# Should My Car Move or Should I? A Model of Residential and Commuting Choices

Christopher Marshall Clapp
Flemington, New Jersey

M.A., University of Virginia, 2007
M.A., Clemson University, 2003
B.A., Clemson University, 2002

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Economics

University of Virginia
August, 2013

## Abstract

Policymakers have been slow to implement price-based congestion policies due in part to how little is known about the effects of policies that influence more than simply an individual's commuting method. An individual can also alter her commute by choosing to travel from a different location. I develop a discrete choice structural model of the joint decisions of family residence and individual commuting modes, given the characteristics of the housing market and commuting options. I use rich individual-level data that allow me to include numerous unobserved heterogeneity terms; this strengthens the validity of my results relative to more aggregate analyses that are often undertaken. I am in the process of using model estimates to simulate the full set of effects of transportation policies that alter the financial and time costs of commuting. These policies include congestion pricing schemes, fuel or carbon taxes, and increased parking fees.

I estimate my model using individual-level Public Use Microdata Sample (PUMS) data from the 2005-2008 American Community Survey (ACS) for the Washington, D.C. metropolitan area. The PUMS data requires that I randomly assign individuals home and work locations, but the Census Bureau has granted me access to precise information on where individuals live and work from the restricted-access version of the ACS that I am currently using to improve the analysis. I augment the information in the ACS with data I have painstakingly assembled on the structure of the transportation network to map each individual's optimal commute from each home and by each commuting method in the choice set. To do this, I use geographic information system (GIS) network analysis. The mappings allow me to create a unique dataset of individual commute options and characteristics that I use to estimate the trade-offs that individuals make among consumption, housing amenities, and leisure when choosing a home and commuting mode pair.

I also develop and plan to implement a methodology that (unlike previous literature) does not require that I treat groups of individuals living together as if they have a single set of preferences. Instead, I use a collective model of the household to account for the fact that spouses rarely commute to the same work location. This allows me to model the interplay between residential and commuting mode choices when spouses consider the proximity of their home to both work locations. I allow family members to have caring preferences, and I treat characteristics of the home as a family public good. The collective model requires observing individual consumption of at least one private good in the household to identify bargaining outcomes, and I use a novel assignable private good: the method and duration of each commute. This work is both an extension of the collective model to the residential choice and travel literatures as well as an application of the collective model to a problem with discrete choices and a rich error structure.

JEL Codes: D13, Q52, R21, R41, and R48

Keywords: Residential Location, Travel Mode Choice, Intra-Household Allocation, Congestion Pricing, Discrete Choice Analysis, Geographic Information Systems

## Acknowledgements

Completing a dissertation requires a great deal of hard work and perseverance on the part of the author and an even greater deal of support from those in both his professional and personal lives. I wish to extend my deepest gratitude to all those who have helped me earn my doctorate.

I must begin by thanking all of the faculty in the Department of Economics at the University of Virginia for the excellent education I have received. I would particularly like to thank my primary advisor, Steven Stern, for all that he has taught me, for always being incredibly generous with his time, and for his unwavering patience along the way. I cannot imagine a better teacher of the discipline. I am also indebted to Leora Friedberg and Amalia Miller for their guidance on numerous aspects of the dissertation, insightful comments on multiple drafts, and general support. I have also benefited from conversations with Wayne-Roy Gayle, Edgar Olsen, Sarah Turner, John Pepper, and numerous seminar speakers who have visited U.Va. over the years. My understanding of Economics, research, and sanity have all benefited from interactions with fellow graduate students who are too numerous to list completely, but include: Justin Ward, Nate Hoover, Paul Landefeld, Jenica Wurm, Kang Jian, Marianne Corbishley, Dusan Curcic, Mariusz Kolczykiewicz, Catherine Alford, Charlie Murry, Ignacio Martinez, and Sarah Tulman.

I would also like to thank Patrick Bayer, Bert Grider, and the staff of the U.S. Census Bureau's Center for Economic Studies (CES) for assistance with restricted-access data and insightful comments. I would like to thank Chris Gist and Kelly Johnston for guidance on the use of ArcGIS to conduct my GIS network analysis, Katherine Holcomb for assistance with the use of Fortran and OpenMP to code my estimation program, and Mila Versteeg for rounding out my dissertation committee. Finally, I would like to recognize all that members of the staff do to keep the Department running smoothly. Thank you, Patty, Debs, Bunny, Michael, Henry, Adam, Dedrick, Linda, and Jack.

I would be remiss if I didn't also thank my family. I am indebted to my wife, Susan, for enduring every setback and celebrating every success with me along the way. I would not have earned my doctorate without her willingness to listen to recaps of my days as a graduate student and unfailing ability to be something to look forward to at the end of even the toughest of days. I also want to thank my parents, Robert and Diane, and brother, David, for their steadfast backing and pride in my accomplishments. Finally, I want to thank my dog, Scottie, for sleeping on my lap while I worked and making days toiling alone in the office actually pleasant.

DISCLAIMER: Any opinions and conclusions expressed herein are those of the author and do not necessarily represent the views of the U.S. Census Bureau. No results in this dissertation are based on restricted-access data, ensuring that no confidential information is disclosed.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Traffic jams are more than just a minor annoyance. American automobile commuters lost an estimated 4.8 billion hours and 3.9 billion gallons of fuel because of congestion in 2009, a cost estimated at $115 billion (Schrank and Lomax 2010). Worse, congestion is not improving. The average annual congestion delay has more than doubled since 1982, the first year for which data is available. The social welfare cost of congestion is likely even greater than these estimates due to losses from uncertainty over commute times and congestion-induced increases in global and local pollution, traffic accidents, and noise. While the optimal amount of congestion is unlikely to be zero, Small (2008) opines that "[v]irtually all economists agree that congestion in cities around the world is greater than [the] optimum." Urban planners traditionally attempt to reduce congestion by increasing capacity: either by expanding roadways or public transit systems. Recent figures show that, in 2006, federal, state, and local governments spent $16.2 billion on new road construction and another $13.8 billion to widen existing roadways, in addition to making $12.8 billion in capital improvements to the nation's mass transit systems (U.S. Department of Transportation 2008). Yet, Duranton and Turner (2009) find that building an additional kilometer of roadway leads to a one-to-one increase in mean daily vehicle kilometers traveled. They also show that the supply of mass transit alternatives has no effect on vehicle kilometers traveled. In other words, the most prevalent policy instruments for reducing congestion do not appear to have their intended effect.

As congestion continues to increase, communities across the country are looking at new congestion pricing policies that place a monetary cost on travel when and where congestion is greatest.[1] Parry et al. (2007) explain that these policies can reduce congestion by internalizing externalities; yet, widespread use of these policies has been difficult for policymakers to implement due to concerns that the policies are regressive in nature and fears that travelers will face increased financial costs without offsetting time savings. Determining whether these, and other, common misgivings about congestion reduction policies are warranted is a daunting task because it requires urban planners to predict the outcome of policies that affect far more than just how people commute. The limitation of simple models of commuting choices is that people make decisions about where to live based on their transportation options. Moreover, many of these decisions are made by households with multiple members facing different commutes.

My research develops a structural model of family residential choice and family member commuting method to inform this discussion. I make three key contributions to the literature. First, I address the endogeneity of residential choice in models of commuting method by explicitly modeling both residential and commuting choices together. Baum-Snow and Kahn (2000), Duranton and Turner (2009), and Bento et al. (2005) all find evidence that residential choice and commuting decisions are inextricably tied, and failure to adequately address this connection results in biased coefficient estimates of either either decision individually.

Second, I estimate my model using PUMS data from the 2005-2009 ACS for the Washington, D.C. metropolitan area. These rich individual-level datasets allow me to include numerous unobserved heterogeneity terms which strengthen the validity of my results relative to more aggregate analyses that are often undertaken. In the current work I present based on the PUMS data, I randomly assign individuals to home and work locations, but

[1] These policies include cordon charges that impose a fee on drivers who travel within or into a congested area and variably priced, managed lanes that prevent congestion by charging an adjustable access toll (Lewis 2008).

work-in-progress uses precise information on where individuals live and work from the restricted-access version of the ACS. The geographic location information, along with data I have assembled on the structure of the transportation network allows me to map each individual's optimal commute from each home and by each commuting method in the choice set. To do this, I use geographic information system (GIS) network analysis. The mappings allow me to create a unique dataset of individual commute options and characteristics that I use to estimate the trade-offs that individuals make among consumption, housing amenities, and leisure when choosing a home and commuting mode pair. This allows me to improve upon the residential choice literature which at best controls for the role of commuting costs in housing decisions with either neighborhood aggregate commute costs (time or distance) or imprecise measures of the head of the household's commute.[2]

Finally, I plan to use model estimates to simulate the full set of effects of transportation policies that alter the financial and time costs of commuting on the joint distribution of residential housing and commuting methods. These policies include congestion pricing schemes, fuel or carbon taxes, and increased parking fees.

My current work focuses on individual decision-makers and results are based on households with a single adult commuter. The paper also shows how the model can be extended to account for households that contain multiple individuals who commute to different work locations, and I plan to estimate a full version of the model that takes advantage of data on both single individuals and cohabiting couples as part of my future research agenda. The extended methodology departs from the current residential choice literature by relaxing the ubiquitous assumption that individuals living together act as if they have a single set of

[2]Langer and Winston (2008), who are also interested in the effects of congestion reduction policies, ask, "How can one estimate the economic effects of road pricing while accounting for its impact on land use?" They posit a methodology similar to the one I develop, but note that it is an ambitious undertaking in explaining their decision to use hedonic methods to answer the question, by saying, "a disaggregate approach for a metropolitan area would model the determinants of a commuter's choice of mode of transportation, departure time, destination, route and residential location and simulate how those choices change in response to an efficient congestion toll. Unfortunately, the data and modeling requirements of a disaggregate approach-especially in determining a commuter's residential location alternatives and their attributes-are formidable." My work attempts to make progress on the formidable task they describe.

preferences.[3] This is a defensible assumption if one considers all the characteristics of a residence to be public goods within the family, but one's commute is a distinctly private good within the family because different individuals within a household often commute to jobs in different parts of their metropolitan area. Browning et al. (1994) find empirical evidence that family members bargain over the consumption of non-durable private goods. It stands to reason that they also bargain over more longstanding decisions such as where they commute from and by what mode. I use the collective household model developed by Chiappori (1988) to explicitly model not only the interplay between residential and commuting mode choices but also how those decisions are made within the family. This added precision will provide a clearer understanding of the value individuals place on commuting characteristics when making residential location decisions than what can be gained from estimates based on unitary models or estimates based on single individual behavior alone.

The next section provides a review of the related literature. Chapter 3 describes my theoretical model for both single individuals and cohabiting partner families. I detail the data used in chapter 4. Chapters 5 and 6 explain my estimation strategy and results for single individuals. The results section in this paper is incomplete, but I outline the tasks that will be completed as part of the published version of this research project. Finally, I offer conclusions in chapter 7. Development of the cohabiting partner family analogs to all sections subsequent to chapter 3 is still in progress and is left for future research.

[3] To the best of my knowledge, a working paper by Chiappori et al. (2012) is the only exception to this rule.

# Chapter 2

# Literature Review

My research draws from four distinct literatures. There is a plethora of work in the transportation, residential location choice, and household behavior literatures that are related to this paper. I discuss insights from each of these areas, as well as highlight ways in which my work advances the given literature, in the subsequent sections. Finally, I conclude the literature review with some background on congestion reduction methods.

## 2.1 Transportation

There has been a great deal of research on what can broadly be categorized as travel demand analysis, and Small and Verhoef (2007) provide a comprehensive overview of the many methods that are used. In terms of the transportation component of my work, I use individual-level data to estimate commuting modes, so I focus on their treatment of disaggregate models of mode choice, much of which is based on random utility maximization (RUM) models. McFadden (2001) provides a historical survey of the methodology of individual travel demand analysis using RUM models. There are several insights from these works that are key to my research.

First, in order to estimate how individuals commute, one needs to find a way to measure the alternative-specific attributes of commutes that an individual did not choose. As

Small and Verhoef (2007) explain, there are two options: either use values reported by individuals in the survey or use engineered values produced from network analysis. Each has shortcomings. The former may be biased because individuals do not know much about the options that they do not choose or because they misreport so as to reinforce the option they do choose. The latter are costly to calculate and are not always accurate. My data reports commute times for chosen options only, so I calculate engineered values using GIS network analysis. Although computationally expensive to calculate, results in Section 4.5 show that they do explain a some of the variation in reported commute times, conditional on commuting by the given method.

Second, the most common methodological tool used for discrete-choice models is the (additively) RUM model developed by McFadden (1974). The basic identically and independently distributed (iid) multinomial probit (MNP) and logit (MNL) versions of this model suffer from the well known independence from irrelevant alternatives (IIA) problem, so subsequent research has relaxed this assumption, commonly with nested MNL or mixed-multinomial logit (MMNL) models. My structural model also relaxes these assumptions, as well as the assumption that an individual, choice specific error enters utility linearly, as there is no strong economic justification for this specification.

Finally, one must be careful when modeling mode choice, as not all explanatory variables can be thought of as independent of the decision of how to commute. Price is often thought of as endogenous in the differentiated products literature because of unobserved characteristics of the alternative that are not included in the model and are correlated with its price. When modeling mode choice, transportation researchers can overcome this problem by including alternative-specific constants, as I do. This is feasible because of the limited number of commuted options, but doing so comes at the cost of precluding variables that vary with alternatives but not individuals.

An endogenous variable of greater concern for models of mode choice is automobile ownership, as the decision to own a car is likely made simultaneously with commuting decision. While adding automobile ownership to the model is beyond the scope of my research, my model still makes several contributions. Such a model can be thought of as the “reduced-form” version of a model that allows individuals to choose whether or not to own a vehicle (Small and Verhoef, 2007, page 29). I discuss the bias that not modeling automobile ownership induces in Section 5.3.4.

Small and Verhoef (2007) argue that the endogeneity of travel characteristics is not a great concern when using disaggregate data because researchers can make the assumption that individuals take those characteristics as given when making travel related decisions. This is true to the extent that individuals have no control over the characteristics of their commute. However, Baum-Snow and Kahn (2000) find evidence that some of the increase in system travelers after the expansion of mass transit systems can be attributed to individuals who move to take advantage of the infrastructure improvements, and Duranton and Turner (2009) indicate that individuals have the same response to the expansion of roadways. Although individuals cannot control, for instance, how fast traffic flows on a given road or where a given subway train stops, they can control which road or which subway line they take by choosing where they live relative to where they work.[1] Additionally, Bento et al. (2005) find evidence of a relationship in the other direction: measures of urban spatial structure have small but significant effects on travel demand. They provide a thorough discussion of the bias that is caused by failure to adequately address this connection. In order to address the biases introduced by the interdependence of commuting characteristics and residential location, I jointly model both residential choice and commuting decisions.

## 2.2 Residential Choice

Early, theoretical work on location decisions is characterized by the assumption that all individuals commute to the same central business district, and a land-rent gradient develops (see, Alonso 1964, Mills 1967, and Muth 1969). Evidence of this gradient can be seen in current empirical research. Bajari and Kahn (2005, 2008) model residential location

[1] To be precise, work location also influences commuting characteristics. See Section 5.3.4 for a discussion of the implications of not including work location decisions in my model.

decisions using a three-step estimation process based on hedonic estimation of home prices. The latter work explicitly controls for commuting costs with the average commute time of individuals who live in the given home's Census tract. They find that willingness to pay to reduce commuting time is slightly less than the household owner's hourly wage at the margin.[2] Langer and Winston (2008), who also use hedonic methods, calculate a marginal willingness to pay of roughly half the average household wage when using the average commute time in the household to measure commuting costs. While these aggregate measures of commuting are useful in hedonic settings where the value of the home is determined by market forces (not just an individual's valuation), they are less satisfactory in a model of individual outcomes. I model the actual commuting options and characteristics that individuals face when choosing a home. I also relax the implicit assumption that all commuters travel to the same area for work and model cities as aspatial urban areas instead of monocentric ones.

An alternative way to model residential location decisions uses a single-crossing assumption about preferences for locality ammenities and costs to model how individuals self-select into neighborhoods. Epple and Sieg (1999), Epple et al. (2001), Banzhaf and Walsh (2008), and Epple et al. (2010) model intra-jurisdictional sorting models in order to test the Tiebout (1956) hypothesis. These community-level studies are the basis for the current location choice literature, but the availability of new data has allowed subsequent studies to extend these jurisdictional-level techniques to the individual residence level.

Bayer and coauthors have pioneered the use of restricted-access Census microdata to explore topics ranging from segregation in housing markets (Bayer et al. 2004) to labor market hiring networks among neighbors (Bayer et al. 2008). Bayer et al. (2005), Bayer et al. (2007), and Bayer and McMillan (2011) estimate equilibrium models of residential choice using household data and the differentiated products methods of Berry et al. (1995). They model differences in household preferences for residential locations, conditional on work location, but focus on the implications of Tiebout (1956) sorting in housing markets,

[2] They note that this estimate is greater than estimates of roughly half the hourly wage commonly found in the transportation literature.

not commuting decisions. They control for the influence of commuting in residential decisions by through the as-the-crow-flies distance to the head of household's job. Bayer and McMillan (2011) finds, for instance, that households are willing to pay about $50 per month to reduce daily commutes by one mile. I improve on their methodology by more accurately modeling the commute faced by individuals in the household. Specifically, I model the duration and mode of the commute, and I allow for heterogeneity in preferences over commuting methods. They also find that their commuting estimates are sensitive to controls for unobserved neighborhood quality, so I develop a flexible way to incorporate neighborhood effects into my model that is outlined in Section 5.1.1.2.

The earliest empirical models of residential choice were developed in the late 1970s (Lerman 1976, McFadden 1978) based on the RUM model. While Lerman (1976) also incorporates commuting decisions (as well as automobile ownership), I know of only one recent paper that models the joint decision of residential location and commuting mode using individual level data. Vega and Reynolds-Feighan (2009) use GIS network analysis to augment individual-level data to estimate a cross-nested logit (CNL) model the joint location and commuting decision. Although they model commutes using GIS techniques for both automotive and mass transit options, they aggregate all of the work locations in their city of analysis (Dublin, Ireland) to four employment centers, resulting in a loss of precision. They caution that a key methodological difficulty that must be overcome in their model is finding a way to limit the size of the choice set (which is a function of the number of homes observed in the data) so that estimation is tractable. They explain that there are two alternatives: sampling from the full choice set or spatially aggregating home location alternatives. The former requires additional assumptions (see McFadden 1978) and the later is problematic because the unit of aggregation is arbitrarily defined.[3] They find that commuters traveling to the center city are responsive to policies that increase travel costs in terms of which mode they choose, but that commuters traveling to suburban work locations are unlikely to switch their commute modes in response to those same policies. They also

[3] I discuss how I handle this issue in Section 4.3.

find evidence that congestion policies are likely to have an effect on residential location decisions, particularly for those who commute to suburban work locations. I am working to improve on their methodology by both more accurately controlling for the commutes individuals face and by accounting for household bargaining over commuting characteristics between spouses. This will allow me to add cohabiting couples to the model, instead of just relying on single individuals for estimation.

## 2.3 Household Behavior[4]

Chiappori (1988), Chiappori (1992), and Browning and Chiappori (1998) provide the basic theory behind the collective model of household behavior. The collective model is one of many attempts to reconcile the inherent contradiction in applying the tools of individual preference theory to a multi-person household. The model does not seek to explain the mechanism behind the household decision process. Instead, it assumes that however household outcomes are determined, the bargaining process is Pareto efficient. In the context of my research, since a household is comprised of multiple individuals who likely have different preferences over household characteristics and methods of commuting, as well as different work locations, it is unlikely that a household will behave as a single decision maker. Instead, members of a household will bargain over the bundle of housing amenities and location characteristics that maximize their own utilities.

Browning et al. (1994) find empirical evidence in support of the collective model but encourage additional analysis. The context of my model provides three key differences from their previous tests. First, housing is a far more expensive, durable good than clothing. My research expands the test of the collective model to a new class of goods. Second, while not heroic, the assumption that men's clothing is exclusive to the husband and women's clothing to the wife leaves room for doubt. I treat commuting method as the assignable good

[4] As the empirical component of my intra-household model is still in progress, the reader may choose to skip this section without fear of missing information that is critical to understanding subsequent sections of the paper.

in my model, which offers advantages over their focus on clothing, namely, that it is observed by the econometrician. Finally, the authors note that "[c]omparing the sharing rules obtained with different supposedly assignable goods would in fact provide additional tests of the actual nature of these goods and of the general collective framework used throughout this paper." Since the choice of housing location and commuting method together determine the time that each spouse must spend commuting, I compare the sharing rules determined from both commuting method and duration.

Xu (2007) and Chiappori and Donni (2009) provide surveys of the state of the literature. They review the application of the inter-household model to numerous empirical contexts (labor supply, household production, etc.). To the best of my knowledge, Chiappori et al. (2012) is the only work that applies the collective model to housing and commuting decisions, finding that failing to model bargaining over household locations would bias model results and effect policy prescriptions. They do so using French census data for the city of Paris, but they do not observe any information on the characteristics of a given home, so they model locations at a more aggregate level (the commune, of which there are 1,300 in Paris) than I do. They model commuting costs as a flexible function of time, but do not allow these costs to vary with commuting mode or allow individuals to choose how they commute. These modeling assumptions reduce the reliability of their estimates relative to my specification and preclude analysis of policies that affect commuting behavior.

## 2.4 Congestion Pricing

Economists have long advocated for use fees that internalize congestion externalities and improve welfare.[5] Lindsey (2006) provides a comprehensive survey of the theoretical literature on road pricing dating back to Adam Smith, but congestion pricing policies have only

[5] Parry et al. (2007) discuss the externalities associated with automotive travel and the policies, ranging from fuel taxes to congestion pricing, that can be used to address those externalities. The discussion is both in terms of efficacy and political feasibility.

more recently begun being implemented and are still not widespread.[6] Lewis (2008) provides an overview of the various forms of congestion pricing policies which I summarize in Table 1.

