 to June 2006, as defined by the pricing element of the Commute Atlanta study. October 2004 to June 2005 is the baseline period of the pricing study, whereas October 2005 to June 2006 is the pricing period. Both periods amount to 546 travel days in total. The intra-regional dataset contains 250,580 trips, totaling up to 1.62 million VMT.

The long-distance travel dataset provides long-distance travel information of 94 households. All the households in the intra-regional dataset are included in the long-distance dataset, except for one household in which one member is a college student living outside of the study area but stays with the parents inside the study area extensively. The long-distance travel data span from January 2004 to June 2006. The longer than intra-regional data period allows for more observations of long-distance tours, which can be viewed as rare events compared to intra-regional trips. The long-distance travel dataset contains 1,006 long-distance tours, totaling up to 0.48 million VMT.

CHAPTER 4

EXPLORATORY DATA ANALYSIS

This chapter conducts exploratory analysis for both intra-regional and long-distance travel behavior, as well as some key indicators of overall travel variability, such as zero-trip days and extremely active travel days, using the Commute Atlanta intra-regional and long-distance datasets as described in Chapter 3. The purposes of exploratory data analysis are twofold: 1) to differentiate between the between-household and within-household information made available by the panel data collected in the Commute Atlanta study, and 2) to visualize patterns in data and identify potential causal relationships. By differentiating the between-household variability from the within-household variability, this chapter will deepen the understanding of the nature of panel data, and therefore provide insight into the design of panel surveys. By visualizing patterns in data and indentifying potential causal relationships, this chapter will form a basis for regression analysis in Chapter 9.

The first four sections of this chapter examine overall travel characteristics, including day-of-week variability, seasonality, and the distributions of zero-trip days and extremely active travel days. Section 4.5 explores the relationships between key travel behavior variables and demographic characteristics. Intra-regional travel and long-distance travel are examined separately in most cases because each has unique characteristics and can be influenced by different factors.

4.1. Day-of-Week Variability

4.1.1. Intra-Regional Travel

The Commute Atlanta intra-regional dataset includes 18 months of data for 95 households. As described in Section 3.2.2, intra-regional travel refers to trips that stay within the 13-county Commute Atlanta study area from beginning to end.

Figure 4.1 illustrates the day-of-week variability in the number of intra-regional trips per day. Figure 4.1 is created by first taking the average number of intra-regional trips per each day of week per household, and then generating bootstrap confidence intervals for the household means. Therefore the widths of the error bars in Figure 4.1 represent between-household variability. The numbers of intra-regional trips are not significantly different among weekdays. However, Friday displays a slightly larger value of number of intra-regional trips than other weekdays. Not surprisingly, the lowest number of intra-regional trips occurs on Sunday.

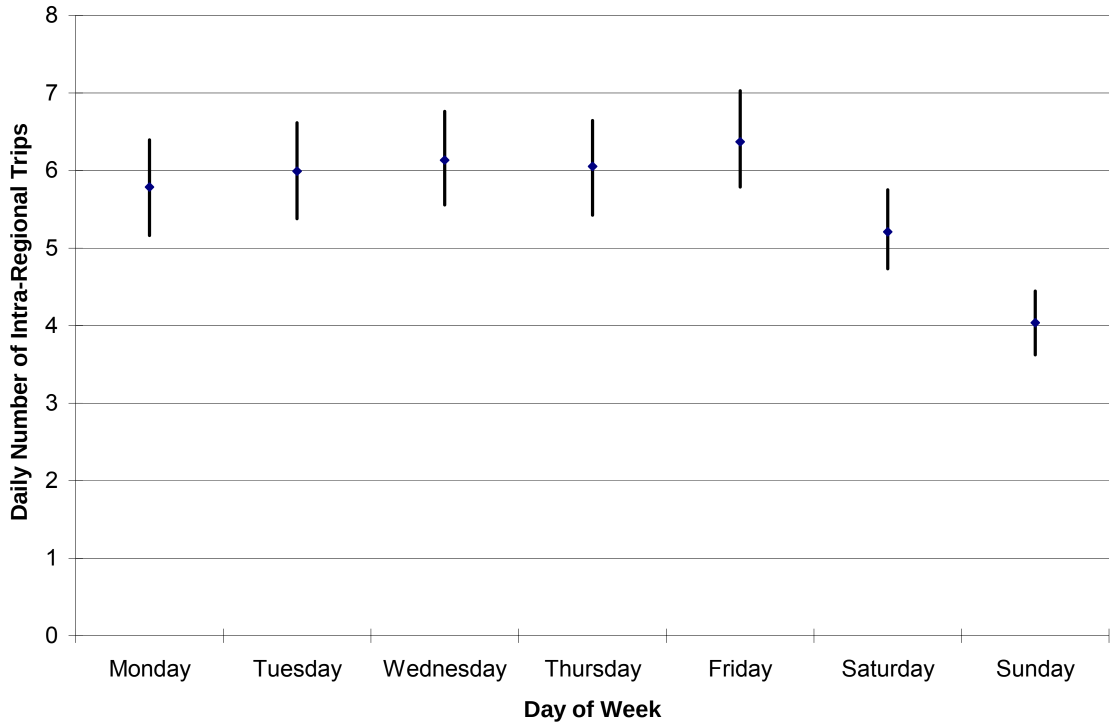

Figure 4.1 Between-Household Day-of-Week Variability in Daily Number of Intra-Regional Trips
Error bars represent 95% bootstrap CI of *household means* of daily number of intra-regional trips.
m=95 households

The confidence intervals in Figure 4.1 are wide, indicating large amount of between-household variability, which can be observed in Figure 4.2. In Figure 4.2, each series of error bars represents an individual household. Three (3) example households are shown to prevent the graph from looking too crowded with all households. The three example households are the 5^{th}, 50^{th}, and 95^{th} percentiles with regard to average number of intra-regional trips per day for *all* days. The width of a bootstrap confidence interval represented by an error bar indicates the variability in the number of intra-regional trips on a day of week within a household. The fluctuations of a series of error bars indicate the rhythm of intra-regional travel frequency by day of week within a household. The differences between each series of error bars show the magnitude of between-household variability in number of intra-regional trips per day by day of week. Figure 4.2 also indicates that different households display different day-of-week profiles with regard to number of intra-regional trips. The trend of higher trip frequency during the week and lower trip frequency during weekends is not necessarily true for each individual household.

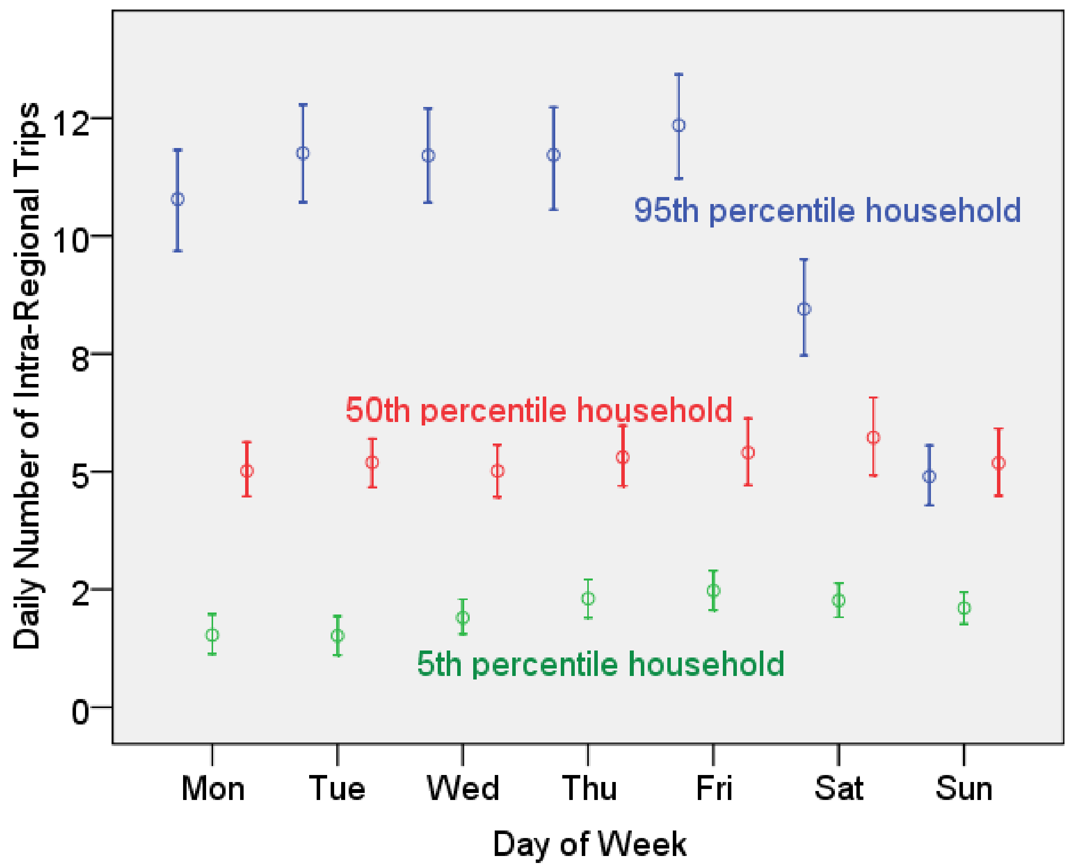

Figure 4.2 Within-Household Day-of-Week Variability with Regard to Daily Number of Intra-Regional Trips
Three series of error bars represent the 5th, 50th, and 95th percentile households with regard to average number of intra-regional trips per day for *all* days.

The VMT associated with intra-regional travel, as shown in Figure 4.3, display similar trends as seen in Figure 4.1. A graph highlighting within-household variability, similar to Figure 4.2 for number of intra-regional trips, is produced for intra-regional VMT and is provided in Figure 4.4.

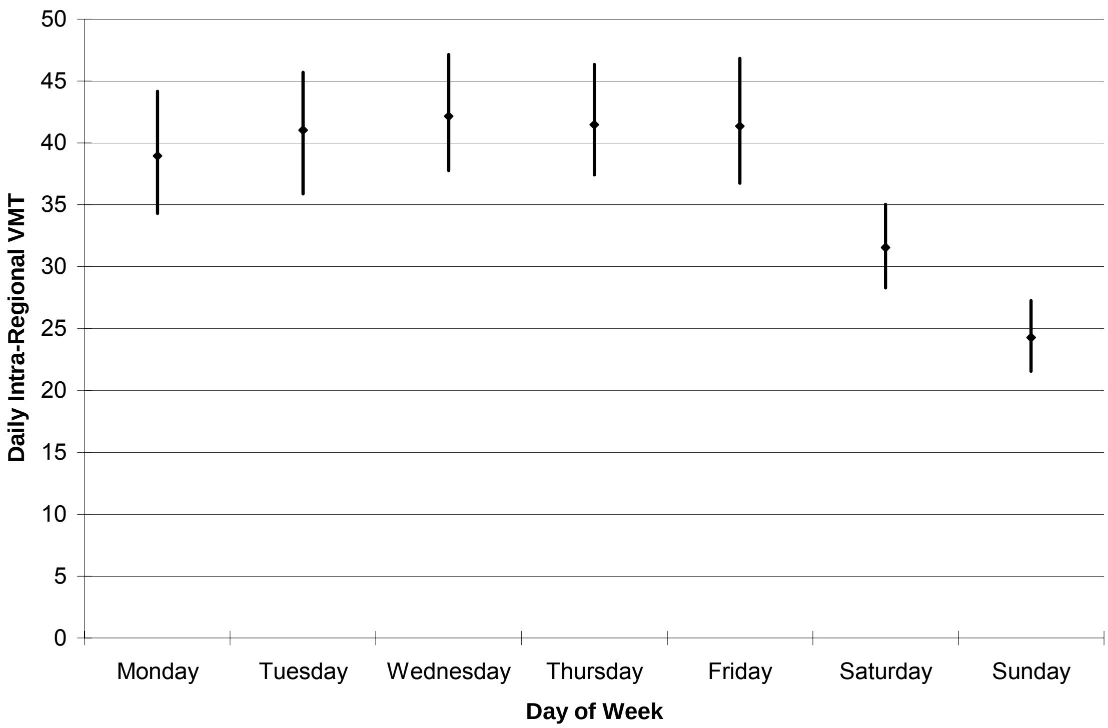

Figure 4.3 Between-Household Day-of-Week Variability in Daily Intra-Regional VMT
Error bars represent 95% bootstrap CI of *household means* of daily intra-regional VMT.
m=95 households

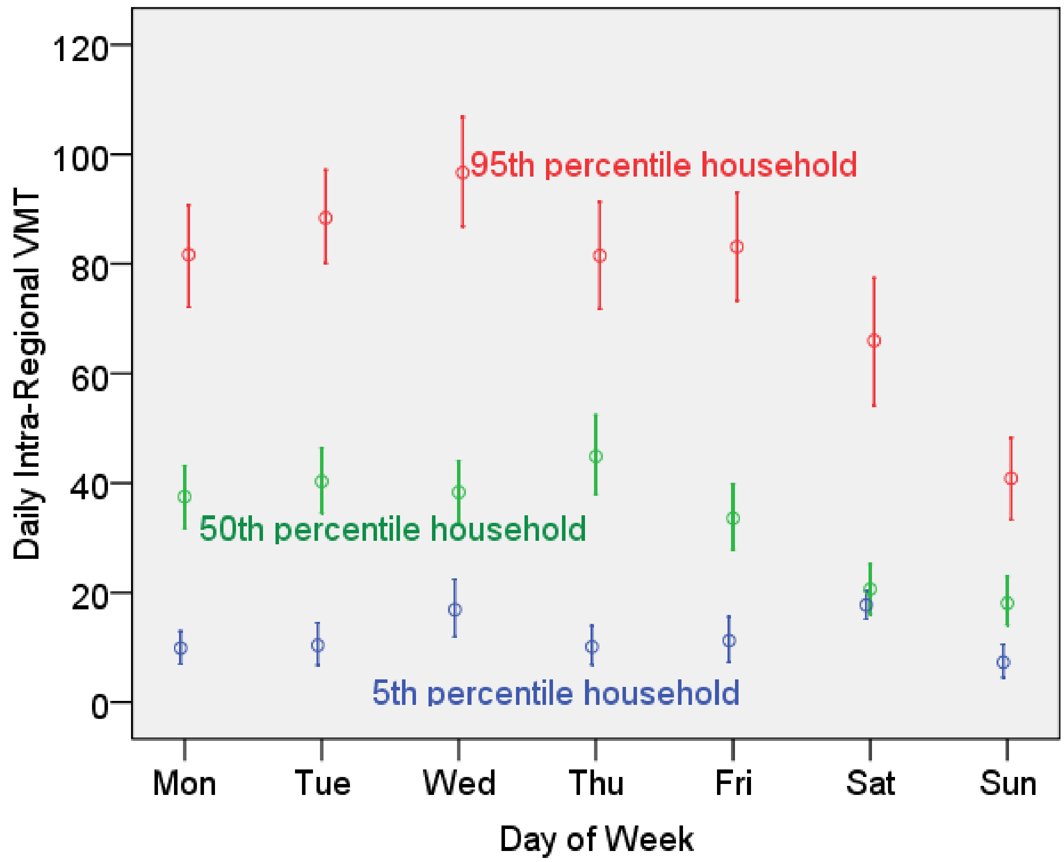

Figure 4.4 Within-Household Day-of-Week Variability with Regard to Daily Intra-Regional VMT
Each series of error bars represents an individual household.
Three series of error bars represent the 5th, 50th, and 95th percentile households with regard to average intra-regional VMT per day for *all* days.

A difference between the day-of-week variability in the number of intra-regional trips and that in intra-regional VMT is that, unlike the number of intra-regional trips, VMT do not peak on Fridays, indicating that households tend to make more trips but shorter trips on Friday. Another difference is that the number of intra-regional trips on Saturday does not seem significantly lower than that on a weekday, but the intra-regional VMT on Saturday does, indicating that households tend to make shorter trips on a Saturday than on a weekday, which is not surprising for trips that stay within the region.

4.1.2. Long-Distance Tours

As described in Section 3.2.3, long-distance tours are defined as a series of consecutive trips that start and end at home, with the farthest destination of all trips being outside of the 13-county study area and more than 50 miles away from home measured in straight-line distance. Because long-distance tours tend to be multi-day events, the numbers and VMT of long-distance tours shown in both figures refer to the tours starting on a particular day of week.

As expected, the day-of-week variability of long-distance travel is significantly different from intra-regional travel. Figure 4.5 shows the average number of long-distance tour per day starting on each day of week. The values are very small, indicating that long-distance tours are rare events. The mean number of long-distance tour starts peaks on Friday and Saturday, as does the variance. The mean number of long-distance tours on Sunday is lower than those on Friday and Saturday which are significantly different than those on Monday through Thursday, but higher than those during the rest of the week. This phenomenon is attributable to the impact of holidays that occur on a Monday. Households are also more likely to make one-day tours on Sunday than multi-day tours. Among all the 149 long-distance tours starting on Sundays, 71 are one-day tours. This is the second highest number of one-day tours starting on a certain day of week, next to 99 one-day tours starting on Saturday among all the 244 long-distance tours starting on Saturday.

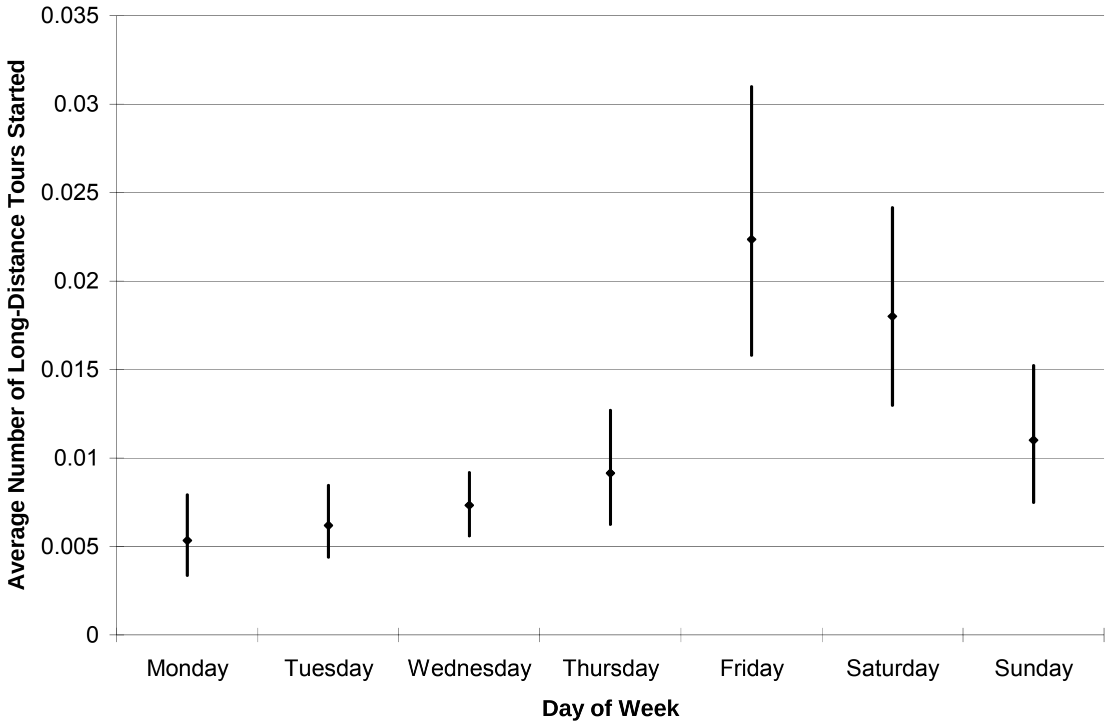

Figure 4.5 Day-of-Week Variability in Number of Long-Distance Tour Starts
Error bars represent 95% bootstrap CI of *household means* of number of long-distance tours per month.
m=94 households

4.2. Seasonality

4.2.1. Intra-Regional Travel

Figure 4.6 and Figure 4.7 present the seasonal variations of daily number of trips and VMT of intra-regional travel as seen in the Commute Atlanta intra-regional travel dataset. The trend of number of trips and that of VMT are similar, indicating that the average distance of a trip is relatively stable throughout the year. Not surprisingly, the daily intra-regional travel does not display a very significant seasonal pattern. The values of both the number of trips and VMT seem low for April and June, in contrast to the higher values in March and May. The low values seen in April could be attributable to the spring break that occurs in early April for schools in the Atlanta area. The low values seen in June could be attributable to the fact that the last day of school occurs in early June in the Atlanta area. Households tend take out-of-town trips when the schools are off, as evidenced in the increase in long-distance travel in April and June that will be examined in the next section, and hence show less intra-regional travel.

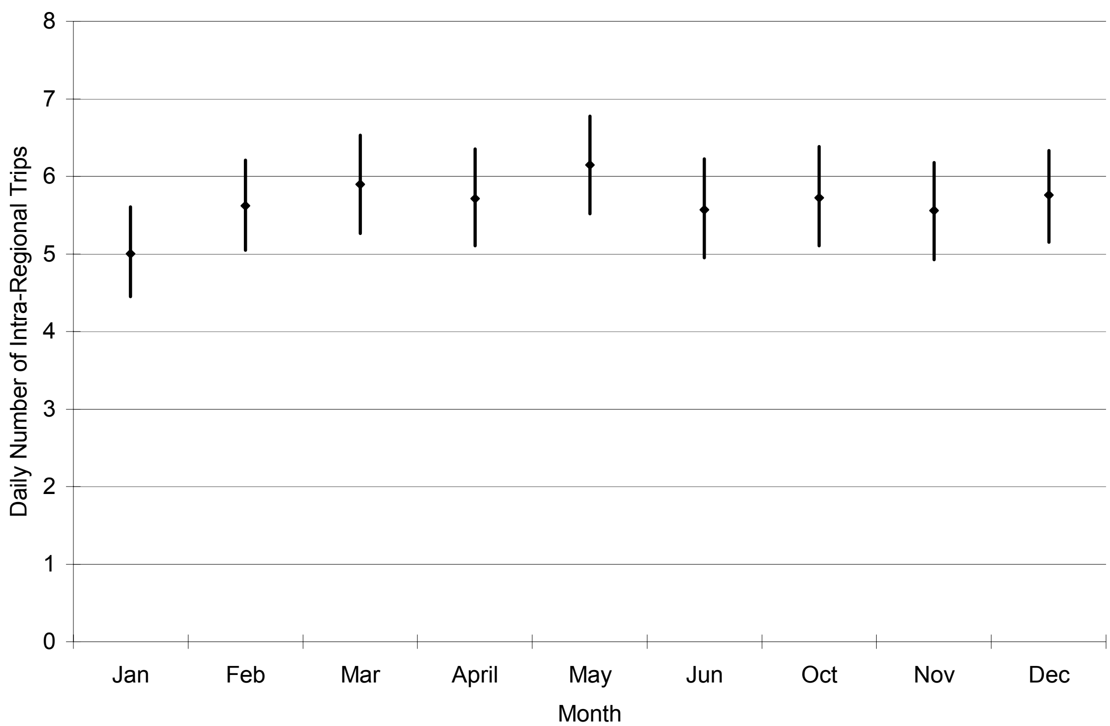

Figure 4.6 Seasonal Effects on Daily Intra-Regional Number of Trips
Error bars represent 95% bootstrap CI of *household means* of number of intra-regional trips per day.
m=95 households

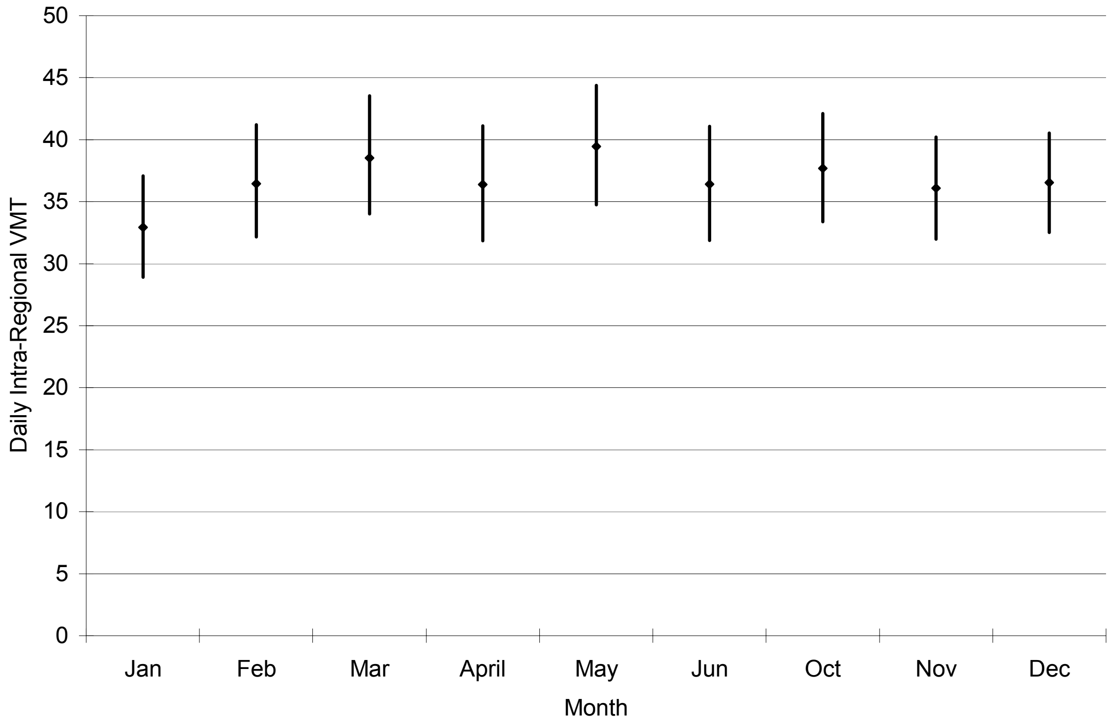

Figure 4.7 Seasonal Effects on Daily Intra-Regional VMT
Error bars represent 95% bootstrap CI of *household means* of daily intra-regional VMT.

The seasonal effects are also compared year to year, as shown in Figure 4.8 and Figure 4.9. The average household daily intra-regional travel during the baseline months of the Commute Atlanta study (October 2004 to June 2005) is compared side by side to that during the pricing months (October 2005 to June 2006). The information revealed in Figure 4.8 and Figure 4.9 will have important implications on sample size analysis for before-and-after studies that will be presented in Chapter 10, because the higher the year-to-year variability in travel is, the more challenging it is to detect a policy impact with a given sample size. Figure 4.8 and Figure 4.9 show that the seasonal trends in both study periods are similar. However, the confidence intervals of the months March, April and May of the pricing period are wider than those of baseline period, indicating the presence of extreme values in those months of the pricing period. The extreme values will have a significant impact on sample size requirements, as will be discussed in Chapter 10.

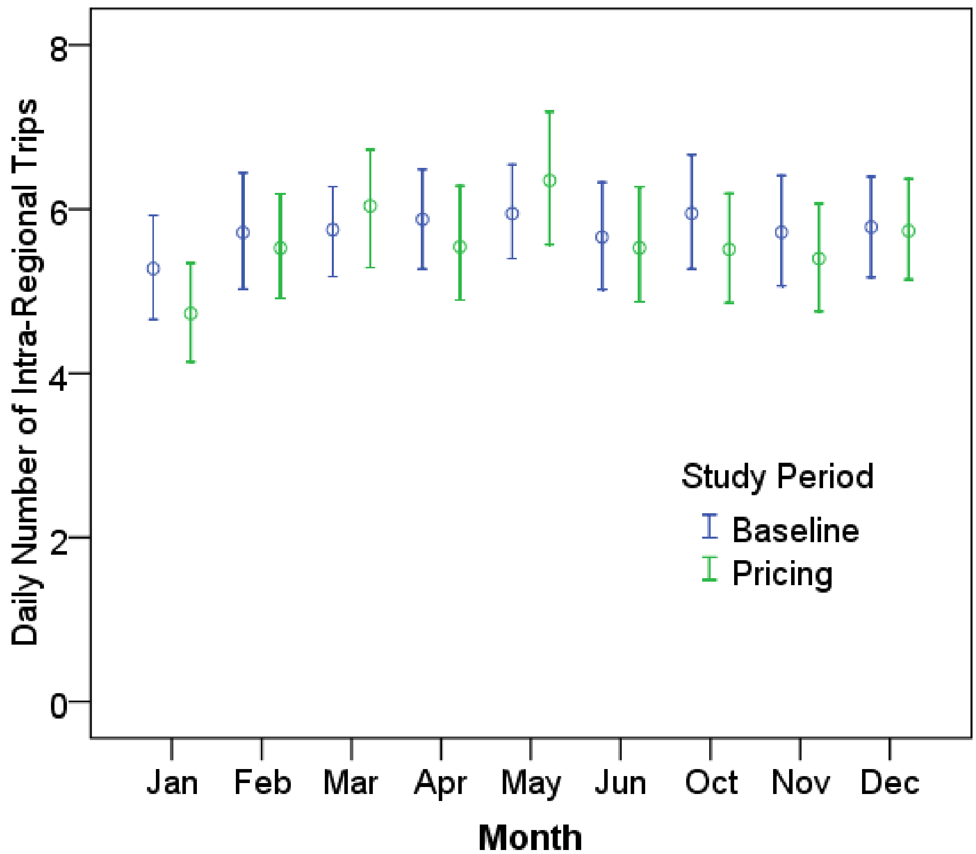

Figure 4.8 Comparison of Daily Number of Intra-Regional Trips between Baseline and Pricing Periods
Error bars represent 95% bootstrap CI of *household means* of daily number of intra-regional trips.

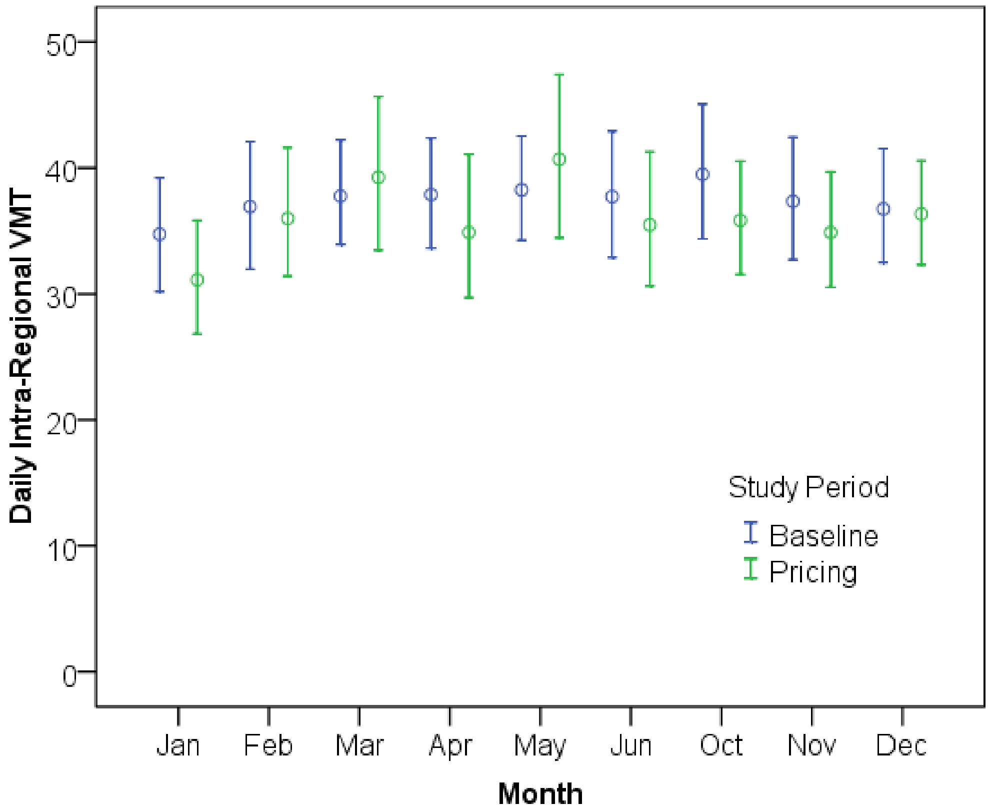

Figure 4.9 Comparison of Daily Intra-Regional VMT between Baseline and Pricing Periods
Error bars represent 95% bootstrap CI of *household means* of daily intra-regional VMT.

4.2.2. Long-Distance Travel

Figure 4.10 shows little seasonal variation of monthly number of long-distance tours. The average number of long-distance tours per household in January appears significantly different than those in April and June, but all other months show very similar numbers of long-distance tours. The wide confidence intervals are associated with the large between-household variability in monthly number of long-distance tours.

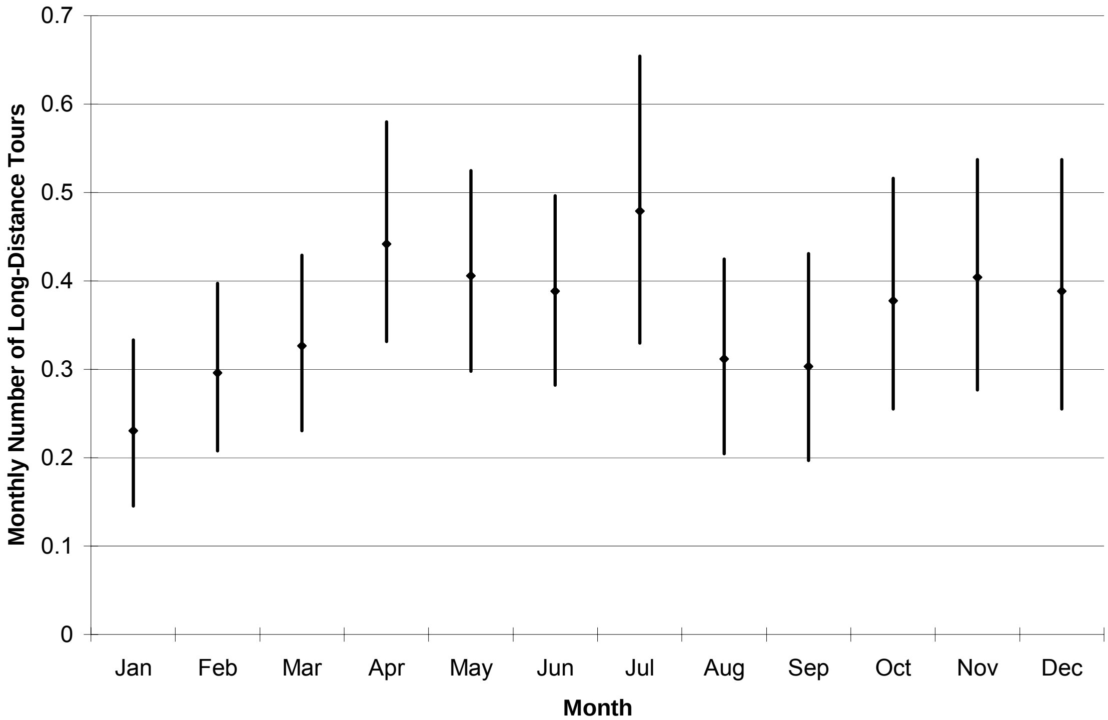

Figure 4.10 Seasonal Effects on Monthly Number of Long-Distance Tours
Error bars represent 95% bootstrap CI of *household means* of number of long-distance tours per month.
m=94 households

Figure 4.11 shows the seasonal variation of monthly VMT of long-distance tours. The month-to-month differences in mean long-distance VMT are larger and more significant than those in number of long-distance tours. Similar to Figure 4.10, January observes the lowest value of long-distance VMT. However, the peak in July is more significant in Figure 4.11 than in Figure 4.10, indicating that households tend to travel not only more frequently in July, but also farther away. The high value of VMT in June is likely attributable to the fact that households tend to take long family vacations after the schools let out.

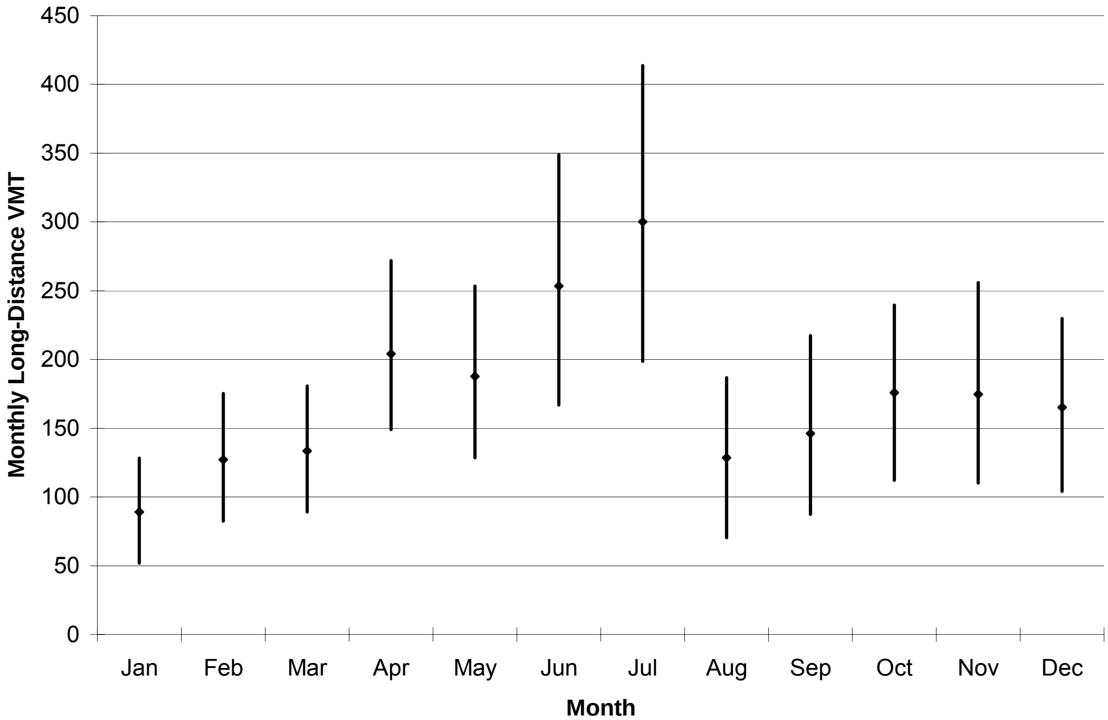

Figure 4.11 Confidence Intervals of Monthly VMT of Long-Distance Tours
m=94 households

4.3. Zero-Trip Days

Based on a meta-analysis of surveys in the U.S. and Europe, Madre, *et al.* (2007) reported that most people will usually stay at home all day at least one day per week, and that a standard one-day, weekday-only travel diary should expect 8% - 12% immobile person-days. Analyzing the occurrence of zero-trip days has important implications on the evaluation of survey quality, because households that do not respond to paper-based travel diaries have higher incidences of zero-trip days (Ogle*, et al.*, 2005). The Commute Atlanta data are based on active tracking, so survey non-response is not an issue, allowing an accurate estimate of the ratio of zero-trip days in household travel behavior.

The Commute Atlanta pricing study that collected18-month of data for 95 households includes 546 days of observations for each household. This should amount to

95*546=54,720 “household-days” of data. With missing data taken into account, the study actually provided 44,173 household-days of data. Among the 44,173 household-days in the sample, 4,959 household-days observed zero trips[1]. This amounts to 11.2% of all the household-days, or slightly less than one day a week. Specifically, 9.3% of all weekdays observed zero trips. This result agrees quite well with conclusions in Madre *et al*. (2007).

Figure 4.12 illustrates the distribution of zero-trip days by day of week. Interestingly, households have a higher propensity not to travel in weekends. On weekdays, Wednesdays observe the least zero-trip days.

[1] The analysis of this section considers all travel, including both intra-regional and long-distance, between October 2004 and June 2005 and October 2005 and June 2006.

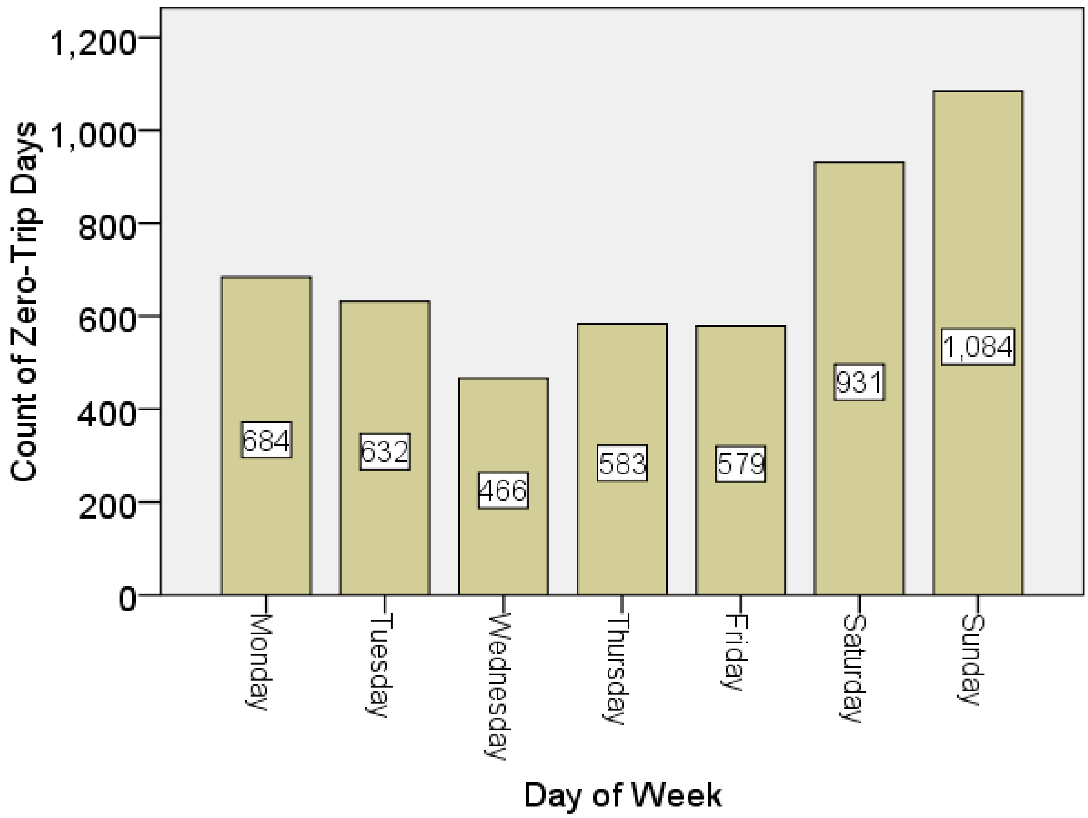

Figure 4.12 Number of Zero-Trip Days by Day of Week

4.4. Extremely Active Travel Days

Approximately 5% of all household-days, as well as all household-weekdays, observed 15 or more trips[1] per day. The 97.5 percentile of number of trips observed in one day is 18. Therefore, an "extremely-active-travel" day is defined as a day when a household made 15 or more trips. Ogle, *et al.* (2005) reported that non-responding households in a paper diary survey have higher incidences of 16+ trip days, and that overall, non-responding households make a higher average number of trips per day per

[1] The analysis of this section considers all travel, including both intra-regional and long-distance, between October 2004 and June 2005 and October 2005 and June 2006.

household. Understanding the non-response bias towards active travel has important implications on evolving travel survey design.

Figure 4.13 shows the distribution of extremely active travel days by day of week. The counts are of 95 households for 18 months. Friday observes more active travel than any other days of the week. Households are also less likely to have an extremely active travel day on weekends than during the week. From Monday to Thursday, the tendency to have extremely active travel varies only very slightly.

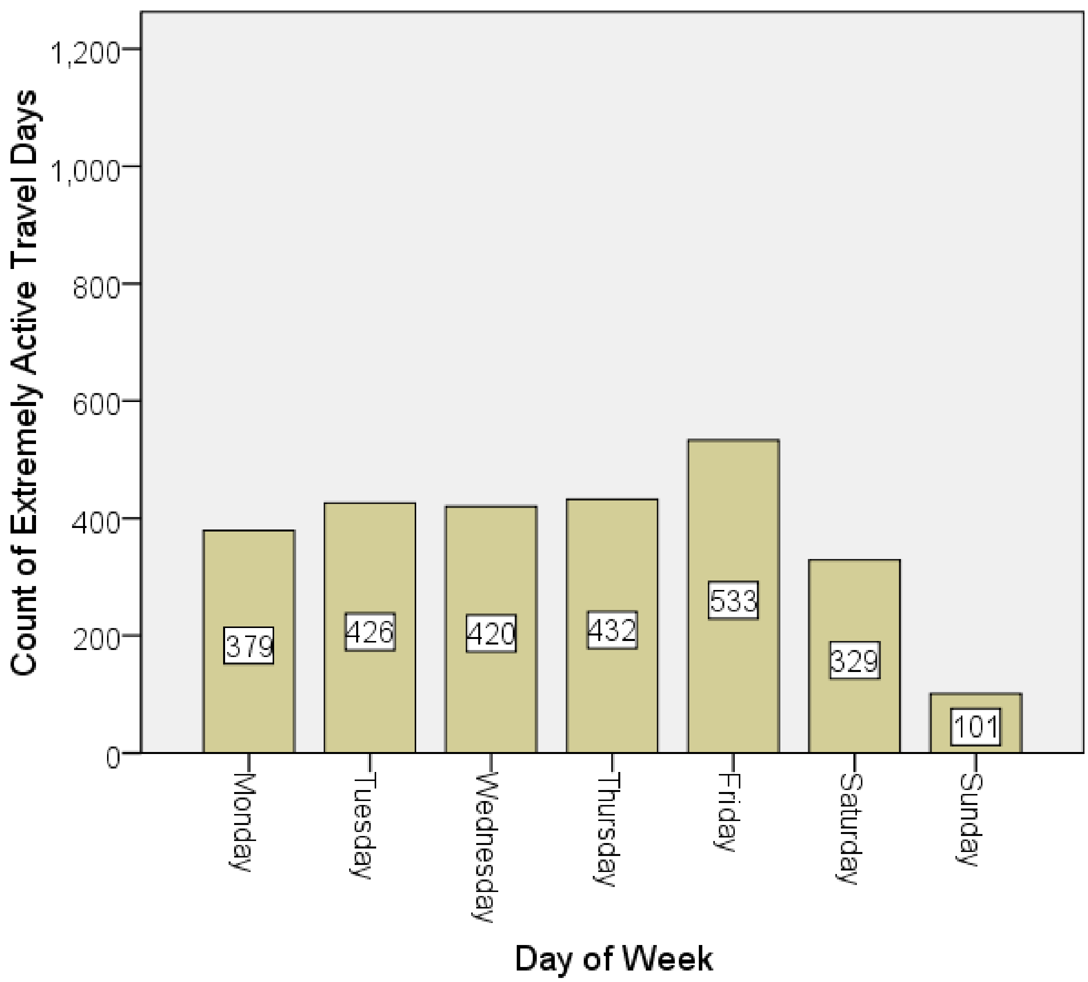

Figure 4.13 Number of Extremely Active Travel Days by Day of Week

4.5. Association between Travel Behavior and Demographic Characteristics

In the Commute Atlanta study, households are recruited based on income, household size and vehicle ownership (Elango*, et al.*, 2007). These variables are also frequently used as control variables in activity-based models (Bradley and Bowman, 2006). In this section, key travel behavior variables will be examined against income, household size and vehicle ownership separately. To provide a basis for subsequent regression analysis, the variable of daily intra-regional VMT is chosen as the proxy for intra-regional travel, and the monthly number of long-distance tours is chosen as the proxy for long-distance travel. These two variables and their association with demographic characteristics are discussed in detail in this section. Relevant graphs for the number of intra-regional trips and long-distance VMT are included in Appendix A, but are not discussed in the text.

4.5.1. Demographic Characteristics and Intra-Regional Travel

This section examines the effects of various demographic characteristics on daily intra-regional VMT. The three major demographic variables assessed in the context are household income, household size, and total number vehicles owned by the household.

4.5.1.1. Household Income

Figure 4.14 shows the association between daily intra-regional VMT and household annual income. The number of observations in each of the four income groups is summarized in Table 4.1, as are the definitions of the income groups. Figure 4.14 indicates a positive correlation between household income and daily intra-regional VMT. Households with annual incomes below $30,000 undertake significantly fewer VMT than other households. Households with annual incomes between $75,000 and $100,000 and households with annual incomes more than $100,000 appear to conduct similar amounts of VMT.

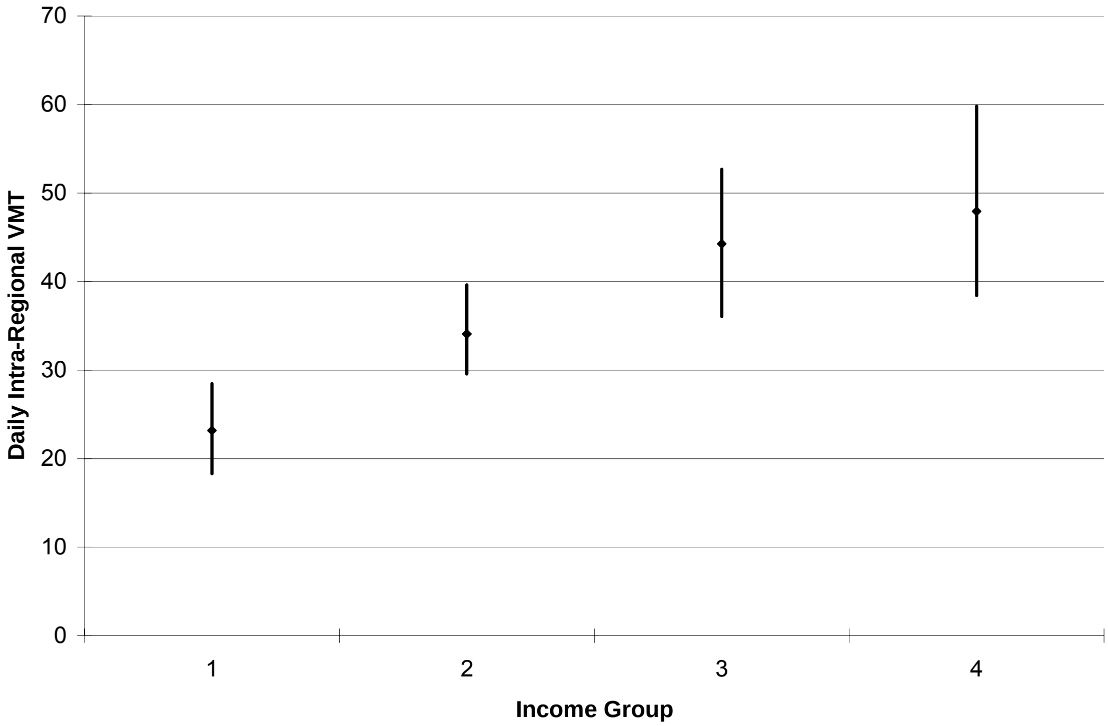

Figure 4.14 Daily Intra-Regional VMT by Income Group
m=95 households

Table 4.1 Number of Observations by Income Group

Income Group		Number of Household-Days	Percent
Code	Definition		
1	<$30k	4095	9.3
2	$30k - $75k	24925	56.4
3	$75k - $100k	5654	12.8
4	≥$100k	9499	21.5
Total		44173	100.0

Figure 4.14 also indicates that, not only do average daily intra-regional VMT increase with household income, but the between-household variability in intra-regional VMT also increase with household income, likely due to the lack of control of other variables, e.g. employment status, household structure, etc.

