Building Causal Connections among Job Accessibility, Employment, Income, and Auto Ownership Using Structural Equation Modeling: A Case Study in Sacramento County

By

Shengyi Gao
B.S. (Hubei University, Wuhan, China) 1985
M.S. (Yunnan University, Kunming, China) 1988

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in
Transportation Technology and Policy
in the
OFFICE OF GRADUATE STUDIES
of the
UNIVERSITY OF CALIFORNIA, DAVIS

Approved

committee in Charge

2006
ACKNOWLEDGEMENTS

I would like to thank the University of California Transportation Center (UCTC) for funding this dissertation through a faculty research grant. I would also like to thank the California Department of Transportation for their support of UCTC.

I am very grateful to my major advisor, mentor, and co-chair Robert A. Johnston for all of the support, encouragement, and inspiration he has provided me during my graduate studies. I would like to give thanks for my mentor and co-chair Patricia L. Mokhtarian for her heuristic discussions and invaluable comments on my dissertation. I would also thank Susan Handy for her important advice in the formation of the dissertation proposal and comments on the final draft of my dissertation. I would also thank Dan Sperling and James Cramer for their advice in the formation of the dissertation proposal.

I would also like to thank my wife, Pin Lu, for all of her love, support, and encouragement during my graduate studies. Thanks also to my daughter, Shuya Gao, for creating a warm family environment for my study.

I would like to thank Gordon Garry, Robert McCrary and Bruce Griesenbeck, Sacramento Area Council of Governments (SACOG), for providing me with SACOG’s travel demand forecasting model and answering my questions about the model. I would like to thank Mike McCoy for providing me with partial financial aid during my graduate studies.
TABLE OF CONTENTS

Chapter One: Introduction......................................................................................1

Chapter Two: A conceptual approach to testing the spatial mismatch hypothesis.....14

Chapter Three: Exploring the causal connections among job accessibility, employment,
income, and auto ownership with structural equation modeling..........................48

Chapter Four: What are the tradeoffs in deleting observations to obtain a multivariate
normal distribution? An empirical study.................................................................90

Chapter Five: Conclusions....................................................................................129
CHAPTER ONE
INTRODUCTION

1.1 IDENTIFICATION AND JUSTIFICATION OF RESEARCH QUESTIONS

After World War II, most American big cities, especially those central cities characterized by strong manufacturing and wholesale sectors in the North-Central and North-East regions, experienced rapid functional and demographic transformations (Kain, 1968; Kasarda, 1985). The functional transformation represented a transition of central cities from mainly serving as production and distribution centers to mainly serving as administrative and professional service centers. Traditionally, the manufacturing sector employed a lot of blue-collar laborers with only entry-level skills, and was spatially close to the labor market. The wholesale sector was characterized by a large number of entry-level jobs as well and was close to the customers to lower the operating costs. These patterns were greatly changed by technological advances, economic growth, and improvements in infrastructure. For example, with the introduction of assembly lines, manufacturing needed more horizontal floorspace than before. Expanding existing plants became costly due to low availability of land in surrounding areas or high land prices. The increase in the efficiency of production as a consequence of the application of new technology lowered the role of laborers in production activities. Whether or not the plants were close to the labor market became less important. From the perspective of economics, moving from central cities to suburbs seemed to be a logical choice for the manufacturing
sector. Like the manufacturing sector, with growth in size, the wholesale sector was also increasingly subject to space limitations. In addition, its operating costs increased due to heavier traffic congestion. For the wholesale sector, moving out of central cities seemed to be the only solution to these problems. Following both population and employment, a number of new retail stores chose to locate in the suburbs instead of the central cities as well.

In contrast to the manufacturing and wholesale sectors, employment in administrative, professional services, and institutions grew in cities. These sectors provided specialized services to their clients and hired persons with high-level job skills. They also consumed less floorspace per employee than the sectors that were characterized by a high proportion of blue-collar workers mentioned above and could pay higher rents for the floorspace in the central cities. Consequentially, changes in the establishments transformed the function of the central cities.