Table 1: Congestion Pricing Policies

| **Type** | **Definition** | **Examples** |
|---|---|---|
| Area Wide | Charges based on congestion level on all roads | None |
| Variable Roadway | Tolls include rush hour fees for particular roads | NJ Turnpike |
| Managed Lanes | Variable tolls for separated lanes within a highway | I-15 & SR-91 (CA) |
| Cordon | Fee to drive within or into a congested area | London |
| Zonal | Cordon charging with adjacent charging zones | Trials in Europe |

Source: Table created by the author using information from Lewis (2008).

He argues for the effectiveness of congestion pricing policies with some impressive measures. The cordon charge introduced in London, England in 2003 reduced traffic in the cordon by 20 percent, increased traffic speeds by 37 percent, and raised more than $100 million in net revenues that were used to improve the city's mass transit system. Leape (2006) reports that the London cordon charge has been such a success that there have been discussions of a nationwide congestion pricing policy. In the United States, managed lanes on SR-91 in California had an average speed of 60 miles per hour during peak hours while congestion in the untolled lanes reduced speeds to under 20 miles per hour.[7]

Small et al. (2005, 2006) perform a thorough analysis of the effects of the congestion pricing mechanism used on SR-91, finding that it does improve motorist welfare due to significant heterogeneity in traveler preferences. This occurs because low-value-of-time commuters are displaced by high-value-of-time commuters who reap large benefits. However, the available data prevents the authors from modeling mode choice, residential location, or time of travel, all of which can be varied by commuters in the long run. My model addresses these concerns, while allowing for a robust set of unobserved heterogeneity pa-

[6] See https://ceprofs.civil.tamu.edu/mburris/pricing.htm for a list of all instances of congestion pricing in practice today. At present, there are less than 50 (broadly defined) examples of congestion pricing on roads around the world.

[7] See Anas and Lindsey (2011) for more information on the effects of several major congestion pricing programs.

rameters their work suggests are important.

# Chapter 3

# Model

This chapter outlines a structural model of residential choice and commuting method that I estimate using Census microdata. The structure allows me to determine the relative importance of housing and neighborhood characteristics on residential choice, including distance to place of employment and access to commuting options. I allow for heterogeneous preferences for those characteristics as well as for commuting methods. I first detail the model as it pertains to a single individual. Then, as an extension I will estimate in future work, I move to the case of a family comprised of two bargaining adults. Finally, I explain how I model the impact of children on the behavior of both of these types of families.

While I advance the literature by treating the choice of residential location and commuting mode as joint in a model with as much geographic detail as I have, I must nevertheless take other decisions as fixed in order to keep the model tractable. I assume that an individual takes her city of residence, family structure, vehicle ownership, and employment as given when deciding among transportation options and residential choices. Additionally, I assume that the locations and hours of employers and schools are independent of residential choices and transportation options. Finally, I assume that there are no household production effects associated with commuting decisions and that both members of a cohabiting couple have the same preference for the well-being of their children. All of these assumptions have the potential to bias my results, although to varying degrees. I discuss the implications of these assumptions in Section 5.3.4.

## 3.1 Single Person Household

The simplest type of family is that of an individual choosing where she alone will reside and how she will commute. I build from the standard labor-leisure framework. An individual has preferences over consumption and leisure and faces both a budget and time constraint. Consumption is defined over a composite good and housing amenities, and leisure takes the form of either time spent away from work or of some fraction of time spent commuting.

### 3.1.1 Preferences

I define a market (indexed by $m$) at the metropolitan level and assume that jobs ($j$) have characteristics that include wages, hours, and location. Given a fixed market and job, an individual ($i$) is faced with the decision of which house to live in ($h$) and by what method to commute ($k$). Preferences are defined over composite consumption ($c_{ihk}$), housing amenities ($\tilde{H}_{ih}$), and leisure ($\tilde{\ell}_{ihk}$) and represented by a utility function as

$$U\left(c_{ihk}, \tilde{H}_{ih}, \tilde{\ell}_{ihk}\right).$$

The aggregate consumption good, $c_{ihk}$, includes all non-housing consumption and savings. As in Bayer et al. (2005), the individual derives utility from many different housing amenities, including characteristics of both the house and the neighborhood. In order to include a rich set of housing characteristics but still keep the utility function tractable, I define $\tilde{H}_{ih}$ as a function of observable housing and neighborhood characteristics ($H_{ih}$) and unobservable characteristics ($\varepsilon_{ih}$),

$$\tilde{H}_{ih} = \exp\left(H_{ih}\gamma^{H} + \varepsilon_{ih}\right).$$

The $\exp(\cdot)$ ensures that the utility function can be evaluated.[1] The observable characteristics, $H_{ih}$, are allowed to vary over both $i$ and $h$ in order to allow for interactions between individual and home-specific observables, but variation over individuals is not necessary for the identification of $\gamma^H$. The error term, $\varepsilon_{ih}$, is necessary to explain cases where an individual chooses to live in a home that is observationally inferior to other homes in her feasible choice set. It can account not only for unobserved characteristics of the home, but also for search and moving costs that might lock an individual into a given home, but that are not explicitly modeled. It is known to the agent but not to the econometrician, thus providing a source of unobserved heterogeneity in the model.

As in McFadden (2001), non-work time has two components,

$$\tilde{\ell}_{ihk} = \ell_{ihk} + (1 - \lambda_{ik})\, t_{ihk}.$$

The $\ell_{ihk}$ term represents pure leisure. Time spent by individual $i$ commuting between home $h$ and job $j$ by method $k$ ($t_{ihk}$) may contain a leisure component that is known to the agent but not the econometrician.[2] This component accounts for heterogeneity in preferences for commuting methods to explain cases where an individual chooses to commute by a method that is more costly, both in terms of time and money, than other feasible methods.[3] It is measured by the random variable $\lambda_{ik}$, which is bounded from below at 0 and varies over individuals and methods of commuting. As such, if $\lambda_{ik} = 0$, time spent commuting by method $k$ is a perfect substitute for pure leisure. Note that if commuting by method $k$ is stressful and work-like, $\lambda_{ik} \cong 1$. A value of $\lambda_{ik} > 1$ means that the individual views commuting to be less enjoyable than work.

[1] In a subsequent section, I specify the utility function with a log transformed Cobb-Douglas functional form. The $\exp(\cdot)$ ensures that $\tilde{H}_{ih} > 0$ so that $\ln(\tilde{H}_{ih})$ can always be evaluated.

[2] Note that I drop the $j$ subscript in all variables that vary over $i$, as jobs are taken as fixed for a given individual.

[3] That heterogeneity in preferences for commuting methods affects utility through leisure is an assumption. This specification is useful because it allows the preference to vary with the duration of the commute. An individual may have an extreme dislike for driving, but may opt to drive if a short commute minimizes the displeasure. This specification is less desirable if the costs or benefits of a given method of commuting are not variable. For instance, if an individual prefers to drive because of the flexibility it allows in running errands after work.

There is nothing in economic theory that requires a lower bound on $\lambda_{ik}$, but $\lambda_{ik} < 0$ does not seem plausible. A value of $\lambda_{ik} < 0$ would mean that the individual would rather commute than engage in general leisure activities. Since traveling by method $k$ is a feasible leisure activity, I restrict $\lambda_{ik}$ to prevent nonsensical preferences.[4]

### 3.1.2 Prices

Individual $i$ takes as given several prices in her market. The price of the aggregate consumption good varies by metropolitan area. A local cost-of-living index, denoted as $p^c_m$, is used to measure this variation. The opportunity cost of owning or renting a home is imputed as in Bayer et al. (2007) and is represented as $p^H_h$.[5] I do not observe savings or wealth, nor does my data allow for a dynamic model, so converting housing stock expenses into flow opportunity costs is necessary, given that a savings motive does not drive housing choice in my model. The average pecuniary cost per mile of commuting via method $k$ in market $m$ is denoted as $p^d_{mk}$, where the $d$ superscript denotes distance.[6]

In the data, there are 12 reported methods of commuting. These methods are condensed to the most relevant options in Table 2, with associated per mile and fixed commuting costs.

Household automobile ownership is is observed only as the number of vehicles available, but not make, model, or year of those vehicles, so I use an average measure of miles per gallon (MPG) to determine the price of commuting by automobile.[7] The number of

[4]This restriction is supported by the the time use literature. Krueger et al. (2008) provide comparisons of how people felt while engaging in different activities. Unsurprisingly, their results indicate that individuals prefer most leisure activities to commuting. Additionally, their results show that commuting and working rank as almost equally unenjoyable activities, with their ordinal rankings varying by survey methodology.

[5]For notational clarity, I capitalize the "*H*" superscript that serves as a label for the price to avoid confusion with the "*h*" subscript that serves as an index.

[6]Fixed costs, such as parking fees and tolls, are assuredly important components of commuting decisions, but I do not observe these costs in the data. The former is not reported by individuals and the latter depends on the exact commuting route, which I do not observe.

[7]Future research will more accurately specify MPG as the sum of the mean MPG of the automotive fleet in the given year and an individual specific error. This will allow me to integrate over the distribution of the error in order to obtain a more accurate measure of automotive commuting costs, as well as allow for correlation with other errors in the model.

Table 2: Commuting Methods and Costs

| | **Method ($k$)** | **Pecuniary Cost Per Mile** |
|---|---|---|
| 1) | Automobile | $p^d_{m,\text{auto}} = \frac{p_{\text{gas}}}{MPG}$ |
| 2) | Carpool | $p^d_{m,\text{pool}} = \frac{p_{m,\text{auto}}}{N^{pool}}$ |
| 3) | Bus | $p^d_{m,bus} = \bar{p}_{\text{m,bus}}$ |
| 4) | Streetcar | $p^d_{m,streetcar} = \bar{p}_{\text{m},streetcar}$ |
| 5) | Subway | $p^d_{m,subway} = \bar{p}_{\text{m},subway}$ |
| 6) | Rail | $p^d_{m,rail} = \bar{p}_{\text{m,rail}}$ |
| 7) | Walk | $p^d_{m,other} = 0$ |

people in the carpooling option is denoted by $N^{pool}$. The $\bar{p}_{mk}$ prices are the average fare per mile for the given system in metropolitan area $m$.

### 3.1.3 Constraints

Individual $i$ faces both a budget constraint and a time constraint. To represent expenditures, I first define $N^H_m$ and $N^K_m$ as the number of homes and commuting methods in market m. I then define an $N^H_m \times 1$ vector, $I_i$, whose $h$th element is 1 if the individual lives in home $h$ and 0 otherwise. Next, I define $d_{ihk}$ as the distance between house $h$ and job $j$ that individual $i$ travels by commuting method $k$. I pack those distances into an $N^K_m \times 1$ vector of commuting distances traveled by individual $i$ from house $h$ by each commuting method, $d_{i\text{h}}$.[8] The budget constraint is defined as

$$p^c_m c_{ihk} + \underset{(1\times N^H_m)}{p^H} \underset{(N^H_m\times 1)}{I_i} + \underset{(1\times N^K_m)}{p^d_m} \underset{(N^K_m\times 1)}{d_{\text{ih}}} = w_{\text{i}} L_{\text{i}},$$

where $w_{\text{i}}$ is individual i's wage, and $L_{\text{i}}$ is the individual's time spent at work. Sample selection criteria guarantee that all individuals are employed, and wages and work hours are taken as fixed.

I denote total time as $T$ and the commuting time by method $k$ as $t_{ihk}$. Individual $i$'s time

[8]Note that the $k$th element of $d_{ih}$ will be 0 for individuals who do not commute by the $k$th method.

constraint is

$$\ell_{ihk} + t_{ihk} + L_{\mathrm{i}} = T.$$

The commuting time by method $k$ ($t_{ihk}$) is treated as a function of a linear index of the characteristics of the commute ($K_{ihk}\gamma^K$) and a measurement error term ($e_{ihk}$) due to the econometrician's uncertainty about the exact route the agent takes, traffic patterns, etc.[9] It is written as

$$t_{ihk} = \exp\left(K_{ihk}\gamma^K + e_{ihk}\right), \tag{3.1}$$

where the $\exp(\cdot)$ ensures that time spent commuting is positive. Note that random, temporary shocks (e.g., accidents, weather, construction) do not affect the agents' long term commuting decisions.

### 3.1.4 Parameterization

I define the utility function with a Cobb-Douglas functional form and make the familiar natural log transformation, which results in

$$U\left(c_{ihk}, \tilde{H}_{ih}, \tilde{\ell}_{ihk}\right) = \alpha_i^c \ln\left(c_{ihk}\right) + \alpha_i^H\left(H_{ih}\gamma^H + \varepsilon_{ih}\right) + \alpha_i^\ell \ln\left(\ell_{ihk} + \left(1-\lambda_{ik}\right)t_{ihk}\right),$$

where $\alpha_i^c$, $\alpha_i^H$, and $\alpha_i^\ell$ are taste parameters over composite consumption, housing amenities, and leisure. I normalize $\alpha_i^c$ to 1 and $\gamma_1^H$ to 0 to ensure identification. The other parameters, $\alpha_i^H$ and $\alpha_i^\ell$, are allowed to vary with observable characteristics ($X_i$) of the individual and contain error terms to capture unobserved heterogeneity in preferences. The

[9]This specification is necessary because my data reports commute times for chosen options only. I use characteristics of the commute calculated using GIS network analysis to impute unobserved commute times based estimates of $\gamma^K$ recovered from observed commute times.

taste parameters are defined as[10]

$$\begin{aligned}
\alpha_i^c &= 1, \\
\alpha_i^H &= \exp\left(X_{\mathrm{i}}\beta^H + \mu_i\right), \\
\alpha_i^\ell &= \exp\left(X_{\mathrm{i}}\beta^\ell + u_i\right).
\end{aligned}$$

### 3.1.5 Choice Problem

Taking labor market decisions, job characteristics, and vehicle ownership as given, the full choice set is a residence and a method of commuting. By choosing a residence, the individual determines the characteristics of both her home and commute options. The joint choice of a residential location and a particular method of commuting determine the individual's consumtion and allocation of time. The former is uniquely determined by the budget constraint, since there is no saving in the model. Similarly, since hours of labor are taken as given, the time constraint determines leisure. In summary, an agent in the model faces the unconstrained choice problem

$$\begin{aligned}
\max_{h_i,k_i} U\left(c_{ihk}, \tilde{H}_{ih}, \tilde{\ell}_{ihk}\right) &= \ln\left(\frac{w_{\mathrm{i}}L_{\mathrm{i}} - p^H I_i - p_m^d d_{\mathrm{ih}}}{p_m^c}\right) + \exp\left(X_{\mathrm{i}}\beta^H + \mu_i\right)\left(H_{ih}\gamma^H + \varepsilon_{ih}\right) \\
&+ \exp\left(X_{\mathrm{i}}\beta^\ell + u_i\right)\ln\left(T - L_{\mathrm{i}} - \lambda_{ik}\exp\left(K_{ihk}\gamma^K + e_{ihk}\right)\right). \qquad (3.2)
\end{aligned}$$

## 3.2 Cohabiting Couple Household[11]

In future work, I will estimate a version of the model that explicitly features a cohabiting or married couple that bargains over housing and commuting choices.

[10] Note that the exponential form guarantees that the utility parameters will be positive, ensuring that "goods are good."

[11] As the empirical component of my intra-household model is still in progress, the reader may choose to skip this section without fear of missing information that is critical to understanding subsequent sections of the paper.

### 3.2.1 Intra-Household Bargaining

Browning et al. (1994), Browning and Chiappori (1998), and Browning et al. (2006) provide a thorough discussion of issues involved in modeling intra-household allocations and the assumptions required for identification and estimation of the collective model of the household. The collective model is a methodology that does not require the econometrician to formally model the household bargaining mechanism. Instead, it assumes only that bargaining is cooperative, so, however it operates, outcomes are efficient.

The beauty of the collective model is that it allows the researcher to model individual behavior despite a severe limitation of most datasets: that private consumption is observed at the household, not the individual level. The collective model does so by transforming the Pareto weight in the standard household welfare maximization problem into what is known as the sharing rule. Although similar, the Pareto weight and sharing rule are not the same. The Pareto weight denotes the influence each member of the household's utility function has on the household welfare function. The sharing rule determines the division of expenditures net of public goods available to each member of the household (Browning et al., 2006). Heuristically, one can think of this method as decomposing household bargaining into a two stage process. First, members of the household pool their resources and use those common funds to pay for public goods within the household (the home, utilities, etc.). Second, they divide the remaining funds according to the sharing rule, then optimize subject to individual budget constraints.

Following the collective model methodology, I assume that the econometrician can perfectly designate goods as public or private in the household setting, at least one private good can be assigned to each family member who consumes it, utility is separable with respect to private consumption, and variables exist that affect the weight each individual's utility receives in the household decision process but do not influence preferences. Specifically, I assume that all housing amenities ($H$) are public goods within the family. I also assume that consumption and leisure are private goods.[12] I treat commuting trips and durations

[12]The leisure time of partners is likely to be complementary, but I am not able to capture this effect, as I do

as assignable private goods, as I observe which family member makes the trip, his or her method of commuting, and the duration of his or her commute.[13]

### 3.2.2 Preferences

When modeling collective decisions, researchers frequently assume that members of the household have interdependent or caring preferences (Becker, 1981). Whereas traditional, egoistic preferences are such that an individual derives utility only from her own consumption, caring preferences imply that an individual's utility is defined over her own consumption, leisure, etc. and her partner's total welfare.[14] When used in an efficient bargaining model, caring preferences moderate results so that outcomes are not heavily skewed to one member of the couple or the other. Chiappori (1997) explains the intuition behind this result: with a very unequal outcome, both partners can increase their utility by agreeing on a more even result. The unfavored partner achieves more utility directly from an increased share and the favored partner achieves more utility indirectly from the unfavored partner's increase in utility.

Caring preferences are both intuitively appealing and empirically supported.[15] However, Browning et al. (2006) shows that the collective model is identified regardless of whether individuals are assumed to have egoistic or caring preferences. More so, Lise and Seitz (2011) explain how the type of preferences assumed change only the interpretation of the sharing rule. With caring preferences, the sharing rule is just that: a measure of how household wealth is divided. With egoistic preferences, the sharing rule is a measure of

---

not observe the extent to which partners spend leisure time together. While I could identify a complementarity effect through the covariation in husband and wife choices of leisure time, I cannot separately identify this effect from similarity in partner preferences for leisure. I leave this for future research.

[13] I do not observe how individuals spend their leisure time, so I have to assume that the duration of an individual's commute does not affect her share of household production responsibilities. This assumption is problematic if, for instance, the partner with the shorter commute is responsible for cooking dinner for the family.

[14] For comparison, an individual with altruistic or paternalistic preferences derives utility from her own consumption, as well as the consumption of her partner. The collective model is generally not identified with altruistic preferences.

[15] Friedberg and Stern (2010) find evidence that members of married couples are willing to trade some of their own utility for an increase in their spouse's utility, which suggests that couples have caring preferences.

both sharing and caring. I explicitly detail the model with caring preferences, then show that egoistic preferences are just a special form of those preferences.

Following the notational convention of Browning et al. (1994), let $s = \{A,B\}$ index members of the household.[16] Define subutility functions for each spouse that are analogous to the utility function given in the maximization problem in Equation 3.2. Formally, these subutility functions are represented as

$$\begin{aligned} v^s\left(c^s_{ihk},\tilde{H}_{ih},\tilde{\ell}^s_{ihk}\right) &= \ln\left(c^s_{ihk}\right)+\exp\left(X^s_{\mathrm{i}}\beta^H+\mu^s_i\right)\left(H_{ih}\gamma^H+\varepsilon_{ih}\right) \\ &+ \exp\left(X^s_{\mathrm{i}}\beta^\ell+u^s_i\right)\ln\left(T-L^s_{\mathrm{i}}-\lambda^s_{ik}\exp\left(K^s_{ihk}\gamma^K+e^s_{ihk}\right)\right) \\ & for\; s=\{A,B\}. \end{aligned}$$

The arguments of the subutility functions are the same as those previously detailed in Section 3.1, save one important difference. Hereafter, the index $i$ should be thought of as an index of families (with the $s$ superscript clarifying individuals within the family). Note that household characteristics $(H_{ih})$ are a public good to the family.[17]

With caring preferences, spouse $s$'s utility $(U^s)$ is a function of both $v^A$ and $v^B$, and it can be written as

$$U^s_i = F^s\left[v^s_i\left(c^s_{ihk},\tilde{H}_{ih},\tilde{\ell}^s_{ihk}\right),v^{-s}\left(c^{-s}_{ihk},\tilde{H}_{ih},\tilde{\ell}^{-s}_{ihk}\right)\right]\; for\; s=\{A,B\}\; and\; -s=\{B,A\},$$

where $F^s$ is spouse $s$'s aggregator function. Egoistic preferences are just a special form of the aggregator function where $F^s[\cdot]$ is such that

$$U^s_i = v^s_i\left(c^s_{ihk},\tilde{H}_{ih},\tilde{\ell}^s_{ihk}\right)\; for\; s=\{A,B\}.$$

[16] The reader can think of $A$ denoting the husband and $B$ the wife if s/he so desires.

[17] This convention does not alter the notation for the previous case of a single person family. In such cases, the $i$ subscript denotes both the family and the individual, as they are one and the same.