4.5.1.2. Household Size

The association between daily intra-regional VMT and household size is presented in Figure 4.15. Table 4.2 summarizes the number of observations by household size. The daily intra-regional VMT generally increase with household size in the Commute Atlanta sample. The differences in daily intra-regional VMT among households with one (1), two (2), and three (3) members are statistically significant, but the differences among households with three (3) and four (4) members are not. The widths of confidence intervals also show a trend of increasing as the household size increases, indicating that households with more members tend to display more variability with regard to daily intra-regional VMT.

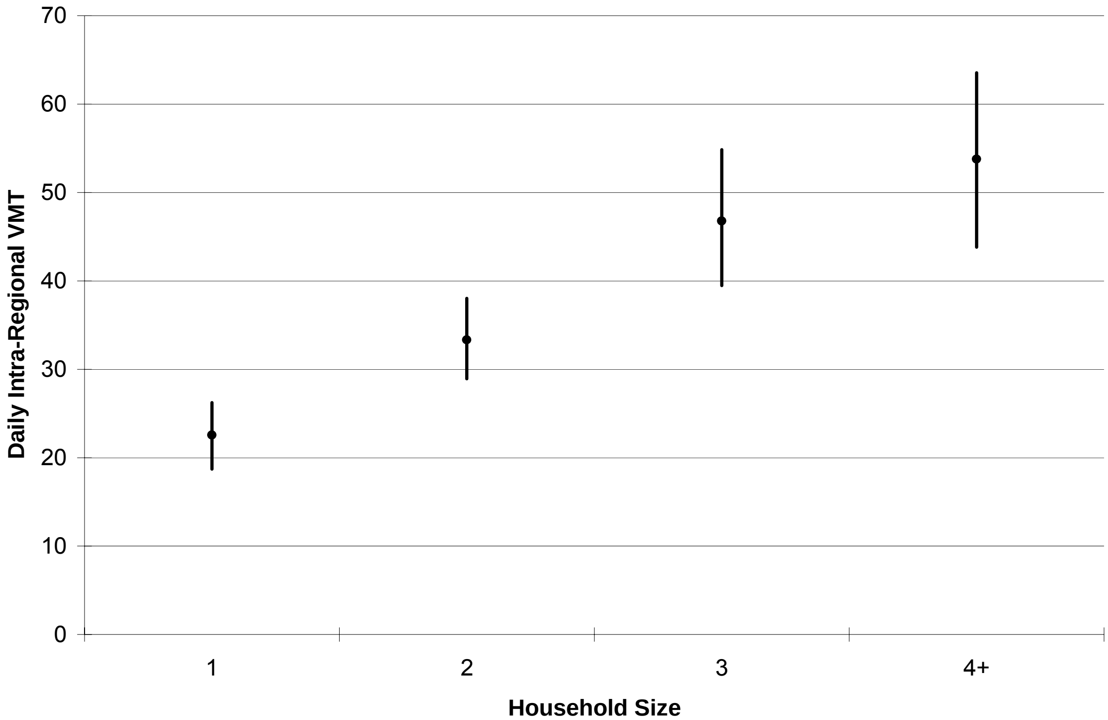

Figure 4.15 Daily Intra-Regional VMT by Household Size
m=95 households

Table 4.2 Number of Observations by Household Size

Household Size	Number of Household-Days	Percent
1	12726	28.8
2	16088	36.4
3	6010	13.6
4+	9349	21.2
Total	44173	100.0

4.5.1.3. Vehicle Ownership

The association between daily intra-regional VMT and household vehicle ownership is presented in Figure 4.16. Not surprisingly, the graph shows a trend of positive correlation. In the Commute Atlanta sample, households that own only one (1) vehicle display significantly lower daily intra-regional VMT, whereas households that own four (4) or more vehicles display significantly higher intra-regional VMT. Households that own two (2) vehicles show similar amount of VMT to households that own three (3) vehicles.

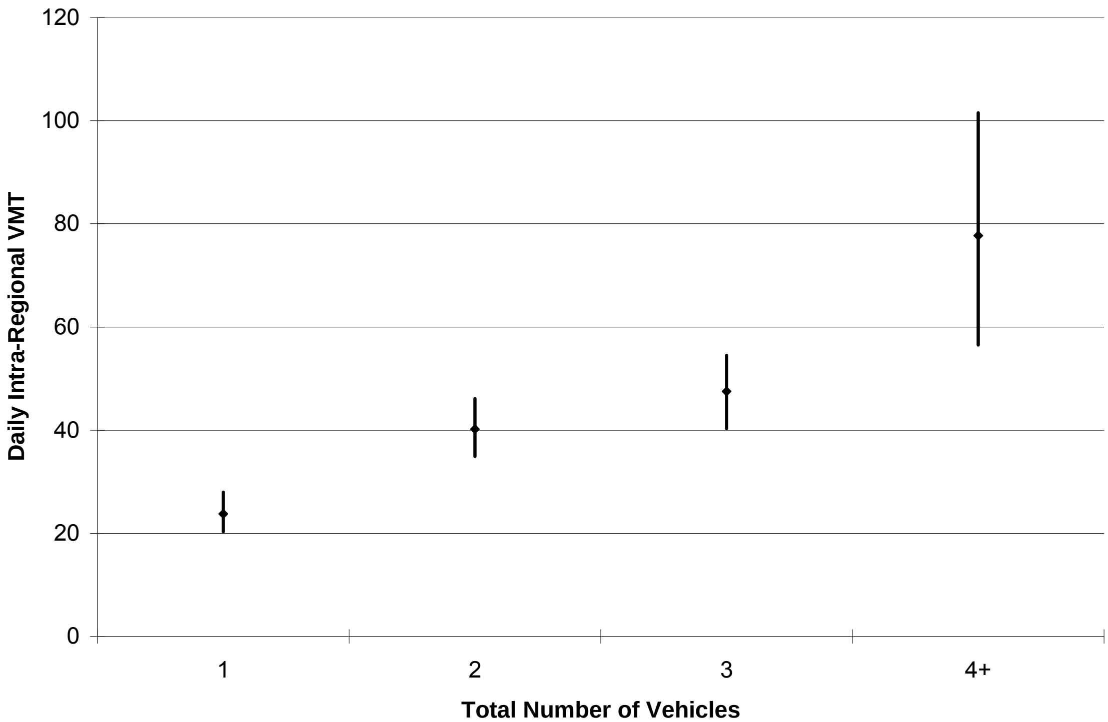

Figure 4.16 Daily Intra-Regional VMT by Total Number of Vehicles Owned
m=95 households

Table 4.3 Number of Observations by Total Number of Vehicles

Total Number of Vehicles	Number of Household-Days	Percent
1	17385	39.4
2	16871	38.2
3	7048	16.0
4+	2869	6.4
Total	44173	100.0

4.5.1.4. Inter-Correlation among Demographic Variables

The positive correlations seen in Figure 4.14, Figure 4.15 and Figure 4.16 should be interpreted with the caveat that there is significant collinearity among the demographic variables. Table 4.4 provides the non-parametric correlation coefficient, Spearman's ρ, between each pair variables among the three demographic variables. All correlations are significant, with the highest correlation coefficient value of 0.615 occurring between household size and total number of vehicles. The significant amount of collinearity implies that remedial measures should be taken into account in the subsequent regression analysis.

Table 4.4 Correlation among Demographic Variables

	Spearman's ρ	Household Size	Income Group
Income Group	Correlation Coefficient	0.412	
	Significance	.000	
Total Number of Vehicles	Correlation Coefficient	0.694	0.484
	Significance	.000	.000

4.5.2. Demographic Characteristics and Long-Distance Travel

The analysis of the association between demographic characteristics and long-distance travel follows the same structure as the previous section with regard to intra-regional travel. The number of long-distance tours per month is analyzed by income group, household size, and vehicle ownership, respectively.

4.5.2.1. Household Income

The association between household average number of long-distance tours per month and annual income is presented in Figure 4.17. The positive correlation is less conspicuous in Figure 4.17 than in the case of intra-regional travel presented in Figure 4.14. Households with annual incomes lower than $75,000, i.e. income groups 1 and 2, display similar monthly number of long-distance tours of about 0.3, whereas households with annual incomes higher than $75,000, i.e. income groups 3 and 4, display similar monthly number of long-distance tours of around 0.5. However, the overall differences are not significant.

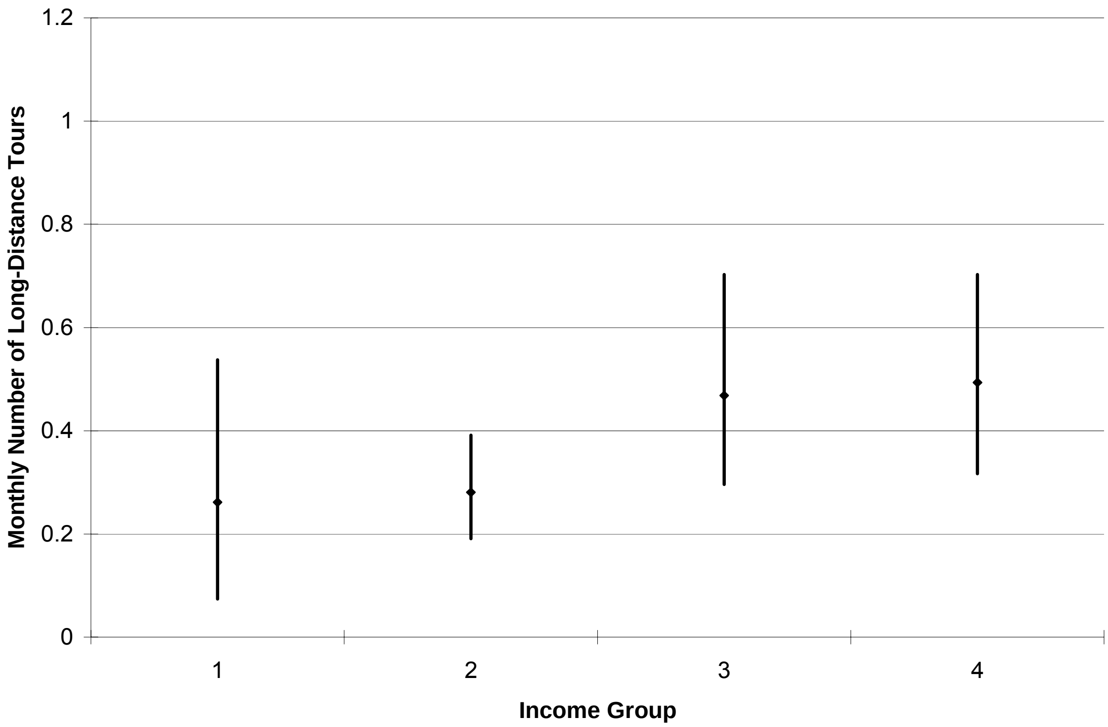

Figure 4.17 Monthly Number of Long-Distance Tours by Income Group
m=94 households

Including both the between-household and within-household information, Figure 4.18 provides more insight into the variability in monthly number of long-distance tours with regard to income. The information presented in Figure 4.18 has a few implications.

First, there is a positive association between monthly number of long-distance tours and household income, as evidenced by the fact that the percent of households that never conducted long-distance tours during the study period within each income group decreases as household income increases. Second, within each income group, there is significant between-household variability. The few households that show significantly larger numbers of long-distance tours within each income group have contributed to the wide confidence intervals in Figure 4.17. Third, the large amount of within-household variability implies that the average number of long-distance tours per month inferred from the Commute Atlanta long-distance dataset may not be reliable. The significant within-household variability in monthly number of long-distance tours accentuates the nature of long-distance travel as rare events, and hence a longer survey period needed to obtain reliable means.

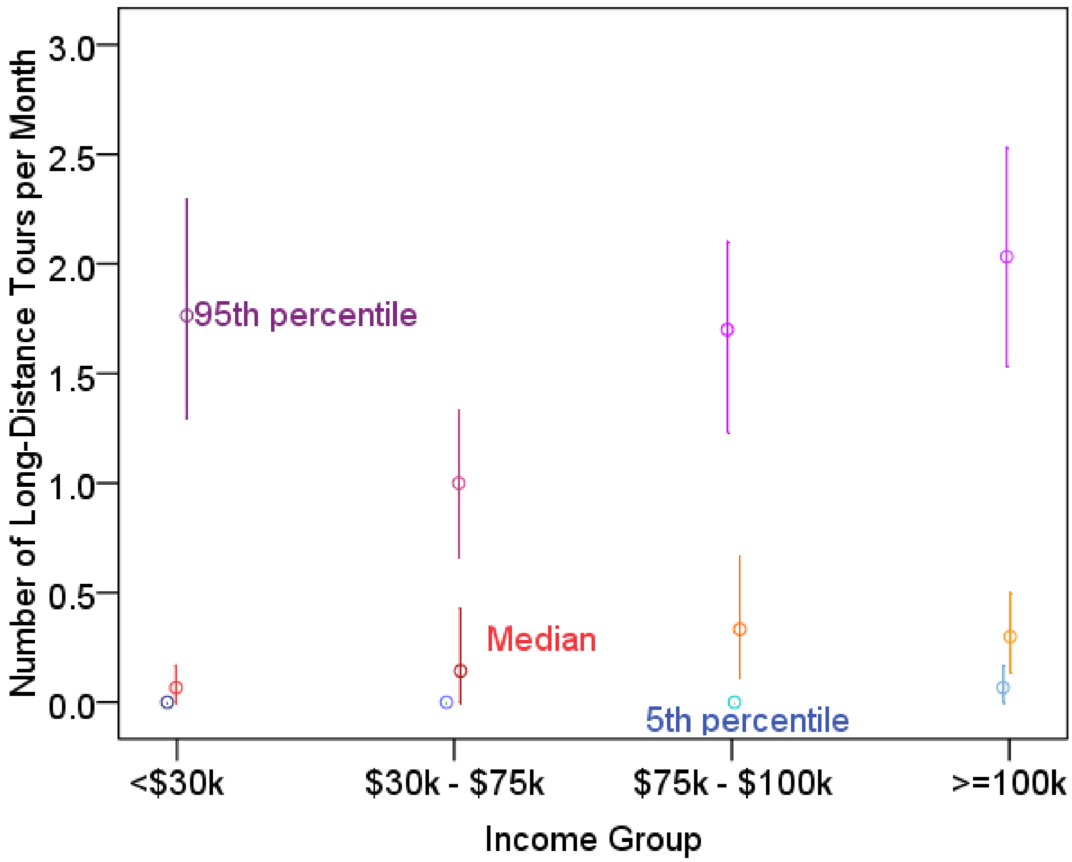

Figure 4.18 Between-Household and Within-Household Variability in Number of Long-Distance Tours per Month by Income Group
Each error bar represents a household

4.5.2.2. Household Size

The association between monthly number of long-distance tours and household size, as shown in Figure 4.19, does not show a significant positive correlation. Single-person households show the smallest average number of long-distance tours per month, but three-person households do not seem to conduct significantly more long-distance tours. Furthermore, all multi-person households (households with two or more people) show similar number of long-distance tours per month.

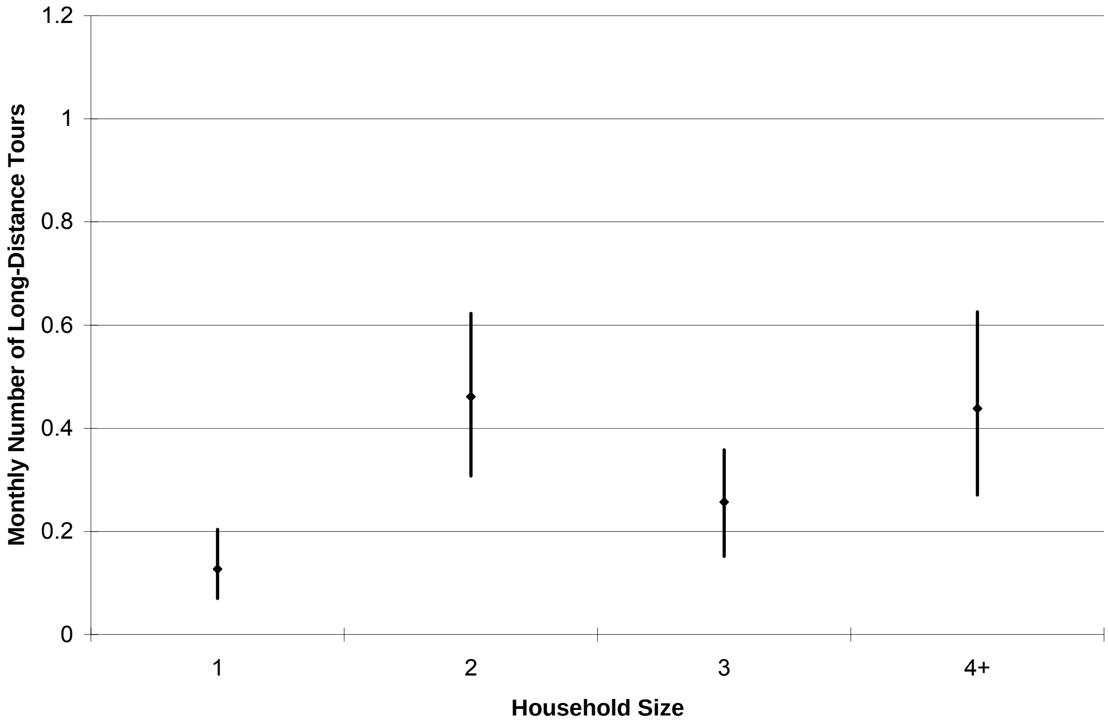

Figure 4.19 Number of Long-Distance Tours per Month by Household Size
m=94 households

4.5.2.3. Vehicle Ownership

Figure 4.20 indicates that the positive correlation between number of long-distance tours per month and total number of vehicles is not significant, similar to the associations seen in Figure 4.17 and Figure 4.19 for household income and household size, respectively. Households that own only one vehicle show a significantly smaller number of long-distance tours per month than other households. Households with two (2) or more vehicles do not show significantly different numbers of long-distance tours per month.

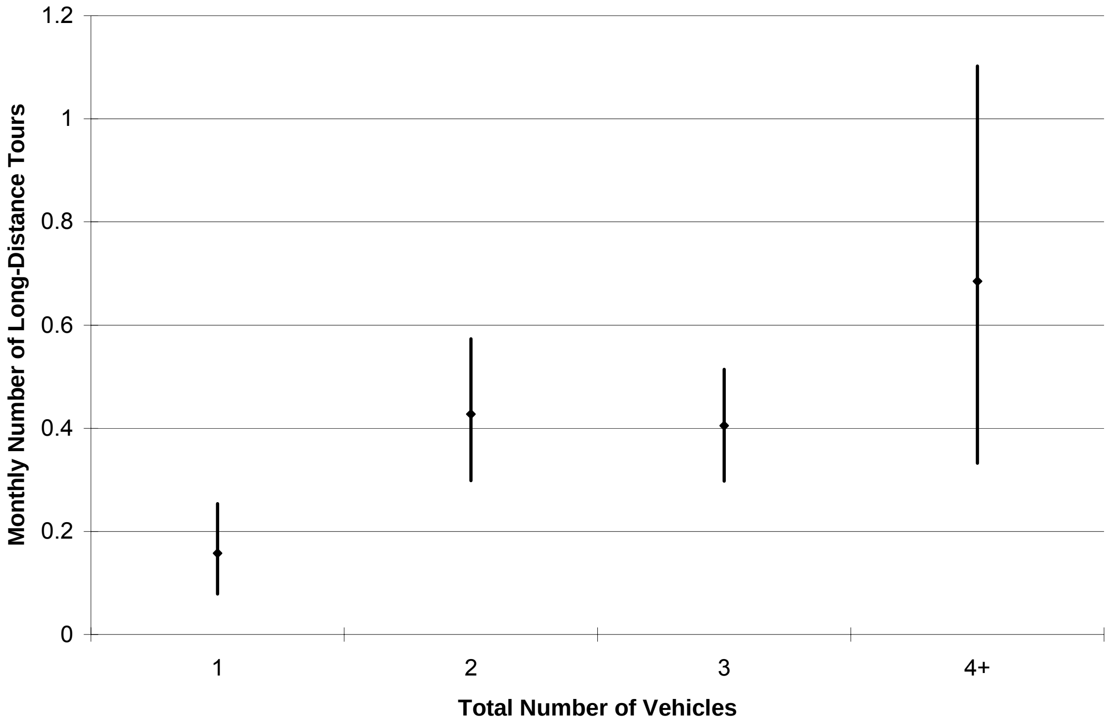

Figure 4.20 Number of Long-Distance Tours per Month by Total Number of Vehicles Owned
m=94 households

To summarize, the associations between long-distance travel and demographic characteristics such as income, household size and vehicle ownership appears less strong than that between intra-regional travel and these demographic characteristics in the Commute Atlanta datasets. The less conspicuous association between long-distance travel and demographic characteristics does not mean that such association does not exist, but only implies that: 1) long-distance tours are rare events and therefore require a larger sample size and a longer survey period, and 2) long-distance travel could be more significantly associated with other factors that are not available in the current dataset, such as the presence of out-of-town relatives.

4.6. Summary

This chapter conducted exploratory analysis of the data to showcase the differences between variability that arises from between-household and within-household sources, and to visualize patterns in intra-regional and long-distance travel against temporal and demographic factors. Perhaps the most important implication of the exploratory data analysis is that the sources of variability in longitudinal data are many - natural temporal rhythms arising from day-of-week and seasonal effects, between-household demographic differences, and unexplained variability within a household, just to name a few. Demographic characteristics of a household tend to change over the course of a long-term panel study, and will often result in changes in travel behavior, as will be discussed in detail in Chapter 5. Exogenous factors such as gasoline prices are also likely to influence travel behavior, and will be adjusted through regression analysis for before-and-after studies in Chapter 10.

As evidenced in this chapter, longitudinal data provide an opportunity to partition these different sources of variability, whereas cross-sectional data cannot. This notion will be formalized statistically in Chapter 7.

CHAPTER 5

TEMPORAL VARIABILITY IN HOUSEHOLD DEMOGRAPHICS AND ITS IMPACTS ON TRAVEL BEHAVIOR STUDIES

This chapter summarizes key household demographic features generally known to be correlated with household travel patterns and discusses how these features were observed to change over time in the Commute Atlanta sample. The potential impact of changing demographics is at the heart of the uncertainty associated with assessment of whether a transportation policy played a role in changing the participants' travel patterns. In the Commute Atlanta study, the policy measure was mileage-based pricing incentives, as described in Chapter 3.

Potential impacts of pricing incentives are overlaid upon the impacts of other pricing elements (i.e. changes in gasoline price), which introduces uncertainty. However, as will be revealed in this chapter, changing household demographic characteristics appears introduce a much more potent effect on the response results and greater uncertainty in pricing response. Furthermore, these changes may have very different impacts on different households as a function of their original and final demographic characteristics. Large portions of the variability simply cannot be explained within the existing sample. The purpose of this chapter is to address the magnitude of within-household variability potentially introduced by changes in household demographic characteristics, and therefore emphasize the importance that analyses of panel data should control for these confounding factors when aiming at quantifying causal effects associated with changes in household travel patterns over time.

5.1. Unstable Household Demographics

Every trip in the Commute Atlanta database is associated with specific vehicle information, specific household information, and specific primary driver information based on the survey information we received. During the study period, participating households underwent continuous demographic changes. Monthly mailings allowed the households to report changes in time, but some uncertainty is introduced give that demographic changes are not always reported at all or reported accurately in a timely manner. Detailed data processing, as described in Chapter 3, helped reveal these unreported household demographic changes, so these additional uncertainties are likely eclipsed by the impact of the household demographic changes occurring during the study period. These demographic changes raised significant challenges in data analysis and required the implementation of a case study approach (Xu*, et al.*, 2009a).

The Commute Atlanta study differs from traditional cross section travel diary studies in that this study was conducted as a longitudinal study spanning multiple years. Household characteristics such as household size, household structure, economic status, and vehicle ownership change over time. The longer the study duration, the more changes take place.

Of the 95 households in the case study, only 28 households remained stable with respect to all six major demographic characteristics: home location, work status, household structure, income, school(s) attended, and vehicle ownership. Two households had as many as eight major demographic changes over the 21-month study period (see Figure 5.1).

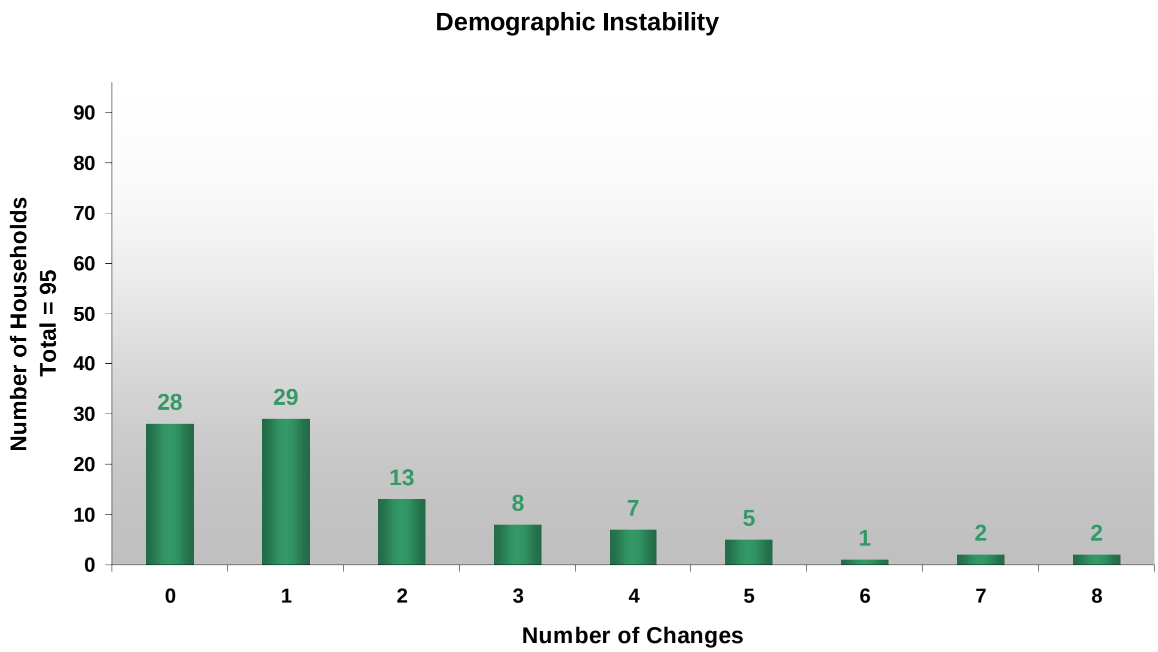

Figure 5.1 Number of Households Experiencing Elements of Instability

Figure 5.2 summarizes the findings of demographic changes across the six categories, wherein each bar represents the percentage of the 95 households that experienced that type of change during the study. When changes are summarized across the six change categories, the most common change experienced was that of vehicle ownership (40%), followed by work status (34%). The bottom part of each bar represents the percentage of the households that experienced a certain change as the only demographic change, whereas the upper part of each bar represents the percentage of the households that experienced a certain change among other types of demographic changes.

The impacts of vehicle ownership changes on household travel appear to be a significant element that will be analyzed using a regression approach in Chapter 9.

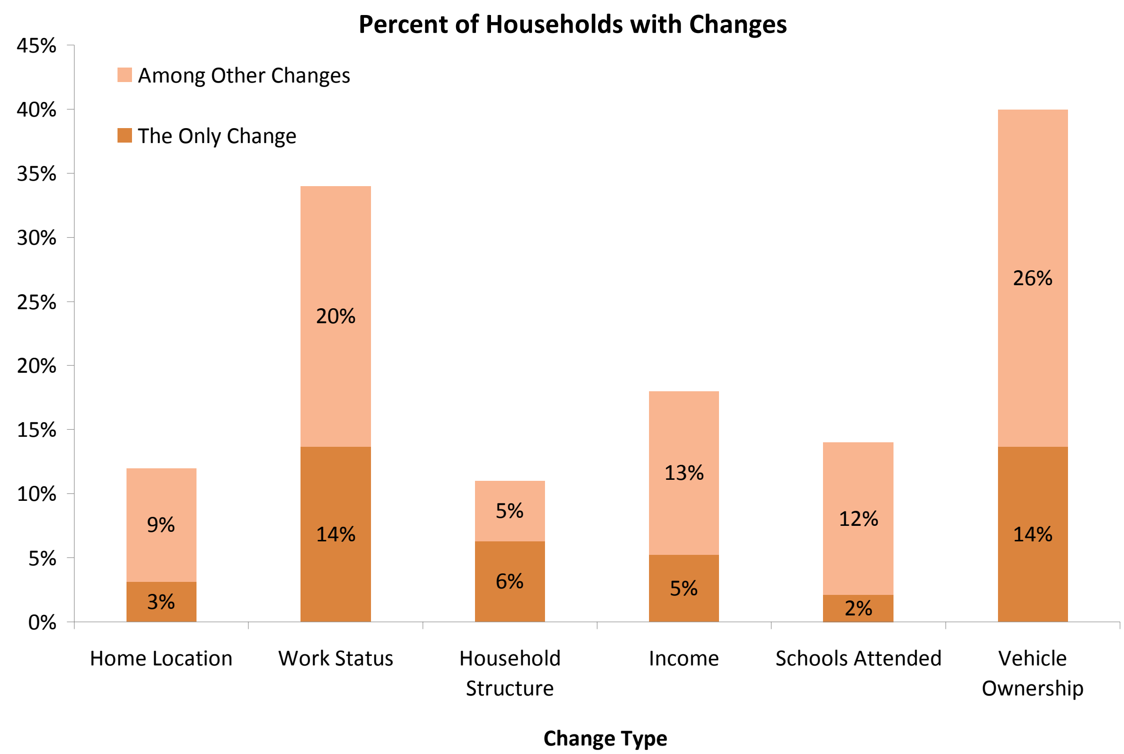

Figure 5.2 Percent of Households Experiencing Each Type of Change

5.2. Impact of Demographic Changes on VMT

Among the households that experienced demographic changes during the study period, the research team determined that the changes reported by 63 households in their monthly surveys could also be confirmed through analysis travel patterns. Of these 63 households that experienced changes, 21 experienced more than one type of demographic change. That is, these 21 households had changes in at least two of the six demographic characteristics. Among the remaining 42 households in which only one type of change occurred, 3 experienced home location changes, 13 underwent job-related changes, 5 exhibited household structure changes, 6 experienced income changes, 2 had school changes, and 13 had vehicle ownership changes. The changes within each demographic category are described in more detail in the following sections.

5.2.1. Home Location Changes Only

One household moved to a location closer to work. Not surprisingly, the research team observed an overall VMT decrease of 48.7% for this household. Based on the analysis of travel patterns by day of week, time of day, and trip purpose, it appears that the change of home locations dominated the reduction. The other two households that changed home location did not display trends that are as clear as the one above. While one could argue that road pricing may incentivize households to move closer to their work, it is not logical to argue that the Commute Atlanta study provided significant incentive. Therefore, any VMT reductions experienced by this household should not be included in an aggregate response analysis.

5.2.2. Work Status Changes Only

The impact of work status changes on travel patterns is probably most obvious among all demographic changes. Almost without exception, total VMT increases when the commute distances increase, or when a new household member starts to work, and vice versa when commute distances decrease or a household member stops working. Pricing incentives are unlikely to have influenced this change. Therefore, any impacts associated with job changes need to be controlled for in any aggregate response analysis.

5.2.3. Income Changes Only

The relationship between the direction of income change and that of VMT change was very interesting. Only one household exhibited a VMT decrease when household income decreased. The self-employed, senior citizen in this one-person household retired from a full time job during the study period (i.e., correlated with work status change). The household income went from between $30,000 and $75,000 to below $30,000 and VMT did decrease accordingly. Another household increased VMT as income increased. In the other four cases, the change in VMT went in the opposite direction of income change. Two (2) households experienced income increase from between $30,000 and $75,000 to between $75,000 and $100,000, but their VMT decreased by 24% and 35% respectively. The other 2 household experienced an income decrease but the VMT in both households increased by approximately 9%. In one household, the income went from above $100,000 to between $30,000 and $75,000. In the other household, the income went from above $100,000 to between $75,000 and $100,000. But, none of these households had annual incomes below $30,000. On the contrary, most households had annual incomes above $75,000. There are not enough households experiencing these changes to draw any conclusions, but additional work in this area seems warranted to determine if there is a direct relationship and if any income threshold levels may play a role in income change vs. VMT change.

5.2.4. Household Structure Changes Only

Among the 5 households in which household structure changes were the only type of change, 4 exhibited intuitive trends. VMT increases when new members join the household and vice versa. In one household, however, VMT increased after a spouse (wife) passed away. However, without further information about this household, such as the health conditions of the husband and the wife prior to passing, it is very difficult to draw any conclusions even for this single case study.

5.2.5. School Attended Changes Only

Changes in schools attended usually coincided with other types of changes, especially when households move or change jobs. There are only two cases in which changes in schools attended were the only type of the changes. In one case the two children in the household went to college and graduate school. Although the household did not report to the research team whether or not the two children moved out of the household, a 12.8% VMT reduction was observed in this household. In the other case, the child changed levels of school from middle school to high school (freshman level, non-driver) and the impact on total household VMT was not discernible.

5.2.6. Vehicle Ownership Changes Only

The VMT changes across the 13 households experiencing vehicle ownership changes exhibit mixed results. The changes range from -39.6% to 105.7%. A vehicle purchase does not necessarily lead to an increase in VMT. Likewise, selling a vehicle does not necessarily come with VMT reductions. The impact of vehicle ownership on total travel and upon specific travel by region and purpose warrants further investigation. To this end, pricing studies need to ensure that driver-vehicle relationships are updated in as timely a manner as possible. Hand-me-down vehicle drivership (selling an older car, changing the old commute vehicle into a general purpose vehicle, and purchasing a new commute vehicle) occurs in the households and can be observed through analysis of changes in travel patterns across vehicles when a vehicle change occurs.

5.2.7. Children Turning 16

Six households had children that turned 16 during the study period. Generally, none of these households were able to reduce their total VMT. Because the sample size is very small, it is basically not possible to single out the impact of children turning the driving age on the household travel patterns. Table 5.1 summarizes the characteristics of these 6 households. The demographic instability of these households makes the analysis especially difficult. More advanced statistical methods will need to be adopted to analyze the impact of children turning 16 on household travel behavior, and will be reported separately in future research. Another major obstacle is that the households did not usually report the percentages of time that each household member drove certain vehicles in a timely manner. Thus, the research team often was not able to tell if the child who turned 16 actually started driving (additional detailed analysis of trip end locations will be required to make this determination). Future studies need to link new drivers to their vehicle activity through special-purpose surveys triggered when the child turns 15 (learner's permit) and 16 (licensure).

Table 5.1 Summary of Households with Children Turning 16

Total VMT Change	Household Structure	Major Change Type	GT Sample Group
-0.5%	Single mother with one child	None	3
2.2%	Single mother with one child	School attended	3
20.5%	Couple with two children	One person started a new job Vehicle turn-over	8
27.1%	Couple with three children	One child moves to college Wife starts full time job One child starts to drive	5
34.2%	Couple with three children	Schools attended Work status and income drop Vehicle turnover	5 in the baseline period 1 in the pricing period
43.9%	Couple with four children	One child moved out to college Income increase Discrepancies between reported and revealed demographic changes	7 in the baseline period 8 in the pricing period

5.3. Overall VMT Changes and Demographic Characteristics

Overall VMT changes during the nine months of the pricing period in all households range from a 12,172 mile decrease to a 9,597 mile increase. As shown in the overall histogram chart in Figure 3, a few cases present extreme VMT reduction or increase. On the quarterly basis, Quarter 1 observes the highest number of households with VMT reductions. This reduction likely resulted from the high gas prices after Hurricane Katrina, which occurred during the Quarter 1 pricing period. Quarter 3 displays the most cross-sectional fluctuations, which could possibly be attributed to the seasonal effects.

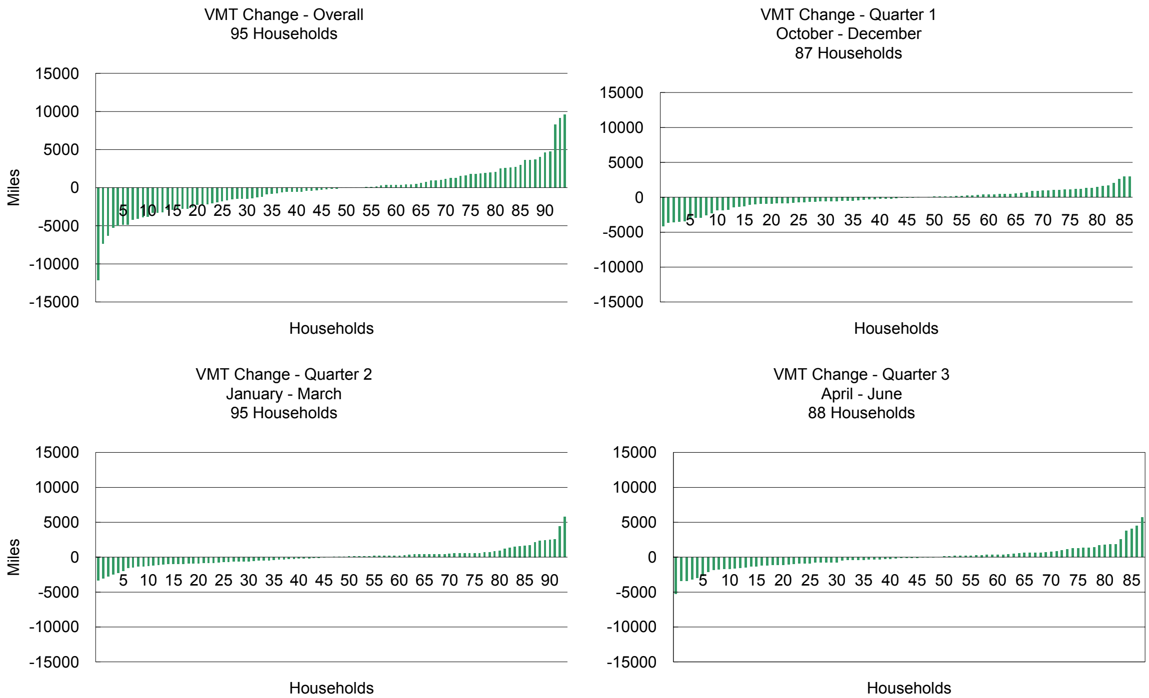

Figure 5.3 VMT Changes between Baseline and Pricing Periods

5.3.1. Households with VMT Reductions

A total of 53 of the 95 households exhibited a reduction in VMT, ranging from -63.1% to -0.2%. In 12 of these 53 households, demographic changes were the likely reason for the reductions. Nine (9) households experienced work status changes, including retirement and job location changes that reduced commute distances. Such changes appeared to be dominant contributors to the VMT reduction. One (1) of the 53 households, composed of a single person in retirement, experienced a significant decrease in income during the study period. The income drop combined with the pricing incentives may be related to the noted VMT reduction of 4.7%, as opposed to the pricing incentives. Changes in household structure were the likely causes for VMT reductions in two households. In one five-person, multi-generational household, the fact that the daughter moved out to college seems to have led to the 28.5% reduction in VMT based upon travel pattern analysis.

Of the remaining households, the research team finds it difficult to attribute VMT reductions to any particular demographic changes. Ten (10) are likely to have been affected by the pricing scheme within the context of their household demographic characteristics. Eight (8) of these 10 households consisted of household members older than 60, 5 of the 10 households were single-person households and 9 of the 10 did not experience any major demographic changes. In the one household that had a demographic change, a rise in income was observed, but did not seem to affect the household's travel pattern significantly (and would be expected to increase rather than decrease travel). The consistent trend of decreasing VMT across quarters for this household makes the research team suspect that the reduction was associated with pricing incentives (the combined effect of gasoline price increase and experimental pricing incentives). In this household, one member started to take transit in the pricing period, contributing to their 23.9% change in VMT. However, this change in travel behavior could as well be attributable to a change in employer policy (e.g. provision of free MARTA passes) that was not reported. The accessibility of the work location to a MARTA station makes the change of travel modes possible for this household, but not for the majority of other households in the study. Future studies will need to incorporate focus groups to ascertain the causal relationship.

The reasons behind the VMT reductions exhibited in 33% of the households are unclear. Among them, 14 households remained demographically-stable during the study period from October 2004 to June 2006. Among the remaining 17 households with major demographic changes, 8 had changes in vehicle ownership. The impact of vehicle ownership changes is not obvious from the current level of analysis and therefore warrants further study. The other 11 households had various changes in other aspects of demographic characteristics, but the impacts were not intuitive. For these households, it is especially important to conduct home interviews so that the research team can learn about the travel decision making process. It would also be helpful if the record for household changes was more precise, providing more details of the exact date of the changes.

5.3.2. Households with Increased VMT

Forty-two (42) households exhibit increases in VMT, ranging from 0.2% to 130.6%. In 11 of the 42 households, demographic changes were the likely reasons for the increase. Ten (10) households experienced work status changes, some including job location changes that increased reduced commute distances and starting new jobs. Such changes appear to be major contributors to the VMT increase. One (1) of the 42 households that experienced a significant household structure change wherein two 2-person households moved in together, causing the two-person household to become a four-person household. Not surprisingly, this is the household that experienced an increase in VMT of 130.6%.

Of the remaining 31 household, it was difficult to explain why the VMT increase was observed. Nine (9) of the 31 households remained stable during the study period, yet the increase in VMT reached as high as 53.1%. Eight (8) of the 31 households had more than one major demographic changes and most often theses changes would affect VMT in different directions. It is difficult to single out the impact of each change. On rare occasions, especially when the research team had detailed records of demographic changes, the impacts of changes on travel patterns could be separated, as showcased in the third example case study presented later. Of the remaining 14 households in which only one demographic change took place, two (2) households changed home locations, one (1) changed work locations, two (2) changed household structures, one (1) changed schools attended, three (3) changed income, and five (5) changed vehicle ownership. These changes did not have impacts on travel patterns that met research expectations and warrant further surveys on the households' travel decision making process.

As mentioned earlier, four (4) out of the five (5) households where a child turned 16 during the study period also displayed VMT increase. Although this is likely the cause of much of the VMT increase, it cannot be attributed to the new driver without much more detailed trip-level analysis as all 4 households experienced multiple demographic changes and it is difficult to single out the effect of driving age children.

5.4. Analysis across Demographic Characteristics

5.4.1. Summary by Sample Groups

During the study period, 76 households remained within the same sample group for the entire 21-month period. To focus on exploring pricing sensitivity in the context of demographic characteristics, these 76 households were considered relatively stable, despite the other various demographic changes these households experienced.

The percentage of households with VMT reductions among all households within the same sample groups tends to decrease as the households become larger in size and have higher incomes. However, this trend is not statistically significant given the small sample size and should not be generalized. Even though these households stayed in the same sample groups, there are plenty underlying demographic changes that could have affected this trend in one direction or the other. Figure 5.4 summarizes changes of VMT in these relatively stable households by sample groups.

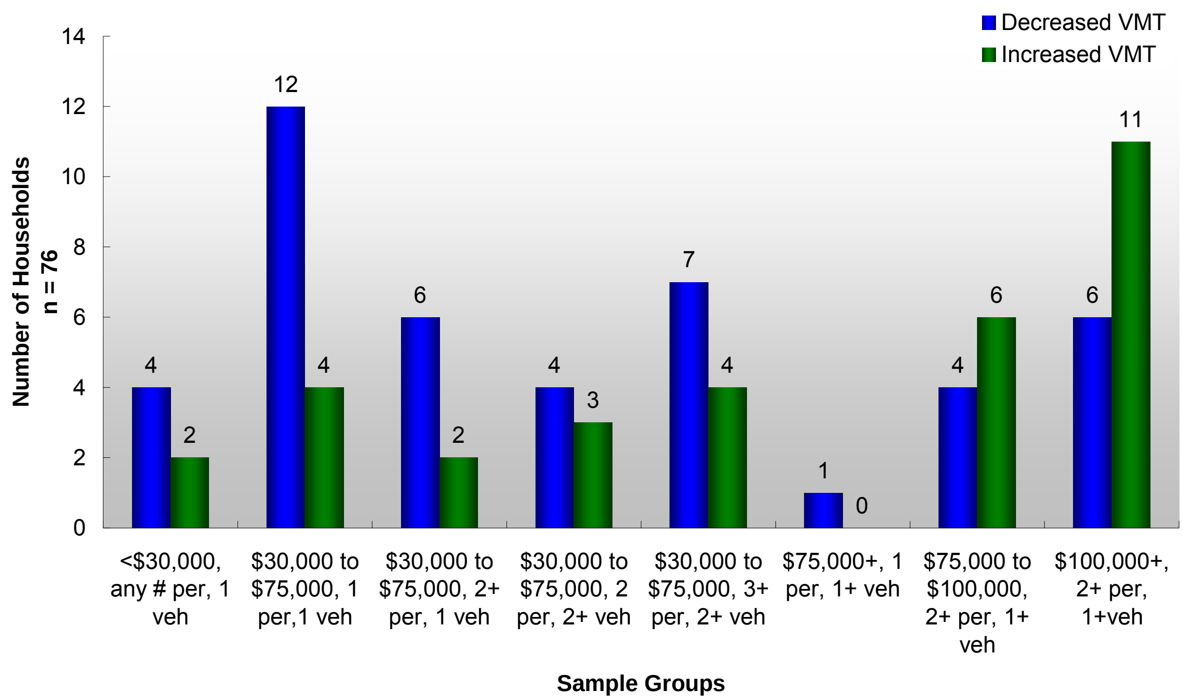

Figure 5.4 VMT Change: Distribution of Relatively Stable Households by Sample Groups

5.4.2. Senior Households

Thirty-three (33) households include household members older than 60, and 19 of these 33 households reduced VMT from the baseline period to the pricing period. Thirteen (13) of the 33 households were stable in terms of demographic changes and 10 of these 13 reduced VMT. Lower income senior households may have responded to the pricing scheme. As the household income increases, senior households may be less responsive to pricing incentives even if they have the flexibility to change their travel patterns. Again, however, the trend is not statistically significant. In addition, trip purpose information may be crucial to evaluating the potential response of senior household travel to economic incentives. Factors such as biomedical needs of the senior household members, their routine social activities such as visits to friends, church and community services, and how independent the senior people are of their children or other types of assistance may play a large role in their response.

5.4.3. Households in which Workers are Self-Employed

Eight (8) households in the sample reported that the workers in the households were self-employed. None of these households reduced VMT, despite the fact that one household experienced a significant decrease in annual income from above $100,000 to between $30,000 and $75,000. No clear trends in travel pattern changes are identifiable, either at the overall level or by time of day and by trip purpose. To better understand the travel patterns for these households, more information is needed, such as the type of business, operating hours and days, business travel needs, and percentage of commercial use VMT for each vehicle (the Commute Atlanta survey questions only identified whether a vehicle was used for business purposes[1]). Identification of each business-related trip, which would be very difficult to undertake in a longitudinal study involving thousands of trips, would help to quantify household travel responses (if any) to pricing.

[1] Elango, *et al.* (2007) pointed out that vehicles reported as used for business purposes display significantly different travel patterns from other household vehicles and need to be analyzed separately.

5.5. A Closer Look at Three Case Studies

5.5.1. Case I: Mileage Reduction due to Work Location Change

This household consists of a single, full-time working mother with 2 children. The participant changed her work location in the middle of March 2006 and her one-way commute distance decreased from around 17 miles to around 7 miles. Figure 5.5 shows the daily average VMT of weekday inside-of-region travel by month during the baseline and pricing periods. Daily average VMT were significantly lower for weekday inside-of-region travel from April to June 2006 after the work location changed. Researchers also observed significantly VMT reduction in October 2005 compared to the same month in 2004. However, a detailed check of the commute pattern revealed that the person only commuted to work 9 days in October 2005 compared to 18 days in October 2004 and was outside-of-town for around 10 days in October 2005. Hence, the lower October result is likely not attributable to pricing.

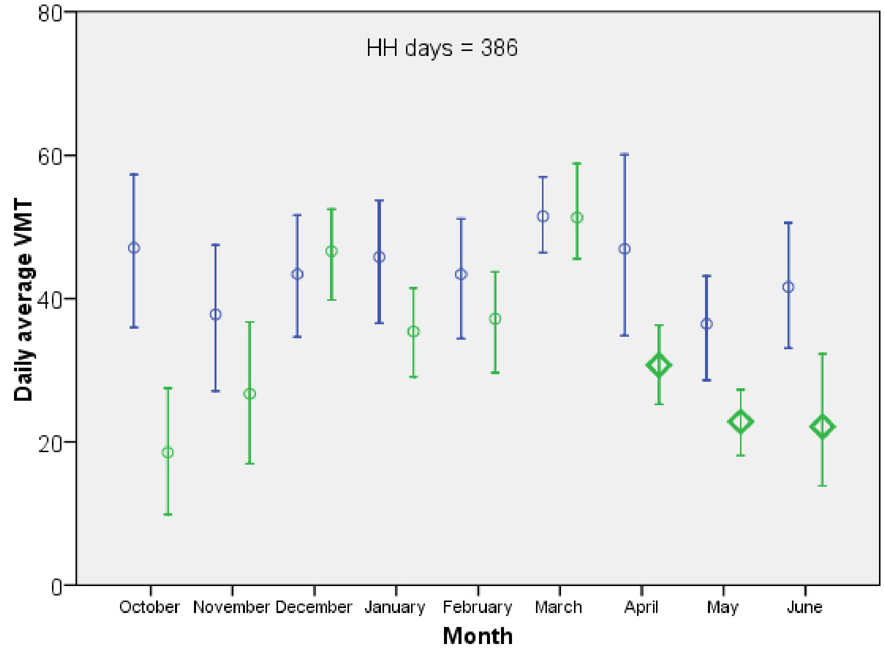

Figure 5.5 Weekday Intra-Regional VMT by Month

5.5.2. Case II: Mileage Reduction for a Retired Household

This household is composed of one low-income ($20k-$29k) retired person. Figure 6 shows the inside-of-region and outside-of-region travel for this household by month, with the net incentives juxtaposed. This household seems responsive to the incentives (including gas incentives and the pricing incentives) since mileage reduced for all the pricing months compared to the baseline months. Researchers observed especially large reductions in October 2005 and April 2006. October 2005 is the first month when pricing initially started and April 2006 is the first month when pricing incentive increased to 15cent/mile. Gasoline prices also experienced spikes during these two months compared to the same months previous year. For outside-of-region travel, the researchers observed a very stable pattern in VMT per month. It looks like that there are two or three destinations that the household visited regularly. These travel needs are hard to eliminate and hence stayed consistent during both baseline and pricing periods.