Demographic changes occurred simultaneously with the functional transformation. Middle-class households, mainly consisting of whites, moved to suburban areas for better housing at a lower price per square foot. The exodus was greatly speeded up by the improvement of transportation infrastructure (in the form of Interstate freeways), the fast growth of private automobile ownership, and the decentralization of jobs (Mouw, 2000; Zikmund II, 1975). In terms of proximity to jobs, especially entry-level jobs, the functional transformation of the cities was advantageous to whites. On the one hand, whites could easily access the jobs in suburban areas due to the proximity of their residences to the jobs. On the other hand, their access to jobs in central cities was not weakened because the new radial freeways allowed them to travel there, also.
African Americans, who were the dominant minority group in most big cities until recently, did not move with the low-skilled jobs (Stoll et al., 2000). Now as then, they generally have fewer years of education and rely on blue-collar jobs more than do whites. They have been and continue to be subject to discrimination in the real estate market, and are not welcomed by Caucasian homeowners in suburban neighborhoods (Bobo and Zubrinsky, 1996; Clark, 1992; Crowder, 2000; Duncan and Duncan, 1957; Farley et al., 1994; Emerson et al., 2001; Farley and Frey, 1994; Taeuber and Taeuber, 1965; Taub et al., 1984; Zubrinsky, 2000; Zubrinsky and Bobo, 1996). Beyond racial discrimination, African Americans are also subject to economic disadvantages because suburban residential zoning features single-family detached housing units and big lots, which are unaffordable to the majority of African American households (Cutler et al., 1999; Fong and Shibuya, 2000; Massey and Denton, 1993; Massey and Fisher, 1999).

Consequently, African Americans are often constrained in ghettos and old neighborhoods in the central cities, and are much less successful in residential attainment in suburban areas than whites (Crowder, 2001; Massey and Fong, 1990; South and Crowder, 1997). Although a number of white households and professionals moved into central cities against the trend, compared with the large flow of white households to the suburbs, this number is trivial (Kasarda, 1985; Mouw, 2000). Central cities are now home to large numbers of African Americans with lower household incomes.

Accompanying the functional and demographic changes was the rise in the unemployment rate of African Americans in the central cities. From the perspective of economists, these simultaneous changes must be causally linked. Because the total number of jobs, especially entry-level jobs, decreased in central cities, African Americans
lost proximity to jobs. Their access to the jobs in the suburbs was severely constrained due to racial discrimination in real estate market, low auto ownership and poor transit service between the central cities and the suburbs. Kain (1968) studied this phenomenon in Detroit and Chicago and proposed the spatial mismatch hypothesis (SMH) to explore the causal relationships among unemployment, residential segregation and decentralization of jobs, especially entry-level jobs. He concluded (p.374) that “racial discrimination in these housing markets, and the serious limitations on the residential choices of Afro-American households it produced, affected the spatial distribution of nonwhite employment and reduced nonwhite employment in both metropolitan areas.” He pointed out that the lack of transportation to overcome the geographic barrier between the jobs and residences was also one of the determinants of high unemployment rates among African Americans in the central cities.

Kain’s (1968) work initiated many empirical studies to test the spatial mismatch hypothesis along different dimensions, one of which is to test causal relationships between the employment ratio and job accessibility. If the effects of job accessibility on African Americans and whites are significantly different, spatial mismatch is suggested to exist. Interestingly, however, the evidence in empirical studies is not always supportive to the SMH (Cooke, 1997; Pastor and Adams, 1996). Ihlanfeldt and Sjoquist (1998) reviewed more than one hundred studies and attributed the inconsistencies of the findings to the neglect of endogeneity, which is caused by the correlation between the observed variables and the unobserved variables (Ramanathan, 2002), when the employment ratio was regressed on job accessibility in linear regression models. In other words, due to the mutual causality between employment and job accessibility, the assumption of the
independence of error terms does not hold. So, linear regression models should not be
employed in this context. It also implies that the findings, regardless of whether they
support or reject the SMH, are not reliable unless the endogeneity is dealt with.

There are only two possible approaches to solving the endogeneity problem
(Ihlandfeldt and Sjoquist, 1990): 1. finding a sample in which the assumption of
independence of error terms is not violated or 2. choosing a modeling method in which
the endogeneity can be incorporated. With respect to the first approach, youth who live
with their parents are thought to consist of such a sample and are used in several
empirical studies (Ihlanfeldt and Sjoquist, 1990; Raphael, 1998). The results in these
studies support the SMH. However, it is easy to see that this approach has two flaws.
First, there is not enough evidence to demonstrate that the residential choices of the
youth’s parents are independent of job accessibility. If their parents’ residential choices
are not independent of job accessibility, the assumption on the independence of the
youth’s residential choice may not be true. Second, even if the assumption is
demonstrated to be true, youths who live with their parents are a specialized population,
from which findings cannot be generalized. Therefore, this approach is not preferred.

Structural equation modeling is widely applied in psychological, economic and
management studies and has been demonstrated to be a powerful tool in disentangling the
complicated correlations between variables in the literature (Bollen, 1989; MacCallum
and Austin, 2000; Shah and Goldstein, 2006). It can incorporate correlations between the
error terms into the covariance matrix and distinguish reciprocal interactions between
endogenous variables. For this reason, a structural equation model does not require the
specialized population of the first approach. Therefore, it is a better approach for
capturing the causal relationships among employment, job accessibility, income, and auto
ownership which exist in the SMH context.