Regardless of the form of individual preferences, the preferences of the household are represented by a welfare function and can be written as

$$W_i \quad = \quad \tilde{\mu}_i U_i^A + (1 - \tilde{\mu}_i) U_i^B,$$

where $\tilde{\mu}_i$ is the Pareto weight for household $i$.[18]

### 3.2.3 Constraints

The household optimization is subject to individual time constraints and a household budget constraint. The individual time constraints are straightforward to deal with (the careful reader will note that they have already been substituted into the third argument of $v^s\left(c^s_{ihk}, \tilde{H}_{ih}, \tilde{\ell}^s_{ihk}\right)$), but the household budget constraint,

$$p^c_m\left(c^A_{ihk} + c^B_{ihk}\right) + p^H I_i + p^d_m\left(d^A_{ih} + d^B_{ih}\right) = w^A_{\mathrm{i}} L^A_{\mathrm{i}} + w^B_{\mathrm{i}} L^B_{\mathrm{i}},$$

presents a problem, as individual consumption, $c^A_{ihk}$ and $c^B_{ihk}$, are not separately observed in the data. [19]

### 3.2.4 Collective Model

Since the utility function is separable in the public and private goods, the Pareto weight is not dependent on the level of public goods chosen. Browning et al. (1994) show that a collective decision can be modeled as a process where each member of the household receives a share of total expenditures (net of expenditures on public goods) and purchases his or her

[18]The tilde breaks from the convention in the literature in order to differentiate the Pareto weight ($\tilde{\mu}_i$) from the error associated with the preference parameter for housing amenities ($\mu_i$).

[19]Note that the budget constraint takes this form because labor market decisions are exogenous. If they were not, individuals would "buy back" their leisure and commuting time from the labor market, so expenditures on those goods would appear on the left-hand side of the equation and total possible household income ($w^A_i T + w^B_i T$) would appear on the right-hand side.

own private goods. The rule for splitting the net expenditures is known as the sharing rule, which is represented by $\rho_i$. Following this procedure generates separate budget constraints for each partner equal to

$$\begin{aligned} p_m c^A_{ihk} + p^d_m d^A_{ih} &= \rho_i \left( w^A_i L^A_{\mathrm{i}} + w^B_{\mathrm{i}} L^B_{\mathrm{i}} - p^H I_i \right), \\ p_m c^B_{ihk} + p^d_m d^B_{ih} &= (1-\rho_i) \left( w^A_i L^A_{\mathrm{i}} + w^B_{\mathrm{i}} L^B_{\mathrm{i}} - p^H I_i \right), \end{aligned}$$

and removes the Pareto weight from the household welfare function. In other words, the sharing rule specification transforms $W_i$ from being a function of the Pareto weight to being a function of the sharing rule. Formally,

$$\begin{aligned} W_i &= \tilde{\mu}_i U^A_i + (1-\tilde{\mu}_i) U^B_i \\ &= U^A_i (\rho_i) + U^B_i (\rho_i), \end{aligned}$$

where $U^s_i$ is a function of $\rho_i$ because the budget constraint equates consumption, so $c^s_{ihk}$ is a function of $\rho_i$. The sharing rule is unobserved by the econometrician. It is assumed to be a function of variables that affect the sharing rule, but not individual preferences, known as sharing shifters or distributional factors (DFs) and an error term that captures unobserved heterogeneity in household bargaining patterns. The DFs are denoted by $Z_i$, the unobserved heterogeneity term is denoted by $\psi_i$, and the sharing rule takes the form

$$\rho_i = \Phi(Z_i \zeta + \psi_i),$$

where the $\Phi(\cdot)$ function is the standard normal cumulative distribution function.[20] It ensures that the sharing rule is bounded between 0 and 1. I include DFs commonly used in the literature, such as the percentage of household income earned by each member, differences

[20] Note that this form is a somewhat arbitrary specification of the sharing rule, but specifying the sharing rule as a function of DFs is not. Since the data I use provides only a repeated cross section of individuals, DFs are required to identify the sharing rule.

in age and education between the spouses, and conditions of relevant marriage markets.[21]

### 3.2.5 Choice Problem

The household maximizes its welfare function with respect to a single residence and a method of commuting for each member. The household optimization problem can formally be stated as

$$\max_{h_i, k_i^A, k_i^B} W_i = U_i^A + U_i^B,$$

subject to

$$\begin{aligned} p_m c_{ihk}^A + p_m^d d_{ih}^A &= \rho_i \left( w_i^A L_{\mathrm{i}}^A + w_{\mathrm{i}}^B L_{\mathrm{i}}^B - p^H I_i \right), \\ p_m c_{ihk}^B + p_m^d d_{ih}^B &= (1-\rho_i) \left( w_i^A L_{\mathrm{i}}^A + w_{\mathrm{i}}^B L_{\mathrm{i}}^B - p^H I_i \right). \end{aligned} \tag{3.3}$$

## 3.3 Households with Children

Children are an important factor in the housing and commuting decisions of both single person and cohabiting couple parents. In order to capture the effect that children have on housing and commuting decisions, I include the presence and characteristics of children and the interaction of these terms with key housing and neighborhood characteristics. For instance, the interaction of local school quality with children in the household will help control for a parental desire to send their children to high quality schools. This specification means that the model treats children like a household public good, so each parent implicitly has the same preferences over their children's well being.

I do not include children in the intra-household bargaining process. Although it is possible to model a household with more than two individuals (see Browning and Chiappori

[21] I define marriage markets based on the individual's age and race to ensure identification.

(1998)), doing so would require estimating additional sharing parameters. As DFs for children are not obvious, these additional sharing parameters are not likely to be identified. Additionally, excluding children from the household bargaining process greatly simplifies the empirical model, as families with children differ from their childless counterparts only in their observable characteristics, not in the specification of their model.

# Chapter 4

# Data

This chapter outlines the main data sources I use and how they are linked. From Equation 3.2 for single person families and Equation 3.3 for cohabiting couple families, it can be seen that to estimate my model I need to observe three outcomes: family housing choice ($h_i$), individual commute method ($k_i^s$), and individual commute time ($t_{ihk}^s$). I also need data on housing characteristics ($H_{ih}$), commute characteristics ($K_{ihk}^s$), and individual characteristics ($X_i^s$). Additionally, in order to recover composite consumption, I need data on the prices of composite consumption ($p_m^c$), homes ($p^H$) and commuting methods ($p_m^d$). No single dataset contains all of this information. In order to construct a dataset that allows me to estimate my model, I combine data from the U.S. Census Bureau's ACS and the U.S. Department of Transportation's National Transportation Atlas Database (NTAD) using GIS mapping software. I also augment that dataset with pricing information from various sources.

The main dataset my analysis is built on is the restricted-access Census microdata versions of the 2000 - 2009 ACS. *However, as I have not yet gained clearance to disclose my analysis using restricted access data, all analysis in this draft is preliminary and based on the 2005 - 2008 ACS PUMS data.*[1] The ACS contains information similar to the Decennial Census Long Form Questionnaire that it replaced after the 2000 Census. It is an annual

[1] I have been approved for Special Sworn Status (SSS) and have gained access to the restricted Census microdata. I travel to the Triangle Census Research Data Center (TCRDC) at Duke University to conduct my estimation. I discuss the shortcomings of the publicly available data in Section 4.4.

sample of one in 40 households in the country.[2] The Census Bureau first began producing ACS data in 2000 to test the survey and officially began producing the survey in 2005, so my data is a repeated cross-section.

There are two key features of ACS data that are important for my research. First, ACS surveys include questions on place of residence, primary commuting method, and commuting duration, as well as individual and household characteristics and linkages that allow the identification of the relationship between members of a household. Second, while the ACS does not contain a great deal of information about individual commute characteristics, it does report the daily commute time and includes information about the place of residence and place of work that allow me to augment the commuting data. The restricted versions of these datasets allow me to identify both the home and work locations of each individual down to the Census block, which provides the geographic precision necessary to calculate unobserved commute characteristics in a meaningful way. I detail the former feature first, then provide more detail on the latter in subsequent sections. I conclude the section by discussing the additional price data I use.

## 4.1 Sample Selection

I begin by defining markets ($m$) using the the Office of Management and Budget's (OMB) definitions of metropolitan areas. The OMB creates these designations for use by federal agencies in statistical analysis. Metropolitan areas are defined as central urban areas and any adjacent counties that have "a high degree of social and economic integration (as measured by commuting to work) with the urban core."[3] The OMB defines Combined Statistical Areas (CSA) to represent contiguous urban areas (ie, Washington, DC and Baltimore, MD) and Core Based Statistical Areas (CBSA) to represent central (ie, Pittsburgh,

[2] For reference, every decade the Long Form sampled one in 6 households. See http://www.census.gov/acs/www/Downloads/handbooks/ACSPUMS.pdf for more information.

[3] See http://www.census.gov/population/metro/ for more information.

PA) or component (ie, Washington, DC) cities.[4]

I restrict all data to the "Washington-Arlington-Alexandria, DC-VA-MD-WV" CBSA (hereafter refereed to as the DC CBSA) in order to keep the estimation tractable.[5] I use this definition of the market for all years of the data even though it was created in 2003 to avoid using a varying definition of the market each year. I choose this CBSA for several reasons. First, it has the second most automotive commuter congestion in the nation according to Schrank and Lomax (2010), so there is a need for the policy analysis I perform. Second, there is a robust mass transit system in the Washington, DC area that allows individuals to respond to a given policy change in multiple ways. This both increases the need for the simulations I perform and allows for the analysis of numerous policy options. Finally, the District offers a great deal of geographic information that is not available nationally which is accessible through the District of Columbia Geographic Information System (DC GIS). Specifically, although the NTAD contains geographic location information for rail systems, it does not have comparable information for bus routes that DC GIS makes available for Washington Metropolitan Area Transit Authority (WMATA) bus lines and stops.

I restrict the sample based on observable characteristics at both the household and individual levels. First, I drop households based on housing unit characteristics that indicate that the residence may not be the family's primary home or that the full financial costs of the home are not accurately reported. Second, I restrict the sample based on household characteristics that indicate that the family's income is in the tails of the income distribution or based on relationships in the household that indicate that the household bargaining

[4]In 2003, the OMB updated the names and definitions of core metropolitan areas, creating, amongst others, the CBSA geography. OMB frequently refers to CBSAs as Metropolitan Statistical Areas (MSA), but since MSAs were defined differently prior to 2003, I use the CBSA moniker for clarity. For a thorough explanation of the changes, see the Missouri Census Data Center website (http://mcdc.missouri.edu/allabout/sumlevs/).

[5]I plan to expand the analysis to include additional metropolitan areas in future work. This will allow me to include measures of commuting and congestion that vary across metropolitan areas in estimation. Ideally, I would define the single market as the the Washington-Baltimore-Northern Virginia, DC-MD-VA-WV CSA, as the Washington, DC and Baltimore, MD residential and labor markets are undoubtedly linked. Doing so would drastically increase the number of Census blocks in the market and the scope of the GIS network analysis (that requires calculating the optimal route between all pairwise combinations of blocks in the market). It is not feasible at this time.

process is too complex to model without additional information. Third, based on individual characteristics, I drop all households that contain an unemployed, military, or part-time employed adult.[6] I also drop households that contain an adult whose job location information is missing or indicates that the individual works outside the geographic scope of the market. Next, I drop households with individuals who commute by methods that are either unavailable in the market (streetcar), occur too infrequently in the data to be modeled as outcomes (bike, commuter rail, ferry, taxi, motorcycle, other), or are beyond the scope of the model (working at home). Finally, I drop individuals who travel for an extremely long time or who cover an implausibly long distance as part of their commute.[7] As I do not observe precise location information in the PUMS data, I do not drop individuals based on commute distances outside the TCRDC.

The percent of the sample dropped for each specific reason is detailed in Table 3. Regardless of whether the reason for the drop is a household or individual level characteristic, I drop the entire household. Column (1) contains the percent of households dropped for the given reason, and column (2) contains the analogous percent of individual dropped. Sample selection results in 5,068 households and 10,731 individuals in the PUMS data. Of those 10,731 individuals, 7,650 are adults. This forms the basis of my sample.

## 4.2 Choice Set

After dropping individuals who commute by unavailable or infrequently observed methods, I model $N^K = 5$ commuting options in DC CBSA: automobile, carpool, Metrorail, Metrobus, and walking. A key shortcoming of the ACS commuting data is that it reports only the primary method of travel one uses to commute, so I treat individuals in the model as if they do not commute by multiple modes.[8]

Table 4 shows the distribution how individuals in the data commute before and after

[6]The "head of household in school" reason in the summary table indicates households where the head of household is enrolled in grade school or a lower grade.

[7]I define such a commute as one with a duration or distance greater than the 99th percentile.

[8]See Appendix A.1.1 for a description of this issue and what I do to mitigate the problem.

Table 3: Percent of Sample Dropped by Reason

| Variable | (1) Percent of Households | (2) Percent of Individuals |
|---|---|---|
| Housing Unit Characteristics ($H_{ih}$) | | |
| Vacant house | 0.045 | 0.000 |
| Mobile home or RV | 0.012 | 0.010 |
| No cash rent | 0.011 | 0.010 |
| Meals included in rent | 0.006 | 0.003 |
| Household Characteristics ($X_i$) | | |
| Net of housing exp. income tails | 0.257 | 0.209 |
| Subfamilies in household | 0.111 | 0.111 |
| Roomate present | 0.026 | 0.015 |
| Under 18 non-children | 0.001 | 0.000 |
| Adult children | 0.118 | 0.060 |
| Child primary wage earner | 0.019 | 0.008 |
| Individual Characteristics ($X_i$) | | |
| Unemployed or not at work | 0.226 | 0.130 |
| Military employment | 0.018 | 0.008 |
| Not full time and year emp. | 0.310 | 0.149 |
| Head of household in school | 0.001 | 0.000 |
| Job location missing | 0.030 | 0.014 |
| Job location outside market | 0.617 | 0.385 |
| Commute Modes ($k$) | | |
| Commute by streetcar | 0.001 | 0.000 |
| Commute by bike | 0.005 | 0.002 |
| Commute by commuter rail | 0.008 | 0.004 |
| Commute by other method | 0.011 | 0.005 |
| Work at home | 0.052 | 0.024 |
| Commute Characteristics ($K_{ihk}$) | | |
| Commute duration in tail | 0.008 | 0.003 |
| HR station (home) distance in tail | 0.000 | 0.000 |
| HR station (job) distance in tail | 0.000 | 0.000 |
| Walk distance in tail | 0.000 | 0.000 |
| Total Observations | 89,110 | 213,870 |
| Selected Observations | 5,068 | 10,731 |

sample selection. Columns (1) and (2) are calculated from the full sample. Column (1)

Table 4: Percent of Commuters by Mode ($k$)

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | All Observations | | Selected Sample | | Random Sample | |
| Variable | Percent | SD | Percent | SD | Percent | SD |
| Automobile | 0.486 | 0.500 | 0.686 | 0.464 | 0.713 | 0.452 |
| Carpool | 0.072 | 0.258 | 0.078 | 0.267 | 0.077 | 0.267 |
| Heavy rail (Metrorail) | 0.053 | 0.223 | 0.035 | 0.183 | 0.024 | 0.153 |
| Bus (Metrobus) | 0.025 | 0.155 | 0.018 | 0.131 | 0.011 | 0.106 |
| Walking | 0.016 | 0.126 | 0.020 | 0.140 | 0.018 | 0.133 |
| Other or not in LF | 0.349 | 0.477 | 0.164 | 0.370 | 0.157 | 0.364 |
| Observations | 133,127 | | 7,650 | | 3,003 | |

contains the percent of individuals who commute by the given method, and column (2) contains the standard deviation. Columns (3) and (4) contain the analogous figures for the selected sample. Households in the selected sample are about 20 percentage points more likely to commute by automobile, most likely because of the income selection criteria and the employment and commuting requirements that shift individuals out of the "other or not in labor force" category. That category includes the five unmodeled commute modes listed in Table 3 and spouses who are not in the labor force.

I define the choice set in my discrete choice model as the $N^K = 5$ commuting options available in the DC CBSA and the $N^H = 5,068$ homes observed in the data. This means that there are potentially $N^K \times N^H = 25,340$ options in an individual's choice set.[9] As Vega and Reynolds-Feighan (2009) explain, the econometrician needs to limit the size of the choice set when dealing with a large number of housing alternatives. I address the issue of choice set size in my model by randomly sampling to reduce the the number of households included in the sample (and thus individuals and housing options as well). Specifically, I randomly select 2,000 households from the sample to form a new selected, random sam-

[9] I say, "potentially" because I allow individuals in the model to commute by automobile only if they own an automobile and limit housing options to homes that individuals can afford.

ple.[10] Doing so allows me to avoid arbitrary spatial aggregation of housing alternatives that would reduce the precision with which I am able to map commuting alternatives.

Columns (5) and (6) pertain to the selected, random sample. As one would expect, random sampling does not change the distribution of commuting methods. After random sampling, there are 1,990 households and 3,003 adult individuals in the sample.

## 4.3 Summary Statistics

Table 5 shows the distribution of housing characteristics before and after sample and random selection. Again, columns (1) and (2) are calculated from the full sample, columns

Table 5: Moments of Housing Characteristics ($H_{ih}$)

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | All Observations | | Selected Sample | | Random Sample | |
| Variable | Mean | SD | Mean | SD | Mean | SD |
| Single family detached | 0.549 | 0.498 | 0.577 | 0.494 | 0.578 | 0.494 |
| Single family attached | 0.185 | 0.388 | 0.178 | 0.383 | 0.183 | 0.387 |
| 2-9 apartments | 0.070 | 0.255 | 0.076 | 0.265 | 0.077 | 0.266 |
| 10-49 apartments | 0.097 | 0.295 | 0.096 | 0.295 | 0.097 | 0.297 |
| 50+ apartments | 0.089 | 0.285 | 0.072 | 0.259 | 0.064 | 0.244 |
| Number of rooms | 6.328 | 2.167 | 6.046 | 2.036 | 6.112 | 1.968 |
| Property age | 33.474 | 19.440 | 29.035 | 19.154 | 28.265 | 18.956 |
| Observations | 85,080 | | 5,095 | | 1,990 | |

(3) and (4) are based on the selected sample, and columns (5) and (6) pertain to the selected, random sample. The selected sample does not differ greatly from the full sample, and again, random sampling does not appear to affect observable characteristics. The first five estimates describe what type of building the home is. The majority of homes in the sample are single family detached homes, as 58 percent of the households in the selected,

[10] As I have access to more observations when using the restricted-access version of the ACS, I randomly select 10,000 households to estimate my model at the TCRDC.

random sample are of that type. Homes in the samples have an average of about six rooms and average about 30 years in age.

Table 6 contains the analogous moments for key individual and household characteristics. The sample is slightly more female than male, and 68 percent of the selected, random

Table 6: Moments of Individual and Household Characteristics ($X_i$)

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | All Observations | | Selected Sample | | Random Sample | |
| Variable | Mean | SD | Mean | SD | Mean | SD |
| Individual Characteristics | | | | | | |
| Male | 0.463 | 0.499 | 0.484 | 0.500 | 0.491 | 0.500 |
| Individual's age | 49.143 | 15.215 | 44.769 | 11.994 | 44.535 | 11.707 |
| Spouse or partner | 0.722 | 0.448 | 0.668 | 0.471 | 0.675 | 0.469 |
| Bachelor's degree (+) | 0.511 | 0.500 | 0.434 | 0.496 | 0.432 | 0.495 |
| Government employee | 0.237 | 0.425 | 0.260 | 0.439 | 0.254 | 0.436 |
| Self employed | 0.091 | 0.287 | 0.064 | 0.244 | 0.071 | 0.256 |
| Household Characteristics | | | | | | |
| Owner occupied | 0.781 | 0.414 | 0.780 | 0.414 | 0.773 | 0.419 |
| Tenure in home | 10.011 | 9.464 | 8.636 | 8.405 | 8.551 | 8.242 |
| Child in home | 0.356 | 0.479 | 0.383 | 0.486 | 0.391 | 0.488 |
| Number of children | 0.649 | 1.022 | 0.709 | 1.053 | 0.711 | 1.034 |
| Number of vehicles | 2.003 | 1.082 | 2.009 | 0.955 | 2.036 | 0.911 |
| Observations | 133,127 | | 7,650 | | 3,003 | |

sample is married or cohabiting with a partner. About a quarter of the sample works for the government, which is unsurprising given the market. Individuals in the selected, random sample average 45 years in age, have lived in their home for just under 9 years, have 0.7 children living in the home, and own 2 cars.

## 4.4 Census Geography Background

Before explaining how I augment the ACS data with characteristics of the commute using GIS, I first provide detail on the Census Geography that forms the basis for the procedure. Census geography is complex because it deals with geographic entities that are both determined by legal boundaries that the Bureau does not control (counties, congressional districts, school districts, etc.) and Census defined summary areas (Census Blocks, PUMAs, etc.) that are used to report statistics at varying levels of aggregation. These geographic entities range in size from the Census block, which is the lowest level of Census geography, to the nation as a whole.[11] In ascending order of size, the geographic entities that are relevant for my analysis are: Census blocks, block groups, tracts, and Public Use Microdata Areas (PUMAs).

Census blocks are defined to sever as the building blocks of all other Census geographies and all land in the United States is assigned to a Census block. They are bounded on all sides by physical features (such as roads or streams) or invisible boundaries (such as city or county limits). They are generally geographically small, but can be large in unpopulated areas. Census blocks are clustered into slightly larger block groups, which in turn are clustered into Census tracts. Tracts are created to contain 4,000 individuals, although they range in size from 1,500 to 8,000 people nationally. They are defined to provide a consistent geographic unit for the Census to use to present aggregate statistics. Finally, PUMAs are areas defined to contain at least 100,000 people and are so created to ensure confidentiality in individual level data.

Table 7 contains the number of these geographies that fall inside the boundaries of the DC CBSA and their mean size.[12] Figure 1 presents the same information visually.