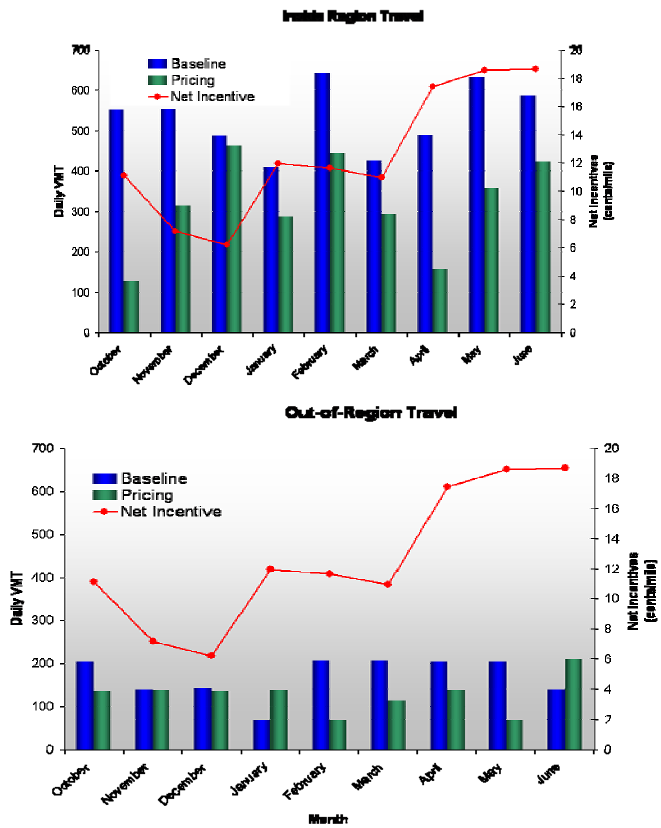

Figure 5.6 Total VMT by Month (Intra-Regional and Outside-of-Region Travel) with Net Incentives Overlay

5.5.3. Case III: Multiple Demographic Changes and the Combined Impact on Travel Patterns

This household went through seven demographic changes during the 21-month study period. This household consists of a two-person household with two vehicles. The two participants in this household were roommates and were not related. Both participants were in the 20-30 age group. Both were in graduate school from the start of the study until May, 2005, and both started working in June, 2005 in downtown Atlanta.

One member moved out of the household (with her vehicle) in early April, 2006. Due to all these changes, the household income fluctuated, from below $30,000 per year, to above $ 75,000, and then to between $30,000 and $75,000 thousand per year after the move-out.

Overall, the household had a 13.1% decrease in VMT, primarily because during the third quarter in Phase II one of the household members moved out and the mileage from this individual no longer contributed to the household total. This major demographic change was the primary factor in the overall 13.1% decrease in VMT. Figure 5.7 illustrates the overall decrease of VMT in the third quarter caused by the household structure change.

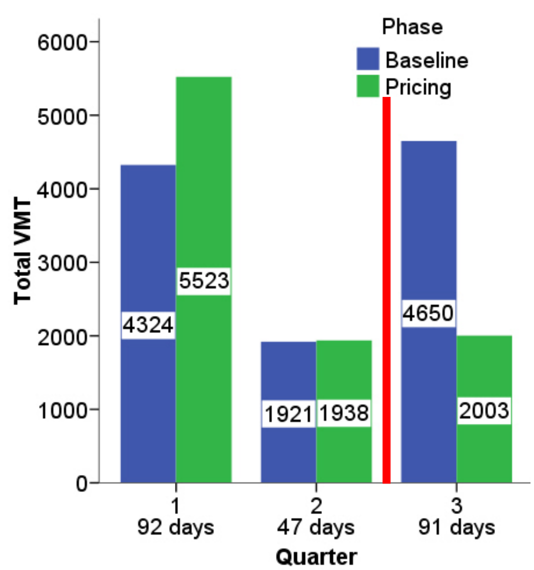

Figure 5.7 VMT Change by Quarter

More detailed analysis of travel patterns by day of week and time of day helps reveal the impact of work status changes on VMT change. There are apparent differences between the lifestyles of a student and a full time worker. When the household members started working, they traveled more in the morning on weekdays and became more active during weekends and holidays. The research team was able to explain the travel pattern changes for this household because this household provided detailed information for every major demographic change.

5.6. Implications on Sample Size Planning for Before-and-After Studies

The total vehicle miles of travel (VMT) within the 95 households for which valid baseline and pricing data were collected decreased by about 3% over the 9-month pricing period. However, the noted reduction is not statistically significant due to large between-household and within-household travel variability. That is not to say that a reduction in vehicle miles of travel did not occur, only that the observed reductions were small, and that researchers could not state with any degree of certainty that the noted reduction in travel was associated with the pricing incentives or any other causal factor. The change was small enough that the reductions could simply be associated with random chance.

The detailed case studies for the 95 households were designed to examine changes in household travel patterns in response to pricing incentives (gasoline price increases coupled with mileage incentives). Each case study was conducted within the context of noted changes in household demographic and employment conditions, such as home location, work status, household structure, income, schools attended, and vehicle ownership. These individual case studies are an interesting read as a stand alone report, and all case studies are presented in the Appendix to the Commute Atlanta Phase II case study report (Xu*, et al.*, 2009a). The biggest finding in this research was the magnitude of variability in household travel behavior. These case studies provided clear and convincing evidence that there is both natural intra-household variability as well as some potentially extreme variability resulting from demographic changes in such panel studies. The clear inference is that much larger sample sizes and improved survey design will be required in longitudinal studies to ascertain how pricing affects travel behavior. Therefore the findings of similar studies[1] that have been conducted should be eyed with caution and that researchers need to be careful in drawing any conclusions on the impact of pricing incentives from these studies.

Based upon analysis of the Commute Atlanta households, mileage incentives had little discernable impact on VMT reduction, due to large inter- and intra- household variability and demographic instability. 10 households were identified in which the pricing incentives may have influenced their travel behavior. These households are generally stable in terms of key demographic characteristics. Nine (9) of these 10 households had annual incomes below $75,000, three (3) of which had incomes below $30,000. Eight (8) of these households consist of people older than 60. Half of these 10 households are single-person households. The reductions observed in these households may be associated with pricing incentives, but other unknown factors could also have contributed to the reductions. Such correlation would need to be verified by stated preference surveys.

Among the demographic changes, the impact of work status changes was most discernible. Home location changes and household structure changes are also important sources of VMT change. Due to the small sample size, it is unclear how changes in income, schools attended, and children turning 16 affect travel behavior. The impact of

---

[1] As mentioned in Chapter 2, these studies include the reports for Minnesota Mileage-Based User Fee Demonstration Project, the Oregon Road User Fee Pilot Program, and Washington Traffic Choice Study

vehicle ownership change on VMT also requires further investigation, as will be demonstrated in Chapter 9.

In conclusion, on the disaggregate level, mileage incentives may have affected some households within the context of gasoline price fluctuations and household demographic changes. However, there are no statistically significant findings that can be reported. Significantly larger samples (much larger numbers of participating households within each demographic stratum, and much larger samples than what have been reported in other relevant studies reported in the literature) need to be coupled with more intensive surveys with improved designs. Experimental controls over households that use their vehicles for business purposes also need to be introduced, possibly as a separate recruitment stratum. Accessibility to viable commuter transit is also an essential control variable in future studies, as households that have viable transit access maybe much more likely to respond to economic incentives. Other confounding factors are many, and the analyses in this chapter has demonstrated the importance of adjusting for such confounding factors in the design and analysis stages of a panel study.

CHAPTER 6

THE TREND OF MEANS VERSUS SURVEY LENGTH

This chapter examines the impacts of survey length and sample size on the precision level of key travel behavior variables for both intra-regional and long-distance travel. Previous studies have shown sample size benefits of conducting short multi-day travel surveys, but no one has studied sample size impacts if the survey spans for multiple months or even years. For example, using the Reading data, Pas (1986) has shown how three-day data from a 75-person sample and two-day data from a 91-person sample give the same level of precision in parameter estimates as a 1-day sample of 136 persons. Additionally, Stopher et al. (2008b) have presented that a 7-day GPS survey offers a 65 percent reduction in sample size needed for a conventional one-day diary survey, and a 15-day GPS survey offers a 70 percent reduction. The length of the Commute Atlanta study has provided an excellent opportunity to not only provide additional evidence that multi-day travel surveys offer sample size savings, but also to expand the existing findings to a wider range of survey lengths.

Using both the original data from the Commute Atlanta study and synthetic data re-sampled from the original dataset, the sample size impacts of various survey lengths are examined for both intra-regional and long-distance travel.

6.1. Intra-Regional Travel

Two inferences were chosen for analysis: number of trips per household per day and VMT per household per day. The analysis pools two survey waves together—baseline and pricing. Each wave is for 9 months (273 travel days).

6.1.1. Observations from the Original Sample

First, the impact of longer survey lengths on inference estimation is first examined in Figure 6.1 and Figure 6.2. These two figures show the results of successively adding the means over the 273 days of the two waves for number and VMT of intra-regional trips per day. In both figures, the results show instability in the mean for the first 60 days and then show a gradual trend to stabilize through the 90th day. After 90 days, the number of intra-regional trips stabilizes to around 5.8 and the intra-regional VMT stabilize to around 37. To understand the instability of the means during the first 60 days, a look at the sampling distribution of the means will be provided through Monte Carlo studies in the subsequent sections.

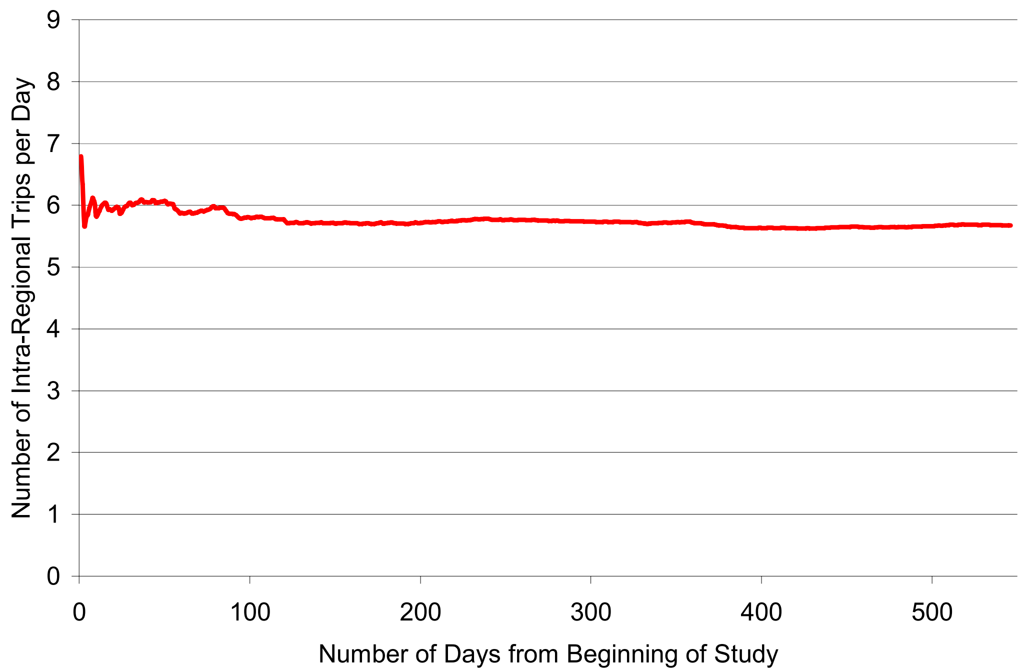

Figure 6.1 Cumulative Mean of Daily Number of Intra-Regional Trips for Different Survey Lengths with the Fixed Start Date (October 1, 2004).
m=95 households, n=546 repeated observations.

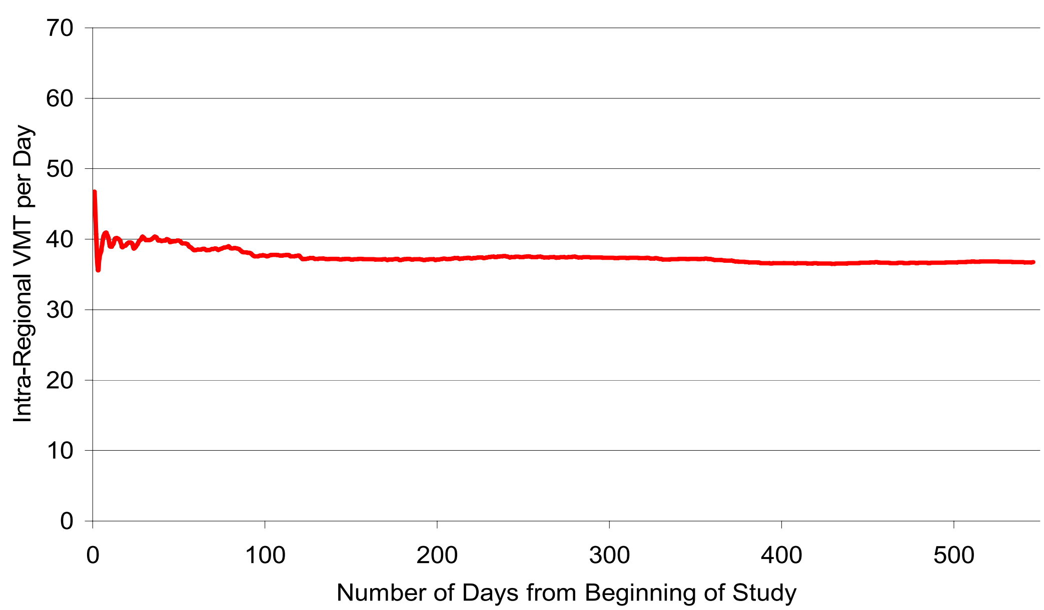

Figure 6.2 Cumulative Mean of Daily Intra-Regional VMT for Different Survey Lengths with the Fixed Start Date (October 1, 2004).
m=95 households, n=546 repeated observations.

6.1.2. Monte Carlo Study of One-Month Samples

To remove the possible seasonal effect as discussed in Section 4.2 on the patterns observed in Figure 6.1 and Figure 6.2, a re-sampling technique is applied to examine the trend of means and confidence intervals as the sample size and sample size vary. The general procedure of resampling is to first randomly draw a household, and then randomly draw a date and include the data from a certain number consecutive days thereafter. This procedure is repeated 1000 times to produce a bootstrap distribution of the means. The first exercise includes one month of data for each drawing of a household. The next section simulates a 2-day travel survey.

Since the transportation modelers and planners are mainly interested in weekday travel, the analysis focuses only on weekdays. The weekday records are extracted from the original dataset to form the pool for resampling. The number of days for each drawing of a household is 20 (5 weekdays×4 weeks in a month). Figure 6.3 and Figure 6.4 show the results of cumulative means for the number of intra-regional trips per weekday and intra-regional VMT per weekday, respectively.

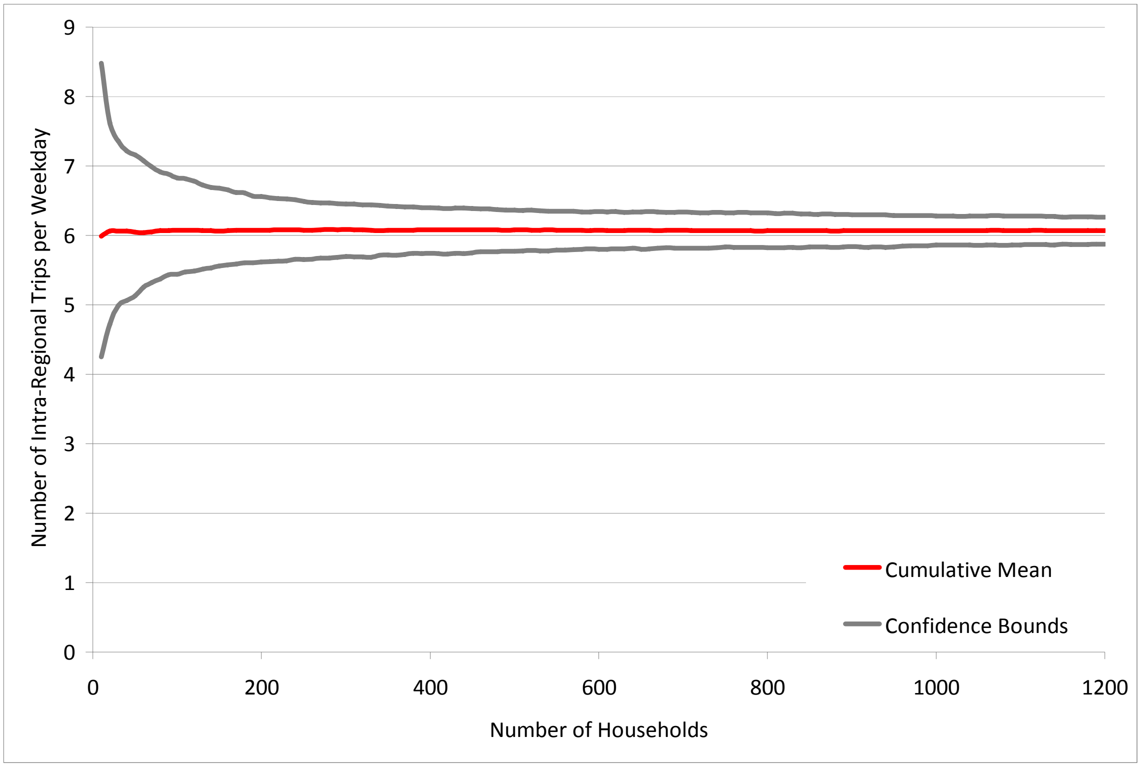

Figure 6.3 Cumulative Mean Number of Intra-Regional Trips per Weekday as Sample Size Increases; Survey Length = 20 Weekdays

The wide confidence interval of number of trips and VMT shown in Figure 6.3 and Figure 6.4 emphasize the instability of the means when sample size is small. In both figures, the confidence intervals start to tighten after 200 households. The number of intra-regional trips per weekday stabilizes to around 6 and the intra-regional VMT per weekday stabilizes to around 41, both higher than the values observed in Figure 6.1and Figure 6.2 where both weekdays and weekend days are included.

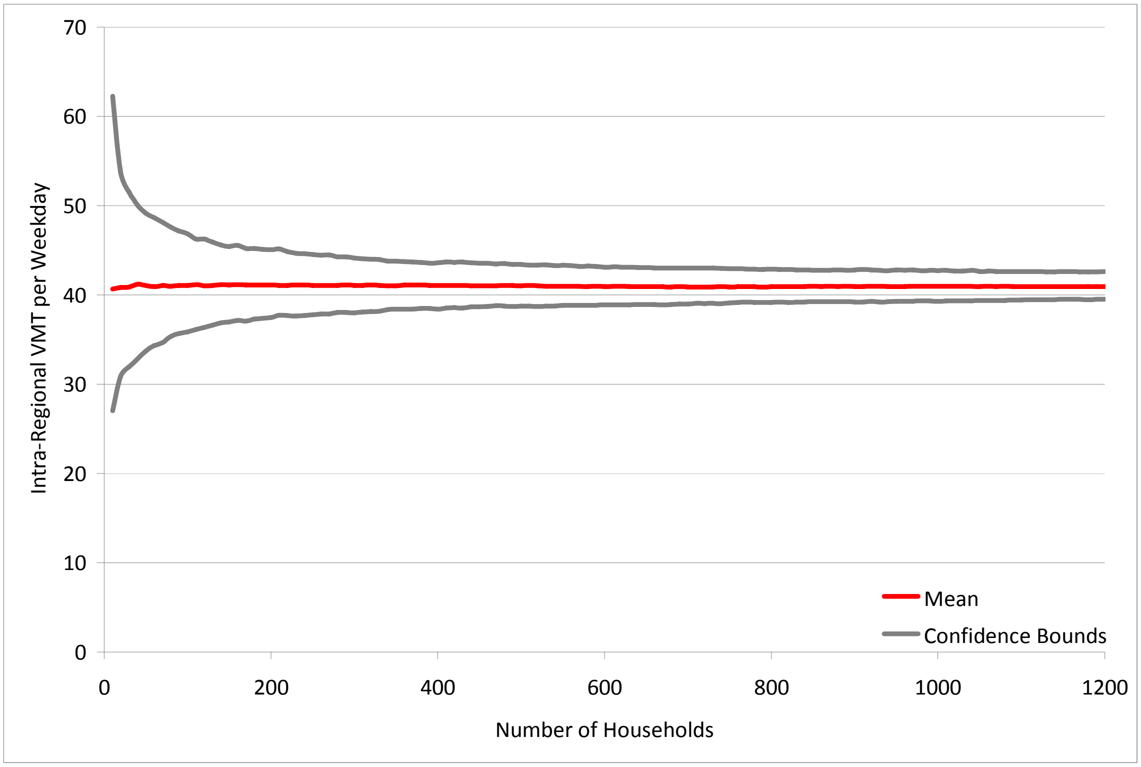

Figure 6.4 Cumulative Mean Intra-Regional VMT per Weekday as Sample Size Increases; Survey Length = 20 Weekdays

6.1.3. Monte Carlo Study of Two-Day Samples

To compare the impact of a longer survey period on the trend of cumulative means to that of conventional two-day travel diaries, two consecutive weekdays are re-sampled randomly at each drawing of a household. The resulting cumulative means of number of intra-regional trips per weekday and intra-regional VMT per weekday are presented in Figure 6.5 and Figure 6.6, respectively.

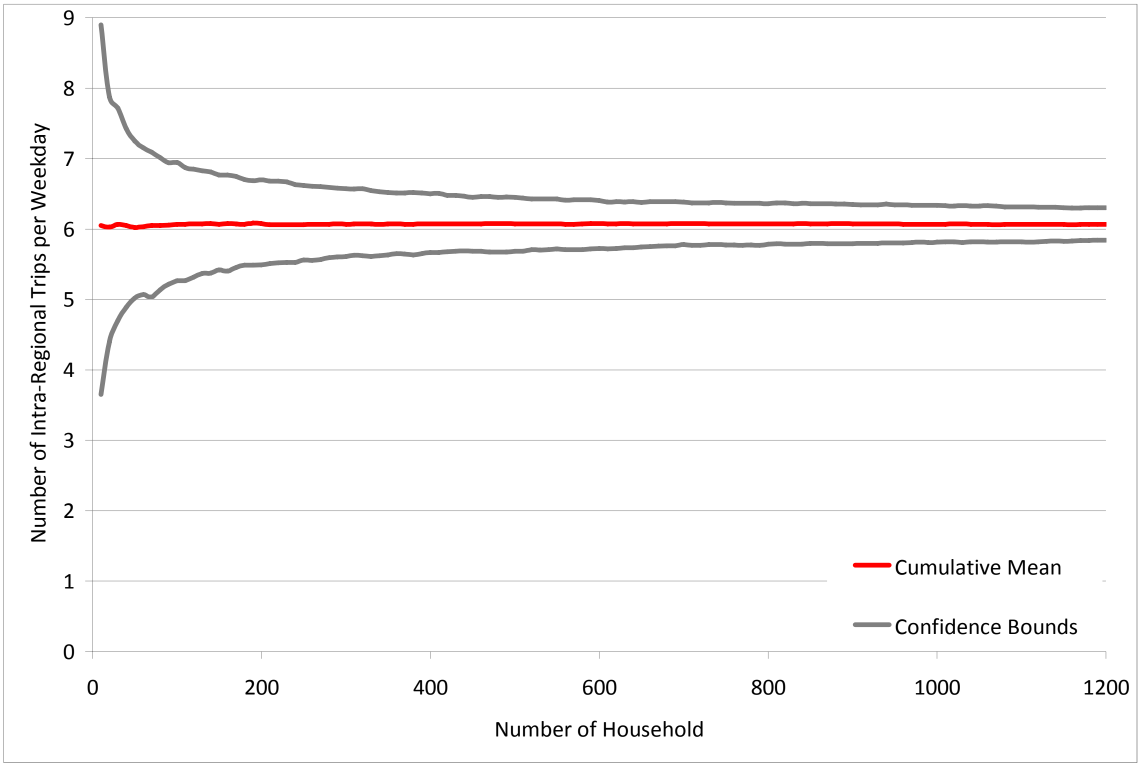

Figure 6.5 Cumulative Mean Intra-Regional Number of Trips per Weekday as Sample Size Increases; Survey Length = 2 Weekdays

When the survey length is 2 weekdays, the cumulative means show similar trends as those when the survey length is 20 weekdays—the confidence intervals are very wide when the number of households is below 200, and start to tighten after the sample size reaches 200 households. The number of intra-regional trips per weekday stabilizes to slightly more than 6 and the intra-regional VMT per weekday stabilize to around 41. If the data conformed to normal distribution and the observations were independent, the widths of the 95% confidence interval shown in Figure 6.5 and Figure 6.6 should be about 10 times as wide as those shown in Figure 6.3 and Figure 6.4 because the survey length is 20 weekdays in the previous case and 2 weekdays in this case. However, the widths of the confidence intervals in Figure 6.5 and Figure 6.6 are only slightly tighter than those in Figure 6.3 and Figure 6.4. For example, with 1200 households, the two-day sample provides a confidence interval of 3.79 for daily intra-regional VMT whereas the 20-day sample gives a confidence interval of 3.11. This phenomenon can be explained by within-household association that will be further explored in Chapter 8. When estimating group averages, positive correlation increases variance (Diggle*, et al.*, 2002). Therefore, as the survey length increases, the widths of the confidence intervals do not decrease at a fast rate as they would if new independent observations, i.e. new households, were collected.

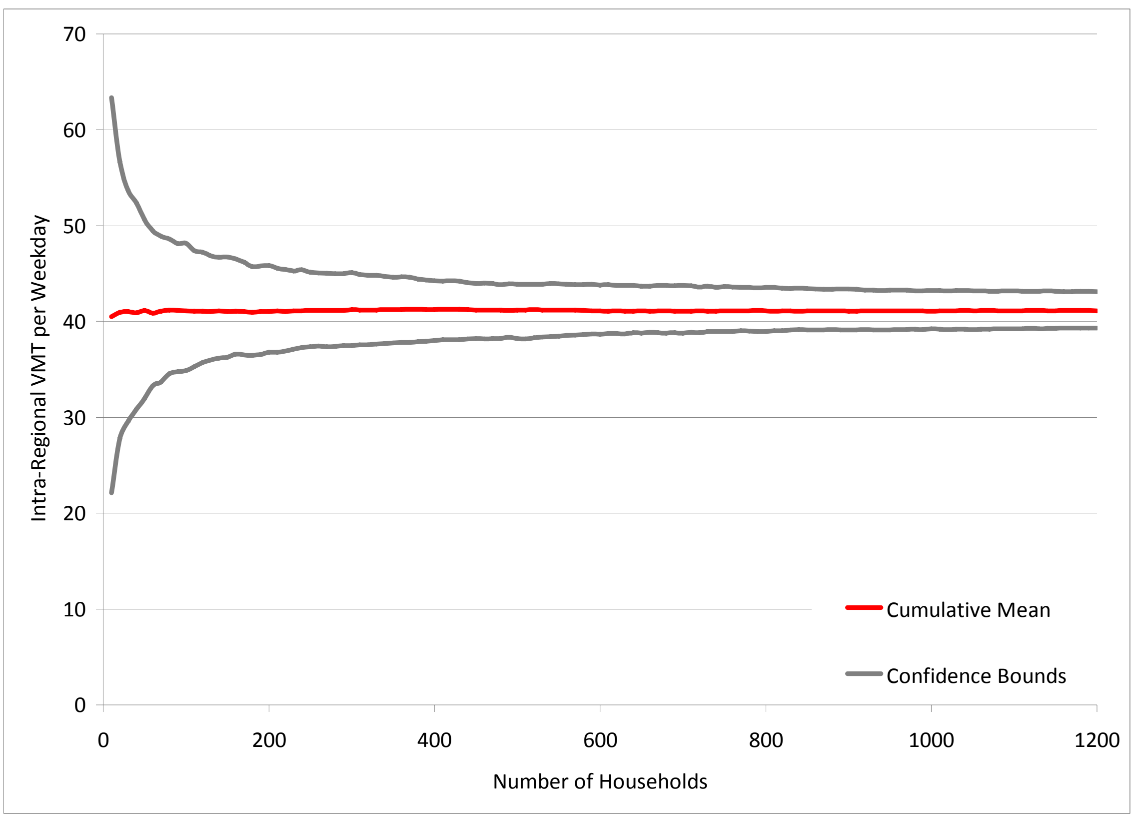

Figure 6.6 Cumulative Mean Intra-Regional VMT per Weekday as Sample Size Increases; Survey Length = 2 Weekdays

Even with high degree of within-household association in weekday intra-regional VMT, the savings on sample size (measured in number of households) associated with conducting a longer survey are obvious, as shown in Figure 6.7. Figure 6.7 overlays Figure 6.4 onto Figure 6.6, and zooms in on the upper confidence bounds. To achieve 10% relative precision of the estimate for the average intra-regional VMT per weekday, the upper confidence bound should be at about 45[1], which would require 250 households for a 2-day survey, or 170 households for a 20-day survey.

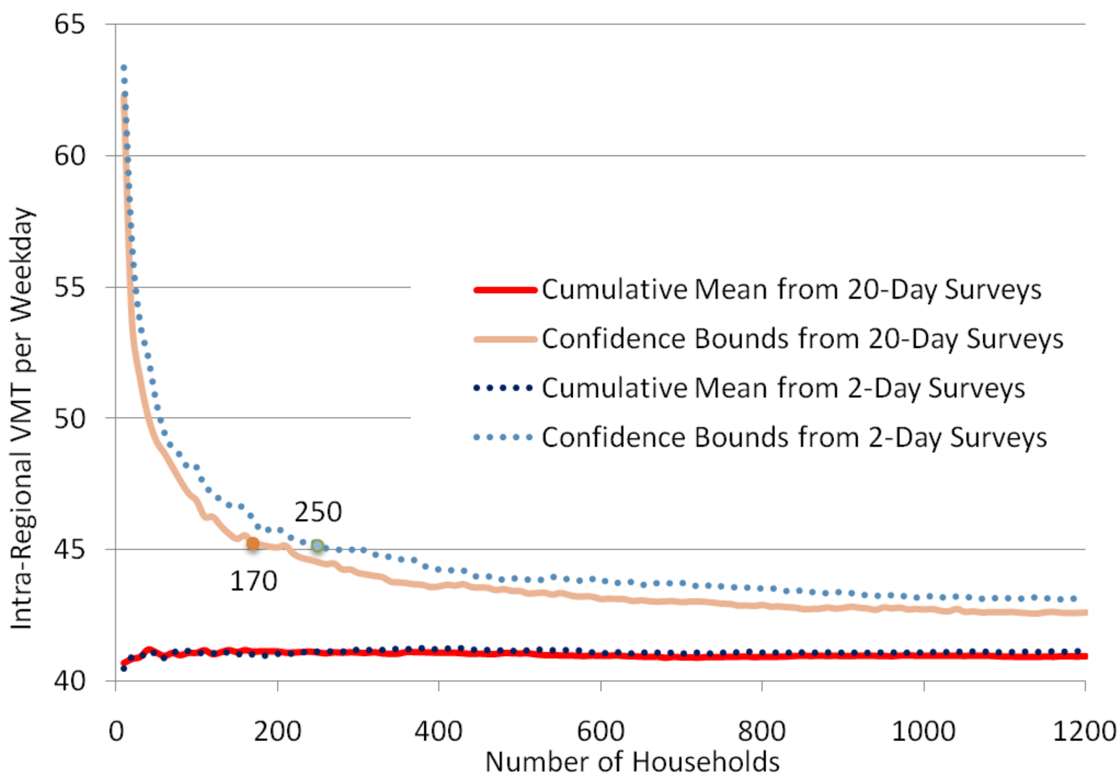

Figure 6.7 Overlay of Cumulative Means from 2-Day and 20-Day Surveys

[1] The sample mean is 41 VMT per weekday, so (45-41)/41 ≈ 10%.

6.2. Long-Distance Tours

The inferences for long-distance travel behavior analysis are number and VMT of long-distance tours per month, given that long-distance travel is a relatively rare event. The dataset for long-distance travel consists of 30 months of data from 94 households, as described in Chapter 3. January, 2004 is the starting month, i.e. Month 1.

6.2.1. Observations from Original Sample

Similar to the analysis of intra-regional travel, the impact of longer survey lengths on inference estimation is examined first, in Figure 6.8 and Figure 6.9. Both the number and VMT of monthly long-distance tours show the tendency to stabilize as the survey length increases, with the fixed start date of the study on January 1, 2004. The seasonal impact is more prominent in long-distance travel than intra-regional travel. As the survey starts in January, 2004, both the number and VMT of long-distance tours show very low values, and then continue to increase steadily through Month 7, which is July, 2004. This trend confirms the seasonality discussed in Section 4.2.2. After the 7th month, the curves still show some evidence of seasonality, but the effect of pooling more data each month becomes dominant, in that the curves tend to level off during the last six months of the survey.

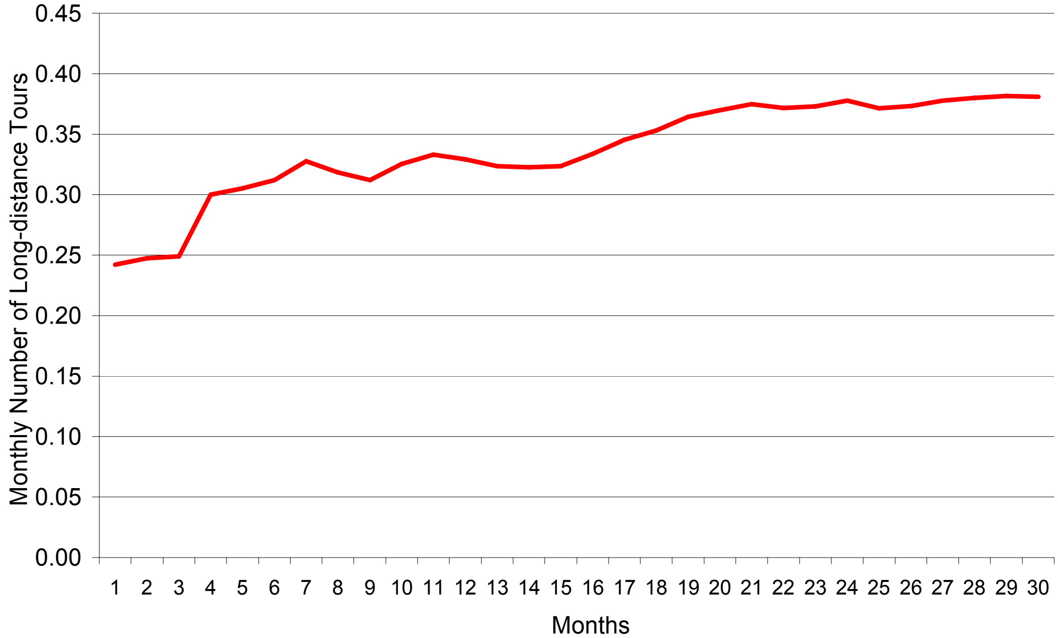

Figure 6.8 Cumulative Mean of Monthly Number of Long-Distance Tours for Different Survey Lengths with Fixed Starting Month of January 2004
m=95 households, maximum *n*=30 repeated observations.

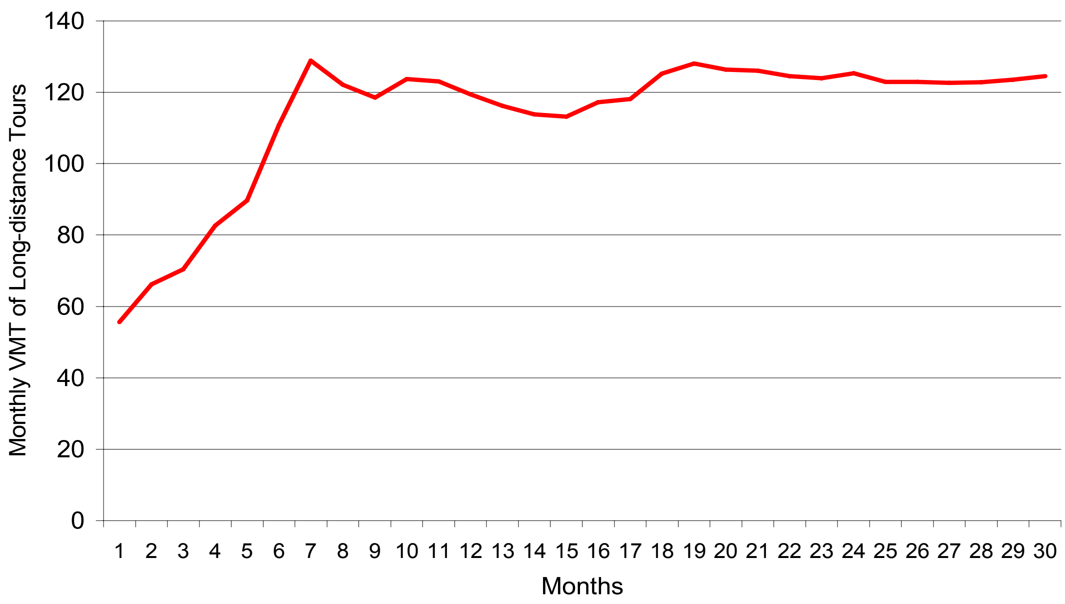

Figure 6.9 Cumulative Mean of Monthly VMT of Long-Distance Tours for Different Survey Lengths with Fixed Starting Month of January 2004
m=95 households, maximum *n*=30 repeated observations.

6.2.2. Monte Carlo Study of Two-Day Samples

The motivation of this analysis is to examine to capability of two-day travel diaries to identify long-distance tours. A synthetic pool of two-day samples is generated the same way as described in Section 6.1.3, except for the fact that weekend days are included, given the nature of long-distance travel.

Table 6.1 summarizes the long-distance tours identified by two-day surveys as the sample size increases. Section 6.1.3 showed that a sample of 200 households with two-day surveys can provide relatively stable means (with relative precision a little more than 10%) of key travel behavior variables with regard to intra-regional travel. In the context of long-distance travel, however, a sample size of 200 household does not seem to provide reliable inferences. As shown in Table 6.1, only two long-distance tours are indentified in the sample provided by the first 200 households, which equates 0.15 long-distance tours and 53 long-distance VMT per month. Both these values are only about half of the actual sample means. When the sample size reaches 400 households, the monthly number of long-distance tours starts to stabilize, but the monthly long-distance VMT continues to fluctuate.

Table 6.1 Long-Distance Tours Identified in Synthetic Two-Day Sample

Number of Households	Number of Long-Distance Tours Identified	Total VMT of Long-Distance Tours Identified	Inferred Monthly Number of Long-Distance Tours[1]	Inferred Monthly Long-Distance VMT[2]
100	1	365	0.15	55
200	2	710	0.15	53
300	4	2,306	0.20	115
400	8	3,497	0.30	131
500	11	4,147	0.33	124
600	13	4,992	0.33	125
700	14	5,362	0.30	115
800	16	5,812	0.30	109
900	19	7,414	0.32	124
1000	20	7,902	0.30	119
1100	23	8,396	0.31	114
1200	25	9,767	0.31	122
1300	25	9,767	0.29	113
1400	26	10,216	0.28	109
1500	31	12,358	0.31	124
1600	33	13,553	0.31	127
1700	35	15,047	0.31	133
1800	36	15,887	0.30	132
1900	38	17,049	0.30	135
2000	40	18,367	0.30	138

The instability of inferences about long-distance travel is only one reason why two-day surveys may not be sufficient for long-distance travel analysis. The short survey period also makes it very difficult to study the distributional properties of long-distance tours over time. A longitudinal design, however, is able to provide insights into temporal distributions, as will be formalized statistically in Chapter 7 and discussed in detail using the Commute Atlanta dataset in Chapter 8.

[1] Monthly number of long-distance tours = number of long-distance tours identified / (number of households × 2 days / 30 days in a month)
[2] Monthly long-distance VMT = total long-distance VMT identified / (number of households × 2 days / 30 days in a month)

6.2.3. Trade-offs between Sample Size and Survey Length

The sample size analysis for long-distance travel takes a slightly different approach than that for intra-regional travel, because long-distance travel is a much rarer event than intra-regional travel. As demonstrated in Section 6.2.2, 2000 two-day samples only identified 40 long-distance tours. It would be unreliable to draw estimates of the means about long-distance travel based on 40 long-distance tours. Therefore, this section takes advantage of the long survey period (30 months) that Commute Atlanta data provided to relate sample size to allowable error. The allowable error *e*, also known as relative precision, dictates that, given the confidence level $1-\alpha$, the required sample size is

$$N = \frac{z_{1-\alpha}^2 CV^2}{e^2}, \tag{6.1}$$

where CV stands for the coefficient of variation and $CV = \frac{Std\ Dev}{Mean}$[1]. Adopting this formula, Figure 6.10 plots the required survey length for each household, based on their average monthly number of long-distance tours. The dots represent the values calculated from the empirical standard deviation and mean, and the line represents the expected survey length assuming the Poisson distribution for each household. The goodness-of-fit of the Poisson distribution will be formally examined in Section 8.1.3, but the Poisson assumption provides a good approximation at this point. The implication of Figure 6.10 is that the number of months required follows the inverse of the mean, i.e.

[1] 15 households in the Commute Atlanta long-distance dataset did not undertake any long-distance tours between January 2004 and June 2006. Given this definition of allowable error, the 15 households have to be excluded from this analysis because the mean number of long-distance tours is zero, and would therefore render the coefficient of variation meaningless.

$1/\lambda$, assuming the Poisson distribution characterized by the mean and variance λ, because (6.1) now becomes

$$N = \frac{z_{1-\alpha}^2}{e^2} \cdot \frac{1}{\lambda}. \tag{6.2}$$

This means that the surveys for households that travel long-distance infrequently should be longer than the ones for those who travel long-distance frequently. This could be very difficult to implement in survey practice. Therefore, for certain demographic groups that do not make frequent long-distance travel, one would wish to recruit more households in exchange for the survey lengths.

One caveat in interpreting Figure 6.10 is not to extrapolate the curve towards zero. It will be very expensive to achieve the relative precision target for households that travel long-distance very rarely or never. The more practical way to reflect long-distance travel behavior in such households is to study their demographic characteristics from a pilot study or previous experience, and to draw a number of households from this demographic group, comparable to the numbers of households in other demographic groups in the survey.

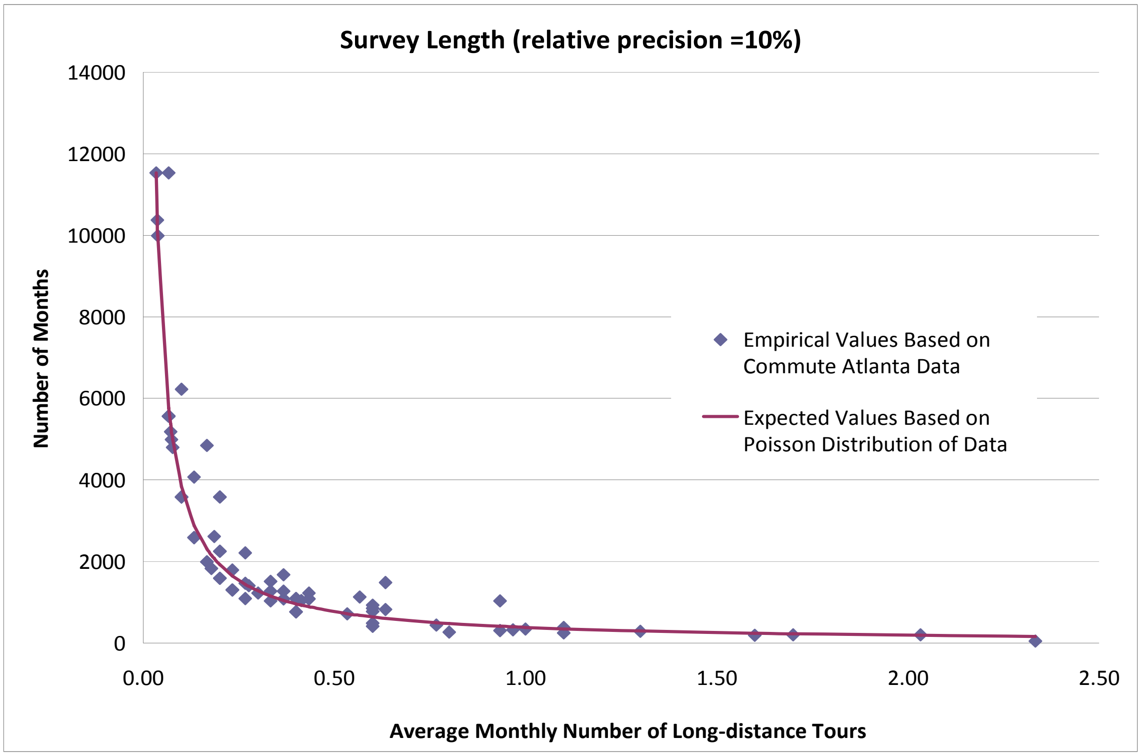

Figure 6.10 Required Survey Lengths to Achieve 10% Relative Precision
m=79 households[1]

The large number of months required for a household, as the average monthly number of long-distance tours decreases, appears prohibiting. However, for a given λ, it is possible to trade the survey length for the sample size. Figure 6.11 shows the number of households required if the survey lasts for 12 months and if the survey lasts for 24 months. According to Figure 6.11, if the survey length is 12 months, 960 households that only travel once every 30 months are needed to obtain a 10% relative precision at a 95% confidence level. The implication is that if through prior experiences and/or pilot studies, the researchers have some understanding of the association between demographic groups and their long-distance travel frequency, the number of households needed within

[1] The 15 households that did not undertake any long-distance tours during the study period were excluded from the 94 households.

each demographic group can be decided by estimating the λ within that demographic group. Such association will be explored in Chapter 9.

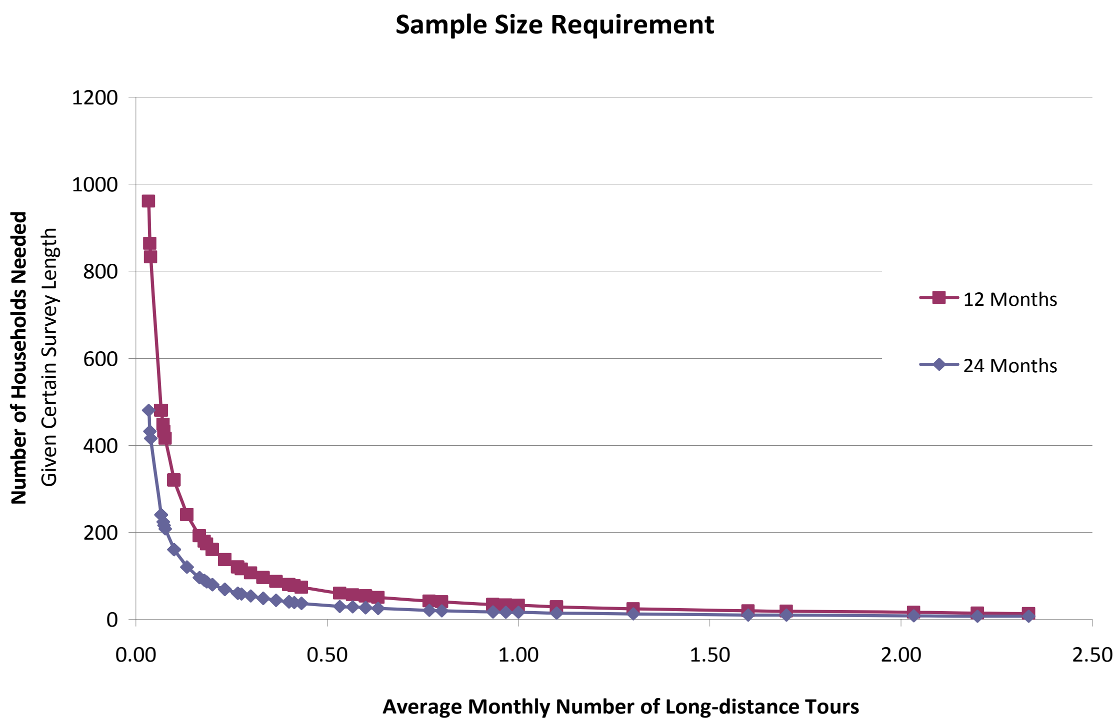

Figure 6.11 Required Numbers of Households Given Certain Survey Length, with Regard to the Average Monthly Number of Long-Distance Tours

6.3. Summary

This chapter examined the impacts of different sample sizes and survey lengths on the reliability of estimated means about key travel behavior variables for both intra-regional and long-distance travel. Resampling techniques were applied to generate random samples for this purpose.

The results for intra-regional travel show that, to obtain a reliable estimate of the population average of a travel behavior variable, a longer monitoring period such as a month provides savings on sample size (number of households), compared to conventional two-day surveys, in addition to potential benefits as reviewed in Chapter 2 that a longer monitoring period may be able to provide information on various trip purposes, route choices and arrival / departure times. To achieve 10% relative precision for estimating the average intra-regional VMT per weekday, a two-day survey would require 250 households and a 20-day survey would require 170 households based on the Monte Carlo analysis of the Commute Atlanta data. Such a trade-off amounts to more than 30% of savings on the number of households needed when the survey length increases from two days to 20 days for obtaining 10% relative precision.

With regard to long-distance travel, which can be considered rare events, a larger sample size is needed to achieve stability in estimated means, compared to the desired sample size for intra-regional travel. The cross-sectional nature of two-day surveys also prevents separating temporal distributions from cross-sectional effects. Readers are referred to subsequent chapters for more detailed discussions on longitudinal and cross-sectional information.

CHAPTER 7

STATISTICAL APPROACH

This chapter introduces the formal statistical approach for the sample size analyses that will be conducted in subsequent chapters for regression models and before-and-after studies. The basis of the statistical approach in this dissertation is the generalized estimating equation (GEE) procedures proposed by Liang and Zeger (1986), which is an extension of generalized linear models (GLMs) to correlated data.

The first section of this chapter introduces the model formulation and notation. The second section discusses the differences between cross-sectional information and longitudinal information, both which are present in panel data and can be at odds if not treated with care. The discussion of cross-sectional and longitudinal information is of essence of this dissertation in that sample size requirements will vary based upon the study objective - whether the longitudinal information is the main concern or the cross-sectional information is. The third section briefly discusses the design efficiency of panel studies, providing a rationale for conducting panel studies from the pure statistical point of view. The fourth section outlines the sample size estimating methodology based on GEE procedures. This methodology is adapted from the work of Rochon (1998).