A critical assumption in structural equation modeling is multivariate normality of
the sample distribution. If a sample is severely multivariate nonnormal, the estimates of
the standard errors of the parameter estimates will be biased, and the model chi-square
statistic will be inflated (Henly, 1993; Muthén and Kaplan, 1985; West et al., 1995).
However, a sample having the multivariate normal distribution is generally not available
in social studies (Micceri, 1989). Therefore, the key issue is not whether a sample is
multivariate normally distributed, but whether the maximum likelihood estimation is still
robust when a sample does not have a multivariate normal distribution.

In this study, I propose two structural equation models as a conceptual approach to
test the SMH through theoretical analysis, and empirically demonstrate that structural
equation modeling is a better approach than linear regression models. I also demonstrate
the effects of deleting observations in a sample on goodness-of-fit, structural coefficients,
and standard errors of the estimated coefficients.

1.2 SHORT DESCRIPTION OF THE STUDY AREA

Sacramento County is the most important county in the Sacramento metropolitan
area in terms of its role in the regional economy. It consists of seven cities and an
unincorporated area. Except for two small cities (in terms of geographic size and
population), the other five cities (Citrus Heights, Elk Grove, Folsom, Rancho Cordova,
Sacramento) are spatially connected into an urbanized area through nine unincorporated
census-designated places (CDPs). The City of Sacramento is the biggest city and the
major job employment center of the county. In 1990, 37% of the retail jobs and 37% of
the non-retail jobs in Sacramento County were within the City of Sacramento. In 2000,
the corresponding percentages were 34% and 37%, respectively. From 1990 and 2000,
the retail jobs increased only by 7% in the City of Sacramento, while by 17% countywide.
The growth of non-retail jobs was 36% in the City of Sacramento and 35% countywide.
In other words, although retail jobs and non-retail jobs grew between 1990 and 2000, the
growth rate of retail jobs in the City of Sacramento was much lower than that of the
county. However, the faster growth of non-retail jobs during the 10-year period in the
City of Sacramento did not change the pattern of the job market, which is dominated by
retail jobs (78% of the total jobs are retail jobs). The CDPs between the cities, besides
being the main suburban residential areas, absorb 43% of the retail jobs and 41% of the
non-retail jobs of the county. These trends in urban employment change and land use
patterns are opposite to those of older and larger cities in the Northern Region and the
same as that of the cities in the Western Region documented by Kasarda (1985) between
the 1970s and 1980s. Therefore, it is interesting to study the relationship among job
accessibility, employment, income, and auto ownership in this metropolitan area.

Furthermore, in the year 2000, Sacramento County had 276 census tracts (CTs) and
792 census block groups (BGs). Only 16 census tracts and 33 block groups were in rural
areas. Therefore, the region can be considered a sample of a typical urbanized area. From
the standpoint of structural equation modeling, the CT sample is large enough for a
reliable estimation, and the BG sample is good for a more accurate (i.e. less biased)
estimation.
1.3 ORGANIZATION OF THE DISSERTATION

The dissertation consists of five chapters. Chapter One is a general introduction to this study. Chapter Two reviews the literature on the spatial mismatch hypothesis (SMH) and proposes two conceptual models for testing the SMH. Chapter Three empirically tests the first conceptual model proposed in Chapter Two. Chapter Four demonstrates the generalizability of the model in Chapter Three and examines the effects of multivariate nonnormality on model fit indices, structural coefficients, and the standard errors of the estimated structural coefficients. Chapter Five summarizes the findings of this study.
References:


CHAPTER TWO
A CONCEPTUAL APPROACH TO TESTING THE
SPATIAL MISMATCH HYPOTHESIS

Abstract

The spatial mismatch hypothesis (SMH) has long been a topic of interest in urban economics and transportation equity analysis. It has been examined from different perspectives and using different approaches, with aggregate data and disaggregate data. The inconsistent findings keep the debate going on. I argue that the findings, regardless of whether they support the SMH or not, come from conceptually incorrect models in which the endogeneity of residence is not taken into account or which use specialized samples, and therefore are either unreliable or subject to strict constraints on generalizability.

I propose two models to integrate job accessibility, employment ratio, income, and auto ownership into a structural equation system. The system allows direct reciprocal interactions and indirect feedback loops, and therefore has the power to address endogeneity and does not require specialized samples.

Key words: income, labor market, mobility
2.1. INTRODUCTION

In the past thirty-five years, the Spatial Mismatch Hypothesis (SMH) has been tested in many empirical studies. These empirical studies have tested different dimensions of the SMH (Ihlandfeldt and Sjoquist, 1998). Papers on the relationship between employment and job accessibility outnumber those on other dimensions. The findings, especially in the 1990s, are thought to be more consistent and support the SMH (Ihlandfeldt and Sjoquist, 1998).