Individual data is not publicly available at the tract level or below. The smallest geographic identifier in publicly available Census microdata is the PUMA, which averages 134 square miles in the DC CBSA. The restricted-access version of the ACS contains geo-

[11] I again refer the interested reader to the Missouri Census Data Center website for more detail on this topic (http://mcdc.missouri.edu/allabout/sumlevs/).

[12] I also include states in the table for reference.

Table 7: Geographies in the DC CBSA

| Variable | (1) Count Sum | (2) Size Mean |
|---|---|---|
| DC CBSA | 1 | 6,030.347 |
| States | 4 | 1,507.587 |
| Census PUMAs | 45 | 134.008 |
| Census Tracts | 1,040 | 5.798 |
| Census Block Groups | 2,979 | 2.024 |
| Census Blocks | 51,972 | 0.116 |

Source: Author's calculations.
Note: The unit of measurement for size is square miles.

graphic information down to the block level. At an average of 0.12 square miles in size in the DC CBSA, Census blocks allow for much greater geographic precision in mapping the locations of individual residential and job locations. This precision is particularly important when mapping locations relative to the commuting infrastructure (such as highways or Metro stations) in the market. Attempting to map the commute between areas that are 134 square miles in size would be an imprecise exercise at best and an impossible exercise for individuals who live and work in the same PUMA. Thus, the geographic precision available in the restricted-access ACS data is essential for creating the GIS data that as accurately as possible approximates the characteristics of both observed and unobserved commutes.[13]

## 4.5 GIS Data Calculation

To perform the GIS network analysis that allows me to calculate the optimal route between a home and job location pair, I begin by constructing a digital representation of the Census

[13]In order to replicate the conditions in the RDC based on the PUMS data for preliminary analysis, I randomly assign households to a population weighted residential block location within their reported PUMA. The residential population weights are based on available block level aggregate population counts. I also randomly assign individuals a job block location within their reported PUMA, but analogous employment density weights are not readily available.

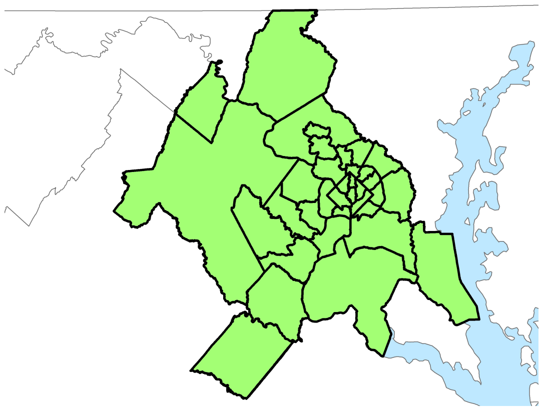

Figure 1a: Census PUMAS

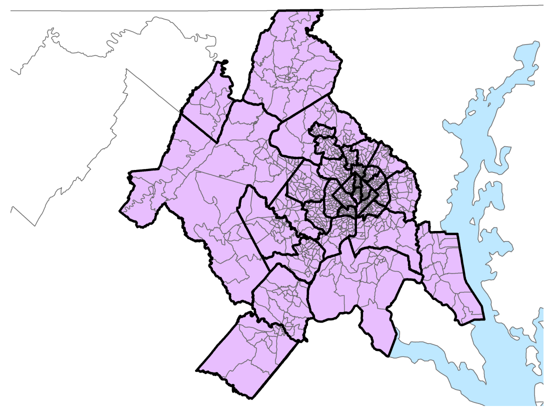

Figure 1b: Census Tracts

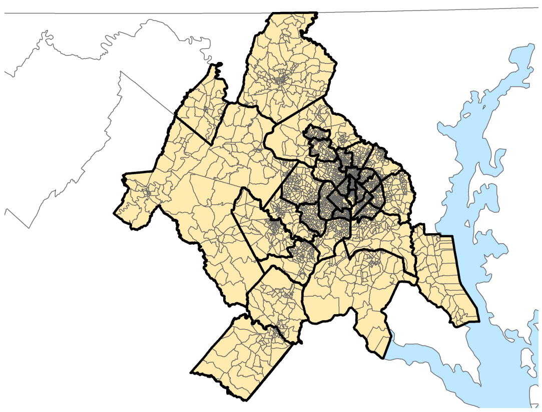

Figure 1c: Census Block Groups

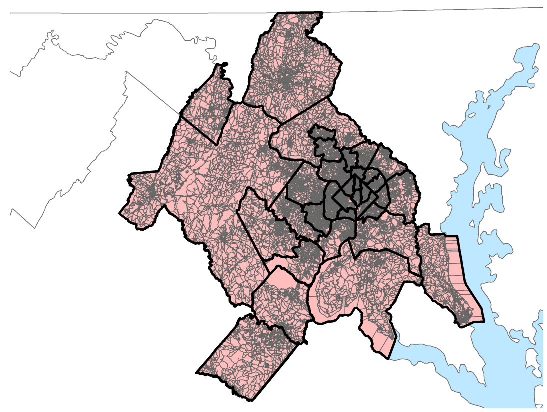

Figure 1d: Census Blocks

Figure 1: Geographies in the DC CBSA

geography. I use the 2009 definition of the CBSA from the Census TIGER/Line® shapefiles to define the market. Since blocks, block groups, and tract definitions are updated every Decennial Census, I use ESRI ArcGIS software to keep all of the 2000 definition of the Census TIGER/Line® block, block group, and tract shapefiles that fall within the boundaries of the DC CBSA. I use block centroids to approximate the exact home or job location. Next, I overlay a street network and a rail network on the Census geographies.[14] The street network data is obtained from ESRI's Data & Maps 9.3 (StreetMap North America). The rail network is created from the locations of rail stations and lines available in the National Transportation Atlas Database (NTAD) for both heavy rail (Metrorail) and com-

[14]I have not yet created the bus network, but plan to do so using bus station and line information from DC GIS. Currently, I use calculations from the automotive network as a proxy for bus commute characteristics.

muter rail (MTA and VRE). Both sources are updated infrequently, so I use one version of the network as the basis for the analysis, as opposed to creating multiple, data year specific networks.[15]

For each job location in the CBSA, I calculate the optimal route from that job location to every home location by every commuting method.[16] I do not observe transfers in the data, so I only need commutes by each given method, not the optimal combination of the methods. Optimal routes are calculated using the ArcGIS OD Cost Matrix Solver, which uses a version of Dijkstra's Shortest Path Tree algorithm to search for the lowest time cost route on a network between two points. The algorithm simultaneously solves forward from the origin and backwards from the destination (in a hierarchical fashion for roads) until the two paths meet.[17] The optimization takes into account turns, stops, and speed limits for automobile travel and stops, transfers at defined hubs, and average speeds for rail travel.[18]

For the calculated optimal route by road travel methods between the home and job locations, I am able to calculate the distance traveled on the network and the predicted travel time if one travels the speed limit. These distances and times should be thought of as similar to the ones an individual would recover from an online mapping website or a GPS, so it is important to note that they do not account for congestion.[19] For rail travel methods,

[15] The street network is based on 2003 TeleAtlas data. Heavy rail (subway) information comes from the 2004 Fixed-Guideway Transit Network database created by the University of Tennessee Center for Transportation Research GIS Group. The commuter rail network is created from data compiled by the Research and Innovative Technology Administration's Bureau of Transportation Statistics (RITA/BTS) for the 2009 NTAD. All sources used were the most current data available at the time of the construction of the network.

[16] Since I do not observe actual home and job locations outside the RDC, and GIS capabilities are limited in the RDC, I have to calculate the routes between all pairwise combinations of locations outside the RDC and import the resulting data. Doing so using Census blocks would require calculating $\frac{51{,}972\times(51{,}972+1)}{2} \approx 1.35\,billion$ automotive routes. This is beyond the GUI capabilities of ArcGIS, but can be accomplished by writing a Python script that accesses the GIS processor and loops over locations. To reduce the dimension of the computational burden and the size of the data, I take advantage of the fact that some block groups and tracts are very large in geographic size relative to their component blocks, while other block groups and tracts are not much larger than their component blocks. I develop algorithm that selects the largest Census geography (block/block group/tract) that will give a reasonably precise measure of location in order to balance computational burden and data size against precision.

[17] See Houde (2012) for technical details of how the algorithm works.

[18] Speed limit information is contained in the street network data. Average rail speeds are approximated based on the author's calculations from Metrorail, MTA, and VRE schedules.

[19] Although possible, I do not repeatedly query an online mapping website and record the resulting data. Small scale experiments with such a process using Google Maps were slower than using GIS network anal-

I calculate the analogous distance and the travel time if one travels the average speed. As discussed in Appendix A.1.1, to control for the fact that the ACS only reports the primary method of travel, I also calculate the as-the-crow-flies distance from both home and job locations to the nearest rail station. Finally, for individuals who commute by walking, I also calculate the as-the-crow-flies distance between locations to provide information about the characteristics of their commute, as there is no geographic network applicable to pedestrians.

Table 8 shows the distribution of key ACS and GIS commuting characteristics before and after sample and random selection. All commute times are in hours per week. The

Table 8: Moments of Commute Characteristics

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | All Observations | | Selected Sample | | Random Sample | |
| Variable | Mean | SD | Mean | SD | Mean | SD |
| ACS Commute Characteristics ($t_{ihk}$) | | | | | | |
| Commute time | 5.369 | 4.109 | 4.890 | 3.773 | 4.716 | 3.663 |
| Auto time | 5.239 | 3.741 | 4.622 | 3.631 | 4.483 | 3.517 |
| Carpool time | 6.279 | 4.053 | 6.111 | 3.642 | 6.040 | 3.634 |
| Metrorail time | 7.607 | 3.453 | 7.684 | 3.534 | 8.148 | 3.915 |
| Metrobus time | 8.082 | 4.848 | 8.194 | 5.357 | 8.333 | 4.859 |
| Walk time | 2.259 | 2.128 | 1.611 | 1.666 | 1.432 | 1.873 |
| GIS Commute Characteristics ($K_{ihk}$) | | | | | | |
| GIS auto commute time | 5.995 | 4.528 | 6.101 | 4.705 | 6.033 | 4.503 |
| GIS Metrorail commute time | 2.294 | 3.214 | 2.247 | 3.211 | 2.116 | 3.201 |
| Metrorail station to home (mi) | 10.807 | 13.013 | 21.508 | 15.272 | 22.245 | 15.298 |
| Metrorail station to job (mi) | 22.329 | 17.527 | 22.980 | 17.608 | 23.656 | 17.051 |
| ATCF distance (mi/week) | 152.570 | 129.872 | 154.147 | 132.429 | 151.551 | 127.704 |
| Observations | 133,127 | | 7,650 | | 3,003 | |

Notes: The unit of measurement for time is hours/week.

average commute time reported in the ACS selected, random sample is 4.7 hours per week. This is similar to the average automotive commuting time of 4.5 hours per week. Com-

ysis, and a mass download of the amount of data I would need would require prior approval from Google to avoid violating their Terms of Service.

muting by carpool results in a longer commute of 6.0 hours per week on average, as would be expected. Commuting by mass transit results in an average commute of a little over 8 hours per week, while walkers have the shortest commutes, on average, likely due to the fact that only those with short distances to travel can plausibly walk to work. The average automotive commute times predicted by the GIS network analysis are about an hour and a half longer than the reported times, but the average Metrorail commute times are much shorter. This is likely due to multi-modal commuters who report their total commute time, but only their primary means of travel. I control for the presence of multi-modal commuters in the model with the distance from the home and work location to the closest transit station. Home and job locations average just over 20 miles from a Metrorail station. Commuters travel an average of about 150 miles per week, implying an average speed across all methods of just over 32 MPH.

Table 9 presents levels-on-levels Ordinary Least Squares (OLS) regressions of reported ACS commute times on predicted GIS commute characteristics, conditional on traveling by the given commuting method. The independent variables are the GIS network time for all modes, save for walking. I use the as-the-crow-flies distance between the home and work locations to inform pedestrian commute times. These simple regressions show that the calculated commute times are all positive, significant predictors of the commute times reported in the ACS. The exception is for walking, likely due to geographic imprecision in the PUMS data. Pedestrians have the shortest commutes, so they are most sensitive to this imprecision, making this result unsurprising. The lack of geographic precision is likely also affecting the controls for distance to the nearest Metrorail station, although it is reassuring that the more accurate distance to the population weighted home location has the correct sign.

Unfortunately, there is a great deal of the variation in commute time around the mean that I am not explaining, as evidenced by the low values of the $R^2$s. This is likely the result of three shortcomings of the estimation. I have already mentioned the first two: lack of geographic precision in the publicly available data and unreported multimodal commuting

Table 9: Baseline Linear Commute Time Regressions

| Variable | (1) Auto $t_{ihk}$ | (2) Carpool $t_{ihk}$ | (3) Metrorail $t_{ihk}$ | (4) Metrobus $t_{ihk}$ | (5) Walk $t_{ihk}$ |
|---|---|---|---|---|---|
| GIS time | 0.226*** | 0.223*** | 0.800*** | 0.630*** | |
| | (0.010) | (0.030) | (0.133) | (0.130) | |
| Metrorail station to home (mi) | | | 0.016 | | |
| | | | (0.080) | | |
| Metrorail station to job (mi) | | | -0.061 | | |
| | | | (0.049) | | |
| ATCF distance (mi/week) | | | | | -0.003** |
| | | | | | (0.002) |
| Constant | 3.205*** | 4.556*** | 5.056*** | 5.594*** | 1.851*** |
| | (0.079) | (0.254) | (0.459) | (0.687) | (0.176) |
| Observations | 5,251 | 593 | 265 | 134 | 152 |
| $R^2$ | 0.089 | 0.085 | 0.140 | 0.151 | 0.028 |

Notes: The unit of measurement for time is hours/week. Single-starred items are statistically significant at the 10 percent level, double-starred items are statistically significant at the 5 percent level, and triple-starred items are statistically significant at the 1 percent level. Standard errors are in parentheses.

methods. More importantly, I have not yet developed an appropriate measure of congestion to include in the model. There is a great deal of congestion in Washington, DC, so this is likely to affect the fit of the model. That the coefficient on Metrorail is much closer to one than the other coefficients supports this hypothesis, as congestion is much less likely to cause delays on subways that run on fixed schedules. Regardless of the deficiency, these regressions show that the GIS network analysis does a reasonable job of modeling commute characteristics.

## 4.6 Pricing Data

I also augment that dataset with pricing information from multiple data sources. Olsen et al. (2012) provides measures of the price of composite consumption ($p_m^c$) in the form of a price index for non-housing goods in the given year. Although I do not have variation in markets that would necessitate the use of this index, my data is a repeated cross section, so I include this measure to smooth variation in prices over time.

I construct the opportunity cost of living in each home ($p^H$) by modifying a procedure outlined in Bayer et al. (2005) and Bayer et al. (2007). The details of this procedure can be found in Appendix A.1.2.

Finally, for the per mile price of each commuting method ($p_{mk}^d$), I use data from the Energy Information Administration (EIA) for gas prices and the National Transportation Database (NTD) for average fares. Gas prices based on the average annual regular reformulated retail gas price in dollars per gallon for the lower Atlantic region. I calculate average fares from the NTD by dividing total annual fares collected by total annual passenger miles for the given mode.

# Chapter 5

# Estimation

I develop an original estimation approach that uses many of the tools described in Train (2009), and estimate my model with the Maximum Simulated Likelihood (MSL) methods of Geweke (1989), Hajivassiliou (1990), Keane (1994) (GHK), and Stern (1997). The individual likelihood contribution is the probability of observing the sample data given the parameters ($\theta$) of the model. Simulation is required to evaluate these probabilities because they contain multidimensional integrals over the joint distribution of the errors that cannot be evaluated analytically.

This chapter proceeds by first explaining the empirical specification for single individuals. Although I do not fully detail the empirical specification for cohabiting couples, the second section provides an overview of how the single person household empirical specification will be modified to account for cohabiting couples. I conclude by discussing how the parameters in the model are identified and the potential biases introduced by the assumptions I make.

## 5.1 Single Person Household Empirical Specification

For single person households, the likelihood contribution involves three dependent variables. Define the observed housing choice of family $i$ as $h$ and the observed commuting method of the head of family $i$ as $k$. Let $P_i = \Pr(h, k, t_{ihk} \mid \theta)$ denote the probability of observing family $i$ living in home $h$ and the head of family $i$ commuting by method $k$ for a duration of $t_{ihk}$ conditional on the parameters in the model. I proceed as follows: first, I define the structure of the errors in my theoretical model. Next, I show that $P_i$ can be decomposed into the product of the joint probability of observing a family living in house $h$ and commuting by method $k$ and the probability of commuting for a duration of $t_{ihk}$. I then detail the estimation routine for each factor separately. Finally, after explicitly defining each of those terms, I am able to write the likelihood function and its simulated analog. Note that, in this section, I drop the $m$ subscripts for notational convenience since I am currently only using the DC CBSA in estimation.

### 5.1.1 Error Structure

There are two types of errors in the single person family utility functions: an idiosyncratic error and unobserved heterogeneity terms. The former, $e_{ihk}$, accounts for the difference between the predicted and observed commute times. The latter comes in three forms: 1) $\mu_i$ and $u_i$ are the unobserved components of preferences for housing amenities and leisure time, 2) $\varepsilon_{ih}$ is the unobserved component of the value of house $h$, and 3) $\lambda_{ik}$ is the unobserved time value of commuting method $k$. These errors are assumed to be known to the agents but not the econometrician. I proceed by first discussing the idiosyncratic error term, then discussing the role the unobserved heterogeneity terms play in estimation.

#### 5.1.1.1 Idiosyncratic Error

Since $t_{ihk}$ is not a choice variable, but rather is determined by the choices of $h$ and $k$, then any deviation in the predicted $t_{ihk}$ from the true travel time is assumed to be idiosyncratic. I assume this error is distributed as $e_{ihk} \sim iidN(0, \sigma_e^2)$.[1] This adds a variance parameter to the model, $\sigma_e^2$.

[1] Although it simplifies simulation of the choice probabilities, the assumption that the $e_{ihk}$ are iid is not necessary for estimation. A more complex correlation structure can be accounted for with a GHK simulator.

#### 5.1.1.2 Unobserved Heterogeneity

Define each of the three types of unobserved heterogeneity errors and their distributions as $\vec{\mu}_i = (\mu_i, u_i) \sim N\left(0, \Omega^{\vec{\mu}}\right)$, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iN^H}) \sim N(0, \Omega^{\varepsilon})$, and $\lambda_i = (\lambda_{i1}, \ldots, \lambda_{iN^K}) \sim LN\left(0, \Omega^{\lambda}\right)$. The latter distribution is chosen to ensure that $\lambda_{ik}$ is bounded below at 0, as is required by the theoretical model. In order to both normalize the model and reduce the computational burden of the estimation routine while still retaining a rich set of covariance terms, I impose structure on $\Omega^{\varepsilon}$ and $\Omega^{\lambda}$. I do so by defining $\varepsilon_i^H$ and $\lambda_i$ as being functions of correlated and idiosyncratic components, in ways that still allow for substantial correlations across related choices. Specifically, I assume that unobserved preferences for homes are correlated for the same individual within and across neighborhoods, but not across homes themselves. Similarly, unobserved preferences for time spent commuting are correlated within commuting method classifications, but neither across classifications nor individual commuting methods. I explain these restrictions in greater detail in the subsequent paragraphs. I detail the specification of $\lambda_i$ first, as it is more straightforward.

Let $\tilde{\lambda}_{i\tilde{k}}$ be an error associated with traveling by commuting method category $\tilde{k}$ and $w_{ik}$ an idiosyncratic error associated with commute method $k$. Formally, assume that $\lambda_{ik} = \exp\left(\tilde{\lambda}_{i\tilde{k}} + w_{ik}\right)$, where $\tilde{\lambda}_{i\tilde{k}} \sim iidN\left(0, \sigma^2_{\tilde{\lambda}}\right)$ and $w_{ik} \sim iidN\left(0, \sigma^2_w\right)$.[2] I assume there are three commuting method categories: personal, mass transit, and other; with the "Car, Truck, or Van" and "Carpool" commuting methods belonging to the first category, the "Bus," "Streetcar," "Subway," and "Rail" commuting methods belonging to the second, and the "Other" category belonging to the last. The intuition behind these classifications is best explained with an example. Individuals who have a high taste for the convenience and flexibility of driving one's own automobile to work (for instance, the ability to park near one's origin and destination) are also likely to have a high taste for the relative convenience and flexibility of carpooling (the ability to park or be picked up and dropped off near one's origin and destination). This would be evidenced in the model by the fact that the errors

[2]Note that this preserves the log-normal distribution of $\lambda_i$ because the sum of two normally distributed random variables is normally distributed, and the exponent of a normally distributed random variable is log-normally distributed.

associated with "Car, Truck, or Van" and "Carpool" would be correlated through their common $\tilde{\lambda}_{i\tilde{k}}$ term. This specification sacrifices some flexibility, but still retains much of the important detail of the model and reduces the number of parameters in $\Omega^{\lambda}$ from 28 to 2.