7.1. Model Formulation and Notation

This research, like most longitudinal analyses, takes the statistical approach of a regression model such as the linear model, or the generalized linear model as described in Diggle, *et al.* (2002)

$$Y_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + \varepsilon_{ij}. \tag{7.1}$$

In this model notation, Y_{ij} represents a response variable, such as the number of trips per day or the VMT of long-distance travel per month, at time t_{ij}, for observation $j=1,\ldots, n_i$ on subject $i=1,\ldots, m$. $x_{ij1}, x_{ij2},\ldots, x_{ijp}$ represent p explanatory variables at t_{ij}. In the context of travel demand modeling, the explanatory variables can be household demographic characteristics, pricing incentives, gas prices, and temporal variables such as days of week and seasons. $\beta_1,\cdots\beta_p$ are unknown regression coefficients. Typically, $x_{ij1}= 1$ for all i and all j, and β_1 is then the intercept term in the linear model. E(Y_{ij}) = μ_{ij} and Var(Y_{ij}) = v_{ij} represent the mean and variance of Y_{ij}. ε_{ij} is a zero-mean random variable that represents the deviation of the response from the model prediction. The distribution of ε_{ij} determines the form of the regression model.

In matrix notation, equation (7.1) takes the form

$$Y_{ij} = \boldsymbol{x_{ij}}'\boldsymbol{\beta} + \varepsilon_{ij}, \tag{7.2}$$

and the regression equation for the i^{th} subject takes the form

$$\boldsymbol{Y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{\varepsilon_i}, \tag{7.3}$$

where: $\boldsymbol{x_{ij}}$ is a vector of length p of explanatory variables $x_{ij1}, x_{ij2},\ldots, x_{ijp}$,

$\boldsymbol{X_i}$ is a $n_i \times p$ matrix with $\boldsymbol{x_{ij}}$ in the j^{th} row,

$\boldsymbol{Y_i}$ is an n_i-vector representing the set of repeated outcomes for subject i,

$\boldsymbol{\beta} = (\beta_1,\cdots\beta_p)$, and

$\boldsymbol{\varepsilon_i} = (\varepsilon_{i1},\cdots,\varepsilon_{in_i})$.

$E(\boldsymbol{Y_i}) = \boldsymbol{\mu_{ij}}$ represents the mean of $\boldsymbol{Y_i}$ whereas Var($\boldsymbol{Y_i}$)= V_i, where the jk element of V_i is the covariance between Y_{ij} and Y_{ik}, denoted by Cov(Y_{ij}, Y_{ik}) = v_{ijk}. $\boldsymbol{R_i}$ denotes the $n_i \times n_i$ correlation matrix of $\boldsymbol{Y_i}$. The responses for all units are denoted by $\boldsymbol{Y} = (\boldsymbol{Y_1}, \cdots, \boldsymbol{Y_m})$, which is an N-vector with $N = \sum_{i=1}^{m} n_i$. Note that a cross-sectional study can be represented using this model form with n_i=1.

It is very important to emphasize that the cross-sectional effect of an explanatory variable, x, on response, Y, can be very different from the longitudinal effect. For example, from a cross-sectional perspective, the daily VMT of a household is probably positively correlated with household size. However, in a longitudinal study, a certain two-person household could very well reduce travel, especially long-distance trips, substantially as a new-born baby arrives. In this case, the daily VMT is negatively correlated with household size. Therefore, it is important to differentiate the cross-sectional effect from the longitudinal effect.

To formalize this idea, let us adopt the notation given above, and begin with a simple model of the form

$$Y_{ij} = \beta_0 + \beta x_{ij} + \varepsilon_{ij}, j = 1, \cdots, n_i; i = 1, \cdots, m. \tag{7.4}$$

If one re-expresses (7.4) as

$$Y_{ij} = \beta_0 + \beta x_{i1} + \beta(x_{ij} - x_{i1}) + \varepsilon_{ij} \ , \tag{7.5}$$

the model will fail to reflect the difference between the cross-sectional effect due to x_{i1} and the longitudinal effect represented by x_{ij}-x_{i1}. Therefore, Diggle, *et al.* (2002) suggested a model of the form

$$Y_{ij} = \beta_0 + \beta_C x_{i1} + \beta_L (x_{ij} - x_{i1}) + \varepsilon_{ij}, \tag{7.6}$$

so that both the cross-sectional effect β_C and the longitudinal effect β_L can be examined separately. Diggle et al. (2002) showed that the least-squares estimate of β derived from the model (7.4) is a biased estimate of β_L.

7.2. Longitudinal and Cross-Sectional Information

The assessment of within-subject changes in the response over time, β_L can only be achieved by a longitudinal study design (Fitzmaurice*, et al.*, 2004). The design of the Commute Atlanta study is to measure an initial cross-sectional sample repeatedly through time, making it possible to make comparisons of longitudinal (or within-subject) and cross-sectional (or between-subject) estimates of changes in the response with respect to endogenous factors such as demographic changes and exogenous factors such as gasoline prices and policy measures. Chapter 5 presented such comparisons qualitatively. This section will elaborate on the concepts of longitudinal and cross-sectional effects that were formalized in the previous section in the context of the empirical experiences of the Commute Atlanta study.

As noted in Chapter 5, the initial sample of the Commute Atlanta study were recruited according to household income, household size and vehicle ownership, but the demographic characteristics of individual households underwent significant changes. For example, initially, the total number of vehicles owned by a household ranges from one (1) to six (6). In the course of the study, the household that owned one (1) vehicle could purchase new vehicles and own three (3) vehicles at the end of the study. As a result, there are two potential sources of information about changes in VMT with changes in vehicle ownership. First, there is cross-sectional (or between-household) information about how VMT differ with regard to vehicle ownership in the initial baseline observations obtained in October 2004, since households enter the study with different total numbers of vehicles. Second, there are longitudinal (or within-household) effects that arise because households are measured repeatedly over time, yielding measurements of VMT when an individual household owns different numbers of vehicles. It is possible that these two sources of effects provide conflicting information about how VMT change with vehicle ownership.

To distinguish these two sources of effects has important implications on transportation planning. Cross-sectional information is crucial to understanding the current travel demand of a region. For an over-simplified example, if transportation planners have a correct estimate of the cross-sectional effects of vehicle ownership on VMT, the overall VMT of a region can be estimated with the distribution of population by vehicle ownership. Such cross-sectional information, however, is not adequate for long-term forecasting, unless transportation planners also have a correct estimate of the longitudinal effects of vehicle ownership on VMT. For illustration purposes, a synthetic numerical example can be adopted. Assume a cross-sectional effect of vehicle ownership on VMT, β_L, of 10%. That is, if household A owns two (2) vehicles and household B owns one (1) vehicle, household A conducts 10% more VMT than household B in a given time period. This effect, however, may not be the rate of increase in VMT when household B owns two (2) vehicles. The longitudinal effect may be 20%. Therefore, if the long-term trend in a region is that most households will increase their total number of vehicles owned, the total VMT will increase by 20% indicated by the longitudinal effect, rather than by 10% indicated by the cross-sectional effect. If the cross-sectional effect was used for the forecasting future VMT trend, the increase would be underestimated. Unfortunately, the conventional two-day travel surveys are by nature cross-sectional, and therefore can only provide estimates of cross-sectional effects.

7.3. Design Efficiency

Design efficiency is defined as $e = \mathrm{Var}(\widehat{\beta}_L)/\mathrm{Var}(\widehat{\beta}_C)$. The smaller the value of e, the more efficient the design is, and therefore, the more information is gained by taking additional measurements on each subject.

Diggle, *et al.* (2002) examined in detail that longitudinal studies tend to be more efficient than cross-sectional studies, even when $\beta_C = \beta_L$. Given the same number of subjects, m, the efficiency measure e takes on different forms according to the forms of the correlation matrix $\boldsymbol{R}$, and is influenced by

- n - the number of repetitions (assuming $n_i = n$ for all i)
- $\delta = \dfrac{\text{the ratio between the averaged within} - \text{subject variation in } x}{\text{the between-subjects variation in } x \text{ at visit 1}}$, and
- ρ - correlation among the repeated observations

Generally, e is a decreasing function of δ and n. The value of ρ can impact e in either directions, based on the values of δ and n.

7.4. Sample Size Requirements

One important implication from the above discussion of efficiency is that, assuming $\beta_C = \beta_L$, given a certain value of e, and a level of precision, the number of subjects needed for a longitudinal study, m_L, can be estimated as

$$m_L = m_C \cdot e. \tag{7.7}$$

This equation demonstrates that, from a pure statistical standpoint, a longitudinal study always reduces the required sample size, given a certain research objective. This conclusion provides theoretical rationale especially for short-term multi-day household travel surveys because in the context of travel demand modeling, the cross-sectional effects of explanatory variables, such as demographic characteristics and land use characteristics, can be seen as approximately the same as the longitudinal effects within the short survey period.

When $\beta_C \neq \beta_L$, the discussion becomes much more complicated than merely the reduction the sample size in a longitudinal study compared to a cross-sectional study. In this case, a cross-sectional study simply cannot reflect β_L, and therefore is not suitable for the study of change. In a longitudinal setting, the number of repeated observations per subject, *n*, needs to be balanced against the sample size, *m*, and may be constrained by practical considerations such as budget, technological availability, and correspondent fatigue.

The remainder of this section reviews the statistical approaches for sample size estimation given different research objectives in the context of travel demand analysis.

7.4.1. Travel Demand Modeling

One important task of this research is to emphasize adopting the appropriate modeling approach for panel surveys. As the GPS technology makes it possible to extend survey lengths, the modeling approach required for data analysis differs from the conventional approaches that have been used for cross-sectional surveys. The sample size analysis, therefore, should be applied based on the modeling technique that accounts for correlations in longitudinal designs.

This research uses the generalized estimating equation (GEE) approach of Liang and Zeger (Liang and Zeger, 1986) to model key travel behavior measurements. GEE is a well known technique for analyzing longitudinal data (Rochon, 1998). It extends the generalized linear model (GLM) of McCaullagh and Nelder (1989) to correlated observations and can be readily implemented in statistical software packages such as SPSS and R. This section first reviews the formulation of GLMs in the cross-sectional situation, and then outlines the basic assumptions of GEE that extend the applications to longitudinal data.

7.4.1.1. Generalized Linear Models

GLMs unify regression models for independent, discrete and continuous responses (McCullagh and Nelder, 1989). In GLMs, the mean response, $\mu_i = E(Y_i)$, is assumed to be related to a vector of covariates, $\boldsymbol{x}$, through

$$h(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} .$$

The function $h(\cdot)$ is called the link function. In linear regression, $h(\mu_i) = \mu_i$; in Poisson regression, $h(\mu_i) = \log(\mu_i)$.

The variance of Y_i is a specified function of its mean, μ_i, namely,

$$Var(Y_i) = v_i = \varphi v(\mu_i),$$

where the known function $v(\cdot)$ is referred to as the variance function and φ is the scaling factor. For some members of the GLM family, φ is a known constant, whereas in others it is an additional parameter to be estimated.

Each class of GLMs corresponds to a member of the exponential family of distributions, with a likelihood function

$$f(y_i) = \exp\{[y_i\theta_i - g(\theta_i)]/\varphi + c(y_i, \varphi)\}, \tag{7.8}$$

where θ_i is known as the natural parameter related to μ_i through $\mu_i = \partial g(\theta_i)/\partial\theta_i$. Some distributions within the exponential family include the Normal distribution, the Poisson distribution, the two-parameter gamma distribution, and the Tweedie distribution as will be discussed in detail in the subsequent chapter. To illustrate (7.8), consider the Poisson distribution, where

$$\theta_i = \log\mu_i,\ g(\theta_i) = \exp(\theta_i),\ c(y_i, \varphi) = -\log(y_i!),\ \varphi = 1.$$

The regression coefficients, $\boldsymbol{\beta}$, can be estimated by solving the estimating equation,

$$\boldsymbol{S(\beta)} = \sum_{i=1}^{m} (\frac{\partial\mu_i}{\partial\boldsymbol{\beta}})' v_i^{-1}\{Y_i - \mu_i(\boldsymbol{\beta})\} = 0, \tag{7.9}$$

where $v_i = Var(Y_i)$. The solution $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate and can be obtained through algorithms available in most statistical software packages.

7.4.1.2. Generalized Estimating Equation Approach

The GEE approach were formalized by Liang and Zeger (1986) to extend the GLM to accommodate correlated data. GEEs are suitable for marginal models to characterize the marginal expectation of a set of outcomes as a function of a set of explanatory variables (Horton and Lipsitz, 1999; Diggle*, et al.*, 2002). The term "marginal" indicates that the model for the mean response depends only on the covariates of interest, and not on any random effects or previous response, in contrast to *mixed effects models*, where the mean response depends on both the covariates and a vector of random effects (Fitzmaurice*, et al.*, 2004). Marginal models are appropriate for travel demand analysis because in such analysis, inferences of the population-average, rather than subject-specific, are the focus. This research will adopt marginal models as the modeling basis for sample size estimation.

A marginal model has the following assumptions (Diggle*, et al.*, 2002):

1. the marginal expectation of the response, $E(Y_{ij}) = \mu_{ij}$, depends on explanatory variables, $\boldsymbol{x_{ij}}$, by $h(\mu_{ij}) = \boldsymbol{x_{ij}}'\beta$ where $h(\cdot)$ is a known link function such as the log for counts (the Poisson distribution);
2. the marginal variance depends on the marginal mean, by $\mathrm{Var}(Y_{ij}) = v(\mu_{ij})\varphi$ where $v(\cdot)$ is a known variance function and ϕ is a scale parameter which may need to be estimated;
3. the correlation between Y_{ij} and Y_{ik} is a function of the marginal means and perhaps of additional parameters $\boldsymbol{\alpha}$, namely, $\mathrm{Corr}(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \boldsymbol{\alpha})$ where $\rho(\cdot)$ is a known function.

In a marginal model, the regression of the response, Y_{ij}, on explanatory variables is modeled separately from within-subject correlation. Therefore, a separate model for the association among observations from each subject must be specified. The estimating procedures using the generalized estimating equations (GEE) approach is discussed in detail in (Liang and Zeger, 1986) and (Prentice and Zhao, 1991). Software packages are available for estimating the coefficients.

7.4.2. Before-and-After Studies

Before and after studies are effective for police analysis. Since household travel surveys are essentially observational studies, it is important to adjust major confounding effects through regression modeling. Consequently, such an adjustment needs to be appropriately acknowledged for sample size estimation.

Rochon (1998) gave a generalized procedure for sample size analysis for longitudinal studies. In the paper and the associated SAS codes Rochon (1998) provided the required input parameters as shown in Table 7.1. Subsequently, each category of parameters will be discussed in detail with applications in travel behavior studies.

Table 7.1 Input Parameters for Sample Size Analysis in Longitudinal Studies

Parameters		Descriptions
Design of the Study	$\boldsymbol{\tau}$	A ($1\times T$) vector indicating the follow-up times.
	$\mathbf{X}$	An ($ST\times r$) design matrix[1] for the study. r is the number of explanatory variables. A ($T\times r$) design matrix $\mathbf{X}_s$ is defined for each of the S subpopulations, whereupon they are stacked on top of each other.
	$\boldsymbol{\mu}$	An ($S\times T$) matrix of expected values μ_{st} under the model, where there are S subpopulations of interest and T time points.
Underlying Stochastic Mechanism	$h(\mu)$	The link function to be used in the analysis, including identity, log, logit, etc.
	$v(\mu)$	The variance function for Y_{ij} to be used in the analysis, taking the possible forms of Gaussian, Poisson, Gamma, and Bernoulli[2].
	ψ	A scale parameter signifying under- ($0<\psi<1$) or over- ($\psi>1$) relative to the underlying stochastic model.
Correlation Structure (See expression in Section7.4.2.3)	ϕ	The autocorrelation parameter.
	θ	The "damping" parameter. The correlation structure becomes the AR(1) model when θ=1 and the exchangeable model when θ=0.
Research Question	$\mathbf{H}$	An ($h\times r$) hypothesis matrix among the elements of linear model parameters $\boldsymbol{\beta}$. The hypothesis is written as $\mathbf{H\beta}=\mathbf{h_0}$.
	$\mathbf{h_0}$	An ($h\times 1$) vector of constant terms under the null hypothesis expressed above.
	α	Type I error rate.
	γ	Type II error rate.
Other Design Considerations	$\boldsymbol{\omega}$	A ($1\times S$) vector providing the relative sizes $\{\omega_s\}$ across the different subpopulations in the asymmetric allocation problem.
	$\boldsymbol{\pi}$	An ($S\times T$) matrix of probabilities $\{\pi_{st}\}$ for staggered entry and attrition.

[1] A design matrix is a matrix of explanatory variables.
[2] The Tweedie distributions are not included in Rochon's algorithm.

7.4.2.1. Design of the Study

Number of Repeated Observations T and Follow-up Times $\boldsymbol{\tau}$

Depending on the context of a study, the number of repeated observations T can be viewed as the number time points collected in the survey, i.e. number of days, weeks, months, or years. Each column of the $1 \times T$ vector $\boldsymbol{\tau}$ indicates the time when the j^{th} observation takes place. For example, for the long-distance travel analysis presented in Section 9.2, there are 30 equally-spaced time points (i.e. months), so $\boldsymbol{\tau} = \{1, 2, \cdots, 30\}$.

Number of Subpopulations S

It is common practice in travel surveys that the entire population is stratified into several subpopulations based on key demographic characteristics such as household size, income, vehicle ownership, etc. Given the growing trend of activity-based modeling approach, the number of subpopulations in the travel survey and modeling practices has increased to provide more precision. For example, the Atlanta Regional Commission uses more than 100 combinations of household size, number of workers, income and age in their forthcoming activity-based model. The larger number of subpopulations is likely to require larger samples.

Design Matrix $\boldsymbol{X}$

As demonstrated in Chapter 5, various confounding effects influence household travel behavior, making it difficult to single out a certain policy effect. For example, response to pricing can be confounded by endogenous factors such as income, household size, and vehicle ownership, and exogenous factors such as gasoline prices and the overall economic trend, as discussed in the literature review. Therefore, it is important to account for the confounding effects in the design stage and include them in the design matrix. For example, a design matrix can take the form of

$$\mathbf{X}_s = \underbrace{\begin{matrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{T1} & \cdots & x_{Tp} \end{matrix}}_{p} \underbrace{\begin{matrix} z_{11} & \cdots & z_{1q} \\ \vdots & \ddots & \vdots \\ z_{T1} & \cdots & z_{Tq} \end{matrix}}_{q}, p+q=r$$

where there are p variables of interest, e.g. policy effects, and q confounding effects. This configuration indicates that if researchers control for all confounding factors with proper subpopulation planning and comprehensive regression variables, a study is able to draw solid conclusions.

This notation implies that each subpopulation should be homogenous. In other words, within each subpopulation, a variable x or z takes the same value for all households at the j^{th} time point, i.e.

$x_{ijk} = x_{jk}$, $i = 1, \cdots, m_s$, $j = 1, \cdots, T$, $k = 1, \cdots, p$, and

$z_{ijl} = z_{jl}$, $i = 1, \cdots, m_s$, $j = 1, \cdots, T$, $l = 1, \cdots, l$,

where m_s is the number of households in subpopulation s. This is difficult to achieve in reality, but is reasonable to assume during the design stage. Such notation therefore requires a careful stratification process and a realistic understanding of the nature of the confounding effects.

In the context of transportation policy analysis, e.g. value pricing, $\boldsymbol{x}_j$ would represent p=1 dummy variables with values

$$x_j = \begin{cases} 1 & \text{if subject receives pricing incentives,} \\ 0 & \text{if subject does not receive incentives.} \end{cases}$$

Alternatively, $\boldsymbol{x}_j$ could represent p=1 variables indicating the amount of the incentives, with example values

$$x_j = \begin{cases} 0 & \text{cents/mile,} \\ 50 & \text{cents/mile,} \\ 100 & \text{cents/mile.} \end{cases}$$

Similarly, confounding effects can be categorical, nominal or continuous, including household demographic characteristics such as household size, income, vehicle ownership and age, external economic variables, such as gas prices, and temporal variables such as season / month of the year and day of week.

Expected Values $\boldsymbol{\mu}$

This matrix specifies the expected values of an outcome variable (i.e. VMT, number of trips, etc) for each subpopulation at each time point. Such specifications require knowledge from pilot studies and the expected smallest detectable difference. When the confounding effects are many and tend to change during the study, as was observed in Commute Atlanta study, regression analysis from previous studies is especially helpful.

<u>7.4.2.2.</u> <u>Underlying Stochastic Mechanism</u>

As discussed in Chapter 6, the widely applied normal distribution does not fit travel behavior data very well. Rather, the Tweedie family of distributions fit daily or monthly VMT data well, and Poisson distribution fits the number of trips / tours data well. For Poisson distribution, the canonical link function is the log link. For Tweedie distributions, both the log link and the identity link could be used. The dispersion parameter ψ can be estimated by data from pilot studies.

7.4.2.3. Correlation Structure

Various common structures may be appropriate for modeling the working correlation matrix among the repeated measures. Horton and Lipsitz (1999) reviewed the definitions and examples of common working correlation models as seen in various software packages. Among the common structures are the exchangeable structure, which takes the form

$$\begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

and the first-order autoregressive, also known as AR(1), which is defined as

$$r_{u,v} = \begin{cases} 1, \mathit{if}\ u = v \\ \rho^{|u-v|}, \mathit{otherwise} \end{cases}$$

where *u* and *v* are the following-up times when two repeated observations are taken.

In travel behavior applications, the correlation between the travel behavior measures taken within the same household may decrease as the time between the two repeated measures is longer, thus violating the assumptions of the exchangeable structure, but the rate of the decrease is smaller than the exponential decay prescribed by the AR(1) model. Therefore, a "damped" exponential correlation structure described in Muñoz, *et al.* (1992) is considered. This structure is defined as

$$r_{uv}(\phi, \theta) = \begin{cases} 1, \mathit{if}\ u = v \\ \phi^{\delta_{uv}^{\theta}}, \mathit{otherwise} \end{cases}$$

where $\delta_{uv} = |\tau_u - \tau_v|$. $|\phi| < 1$ is the correlation between observations separated by one unit of time. $\theta \geq 0$ is a 'damping' parameter that permits attenuation ($0 < \theta < 1$) or acceleration ($\theta > 1$) of the exponential decay of correlation. This research takes $0 < \theta < 1$ to reflect common knowledge in travel behavior studies.

7.4.2.4. Research Question

In most before-and-after travel behavior studies, there is only one targeted policy effect, such as the mileage-based pricing incentive, or a toll road facility. Therefore, in the null hypothesis $\mathbf{H\beta = h_0}$, $\mathbf{h_0}$ is a h=1 parameter, and $\mathbf{H}$ is a $(1 \times r)$ vector $[\underbrace{0 \ \cdots \ 0}_{r-1} \ 1]$.

As common practice (Muthen and Muthen, 2002), the Type I error rate α will be set to 0.05 and the Type II error rate γ will be set to 0.2.

7.4.2.5. Other Design Considerations

Differential Allocation $\boldsymbol{\omega}$

In travel behavior studies, differential sample allocation across the subpopulations is often desirable. For example, practitioners usually tend to favor the treatment group and keep the control group size relatively small. The vector $\boldsymbol{\omega}$ characterizes the proportions of samples in each subpopulation.

Suppose that there are m households in the first subpopulation. The number of households in the s^{th} population can be expressed as $m\omega_s$, where ω_1=1 and ω_s for $s \geq 2$ are the known numbers of individuals required in the s^{th} subpopulation relative to the first.

Staggered Entry, Administrative Censoring and Dropout $\boldsymbol{\pi}$

The issue of missing data is common in longitudinal travel studies, especially surveys with a GPS component. This issue arises for a variety of reasons. Staggered entry could occur because households are typically recruited over a period of time. Or, some members of a household may move out and form a second household and stay in the study. Administrative censoring (Frangakis and Rubin, 2001) occurs when the research team loses contact of a household and cannot obtain their updated demographic information, or when there are equipment issues. Households could also choose to opt out during the study.

Each element π_{st} in the matrix $\boldsymbol{\pi}$ represents the proportion of the sample in the s^{th} subpopulation expected to provide the first t evaluations before being censored, lost to follow-up or dropout. The derivation of the $\{\pi_{st}\}$ will be illustrated in Section 10.6 with the numerical example.

CHAPTER 8

DATA CHARACTERISTICS AS INPUTS FOR THE GENERALIZED ESTIMATING EQUATION PROCEDURES

The previous chapter reviewed the statistical formulation of the generalized estimating equations (GEE) procedures for longitudinal data analysis. This chapter characterizes the properties of the intra-regional and long-distance travel data that are needed to specify the three assumptions for the GEE procedures (Diggle*, et al.*, 2002) outlined in Chapter 7, namely:

1. the marginal expectation of the response, $\mathrm{E}(Y_{ij}) = \mu_{ij}$, depends on explanatory variables, $\boldsymbol{x_{ij}}$, by $h(\mu_{ij}) = \boldsymbol{x_{ij}}'\beta$ where $h(\cdot)$ is a known link function such as the log for counts, i.e. number of intra-regional trips and number of long-distance tours (the Poisson distribution);
2. the marginal variance depends on the marginal mean, by $\mathrm{Var}(Y_{ij}) = v(\mu_{ij})\varphi$ where $v(\cdot)$ is a known variance function and φ is a scale parameter which may need to be estimated;
3. the correlation between Y_{ij} and Y_{ik} is a function of the marginal means and perhaps of additional parameters $\boldsymbol{\alpha}$, namely, $\mathrm{Corr}(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \boldsymbol{\alpha})$ where $\rho(\cdot)$ is a known function.

The first two assumptions correspond to the standard generalized linear model as described in Section 7.4.1.1, and the third assumption, the within-subject association among the repeated responses, represents the main extension of generalized linear models to longitudinal data (Fitzmaurice*, et al.*, 2004). The GEE approach for marginal models do not require distributional assumptions for the observations, but a basic understanding of the range and nature of the outcome variable is beneficial to choosing the link and variance functions (Hardin and Hilbe, 2002). Consequently, the first section of this chapter examines the distributional properties in response to the first two assumptions, and the second section explores the within-subject association among repeated observations.

8.1. Data Distribution

Measurements of the magnitude of travel, such as daily intra-regional VMT and monthly long-distance VMT, have the property that they are positive and continuous, except for the possibility of exact zeroes when travel does not occur. Data with exact zeros are common (Smyth, 1996), for example, weather variables such as rainfall and wind speed, and population sizes in ecology (Perry and Taylor, 1985). In travel behavior studies, zero values occur when households do not make any trips in a particular time period. As noted in Section 4.3, about 11% of all household-days in the Commute Atlanta Pricing study data set observed no trips.

Such data cannot be modeled by the four most common distributions used in generalized linear models discussed in McCullagh and Nelder (1989), namely the normal, Poisson, gamma and inverse-Gaussian, but can be modeled using a family of exponential distributions intermediate between the Poisson and the gamma families, namely the Tweedie distributions.

Tweedie distributions have been applied to a wide array of fields such as actuarial studies, meteorology and climatology, and biomedical applications. For example, Jørgensen and De Souza (1994) used the Tweedie distributions to model the total costs of insurance claims, where the authors assumed Poisson arrival of claims and gamma distributed costs for individual claims. Dunn (2003) modeled precipitation, where the occurrence of precipitation can be modeled as the Poisson distribution and the amount of precipitation at each occurrence can be modeled as the gamma distribution. Smyth (1996) provides an example of studying the relationship between children's gastro-esophageal reflux and sleeping positions using the Tweedie distributions.

The transportation field is not generally familiar with the Tweedie distributions. Intuitively, however, travel behavior data can be modeled using these distributions, considering total VMT in a given time frame as a sum of VMT by several trips or tours. The data will include exact zeros when households do not travel during that time period.

This chapter will first introduce the mathematical formulation of the Tweedie family of distributions, using long-distance travel as an illustration. Then, the suitability of Tweedie distributions to both intra-regional and long-distance travels are formally examined respectively.

8.1.1. The Formulation of Tweedie Family of Distributions

Assume any long-distance tour i results in a distance measured in miles R_i, and that each R_i has a gamma distribution Gam($-\alpha, \gamma$), where the mean is $-\alpha\gamma$ and the variance is $-\alpha\gamma^2$. Then, assume the number tours in any one month, say N, has a

Poisson distribution. This implies that there will be months with no long-distance tours when $N = 0$. The total monthly VMT of long-distance tours, Y, can be found as the Poisson sum of the gamma random variables so that

$$Y = R_1 + R_2 + \cdots + R_N ,$$

where N has a Poisson distribution with mean and variance λ.

An identical argument can be applied to daily VMT of intra-regional trips, when R_i could refer to the intra-regional trips recorded on any one day, and Y the total daily VMT.

The distribution of Y has been widely called a compound Poisson distribution (Feller 1968; Bar-Lev and Stramer 1987; Jørgensen and De Souza 1994; Smyth and Jørgensen 2002). The resulting probability function can be written as

$$\log f_p(y;\mu,\varphi) = \begin{cases} -\lambda, & \text{for } y = 0 \\ -\dfrac{y}{\gamma} - \lambda - \log y + \log W(y,\varphi,p), & \text{for } y > 0 \end{cases} \tag{8.1}$$

where

$\gamma = \varphi\,(\text{p-1})\,\mu^{\text{p-1}}$

$\lambda = \mu^{\text{2-p}}/[\varphi\ (\text{2-p})]$, and

$$W(y,\varphi,p) = \sum_{j=1}^{\infty} \frac{y^{-j\alpha}(p-1)^{\alpha j}}{\varphi^{j(1-\alpha)}(2-p)^j\, j!\,\Gamma(-j\alpha)}$$

The mean of the distribution is μ and the variance is $\text{var}[Y] = \psi\mu^p$, where $\psi > 0$ is a dispersion parameter and $1 < p < 2$ is the index which determines the compound Poisson distribution. Importantly, the probability of recording no long-distance tours in a given month is:

$$\Pr(Y=0)=\exp(-\lambda)=\exp\left\{-\frac{\mu^{2-p}}{\varphi(2-p)}\right\}$$

The compound Poisson distributions belong to the class of distributions known as the Tweedie family of distributions, named by Jørgensen (1987; 1997) after Tweedie (1984). Tweedie distributions have a variance of the form $\mathrm{var}[Y]=\varphi\mu^p$ for $p \notin (0,1)$. The index *p* defines the type of distributions as follows (Dunn, 2003):

- For $p<0$, the data *y* are supported on the whole real line. However, applications for these distributions are unknown
- For $p=0$, the distributions become the normal distribution
- For $p=1$ with $\varphi=1$, the distributions become the Poisson distribution
- For $1<p<2$, the distributions are the compound Poisson distributions as discussed
- For $p=2$, the distributions are the gamma distribution
- For $p>2$, the distributions have a similar shape to the gamma but are progressively more right-skewed as *p* gets larger

As shown in equation (8.2), mathematically, the Tweedie distributions are best characterized using the (μ,φ,p) parameterization. In transportation applications, parameterization in terms of (λ,γ,α) is more straightforward, where λ refers to the mean number of long-distance tours per month, γ to the shape parameter of the distribution of number of tours, and $-\alpha\gamma$ to the mean distance of long-distance tours per tour. The transformation between the two sets of parameterizations are summarized as

$\lambda = \mu^{2\text{-p}} / [\varphi \ (2\text{-p})]$

$\gamma = \varphi \ (\text{p-1}) \, \mu^{\text{p-1}}$

$\alpha = (2\text{-}p) / (1\text{-}p)$

Because the Tweedie distributions belong to the exponential family of distributions, upon which GLMs are based, there is a framework readily available for fitting models based on these distributions and for diagnostic testing (McCullagh and Nelder, 1989).

Although no closed forms exist for evaluating the density function or cumulative distribution functions of the Tweedie distributions, programs exists to provide fast and accurate algorithms for estimating parameter values (Dunn and Smyth, 2005) and to allow the computation of quantile residuals (Dunn and Smyth, 1996) in diagnostic analysis.

8.1.2. Intra-Regional Travel Data Distributions

8.1.2.1. Daily Intra-Regional VMT

A probability density plot can provide a visual summary of how the random variable is distributed, including symmetry, skewness and disperseness, among many others (Kvam and Vidakovic, 2007). Kernel density estimation is a statistical method that disperses the probability mass of each observation smoothly, so that a good visual presentation of the probability distribution function can be obtained. Figure 8.1 shows the kernel density estimation of the daily VMT of intra-regional trips. The density plot indicates that the data are heavily right skewed and that there is a high concentration of observations at or around zero. However, this plot does not fully reflect the nature of the data distribution in that kernel density plot is suitable for continuous data, whereas the daily VMT data are continuous but with exact zeros. This nature of the data is reflected inconspicuously by the dual-peak shape of the density plot. The kernel density plot suggests that the data distribution resembles a Tweedie distribution, but further examination is need.

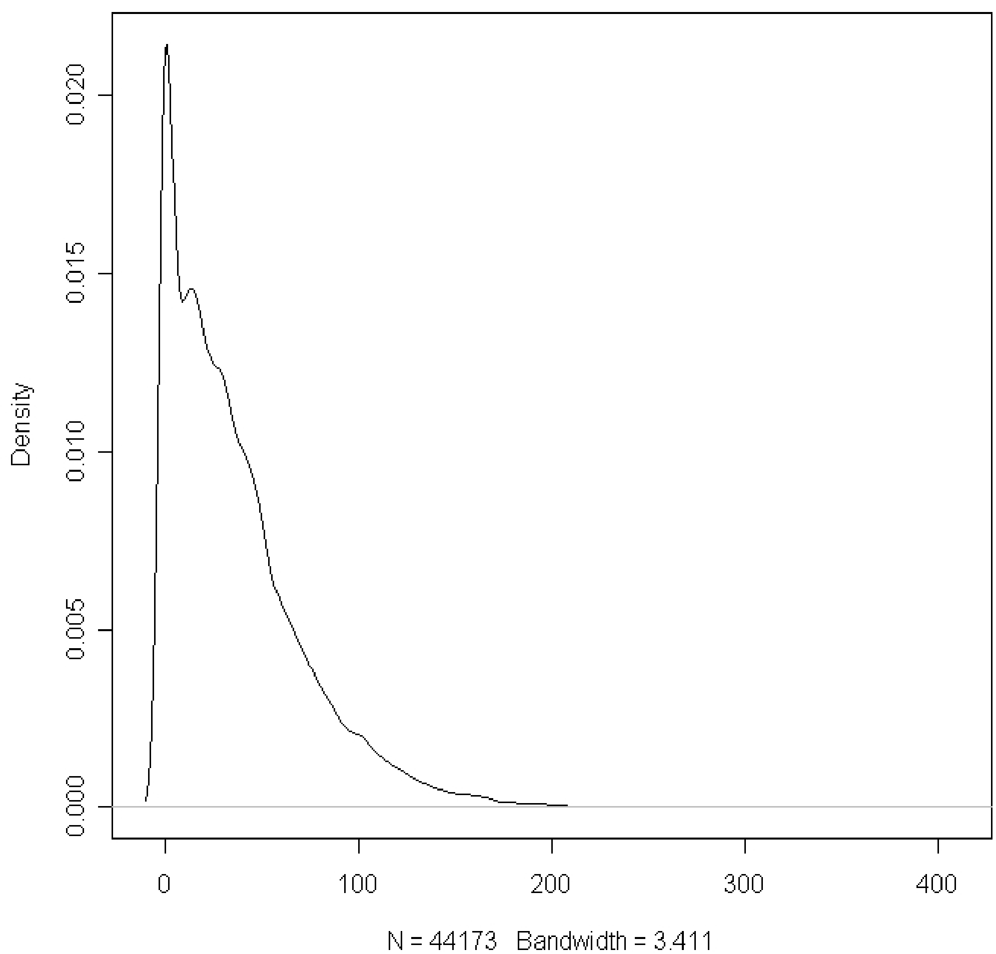

Figure 8.1 Kernel Density Estimation of Daily Intra-Regional VMT

To estimate the Tweedie parameters for the daily intra-regional VMT data, the Tweedie package in the R statistical software is employed. The maximum likelihood estimate of μ is the sample mean, 36.73 miles. The Tweedie package in R produces the profile likelihood plot (Dunn and Smyth, 2005) for selecting the value of p is shown in Figure 8.2. The resulting estimate of p is 1.43, and the estimate of φ is 6.95.

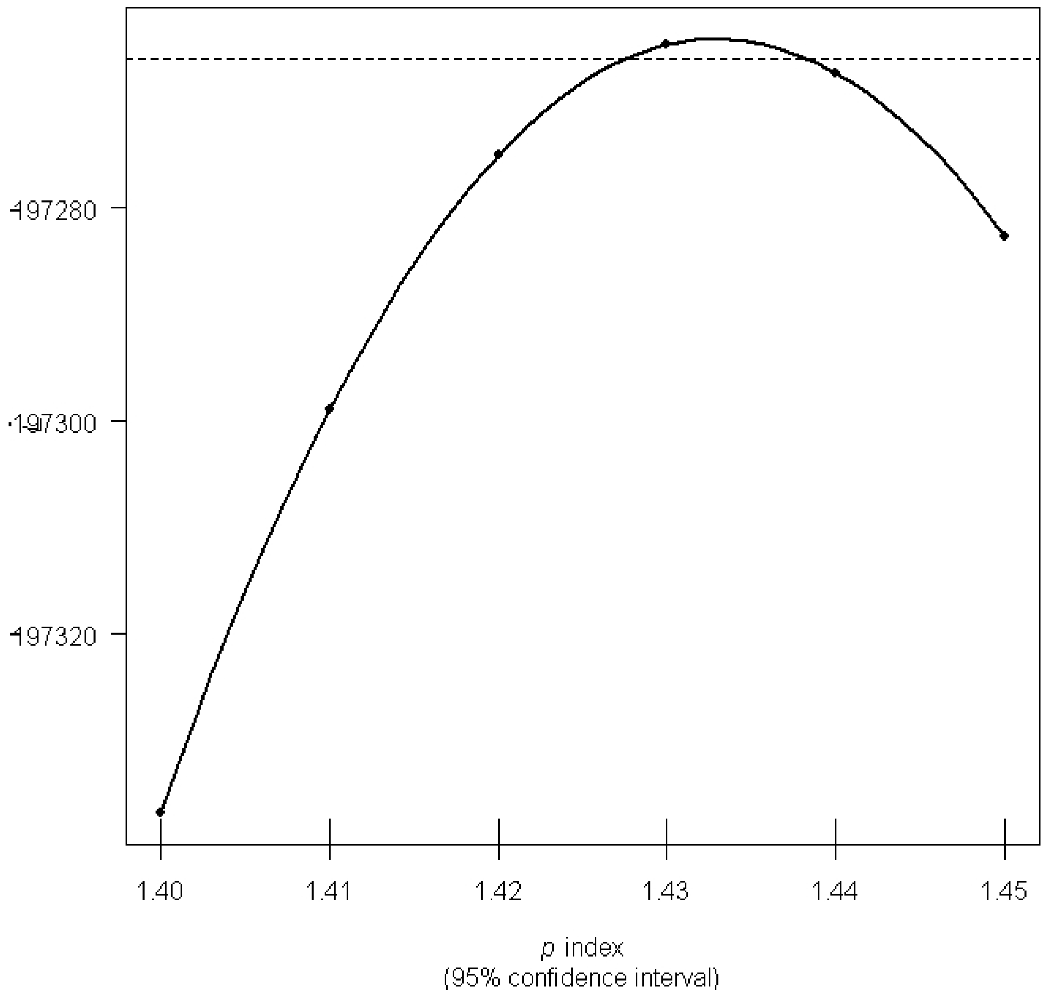

Figure 8.2 The Profile Likelihood Functions Showing the Maximum (log-) Likelihood Values of p for Daily Intra-Regional VMT per Household
The horizontal line indicates the approximate 95% confidence interval for p.

With the estimated Tweedie parameters, the density of the daily intra-regional VMT data can be revisited. The Tweedie probability density function is plotted in Figure

8.3. The point mass at $x = 0$ is characteristic of daily intra-regional VMT, indicating the probability of households not undertaking any intra-regional travel on a given day.

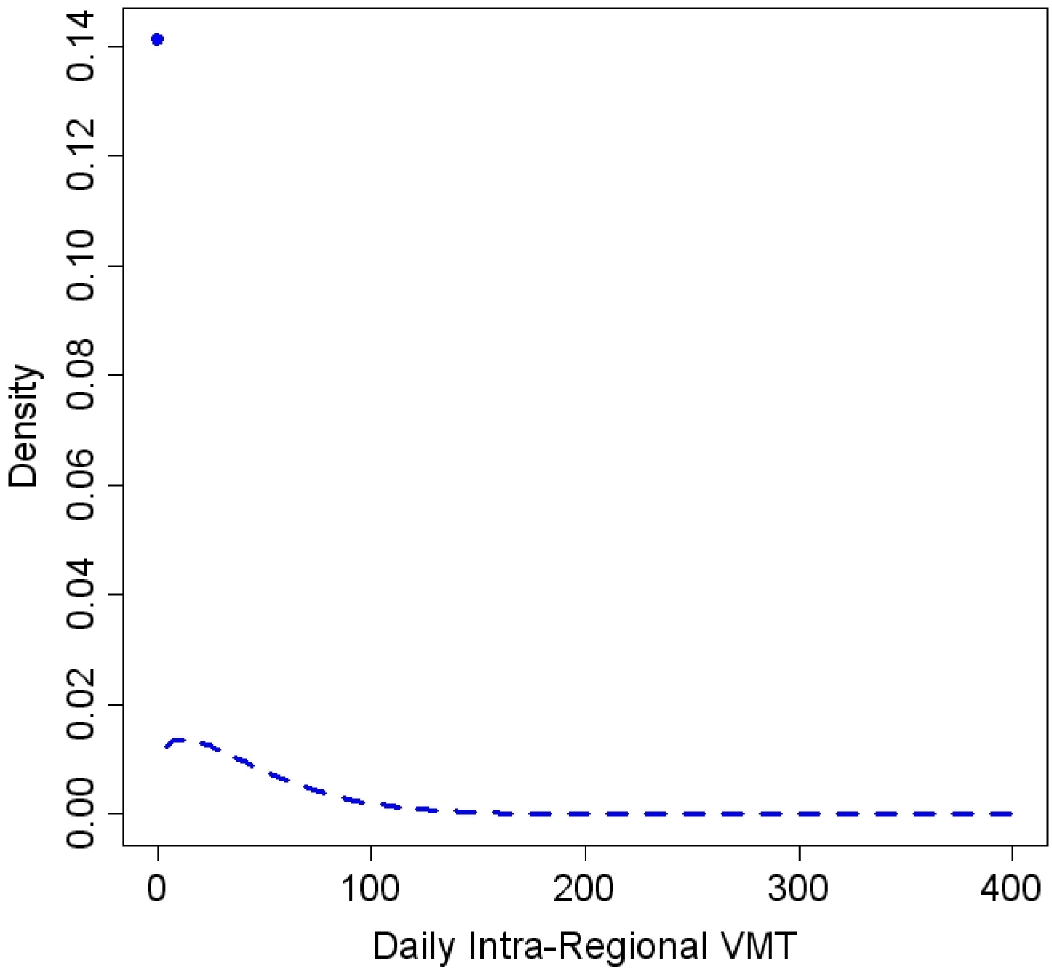

Figure 8.3 Tweedie Probability Density Function for Daily Intra-Regional VMT

To assess the quality of the fitted distributions, quantile residuals can be used, as defined by Dunn and Smyth (1996). Quantile residuals are produced by inverting the estimated distribution function for each observation to obtain standard normal residuals (Dunn and Smyth, 1996). A typical QQ-plot of these quantile residuals is shown in Figure 8.4, indicating that the quality of the fitted distribution on the daily level is good.

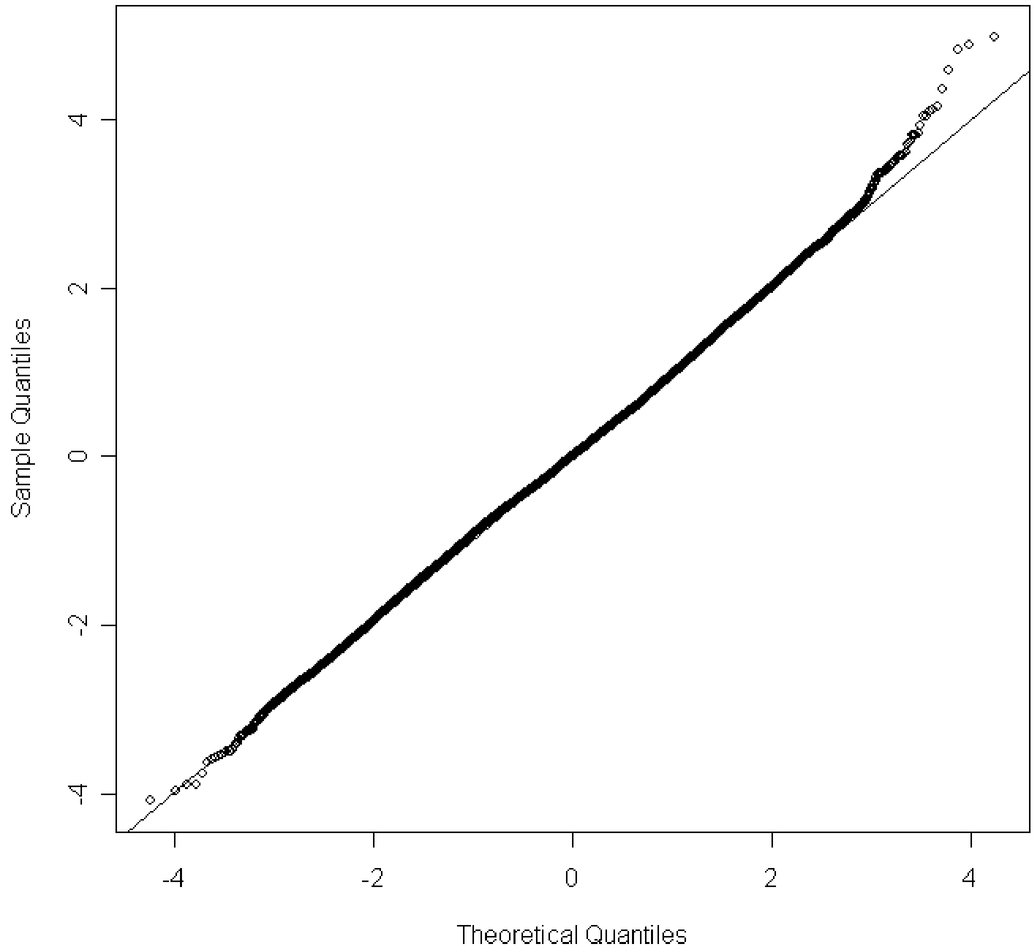

Figure 8.4 The Q-Q Plot of the Quantile Residuals after Fitting a Tweedie Distribution to Daily Intra-Regional VMT

After confirming the goodness-of-fit statistically by using the QQ-plot, it is of interest to examine how Tweedie distributions reflect the intra-regional travel data practically. Section 8.1.1 noted that the mathematical parameterization (μ, φ, p) can be translated into practical parameterization (λ, γ, α). Table 8.1 summarizes the maximum likelihood estimators of the mathematical parameters and the practical estimators. Despite the good quality of fit for daily intra-regional VMT shown in Figure 8.4, the interpreted estimators on the trip level are less than satisfactory because the estimated mean number of trips per day is 1.97, much lower than the actual sample mean 5.67, and the estimated distance per trip is 18.66 miles, much higher than the sample mean 6.48. These results indicate that the Tweedie distributions may not model daily intra-regional VMT very well on the trip level. This observation suggests that the adoption of Tweedie distribution may not be extended to estimate number of trips and distance per trip for studies that are mainly concerned with such trip-level travel behavior. Nevertheless, Tweedie distributions can be useful for studies that are mainly concerned about the total VMT per time period, such as fuel consumption studies. It is also important to note, as stated at the beginning of this chapter, that the GEE procedure do not require distributional assumptions of the data. The goodness-of-fit of the Tweedie distributions only provide confidence in specifying the canonical link and variance functions that are associated with the Tweedie distributions when the GEE procedures are carried out in the next chapter.

Table 8.1 The Maximum Likelihood Estimates of Household Daily Intra-Regional VMT

Mathematical parameters	Estimators	Practical parameters	Interpretation	Estimators
$\hat{\mu}$	36.73	$\hat{\lambda}$	The mean number of trips per day	1.97
$\hat{p}$	1.43	$\hat{\gamma}$	The shape of the distance per trip gamma distribution	14.07
$\hat{\varphi}$	6.95	$-\hat{\alpha}\hat{\gamma}$	The distance per trip	18.66

8.1.2.2. Daily Number of Intra-Regional Trips

The underlying reason for the deviation of the Tweedie distributions with regard to number of trips and distance per trip from the actual sample means will be explained in this section, where the Poisson assumption of daily number of intra-regional trips is scrutinized.

The Poisson assumption dictates that the mean and variance are equal. However, this is inconsistent with the empirical evidence in daily number of intra-regional trips. Table 8.2 gives the ratio of the sample variances to the means of the counts for each month. Extra-Poisson variation is evident as the variance-to-mean ratios range from 3.15 to 4.50. This phenomenon is often referred to as over-dispersion (Diggle*, et al.*, 2002). The degree of over-dispersion in the daily number of intra-regional trips is not as high as the biomedical examples given by Diggle *et al.* (2002), but nevertheless indicates the violation of Poisson assumption. The extra-Poisson variation implies that there are more large values of daily number of intra-regional trips than one would expect if the data are Poisson-distributed. This could explain why the estimated number of trips by the Tweedie distribution is smaller than the sample mean.

Table 8.2 Variance to Mean Ratios for Daily Number of Intra-Regional Trips

	Jan	Feb	Mar	Apr	May	Jun	Oct	Nov	Dec	Overall
Baseline	3.69	3.47	3.15	3.43	3.28	3.91	3.75	3.71	3.80	3.80
Pricing	4.06	3.71	3.75	4.20	4.47	4.50	3.82	3.64	3.76	

8.1.3. Long-Distance Travel Data Distributions

8.1.3.1. Monthly VMT of Long-Distance Tours

The same procedure for estimating Tweedie distribution parameters for daily intra-regional VMT applies to monthly VMT of long-distance tours. The maximum likelihood estimate of μ for the monthly VMT of long-distance tours is the sample mean, 178.14 miles. The maximum likelihood estimates of p and ϕ are calculated using the Tweedie package in R. The profile likelihood plot for selecting the value of p is shown in Figure 8.5. The resulting estimate of p is 1.36, and the estimate of ϕ is 152.46.