Similar to $\tilde{k}$, let $\tilde{h}$ index neighborhoods and $\tilde{\varepsilon}_{i\tilde{h}}$ be the component of the error associated with neighborhood $\tilde{h}$. Assume that $\varepsilon_{ih} = \tilde{\varepsilon}_{i\tilde{h}} + \upsilon_{ih}$, where the first term is allowed to be correlated with other members of its group and the second term is idiosyncratic: $\tilde{\varepsilon}_i \sim N\left(0, \Omega^{\tilde{\varepsilon}}\right)$ and $\upsilon_{ih} \sim iidN\left(0, \sigma^2_{\upsilon}\right)$.[3] Previous studies have defined neighborhoods based on Census geography at either the Census Block, Block Group, or Census Tract level.[4] There are 51,972 Census Blocks, 2,979 Block Groups, and 1,040 Census Tracts in the DC CBSA. Estimation of the $\frac{N^{\tilde{H}}\left(N^{\tilde{H}}+1\right)}{2}$ elements in $\Omega^{\tilde{\varepsilon}}$ at even the Census Tract level is computationally infeasible. In order to allow for covariation in neighborhood unobservable characteristics in an estimable manner, I define the correlation between any two given neighborhoods as being a decaying function of the distance between those neighborhoods. The intuition behind this specification is that the unobservable characteristics of two neighborhoods that are one mile apart should be more closely correlated than the unobservable characteristics of two neighborhoods that are five miles apart, and beyond a threshold distance, there should be no correlation. This specification assumes that there is an underlying continuum of unobservable neighborhood characteristics that dies out as distance from the given neighborhood increases, as opposed to a discrete change in unobservable characteristics when one crosses from the given neighborhood to "the other side of the tracks." I define neighborhoods at the Census Tract level and let $\tilde{h}$ and $\tilde{j}$ index neighborhoods. I also define $\tilde{d}_{\tilde{h}\tilde{j}}$ to be the "as-the-crow-flies" distance between the given neighborhoods. I define a spline function that weights the correlation between the $\tilde{h}$th and $\tilde{j}$th neighborhoods as

[3]Since $N^H$ is large, assuming that $\upsilon_{ih} \sim iidN(0, \Omega^{\upsilon})$ would be intractable because it would mean that there are $N^H$ variance parameters to estimate.

[4]Bayer et al. (2008) find evidence of neighborhood effects in hiring networks at the Census block level. Bayer et al. (2004) uses Census block groups to define neighborhoods when examining racial segregation in housing markets. Bayer et al. (2007) uses school attendance zones, as well as including controls at both the Census block and block group levels. Kiel and Zabel (2008) find that multiple definitions of a neighborhood, including Census tracts, are jointly relevant in hedonic equations.

$$\begin{aligned}\delta_{\tilde{h}\tilde{j}}\left(\tilde{d}_{\tilde{h}\tilde{j}}\right) &= 1\left(\tilde{d}_{\tilde{h}\tilde{j}}=0\right)\delta_0+1\left(0<\tilde{d}_{\tilde{h}\tilde{j}}\leq 1\right)\tilde{d}_{\tilde{h}\tilde{j}}\delta_1 \\ &+ 1\left(1<\tilde{d}_{\tilde{h}\tilde{j}}\leq 3\right)\tilde{d}_{\tilde{h}\tilde{j}}\delta_2+1\left(3<\tilde{d}_{\tilde{h}\tilde{j}}\leq 5\right)\tilde{d}_{\tilde{h}\tilde{j}}\delta_3,\end{aligned}$$

where $1(\cdot)$ is an indicator function equal to 1 if the argument is true and 0 otherwise. Note that $\tilde{d}_{\tilde{h}\tilde{j}}=0$ implies that $\tilde{h}=\tilde{j}$. Since manipulating a $1040\times 1040$ matrix is computationally costly, I do not estimate $\Omega^{\tilde{\varepsilon}}$. Instead, I allow neighborhood unobservables to be correlated by calculating $\tilde{\varepsilon}_{i\tilde{h}}$ as a weighted sum of the standard normal errors associated with each neighborhood. By defining $\eta^{\tilde{\varepsilon}}_{i\tilde{j}}\sim iidN(0,1)$, the definition of $\tilde{\varepsilon}_{i\tilde{h}}$ can be stated formally as $\tilde{\varepsilon}_{i\tilde{h}}=\sum_{\tilde{j}=1}^{1040}\delta^{\tilde{\varepsilon}}_{\tilde{h}\tilde{j}}\eta^{\tilde{\varepsilon}}_{i\tilde{j}}$. This specification reduces the number of variance/covariance parameters in $\Omega^{\tilde{\varepsilon}}$ to be estimated to four (the elements of the vector $\delta$).

After specifying the errors in this way, there are five vectors of unobserved heterogeneity errors. To keep subsequent notation compact, I define a vector of the unobserved heterogeneity terms as $\xi_i=\left(\vec{\mu}_i,\tilde{\varepsilon}_i,\upsilon_i,\tilde{\lambda}_i,w_i\right)$. Let $\theta=\left\{\beta^H,\beta^\ell,\gamma^H,\gamma^K,\sigma_e,\Omega^{\vec{\mu}},\delta,\sigma_\upsilon,\sigma_{\tilde{\lambda}},\sigma_w\right\}$ be the full set of parameters to be estimated. After imposing structure on the errors in my model, I am able to reduce the total number of variance/covariance parameters to 11.

### 5.1.2 Joint Probability of Observing $t_{ihk}$, $h$, and $k$

Next I use the error structure to define the probability of observing the sample data. The probability of interest is $P_i=\Pr(h,k,t_{ihk}\mid\theta)$. Using the law of total probability, I write this probability as

$$P_i = \Pr(t_{ihk}\mid\theta)\Pr(h,k\mid\theta,e_{ihk}), \tag{5.1}$$

since $e_{ihk}$ is the only random component of $t_{ihk}$. The first factor is the probability of observing individual $i$ commuting for a duration of $t_{ihk}$, and the second is the probability of observing individual $i$ living in house $h$ and commuting by method $k$. For notational compactness, I define $\Pr(t_{ihk} \mid \theta) \equiv P_i^t$ and $\Pr(h,k \mid \theta, e_{ihk}) \equiv P_i^{HK}$.

#### 5.1.2.1 Probability of Observing $t_{ihk}$

Recall from Equation 3.1 that the observed commute time is a function of both a linear index of the characteristics of individual $i$'s commute and an error term. The probability that the individual's observed commuting time is equal to the commuting time the model predicts is the probability that this equality holds for the observed home and commuting method: $P_i^t = \Pr\left(t_{ihk} = \exp\left(K_{ihk}\gamma^K + e_{ihk}\right)\ for\, h\, and\, k\right)$. Explicitly, this is

$$P_i^t = \frac{1}{\sigma_e t_{ihk}} \phi\left(\frac{\ln(t_{ihk}) - K_{ihk}\gamma^K}{\sigma_e}\right),$$

where $\phi(\cdot)$ is the standard normal probability distribution function (PDF). Note that there is one such condition for the observed $h$ and $k$ for each individual, as I do not observe commute times for home and commuting method alternatives that individuals did not choose.

#### 5.1.2.2 Conditional Probability of Observing $h$ and $k$

I outline the empirical specification of $P_i^{HK}$ for an arbitrary $N^H$ homes and $N^K$ commuting options in individual $i$'s market. $P_i^{HK}$ is a statement about the joint probability that $N^H N^K - 1$ optimality conditions hold for each individual, conditional on $e_{ihk} = \ln(t_{ihk}) - K_{ihk}\gamma^K$. I proceed by first defining the optimality conditions as functions of the errors and the data. Then, I condition on a subset of the errors and recast the optimality conditions in a tractable form for estimation.

**Optimality Conditions** Observing an individual living in house $h$ and commuting by method $k$ implies that $U_{ihk} > U_{ih'k'}\ \forall\ (h',k') \neq (h,k)$.[5] After algebraic manipulation, it

[5] Optimality conditions can take one of three forms. Either the individual prefers the observed combination of $h$ and $k$ to

1. Living in another house and commuting by another method ($U_{ihk} > U_{ih'k'}\ \forall h' \neq h\, \&\, k' \neq k$),

can be shown that these conditions are equivalent to $e_{ih'k'} > f(\xi_i, D_{ih'k'}) \; \forall \; (h',k') \neq (h,k)$, where $f(\cdot)$ is defined to compactly represent the optimality condition as a function of the errors and data; $\xi_i$ was defined in Section 5.1.1.2 as the vector of unobserved heterogeneity terms; and $D_{ih'k'} = \{c_{ih'k'}, H_{ih'}, K_{ih'k'}\}$ is the set of data that varies over conditions and is used for notational convenience. Explicitly,[6]

$$\begin{aligned} f(\xi_i, D_{ih'k'}) &= \ln\Bigg(T - L_i \\ &- \frac{T - L_i - \exp\left(\tilde{\lambda}_{i\tilde{k}} + w_{ik}\right) t_{ihk}}{\exp\left(\frac{\ln\left(\frac{c_{ih'k'}}{c_{ihk}}\right) - \exp(X_i\beta^H + \mu_i)\left[\left(H_{ih} - H_{ih'}\right)\gamma^H + \tilde{\varepsilon}_{i\tilde{h}} + \upsilon_{ih} - \tilde{\varepsilon}_{i\tilde{h}'} - \upsilon_{ih'}\right]}{\exp\left(X_i\beta^\ell + u_i\right)}\right)}\Bigg) \\ &- K_{ih'k'}\gamma^K - \tilde{\lambda}_{i\tilde{k}'} - w_{ik'}. \end{aligned}$$

**Bounds of Integration** Evaluation of the joint probability that all of these optimality conditions holds cannot be accomplished analytically or numerically. Instead, I proceed by conditioning on $e_{ihk}$ and $\xi_i$ to make the problem tractable, then evaluating the multi-dimensional integrals that result using simulation methods. This requires determining the region over which each of the errors in $\xi_i$ are integrated. Placing bounds on the errors being integrated is necessary to avoid situations where draws of the simulated errors are such that no values of the remaining, unintegrated, errors (the $e_{ih'k'}$) are consistent with the data. Figure 2 provides a general, graphical representation of this situation, abstracted to two dimensions. The amorphous, shaded region depicts the values of the errors that are consistent with individual i choosing home h and commuting method k. The errors in $\xi_i$ must be drawn such that $B < \xi_i < A$, otherwise no value of $e_{ih'k'}$ is consistent with what is

2. Living in another house and commuting by the observed method ($U_{ihk} > U_{ih'k} \, \forall \, h' \neq h$), or
3. Living in the observed house and commuting by another method ($U_{ihk} > U_{ihk'} \, \forall \, k' \neq k$),

so the notation $(h',k') \neq (h,k)$ is equivalent to $h' \neq h$ and/or $k' \neq k$. Regardless of which of the three conditions is relevant for the given combination of $h'$ and $k'$, I can express the optimality condition as a function of $e_{ih'k'}$.

[6]Note that I replace $\varepsilon_{ih} = \tilde{\varepsilon}_{i\tilde{h}} + \upsilon_{ih}$ and $\lambda_{ik} = \exp\left(\tilde{\lambda}_{i\tilde{k}} + w_{ik}\right)$ when defining this function. Also note that I do not replace $t_{ihk}$ as it is observed directly in the data for the individual's chosen home and commuting method.

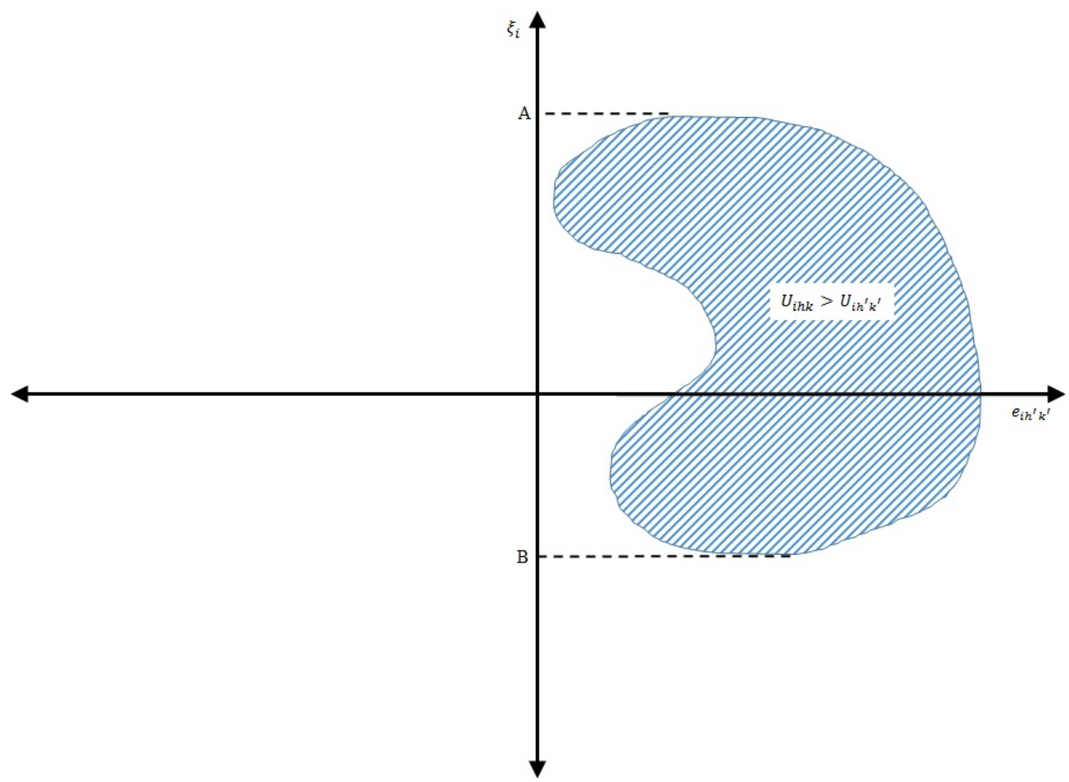


Figure 2: Intuition for Bounds of Integration

observed in the data.

Specifically, the the value of $e_{ihk}$ is fixed by the estimation of $P_i^t$ (by the relationship that $e_{ihk} = \ln(t_{ihk}) - K_{ihk}\gamma^K$), and all of the errors in $\xi_i$, save for $w_{ik}$, are integrated over their full distributions. Simulating $w_{ik}$ from an untruncated distribution in this manner would be problematic, as there are some feasible values of $w_{ik}$ for which no value of a given $e_{ih'k'}$ could explain the observed outcomes. This would occur when the random components in the leisure term associated with the observed choices ($\tilde{\ell}_{ihk}$) are such that $\tilde{\ell}_{ihk} \to 0$. Leisure enters the utility function as the argument of a natural log (see Equation 3.2), so as leisure goes to zero, the utility of the given choices goes to negative infinity. This is not a problem for an unobserved combination of a home and commuting method, as it ensures that $U_{ihk} > U_{ih'k'}$ for the given $(h',k') \neq (h,k)$, however, such a situation is a concern for the observed combination of choices because the model cannot explain an individual choosing options that result in a utility of negative infinity. To ensure that the the optimality conditions can be evaluated, it must be the case that leisure is positive for the observed choices. This necessitates a bound on $w_{ik}$. It can be shown that the condition that $\tilde{\ell}_{ihk} > 0\, for\, h\, and\, k$ is equivalent to the condition that $w_{ik} < B_i^w$ where

$$B_i^w = \ln(T - L_{\mathrm{i}}) - K_{ihk}\gamma^K - \tilde{\lambda}_{i\tilde{k}} - e_{ihk}\, for\, h\, and\, k.$$

Proof that bounding the other errors in $\xi_i$ is unnecessary is straightforward, as the $e_{ih'k'}$ errors enter the optimality condition linearly. Since the support of $e_{ih'k'}$ is the real line, any observed outcome can be justified by an $e_{ih'k'}$ in the appropriate range, so long as $f(\xi_i, D_{ih'k'})$ can be evaluated. The restriction that $w_{ik} < B_i^w$ ensures that the numerator of the second term in the argument of the natural log of $f(\xi_i, D_{ih'k'})$ is positive. The only concern, then, is that the denominator of the second term in the argument of the natural log in $f(\xi_i, D_{ih'k'})$ is such that

$$T - L_i - \frac{T - L_i - \exp\left(\tilde{\lambda}_{i\tilde{k}} + w_{ik}\right) t_{ihk}}{\exp\left(\frac{\ln\left(\frac{c_{ih'k'}}{c_{ihk}}\right) - \exp(X_i\beta^H + \mu_i)\left[\left(H_{ih} - H_{ih'}\right)\gamma^H + \tilde{\varepsilon}_{i\tilde{h}} + \upsilon_{ih} - \tilde{\varepsilon}_{i\tilde{h}'} - \upsilon_{ih'}\right]}{\exp\left(X_i\beta^\ell + u_i\right)}\right)} < 0.$$

This occurs as the denominator goes to zero $(\exp(\cdot) \to 0)$. With algebraic manipulation, it can be shown that this is a corner solution. It only occurs when $U_{ihk} > U_{ih'k'}$ for the given $(h', k') \neq (h, k)$ regardless of the value of $e_{ih'k'}$. Intuitively, this means the utility from composite consumption and housing amenities associated with the observed choice is great enough that that it doesn't matter how little time it takes to commute from the unobserved home and/or by the unobserved commuting method, the individual will always choose the observed combination. When this is the case, I do not need to calculate $f(\xi_i, D_{ih'k'})$ to evaluate the probability of the given optimality condition holding.

I define $\bar{\phi}(\xi_i)$ as the joint distribution of $\xi_i$ and $B_i$ as the upper bound on $\xi_i$. The overbar on $\bar{\phi}$ denotes that the distribution is truncated for some elements of $\xi_i$, namely the $w_{ik}$. Similarly, the bound on the joint distribution of $\xi_i$ is only binding for $w_{ik}$ so

$$B_i = \begin{cases} B_i^w & if\, \xi_{ij} = w_{ik} \\ \infty & else. \end{cases}$$

The probability that $\xi_i$ is less than $B_i$ is equal to the $\Pr(w_{ik} < B_i^w)$. I define this probability as $P_i^B$ where

$$P_i^B = \Phi\left(\frac{B_i^w}{\sigma_w}\right). \tag{5.2}$$

**Joint Probability** After integrating over the errors in $\xi_i$, the optimality conditions can be written in a form that is tractable for estimation. The probability of interest, $P_i^{HK}$, is the probability that individual $i$ chooses house $h$ and commuting method $k$. It is expressed as

$$P_i^{HK} = \Pr\left(f(\xi_i, D_{ih'k'}) < e_{ih'k'} \,\forall\, \left(h', k'\right) \neq (h, k) \mid e_{ihk}\right).$$

Using the law of total probability and the assumption that the $e_{ih'k'}$s are independent of both each other and the the other errors in $\xi_i$, $P_i^{HK}$ can be written as the product of $N^H N^K - 1$ conditional probabilities, so

$$P_i^{HK} = \prod_{(h',k')\neq(h,k)} \Pr\left(f\left(\xi_i, D_{ih'k'}\right) < e_{ih'k'} \mid e_{ihk}, \xi_i < B_i\right) P_i^B.$$

After integrating over the joint distribution of the errors in $\xi_i$, conditioning on $e_{ihk} = \ln\left(t_{ihk}\right) - K_{ihk}\gamma^K$, and replacing the remaining $e_{ih'k'}$s with their standard normal component according to the relationship $e_{ih'k'} = \sigma_e \eta_{ih'k'}$, I write the joint probability of observing a family living in house $h$ and commuting by method $k$ in a form that is tractable for estimation as

$$P_i^{HK} = \int_{-\infty}^{B_i} \prod_{(h',k')\neq(h,k)} \left[1 - \Phi\left(\frac{f\left(\xi_i, D_{ih'k'}\right)}{\sigma_e}\right)\right] P_i^B \bar{\phi}\left(\xi_i\right) d\xi_i. \tag{5.3}$$

### 5.1.3 Likelihood Function

Recall that $P_i = \Pr(h, k, t_{ihk} \mid \theta)$ is the probability of observing family $i$ living in home $h$, commuting by method $k$, and commuting for a duration of $t_{ihk}$ conditional on the parameters in the model, and $P_i = P_i^t P_i^{HK}$. Assume that there are $N$ total families in the data. The log likelihood function is

$$\ln L\left(\theta\right) = \sum_{i=1}^{N} \ln\left(P_i^t P_i^{HK}\right). \tag{5.4}$$

### 5.1.4 Simulation

Evaluation of the multidimensional integrals in $L\left(\theta\right)$ is not possible analytically or numerically, so I use a GHK simulator to evaluate the choice probabilities. Following Stern

(1997), I compute $N^R$ draws of $\xi_{ir}\varepsilon B_i$ from $\bar{\phi}(\xi_i)$. I define $P_{ir}^B = \Pr\left(w_{ikr} < B_i^w\left(\tilde{\lambda}_{i\tilde{k}r}\right)\right)$ as the simulated analog to $P_i^B$ (Equation 5.2). I replace the analytical likelihood contribution of $P_i^{HK}$ (Equation 5.3) with its unbiased simulated analog as

$$\begin{aligned} P_i^{HKR} &= \frac{1}{N^R}\sum_{r=1}^{N^R}\prod_{(h',k')\neq(h,k)}\left[1-\Phi\left(\frac{f\left(\xi_{ir},D_{ih'k'}\right)}{\sigma_e}\right)\right]P_{ir}^B \\ &= \frac{1}{N^R}\sum_{r=1}^{N^R}P_{ir}^{HKR}. \end{aligned}$$

The simulated likelihood function is

$$\ln L(\theta) = \sum_{i=1}^{N}\ln\left(\left[\frac{1}{N^R}\sum_{r=1}^{N^R}P_{ir}^{HKR}\right]P_i^t\right), \tag{5.5}$$

and estimation proceeds by MSL. I maximize the simulated likelihood function using the optimization routine outlined in Berndt, Hall, Hall, and Hausman (1974) which is commonly refereed to as the BHHH algorithm. I implement my estimation routine using the Fortran programing language and the OpenMP application programming interface. The latter allows me to use parallel processing to decrease the computational burden of the routine. [7]

## 5.2 Cohabiting Couple Household Empirical Specification

Work to define the empirical specification for cohabiting couples households is ongoing. The family likelihood contribution is the probability of observing a joint home location and individual commuting outcomes for each partner. This results in five dependent variables.

[7] As Fortran is not officially supported by the Census Bureau, I use the GNU Fortran (GFortran) compiler that is included in the Unix distribution that is run on the Census servers as part of the GNU Compiler Collection (GCC). As such, I do not have access to any commercial libraries of numerical analysis functions, so the bulk of my estimation routine was written from scratch.

Formally, let $P_i = \Pr(h, k^A, t_{ihk}^A, k^B, t_{ihk}^B \mid \theta)$ denote the probability of observing family $i$ living in home $h$, partner $A$ commuting by method $k^A$ for a duration of $t_{ihk}^A$, and partner $B$ commuting by method $k^B$ for a duration of $t_{ihk}^B$conditional on the parameters in the model.

The empirical specification builds on its analog for single individuals, explained in Section 5.1. I specify the aggregator function such that individuals have egoistic preferences (see Section 3.2.2), modify the error structure to allow for distinct, but correlated, commuting and preference errors for each partner, add the error associated with the sharing rule to the error structure, write the probability of observing the sample data based on this new error structure in a tractable form, and finally modify the GHK simulator accordingly. So long as the probability of interest can be written in a tractable form, this is a straightforward extension of the single individual household specification.

## 5.3 Identification

### 5.3.1 Why a structural model?

In thinking about identification, it is important to first explain the need for the econometric sophistication used in my model. Quite simply, this is due to the fact that a randomized, controlled experiment that would address my research question would be impossible to implement, and no natural experiment exists that would allow me to disentangle the separate effect each of the many factors and motivations in my model have on observed responses.[8] To do so, a natural experiment would have to impact individuals in such a way that their responses would be through one of the channels I am modeling, but no others. With such interconnected decisions as residential location and commuting method, both of which are influenced by a multitude of factors, such a natural experiment is hard to imagine.

[8] Being able to explain why individuals react the ways they do is important for predicting responses to the policies I am investigating. For instance, Baum-Snow and Kahn (2000) examine how much expansions to rail transit systems cause individuals to switch from other commuting methods to commuting by rail. They find evidence of an increase in rail transit use in areas near expansions, but they cannot determine what percentage of that increase is due to new riders and what percentage can be attributed to former rail commuters who moved from another location to take advantage of the infrastructure improvements.

While my structural model comes at a cost in terms of both implementation and understanding, it also yields important benefits. Estimation of preference parameters allows me to perform simulations that address a myriad of questions about the effects of proposed policies that have not yet been widely implemented. The model also allows for the extension of the collective model to a new arena in order to account for the fact that spouses behave differently than single individuals when making housing and commuting decisions. Addressing this issue directly would not be possible without a structural model.

### 5.3.2 Exclusion Restrictions

The main concern with a model of the joint decision of where to live and how to commute is that both decisions are made simultaneously. The housing location decision pins down where an individual is commuting from. Conversely, the availability and characteristics of commuting options are characteristics of the home themselves. In order to separately identify each effect without relying on functional form assumptions, I need at least one variable that exogenously affects each given decision alone. I use intrinsic, physical characteristics of the home that are observed in the data (e.g. number of rooms and property age) as an exclusion restriction to help identify the parameters pertaining to the commuting mode decision.[9] I do not allow an individual's commute to factor into the decision to purchase a home beyond its effect on leisure, so I exclude commute characteristics from the housing equation to identify the parameters relevant to the decision to purchase a home.

### 5.3.3 Identifying Variation

I encourage the reader to refer back to Equation 3.2, the individual's full choice problem for single person families, and Section 5.1.1, that details the model's error structure, while reading this section.[10] I begin by discussing the identification of the commute time pa-

[9] This is a valid restriction so long as the characteristics of homes vary with location. In other words, this restriction fails if one can buy an identical home in every location.

[10] As the empirical component of my intra-household model is still in progress, I discuss identification only in the context of single person families. Extension of the intuition in this section to cohabiting couple families

rameters, which are the coefficients in the commute time equations ($\gamma^K$) and the standard deviation of the commute time error ($\sigma_e$). The amount of time an individual reports taking to travel from her home to her job depends on the distance between the two locations and the speed the individual travels. The GIS commute characteristics I produce (GIS predicted times) are used to capture the effects of these factors on commute time. The commute characteristic parameters, $\gamma^K$, are identified by the covariation of commute characteristics ($K_{ihk}$) with the commute time the individual reports to the ACS ($t_{ihk}$). There is no guarantee that the individual will choose to travel the exact route mapped by the GIS algorithm, and even after conditioning on route, the characteristics are not perfect descriptors because of congestion, speeding, variation in mass transit schedules, and measurement error in the network data. This means that model-predicted commute times will deviate from the observed times. Variation in these deviations identifies the standard deviation of the commute time error, $\sigma_e$.

The error associated with commute time, $e_{ihk}$, is necessary, but not sufficient, to explain why individuals do not always commute by what the model determines is the optimal method. Individuals choose commutes based on considerations other than financial and time costs. Some individuals in the data choose to commute by a method that is more expensive, both in terms of money and time, than a given alternative. This can be explained by the individual having a high preference for the costlier method, be it because an automotive commuter enjoys listening to music in his car or a mass transit commuter enjoys reading the paper on the subway.[11] The method-specific $\lambda_{ik}$ error accounts for these preferences. It is separately identified from $e_{ihk}$ by the exclusion restriction that $e_{ihk}$ varies with homes, because of error in predicting commute-location-specific-times, but $\lambda_{ik}$ does not. The variance-covariance parameters in $\Omega^{\lambda}$ are identified as in other polychotomous discrete choice models (see Bunch, 1991). The intuition for the identification of these parameters is that if an individual does not choose the commuting method that results in the

applies similar logic.

[11]Note that it can also be explained by the individual having a low preference for the unchosen option because she finds driving on congested roads to be stressful or because she does not like to stuff herself into a crowded bus.

greatest utility according to the model (for the purposes of exposition, say automobile), then the unobserved preference for automotive commuting, $\lambda_{i,auto}$, must be such that it was not the best option ($\lambda_{i,auto}$ is large relative to other $\lambda_{ik}$). If when individuals do not select commuting by car, they frequently do not select another given method (say, carpool), then there is a positive correlation between the unobserved preference for those commuting methods. Alternatively, if individuals do frequently select carpool when the (hypothetical) model-predicted best option of automotive commuting is not chosen, then there is a negative correlation between commuting by car and by carpool.

Now I consider the parameters involved in the housing choice, $\gamma^H$ and $\Omega^\varepsilon$. Individuals choose a home based on its intrinsic characteristics (e.g. number of rooms), locational characteristics (e.g. proximity to mass transit), and cost. The covariation of observable housing characteristics and the observed housing choice identifies the $\gamma^H$ parameters. There are assuredly additional characteristics of the home that the econometrician does not observe. An individual may prefer an open floor plan and choose a large home with few rooms. An individual may select a home because it is close to family members (or select a home that is on the other side of town). The housing-specific error term, $\varepsilon_{ih}$, is necessary to explain cases where an individual selects a home that is observationally inferior to other homes in her feasible choice set. The variance-covariance parameters in $\Omega^\varepsilon$ are identified as in other polychotomous discrete choice models (again, see Bunch, 1991). The intuition in this case is similar to the intuition for identifying the parameters in $\Omega^\lambda$. If an individual does not choose the home with the highest observable quality ($H_{ih}\gamma^H$) she can afford, the unobserved preference for that home, $\varepsilon_{ih}$, must be such that it was not the best option ($\varepsilon_{ih}$ is negative or relatively small if positive). If when individuals do not select that home, they frequently also do not select another given home, then there is a positive correlation between the unobserved quality of those homes. If, on the other hand, individuals do frequently select the other given home, then there is negative correlation between the errors.

Finally, while the errors mentioned previously are necessary to explain deviations from the predicted optimal housing and commuting methods separately, the joint decision of housing and commuting method needs to be explained as well. The random preference parameters, $\alpha_i^c$, $\alpha_i^H$, and $\alpha_i^\ell$, are necessary to explain deviations from the predicted joint decision. As I am modeling a discrete choice, I must normalize one of the parameters, as the level and scale of utility are irrelevant. I do so by setting $\alpha_i^c = 1$, which addresses the issue and is equivalent to fixing one of the variance terms.[12] The remaining parameters account for the fact that even if two individuals value all homes and commutes the same, they may be observed living in two distinctly different homes and commuting by different methods. This would occur if they had different relative preferences for composite consumption, housing amenities, and leisure time. I provide intuition with three illustrative examples. In all, I assume that two individuals agree in their valuations of housing and commuting options, and they both commute to the same location.

1. Assume that these two individuals live in homes that are identical in every way, save location. The first lives in a home that is closer to their shared job location, so he has a shorter commute, but that commute is more financially costly than the commute taken by the second individual. The former has a greater preference for leisure relative to composite consumption than the later, so $\alpha_1^\ell > \alpha_2^\ell$.

2. Now assume that the two individuals are neighbors as well as coworkers, so they have identical commutes, both in terms of time and financial costs. If the first individual lives in a better, more expensive home than the second, then he prefers consumption of housing amenities to composite consumption, so $\alpha_1^H > \alpha_2^H$.

3. Finally, assume that these two individuals live in homes that are of equal cost and commute by methods of identical financial cost. The first lives in a downtown apartment that is close to their shared job location. The second lives in a suburban home that is farther from work, but has more housing amenities than the downtown apartment. The former has a greater preference for leisure relative to housing consumption than the later. This indicates that $\frac{\alpha_1^\ell}{\alpha_1^H} > \frac{\alpha_2^\ell}{\alpha_2^H}$.

[12] See Train (2009) for an excellent treatment of the subject.

Regardless of case, the covariation of the observable individual characteristics ($X_i$) and consumption of housing ($\tilde{H}_{ih}$) with housing and commuting outcomes identifies the $\beta^H$ parameters. Similarly, the covariation of the observable individual characteristics ($X_i$) and leisure ($\tilde{\ell}_{ihk}$) with outcomes identifies the $\beta^\ell$ parameters. The individual observables will not perfectly predict the preference parameters, hence the inclusion of error terms associated with the individual's preference for housing amenities ($\mu_i$) and leisure ($u_i$) in the model. Correlation between higher order moments of the deviations and higher order moments of the consumption of housing and leisure identifies the variance parameters in $\Omega^{\vec{\mu}}$.

### 5.3.4 Threats to Identification

Although this work advances the literature in several important ways, assumptions are necessary to keep the model tractable. As stated earlier, I assume that an individual takes her city of residence, family structure, vehicle ownership, and employment as given; the locations and hours of firms and schools are independent of residential choices and transportation options; there are no household production effects; and both members of a cohabiting couple have the same preferences for the well-being of their children. I discuss the potential bias that each of these assumptions introduces in the remainder of this section.

Defining the local residential market as closed at the metropolitan level is necessary to limit an individual's choice set when searching for a home. It has the potential to bias results to the degree that individuals select their city of residence based on characteristics of the residential or commuting markets in the city. For instance, if an individual chose to locate in a city because of a lack of congestion or the availability of a particular commuting option, my model would understate the preference that the individual has for those amenities. On the other hand, if employment opportunities alone drive the choice of city, then this source of bias might arise only if firms choose locations on the same basis, which I will not be able to model.

Excluding family structure decisions, such as marriage and fertility, are another possible source of bias. For example, Dettling and Kearney (2011) find that changes in house prices have differential effects on the birth rates of home owners and non-owners. If an individual decides to have children because their home location is more conducive to raising children, my model will overstate the impact of those children on the individual's value of the given housing amenities. A similar logic applies to the sign of the bias that children might cause on commuting amenities (e.g. a shorter or more flexible commute). These concerns are an interesting topic for future research.

Ignoring automobile ownership decisions is a more problematic assumption in the direct context of my model. My model removes commuting by car from the choice set of a household that does not own an automobile, but an individual who does not own a car may do so because she has a high distaste for commuting by car. My model will understate this individual's distaste for commuting by car, but explicitly modeling automotive ownership decisions is not supported by the available data. I observe very little about automobile ownership: only how many vehicles are available for use by members of the household. Fortunately, concern over this bias is mitigated by the fact that the automobile ownership rate is quite high: 87.4% of the families in my sample have at least one car per adult in the family.

Assuming that labor market decisions are exogenous also is not benign. In my model, I treat individuals as searching for a place to live subsequent to finding a job. However, the converse could also be true. This causes bias if, for example, an individual with a high distaste for commuting trades proximity to her home for wages when accepting employment. If so, my model would return a biased estimate of this individual's aversion to commuting, as it will explain some of the residential choice as a function of low wages preventing the individual from being able to afford a long commute, understating the individual's distaste for commuting. It is important to note that all of the residential choice studies I cite in Section 2.2 make a similar assumption. The alternative would require modeling job search behavior, which is not possible given the available data, as I observe only minimal characteristics of the individual's current job.

The assumption that the locations and hours of firms and schools are independent of my choice variables is implausible. Both are likely to locate in response to the distribution of residential housing and factor local commuting conditions into their decision of how to set their hours of operation. Again, I provide an example of how this might lead to bias. If firms locate close to neighborhoods where a critical mass of individuals reside, my model will overstate the aversion those workers have to commuting long distances. I justify this assumption similar to the justification for the assumption that the agents in a perfectly competitive market are price takers by assuming that any individual's choice of residence and method of commuting can neither influence where nor when firms and schools operate.

Assuming that there are no household production effects associated with commuting decisions is problematic if a long commute relative to that of an individual's partner results in an offsetting set of household production responsibilities. If that is the case, it would make long commutes relatively more desirable than they otherwise would be, and my model will understate aversion to commuting. This concern is beyond the scope of the available data, as there is no one source of data on individual residential, commuting, and time use decisions.

Finally, the assumption that parents have the same preferences for the well-being of their children is a possible source of bias, as there is evidence that parents do not have shared preferences over children. Perhaps the most well-known example is Lundberg et al. (1997), which shows that a policy shift that resulted in a change in the recipient of child benefits from husbands to wives resulted in a change in the demand for children's clothing. If parents have different preferences for their children, my model will produce biased estimates of the sharing rule. If the parent with a high taste for her children's well-being gives up some of her private consumption to attain better amenities for the children, then her sharing rule parameter would be biased down and her partner's (with a relatively low taste for their children's well-being) would be biased up. It is not clear whether the bias will be positive or negative, as it is not clear whether bargaining would result in the parent with the high preference gaining more for her children or being compensated with more personal consumption for allowing the children to less well-off. Blundell et al. (2005) show that the collective model can be used in instances when members of the household have different preferences over the public good, but this task is left for future research. While feasible from the point of view of the collective model, the requirements for estimation of my model are daunting. Partners having different preferences over the public good implies a different unobserved heterogeneity term associated with each house for each partner. This complicates the estimation routine and adds to the number of parameters to be estimated.

Despite these shortcomings, it is important to remember that that my model makes several key advances by jointly modeling residential choice and commuting method at the individual level in a way that allows for a rich heterogeneity structure and incorporates collective household decisions. I remind the reader that Langer and Winston (2008) propose a joint model of residential choice and commuting mode similar to the one I outline but opt for a different research design because "the data and modeling requirements of a disaggregate approach... are formidable."

# Chapter 6

# Results

This chapter summarizes the results from my estimated model. I proceed by first presenting the parameter estimates and standard errors. Since my model contains discrete outcomes, the parameter estimates cannot be interpreted as the effect of the explanatory variable on the outcome, so I also present accompanying marginal effects. The second section compares aggregate moments generated from the model with the true moments found in the data, and the third section more formally tests the model using several different specification tests. Finally, I discuss and perform policy simulations.

Although I have been able to estimate my model, I have not yet completed the calculation of all of the aforementioned results. Where necessary, I outline what will be calculated and presented as part of the published version of this research project.

## 6.1 Model Parameter Estimates

I present preliminary estimates from the model based on the PUMS data that lacks geographic precision in terms of home and job locations (see Section 4.4 for more details). I am in the process of calculating marginal effects, so these coefficients are estimates of utility parameters. They can be interpreted as affecting utility, but not the probability of choosing a particular home or commuting option. Although imperfect, these estimates are interesting, both at face value because they indicate how observables affect utility, and because they provide baseline against which to compare estimates based on restricted-access data. They are also useful as they provide evidence that the Fortran code that executes the estimation routine functions properly, both through examples of sensible values for parameters that should not be directly affected by the lack of geographic precision in the data and instances of insignificant values for coefficients that are based on noisy geographic data.

I estimate the model based on single individuals in the selected, random sample. There are 1,990 households in the selected, random sample, of which 973 are headed by single individuals.

### 6.1.1 Housing Consumption Parameter Estimates

Recall from Equation 3.2 and Section 5.1.1 that the housing parameters are $\gamma^H$, $\delta$, and $\sigma_\upsilon$. Estimates of these parameters and the associated standard errors are included in Table 10.

The first four parameter estimates describe the type of building the individual lives in. The baseline, omitted category is an apartment building with less than ten units. Surprisingly, single-family-attached homes are preferred to single-family-detached homes, given the negative coefficient on the former and the positive coefficient associated with the latter. This may be due to the lack of geography available in the data, as attached homes are more likely to be located in areas dense with other urban amenities like as restaurants, shops, etc. Attached homes are are preferred to apartments of all sizes, as evidenced by the positive, significant coefficient estimates of the second parameter and the negative coefficients for the third and fourth parameters. The estimates also indicate that smaller apartment buildings are preferred to buildings that contain more units. These are believable estimates, as are the results that that living in a home with more rooms increases the utility one gets from that home and that older homes decrease utility (although the property age topcode parameter indicates that this effect reverses in sign past some critical age).

Table 10: Housing Consumption Parameter Estimates

| Variable | Estimate | Std. Error |
|---|---|---|
| Housing Characteristic Parameters ($\gamma^H$) | | |
| Single-family home-detached | -0.190*** | 0.011 |
| Single-family home-attached | 0.053*** | 0.012 |
| 10-49 apartments | -0.335*** | 0.013 |
| 50+ apartments | -0.210*** | 0.007 |
| Number of rooms | 0.169*** | 0.005 |
| Property age | -1.812*** | 0.030 |
| Property age topcoded | 0.628*** | 0.046 |
| Second Moment Parameters ($\Omega^{\varepsilon}$) | | |
| Same neighborhood ($\delta_0$) | -0.093*** | 0.003 |
| Neighborhoods within 1 mile ($\delta_1$) | -0.025*** | 0.003 |
| Std. dev. of housing amenity error ($\sigma_\upsilon$) | 14.921*** | 0.036 |
| Observations | 973 | |

Notes: Single-starred items are statistically significant at the 10 percent level, double-starred items are statistically significant at the 5 percent level, and triple-starred items are statistically significant at the 1 percent level.

There are very small, negative effects for the $\delta$ parameters that govern how correlated the unobserved values of neighborhoods in close proximity to one another are. This is unsurprising when using PUMS data that is requires randomly assign homes to neighborhoods. The standard deviation of the idiosyncratic housing error is large relative to the parameter estimates and indicates the amount of variation in housing characteristics that is not explained by the observed characteristics. While adding geographic precision will surely help improve the fit of the model, I am also working to include more housing and neighborhood measures to address this issue. All parameters are precisely estimated based on the low values of the standard errors. This is a common feature of nonlinear models, and I am working to construct more informative measures of fit (see Sections 6.2 and 6.3).

### 6.1.2 Leisure Parameter Estimates

The commute time parameters are $\gamma^K$ and $\sigma_e$, and the commute mode preference parameters are $\sigma_{\tilde{\lambda}}$ and $\sigma_w$. Estimates of these parameters appear in Table 11. The first 15 parameters

Table 11: Leisure Parameter Estimates

| Variable | Estimate | Std. Error |
|---|---|---|
| Commute Characteristic Parameters ($\gamma^K$) | | |
| Auto GIS time | 0.267*** | 0.000 |
| Carpool constant | 3.718*** | 0.003 |
| Carpool GIS time | 0.118*** | 0.000 |
| Metrorail constant | 3.445*** | 0.004 |
| Metrorail GIS time | -0.026*** | 0.003 |
| Home: miles to closest Metro station | -0.078*** | 0.003 |
| Home: miles to closest Metro station squared | 1.919*** | 0.035 |
| Job: miles to closest Metro station | -0.025*** | 0.005 |
| Job: miles to closest Metro station squared | 1.956*** | 0.027 |
| Metrobus constant | 3.621*** | 0.009 |
| Metrobus GIS time | 0.184*** | 0.001 |
| Walk constant | 2.263*** | 0.005 |
| Walk distance | 0.328*** | 0.000 |
| Walk distance squared | -1.439*** | 0.004 |
| Second Moment Parameters ($\sigma_e$ and $\Omega^\lambda$) | | |
| Std. dev. of commute time measurement error ($\sigma_e$) | 1.225*** | 0.001 |
| Std. dev. of mode category preference error ($\sigma_{\tilde{\lambda}}$) | 0.000*** | 0.000 |
| Std. dev. of mode preference error ($\sigma_w$) | 0.000*** | 0.000 |
| Observations | 973 | |

Notes: Single-starred items are statistically significant at the 10 percent level, double-starred items are statistically significant at the 5 percent level, and triple-starred items are statistically significant at the 1 percent level.

are from the commute time equation. For each commuting method (save walking), I include a mode specific constant and the GIS predicted commute time. I normalize the automobile constant to 0 for identification. I control for multimodal commuters who report commuting by subway by including a quadratic function of the distance to the nearest Metro station from both the home and work location in the Metrorail commute time equation. Finally, as in Table 9, I use the as-the-crow-flies distance between an individual's home and job to inform the walking commute time.

As would be expected, increasing the GIS predicted commute time increases reported automobile commute times. The estimate of 0.267 is similar in magnitude to the estimate of 0.226 from the baseline linear commute time regressions (see Table 9), although the latter specification contained a constant. The analogous coefficients on carpooling and Metrobus are also positive, but it neither is similar in magnitude to the baseline specification. The Metrorail coefficient is of the wrong sign, but the distance to the Metrorail station effects indicate that time is a convex function of distance. This is consistent with individuals who need to travel greater distances to catch the subway traveling to the station by faster methods (ie, driving instead of walking), however, these results must be interpreted with caution given the lack of the geographic precision in the data. As-the-crow-flies distance is a positive predictor of walk times, although the convex relationship between the two indicates that it is only a positive predictor for short distances. Finally, the standard deviation of the idiosyncratic commute time error, $e$, is the amount of variation in commute times that is not being explained by the model.