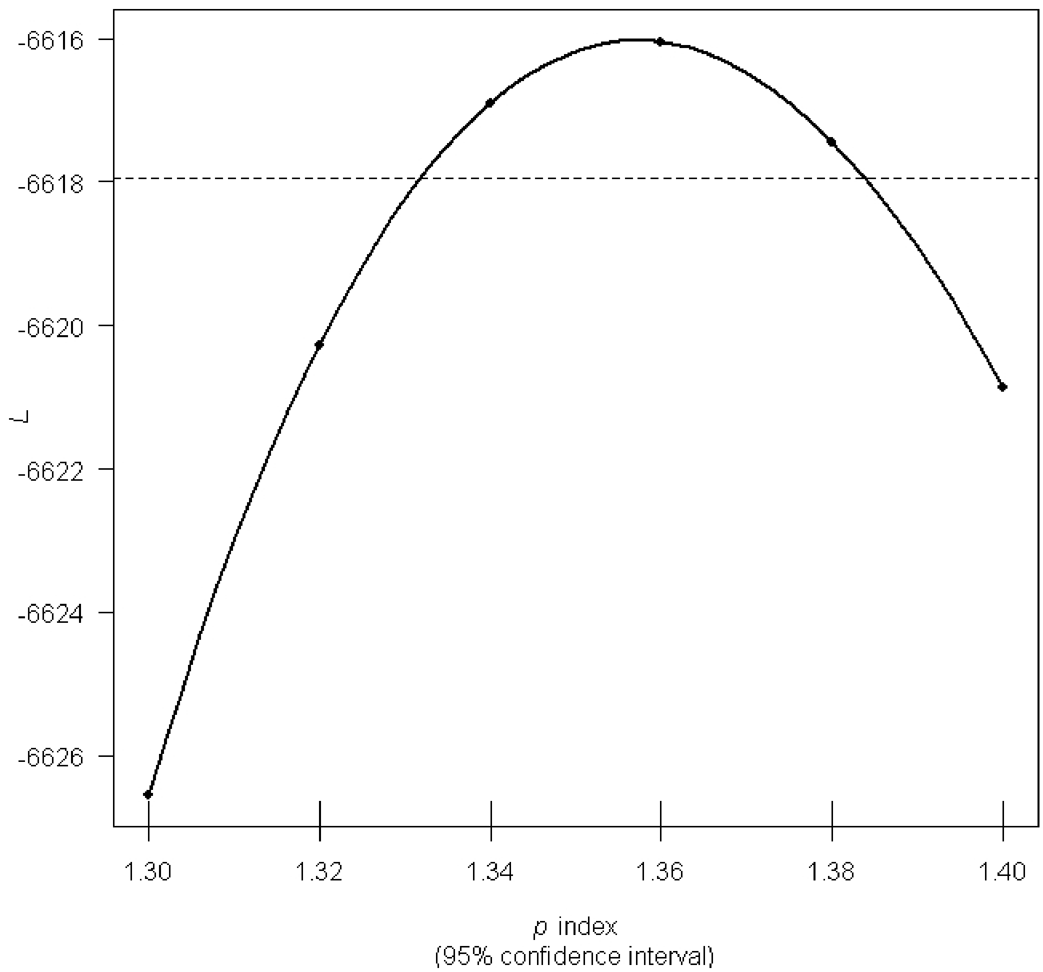

Figure 8.5 The Profile Likelihood Functions Showing the Maximum (log-) Likelihood Values of p for Monthly Long-Distance VMT per Household
The horizontal line indicates the approximate 95% confidence interval for p.

Similar to the case in intra-regional travel, the estimated Tweedie distribution fits monthly long-distance VMT well, as shown by the quantile residual QQ-plot in Figure 8.6.

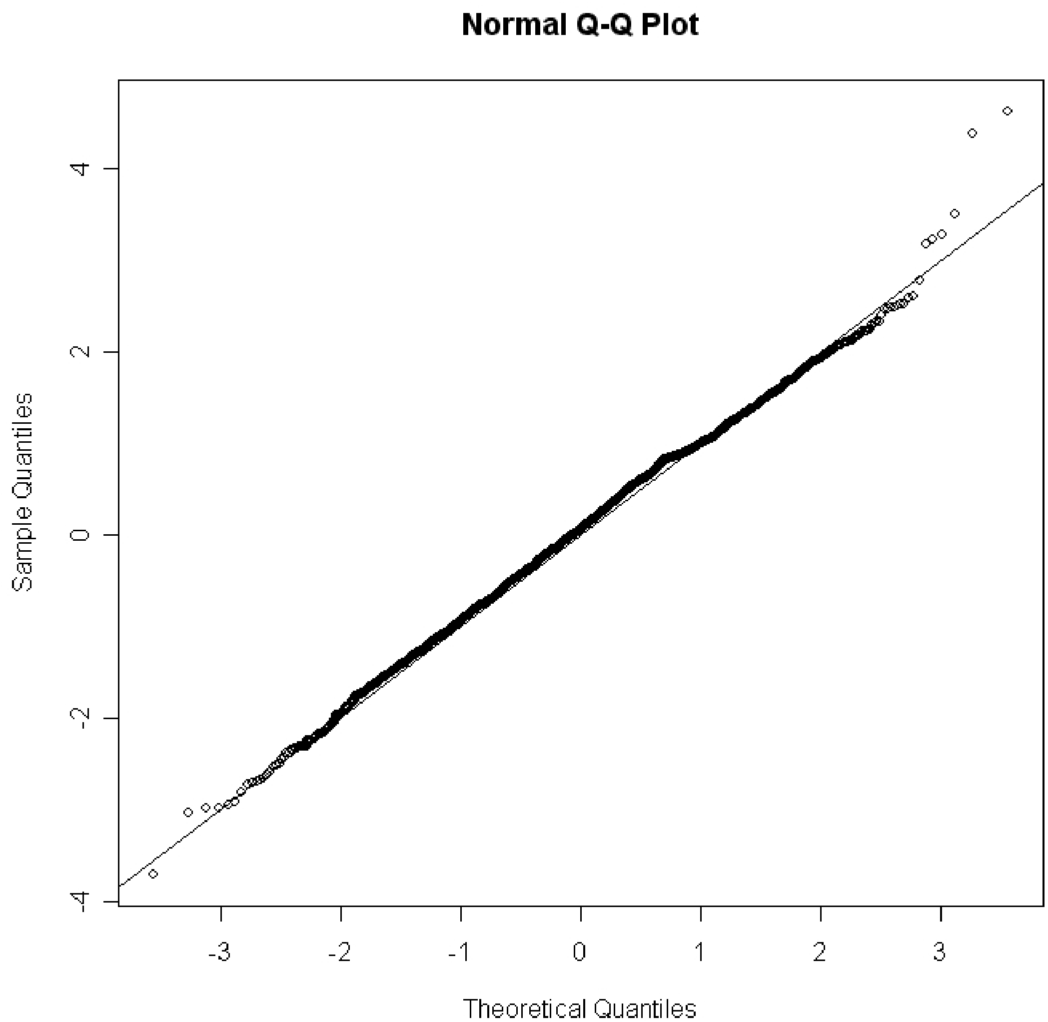

Figure 8.6 The Q-Q Plot of the Quantile Residuals after Fitting a Tweedie Distribution to Monthly Long-Distance VMT

Table 8.3 summarizes the maximum likelihood estimators of the mathematical parameters and the practical estimators. Similar to the case with intra-regional travel, the Tweedie distributions perform less than satisfactory in modeling tour-level long-distance travel activities. The estimated mean number of long-distance tours per month is 0.29, when the sample mean is 0.37, and the estimated distance per tour is 617 miles, higher than the sample mean 471.64 miles. The commonality between the intra-regional travel and long-distance travel models is that the model tends to underestimate the mean number of events (trips or tours) within a time frame (daily or monthly), and therefore overestimate the distance traveled per event (a trip or a tour). However, the estimated values of monthly number of long-distance tours and distance per tours are closer to the actual sample means than the estimated values of daily number of intra-regional trips. This can be explained by a lower degree of over-dispersion in the monthly number of long-distance tours, as will be discussed next.

Table 8.3 The Maximum Likelihood Estimators of Monthly Long-Distance VMT per Household

Mathematical parameters	Estimators	Practical parameters	Interpretation	Estimators
$\hat{\mu}$	176.07	$\hat{\lambda}$	The mean number of tours per month	0.29
$\hat{p}$	1.36	$\hat{\gamma}$	The shape of the distance per tour gamma distribution	342.75
$\hat{\varphi}$	151.57	$-\hat{\alpha}\hat{\gamma}$	The distance per tour	617.34

8.1.3.2. Monthly Number of Long-Distance Tours

Based on the formulation of Tweedie distributions, the monthly number of long-distance tours should follow the Poisson distribution. However, as shown in Figure 8.7, the Poisson distribution does not fit the monthly number of long-distance tours very well.

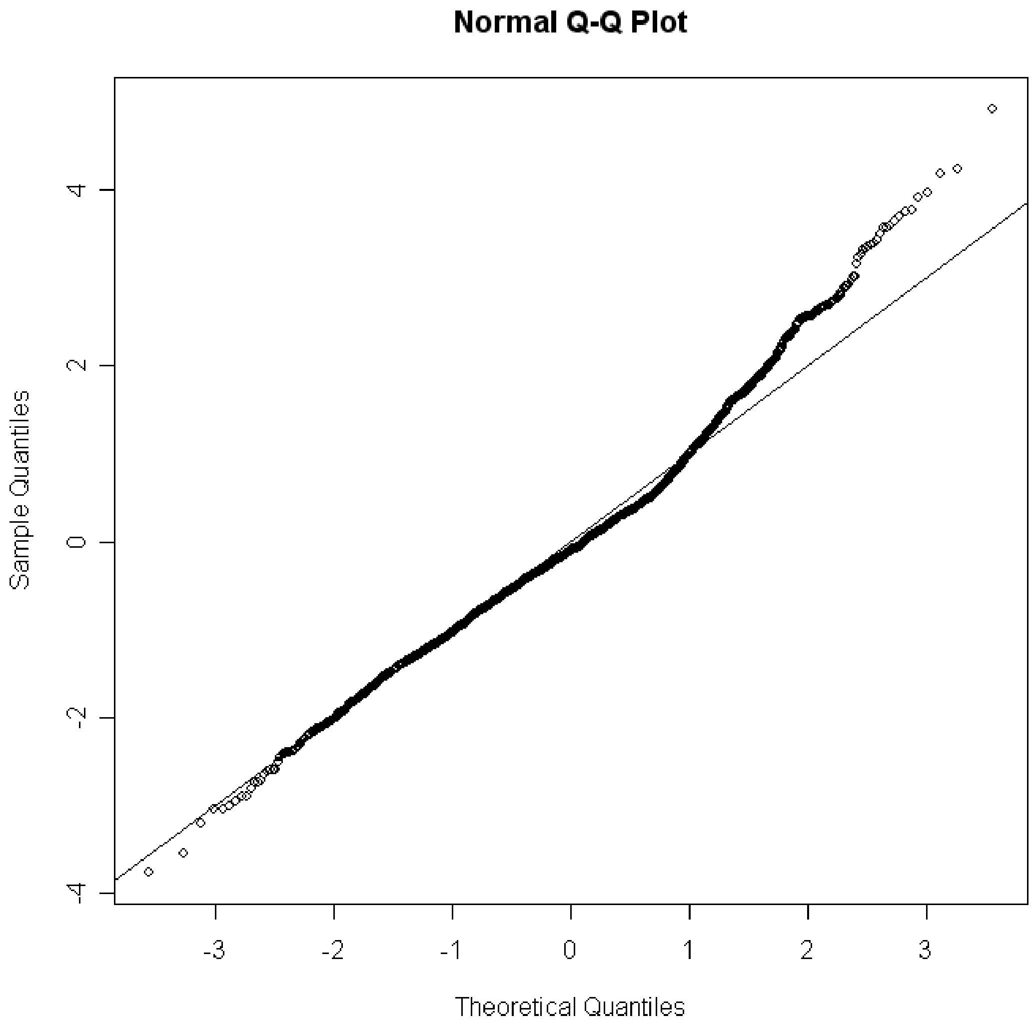

Figure 8.7 The Q-Q Plot of the Quantile Residuals after Fitting a Poisson Distribution to Monthly Number of Long-Distance Tours

The deviation of the empirical distribution from the Poisson assumption is also manifest in the ratio of the sample variance to the mean of the monthly number of long-distance tours. According to the Poisson assumption, the ratio should be 1. However, the data has a variance-to-mean ratio of 1.53. A explanation for over-dispersion (Diggle*, et al.*, 2002) is to assume that given a mean monthly number of long-distance tours μ_i for a household, the Y_{ij}'s are independent Poisson variables with mean and variance equal to

μ_i. The over-dispersion arises because the μ_i's are assumed to vary across households according to a gamma distribution with mean μ and variance $\psi\mu^2$, which determines that the distribution of Y_{ij} has mean μ and variance $\mu + \psi\mu^2$.

To test the first part of the aforementioned assumption, that the Y_{ij}'s are independent Poisson variables with mean and variance equal to μ_i, the distribution of the monthly number of long-distance tours is studied household by household. The chi-square test shows that 7 out of the 94 households have *p*-values smaller than 0.05, indicating that the null hypothesis is rejected for only about 15% of all the households. The independence of Y_{ij}'s within a household can be further evidenced in Section 8.2.2. The distribution of the mean monthly numbers of long-distance tours by household, μ_i, is examined against the gamma distribution to test the second part of the assumption, as shown in Figure 8.8. The plot indicates that the gamma assumption of the μ_i's is acceptable. These findings help explain the deviation of monthly number of long-distance tours from the Poisson distribution.

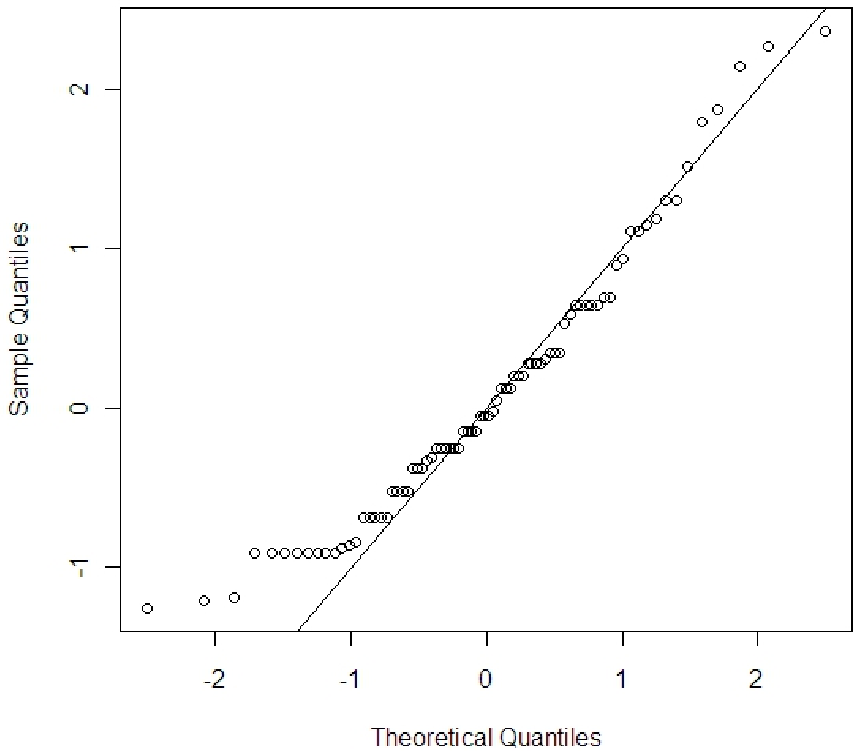

Figure 8.8 The Q-Q Plot the Quantile Residuals after Fitting a Gamma Distribution to the Mean Monthly Number of Long-Distance Tours by Household

It should be noted that most count data are over-dispersed, and the degree of over-dispersion in the monthly number of long-distance tours is modest (Cameron and Trivedi, 1998). In biomedical studies, the variance-to-mean ratio can sometimes be as large as 20; see, for example, (Diggle*, et al.*, 2002). The variance to mean ratio of 1.53 for long-distance travel is smaller than that for intra-regional travel. Therefore, in the regression analysis in the next chapter, this modest over-dispersion will be ignored. Nevertheless, the examination of the Poisson assumption carried out in this section is an indispensible step that ensures the appropriateness of specifying Poisson regression for count data in Section 9.2.

8.2. Within-Household Association

This section explores the degree of association in a longitudinal travel data set. The lack of independence among observations is characteristic of longitudinal data. As mentioned at the beginning of this chapter, the GEE procedures for longitudinal data analysis requires the specification of the degree and structure of within-subject association. Chapter 9 will note the impact of different correlation structures on the goodness-of-fit of a model. In Chapter 9, it will become clear how the assumed specification of within-household association would imply different sample size requirements for before-and-after studies. The purpose of this section is to characterize the within-household association with regard to intra-regional and long-distance travel in preparation for the subsequent regression analysis and sample size planning analysis.

Graphical displays, specifically, scatterplot matrices are normally the first step to assess the degree of association. If each scatterplot in the matrix appears like a sample from the bivariate normal distribution, the association can then be characterized with a correlation matrix, comprised of a correlation coefficient for each plot (Diggle*, et al.*, 2002).

8.2.1. Intra-Regional Travel

To examine the correlation structure, the number and VMT of intra-regional trips are summed by month. As described in Chapter 3, there are 18 months of data altogether.

Graphically, the correlation of intra-regional travel behavior among these 18 months can be studied using a scatterplot matrix in which y_{ij} is plotted against y_{ik} for all

$j < k = 1, \cdots, 18$ (Diggle*, et al.*, 2002), thus producing 153[1] scatterplots of responses from households at different months. The scatterplot matrices for number of intra-regional trips and intra-regional VMT are shown in Figure B.1 and Figure B.2, respectively. Appendix B also includes a detailed description of the scatterplot matrices.

The most important information reflected in Figure B.1 and Figure B.2 is that each scatterplot reasonably resembles a sample from the bivariate normal distribution. Therefore, it is appropriate to summarize the association with correlation matrices as shown in Table 8.4 and Table 8.5, comprised of a Pearson correlation coefficient for each plot.

All the correlation coefficients in both tables are significant at the 0.05 level. The amount of association is substantial, ranging from almost 0.9 between adjacent months to around 0.6 between months that are more than a year apart for the number of daily intra-regional trips, and from 0.45 to 0.9 for daily intra-regional VMT. The correlations show some tendency to decrease with increasing time lag, but the rate of decrease is slow. The italic numbers highlight correlations in the months that are exactly one year apart. The correlations between the same months of different years appear higher than those between the months that are 11 months apart (the diagonal line above the line of red numbers) in most cases, underscoring the impact of seasonality in travel.

[1] 153=18 choose 2

Table 8.4 Estimated Correlation Matrix for Number of Intra-Regional Trips
Entries are Corr(Y_{ij}, Y_{ik}), $1 \le t_{ij} < t_{ik} \le 21$ months .

	t_{ij}																
t_{ik}	1	2	3	4	5	6	7	8	9	13	14	15	16	17	18	19	20
2	.85																
3	.82	.86															
4	.87	.88	.88														
5	.82	.77	.78	.86													
6	.85	.80	.85	.84	.88												
7	.83	.83	.85	.83	.77	.88											
8	.83	.82	.85	.86	.79	.82	.89										
9	.86	.79	.78	.79	.80	.83	.83	.85									
13	*.86*	.75	.81	.81	.83	.87	.82	.83	.84								
14	.73	*.82*	.79	.77	.80	.82	.78	.74	.80	.83							
15	.78	.74	*.82*	.79	.80	.86	.81	.78	.81	.86	.89						
16	.79	.73	.78	*.77*	.79	.81	.77	.75	.78	.87	.87	.86					
17	.70	.74	.74	.75	*.81*	.79	.74	.76	.78	.79	.88	.87	.86				
18	.80	.83	.82	.78	.75	*.81*	.78	.72	.77	.80	.86	.85	.85	.88			
19	.78	.73	.77	.77	.76	.85	*.82*	.76	.83	.79	.79	.84	.80	.81	.86		
20	.70	.70	.69	.59	.64	.71	.65	*.71*	.68	.74	.77	.72	.74	.79	.82	.79	
21	.72	.58	.61	.64	.71	.77	.68	.68	*.74*	.80	.71	.72	.75	.68	.70	.82	.81

Table 8.5 Estimated Correlation Matrix for Intra-Regional VMT
Entries are Corr(Y_{ij}, Y_{ik}), $1 \le t_{ij} < t_{ik} \le 21$ months .

	t_{ij}																
t_{ik}	1	2	3	4	5	6	7	8	9	13	14	15	16	17	18	19	20
2	.86																
3	.79	.82															
4	.85	.85	.85														
5	.80	.80	.78	.87													
6	.85	.78	.83	.84	.86												
7	.85	.81	.84	.83	.78	.85											
8	.82	.80	.82	.83	.81	.82	.90										
9	.78	.69	.63	.70	.76	.73	.79	.83									
13	*.75*	.72	.73	.74	.78	.75	.78	.81	.80								
14	.65	*.70*	.70	.66	.75	.66	.71	.72	.77	.87							
15	.62	.63	*.72*	.70	.74	.68	.71	.73	.76	.84	.88						
16	.72	.67	.72	*.70*	.79	.73	.76	.76	.78	.87	.82	.85					
17	.62	.66	.64	.62	*.75*	.70	.66	.74	.74	.78	.86	.87	.82				
18	.76	.72	.74	.69	.71	*.73*	.72	.72	.77	.83	.85	.90	.84	.89			
19	.66	.60	.56	.65	.68	.68	*.71*	.71	.81	.78	.75	.78	.79	.83	.90		
20	.60	.57	.54	.55	.62	.64	.60	*.66*	.70	.73	.73	.71	.69	.83	.87	.88	
21	.61	.50	.45	.60	.70	.64	.60	.63	*.78*	.75	.67	.64	.69	.65	.72	.83	.81

8.2.2. Long-Distance Tours

As described in Chapter 3, there are 30 months of long-distance data. Therefore there would be 435 (30 choose 2) scatterplots, making a scatterplot matrix too crowded. Therefore, only a selected number of scatterplots are presented to illustrate the correlations.

Figure 8.9 shows the correlation between the numbers of long-distance tours in December, 2004 and January, 2005. The scatterplot does not resemble a sample from bivariate normal distribution. Therefore, a non-parametric correlation coefficient, namely, Spearman’s rho, is used to quantify the degree of association. The estimate of the correlation coefficient is 0.48. However, a closer look at the scatterplot shown in

Figure 8.9 reveals that the correlation coefficient could be misleading, because 66 of the 93 households shown in the scatterplot did not undertake any long-distance tours at all in December 2004 or January 2005. The large number of households with zero long-distance tours may have contributed to the seemingly large correlation coefficient. If one disregards the 66 zero values in Figure 8.9, little to none correlation can be detected.

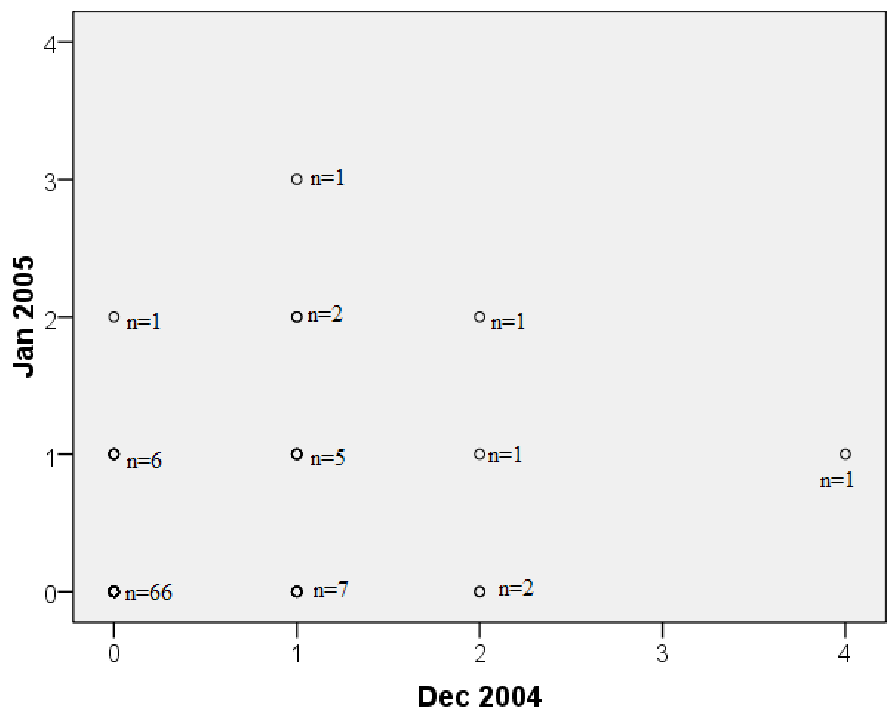

Figure 8.9 Scatterplot of Number of Long-Distance Tours: December 2004 versus January 2005
Total valid number of households for both months = 93

Figure 8.10 shows the correlations of long-distance VMT between December 2004 and January 2005. Again, the scatterplots do not resemble a sample from the bivariate normal distribution. The nonparametric correlation coefficient is 0.49 for the long-distance VMT between December 2004 and January 2005, but could be misleading for the same reason presented for the interpretation of Figure 8.9.

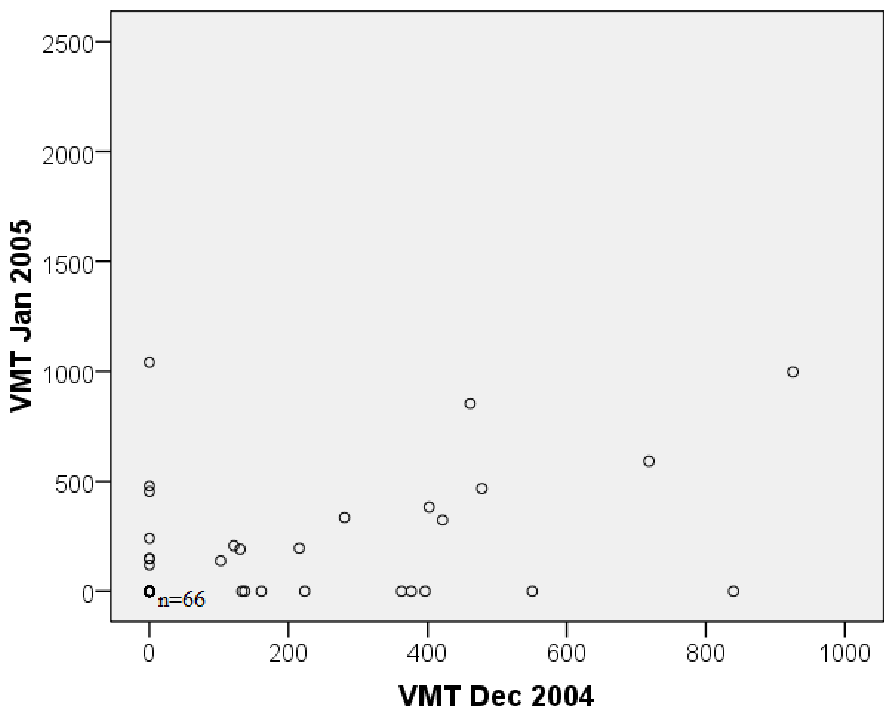

Figure 8.10 Scatterplot of Long-Distance VMT: December 2004 versus January 2005
Total valid number households for both months = 93

The correlation matrices for monthly number of long-distance tours and monthly long-distance VMT, comprised of a Spearman's ρ for each pair of months, are not included in the main body of this document due to their sheer size, but are given in Table B.1 and Table B.2, respectively. The information presented in these two tables indicates that, not surprisingly, the degree of within-household association with regard to long-distance travel is much lower than that with regard to intra-regional travel. In some cases, the correlations are not significant at the 0.05 level.

The little to none positive correlation implies larger within-household variance over time, and hence a smaller chance to detect a significant change in long-distance travel behavior within a household, given a set sample size. This effect will become clear in the next chapter, when cross-sectional and longitudinal impacts of demographic characteristics on monthly number of long-distance tours are estimated and compared.

8.3. Summary

This chapter explored the data properties required by the GEE procedures, in preparation for the regression analysis in the next chapter as a basis for sample size estimation.

Section 8.1 examined the distributional properties of intra-regional and long-distance travel variables and found that Tweedie distributions provide a good approximation for the distributions of VMT per time unit, i.e. intra-regional VMT per day and long-distance VMT per month. The canonical link function of Tweedie distributions, the log link, and the corresponding variance functions will therefore be assumed in the next chapter to model intra-regional VMT per day. Section 8.1 also found that, with proper consideration of over-dispersion, the Poisson distribution provides a basis for daily number of intra-regional trips and monthly number of long-distance tours. The Poisson distribution may approximate the count of long-distance tours better than the count of intra-regional trips. The next chapter will adopt Poisson regression to model the monthly number of long-distance tours.

Section 8.2 explored the within-household association in intra-regional and long-distance travel. Intra-regional travel was found to have a higher degree of association than long-distance travel. In both cases, the degree of association appears to decrease slowly as the time lag between two observations increases.

CHAPTER 9

SAMPLE SIZE ANALYSIS FOR GENERALIZED LINEAR MODELS OF KEY TRAVEL BEHAVIOR VARIABLES

This chapter relates sample size analysis to the requirements of regression analysis. The Transportation Research Board Travel Survey Methods Committee 2009 Research Needs Statements (Travel Survey Methods Committee) emphasized the need to relate sample size calculations to regression analysis rather than the maximum allowable error of certain statistical inferences, most often the mean, on travel behavior variables, such as number of trips per day and total VMT per day. The Travel Survey Methods Committee noted that the sample size requirements for statistical methods used in the development of regression models in travel demand forecasting to detect whether an explanatory variable has a significant impact on travel behavior may be very different from the sample sizes calculated to meet the maximum allowable error requirements on the mean value of a travel behavior variable. Therefore, research into the sample size requirements for common transportation-related regression analysis that are employed based on survey data is warranted.

In particular, this chapter relates sample size estimation to regression analysis for longitudinal data, accounting for the within-household variability and correlation revealed through GPS-based panel surveys. GPS-based panel survey data impose a significant challenge to the aforementioned need for sample size analysis, because a different modeling approach is needed for longitudinal data analysis, as outlined in

Chapter 7. Therefore the objectives of this chapter are twofold: 1) to develop regression models adopting the generalized estimating equation (GEE) procedures appropriate for longitudinal data, and 2) to relate sample size requirements to the regression analysis conducted in the first step.

The goal of the regression analyses in this chapter is not to build advanced models, but to use simple but reasonable models as the bases for sample size analyses. The small number of households in the Commute Atlanta datasets poses significant limitation on how much cross-sectional information can be identified in regression models, but the extended survey length makes it possible to identify longitudinal effects. Consequently, this chapter underscores the need to use models that suit the longitudinal nature and the data distributions of intra-regional VMT and number of long-distance tours. By doing so, the chapter highlights the differences between cross-sectional and longitudinal studies and their respective strengths, and their different sample size requirements.

Section 7.2 provided detailed discussions on cross-sectional and longitudinal effects with a hypothesized example. The modeling exercises in this chapter will illustrate the concepts outlined in Section 7.2 with empirical evidence from the Commute Atlanta data. Section 9.1 focuses on daily intra-regional VMT, whereas Section 9.2 will focus on long-distance tours.

9.1. Intra-Regional Travel

9.1.1. Model Estimation

The goals of model estimation in this section are: 1) to illustrate the differences between cross-sectional and longitudinal effects, and 2) to form a basis for the subsequent sample size analysis. Hence, the model set-up is simplified, choosing fewer variables than those included in a practical travel demand model, for a concise discussion.

Daily intra-regional VMT are regressed against two variables: household size and number of vehicles per adult. A variable indicating household income - income per adult - was included in initial analysis, but is not included in the model in this section because the coefficient for the income per adult variable is not significant, most likely for two reasons. First, the impacts of household income on daily intra-regional VMT is likely to differ by segments of income levels, and therefore would be inappropriate to be modeled with a linear relationship. Second, income per adult and number of vehicles per adult are positively correlated with a Spearman's ρ value of 0.34, and this positive correlation is statistically significant at the 0.05 level. Kutner *et al.* (2005) noted that correlations among explanatory variables could inflate the variability of estimated regression coefficients and therefore render the coefficients not statistically significant. In future model developments, specifications of the income variable reflecting market segments, rather than a linear approach, will be adopted.

As pointed out in Chapter 4, total number of vehicles owned is positively correlated with household size, as evidenced by a Spearman's ρ value of 0.69. Therefore, the total number of vehicles is standardized by number of adults[1] in the household to

[1] Another approach to account for the correlation between household size and total number of vehicles owned is to use number of vehicles per driver as an explanatory variable. The number of adults and the number of drivers are generally comparable in a household, but there are 11 households in the Commute Atlanta data where the two numbers are not equal. Six of these households had more drivers than adults because some children turned 15 and could obtain learner's permits. The remaining five households had more adults than drivers because some household members were too old to drive. The differences in model goodness-of-fit between using number of vehicles per adult and number of vehicles per driver will be addressed in future research.

account for the positive correlations between total number of vehicles and household size.

To demonstrate the distinctions between cross-sectional and longitudinal effects, this section first constructs a model that assumes that the cross-sectional and longitudinal effects of household demographic characteristics on daily intra-regional VMT are about the same, i.e. $\beta_L = \beta_C$. Subsequently, a second model will be constructed to relax the assumption that $\beta_L = \beta_C$, adopting the form given in Equation (7.6).

9.1.1.1. Assuming $\beta_C = \beta_L$

Based on the discussions in Chapter 8, the chosen model type is the Tweedie distribution with log link. Hence, the model assumes the form

$$\log E(Y_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2}$$

where x_{ij1} and x_{ij2} are household size and number of vehicles per adult, respectively.

The GENLIN command in SPSS gives estimation results as listed in Table 9.1. As expected, household size and number of vehicles per adult have significant positive association with daily intra-regional VMT.

Table 9.1 Parameter Estimates for Modeling Daily Intra-Regional VMT Assuming Longitudinal Effects and Cross-Sectional Effects are Equal (Model 1)

Parameter	β	Standard Error	95% Wald Confidence Interval		Hypothesis Test
			Lower	Upper	*p*-value
Intercept	2.512	.209	2.103	2.922	.000
Household Size	.238	.054	.133	.344	.000
Number of Vehicles per Adult	.453	.100	.256	.650	.000

9.1.1.2. Differentiating β_C and β_L

As described in Chapter 5, households underwent significant demographic changes during the course of the Commute Atlanta study. With regard to the two primary variables - household size, and number of vehicles per adult, the number of households that underwent changes are summarized in Table 9.2. About 30% of all households underwent changes with regard to number of vehicles per adult. The number of households that experienced changes in household size is smaller than those that experienced the other two types of changes, but nevertheless amounts to 14% and warrants analysis on longitudinal effects.

Table 9.2 Number of Households that Underwent Changes

	Household Size	Number of Vehicles per Adult
No Change	82	64
Increase	7	14
Decrease	6	17

To differentiate β_L and β_C, the values of all three variables for each household at the beginning of the study are taken to evaluate the cross-sectional effects, and, based on Equation (7.6), the changes for the p^{th} variable are calculated as:

$$x_{change} = x_{ij} - x_{i1},$$

where x_{i1} is the initial value of the p^{th} variable, and x_{ij} is the value of the pth variable at the j^{th} observation; p=1 or 2.

The previous model is now re-estimated with both β_L's and β_C's. Table 9.3 shows the results.

Table 9.3 Parameter Estimates for GEE of Intra-Regional VMT Contrasting Cross-Sectional and Longitudinal Effects (Model 2)

Parameter	$\boldsymbol{\beta}$	Standard Error	95% Wald Confidence Interval		Hypothesis Test
			Lower	Upper	*p*-value
(Intercept)	2.811	.193	2.431	3.190	.000
Initial Household Size	.246	.044	.160	.332	.000
Initial Number of Vehicles / Adult	.159	.137	-.110	.428	.247
Household Size Change	.240	.073	.097	.383	.001
Number of Vehicles Change / Adult	.465	.115	.240	.689	.000

Comparing Table 9.3 to Table 9.1, the distinctions between cross-sectional and longitudinal effects become evident. First, the cross-sectional and longitudinal effects household size on daily intra-regional VMT are of similar magnitude. The significance of longitudinal effects coincides with the case study findings described in Chapter 5.

Second, the cross-sectional and longitudinal effects of number of vehicles per adult appear to be different. The regression coefficient of the longitudinal effects is 0.47 with a *p*-value of 0. The regression coefficient of the cross-sectional effects is not statistically significant given the Commute Atlanta data. This is not to say households with different numbers of vehicles per adult conduct the same amount of intra-regional travel in the general population. The non-significance of the cross-sectional effects is likely attributable to the small number of households (95) in the data. Also, if the number of vehicles per driver were used as an explanatory variable instead of the number of vehicles per adult, the results might be different. As mentioned earlier, such specifications about vehicle ownership for the regression model, together with the inclusion of income, will be analyzed in future research and reported separately. In comparison, in Model 1 where cross-sectional and longitudinal effects are assumed to be the same, the regression coefficient is 0.45, slightly under-stating the longitudinal effect and possibly over-stating the cross-sectional effects in the Commute Atlanta data.

It is common knowledge that the presence of children influences a household's travel behavior. However, there are only 4 households in the Commute Atlanta sample that experienced changes in the number of children. A much larger sample or a targeted sample[1] will be needed for the analysis on the impact of change in number of children on household travel.

The corrected quasi likelihood under independence model criterion (QICC) can be used as the measure of goodness of fit to compare the two models. The QICC criterion can be used to choose between two sets of model terms, given a correlation structure and is in a small-is-better form (SPSS Inc., 2007), meaning that the model with the smaller value of QICC performs better, but there is no formal statistical test available to determine if the difference between two models is statistically significant. The QICC is 2,752,550 for Model 1, and 2,750,450 for Model 2 indicating that Model 2 does not perform worse than Model 1. More importantly, Model 2 is able to distinguish

[1] For example, a study may wish to recruit newly wedded couples that plan to have children soon to examine the effects of adding a child on travel behavior.

longitudinal and cross-sectional effects. Hence, Model 2 is chosen for the analysis in the next section.

9.1.2. Sample Size Analysis

The small number of households in the Commute Atlanta data poses a limitation on the extent of the sample size analysis for model estimation. The significance of coefficients for explanatory variables is confined to only the prevalent effects observed in the 95 households. Therefore, it is not the intention of such analysis to extrapolate the sample size desired to obtain significance of coefficients for explanatory variables that are not evident in the existing dataset. Rather, the analysis in the section intends to provide a picture of how the coefficients of explanatory variables would change if the number of households is even smaller, or the survey length is shorter with a small number of households, and to set a method that can be used with larger samples in the future. By doing so, this section will demonstrate the trade-offs between detecting cross-sectional effects and longitudinal effects.

Various combinations of number of households and number of months are used to test the performance of Model 2. Households are randomly drawn from the original 95 households. This analysis uses 100%, 75%, and 50% of all households each time. Once a set of households are drawn, the data for each household is carved out for a number of months counted from the beginning of the intra-regional VMT study (October 2004). As noted in the data descriptions in Chapter 3, due to equipment issues, the data for each household tend to be incomplete during random months. About 10% of all data were missing due to equipment issues. The impact of these missing data on this analysis is that some households won't be included in the analysis if they are missing data in the months selected. Exactly how many households are missing the first few months of data for each trial is completely random.

First, all 95 households are included. For the first three (3) months of the study, data are missing for nine (9) households, and hence the 86 households in the 6th row of the table. Next, about 75% of the 95 households were randomly selected, resulting in 79 households. Similarly, data are missing for seven (7) of these 79 households for the first three (3) months, and hence the 72 households in the 12th row. Finally, about 50% of the 95 households are randomly chosen, resulting in 43 households. Data are missing for four (4) of these 43 households for the first three (3) months. The parameter estimates and corresponding *p*-values are listed in Table 9.4.

Table 9.4 Model Performance with Subsets of Sample: Intra-Regional VMT (Model 2)

Number of Households	Number of Months Elapsed from October 2004	Cross-Sectional Effects				Longitudinal Effects			
		Initial Household Size		Initial Number of Vehicles per Adult		Change in Household Size		Change in Number of Vehicles per Adult	
		Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value
95	18	.201	.001	.222	.139	.188	.012	.491	.000
95	15	.191	.000	.309	.023	.177	.009	.501	.000
95	12	.178	.000	.348	.005	.160	.024	.386	.000
95	9	.180	.000	.348	.003	.101	.036	.358	.023
95	6	.169	.000	.350	.005	.149	.011	.484	.021
86	3	.183	.000	.260	.068	.116	.431	.619	.057
79	18	.195	.000	.454	.006	.167	.074	.540	.001
79	15	.186	.000	.521	.000	.140	.085	.505	.000
79	12	.197	.000	.498	.000	.123	.143	.411	.005
79	9	.197	.000	.484	.000	.067	.157	.361	.039
79	6	.189	.000	.485	.000	.112	.054	.502	.045
72	3	.189	.000	.364	.037	.230	.254	.693	.059
43	18	.193	.000	.391	.060	.203	.195	.392	.006
43	15	.178	.000	.355	.049	.237	.089	.528	.007
43	12	.180	.000	.324	.038	.275	.058	.635	.029
43	9	.197	.000	.352	.039	.196	.300	.537	.104
43	6	.196	.000	.305	.093	.192	.405	.789	.056
39	3	.190	.000	.233	.325	.279	.410	.876	.080

The first two coefficients indicate the cross-sectional effects. The initial household size appears stable and stays significant across all subsets. The coefficient of initial number of vehicles per adult shows significant sampling variability. When the sample size is only less than 50% of the original sample, the *p*-values for this coefficient are all larger than the *p*-values when the sample size is larger. This implies that, not surprisingly, a larger sample helps identify cross-sectional effects. However, as the survey length increases, the *p*-values do not always decrease. It implies that, when there is a significant amount of longitudinal variation, which is the case for the working sample, it could be more difficult to differentiate cross-sectional effects as the length of study increases.

The last two coefficients present longitudinal effects. Generally, the *p*-values of all longitudinal effects increase as the number of households decreases *and* as the number of months decreases, as expected. This trend is particularly conspicuous with regard to the coefficient of the number of vehicles per adult, with a few exceptions[1] when the sample size is only 50% of the original sample.

Based on these observations, one could argue that even with a relatively small sample (about 100 households), it is still feasible to observe the impact of some demographic changes (e.g. vehicle ownership) if the study is long enough (preferably more than a year). It is desirable to recruit a larger longitudinal sample, though, because the impact of other demographic changes such as income per adult is less straightforward, and a small sample may provide misleading results, even if the coefficient is statistically significant. For example, β_4 is significant when the survey length is 15 months counting from October 2004 with the original 95 households, but becomes non-significant when the survey length is 18 months. This phenomenon implies large amount of longitudinal variability in the data and that the significance of β_4 with 15 months of data may not be reliable.

If the main objective of a study is to study cross-sectional effects, it may be more beneficial to shorten the study and to expand the number of households, unless both sets of effects are needed for travel demand forecasting. As reviewed in Chapter 2, one of the objectives of the new generation of activity-based models is to forecast long-term travel trends with regard to population aging and other demographic shifts. To achieve this goal, activity-based models need input data that reflect both cross-sectional and

[1] The exceptions are likely because the small sample size (about 43 households) renders the model coefficients unstable to extreme values.

longitudinal information. Given limited resources, a combination of a large sample of short surveys and a small sample of long-range panel surveys are probably optimal for planning purposes.

9.2. Modeling Long-Distance Travel

9.2.1. Model Estimation

This section first builds a model for monthly number of long-distance tours, since this is usually the focus of long-distance trip generation modeling. While building the model will provide insights into long-distance travel behavior, the major purpose of this section is to provide a basis for the next section: sample size analysis. Therefore, as with the intra-regional travel, the model estimation process will purposefully select only a small number of variables. The behavioral interpretations of the model coefficients will also be discussed.

9.2.1.1. Selecting variables

Based on the exploratory data analysis about demographic characteristics presented in Chapter 4, households are categorized into 3 groups: category 1 - multi-person with high income (≥ $75,000 annual income), category 2 - multi-person with low income (<$75,000 annual income), and category 3 - single-person households, as summarized in Table 9.5. These categories are included in the model. Based on the examination of seasonal effects in Chapter 5, the variable indicating the months is also included in the model.

Table 9.5 Summary of Household Categories

Category	Description	Number of Households	Percent
1	Multi-person with high income (≥ \$75,000 annual income)	32	34%
2	Multi-person with low income (<\$75,000 annual income)	37	40%
3	Single-person	25	26%
Total		94[1]	100%

Table 9.6 presents the parameter estimates for the above mentioned variables. The interpretations of the coefficients will be discussed in detail in the next section. At this stage, the point of interest is that the value of the coefficient for the dummy variable "category 2" is not significantly different from 0 compared to reference variable "category 1", indicating that the monthly number of long-distance tours conducted by multi-person households with high income is not significantly different from that conducted by those with low income. This result is contradictory to research expectations.

[1] One household is excluded for the analysis on long-distance travel. This household has a college student who is attending an out-of-town college with a home located at the college, but this student comes back to stay in Atlanta for extended periods. Consequently, all the trips this student makes in Atlanta while staying with the parents are identified as segments of long-distance tours, based on the criteria outlined in Section 3.2.3.

Table 9.6 Parameter Estimates for Modeling Monthly Number of Long-Distance Tours Assuming Longitudinal Effects and Cross-Sectional Effects are Equal (Model 1)

Parameter		**β**	Std. Error	Hypothesis Test
				p-value
Intercept		-1.193	.2169	.000
Month	December	.518	.1683	.002
	November	.559	.1584	.000
	October	.483	.1947	.013
	September	.270	.2075	.193
	August	.304	.1740	.081
	July	.711	.2021	.000
	June	.535	.1739	.002
	May	.579	.1940	.003
	April	.662	.1621	.000
	March	.341	.1624	.036
	February	.247	.1760	.161
	January	Set to 0	NA	NA
Household Category	Single-Person	-1.313	.3528	.000
	Multi-Person Low Income	-.104	.2204	.637
	Multi-person High Income	Set to 0	NA	NA

This contradiction leads to a second look at the model assumptions. As mentioned before, the marginal model approach assumes that $\beta_C = \beta_L$. However, this assumption may not hold with the empirical experience. To examine the validity of this assumption, cross-sectional effects and longitudinal effects are distinguished by taking the model form (7.6). The next step is to incorporate the longitudinal effects into the model. There are only 4 households that experienced changes in their single- or multi-person household status. Therefore, the longitudinal effect of being a single- or multi-person household is not examined due to the small number of households that experienced such changes. The change of income category is calculated as

$$\text{Income category change} = x_{ij} - x_{i1} \tag{9.1}$$

where

$$x\text{ (income category)} = \begin{cases} 1 & \text{annual income} < \$75\text{K} \\ 2 & \text{annual income} \geq \$75\text{K} \end{cases}$$

i indicates the household ID, and

$j = 1, 2, \ldots, 30$ indicates the number of the month relative to the beginning of the study.

Thus, the variable "income category change" takes three values: 1 - indicating that a household changed from low income category to high income category, 0 - indicating that a household stayed in the same income category throughout the study, and -1 - indicating that a household changed from high income category to low income category. Six (6) households changed from high income to low income category, and four (4) changed from low income category to high income category. Table 9.7 summarizes the information described in this paragraph.

Table 9.7 Summary of the Longitudinal Variable "Income Category Change"

Value	Description	Number of Households	Percent
1	a household changed from low income category to high income category	4	4%
0	a household stayed in the same income category throughout the study	84	90%
-1	a household changed from high income category to low income category	6	6%
Total		94	100%

The coefficients in this new model that incorporates this income change variable are summarized in Table 9.8. The results show that the coefficient for the income change is not significant with a *p*-value of 0.77. On the contrary, the coefficient for the variable indicating whether a household belongs to category 2 at the beginning of the study is significant at the 0.10 level, though not at the 0.05 level. Controlling for changes in income categories, the cross-sectional effects of income among multi-person household become significant in Model 2, even though the longitudinal effect is not significant in the sample. There may be a significant lag between the change in income levels and a change in long-distance travel behavior, and the incorporation of the lag variable will be investigated in future models.

Table 9.8 Parameter Estimates for GEE of Monthly Number of Long-Distance Tours Contrasting Cross-Sectional and Longitudinal Effects (Model 2)

Parameter		β	Std. Error	Hypothesis Test
				p-value
Intercept		-1.057	.2092	.000
Month	December	.516	.002	.002
	November	.559	.000	.000
	October	.482	.013	.013
	September	.268	.195	.195
	August	.301	.083	.083
	July	.708	.000	.000
	June	.524	.003	.003
	May	.566	.004	.004
	April	.649	.000	.000
	March	.336	.038	.038
	February	.246	.162	.162
	January	Set to 0	NA	NA
Initial Household Category	Single-Person	-1.351	.3504	.000
	Multi-Person Low Income	-.424	.2521	.092
	Multi-Person High Income	Set to 0	NA	NA
Change in Income Category		.065	.2193	.767

After confirming the non-significance of longitudinal effects through Model 2, the monthly numbers of long-distance tours are regressed against the category status of each household at the beginning of the study to simplify the model. The parameter estimates are presented in Table 9.9. This model that is only concerned with cross-sectional effects is referred to as Model 3.

Table 9.9 Parameter Estimates for Examining Cross-Sectional Effects (Model 3)

Parameter		***β***	Std. Error	95% Wald Confidence Interval		Hypothesis Test
				Lower	Upper	*p*-value
Intercept		-1.053	.2071	-1.459	-.647	.000
Month	December	.517	.1685	.186	.847	.002
	November	.558	.1581	.249	.868	.000
	October	.482	.1945	.101	.863	.013
	September	.268	.2068	-.137	.673	.195
	August	.301	.1738	-.039	.642	.083
	July	.709	.2010	.315	1.103	.000
	June	.524	.1740	.183	.865	.003
	May	.567	.1945	.185	.948	.004
	April	.650	.1627	.331	.969	.000
	March	.336	.1620	.019	.654	.038
	February	.246	.1758	-.099	.590	.162
	January	Set to 0	NA	NA	NA	NA
Initial Household Category	Single-person	-1.357	.3495	-2.042	-.672	.000
	Multi-person Low Income	-.434	.2435	-.911	.043	.075
	Multi-person High Income	Set to 0	NA	NA	NA	NA

Comparing Table 9.9 to Table 9.6, the values for all coefficients are relatively robust, except for the value of the dummy variable for “category 2”. Using the initial status of each household, the new model shows a more significant difference in the monthly numbers of long-distance tours between multi-person households with low income and those with high income, even though not significant enough at the 95% confidence level (p-value=0.075 > 0.05, but < p-value of 0.637 in Model 1). This difference represents the cross-sectional effect of income within multi-person households. This model has a QICC value of 3826, slightly smaller than the QICC value of 3831 for Model 2, indicating that dropping the longitudinal information does not impact model performance negatively.

The results from Model 2 and Model 3 provide insight into the non-significance of income level in Model 1. Model 1 assumes that $\beta_C = \beta_L$ for household income among multi-person households, but Model 2 and Model 3 indicate that this assumption may be incorrect. Model 2 has shown that, with 90% confidence, income level has significant cross-sectional impact on the monthly number of long-distance tours per household. However, the data do not have enough evidence to reject the hypothesis that the longitudinal impact of income levels is not significant. To put it in mathematical terms, Model 2 showed that, with regard to income levels among multi-person households in the Commute Atlanta long-distance dataset, $\beta_C \neq 0$, but $\beta_L = 0$; therefore, $\beta_C \neq \beta_L$.

9.2.1.2. Discussion of model coefficients

Section 9.2.1.1 shows that Model 3 performs the best. This section discusses the interpretations of the model coefficients.

Model 3 takes the form:

$$\begin{aligned}\log E(Y_{ij}) = \beta_0 &+ (\beta_{Feb}x_{j1} + \beta_{Mar}x_{j2} + \beta_{Apr}x_{j3} + \beta_{May}x_{j4} + \beta_{Jun}x_{j5} + \beta_{Jul}x_{j6} + \beta_{Aug}x_{j7} \\ &+ \beta_{Sep}x_{j8} + \beta_{Oct}x_{j9} + \beta_{Nov}x_{j10} + \beta_{Dec}x_{j11}) + (\beta_{Cat2}x_{i12} + \beta_{Cat3}x_{i13})\end{aligned} \tag{9.2}$$

where

$\beta_{Feb},\ldots,\beta_{Dec}$ are the coefficients indicating the impact of a particular month on the numbers of long-distance tours, with respect to January (set as reference purposefully, based on the data patterns seen in Chapter 4),

accordingly, $x_{j1},\ldots,x_{j11}$ [1] are dummy variables indicating whether an observation occurred on a specific month,

β_{Cat2} and β_{Cat3} are the coefficients indicating the cross-sectional impact of a demographic group[2] on the numbers of long-distance tours, with respect to Category 1 (set as reference), and

x_{i12} and x_{i13} [3] are dummy variables indicating whether a household belongs to a specific category at the beginning of the study.

Because the Poisson model is estimated with the link function $h(\mu_i) = \log(\mu_i)$, the interpretation of the regression coefficients requires examining the linear estimates and the exponential estimates simultaneously to obtain sensible understanding of the model. Revisiting Table 9.9, Table 9.10 includes the exponential estimates of the regression coefficients for practical interpretations.

The intercept gives the expected number of long-distance tours for the reference group - the multi-person households with high income in January. The model estimates that the mean number of long-distance tours conducted by multi-person households with high income in January is 0.35.

The coefficients for the month dummy variables all show positive signs, confirming that January is the least active month for long-distance travel. The *p*-values

[1] The footnote *i* is dropped because each household has the same set of variables for the month of the year.
[2] The demographic groups are listed in Table 9.5.
[3] The footnote *j* is dropped because each household has the same initial category assignment throughout the study.

show that the number of long-distance tours in January is significantly different from most other months except February, August and September. The similar values of β's for the months from April to July, and from November to December imply that the number of long-distance tours are not different among these months.

The coefficients for the household group dummy variables both show negative signs, indicating that multi-person households with high income are the most likely to take long-distance tours among all three household groups. The coefficient for multi-person households with low income has a *p*-value of 0.075, which is not significant at the 0.05 level, but is significant at the 0.10 level. The exponential estimates in Table 9.10 imply that, in January month, a multi-person household with low income takes 0.648 times as many long-distance tours as a household in the reference group, whereas a single-person household takes about 0.257 times as many long-distance tours as a household in the reference group.

Table 9.10 Regression Coefficients and their Exponential Estimates for Model 3

Parameter		β	p-value	$\exp(\beta)$
Intercept		-1.053	.000	.349
Month	December	0.517	0.002	1.676
	November	0.558	0	1.748
	October	0.482	0.013	1.619
	September	0.268	0.195	1.307
	August	0.301	0.083	1.352
	July	0.709	0	2.032
	June	0.524	0.003	1.689
	May	0.567	0.004	1.762
	April	0.65	0	1.916
	March	0.336	0.038	1.4
	February	0.246	0.162	1.279
	January	Set to 0	NA	1
Household Category	Single-person	-1.357	.000	.257
	Multi-person Low Income	-.434	.075	.648
	Multi-person High Income	Set to 0	NA	1

9.2.2. Sample Size Analysis

9.2.2.1. Length of Study

Section 9.2.1 has shown that demographic characteristics have different cross-sectional and longitudinal effects on the monthly numbers of long-distance tours a household takes. While the cross-sectional effects are significant with regard to household income and the numbers of household members, the longitudinal effects of income are inconspicuous, probably due to the small sample size. Or, there may be a lag effect between changes in income levels and changes in the monthly number of long-distance tours. Various other factors could also explain this phenomenon, as discussed below.

First, the long-distance travel behavior may be similar to short-distance travel behavior in that they both can be categorized as mandatory and discretionary. In the case of long-distance travel, mandatory demand does not necessarily mean work-related demand. An annual visit to the out-of-town parents could very well be “mandatory”, and such demand is unlikely to change as household demographic characteristics change.

Second, household income is specified as a categorical variable for the long-distance regression analysis. This means that a household whose income changed from $74,000 per year to $76,000 a year will be treated the same as a household whose income changed from $30,000 to $ 90,000. To measure income change in this fashion could have undermined the model’s ability to correctly estimate the longitudinal effects of income changes on long-distance travel. The specification of the model will be refined in future research.

A final observation is that this dataset only records auto tours. The change of long-distance travel by other modes of travel, e.g. air, is not captured. Households could very well increase long-distance travel as income and/or the number of household members increase, but they may choose air travel over auto, rendering constant or even reduced numbers of long-distance tours collected in the dataset, and therefore making the longitudinal effects not statistically significant. Future research could assess air travel by identifying trips to the Atlanta Hartsfield-Jackson airport.

The changes concerning children are likely to have a significant impact on long-distance travel behavior, too. For example, when a household has a new-born baby, it is likely to reduce long-distance travel for a couple of years. When a child reaches school age, the household’s long-distance travel decisions are likely to be constrained by the school schedule. However, the dataset does not include enough households that experienced such changes for these analyses.

These observations suggest that a much longer study period and much larger samples are likely required to obtain longitudinal information with regard to household long-distance travel behavior, because long-distance travel is a rare event compared to intra-regional travel, and may change only gradually over time in response to demographic and/or other changes. The need for a large longitudinal sample for long-distance travel is especially compelling in the context of policy concerns such as fuel consumption and high speed rail construction.

9.2.2.2. Sample Size Planning

To test the impacts sample size and survey lengths on model performance, various subsets of the data are used to estimate Model 3 as described in Section 9.2.1. The method to select subsets of the sample is the same as described in Section 9.1.2.

β_{Cat2}, whose p-values are plotted against number of months counting from October 2004 in Figure 9.1, is given special attention because of the behavioral implication it represents - the association between income levels and the frequency of long-distance travel, and because of the borderline *p*-value - larger than 0.05 but smaller than 0.10.

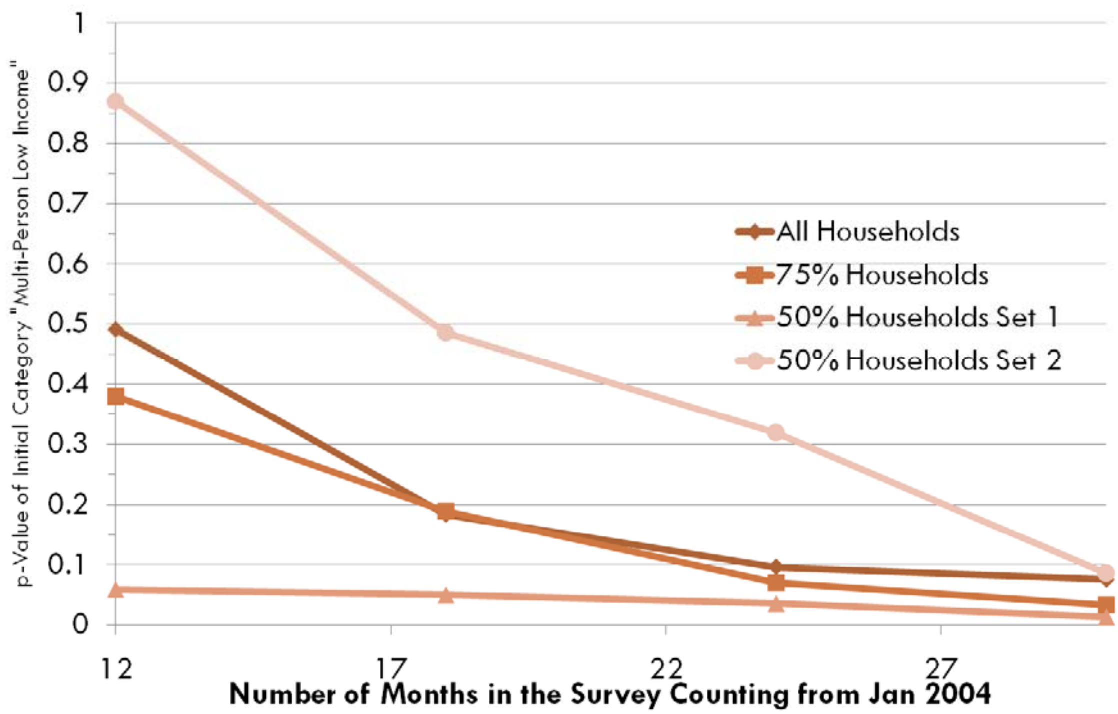

Figure 9.1 The Trend of p-Values for β_{Cat2} as the Number of Months Increases, Counting from October 2004

With the same number of households, all p-values increase as the number of months from the fixed starting month (January 2004) decreases, as expected. The model effects with different numbers of households, however, show less straightforward results. For example, in trials with 30 months of data, the p-value for β_{Cat2} decreases as the number of households decreases from all 94 households to 75% of all households, but the p-values for the two subsets of 50% of all households are very different. This phenomenon suggests that there is a lot of variability across demographic groups, in that some households behave more in an "expected" way in terms of long-distance travel, while other households tend to defy the trend. This observation implies that the current sample does not contain enough households for the targeted regression effect.

9.3. Summary

This chapter has underscored a few important points regarding longitudinal studies, as summarized below.

It is important to adopt a suitable model form for longitudinal data. The regression analyses conducted for both daily intra-regional VMT and monthly number of long-distance tours suggest that the longitudinal effects and cross-sectional effects may be different for certain variables. To assume that longitudinal effects and cross-sectional effects are the same could lead to biased conclusions in modeling practices (Diggle*, et al.*, 2002).

Long-distance travel and intra-regional travel may be driven by very different factors, and therefore may present different patterns in response to changes in household demographics. A much larger sample size and a longer survey period are needed for modeling long-distance travel than for intra-regional travel.

The strength of a longitudinal study lies in the study of change (Fitzmaurice*, et al.*, 2004). For intra-regional travel, this chapter has demonstrated that a small sample size (about 100 households) with a long study period (preferably more than a year) could present significant longitudinal effects for some variables such as household size and number of vehicles per adult. It would require a large number of households and a long survey (preferably more than a year) to detect *both* the cross-sectional and longitudinal effects. This would require many resources that may not be available at most MPOs. If, however, the main purpose of a study is to examine how households would change their intra-regional travel behavior as their demographic characteristics change, a small number of households combined with an extended survey period may be sufficient.

Therefore, a combination of a large-sample one-day or two-day travel survey and a small-sample longitudinal monitoring program appears ideal to examine both the cross-sectional variability in a region and the trends over time. Further, the longitudinal surveys can be a combination of continuous monitoring over several months, and a multi-wave panel over a year consisting of two-day surveys in each wave.

CHAPTER 10

SAMPLE SIZE ANALYSIS FOR BEFORE-AND-AFTER STUDIES

As outlined in Chapter 7, sample size calculation algorithms for longitudinal studies are applied frequently in biomedical research, but not in travel behavior studies. The task of this chapter is to adapt existing algorithms to travel behavior studies. The challenges are two-fold: 1) to translate the existing statistical methods into practical procedures that can be applied to transportation policy evaluation, and 2) to estimate the inputs needed in these procedures based on experiences from the Commute Atlanta study and applications of reasonable assumptions.

This chapter uses the Commute Atlanta study as an example for sample size analysis. The first section of this chapter specifies the inputs required for the sample size analysis as outlined in Chapter 7. The second section of this chapter examines the sensitivity of estimated sample size requirements to the eliminations of extreme values in the data. The third section presents the relationship between the minimum sample size and the degree of within-household correlation in the data. The fourth section of this chapter analyzes sample size requirement as the length of the survey varies. The fifth section examines the relationship between the required sample size and the magnitude of the demand elasticity expected due to a policy change. The sixth section extends the discussions to the missing data issue. A summary of findings is given at the end this chapter.

10.1. Input Parameters for Sample Size Analysis

In the Commute Atlanta study, household VMT were summarized once every month from October 2004 to June 2005 as the baseline period and from Oct 2005 to June 2006 as the pricing period. Therefore, there are T=18 time points and the vector $\boldsymbol{\tau} = [1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 13 \quad 14 \quad 15 \quad 16 \quad 17 \quad 18 \quad 19 \quad 20 \quad 21]$.

10.1.1. Design Matrix

Three sets of variables are included in the design matrix: a (1×12) vector of dummy variables indicating month of year to control for seasonal effects, a scalar variable z to measure gasoline price changes with regard to the beginning of the study, and a dummy variable indicating the application of pricing incentives. The variable z is calculated as

$$z = \log\,(\text{gasoline price}_j/\text{gasoline price}_0) \tag{10.1}$$

where gasoline price$_0$ is the gasoline price at the beginning of the study. Equation (10.1) takes the log form commonly adopted in studies on gasoline price elasticity (Puller and Greening, 1999; Parry and Small, 2005). To put in practical terms, the log form implies a non-linear relationship household VMT decrease and gasoline price increase. The calculations use the lower Atlantic gasoline price data provided by the Energy Information Administration (2010).

Arguably, it is likely impossible to predict gasoline price changes at the design stage. However, this analysis includes gasoline price changes as a variable to emphasize the importance of considering such external variables when designing before-and-after studies. When changes in gasoline price and other exogenous variables cannot be controlled for in practice, researchers need to increase sample size to account for the background variability. Such matrix is provided in Table 10.1.

Table 10.1 Design Matrix

T	X										
	Jan	Feb	Mar	Apr	May	Jun	Oct	Nov	Dec	Log(*z*)	Pricing Incentives Applied
1	0	0	0	0	0	0	1	0	0	0.00	0
2	0	0	0	0	0	0	0	1	0	0.00	0
3	0	0	0	0	0	0	0	0	1	-0.06	0
4	1	0	0	0	0	0	0	0	0	-0.07	0
5	0	1	0	0	0	0	0	0	0	-0.04	0
6	0	0	1	0	0	0	0	0	0	0.05	0
7	0	0	0	1	0	0	0	0	0	0.13	0
8	0	0	0	0	1	0	0	0	0	0.09	0
9	0	0	0	0	0	1	0	0	0	0.08	0
10	0	0	0	0	0	0	1	0	0	0.36	1
11	0	0	0	0	0	0	0	1	0	0.15	1
12	0	0	0	0	0	0	0	0	1	0.10	1
13	1	0	0	0	0	0	0	0	0	0.17	1
14	0	1	0	0	0	0	0	0	0	0.15	1
15	0	0	1	0	0	0	0	0	0	0.21	1
16	0	0	0	1	0	0	0	0	0	0.34	1
17	0	0	0	0	1	0	0	0	0	0.37	1
18	0	0	0	0	0	1	0	0	0	0.36	1

10.1.2. Stochastic Properties of Outcome Variable

Chapter 8 presented the suitability of Tweedie distributions to the daily intra-regional VMT data. However, the algorithms developed by Rochon (1998) do not include the Tweedie distributions as an option of underlying distributions. An alternative is to model weekday VMT with gamma distribution, since there are very few zero values of daily VMT in the weekday travel dataset.

Gamma distribution fits the monthly weekday intra-regional VMT data well, as shown in Figure 10.1 (Kleinman and Horton, 2010). The goodness-of-fit is further confirmed with the Q-Q plot shown in Figure 10.2. The scale parameter estimate of the gamma distribution is 393 and the shape parameter estimate is 2.17, resulting in the expected variance of 335,413. The empirical variance is 335,594, very close to the expected variance. Therefore, the parameter ψ indicating under- or over-dispersion is set to 1.

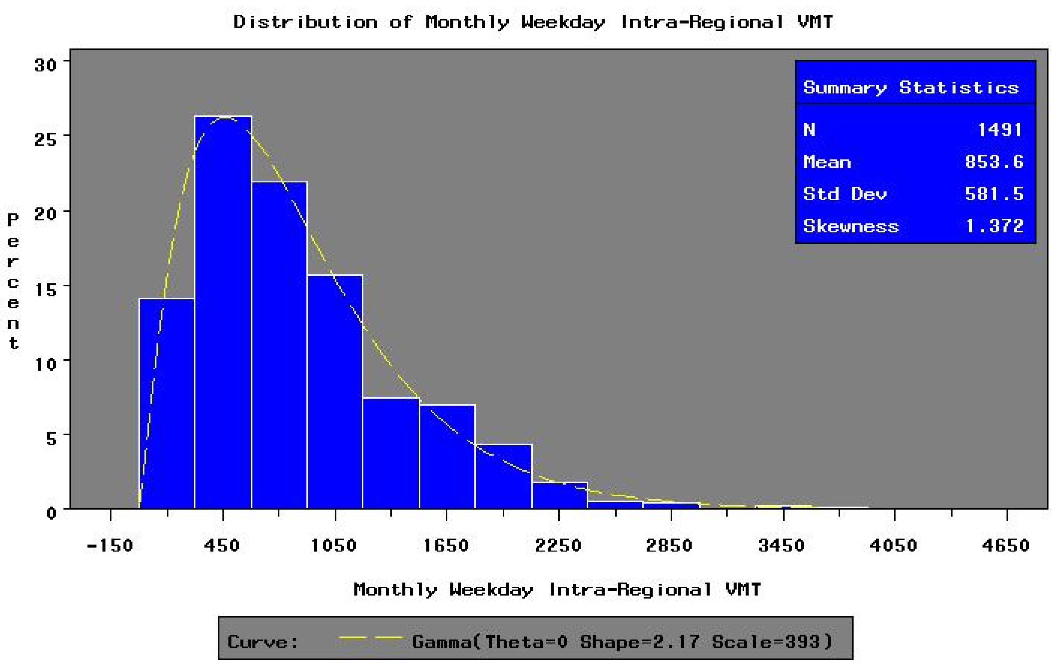

Figure 10.1 Histogram of monthly weekday intra-regional VMT with estimated gamma density curve

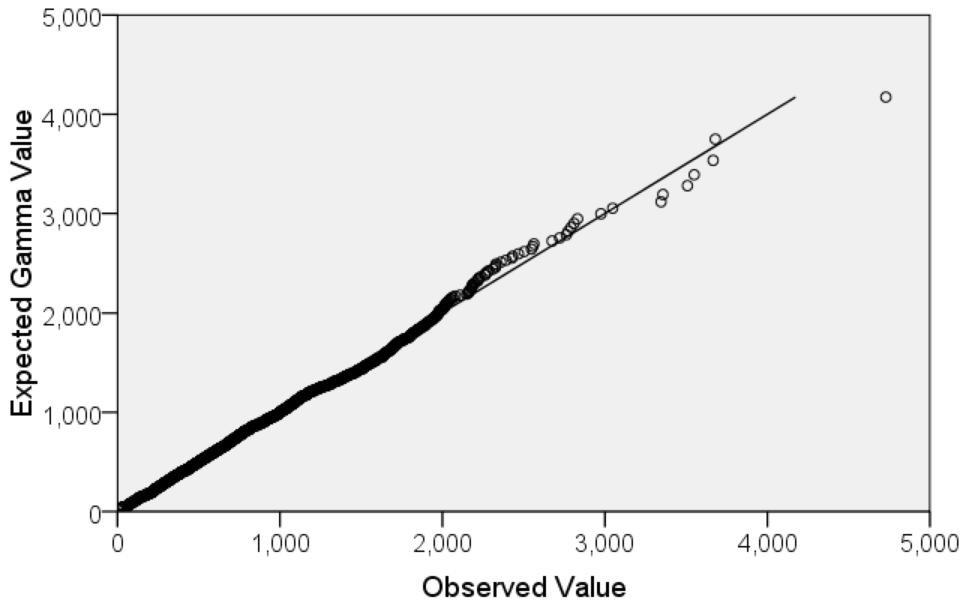

Figure 10.2 Gamma Q-Q Plot of monthly weekday intra-regional VMT

In comparison to Section 8.1.2, where Tweedie distributions were shown to fit daily intra-regional VMT data well, monthly weekday intra-regional VMT data rarely have zero values, so the suitability of gamma distribution is not surprising. The log link is adopted as the norm for gamma distributed data.

10.1.3. Correlation Structure

As discussed in Chapter 7, the exchangeable correlation structure is a reasonable assumption. The GENLIN command of SPSS gives the correlation estimate of 0.721. The sample size requirements will first be analyzed under the assumption of the exchangeable correlation structure.

The discussions will also consider "damped" correlation structures[1] (Muñoz*, et al.*, 1992) with $\theta = 0.5$ and $\phi = 0.75$, and with $\theta = 0.5$ and $\phi = 0.85$. The assumption of the

[1] Please refer to Chapter 7 for the specifications of the damped correlation structures.

damped correlation structures reflects the trend observed in Table 8.5, where the decrease in correlation when two repeated observations are further apart in time, but the θ value of 0.5 implies that the rate of decrease is slower than that of an autoregressive correlation structure.

10.1.4. Expected Outcomes

Consider the number of households needed under the design of the pricing element in the Commute Atlanta study. For simplicity, the sample will not be divided into subpopulations in this initial stage, hence S=1. The expected outcome values, as summarized in Table 10.2, are estimated through the Commute Atlanta study intra-regional dataset with adjustments for pricing impacts, assuming that all households respond to the pricing policy with similar magnitude of VMT reduction[1]. The basic idea is to impose pricing effects on the observed values. As discussed in Chapter 5, the case studies of every household in this study have indicated that the mileage-based pricing incentives seem to have influenced the travel behavior of some households at various levels, but the impact is not significant in the entire sample. Therefore, this analysis needs to impose a pricing impact to estimate the mean monthly weekday intra-regional VMT values for the pricing period. The detailed procedures upon which the calculation of the pricing impact is based are described in detail below.

In Table 10.2, the expected sample mean for the pricing period is obtained through multiplying the observed sample mean by the amount of the pricing impact. The

[1] In practice, some households may be more sensitive to pricing incentives than others, but the small number of households in the Commute Atlanta study was not able to provide evidence on how much the differences may be between different demographic groups. Therefore, a single impact value is assumed for all households. Future research is warranted to investigate the differential pricing impacts and the associated sample size requirements in stratified sampling.

amount of the pricing impact, i.e. coefficient for the variable indicating the policy measure, is decided based on the amount of the pricing incentives and the price elasticity of travel demand, described as the following.

1. The absolute amount of pricing incentives.

   In the Commute Atlanta study, the largest value of pricing incentives was 15 cents per mile, which is very close to the optimal VMT tax of 14 cents per mile suggested by Parry and Small (2005). This equates to about 290 cents per gallon of gasoline.

2. The amount of pricing incentives relative to gasoline prices

   The average gasoline price during the study period from October 2004 to June 2006 is 234 cents per gallon. Therefore, applying the mileage-based pricing incentive of 15 cents per mile is equivalent to setting the gasoline price by 2.24 times as high as the original average price[1].

3. Gasoline Price Elasticity

   The price elasticity of travel demand is estimated based on gasoline demand elasticity with respect to VMT. The literature review identified the work of Parry and Small (2005), who adopted the value of -0.22 for the VMT portion of the gasoline demand elasticity in their analysis of gasoline taxation. This value will be employed for the initial analysis. In Section 10.5, a range of assumed gasoline price elasticity will be tested for the corresponding sample size requirements.

[1] This is to assume that households respond to the pricing incentive in the Commute Atlanta pricing study similarly to they would with out-of-pocket gasoline costs.

4. Corresponding Percentage Change in VMT

Based on the gasoline price elasticity value of -0.22, a 15-cent-per-mile VMT charge would calculate to a coefficient value of -0.18 for the policy measure indicator[1]. As mentioned in Section 10.1.2, the data distribution is assumed to be gamma and modeled with a log link. Hence, the -0.18 coefficient translates to an exponential coefficient value of 0.83[2], meaning that weekday intra-regional VMT is expected to be 83% of the baseline amount during the pricing period[3].

---

[1] -0.18=-0.22*ln(2.24)

[2] 0.83=exp(-0.18)

[3] Please refer to Section 9.2.1.2 for a similar discussion on the interpretation of coefficients based on a log link.

Table 10.2 Baseline Means and Expected Pricing Period Means of Monthly Weekday Intra-Regional VMT at Baseline and Pricing Periods

Observation Number	Follow-up Times	Pricing	Observed Sample Mean	Expected Sample Mean Had Pricing Effects Been Present[1]
1	1	0	906	
2	2	0	853	
3	3	0	902	
4	4	0	821	
5	5	0	755	
6	6	0	901	
7	7	0	868	
8	8	0	910	
9	9	0	914	
10	13	1	817	683
11	14	1	817	682
12	15	1	866	723
13	16	1	724	605
14	17	1	762	636
15	18	1	943	787
16	19	1	764	638
17	20	1	978	817
18	21	1	841	702

The above procedures demonstrated the calculation of expected sample means under the assumption that the gasoline price elasticity with regard to VMT is -0.22. In the remainder of this chapter, sample size requirements will be estimated for two other gasoline price elasticity values: -0.11 and -0.30.

To test the null hypothesis that households do not adjust their weekday intra-regional VMT to the pricing incentives, the hypothesis matrix would be

$$\mathbf{H} = [\mathbf{0}_{(1\times 10)}\ 1]$$

with $\mathbf{h_0} = 0$.

[1] Assuming gasoline price elasticity with regard to VMT is -0.22 as recommended by Parry and Small (2005).

10.2. Sensitivity to Extreme Values

The case-by-case studies (Xu*, et al.*, 2009a) of the 95 households in the Commute Atlanta sample indicate significant variability between-household and within-household. Extreme values exist for some households in certain months, often associated with background changes that a travel survey is not able to control for. Such changes could be a vehicle breaking down that limits the amount of travel in a household during a certain period of time, or a student starting a job that requires a significant amount of travel. Extreme values could have a significant impact on the estimated minimum sample size, so this section tests the sensitivity of sample size requirements to the elimination of extreme values.

In this analysis, an extreme value is determined if an observation for a certain household in a certain month has a relative deviation that is above or below a given percentile among all household-month observations. The relative deviation is defined as:

$$\text{Relative deviation} = \frac{\text{Observed Value - Predicted Value}}{\text{Predicted Value}}$$

The predicted values are calculated based on the regression coefficients for the variables in the design matrix exhibited in Table 10.1, estimated through the GEE algorithms developed by Rochon (1998). Two scenarios are tested: 1) observations with the highest 1% and lowest 1% relative deviations are trimmed, and 2) observations with the highest 2.5% and lowest 2.5% relative deviations are trimmed. The resulting minimum sample sizes are shown in Figure 10.3, in comparison with the minimum sample size based on the original (untrimmed) data. Figure 10.3 indicates that to include extreme values in the pricing study would require a very large sample size of 2,263 households. Eliminating one percent of the household-month observations on each extreme side would result in a smaller sample size of 1,671 households, but a study with this sample size would still be very expensive. Eliminating 2.5 percent of the observations on each side (5% in total) would result in a much smaller sample size of 528 households. A study with 528 households would be reasonable in practice, and the subsequent analyses will be based on this trimmed sample in which 2.5% of the observations are eliminated on each side.

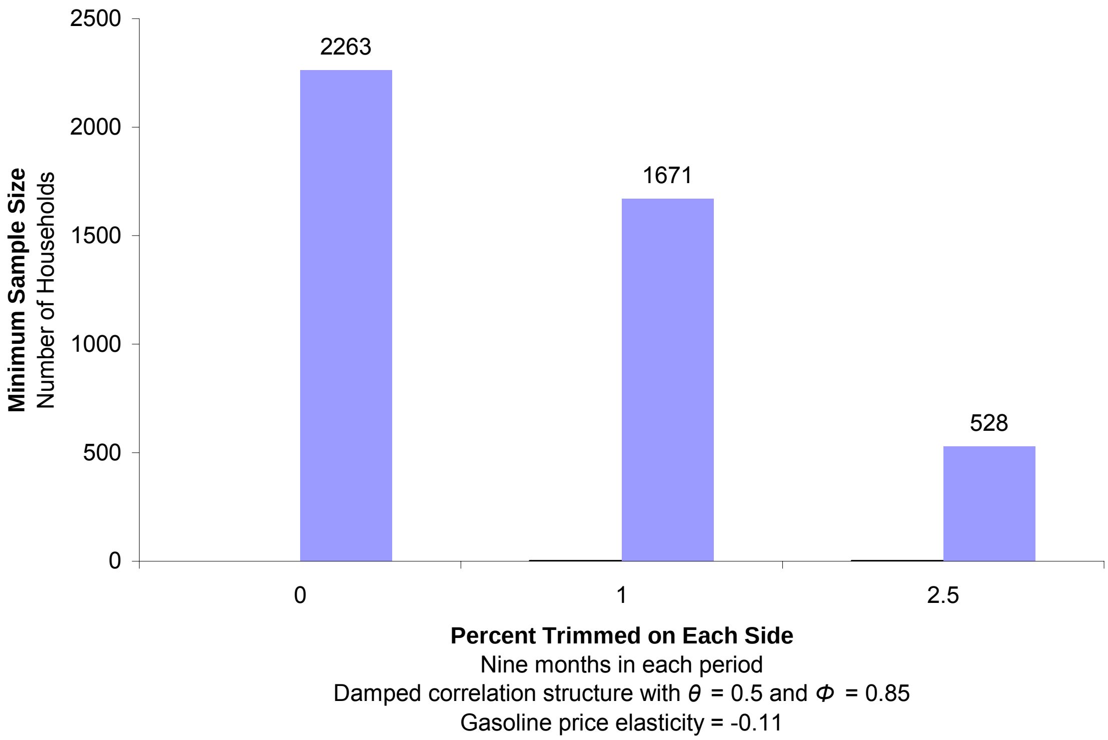

Figure 10.3 Required Sample Size in Relation to Eliminating Extreme Values
Assumptions: damped correlation structure with $\theta = 0.5$ and $\phi = 0.85$

It is important for practitioners to exercise extreme caution in considering extreme values in the design stages and in the analytical stages of consumer response research. One the one hand, it is important to eliminate true outliers from analyses when such households do not represent normal behavioral change, unduly influence the model outputs, and can for a specific reason be removed and modeled separately. On the other hand, when model developers remove observations with extreme values based purely upon standard procedures without studying the potential causes for such extreme values, they could remove valid data that represent the behavior of a true subset of the population from the analytical dataset, and therefore run the risk of deriving biased model estimators. This dissertation confirmed the appropriateness to remove extreme values based on the case-by-case studies of household demographic information (Xu*, et al.*, 2009a) to ensure the eliminated observations are true outliers.

In the design stage, identifying and planning ahead for potential deviations from expected behavioral change within sub-populations is critical. For example, the Commute Atlanta study found that households using a vehicle that they identify a vehicle as being used in part "for commercial purposes" exhibited significantly elevated travel patterns, which may or may not be influenced by pricing (Elango*, et al.*, 2007). Household groups that appear to remain uninfluenced by a pricing stimulus for some specific reason should be identified and sampled separately in future studies. One-time changes in household behavior may also unduly influence sample size estimation. For example, the return of a child from college for spring break may significantly influence household travel for one month and mileage within this household may increase significantly rather than decrease slightly in response to pricing. By including such single-point anomalies in the distribution used to establish sample size requirements, significantly inflated sample size estimates could result.

However, when extreme values do not occur on a random basis, simply eliminating extreme values from any analysis without thorough investigation could create significant biases in predictions. That is, the "rare events" that led to the extreme values in the Commute Atlanta data may not be very rare and larger sample sizes may therefore be needed to accommodate observations that do not conform to the research expectations of a study.

In planning future before-and-after studies designed to assess consumer response to policy initiatives, it will be important to collect more detailed post-experiment survey data, e.g. via interactive focus groups and personal interviews, especially from those households that exhibit large deviations from expected behavior. Researchers need to ensure that there is a reason for removing these households from any analytical dataset used in establishing sample size criteria or else risk under-sampling in future endeavors.

10.3. Sample Size versus Assumptions on Correlation Structures

The degree of within-household association plays an important role in sample size requirements for longitudinal studies; generally, the required sample size increases as the within-household association increases, when estimating a change over time (Diggle*, et al.*, 2002). Based on the correlation structure of intra-regional VMT in the Commute Atlanta sample, as examined in Chapter 8, a damped correlation structure with $\theta = 0.5$ and $\phi = 0.85$ would fit the data well. A more commonly seen structure, the exchangeable correlation structure[1] with $\phi = 0.72$, would also be a reasonable assumption. It is also acknowledged that the degree of within-household association could very well differ across demographic groups. For example, a two-person household in which both members have inflexible work schedules is likely to show more correlation in intra-regional VMT than a two-person household in which both members are retired. Therefore, studies targeting different demographic groups should consider different

[1] The exchangeable correlation structure is a special case of the damped correlation structure when $\theta = 0$.

correlation structures among observations within a household. To examine the sensitivity of sample size requirement to correlation structures, three scenarios are specified: 1) θ= 0, ϕ= 0.72, 2) θ= 0.5, ϕ= 0.85, and 3) θ= 0.5, ϕ= 0.75. The degrees of within-household association under scenarios 1) and 2) are comparable, and scenario 3) assumes a lower degree of within-household association. The corresponding sample size requirements are shown in Figure 10.4.

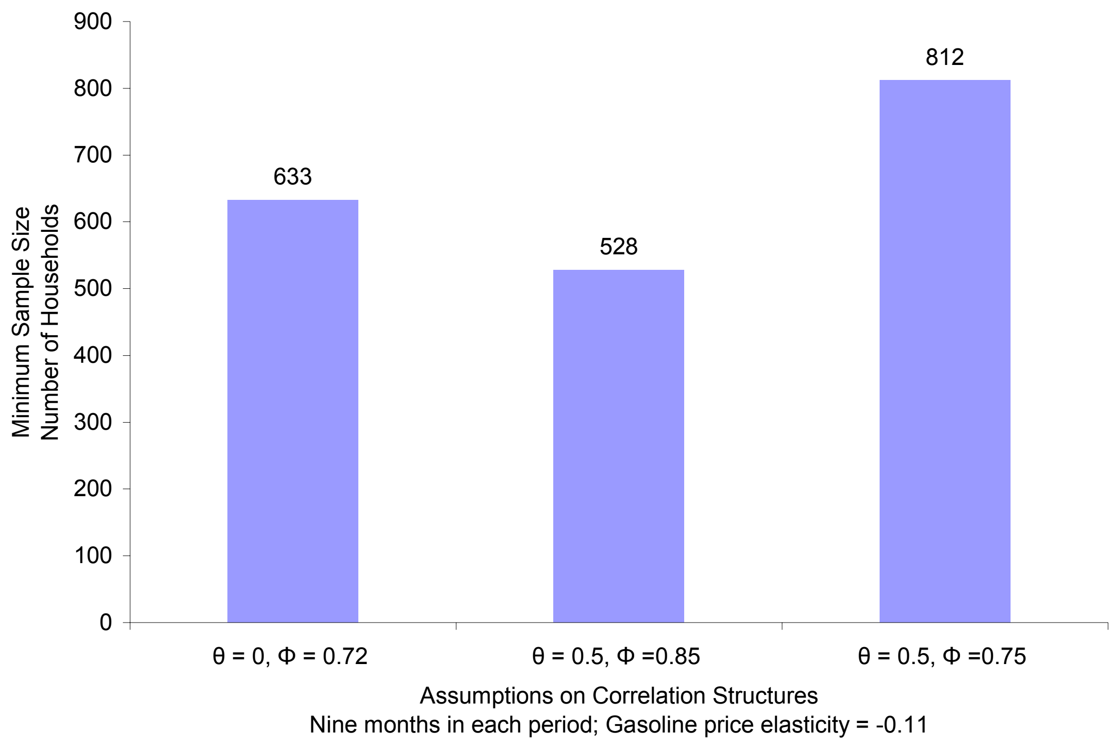

Figure 10.4 Sample Size versus Correlation Structures

10.4. Sample Size versus Number of Survey Months

In this section, the number of repeated observations and the follow-up times are varied to test the impact on required sample size. The pricing period months are a year apart from baseline months to control for seasonality. The number of months elapsed from October in each period is reduced for sample size analysis. The results are presented in Figure 10.5.

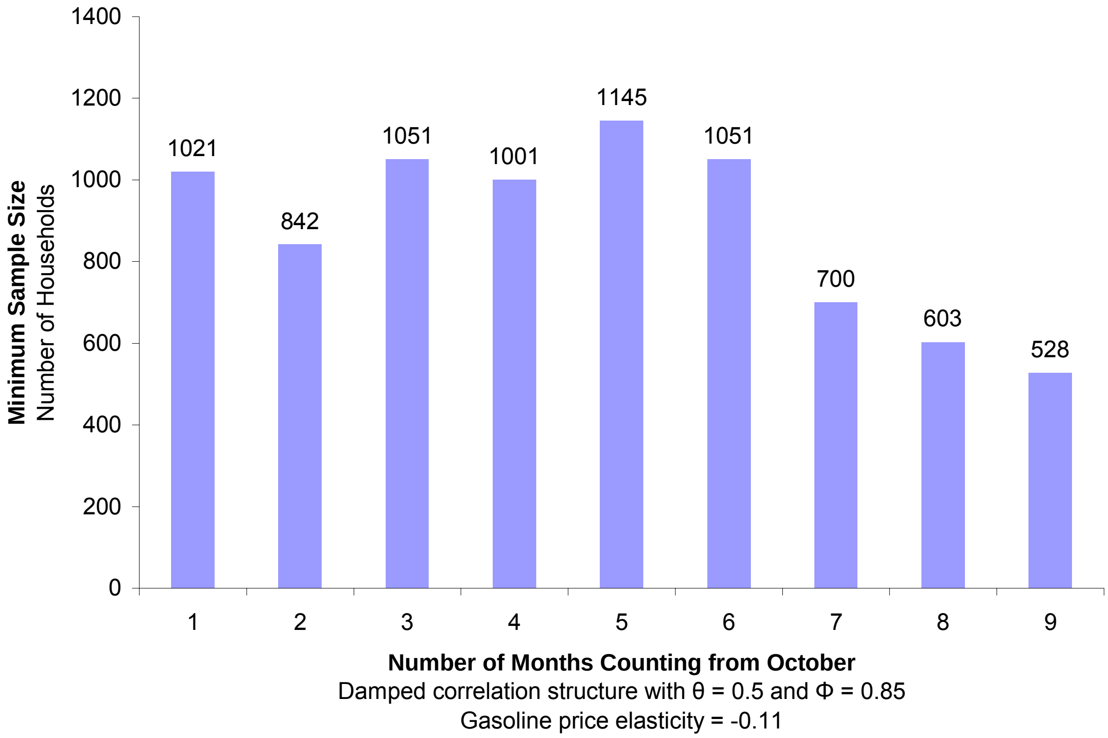

Figure 10.5 Sample Size Requirements as the Number of Survey Months in Each Period Increases with Fixed Starting Month in October

Figure 10.5 shows a general decreasing trend of minimum sample size as the number of months in each period increases, with the starting months fixed at October given the Commute Atlanta study set up. The decreasing trend of minimum trend is expected, but it is important to acknowledge that this graph also reflects the seasonal effects intrinsic to the study design in which the starting month was October. The graphs shows that the minimum sample size fluctuates as the number of months elapsed from October increased from one to five, before the sample size starts to steadily decrease as the number of months elapsed from October increased from six to nine. The non-decreasing sample size depicted by the first five columns in Figure 10.5 reflects the high year-to-year variability in household travel during the holiday season (November and December) and the potential impact of inclement weather in winter months (January and February). During these months, household are likely to change travel decisions driven by factors other than pricing incentives, rendering the estimation of sample size for detecting policy impacts unstable.

10.5. Sample Size versus Expected Policy Impact

In the previous section, the expected pricing impact on household VMT is set according to the assumption that price elasticity is -0.11, applied to all demographic groups equally. This is to assume the price elasticity with regard to VMT in the Commute Atlanta study is only half the amount Parry and Small (2005) recommended for gasoline price elasticity with regard to VMT, based on a comprehensive literature review. The rationale for this assumption was that, given the setup of the study, households did not incur any out-of-pocket costs with the mileage-based pricing incentives. When households do not experience a direct impact from VMT pricing, the absolute value of price elasticity could be smaller than the assumed price elasticity of gasoline demand. In different regions and study settings, in an area where residents have ample alternative transportation options, households may be more sensitive to pricing incentives. Therefore, a range of price elasticity is adopted for sample size estimation - a larger price elasticity of -0.30, the elasticity value of -0.22 recommended by Parry and Small (2005), and half the recommended elasticity value, -0.11. The results are presented in Figure 10.6. Because the damped correlation structure with θ=0.5 and ϕ=0.85 approximates the within-household association in Commute Atlanta data the best among the three scenarios presented in Figure 10.4, parameters θ=0.5 and ϕ=0.85 are chosen for this and subsequent analyses.

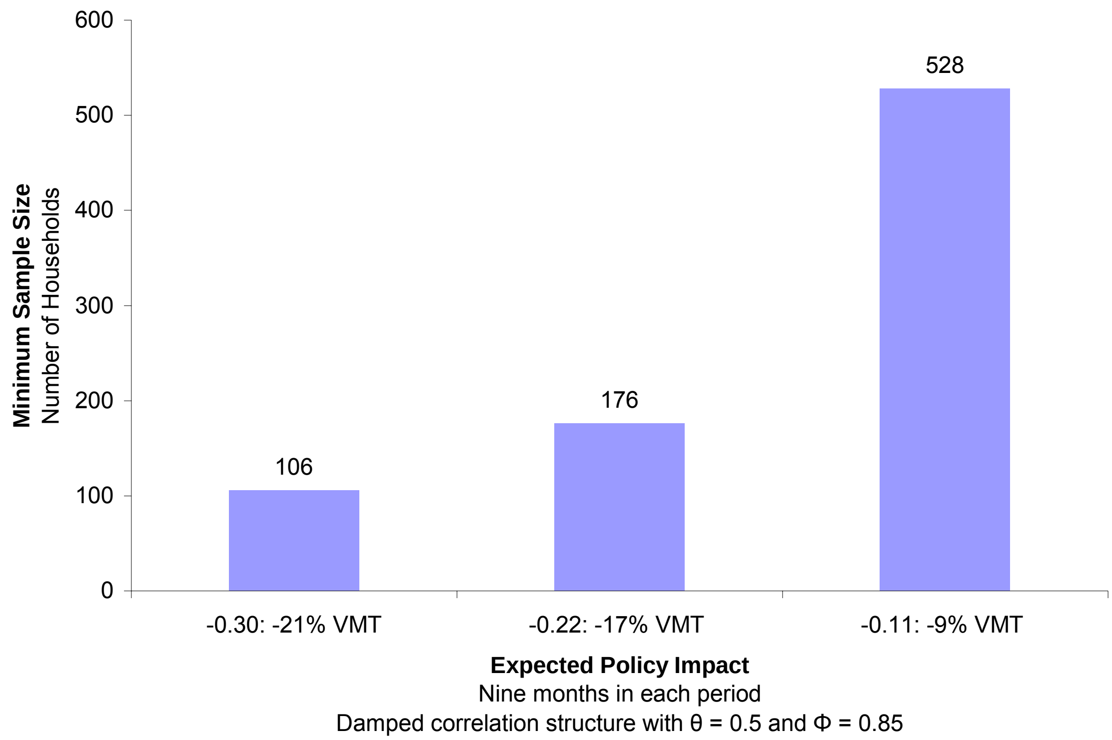

Figure 10.6 Sample Size versus Expected Policy Impact

Figure 10.6 shows that, not surprisingly, the larger the absolute value of price elasticity, the smaller the sample size is required. As the smallest VMT difference that needs to be detected decreases, the required sample size quickly increases. For example, under the design scheme where the baseline period and the pricing period each lasts for 9 months, the required sample size is 106 if the price elasticity is -0.30, or a 21% decrease in VMT is expected, whereas the required sample size is 176 if the price elasticity is -

0.22, or a 17% decrease in VMT is expected, but a 528-household sample is required if the price elasticity is -0.11, corresponding to a 9% decrease in VMT.

10.6. Missing Data

Missing data is a very common issue in practice. In GPS-based panel surveys, missing data mainly arises from equipment issues and participants dropout. Borrowing similar concepts from clinical trials, missing data arising from equipment issues can be viewed approximately as “staggered entry”, i.e. household travel data are accrued over a period of time as the equipment issues get resolved. Participating households may also be lost to follow-up due to dropout and various other reasons, sometimes referred to as panel attrition, e.g. (Brownstone*, et al.*, 1999). In the Commute Atlanta study, the rate of equipment issues is about 11%. At the beginning of the sample recruitment in 2003, there were 268 households in the sample. At the end of the pricing study in 2006, there were only 95 households with valid travel data and demographic information. This amounts to a loss-to-follow-up rate of 1.8% per month. Given such information, the $\{\pi_t\}$ as called for in Table 7.1 is outlined in Table 10.3. The procedures are described in detail below.

Table 10.3 Procedures of Deriving the $\{\pi_t\}$

Time Point in Pricing Period	Pr[$\Lambda \geq t$]			(d) $\pi_t = \Pr[\Lambda = t]$
	(a) Staggered entry	(b) Loss to follow-up	(c) Joint	
1	1.000	1.000	1.000	0.126
2	0.890	0.982	0.874	0.110
3	0.793	0.964	0.765	0.096
4	0.706	0.947	0.669	0.084
5	0.629	0.930	0.585	0.073
6	0.560	0.914	0.511	0.064
7	0.498	0.897	0.447	0.056
8	0.444	0.881	0.391	0.049
9	0.395	0.865	0.342	0.342

Define Λ to be a random variable that signifies the number of months observed in the pricing period. All households are assumed to provide baseline data. Column (a) provides Pr[$\Lambda \geq t$] that arises from the staggered entry considerations due to equipment issues. All households are expected to provide at least one month of data in the pricing period. Given the equipment issue rate of 11%, the probability of households providing 2 or more months of data is 0.890, the probability of providing at least 3 months of data is 0.890×0.890=0.793, and so on. Column (b) considers Pr[$\Lambda \geq t$] that arises from loss to follow-up. Similarly, all households are expected to provide at least one month of data in the pricing period. The probability of households providing 2 or more months of data is 0.982, the probability of households providing at least 3 months of data is 0.982×0.982=0.964, and so on. The joint probability of providing at least t months of data, as shown in column (c), is the product of columns (a) and (b). The probability of providing the first t observations only is computed as $\pi_t = \Pr[\Lambda \geq t] - \Pr[\Lambda \geq t+1]$, and is given by column (d).

With the same design specifications of the Commute Atlanta study, assuming price elasticity of -0.11 and correlation parameters of $\theta = 0.5$ and $\phi = 0.85$, the consideration of missing data indicates that 727 households need to be recruited per group. This amounts to a 38% increase in required sample size compared to the design with the consideration of missing data over a 21-month period[1].

10.7. Summary of Sample Size Analysis for Before-and-After Studies

The analysis in this chapter has several implications on sample size planning for before-and-after studies.

1. The required sample size assuming the same design as the Commute Atlanta study with expected price elasticity of -0.11 is 528 households, much larger than the originally planned sample size.
2. A larger sample size than 528 would be required if the number of months in each period were smaller than 9 months, with fixed starting month of October.
3. The consideration of missing data will increase the required sample size by about 38% compared to the sample size requirement without such consideration, based on the rate of missing data in the Commute Atlanta study.

The analyses in this chapter did not stratify the population into distinct demographic groups. The small number of households in the Commute Atlanta sample does not provide enough data for analyses on individual demographic groups. Stratification may increase or decrease the required sample size, depending on the

[1] From October 2004 to June 2006.

relative magnitude of between-stratum and within-stratum variability. Proper stratification where between-stratum variability is much more significant than within-stratum variability can lead to overall smaller samples. Further research is warranted when data from similar before-and-after studies with larger sample sizes become available.

The analyses in this chapter also assumed that the pricing incentives have the same amount of impact on household travel behavior across all demographic groups. This assumption will likely need to be relaxed in future research, because some households are likely to be more sensitive to a policy measure than others, as seen in the household responses to the pricing incentives in the Commute Atlanta study (Xu*, et al.*, 2009b). This chapter demonstrated how the required sample size would increase as the expected magnitude of policy impact decreases, implying that the demographic group in which a smaller policy impact is expected would require a larger sample size. This notion will need to be confirmed with empirical evidence in the future.

CHAPTER 11

CONCLUSIONS

This chapter discusses the implications and contributions of this dissertation. The first section summarizes the research findings. The second section offers some suggestions for future research directions.

11.1. Summary of Research Findings

The purpose of this dissertation was to develop a framework and methods to estimate sample size requirements for global positioning system (GPS) based panel travel surveys. The goals of the research were threefold:

1. Evaluate the trade-offs between sample size and length of study for obtaining reliable estimates of the means of travel behavior variables
2. Relate sample size requirements to regression analysis suited for longitudinal travel data
3. Estimate minimum sample size for before-and-after studies in the context of transportation policy evaluation

Given these goals, the following technical tasks had to be accomplished:

- Differentiate the between-household (cross-sectional) and within-household (longitudinal) information in household travel behavior
- Characterize the variability of travel behavior associated with natural temporal rhythms and demographic changes

- Adapt generalized estimating equation (GEE) procedures to the analysis of GPS-based panel data for travel behavior studies
- Explore the distributional properties and correlation structures of key travel behavior variables
- Apply GEE procedures to regression analyses and before-and-after studies of travel behavior variables as the basis for sample size estimation for such analyses

This dissertation utilizes the data collected in the Commute Atlanta study to accomplish the aforementioned technical tasks, and hence, the overall goals. Intra-regional travel behavior and long-distance travel behavior were examined separately. The intra-regional travel dataset includes 95 households, the monitoring period spanning from October 2004 to June 2005, and from October 2005 to June 2006, providing 250,580 intra-regional trips totaling 1.62 million VMT. The long-distance travel dataset covers the period from January 2004 to June 2006 for 94 households, providing 0.48 million VMT from 1,006 long-distance tours. The uniqueness of GPS-based panel data is assessed through a literature review, exploratory data analysis using the bootstrap method, and a detailed study on household demographic variability using a case study approach.

Based upon a literature review of current trends in travel demand forecasting to adopt activity-based models and existing GPS-based travel surveys, the GPS technology enables extended monitoring periods for travel surveys without increasing respondent burden. This capability can improve the accuracy of activity-based models and associated transportation policy evaluation, because the longitudinal information provided by panel surveys is the only means to potentially establish causal relationships.

However, the current travel demand models have not taken full advantage of the strengths of GPS-based panel surveys, possibly due to the higher cost of such surveys. Therefore, a procedure to estimate sample size for GPS-based panel surveys for travel behavior studies is needed to ensure the cost-effectiveness of such studies.