The standard deviations of $\tilde{\lambda}$ and $w$ are close to zero, indicating that individuals view time spent commuting as a close substitute for leisure time.

### 6.1.3 Taste for Housing Consumption / Leisure Parameter Estimates

The preference parameters are those included in the $\alpha$s: $\beta^H$, $\beta^\ell$ and $\Omega^{\vec{\mu}}$. Recall that the taste parameter that governs the relative weight the individual places on composite consumption ($\alpha_i^c$) is normalized to one for identification. I present the parameters that govern relative taste for both housing and leisure in Table 12.

The leisure constant indicates that leisure dominates consumption of both housing and all-other-goods. The coefficients on the observable characteristics are similar in magnitude. Men prefer both more housing consumption and more leisure than women. As individuals

Table 12: Taste Parameter Estimates

| | Housing ($\alpha^H$) | | Leisure ($\alpha^\ell$) | |
|---|---|---|---|---|
| Variable | Estimates | Std. Error | Estimates | Std. Error |
| Individual Characteristic Parameters | $\beta^H$ | | $\beta^\ell$ | |
| Constant | 0.000[a] | - | 4.514*** | 0.007 |
| Male | 0.586*** | 0.005 | 0.591*** | 0.004 |
| ln(Age) | -0.703*** | 0.000 | -0.742*** | 0.001 |
| Second Moment Parameters ($\Omega^{\vec{\mu}}$) | | | | |
| Variances of taste parameter errors | 0.004*** | 0.000 | 0.042*** | 0.001 |
| Covariance of taste parameter errors | 0.013*** | 0.001 | - | - |
| Observations | | 973 | | |

[a] Parameter normalized to zero to ensure identification.
Notes: Single-starred items are statistically significant at the 10 percent level, double-starred items are statistically significant at the 5 percent level, and triple-starred items are statistically significant at the 1 percent level.

age, they place less value on housing amenities and leisure (and more on composite consumption). These parameters results are not intuitive and require further examination. The standard deviation of $\mu_i$ ($u_i$) is 0.004 (0.042). It indicates the amount of the preference for housing (leisure) that is not explained by individual observables.

## 6.2 Predicted Outcomes

In order to test how well the model performs, I compare three types of predicted, aggregate moments, evaluated at the estimated values of the parameters, $\hat{\theta}$, with their real-world counterparts from the data. Each of the moments corresponds to one of the modeled outcomes: commute time, commute mode, or housing choice. I begin with commute time, as it is the simplest outcome to calculate and is an input into other predicted outcomes. Next, I present moments relating to commute mode choice and finally housing choice.

To clarify the notation used in this section, recall that the observed housing choice of family $i$ is $h$, the observed commuting method of the head of family $i$ is $k$, and the head of family $i$ commutes for a duration of $t_{ihk}$. I define $\hat{h}$ as a possible home (from the affordable choice set of homes) family $i$ could live in, $\hat{k}$ as a potential method (again, from the feasible choice set) the head of family $i$ could use to get to work, and $\hat{t}_{i\hat{h}\hat{k}}$ as the predicted time it would take her to get to work from the given home by the given method.

### 6.2.1 Commute Time

To measure how well the model predicts commute times, I compare predicted commute times to the distribution of actual commute times from the data. Recall from Section 5.1.2.1 that $t_{ihk} = \exp\left(K_{ihk}\gamma^K + e_{ihk}\right)$, so the predicted commute time is $\hat{t}_{i\hat{h}\hat{k}} = \exp\left(K_{i\hat{h}\hat{k}}\hat{\gamma}^K\right)$. The commute characteristic parameters, $\gamma^K$, are estimated in the $P_i^t$ equation which uses the observed commute time as the response variable. The $\gamma^K$ are identified from the covariation of the commute characteristics ($K_{ihk}$) and that response variable, but they also appear elsewhere in the model. They are used in the $P_i^{HK}$ equation to proxy for unobserved commute times, so accurate estimates of commute times are an important input into the residential and commute mode choice parts of the model. I present predicted aggregate measures by commute method in Table 13. After the first row displays the mean and standard deviation of commute time by method from the data, the second presents the analogs to those moments based on predicted commute times from the observed home and by the observed method ($\hat{t}_{i\hat{h}\hat{k}}\, for\, \left(\hat{h},\hat{k}\right) = (h,k)$). The table shows that the model under-predicts automotive commute times and over-predicts the other commute times. The model does not do a reasonable job of matching the first and second moments of commute times (likely, in part, due to the lack of geographic precision in the PUMS data). Of particular concern, predicted Metrorail times are off by many orders of magnitude. I am in the process of refining the specification based on the restricted-access data in order to address this issue for the published version of this project.

Table 13: Observed vs. Predicted Commute Times

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
|---|---|---|---|---|---|---|---|---|---|---|
| | Auto | | Carpool | | Metrorail | | Metrobus | | Walk | |
| Variable | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Observed time ($t_{ihk}$) | 4.357 | 2.949 | 7.949 | 4.417 | 7.651 | 2.642 | 9.120 | 4.666 | 1.316 | 0.831 |
| Predicted time ($\hat{t}_{ihk}$) | 2.058 | 1.387 | 8.390 | 0.906 | 46.722 | 63.412 | 20.818 | 3.330 | 3.891 | 0.338 |

Notes: The unit of measurement for time is hours/week.

### 6.2.2 Commute Mode Choice

The probability an individual chooses a given $\left(\hat{h},\hat{k}\right)$ pair is the predicted analog to Equation 5.3, the joint probability of observing a family living in house $h$ and commuting by method $k$. I define this predicted probability as $\hat{P}^{HK}_{i\hat{h}\hat{k}} = \Pr(\hat{h},\hat{k} \mid \hat{\theta},\hat{t}_{i\hat{h}\hat{k}})$.[1] Summing this probability over homes for a given $\hat{k}$ gives the predicted probability that the commuter from family $i$ commutes by method $\hat{k}$: $\hat{P}^{K}_{i\hat{k}} = \Pr(\hat{k} \mid \hat{\theta},\hat{t}_{i\hat{h}\hat{k}}) = \sum_{\hat{h}=1}^{N^H} \hat{P}^{HK}_{i\hat{h}\hat{k}}$. I take the mean of these probabilities over individuals by commute mode to calculate the average predicted probability of commuting by the given method. Formally, I calculate the average predicted probabilities as $\bar{P}^{K}_{\hat{k}} = \frac{1}{N^I}\sum_{i=1}^{N^I} \hat{P}^{K}_{i\hat{k}} \,\forall\, \hat{k}$.

I am in the process of calculating these probabilities and will present and analyze them as part of the published version of this research project.

### 6.2.3 Housing Choice

As analogous aggregate measures to the ones detailed for commute mode choices (that aggregate $\hat{P}^{HK}_{i\hat{h}\hat{k}}$ to show the average probability each individual lives in each home) are not meaningful because each home can house only one family, I aggregate in a different manner. I am interested in how policies that affect commutes influence the distribution of housing locations, so I use geographic information on housing and job locations to calculate the aggregate probability that individuals live within a given distance range from their work locations. I index these ranges with $l$ and define the bounds of these ranges as $\left(\overleftarrow{d_l},\overrightarrow{d_l}\right)$. Recall from Section 3.1.3 that $d_{ihk}$ is the distance between house $h$ and job $j$ that individual $i$ travels by commuting method $k$. I define an indicator function that is equal to 1 if $d_{\hat{h}j}$ falls within range $l$: $1\left(\overleftarrow{d_l} < d_{\hat{h}j} < \overrightarrow{d_l}\right)$. Finally, I can write the individual probability of interest as $\hat{P}^{H}_{il} = \Pr(\overleftarrow{d_l} < d_{\hat{h}j} < \overrightarrow{d_l} \mid \hat{\theta},\hat{t}_{i\hat{h}\hat{k}}) = \sum_{\hat{h}=1}^{N^H}\sum_{\hat{k}=1}^{N^K} \hat{P}^{HK}_{i\hat{h}\hat{k}} 1\left(\overleftarrow{d_l} < d_{\hat{h}j} < \overrightarrow{d_l}\right) \forall\, l$, and it's aggregate analog as $\bar{P}^{H}_{l} = \frac{1}{N^I}\sum_{i=1}^{N^I} \hat{P}^{H}_{i\hat{d}} \,\forall\, l$.

I am in the process of calculating these probabilities, as well as developing other mea-

[1] Note that this notation means that I use the predicted value of $\hat{t}_{ihk}$, not the observed commute time, in calculating the predicted probability of observing $(h,k)$. For all other $\left(\hat{h},\hat{k}\right)$, this is the only option.

sures of interest, for the published version of this research project.

## 6.3 Specification Tests

To assess the accuracy of the model, I conduct a several specification tests. In this section, I explain how I conduct a chi-square goodness-of-fit test to asses how well the model performs. I am also working to develop additional specification tests including Wald tests of whether relevant subsets of the parameters are jointly equal to zero and Lagrange Multiplier tests to confirm that the model is properly specified.

### 6.3.1 Chi-Square Goodness-of-Fit Test

I use chi-square goodness-of-fit tests to determine how well the model reflects the data. I perform two tests that relate to the previously outlined outcome probabilities. First, I test the null hypothesis that the observed and predicted proportion individuals commuting by each mode are identical. Formally, this test is

$$\begin{aligned} H_0 : \bar{P}_k^K N^I &= \bar{P}_{\hat{k}}^K N^I \\ H_A : \bar{P}_k^K N^I &\neq \bar{P}_{\hat{k}}^K N^I \, for \, each \, k. \end{aligned}$$

The $\chi^2$ statistic for each commuting method is

$$\frac{\left[N^I \left(\bar{P}_k^K N^I - \bar{P}_{\hat{k}}^K N^I\right)\right]^2}{\bar{P}_{\hat{k}}^K N^I},$$

which has a $\chi^2$ distribution with $N^K - 1$ degrees of freedom.

I also perform a similar test on the observed and predicted average probability of living within a given distance range from work. The test and $\chi^2$ statistic are defined similarly to the previous case as

$$\begin{aligned} H_0 : \bar{P}_l^H \left(d_{hj}\right) N^I &= \bar{P}_l^H \left(d_{\hat{h}j}\right) N^I \\ H_A : \bar{P}_l^H \left(d_{hj}\right) N^I &\neq \bar{P}_l^H \left(d_{\hat{h}j}\right) N^I \, for\, each\, l, \end{aligned}$$

and

$$\frac{\left[N^I \left(\bar{P}_l^H \left(d_{hj}\right) - \bar{P}_l^H \left(d_{\hat{h}j}\right)\right)\right]^2}{\bar{P}_l^H \left(d_{\hat{h}j}\right) N^I}.$$

Letting $N^d$ denote the number of distance ranges indexed by $l$, the $\chi^2$ statistic follows a distribution with $N^d - 1$ degrees of freedom. As I have not yet calculated the predicted probabilities, I am unable to present the results of these tests in this dissertation, however they will be completed as part of the published version of the research project.

## 6.4 Policy Simulations

Congestion is the result of the nature of impure public goods (roads are non-excludable, but are rival) that causes them to be provided by the government at zero marginal cost. The rivalry leads to external costs because each additional driver on the road imposes costs on her fellow commuters that she does not fully bear. Direct quotas and Pigouvian taxes on vehicle miles traveled during congested times of the day are politically infeasible first best solutions (Parry et al. (2007)). Congestion pricing has been gaining traction as a more feasible alternative means of reducing congestion, and Shoup (1997) advocates applying congestion pricing principles to public street parking to reduce congestion, amongst other benefits.[2] All of these policies have the potential to influence both the monetary and time costs of commuting. I seek to better inform the discussion of ways to reduce congestion by performing simulations that illuminate the response to shifts in costs caused by a given policy. My model allows me to account for the response to policy shifts both in terms of

[2] I explain the details of several congestion pricing policies in Section 2.4.

the distribution of commuting method and residential location decisions.

I am in the process of using my model estimates to performing these comparative statics. I plan to conduct policy experiments based on numerous policies. Based on the parameters of a given policy, I first alter the pecuniary and time costs of the commutes faced by individuals in my model. Then I allow for three types of responses. First, I allow for a short-term response in terms of mode choice only. Second, I allow individuals to switch to a different commuting mode and/or move to a new residence in the medium term. Finally, in the long term, I also allow the housing stock to respond. The first two responses require only that I include a measure of congestion that feeds-back individual responses into the model.[3] The later requires estimating an additional housing stock equation. Although I do not estimate an equilibrium model, I can perform equilibrium comparative statics using the following algorithm:

1. Change the cost inputs based on the parameters of the given policy,
2. Calculate the distribution of $k$ that the model predicts with the new costs (do the same for $h$ in the later two scenarios) ,
3. Recalculate $t_{ihk}$ based on the new distribution of $k$ (and $h$, where applicable),
4. Recalculate $p^H$ based on the new distributions of $k$ and $h$ (in the later two scenarios), then
5. Repeating the previous steps until the process converges to a state where individuals no longer change their commuting mode or housing location.

The results of these simulations will allow me to determine not only the effects of a given policy on congestion, but also how much of that effect is due to mode switching and how much is due to individuals moving to new residential locations. I can also analyze which

[3] I have experimented with aggregate measures of congestion to capture this effect, but they are not identified by time variation alone with only one market included in my estimation. Instead, I plan to develop a measure of congestion based on the number of individuals commuting from an area around a given home to an area around a given individual's job location.

individuals are affected by the given policy to determine whether the policy is regressive in nature. This is an important consideration that is often cited by opponents of congestion pricing policies (Parry et al. 2007, Lewis 2008).

# Chapter 7

# Conclusions

My research develops a structural model of family residential choice and family member commuting that makes contributions to both the transportation and residential choice literatures. I do so by addressing the endogeneity of residential choice in analysis of commuting behavior with an individual-level model that has a rich unobserved heterogeneity structure. I have gained permission to and am currently estimating the model using restricted-access ACS data. The restricted-access data contains geographic precision that allows me to use GIS network analysis to painstakingly model the optimal commute between each pairwise combination of home and job locations by each commuting method observed in the data. I have obtained reasonable, preliminary estimates using PUMS data that indicate that the Fortran code I have written to optimize my likelihood function works. Finally, I outline policy simulations that are directly relevant to an emerging policy, the effects of which we do not yet fully understand, that has the potential to drastically reshape the urban environment in this country.

I also develop and am in the process of implementing a methodology that relaxes one of the most untenable assumptions in the residential choice literature: that family members living under the same roof share the same preferences for the characteristics of that home. Future work will focus on improving the fit of my model, conducting the policy simulations I outline, and adding cohabiting couples to my estimation routine.

# Bibliography

**Alonso, William**, *Location and Land Use: Toward a General Theory of Land Rent*, Cambridge, Mass.: Harvard Univ. Pr., 1964. 7

**Anas, Alex and Robin Lindsey**, "Reducing Urban Road Transportation Externalities: Road Pricing in Theory and in Practice," *Review of Environmental Economics and Policy*, 2011, *5* (1), 66–88. 12

**Arévalo, Raquel and Javier Ruiz-Castillo**, "On the Imputation of Rental Prices to Owner-Occupied Housing," *Journal of the European Economic Association*, 2006, *4* (4), 830–861. 91, 92

**Bajari, Patrick and Matthew E Kahn**, "Estimating Housing Demand With an Application to Explaining Racial Segregation in Cities," *Journal of Business & Economic Statistics*, 2005, *23* (1), 20–33. 7

_ **and Matthew E. Kahn**, "Estimating Hedonic Models of Consumer Demand with an Application to Urban Sprawl," in Andrea Baranzini, José Ramirez, Caroline Schaerer, and Philippe Thalmann, eds., *Hedonic Methods in Housing Markets*, Springer New York, 2008, pp. 129–155. 7

**Banzhaf, H. Spencer and Omar Farooque**, "Interjurisdictional Housing Prices and Spatial Amenities: Which Measures of Housing Prices Reflect Local Public Goods?," Working Paper 17809, National Bureau of Economic Research February 2012. 90

**Banzhaf, H.S. and R.P. Walsh**, "Do People Vote with Their Feet? An Empirical Test of Tiebout's Mechanism," *American Economic Review*, 2008, *98* (3), 843–863. 8

**Baum-Snow, N. and M.E. Kahn**, "The Effects of New Public Projects to Expand Urban Rail Transit," *Journal of Public Economics*, 2000, *77* (2), 241–263. 2, 7, 56

**Bayer, Patrick and Robert McMillan**, "Tiebout Sorting and Neighborhood Stratification," Working Paper 17364, National Bureau of Economic Research August 2011. 8, 9

**_ , Fernando Ferreira, and Robert McMillan**, "A Unified Framework for Measuring Preferences for Schools and Neighborhoods," *Journal of Political Economy*, 2007, *115*, 588–638. 8, 17, 43, 47, 88, 89, 90, 95

**_ , Robert McMillan, and Kim Rueben**, "An Equilibrium Model of Sorting in an Urban Housing Market," Working Paper 10865, National Bureau of Economic Research November 2005. 8, 15, 43, 88, 95

**_ , _ , and Kim S. Rueben**, "What drives racial segregation? New evidence using Census microdata," *Journal of Urban Economics*, November 2004, *56* (3), 514–535. 8, 47

**_ , Stephen L. Ross, and Giorgio Topa**, "Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes," *Journal of Political Economy*, December 2008, *116* (6), 1150–1196. 8, 47

**Becker, Gary S**, "A treatise on the family," 1981. 22

**Bento, Antonio M., Maureen L. Cropper, Ahmed Mushfiq Mobarak, and Katja Vinha**, "The Effects of Urban Spatial Structure on Travel Demand in the United States," *Review of Economics and Statistics*, 2005, *87* (3), 466–478. 2, 7

**Berndt, E.K., B.H. Hall, and R.E. Hall**, "Estimation and Inference in Nonlinear Structural Models," in "Annals of Economic and Social Measurement, Volume 3, number 4," NBER Books, 1974, pp. 103–116. 55

**Berry, Steven, James Levinsohn, and Ariel Pakes**, "Automobile Prices in Market Equilibrium," *Econometrica*, 1995, *63* (4), pp. 841–890. 8

**Blundell, Richard, Pierre-André Chiappori, and Costas Meghir**, "Collective Labor Supply with Children," *Journal of Political Economy*, 2005, *113* (6), 1277–1306. 64

**Browning, Martin and Pierre-André Chiappori**, "Efficient Intra-Household Allocations: A General Characterization and Empirical Tests," *Econometrica*, 1998, *66* (6), 1241–1278. 10, 21, 26

**_ , Francois Bourguignon, Pierre-André Chiappori, and Valerie Lechene**, "Income and Outcomes: A Structural Model of Intrahousehold Allocation," *Journal of Political Economy*, 1994, *102* (6), 1067–1096. 4, 10, 21, 23, 24

**_ , Pierre-André Chiappori, and Valerie Lechene**, "Collective and Unitary Models: A Clarification," *Review of Economics of the Household*, 03 2006, *4* (1), 5–14. 21, 22

**Bunch, David S.**, "Estimability in the multinomial probit model," *Transportation Research Part B: Methodological*, 1991, *25* (1), 1 – 12. 58, 59

**Chiappori, Pierre-André**, "Rational Household Labor Supply," *Econometrica*, 1988, *56* (1), 63–90. 4, 10

_ , "Collective Labor Supply and Welfare," *Journal of Political Economy*, 1992, *100* (3), 437–467. 10

**Chiappori, Pierre-Andre**, *Intrahousehold Resource Allocation in Developing Countries*, Johns Hopkins University Press, 22

**Chiappori, Pierre-André and Olivier Donni**, "Non-unitary Models of Household Behavior: A Survey of the Literature," IZA Discussion Papers 4603, Institute for the Study of Labor (IZA) November 2009. 11

— , **André de Palma, and Nathalie Picard**, "Couple Residential Location and Spouses Workplaces," *Columbia University, Department of Economics, mimeo*, 2012. 4, 11

**Dettling, Lisa J. and Melissa Schettini Kearney**, "House Prices and Birth Rates: The Impact of the Real Estate Market on the Decision to Have a Baby," Working Paper 17485, National Bureau of Economic Research October 2011. 62

**Duranton, Gilles and Matthew A. Turner**, "The Fundamental Law of Road Congestion: Evidence from US cities," NBER Working Papers 15376, National Bureau of Economic Research, Inc September 2009. 1, 2, 7

**Epple, D., T. Romer, and H. Sieg**, "Interjurisdictional Sorting and Majority Rule: An Empirical Analysis," *Econometrica*, 2001, *69* (6), 1437–1465. 8

**Epple, Dennis and Holger Sieg**, "Estimating Equilibrium Models of Local Jurisdictions," *Journal of Political Economy*, 1999, *107* (4), 645–681. 8

— , **Brett Gordon, and Holger Sieg**, "A New Approach to Estimating the Production Function for Housing," *The American Economic Review*, 2010, *100* (3), 905–924. 8

**Friedberg, Leora and Steven Stern**, "Marriage, Divorce, and Asymmetric Information," *Under Revision for the International Economic Review*, 2010. 22

**Geweke, John**, "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 1989, *57* (6), pp. 1317–1339. 44

**Goodman, Allen C. and Masahiro Kawai**, "Length-of-Residence Discounts and Rental Housing Demand: Theory and Evidence," *Land Economics*, 1985, *61* (2), pp. 93–105. 91, 92

**Goodman Jr., John L. and John B. Ittner**, "The accuracy of home owners' estimates of house value," *Journal of Housing Economics*, 1992, *2* (4), 339 – 357. 90