The exploratory data analysis in this dissertation employed bootstrap techniques, and revealed the sources and magnitude of variability intrinsic to GPS-based panel data that arises from temporal factors and household demographic characteristics. The exploratory data analysis provided graphic presentations of the between-household and within-household variability.

The travel variability associated with household demographic changes was examined using a case study approach. Experiences from the Commute Atlanta study showed that more than 70% percent of the households participating in the pricing study underwent one ore more major demographic changes over the study period between October 2004 and June 2006. The unstable household demographic characteristics will need to be controlled for in transportation policy evaluation.

With regard to obtaining reliable estimates of population averages for travel behavior variables, intra-regional travel and long-distance travel require very different survey schemes. Applying re-sampling techniques, this dissertation found that a two-day survey with about 250 households and a 20-day survey with about 170 households will provide 10% relative precision in the estimate for average weekday intra-regional VMT. Considering the trade-offs between the number of households needed and the length of the survey, practitioners will have to compare the incremental costs of adding a household versus extending the length of the survey to achieve optimal cost-effectiveness. With the advancement of GPS technologies coupled with web-based survey applications, extending the length of a survey would incur marginal cost, even though a GPS-based survey is generally more expensive than a paper diary survey per household. Given the savings on number of households, a GPS-based survey with extended monitoring period (e.g. a month) may be more cost-effective than a paper diary survey with a larger number of households. In addition, GPS-based surveys can provide more accurate information on household travel behavior, as presented in the literature review of this dissertation. For long-distance travel, often considered as rare events, a much longer monitoring period and a much larger sample size than those available from the Commute Atlanta sample are needed to achieve a 10% relative precision in monthly number of long-distance tours. The monthly number of long-distance tours for each household approximates the Poisson distribution, implying that a longer monitoring period is needed for households that undertake long-distance travel less frequently. For example, for households that only travel long-distance once every 30 months (0.03 long-distance tours per month), a survey will need to monitor 480 households for two (2) years to achieve 10% relative precision, whereas for households that on average travel long-distance twice every month, a survey will only need to monitor seven (7) households for two (2) years to achieve the same 10% relative precision.

The effort to relate sample size estimation to regression analysis and before-and-after studies requires the generalized estimating equation (GEE) procedure. The GEE procedure is an extension of the generalized linear models (GLMs) to correlated data. To fit the data to GLMs using the GEE procedure, the distributions and correlation structures of intra-regional and long-distance travel variables were examined.

Daily intra-regional VMT and monthly long-distance VMT were found to approximate Tweedie distributions. The unique feature of Tweedie distributions is that these distributions can model non-negative continuous data with exact zeros. In transportation applications, Tweedie distributions can be considered as compound Poisson distributions, where the outcome variables (e.g. daily intra-regional VMT and monthly long-distance VMT) are the Poisson sum of gamma variables. Under the assumption of Tweedie distributions, the monthly number of long-distance tours should be Poisson distributed. The formal statistical analyses in this dissertation supported the Poisson assumption of the monthly number of long-distance tours.

Relating sample size estimation to regression analysis, it is important to adopt a suitable model form for longitudinal data differentiating the cross-sectional and longitudinal effects. It is possible to examine how households would change their travel behavior as their demographic characteristics change using a small number of households combined with an extended survey period. It would require a large number of households and a long survey (preferably more than a year) to detect *both* the cross-sectional and longitudinal effects.

This dissertation adapted the work of Rochon (1998) to estimate the minimum sample size for before-and-after studies in the context of transportation policy evaluation. The results showed that the minimum sample size is sensitive to extreme values in the data. In general, the required sample size decreases as the number of months in the "before" and "after" periods increases, even though seasonal variability could interfere with the decreasing trend of sample size when the number of months is small, e.g. less than six (6) months. This decreasing trend of minimum sample size as the length of study increases implies that a panel survey consisting of two two-day waves one year apart may require a very large number of households in the sample, and that the minimum sample size may be very sensitive to temporal variability. The results also showed that the minimum sample size increases as the degree of within-household correlation in the data decreases. Further, the minimum sample size was found to be negatively correlated with the expected size of the policy impact. Assuming a “damped” correlation structure with $\theta = 0.5$ and $\phi = 0.85$, the Commute Atlanta study would have required 528 households, expecting a 9% reduction in weekday intra-regional VMT. Finally, the consideration of missing data due to equipment issues and sample attrition would increase the required sample size by about 37% percent, everything else held equal, based on the rate of missing data in the Commute Atlanta data under the design scheme where gasoline prices are controlled for without the control group. These findings are generally in agreement with sample size guidelines developed in biomedical research. The caveat is that the number of households estimated through these sample size analyses are confounded by the representativeness of the Commute Atlanta data, and may have limited applicability to other studies. Nevertheless, these results indicate that a panel survey with a sample size of a few hundred households is needed for policy evaluation. The results also depict how the minimum sample size would vary with the assumed correlation structure and expected policy impact, presenting the required sample size if assumptions for other studies differ from the experiences with the Commute Atlanta data.

To summarize, this dissertation demonstrated that implementing a GPS-based panel travel survey would require special consideration of the nature of longitudinal data at both the design and analysis stages. The variability of longitudinal travel data arise from various sources, so it is important to partition the variability into within-household and between-household portions for the regression analysis in travel demand forecasting and before-and-after studies in the context of transportation policy evaluation. Given the unstable household demographic characteristics described in this dissertation, and other factors such as fluctuations in gasoline prices that may influence household travel demand, it is important that exogenous and endogenous confounding effects be adequately acknowledged and adjusted through regression modeling, and further appropriately reflected for sample size estimation in before-and-after studies.

This dissertation has emphasized that the main goal of a panel survey would be to characterize within-household travel behavior changes in response to policy measures, economic trends or demographic shifts over time. This strength of GPS-based panel surveys implies that such surveys would not replace conventional one-day or two-day cross-sectional surveys, but to complement the cross-sectional surveys so that dynamic changes can be accurately estimated for future trends. The purpose of this dissertation was not to prescribe a fixed sample size for all GPS-based panel surveys, but to establish a framework and methods so that different regions and agencies could adopt these methods to estimate the minimum sample size given the data characteristics and study objectives that suit their situations.

11.2. Limitations and Future Research

The small number of households in the Commute Atlanta study was the most important limiting factor in this dissertation research. The results from this research may have limited transferability to other metropolitan areas. The methodology and results from this study will need to be validated and updated with data from other regions, when such data become available.

Due to the small number of households in the Commute Atlanta study, it was difficult to extract useful information for any demographic group. The largest number of households in a demographic stratum in which household income, household size and total number of vehicles owned were controlled for was 18. Therefore, most analyses in this dissertation were not able to consider refined sample stratification schemes. For example, the interaction between temporal factors and demographic characteristics were not analyzed. Additionally, in regression analysis for intra-regional VMT, the income variable was not included in the model because the incorporation of household income would require specifying market segments. Finally, the impact of pricing policy was assumed to be identical across the entire sample in the analyses for before-and-after studies, when households in different demographic groups would likely respond to pricing incentives with varying degrees of sensitivity. These aspects and their impacts on sample size requirements will need to be evaluated with information from a much larger sample.

Compared to cross-sectional surveys, the design of a panel survey is more complex in that there are practically endless compositions of a panel, specified by the non-inclusive list of considerations below:

- The length of the entire study
- The number of repeated measurements, or waves
- The length of the continuous monitoring period in each wave
- The placement of each wave in time

- The length of intervals between waves
- Whether the waves should be equally spaced

Future research can perform refined analyses to evaluate the impact of the aforementioned design considerations on sample size requirement. In addition to following statistical guidelines, researchers and practitioners will also need to identify the most cost-effective design based on incremental costs per observation and per household.

APPENDIX A

BETWEEN-HOUSEHOLD AND WITHIN-HOUSEHOLD VARIABILITY

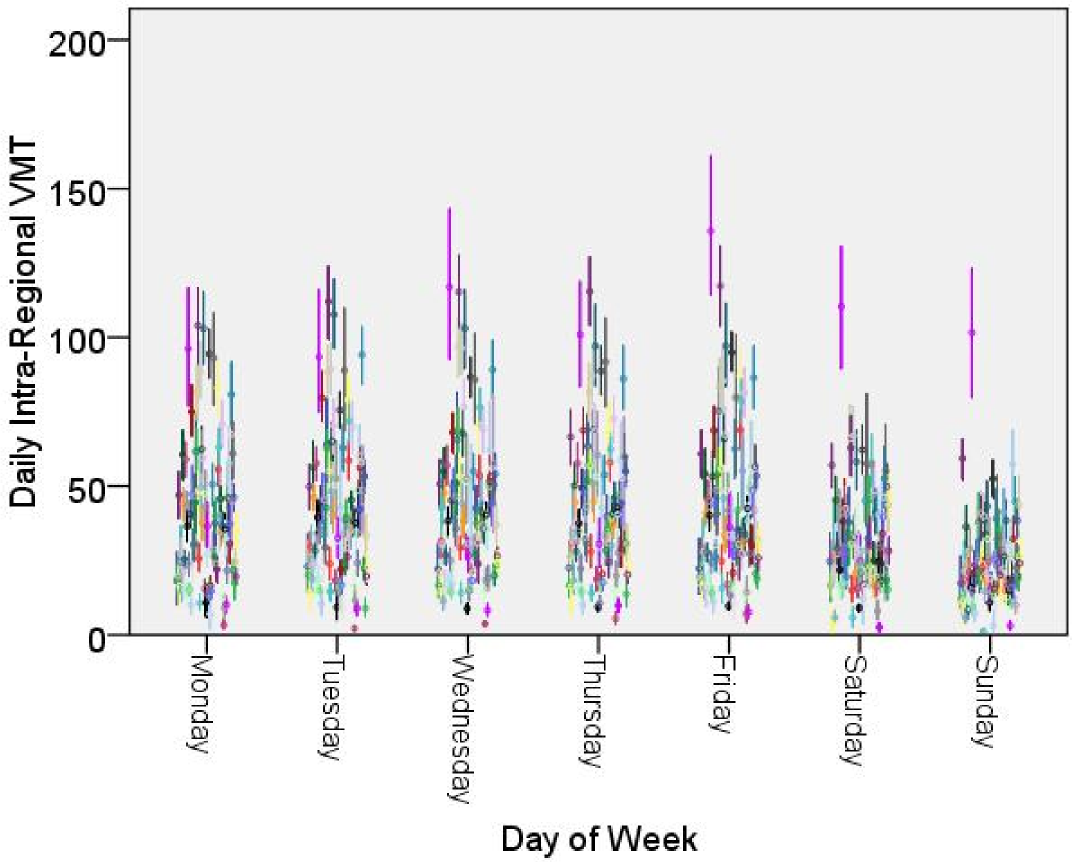

Figure A.1 Between-Household and Within-Household Day-of-Week Variability with Regard to Daily Intra-Regional VMT.
Each series of error bars represents an individual household.

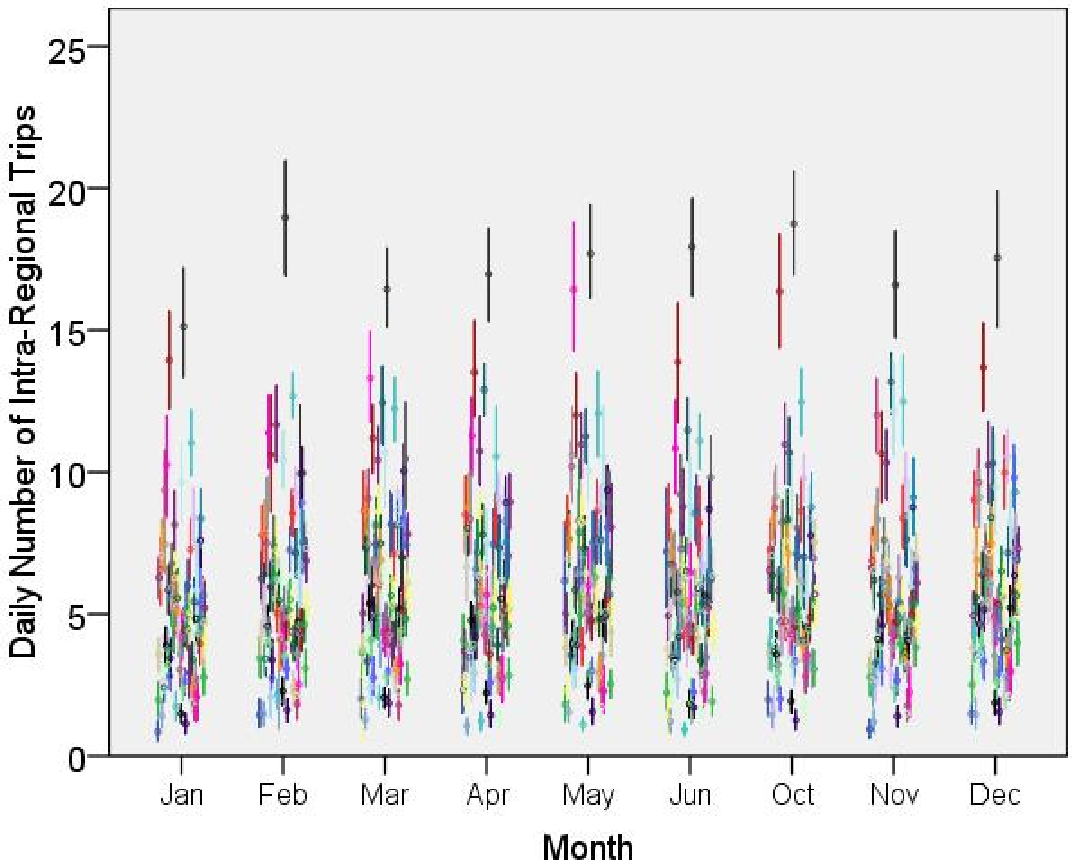

Figure A.2 Between-Household and Within-Household Seasonal Variability with Regard to Number of Intra-Regional Trips per Day
Each series of error bars represents an individual household.

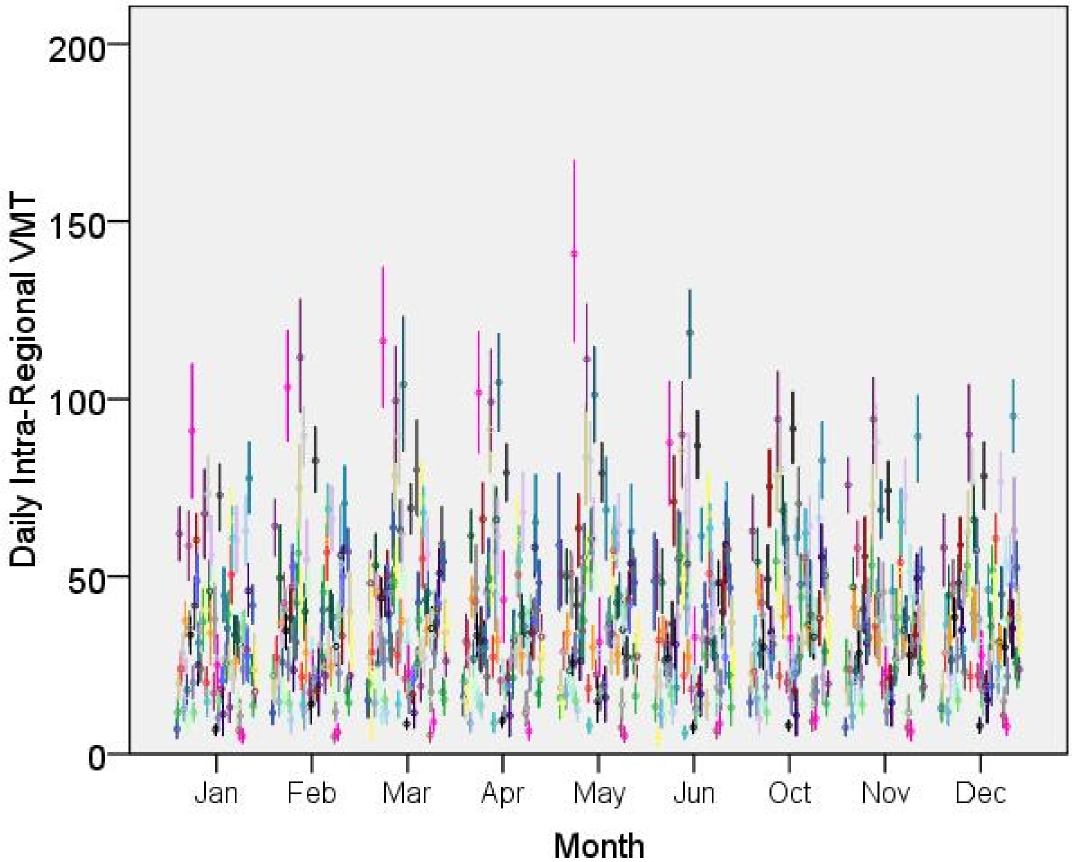

Figure A.3 Between-Household and Within-Household Seasonal Variability with Regard to Daily Intra-Regional VMT
Each series of error bars represents an individual household.

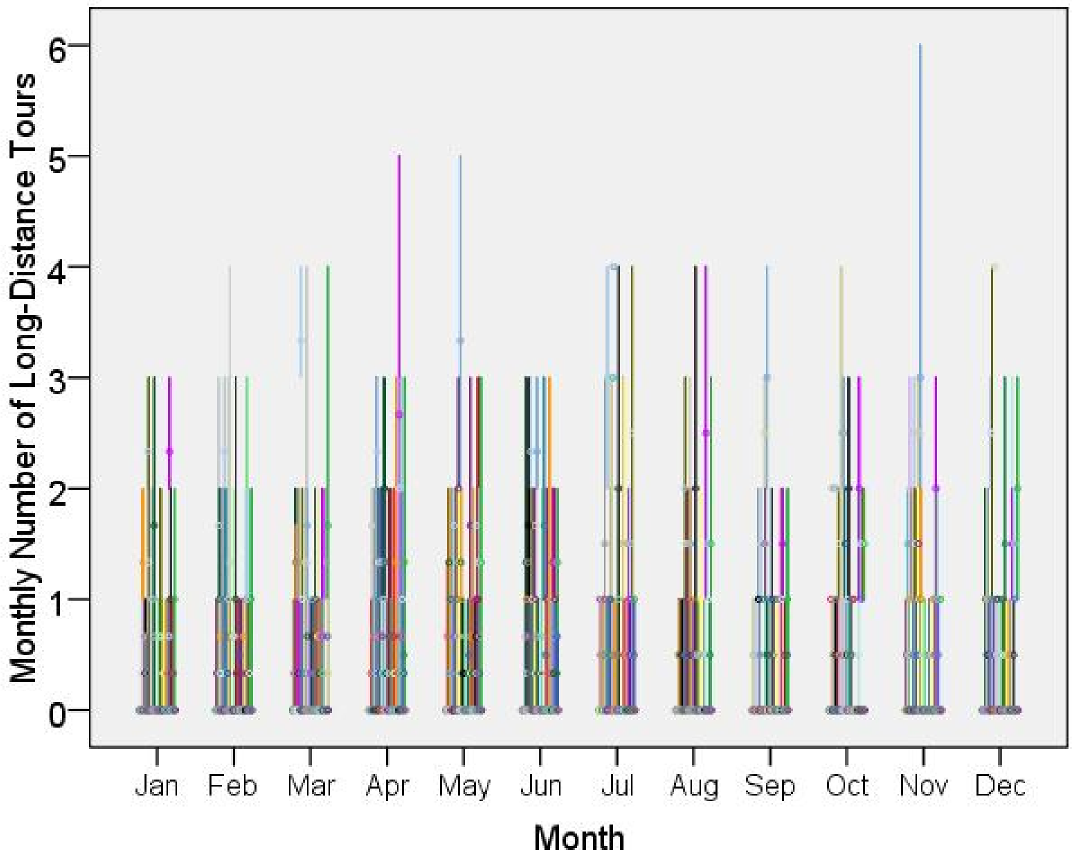

Figure A.4 Between-Household and Within-Household Seasonal Variability with regard to Number of Long-Distance Tours per Month.
Each series of error bars represents an individual household.

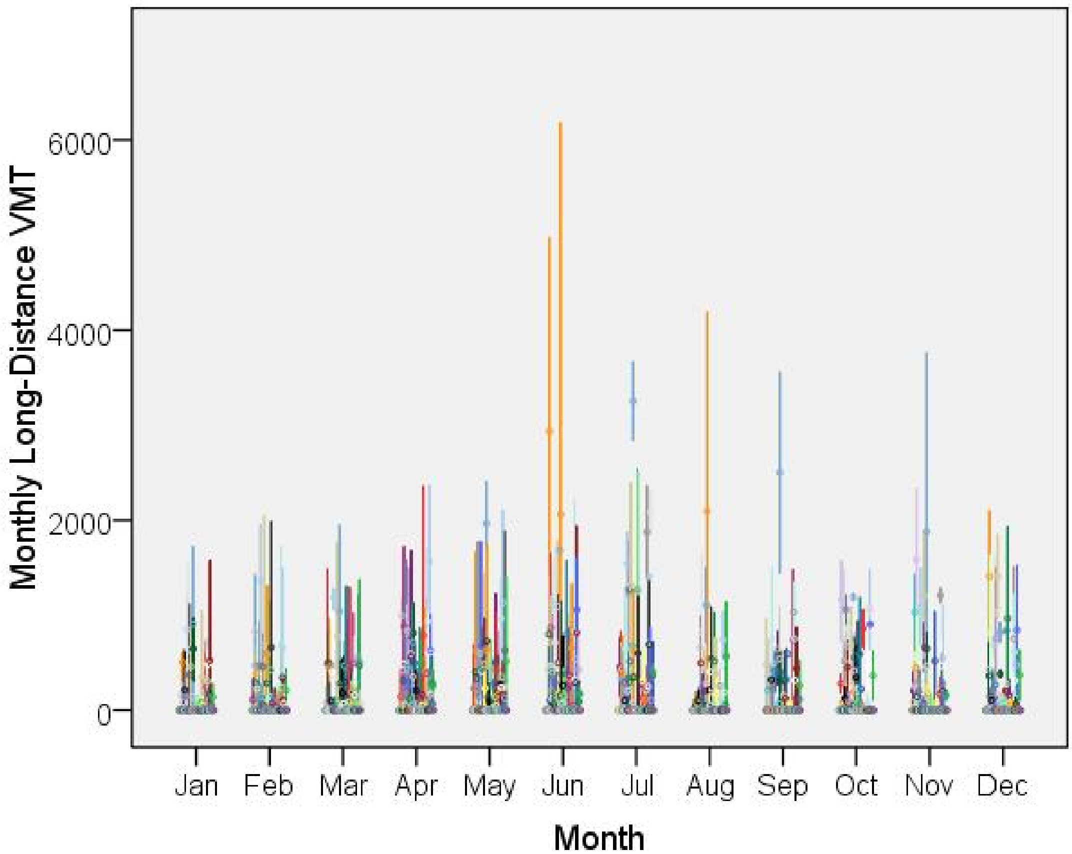

Figure A.5 Between-Household and Within-Household Seasonal Variability with Regard to Monthly Long-Distance Tours VMT.
Each series of error bars represents an individual household.

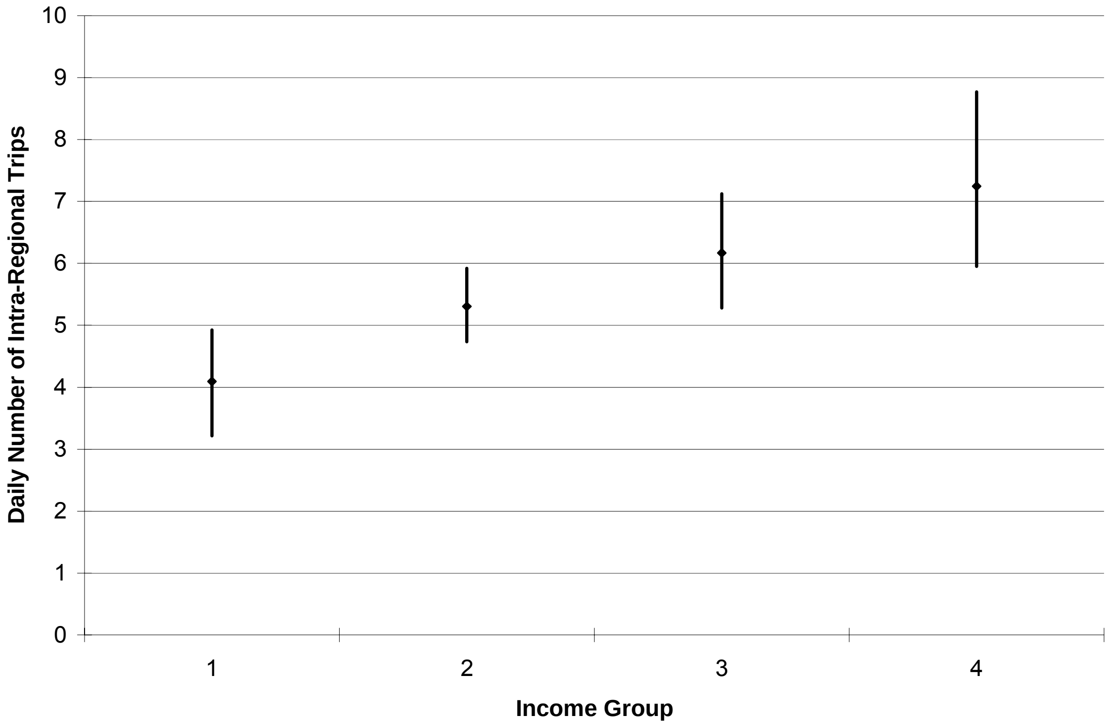

Figure A.6 Association between Daily Intra-Regional Number of Trips and Household Income
Error bars represent 95% bootstrap CI of *household means* of number of intra-regional trips per day.

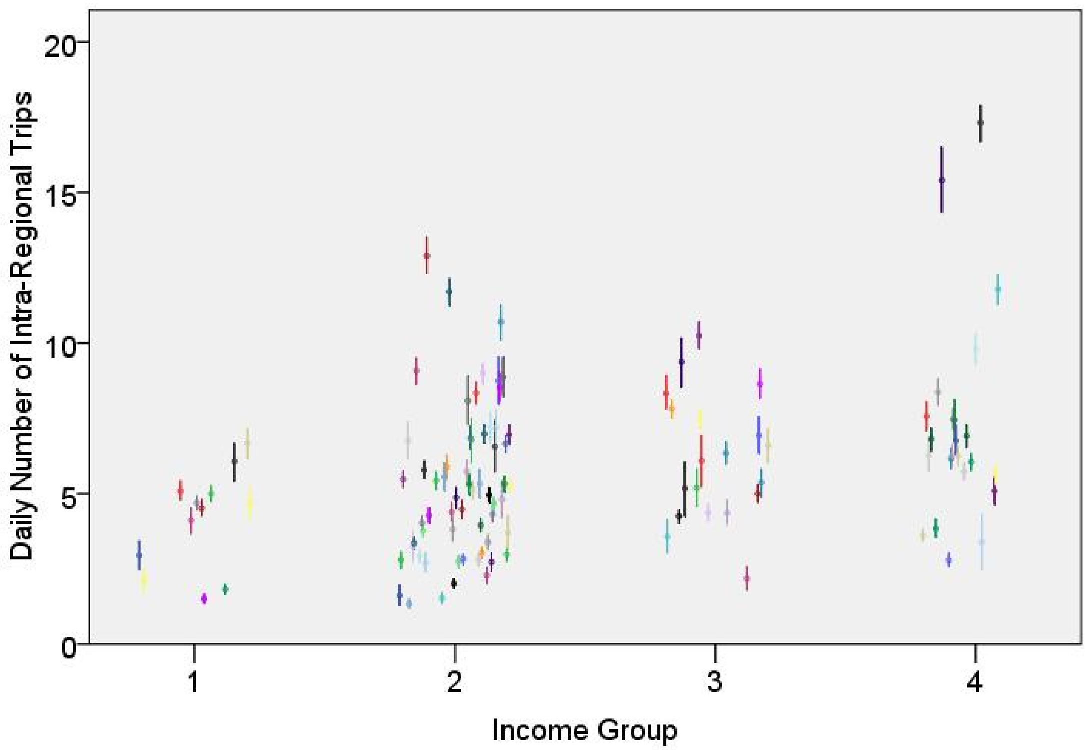

Figure A.7 Between-Household and Within-Household Variability in Number of Intra-Regional Trips per Day by Income Group
Each error bar represents a household.

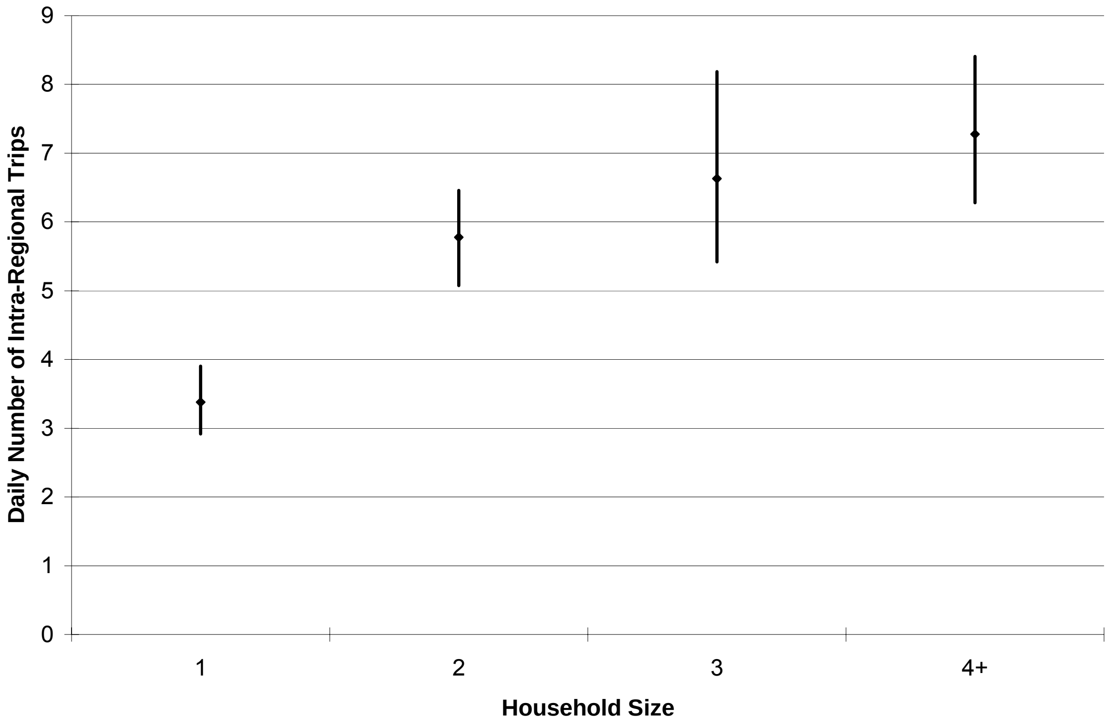

Figure A.8 Association between Daily Intra-Regional Number of Trips and Household Size
Error bars represent 95% bootstrap CI of *household means* of number of intra-regional trips per day.

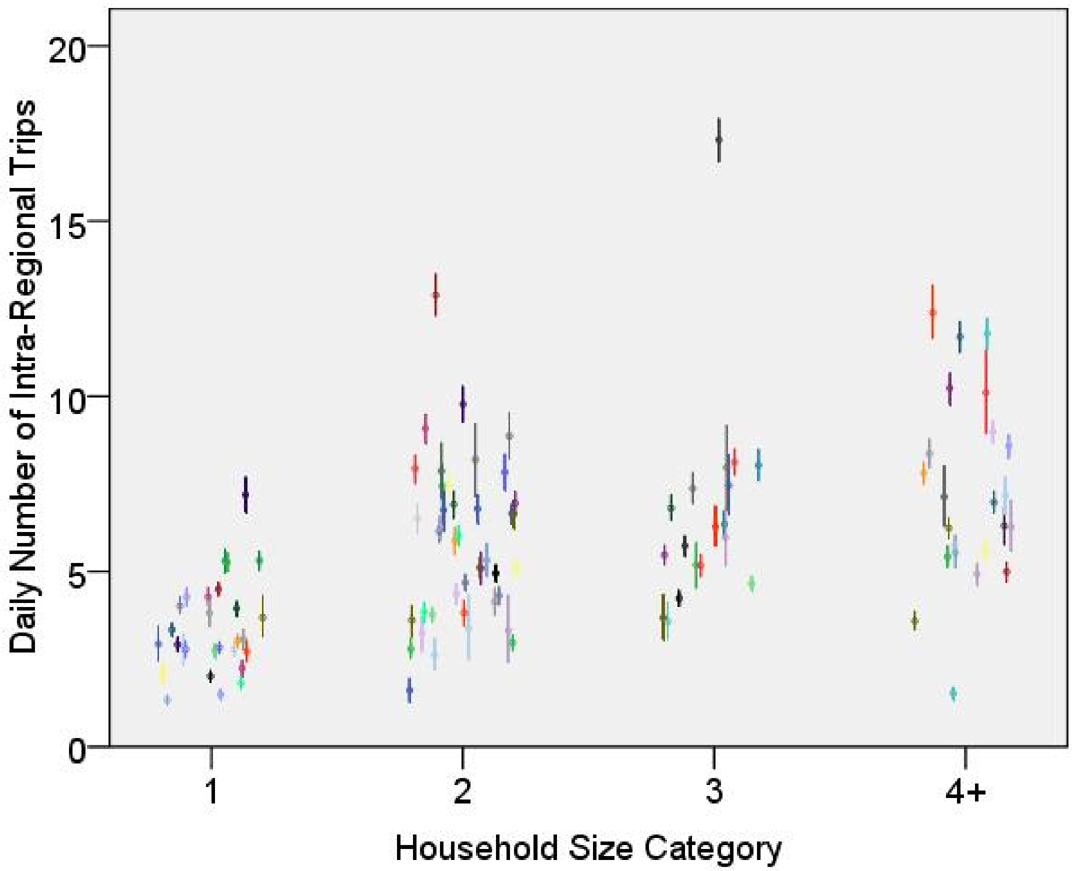

Figure A.9 Between-Household and Within-Household Variability in Number of Intra-Regional Trips per Day by Household Size
Each error bar represents a household.

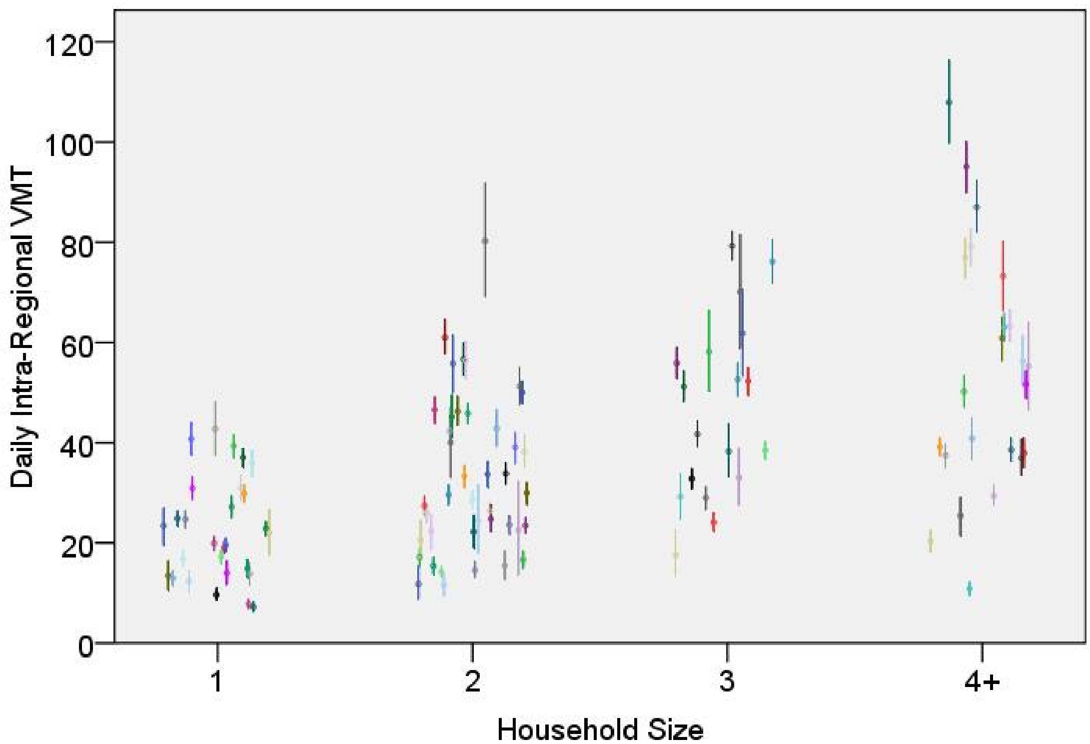

Figure A.10 Between-Household and Within-Household Variability in Daily Intra-Regional VMT by Household Size
Each error bar represents a household.

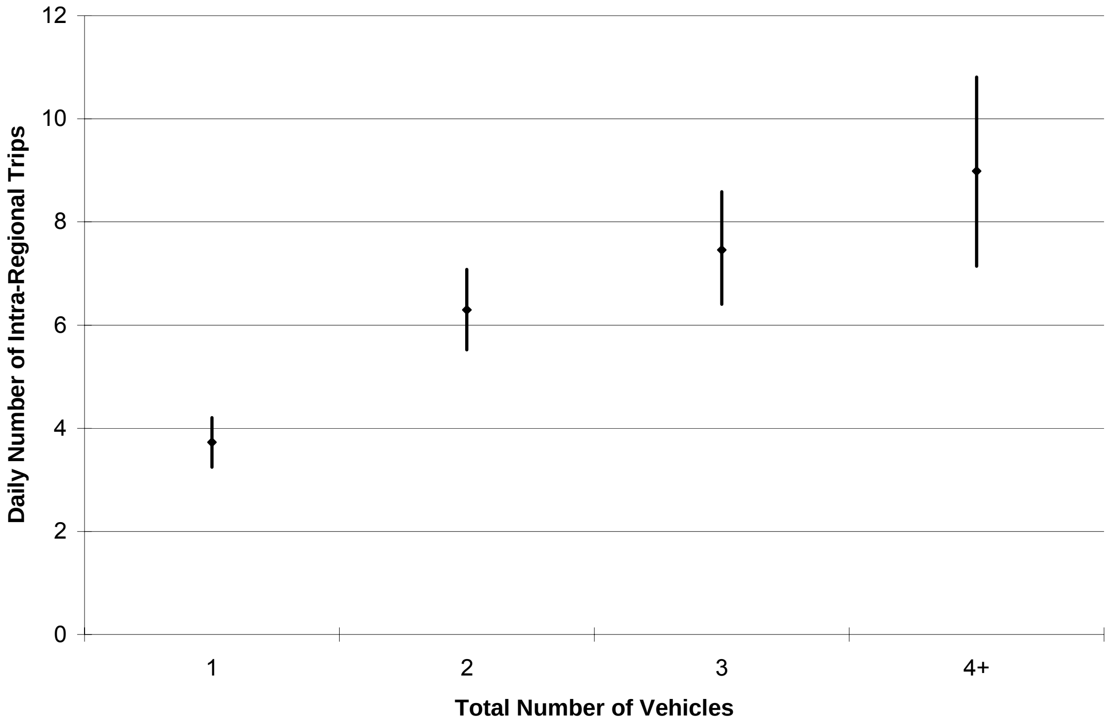

Figure A.11 Association between Daily Intra-Regional Number of Trips and Total Number of Vehicles Owned
Error bars represent 95% bootstrap CI of *household means* of number of intra-regional trips per day.

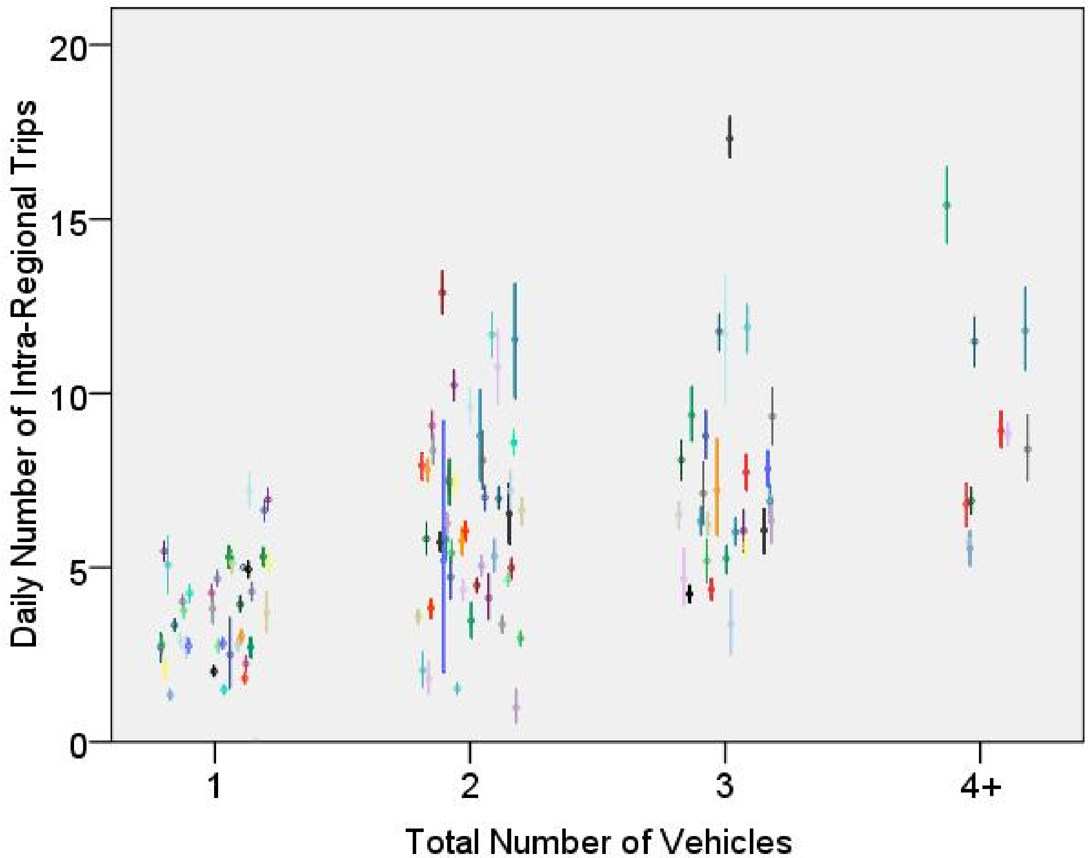

Figure A.12 Between-Household and Within-Household Variability in Number of Intra-Regional Trips per Day by Total Number of Vehicles Owned
Each error bar represents a household.

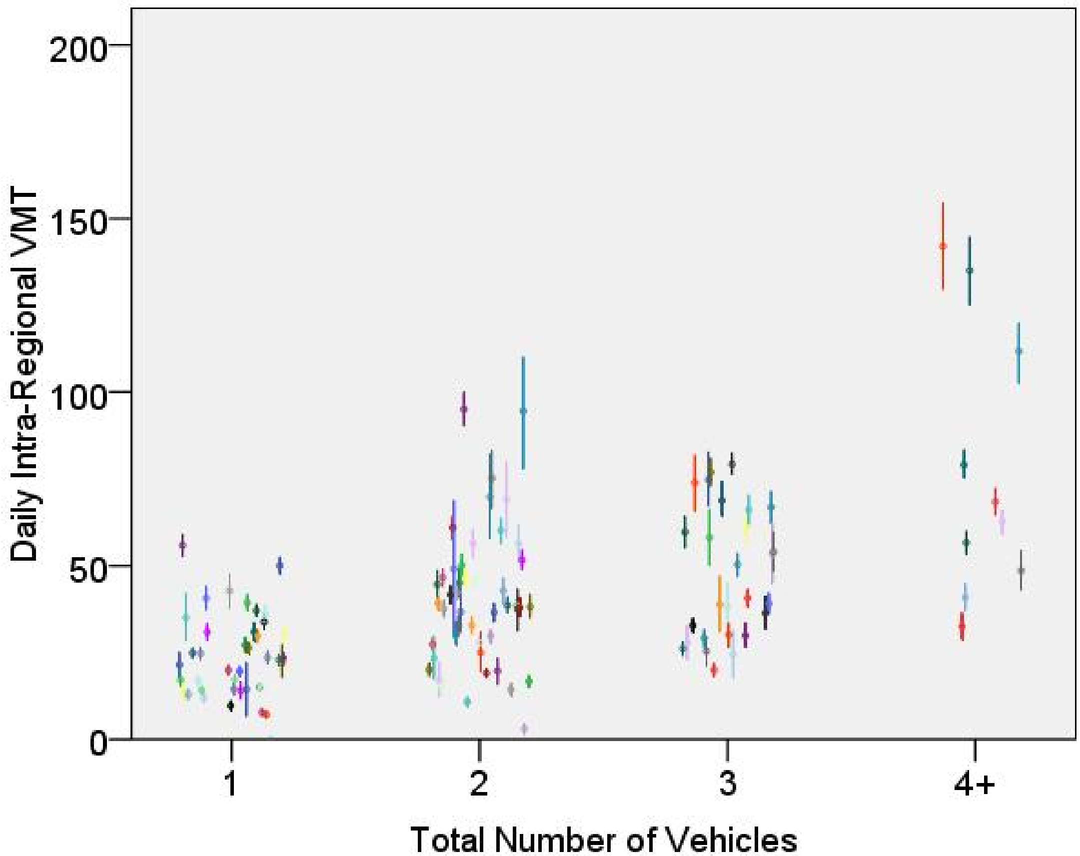

Figure A.13 Between-Household and Within-Household Variability in Daily Intra-Regional VMT by Total Number of Vehicles Owned
Each error bar represents a household.

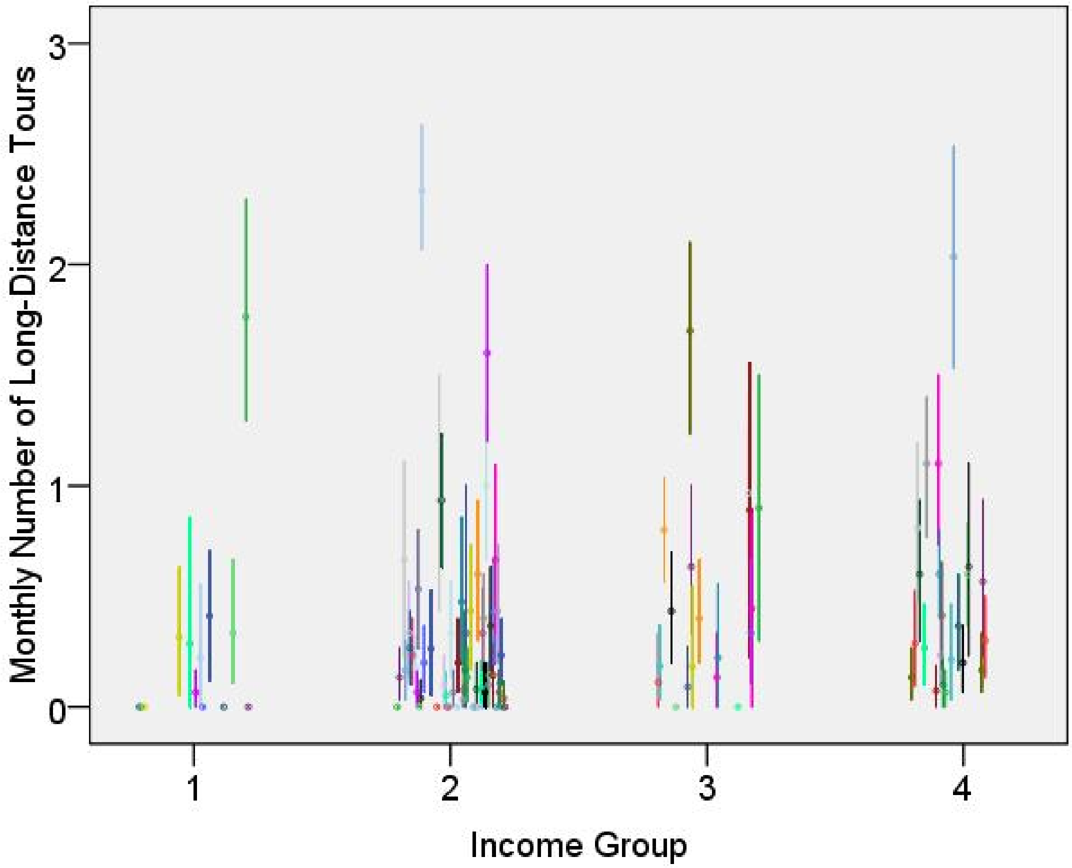

Figure A.14 Between-Household and Within-Household Variability in Number of Long-Distance Tours per Month by Income Group
Each error bar represents a household.

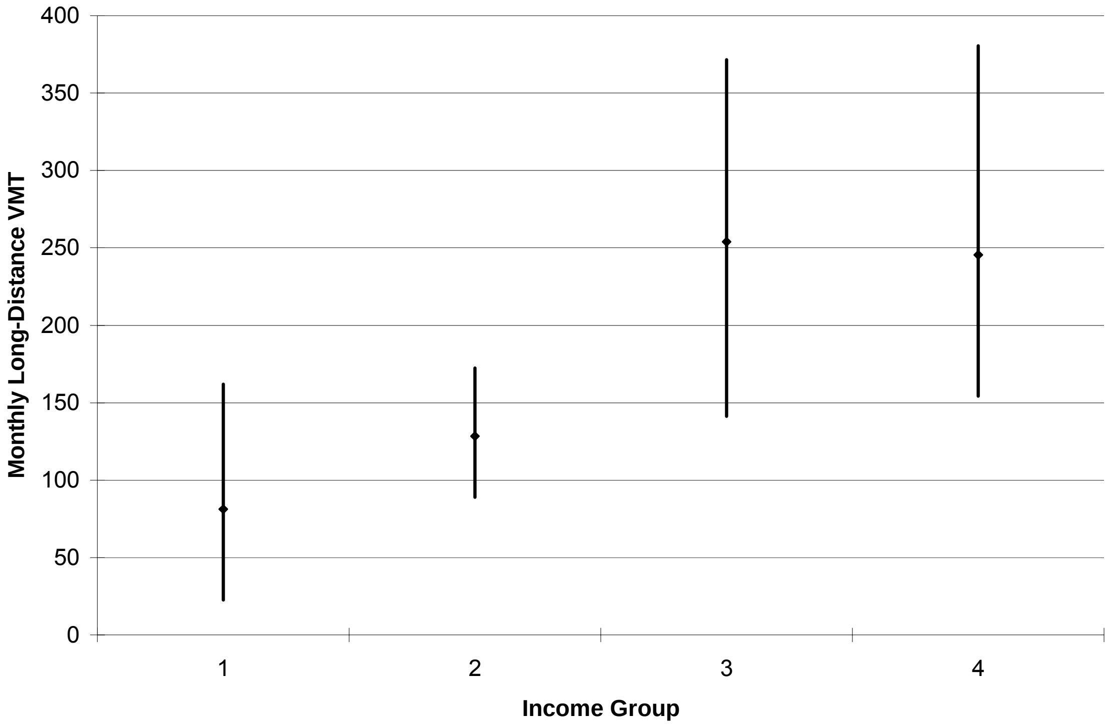

Figure A.15 Association between Monthly Long-Distance VMT and Income
Error bars represent 95% bootstrap CI of *household means* of monthly long-distance VMT.

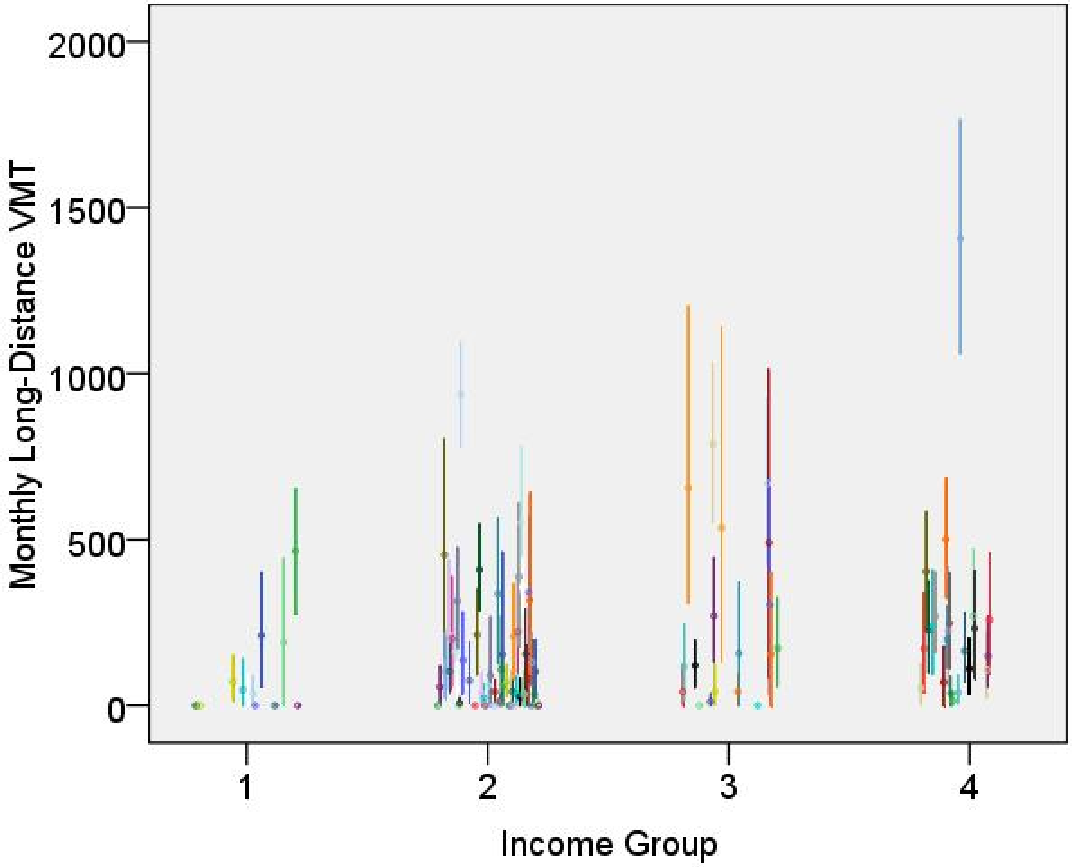

Figure A.16 Between-Household and Within-Household Variability in Monthly Long-Distance VMT by Income Group
Each error bar represents a household.

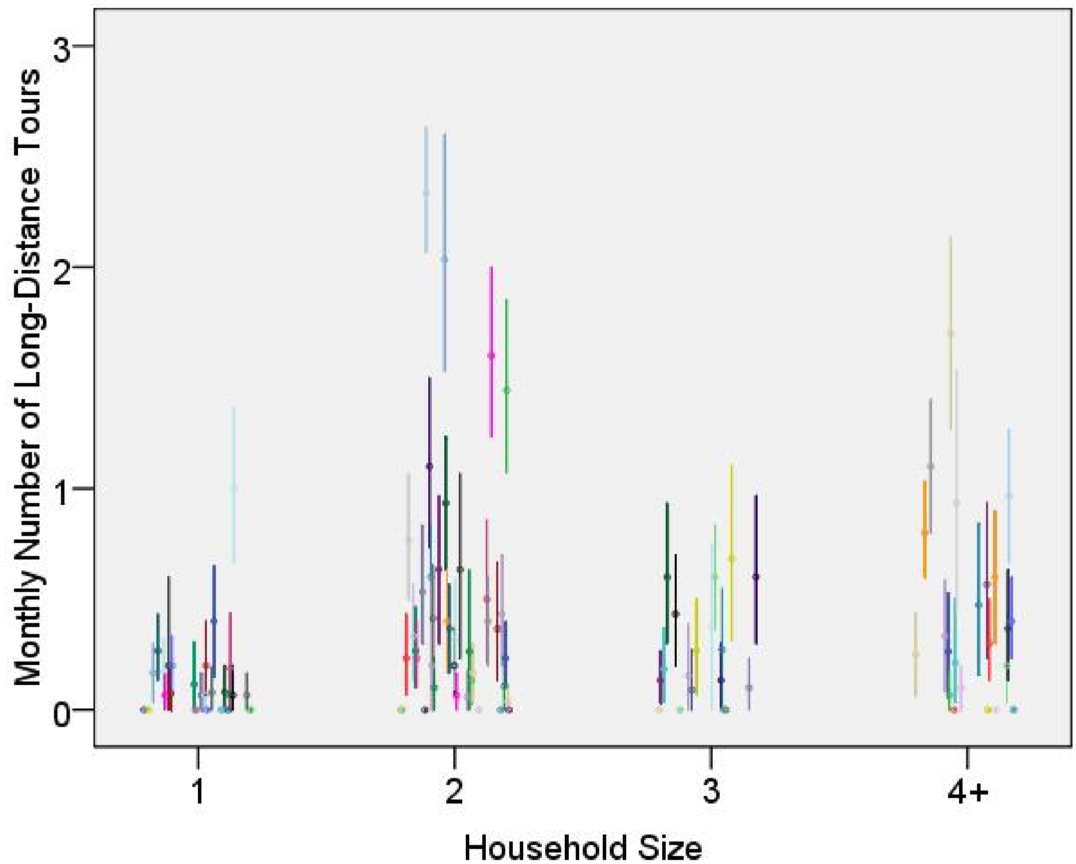

Figure A.17 Between-Household and Within-Household Variability in Number of Long-Distance Tours per Month by Household Size
Each error bar represents a household

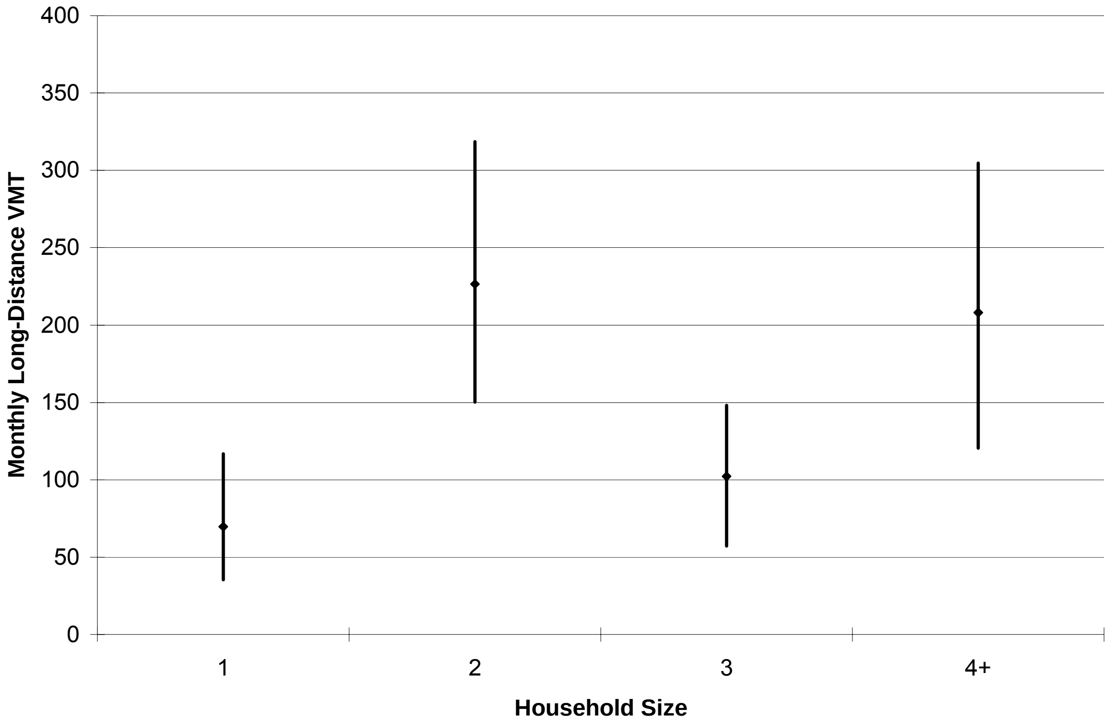

Figure A.18 Association between Monthly Long-Distance VMT and Household Size
Error bars represent 95% bootstrap CI of *household means* of monthly long-distance VMT.

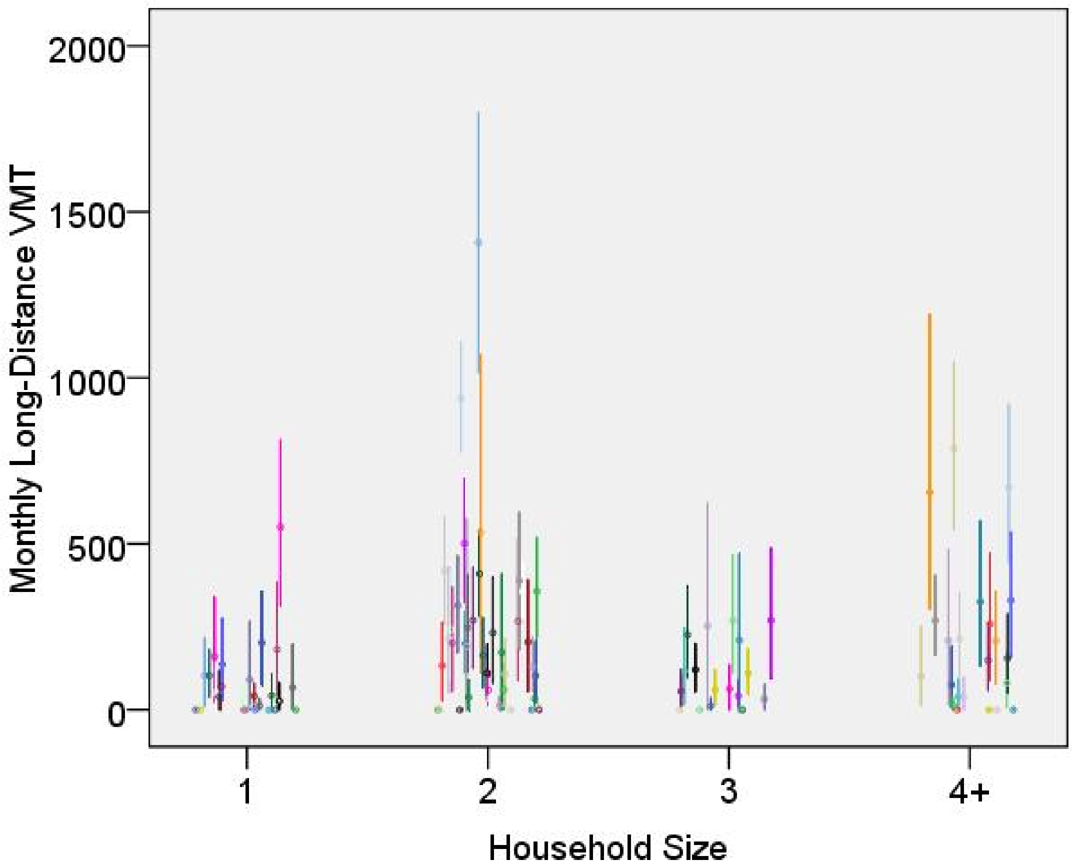

Figure A.19 Between-Household and Within-Household Variability in Monthly Long-Distance VMT by Household Size
Each error bar represents a household.

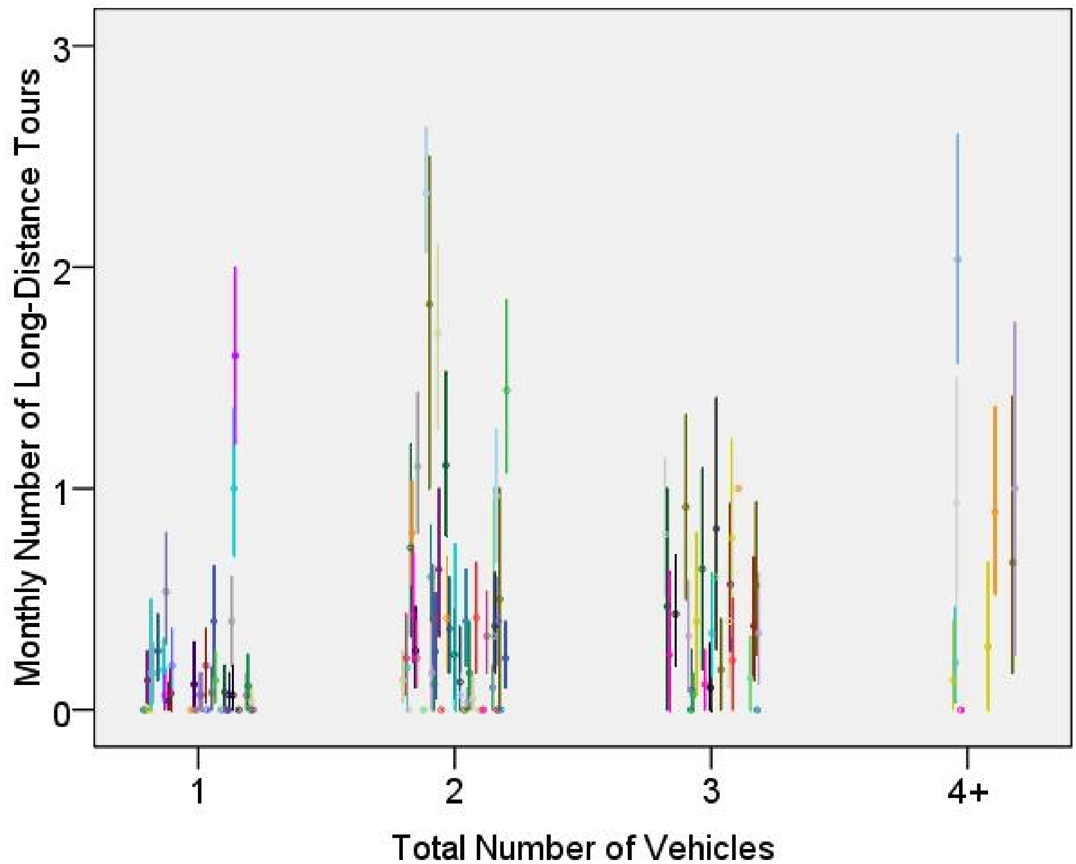

Figure A.20 Between-Household and Within-Household Variability in Number of Long-Distance Tours per Month by Total Number of Vehicles Owned
Each error bar represents a household.

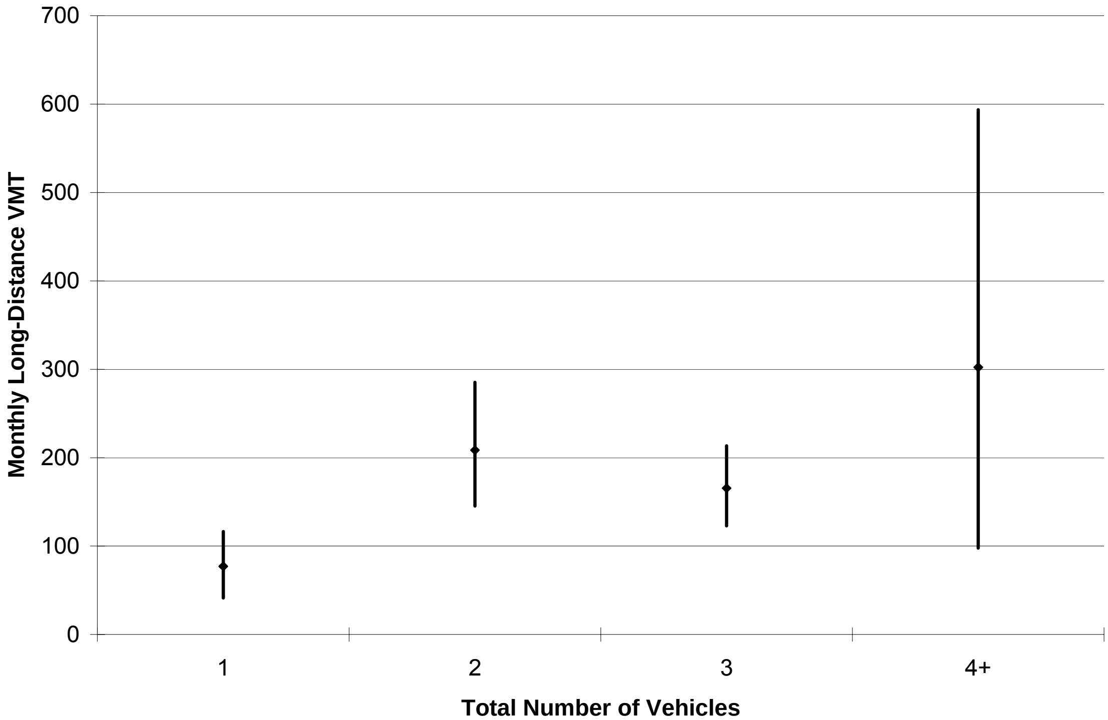

Figure A.21 Association between Monthly Long-Distance VMT and Total Number of Vehicles Owned
Error bars represent 95% bootstrap CI of *household means* of monthly long-distance VMT.

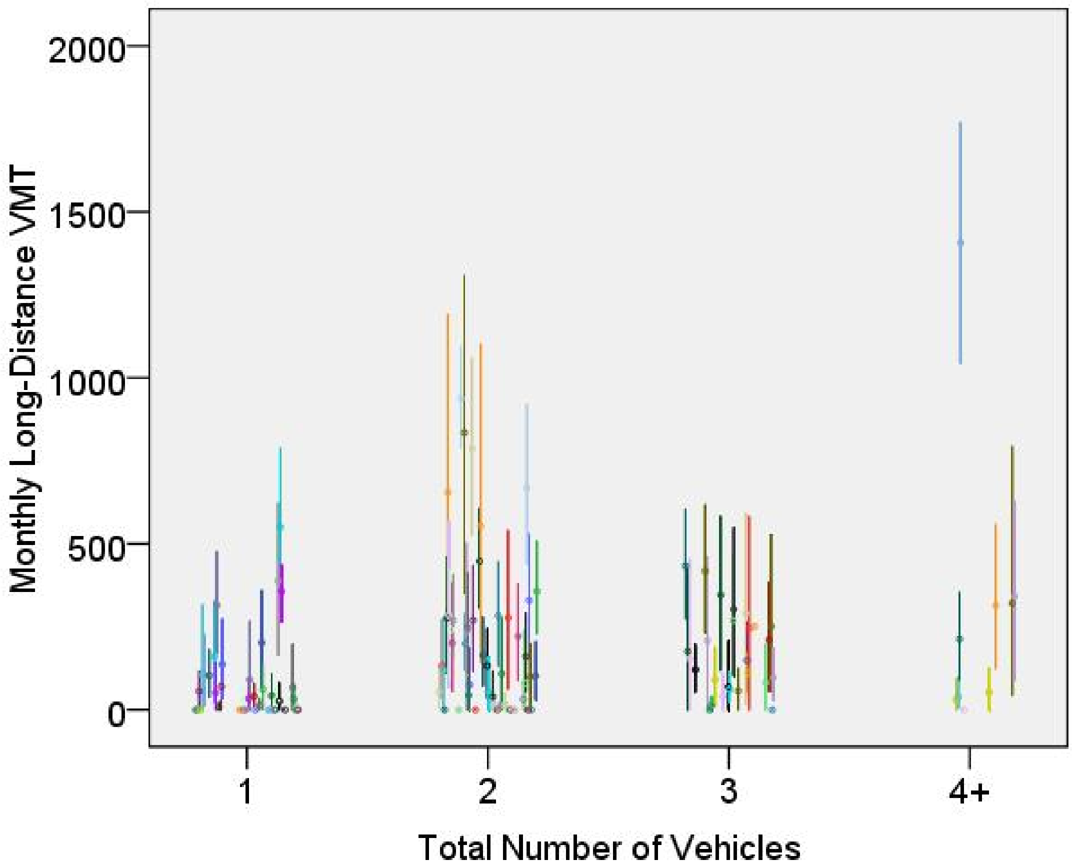

Figure A.22 Between-Household and Within-Household Variability in Monthly Long-Distance VMT by Total Number of Vehicles Owned
Each error bar represents a household.

APPENDIX B

WITHIN-HOUSEHOLD ASSOCIATION

Figure B.1 and Figure B.2 show each of the 18 choose 2 scatterplots of responses from households at different times. The months are labeled relative to the beginning of the study, so month 1 is the first month (October, 2004), and month 21 is the last month (June, 2006) of the study. The red lines divide the matrices into three major sections – the upper left section illustrates the correlations among months within the baseline period; the lower right section represents the correlations among months within the pricing period; the lower left section represents the correlations among months between the baseline and the pricing periods.

Notice from the main diagonal of both scatterplot matrices that there is substantial positive correlation between repeated observations on the same household that are one month apart. The degree of correlation decreases as the observations are moved farther from one another in time, corresponding to moving farther from the diagonal. While this phenomenon is true to both the number of trips and the VMT, the degree of correlation in the VMT decreases faster than that of the number of trips. Notice also that the correlations in both figures are reasonably consistent along a diagonal in the matrices. This phenomenon is more prominent in Figure B.1. This indicates that the correlation of number of trips depends more strongly on the time between observations than on their absolute times. The correlation of VMT, however, seems to depend both on the time between observations and on absolute times, which could be associated with some explanatory variables such as gas prices and pricing incentives.

The red ovals in Figure B.1 and Figure B.2 highlight correlations in the months that are exactly one year apart. In other words, the scatterplots in the red ovals depict the impact of seasonality in travel. If seasonality is prominent, one should expect to see stronger correlations in the scatterplots highlighted by the red ovals. It is difficult to identify stronger correlations between the same months from the scatterplots, but the correlation matrices in Section 8.2 were able to reflect the presence of seasonality.

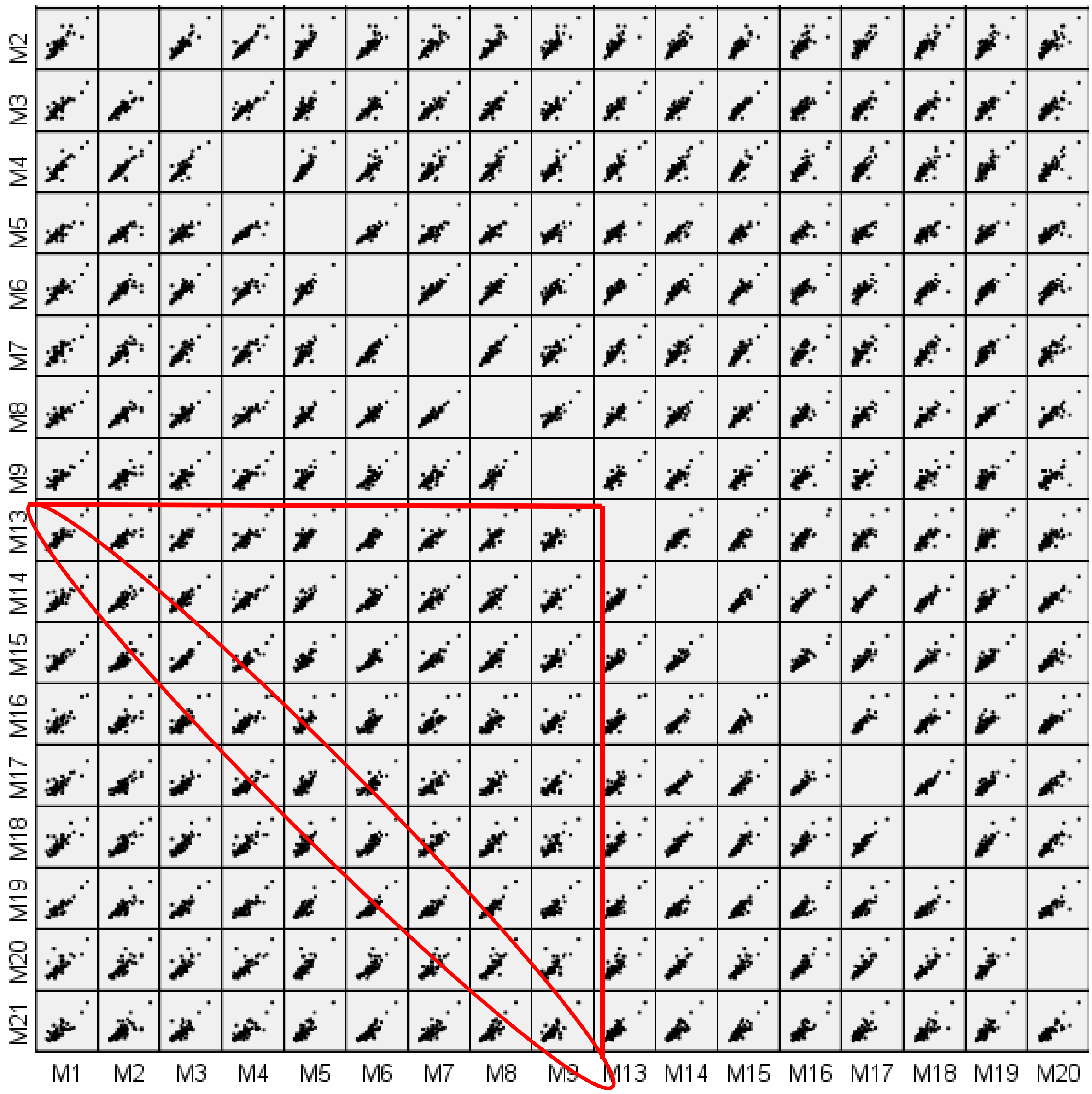

Figure B.1 Scatterplot Matrix of Monthly Number of Intra-Regional Trips.
Axis lables are months relative to beginning of study.

Figure B.2 Scatterplot Matrix of Monthly Intra-Regional VMT
Axis labels are months relative to beginning of study.

Table B.1 Estimated Correlation Matrix for Monthly Number of Long-Distance Tours

Entries are Spearman's $\rho(Y_{ij}, Y_{ik})$, $1 \le t_{ij} < t_{ik} \le 30$ months[1].

	t_{ij}																												
t_{ik}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
2	.48*																												
3	.54*	.52*																											
4	.32*	.45*	.28*																										
5	.48*	.45*	.43*	.42*																									
6	.58*	.54*	.50*	.42*	.47*																								
7	.43*	.57*	.46*	.46*	.61*	.60*																							
8	.49*	.53*	.38*	.44*	.58*	.47*	.52*																						
9	.39*	.32*	.46*	.43*	.50*	.64*	.52*	.54*																					
10	.28*	.36*	.31*	.50*	.21*	.38*	.28*	.47*	.31*																				
11	.37*	.27*	.31*	.44*	.42*	.34*	.45*	.39*	.42*	.35*																			
12	.37*	.29*	.27*	.54*	.32*	.36*	.37*	.45*	.31*	.47*	.53*																		
13	.52*	.55*	.38*	.49*	.38*	.47*	.38*	.40*	.34*	.40*	.40*	.48*																	
14	.19	.18	.03	.26*	.23*	.16	.25*	.26*	.11	.28*	.34*	.45*	.52*																
15	.20	.18	.22*	.17	.20	.29*	.20*	.18	.20	.30*	.13	.30*	.43*	.32*															
16	.13	.22*	.18	.39*	.29*	.35*	.35*	.35*	.37*	.56*	.33*	.48*	.41*	.39*	.35*														
17	.29*	.23*	.14	.32*	.35*	.26*	.24*	.43*	.29*	.24*	.34*	.30*	.37*	.38*	.37*	.46*													
18	.26*	.21*	.28*	.33*	.38*	.35*	.21*	.47*	.39*	.33*	.29*	.42*	.31*	.28*	.47*	.31*	.46*												
19	.23*	.26*	.13	.20	.07	.25*	.22*	.34*	.18	.32*	.29*	.19	.28*	.31*	.29*	.39*	.35*	.18											
20	.11	.01	.13	.15	.10	.06	.12	.15	.17	.19	.22*	.24*	.25*	.25*	.29*	.20	.39*	.31*	.30*										
21	.25*	.16	.07	.35*	.15	.28*	.12	.23*	.39*	.26*	.25*	.21*	.29*	.26*	.22*	.26*	.20	.30*	.40*	.40*									
22	.23*	.44*	.27*	.62*	.45*	.30*	.44*	.51*	.37*	.65*	.47*	.45*	.47*	.31*	.21*	.55*	.47*	.29*	.40*	.26*	.34*								
23	.29*	.26*	.26*	.27*	.19	.26*	.19	.22*	.28*	.27*	.35*	.36*	.39*	.26*	.17	.27*	.21*	.35*	.24*	.18	.26*	.19							
24	.38*	.31*	.23*	.44*	.21*	.38*	.29*	.38*	.32*	.50*	.31*	.44*	.42*	.39*	.37*	.41*	.36*	.53*	.44*	.26*	.45*	.35*	.46*						

[1] *: Correlation is significant at 0.05 level (two-tailed)

Table B.1 (continued)

25	.20	.16	.30*	.15	.09	.21*	.12	.27*	.29*	.27*	.37*	.42*	.29*	.23*	.21*	.27*	.12	.32*	.21*	.31*	.20	.18	.52*	.31*					
26	.24*	.31*	.12	.22*	.28*	.12	.22*	.40*	.16	.19	.24*	.36*	.28*	.27*	.32*	.24*	.34*	.35*	.24*	.29*	.48*	.36*	.25*	.31*	.19				
27	.19	.13	.23*	.01	.18	.24*	.16	.25*	.31*	.21*	.21*	.30*	.34*	.29*	.45*	.30*	.25*	.37*	.33*	.30*	.33*	.26*	.35*	.25*	.42*	.39**			
28	.07	.04	.03	.24*	.20	.08	.27*	.18	.12	.12	.22*	.12	.28*	.47*	.31*	.25*	.28*	.34*	.30*	.22*	.26*	.24*	.20	.39*	.32*	.30*	.24*		
29	.22*	.18	.16	.15	.27*	.10	.13	.27*	.13	.09	.16	.22*	.33*	.41*	.33*	.27*	.58*	.33*	.25*	.43*	.21*	.20	.15	.25*	.18	.23*	.31*	.42*	
30	.07	.17	.16	.22*	.14	.32*	.19	.27*	.26*	.39*	.34*	.43*	.30*	.30*	.49*	.48*	.35*	.32*	.26*	.21	.26*	.39*	.05	.17	.33*	.24*	.38*	.32*	.46*

Table B.2 Estimated Correlation Matrix for Monthly Long-Distance VMT

Entries are Spearman's $\rho(Y_{ij}, Y_{ik})$, $1 \le t_{ij} < t_{ik} \le 30$ months[1].

	t_{ij}																												
t_{ik}	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**
2	.50*																												
3	.52*	.49*																											
4	.25*	.33*	.20																										
5	.45*	.43*	.41*	.37*																									
6	.49*	.45*	.40*	.36*	.36*																								
7	.42*	.51*	.33*	.34*	.57*	.52*																							
8	.49*	.52*	.34*	.32*	.58*	.36*	.46*																						
9	.43*	.27*	.45*	.33*	.50*	.59*	.44*	.49*																					
10	.19	.24*	.19	.48*	.14	.27*	.10	.36*	.16																				
11	.35*	.20	.29*	.42*	.39*	.30*	.35*	.29*	.38*	.23*																			
12	.35*	.28*	.22*	.49*	.28*	.36*	.29*	.40*	.23*	.39*	.47*																		
13	.46*	.51*	.33*	.41*	.32*	.45*	.27*	.36*	.29*	.28*	.35*	.49*																	
14	.17	.17	-.01	.23*	.22*	.16	.22*	.22*	.06	.22*	.35*	.45*	.52*																

[1] *: Correlation is significant at 0.05 level (two-tailed)

Table B.2 (continued)

15	.17	.16	.19	.16	.21*	.32*	.16	.14	.17	.24*	.11	.31*	.42*	.28*															
16	.06	.15	.15	.34*	.29*	.38*	.28*	.26*	.25*	.53*	.26*	.43*	.33*	.33*	.33*														
17	.29*	.24*	.11	.20	.31*	.26*	.22*	.41*	.24*	.24*	.28*	.29*	.37*	.36*	.29*	.46*													
18	.22*	.18	.20*	.22*	.34*	.23*	.17	.42*	.22*	.25*	.20	.41*	.29*	.28*	.45*	.32*	.42*												
19	.28*	.29*	.18	.21*	.09	.28*	.25*	.38*	.17	.34*	.34*	.22*	.28*	.29*	.15	.34*	.32*	.10											
20	.12	.03	.15	.12	.11	.08	.08	.15	.16	.13	.24*	.25*	.25*	.23*	.21*	.17	.34*	.27*	.21*										
21	.28*	.16	.06	.35*	.17	.28*	.09	.20	.39*	.19	.31*	.26*	.26*	.25*	.13	.16	.17	.18	.30*	.24*									
22	.19	.40*	.24*	.53*	.41*	.30*	.35*	.46*	.29*	.61*	.43*	.43*	.46*	.36*	.18	.54*	.52*	.29*	.43*	.23*	.31*								
23	.27*	.23*	.21*	.20	.15	.24*	.12	.13	.19	.21*	.36*	.28*	.34*	.24*	.15	.23*	.16	.26*	.24*	.10	.26*	.18							
24	.39*	.29*	.20*	.42*	.18	.39*	.25*	.31*	.26*	.46*	.28*	.47*	.39*	.32*	.33*	.40*	.33*	.50*	.43*	.23*	.42*	.35*	.42*						
25	.15	.16	.29*	.10	.10	.20	.02	.22*	.25*	.15	.35*	.38*	.29*	.24*	.23*	.20	.13	.29*	.18	.23*	.15	.16	.39*	.22*					
26	.31*	.40*	.16	.20*	.36*	.15	.27**	.46*	.17	.14	.26*	.36*	.31*	.25*	.26*	.17	.36*	.41*	.21*	.19	.43*	.38*	.18	.33*	.16				
27	.22*	.12	.23*	-.05	.21*	.20	.10	.20	.25*	.13	.22*	.25*	.30*	.23*	.46*	.22*	.25*	.26*	.23*	.17	.27*	.25*	.18	.17	.34*	.30*			
28	.05	.04	-.01	.28*	.21*	.11	.26*	.18	.15	.14	.21*	.14	.28*	.44*	.26*	.25*	.19	.29*	.26*	.12	.24*	.22*	.13	.34*	.24*	.29*	.15		
29	.24*	.23*	.16	.12	.35*	.11	.15	.28*	.12	.07	.16	.26*	.32*	.37*	.24*	.27*	.47*	.36*	.15	.42*	.16	.20	.16	.25*	.14	.22*	.33*	.32*	
30	.04	.15	.10	.20	.11	.34*	.08	.23*	.19	.33*	.28*	.39*	.33*	.30*	.44*	.47*	.34*	.26*	.24*	.18	.27*	.38*	-.04	.16	.28*	.23*	.33*	.36*	.39*

REFERENCES

Abou-Zeid, M., M. Ben-Akiva, K. Tierney, K. Buckeye and J. Buxbaum (2008). "Minnesota Pay-as-You-Drive Pricing Experiment." Transportation Research Record: Journal of the Transportation Research Board **2079**: 8-14.

Blossfeld, H.-P. and G. Rohwer (2002). Techniques of Event History Modeling: New Approaches to Causal Analysis. Mahwah, NJ, Lawrence Erlbaum Associates.

Bradley, M. and J. Bowman (2006). "A Summary of Design Features of Activity-Based Microsimulation Models for US MPOs." White Paper for the Conference on Innovations in Travel Demand Modeling. Austin, TX.

Brownstone, D., T. Golob and C. Kazimi (1999). Modeling Non-Ignorable Attrition and Measurement Error in Panel Surveys: An Application to Travel Demand Modeling. International Conference on Survey Nonresponse. Portland, OR, University of California, Irvine, Dept. of Economics, Institute of Transportation Studies.

Buxbaum, J. (2008). Minnesota: Variabilization of Fixed Auto Costs. Value Pricing Pilot Program Projects Report: October - December 2008 Quarterly, Cambridge Systematics, Inc.

Cambridge Systematics, Inc. (1996). Travel Survey Manual.

Cambridge Systematics, Inc. (2008). Review of Activity-Based Models, Florida Department of Transportation.

Cameron, A. C. and P. K. Trivedi (1998). Regression Analysis of Count Data. Cambridge, UK, Cambridge University Press.

Davidson, W., R. Donnelly, P. Vovsha, J. Freedman, S. Ruegg, J. Hicks, J. Castiglione and R. Picado (2007). "Synthesis of First Practices and Operational Research Approaches in Activity-Based Travel Demand Modeling." Transportation Research Part A: Policy and Practice **41**(5): 464-488.

Davison, A. C. and D. V. Hinkley (1997). Bootstrap Methods and Their Application. Cambridge, Cambridge University Press.

Diggle, P. J., P. J. Heagerty, K.-Y. Liang and S. L. Zeger (2002). Analysis of Longitudinal Data, New York, Oxford University Press.

Dunn, P. (2003). "Occurrence and Quantity of Precipitation Can Be Modelled Simultaneously." International Journal of Climatology **24**: 1231-1239.

Dunn, P. and G. Smyth (1996). "Randomized Quantile Residuals." Journal of Computational and Graphical Statistics **5**(3): 236-244.

Dunn, P. and G. Smyth (2005). "Series Evaluation of Tweedie Exponential Dispersion Model Densities." Statistics and Computing **15**(4): 267-280.

Efron, B. and R. J. Tibshirani (1993). An Introduction to the Bootstrap. New York, Chapman & Hall.

Elango, V. and R. Guensler (2010). A Methodology to Identify Activity for Passively Collected GPS Data. Manuscript submitted to the 89th Transportation Research Board Annual Meeting.

Elango, V., R. Guensler and J. Ogle (2007). "Day-to-Day Travel Variability in the Commute Atlanta, Georgia, Study." Transportation Research Record: Journal of the Transportation Research Board **2014**: 39-49.

Energy Information Administration (2010). Lower Atlantic Weekly Retail Gasoline Historic Prices.

Fitzmaurice, G., N. Laird and J. Ware (2004). Applied Longitudinal Analysis, Wiley-IEEE.

Fraenkel, J. and N. Wallen (2008). How to Design and Evaluate Research in Education 7 Edition, McGraw Hill Higher Education.

Frangakis, C. and D. Rubin (2001). "Addressing an Idiosyncrasy in Estimating Survival Curves Using Double Sampling in the Presence of Self-Selected Right Censoring." Biometrics **57**(2): 333-342.

Hardin, J. W. and J. M. Hilbe (2002). Generalized Estimating Equations. Boca Raton, FL, Chapman & Hall/CRC.

Horton, N. and S. Lipsitz (1999). "Review of Software to Fit Generalized Estimating Equation Regression Models." The American Statistician **53**(2).

Hu, P. S. and T. R. Reuscher (2004). Summary of Travel Trends 2001 National Household Travel Survey. Oak Ridge, Tennessee, Oak Ridge National Laboratory.

Jones, P. and M. Clarke (1988). "The Significance and Measurement of Variability in Travel Behaviour." Transportation **15**(1): 65-87.

Jørgensen, B. (1987). "Exponential Dispersion Models (with Discussion)." Journal of the Royal Statistical Society, Series B **49**: 127–162.

Jørgensen, B. (1997). The Theory of Dispersion Models, Chapman & Hall London.

Jørgensen, B. and M. De Souza (1994). "Fitting Tweedie's Compound Poisson Model to Insurance Claims Data." Scandinavian Actuarial Journal **1**: 69–93.

Kitamura, R. (1990). "Panel Analysis in Transportation Planning: An Overview." Transportation Research **24**(6): 401–415.

Kleinman, K. and N. J. Horton (2010). SAS and R: Data Management, Statistical Analysis, and Graphics. Boca Raton, FL, CRC Press.

Kochan, B., T. Bellemans, D. Janssens and G. Wets (2008). Dynamic Activity-Travel Diary Data Collection Using a Global Positioning System–Enabled Personal Digital Assistant. Innovations in Travel Demand Modeling, Volume 2: Papers. Austin, TX, Transportation Research Board**:** 94-97.

Kuppam, A., M. Outwater, M. Bradley, L. Blain, R. Tung and S. Yan (2005). Application of Time-of-Day Choice Models Using Emme/2–Washington State DOT Congestion Relief Analysis.

Kuppam, A., M. Outwater and R. Hranac (2008). Innovative Methods for Pricing Studies. Innovations in Travel Demand Modeling, Volume 2: Papers. Austin, Texas, Transportation Research Board. **2:** 142-149.

Kutner, M. H., C. J. Nachtsheim, J. Neter and W. Li (2005). Applied Linear Statistical Models, McGraw Hill.

Kvam, P. and B. Vidakovic (2007). Nonparametric Statistics with Applications to Science and Engineering Hoboken, New Jersey, Wiley-Interscience.

Lee-Gosselin, M. (2002). Some Reflections on GPS-Supported Travel Survey Methods in an Increasingly ICT-Rich Environment. Workshop on ICT, Innovation and the Transport System. Arlington.

Li, H. (2004). "Investigating Morning Commute Route Choice Behavior Using Global Positioning Systems and Multi-Day Travel Day." Diss. Georgia Institute of Technology.

Li, H., R. Guensler and J. Ogle (2005). "An Analysis of Morning Commute Route Choice Patterns Using Gps Based Vehicle Activity Data." Transportation Research Record: Journal of the Transportation Research Board **1926**: 162-170.

Li, H., R. Guensler, J. Ogle, J. Jun and V. Elango (2007). Descriptive Analysis of Long-Distance Travel by Personal Vehicles Using 2004 Commute Atlanta Data. TRB 86th Annual Meeting Compendium of Papers CD-ROM. Transportation Research Board, Washington, DC.

Liang, K.-Y. and S. L. Zeger (1986). "Longitudinal Data Analysis Using Generalized Linear Models." Biometrika **73**(1): 13-22.

Liu, G. and K. Liang (1997). "Sample Size Calculations for Studies with Correlated Observations." Biometrics **53**(3): 937-947.

Mace, A. (1964). Sample-Size Determination. New York, Reinhold

Madre, J.-L., K. W. Axhausen and W. Brög (2007). "Immobility in Travel Diary Surveys." Transportation **34**: 107-128.

Markus, G. (1979). Analyzing Panel Data, Newbury, CA, Sage Publications.

McCullagh, P. and J. A. Nelder (1989). Generalized Linear Models. New York, Chapman and Hall.

Moser, C. and G. Kalton (1971). Survey Methods in Social Investigation, Heinemann Educational Books.

Muñoz, A., V. Carey, J. P. Schouten, M. Segal and B. Rosner (1992). "A Parametric Family of Correlation Structures for the Analysis of Longitudinal Data." Biometrics **48**(3): 733-742.

Murakami, E., S. Greaves and T. Ruiz (2006). Moving Panel Surveys from Concept to Implementation. Travel Survey Methods: Quality and Future Directions. P. Stopher and C. Stecher**:** 399-412.

Muthen, L. and B. Muthen (2002). "How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power." Structural Equation Modeling **9**(4): 599-620.

NIST/SEMATECH (2003). "E-Handbook of Statistical Methods." http://www.itl.nist.gov/div898/handbook/pri/section3/pri332.htm. (Accessed May 25, 2009).

Ogle, J. (2005). "Quantitative Assessment of Driver Speeding Behavior Using Instrumented Vehicles." Diss. Georgia Institute of Technology.

Ogle, J., R. Guensler and V. Elango (2005). "Commute Atlanta Value Pricing Program: Recruitment Methods and Travel Diary Response Rates." Transportation Research Record **1931**: 28-37.

Ogle, J., J. Ko, H. Li and R. Guensler (2006). Analysis of under-Reporting and Other Errors in Commute Atlanta Travel Diaries Using Instrumented Vehicle Data. Manuscript.

Ostrom, C., Jr. (1978). Time Series Analysis: Regression Techniques. Beverly Hills, CA, Sage Publications.

Parry, I. W. H. and K. A. Small (2005). "Does Britain or the United States Have the Right Gasoline Tax?" American Economic Review **95**(4): 1276-1289.

Parsons Brinckerhoff Quade & Douglas, Inc. (2000). Executive Summary: General Final Report for the Regional Travel - Household Interview Survey. New York, NY.

Pas, E. (1986). "Multiday Samples, Parameter Estimation Precision, and Data Collection Costs for Least Squares Regression, Trip-Generation Models." Environment and Planning A **18**(1): 73-87.

PB Consult Inc. and The Gallop Corporation (2005). The Integrated Regional Model Project. Vision Phase - Final Report. Denver, CO, Denver Regional Council of Governments.

Pendyala, R. (1999). "Measuring Day-to-Day Variability in Travel Behavior Using Gps Data." Washington, DC, FHWA. http://www.fhwa.dot.gov/ohim/gps. (Accessed April 5, 2010).

Perry, J. and L. Taylor (1985). "Ades: New Ecological Families of Species-Specific Frequency Distributions That Describe Repeated Spatial Samples with an Intrinsic Power-Law Variance-Mean Property." Journal of Animal Ecology **54**(3): 931-953.

Pierce, B., J. Casas and G. Giaimo (2003). "Estimating Trip Rate under-Reporting: Preliminary Results from the Ohio Household Travel Survey." Transportation Research Board Annual Meeting, Washington, DC.

Prentice, R. L. and L. P. Zhao (1991). "Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses." Biometrics **47**(3): 825-839.

Puget Sound Regional Council (2008). "Traffic Choice Study - Summary Report." Seattle, WA

Puller, S. L. and L. A. Greening (1999). "Household Adjustment to Gasoline Price Change: An Analysis Using 9 Years of Us Survey Data." Energy Economics **21**(1): 37-52.

Rochon, J. (1998). "Application of Gee Procedures for Sample Size Calculations in Repeated Measures Experiments." Statistics in Medicine **17**(14): 1643-1658.

Rufolo, A. and T. Kimpel (2008). "Responses to Oregon's Experiment in Road Pricing." Transportation Research Record: Journal of the Transportation Research Board **2079**: 1-7.

Schönfelder, S., H. Li, R. Guensler, J. Ogle and K. Axhausen (2006). "Analysis of Commute Atlanta Instrumented Vehicle GPS Data: Destination Choice Behavior and Activity Spaces." TRB 85th Annual Meeting Compendium of Papers CD-ROM. Washington, DC, Transportation Research Board.

Smyth, G. (1996). "Regression Analysis of Quantity Data with Exact Zeros." Proceedings of the Second Australi-Japan Workshop on Stochastic Models in Engineering, Technology and Management, Technology Management Centre, University of Queensland.

SPSS Inc. (2007). SPSS Tutorial.

Stopher, P. (2008a). "Collecting and Processing Data from Mobile Technologies." 8th International Conference on Survey Methods in Transport. Annecy, France.

Stopher, P. (2008b). "The Travel Survey Toolkit: Where to from Here?" 8th International Conference on Travel Survey Methods. Annecy, France.

Stopher, P., E. Clifford and M. Montes (2008a). "Variability of Travel over Multiple Days: Analysis of Three Panel Waves." Transportation Research Record: Journal of the Transportation Research Board **2054**: 56-63.

Stopher, P., C. FitzGerald, J. Zhang and T. Bretin (2007). "Analysis of 28-Day Global Positioning System Panel Survey: Findings on Variability of Travel." Transportation Research Record: Journal of the Transportation Research Board **2014**: 17-26.

Stopher, P. and S. Greaves (2007). "Guidelines for Samplers: Measuring a Change in Behaviour from before and after Surveys." Transportation **34**(1): 1-16.

Stopher, P., K. Kockelman, S. Greaves and E. Clifford (2008b). "Sample Size Requirements for Multi-Day Travel Surveys: Some Findings." 8th International Conference on Travel Survey Methods. Annecy, France.

Travel Survey Methods Committee (2009) "Research Needs Statement."

Tweedie, M. (1984). "An Index Which Distinguishes between Some Important Exponential Families." Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference, Indian Statistical Institute, Calcutta: 579–604.

Vovsha, P. and M. Bradley (2004). "Hybrid Discrete Choice Departure-Time and Duration Model for Scheduling Travel Tours." Transportation Research Record: Journal of the Transportation Research Board **1894**: 46-56.

Vovsha, P. and M. Bradley (2006). "Advanced Activity-Based Models in Context of Planning Decisions." Transportation Research Record: Journal of the Transportation Research Board **1981**: 34-41.

Vovsha, P., W. Davidson and R. Donnelly (2005). "Making the State of the Art the State of the Practice: Advanced Modeling Techniques for Road Pricing." Expert Forum on Road Pricing and Travel Demand Modeling Proceedings. Alexandria, VA. 95-122.

Wolf, J., M. Oliveira and M. Thompson (2003). "Impact of Underreporting on Mileage and Travel Time Estimates: Results from Global Positioning System-Enhanced Household Travel Survey." Transportation Research Record: Journal of the Transportation Research Board **1854**: 189-198.

Xu, Y., L. Zuyeva, D. Kall, V. Elango and R. Guensler (2009a). Commute Atlanta Value Pricing Phase II Case Study Report. Atlanta, GA, Georgia Department of Transportation**:** 834.

Xu, Y., L. Zuyeva, D. Kall, V. Elango and R. Guensler (2009b). "Mileage-Based Value Pricing: Phase Ii Case Study Implications of Commute Atlanta Project." TRB 88th Annual Meeting Compendium of Papers CD-ROM. Washington, DC, Transportation Research Board.

Ye, X., R. Pendyala and G. Gottardi (2007). "An Exploration of the Relationship between Mode Choice and Complexity of Trip Chaining Patterns." Transportation Research Part B **41**(1): 96-113.

Yee, J. and D. Niemeier (1996) "Advantages and Disadvantages: Longitudinal Vs. Repeated Cross-Section Surveys." Project Battelle **94**.

Zumkeller, D., J. Madre, B. Chlond and J. Armoogum (2006). "Panel Surveys." Travel Survey Methods: Quality and Future Directions. P. Stopher and C. Stecher, Elsevier**:** 363-398.