**Hajivassiliou, Vassilis A.**, "Smooth Simulation Estimation of Panel Data LDV Models," 1990. Unpub. msc. Yale U., 1990. 44

**Houde, Jean-François**, "Spatial Differentiation and Vertical Mergers in Retail Markets for Gasoline," *American Economic Review*, September 2012, *102* (5), 2147–82. 39

**Keane, Michael P.**, "A Computationally Practical Simulation Estimator for Panel Data," *Econometrica*, 1994, *62* (1), pp. 95–116. 44

**Kiel, Katherine A. and Jeffrey E. Zabel**, "The Accuracy of Owner-Provided House Values: The 1978-1991 American Housing Survey," *SSRN eLibrary*, Summer 1999, *27* (2). 90

**_ and _** , "Location, location, location: The 3L Approach to house price determination," *Journal of Housing Economics*, 2008, *17* (2), 175 – 190. 47

**Krueger, Alan B., Daniel Kahneman, David Schkade, Norbert Schwarz, and Arthur A. Stone**, "National Time Accounting: The Currency of Life," Working Papers 523, Princeton University, Department of Economics, Center for Economic Policy Studies. April 2008. 17

**Langer, Ashley and Clifford Winston**, *Brookings-Wharton Papers on Urban Affairs 2008*, The Brookings Institution Press, 3, 8, 64

**Leape, Jonathan**, "The London Congestion Charge," *The Journal of Economic Perspectives*, 2006, *20* (4), pp. 157–176. 12

**Lerman, S.R.**, "Location, Housing, Automobile Ownership, and Mode to Work: A Joint Choice Model," *Transportation Research Record*, 1976, *610*, 6–11. 9

**Lewis, David**, "America's Traffic Congestion Problem: Toward a Framework for Nationwide Reform," *Brookings Institute Discussion Paper*, 2008, *2008-06.* 2, 12, 77

**Lindsey, R.**, "Do Economists Reach a Conclusion?," *Econ Journal Watch*, 2006, *3* (2), 292–379. 11

**Lise, Jeremy and Shannon Seitz**, "Consumption Inequality and Intra-household Allocations," *The Review of Economic Studies*, 2011, *78* (1), 328–355. 22

**Lundberg, Shelly J., Robert A. Pollak, and Terence J. Wales**, "Do Husbands and Wives Pool Their Resources? Evidence from the United Kingdom Child Benefit," *The Journal of Human Resources*, 1997, *32* (3), pp. 463–480. 63

**Malpezzi, Stephen**, *Hedonic Pricing Models: A Selective and Applied Review*, Blackwell Science Ltd, 95

**Marshall, Robert C. and J. Luis Guasch**, "OCCUPANCY DISCOUNTS IN THE U.S. RENTAL HOUSING MARKET*," *Oxford Bulletin of Economics and Statistics*, 1983, *45* (4), 357–378. 91

**McFadden, D.**, *Conditional Logit Analysis of Qualitative Choice Behavior*, Academic Press: New York, 6

**McFadden, Daniel**, "Modelling the Choice of Residential Location," in A Karlqvist, L Lundqvist, F Snickars, and J Weibull, eds., *Spatial Interaction Theory and Planning Models*, Vol. 3, Amsterdam: North Holland, 1978, pp. 75–96. 9

_ , "Disaggregate Behavioral Travel Demand's RUM Side: A 30-Year Retrospective," in David Heshner, ed., *The Leading Edge of Travel Behavior Research*, Oxford: Pergamon Press, 2001. 5, 16

**Mills, Edwin S.**, "An Aggregative Model of Resource Allocation in a Metropolitan Area," *The American Economic Review*, 1967, *57* (2), pp. 197–210. 7

**Muth, Richard**, *Cities and Housing: The Spatial Pattern of Urban Residential Land Use*, Chicago: The University of Chicago Press, 1969. 7

**Olsen, Edgar O., Dirk W. Early, and Paul E. Carrillo**, "A Panel of Price Indices for Housing Services, Other Goods, and All Goods for All Areas in the United States 1982-2010," Virginia Economics Online Papers 402, University of Virginia, Department of Economics June 2012. 43

**Parry, Ian W. H., Margaret Walls, and Winston Harrington**, "Automobile Externalities and Policies," *Journal of Economic Literature*, June 2007, *45* (2), 373–399. 2, 11, 75, 77

**Pisarski, Alan**, "Commuting in America III: The Third National Report on Commuting Patterns and Trends," Technical Report, Transportation Research Board 2006. 87

**Polzin, Steven and Xuehao Chu**, "Public Transit in America: Results from the 2001 National Household Travel Survey," Technical Report, National Center for Transit Research at the Center for Urban Transportation Research at the University of South Florida 2005. 87

**Reschovsky, Clara**, "Journey to Work: 2000," Technical Report C2KBR-33, United States Census Bureau 2004. 87

**Schrank, David and Tim Lomax**, "The 2010 Urban Mobility Report," Technical Report, Texas Transportation Institute 2010. Texas Transportation Institute, Texas A&M University System. 1, 30

**Shoup, Donald C.**, "The High Cost of Free Parking," *Journal of Planning Education and Research*, 1997, *17* (1), 3–20. 75

**Small, Kenneth A.**, "Urban Transportation," *The Concise Encyclopedia of Economics*, 2008, pp. 507–510. 1

_ **and Erik T. Verhoef**, *The Economics of Urban Transportation*, 2nd ed., Routledge, 2007. 5, 6, 7

_ **, Clifford Winston, and Jia Yan**, "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability," *Econometrica*, 2005, *73* (4), 1367–1382. 12

**_ , _ , and _** , “Differentiated Road Pricing, Express Lanes, and Carpools: Exploiting Heterogeneous Preferences in Policy Design,” *Brookings-Wharton Papers on Urban Affairs*, 2006, pp. pp. 53–96. 12

**Stern, Steven**, “Simulation-Based Estimation,” *Journal of Economic Literature*, 1997, *35* (4), 2006–2039. 44, 54

**Tiebout, Charles M.**, “A Pure Theory of Local Expenditures,” *Journal of Political Economy*, 1956, *64* (5), 416–424. 8

**Train, K.E.**, *Discrete Choice Methods with Simulation*, 2nd ed., Cambridge University Press, 2009. 44, 60

**U.S. Department of Transportation**, “2008 Status of the Nation’s Highways, Bridges, and Transit: Conditions and Performance,” Technical Report 2008. 1

**Vega, Amaya and Aisling Reynolds-Feighan**, “A methodological framework for the study of residential location and travel-to-work mode choice under central and suburban employment destination patterns,” *Transportation Research Part A: Policy and Practice*, 2009, *43* (4), 401 – 419. 9, 33

**Xu, Zeyu**, “A Survey on Intra-Household Models and Evidence,” 2007. 11

# Appendix A

# Appendix

## A.1 Data

### A.1.1 Census Commuting Questions

Pisarski (2006) provides a list of shortcomings in the Census journey-to-work data that begins with the fact that the data contains no information "about aspects of trips using more than one mode of travel to get to work." According to the Census 2000 Brief "Journey to Work: 2000," Census data report the "usual means of transportation to work." When a person usually commutes via multiple transportation methods, only the method that covers the greatest distance is recorded (Reschovsky 2004). Given the prevalence of park-and-rides, transfer passes, and bike racks on buses, it is clear that multimodal travel is a reality in modern commuting, but there is little data available on this type of behavior. The National Household Travel Survey (NHTS) is the only national survey that measures mode transfers. Polzin and Chu (2005) calculate that 20 percent of all daily travel trips on transit are multimodal based on NHTS data, but the authors cannot reconcile this estimate of the prevalence of transfers with aggregate counts of the number of individuals who board transit vehicles reported by the NTD and the American Public Transportation Association (APTA). This discrepancy suggests that the NHTS figure may be an underestimate.[1]

While my research does not address this concern directly, and I am not able to identify the exact secondary means of transportation to work, I can control for the effect of multi-modal travel on commute time for some modes. To do this, I calculate the as-the-crow-flies distance from the fixed locations where individuals can enter and exit transit systems to their home and office locations. I assume that all commuters travel to the closest transit station to their home and exit at the closest transit station to their work location when traveling to work. To account for the fact that, for instance, those who live close to a station most likely walk or bike there, and beyond some threshold distance, individuals likely take the bus or drive to the station, I allow for differential effects by distance. Unfortunately, this control only works for multimodal commuters who report a form of rail as their primary means of commuting. I cannot apply a similar control to those who report commuting by road as their primary means of commuting.

### A.1.2 Opportunity Cost of Housing

The ACS includes two types of housing costs depending on the tenure type of the family being surveyed (ie, home owner or renter). Home-owners self-report a measure of total property value and renters report their monthly rent. Bayer et al. (2005) and Bayer et al. (2007) explain that there are three concerns with interpreting this data as a continuous measure of the opportunity cost of living in the given home. First, both the property value and rent variables may not reflect the true market value of the home. For home owners, the likely culprits are misreporting and overestimating the value of one's home. The real estate market is fluid and keeping up with it is costly, so home owners may not be savvy to the current market value of their home if they did not purchase it in the recent past (or if they have no intention of selling it in the near future). They may also have a more optimistic outlook on the value of their home than is warranted. While renters are much more likely

[1]Note that the NHTS figure reports the number of transfers on daily travel trips of all types, not just trips made commuting to and from work. It is not clear whether controlling for this distinction biases the estimate up or down.

to know their monthly rent, that rent may reflect a tenure discount if they have lived in the home for an extended period of time. A second issue is that, property values are reported in intervals in surveys prior to 2008 and are top-coded in all years. Finally, one must also adjust the owner and renter home value measures to be compatible across tenure types, as home values reflect the present discounted value of the flow of value from the home and rents reflect the stock value of the home.

This section details how I account for these issues and construct a consistent measure of the opportunity cost of living in each home from the available data. I primarily build on the data cleaning procedures of the Bayer et al. papers, however, since both studies use data from the 1990 Decennial Census for the San Francisco Bay Area, I modify their methodology to more appropriately fit my model and data. While the questions asked in recent Decennial Censuses and the ACS are remarkably similar, my data differs in three key ways. First, my data is a repeated cross section that spans multiple years as opposed to just one. Second, my data pertains to a different metropolitan area. This is noteworthy because of institutional differences in the way property taxes are assessed. Finally, property values are reported categorically in Census products prior to 2008 and continuously thereafter.

I proceed by detailing adjustments made to property values, rent values, and tenure type.

#### A.1.2.1 Property Value

Home-owners are asked to self-report the value of their home and property and this data is reported as a categorical, top-coded variable. Bayer et al. (2007) find that owners frequently report their home's purchase price, not its current value. There is evidence that this effect is present in my data as well. Homes sold within the previous year have, on average, reported values that are 10% higher than observationally equivalent homes purchased between 20 and 30 years earlier, all else equal.[2] In addition to misreporting, it has been shown in the literature that home owners frequently overestimate the value of their homes using

[2] All calculations reported in this appendix are based on 2005-2008 ACS PUMS data.

comparisons of self-reported and housing transactions data (see, for instance, Goodman Jr. and Ittner (1992), Kiel and Zabel (1999)). While I cannot determine the prevalence of overestimation of home prices due to the lack of transactions data, Banzhaf and Farooque (2012) find that price indices based on self-reported home values are highly correlated with those based on transactions data and are a practical alternative to more accurate, but less available, transactions data. To correct for the differential effects of misreporting across different categories of the family's tenure in the home and account for the overestimation of home values in my self-reported data, I estimate a house value hedonic at the community level using interval regression and use this regression to predict a continuous variable from the categorical, top-coded data. Doing so at the community level is equivalent to computing a price index (see Banzhaf and Farooque (2012), footnote 10), so this measure should perform as well as one based on unavailable, but more precise transactions data.

Formally, I interval regress log home value on tenure categories, annual property taxes paid, their interactions, housing characteristics, and year indicators. I do so separately by Public Use Microdata Areas (PUMAs).[3] I estimate the following equation

$$\ln(V_h) = \alpha_1 tenure_h + \alpha_2 \ln(tax_h) + \alpha_3 (tenure_h \times \ln(tax_h)) + \alpha_4 H_h + \alpha_5 year_h + \omega_h^V,$$

where $V_h$ is the self-reported house value, $tenure_h$ is a categorical measure of the length of time the family has resided in home $h$, $tax_h$ is the self reported property taxes paid, $H_h$ is the set of housing characteristics, $year_h$ is an indicator for the year the data was collected, and $\omega_h^V$ is an error. Bayer et al. (2007) are able to use the rules associated with Proposition 13 to transform property taxes paid into an estimate of the home's current value. I depart from their framework by including property taxes paid instead of this estimate, which I cannot easily calculate because property tax laws vary over time and with geography in my sample.[4] However, since I am running separate regressions at the PUMA level, tax

[3] Estimates based on regressions at the PUMA and year level did not substantially improve results.

[4] Washington, DC assess property taxes at the district level. Virginia assess property taxes at the county, city, or town level, and Maryland assess property taxes at the county or city level.

laws should be close to consistent by regression, although rates will undoubtedly vary over time. If this is the case, $\alpha_2$ will have predictive power so long as homes that have higher property taxes have higher values and it will return a linear approximation to the property tax cost in the PUMA. To the extent that multiple jurisdictions may exist in a given PUMA, $\alpha_2$ will return a weighted average of these costs. To reduce the influence that misreporting associated with longer tenured homes has on the fitted values of the hedonic, I interact the tenure and property tax rates. Finally, I replace $V_h$ with $\hat{V}_h$ in subsequent steps to correct for misreporting.

#### A.1.2.2 Rental Value

The existence of substantial tenure discounts in the rents of residents based on their length-of-residence in a given home is a well known phenomenon in the literature. For example, Marshall and Guasch (1983) are unable to reject the existence of such discounts. Goodman and Kawai (1985) find that the rent of recent movers is between 4% and 11% greater than that of all renters, depending on specification. More recently, Arévalo and Ruiz-Castillo (2006) report discounts in Spanish housing markets ranging from 3.2% to 83.5%, depending on the length-of-residence (up to 25 years). Discounts in line with these estimates exist in my data: renters who are in the second year of their lease receive a 4% discount relative to renters in the first year of their lease, all else equal. This discount increases to 50% for individuals who have lived in their residence for between 20 and 30 years.

Tenure discounts are believed to be the result of unobserved heterogeneity due to depreciation and/or state dependence due to match quality between the landlord and the tenant. The first explanation posits that if landlords postpone performing maintenance or reconditioning a home until tenant turnover, homes with longstanding tenants will be of lower quality than those available in the market. To the extent that the available information in the data does not accurately measure the quality of a unit (for instance, the data provides the number of bedrooms, but not how recently the carpet in those bedrooms was replaced), this depreciation will be unobserved and explains the existence of a tenure discount as a means of accounting for quality differences. An alternative explanation is due to state dependence. Arévalo and Ruiz-Castillo (2006) explain that turnover is costly not only for the tenant, but also the landlord (the costs of filling a vacancy include advertising costs, forgone rent, etc.). In addition, landlords may want to retain "good" tenants who treat the unit well and coexist with their neighbors. Landlords may do so by offering a discount to tenants who reveal themselves to be of high quality (see Goodman and Kawai (1985) for a theoretical model). Note that it is not possible to determine which phenomena leads to tenure discounts and both are likely to play a role in their existence.[5]

These discounts are of consequence when I construct a measure of market rent for every home in my sample. Doing so requires capturing the unmodeled dynamics that generate tenure discounts. The salient question is: what rent would a family in the model pay in each home *other* than the one the family lives in? There are four ways to construct a measure of unobserved rents, either as

1. The reported rent in the given home, which includes any tenure discounts that the current tenant has accrued or

2. An estimate of the rent in the given home that excludes the current tenant's tenure discount and instead includes an estimated tenure discount based on how long the family has lived in its observed home or

3. An estimate of the rent in the given home that excludes all tenure discounts or

4. An estimate of the rent in the given home that includes an estimated tenure discount based on what the current tenant has accrued.

How one interprets the cause of tenure discounts can help to guide the decision of which method is best, but they all have their drawbacks.

[5]Ideally, I would be able to model home choice as a dynamic programing problem where individuals choose their optimal home in each period given their expectations about future utility flows from the home (net of ownership or rental costs). With multiple observations on renters and homes, I would be able to separately identify the cause of the tenure discount and adjust the rent each family would face at each home accordingly. Unfortunately, the data preclude a dynamic model, as they are cross-sectional in nature.

The first method of constructing a measure of unobserved rents is to naively assume that the observed rent in the given home is the market rent. This is not sensible because it implies strong assumptions about the nature of both sources of tenure discounts. If the discount is entirely due to unobserved depreciation, using the observed rent without adjusting for duration of tenure implicitly assumes that landlords do not perform maintenance on the apartment before new tenants move in. However, this assumption negates the explanation for why unobserved depreciation results in a tenure discount that is not captured by controlling for the age of the housing structure. If, on the other hand, the discounts are entirely due to state dependence, this method of construction assumes that the discount is solely the result of the characteristics of the landlord (not the quality of the match between the tenant and landlord). Again, this negates the explanation for why state dependence results in a tenure discount: the landlord would just offer a low rent to new tenants to quickly fill his apartment if he was unconcerned with match quality.

The second means of constructing the rent measure equalizes the family's discount across all homes. The method is paramount to assuming that the family, however many years ago it was searching for its current home, faced the options that currently exist in the data and made a decision about where to live. The benefit of "turning back the clock" in this way is that doing so not imply any assumptions about the nature of the unobserved depreciation or state dependence that generates the discounts. However, this method is difficult to implement in practice because it would require adjusting the time dependent observable characteristics of the family members (such as age and marital status) back to what they were when the family last moved. Additionally, it is not an accurate representation of the decision a family thinking about moving in the given period faces, as it assumes that the family has perfect foresight and decides where to move once and stays there. The bias that using this method would introduce into my model depends on the reason tenure discounts exist. To the extent that tenure discounts are due to unobserved depreciation (and the discount prices this depreciation appropriately), the family would be indifferent between moving to a higher cost, recently maintained home and staying in their depreciated home with a discount, so no bias would be introduced. To the extent that tenure discounts are due to state dependence, families in my model would be more apt to move because they would not forfeit their accrued tenure discount by moving. If this is the case, this method would overstate the response to a policy that shifts the distribution of housing.

The third method of constructing the rent measure removes tenure discounts from all homes. It is equivalent to assuming that the family is living in the current home and considering moving to the other homes in the choice set. Whether consciously or not, this is a choice that families make each period. If the cause of tenure discounts is entirely due to unobserved depreciation, this means of constructing the rent measure assumes that landlords perform maintenance on homes before new tenants move in and adjust rents accordingly, so it is consistent with the unobserved depreciation theory. If discounts exist because of state dependence, this method assumes that they are the result of a good match between the both the tenant and landlord, which cannot be known to either the landlord or the tenant when they first sign a lease. This method is also consistent with the state dependence theory. While all of these implications are reasonable, the drawback of this method is that the family's housing history is endogenous because the discount is not removed from the home the family is currently living in.[6] This means that my model would understate a family's willingness to move in response to a policy shift because doing so would mean forfeiting an accrued tenure discount. Again, to the extent that tenure discounts are due to appropriately priced depreciation, this concern would be mitigated because the family would be indifferent between moving to a higher cost, recently maintained home and staying in their depreciated home with a discount. However, if the discount is caused by state dependence, this issue would be of greater concern.

The fourth method is similar to the first, but smooths the tenure discount by using the aggregate market discount instead of the individual home/landlord/renter discount. As all four methods of construction have drawbacks, I proceed by following the fourth method

[6]This problem could be ameliorated by adjusting rents for tenants in their observed homes to reflect a market rent that excludes tenure discounts, but doing so would mean that some individuals would not be able to afford the home they live in.

because it mitigates the extreme assumptions of other methods, it can be implemented without requiring ad-hoc adjustments to the time dependent observable characteristics of the family members, and it most closely reflects the standard in the literature set forth by Bayer et al. (2005) and Bayer et al. (2007). To model the market rents as the rent associated with the home reported in the data after smoothing with an aggregate measure of the tenure discount, I regress

$$\ln(R_h) \quad = \quad \beta_1 tenure_h + \beta_2 H_h + \beta_3 year_h + \omega_h^R,$$

where $R_h$ is the self reported gross rent, $tenure_h$, $year_h$, and $H_h$ are as described in the home value equation, and $\omega_h^R$ is an error.[7] Again, I run separate regressions by PUMA. I use $\hat{R}_h$ as a measure of the market rent in the given home.

#### A.1.2.3 Adjustment for Tenure Type

Data on owner and renter home costs in the data are not compatible because home values represent the present discounted value of a flow of services from the home and rents are the stock value of those services. In order to calculate the opportunity cost of living in an owner-occupied home, I regress

$$\ln(\pi_h) \quad = \quad \gamma_1 o_h + \gamma_2 H_h + \gamma_3 year_h + \omega_h^p,$$

by PUMA, where

$$\pi_h = \begin{cases} \hat{V}_h & if\, o_h = 1 \\ \hat{R}_h & else, \end{cases}$$

[7]To be consistent with home values, I use contract rents instead of gross rents (gross includes applicable home utility costs such as heat, electricity, etc.). In order to account for the effects of the inclusion of utilities in rental costs in some homes, but not in others, I follow Malpezzi (2008) and include indicators for the inclusion of a given utility in the rent in $H_h$.

$$o_h = \begin{cases} 1 & if\, h\, is\, owner-occupied \\ 0 & else, \end{cases}$$

$H_h$ and $year_h$ are as described in the previous equations, and $\omega_h^p$ is an error. I then use the estimate of $\gamma_1$ to convert home values to a measure of the rent the family would pay if it were renting the home. After making these adjustments to the data, the price of housing as defined in my model, is

$$p_h^H = \exp\left(\hat{\pi}_h\left(o_h = 0\right)\right).$$