However, two important research questions have not been solved yet. The first one is the endogeneity issue. In the spatial mismatch context, the endogeneity of residence and of auto ownership are not taken into account in the classic multiple regression models and logistic regression models that are typically used in these studies. The endogeneity bias lowers the reliability of the estimates. The second problem is a sampling bias in some studies, which undermines the generalizability of the results.

This paper is motivated by the need for a generalizable method that can overcome these two problems. Rather than using conventional multiple regression and logistic regression, I propose two new structural equation models.

2.2. LITERATURE REVIEW

The Spatial Mismatch Hypothesis is the idea that, as jobs, especially low-skilled jobs, decentralize from the urban center, many poor urban blacks are left without easy access to employment (Holzer, 1991). The farther jobs move from impoverished neighborhoods, the more difficult it becomes for blacks to find work. Kain (1968) first discussed this in an empirical study of Chicago and Detroit. He found that the
employment rate of inner-city blacks was negatively associated with the distance between their residence and employment opportunities.

Since Kain’s paper, many empirical studies have been conducted to test the impacts of isolation of poor neighborhoods from low wage, low skill jobs. A key question for Kain is whether blacks who are segregated in the housing market are more subject to the impacts of job decentralization than whites. Although Kain does not use the term, his findings are the basis for what is termed the Spatial Mismatch Hypothesis by other researchers (Arnott, 1998; Ellwood, 1986; Holzer, 1991; Ihlanfeldt and Sjoquist, 1998; Kasarda and Ting, 1996; Ong and Miller, 2003; Raphael, 1998). Kain used the term, but did not explicitly define it, in his later review (1992). The following was thought by Arnott (1998) to be Kain’s (1992, p.371) definition of the spatial mismatch hypothesis: “Serious limitations on black residential choice, combined with the steady dispersal of jobs from central cities, are responsible for the low rates of employment and low earnings of Afro-American workers.”

Ihlanfeldt and Sjoquist (1998, p.851) summarize the spatial mismatch hypothesis: “A simple statement of the SMH is that there are fewer jobs per worker in or near black areas than white areas. As a result, blacks may have greater difficulty in finding work, may be paid less, or may have to make a longer commute in comparison with whites with similar job credentials. Underlying the SMH are the following premises: (1) The demand for labor has shifted away from neighborhoods where blacks are concentrated in favor of high-growth, mostly suburban areas; (2) racial discrimination in housing and mortgage markets has prevented blacks from moving to where job growth exists; and (3) customer discrimination against blacks, poor information about distant job openings, limited public
transportation linkages between black neighborhoods and areas of job growth, and possibly other factors have restricted the ability of blacks to commute to and work in job-rich areas.”

This hypothesis has been tested by many empirical studies (Cooke, 1993, 1996, 1997; Ellwood, 1986; Holloway, 1996; Holzer et al., 1994; Hughes and Madden, 1991; Ihlanfeldt and Sjoquist, 1989, 1990a, 1990b, 1991; Immergluck, 1998; Kasarda and Ting, 1996; Master, 1974; McLafferty et al., 1992; McLafferty and Preston, 1996; Mooney, 1969; Ong and Miller, 2003; Raphael, 1998; Taylor and Ong, 1995; Thompson, 1997; Yinger, 1995; Wyly, 1996;). The empirical studies have tested different dimensions of the SMH such as travel time/distance comparisons between whites and blacks, correlations between employment rates and job accessibility, and income comparisons of blacks who reside in central cities and suburban areas. The papers on the relationship between employment and job accessibility outnumber those on other dimensions. This paper will discuss the empirical studies along this dimension.

Table 2.1 is a short list of the literature exploring the relationship between employment and job accessibility. From the table, we can see that the majority of the empirical studies are conducted with ordinary least squares (OLS) multiple regression or logistic regression in which employment-related variables were used as dependent variables and job accessibility as one of the explanatory variables. Some results are consistent with the SMH, and some are not. In their literature reviews, Kain (1992) and Ihlanfeldt and Sjoquist (1998) argue that the main reasons why some research does not support the spatial mismatch hypothesis are improper measures of job accessibility (for example, Ellwood, 1986) and failure to take into account the endogeneity of residence,
Table 2.1 Summary of literature on spatial mismatch (relationship between job accessibility and employment)

<table>
<thead>
<tr>
<th>Paper</th>
<th>Data and unit of analysis</th>
<th>Sample</th>
<th>Research Method</th>
<th>Dependent Variable</th>
<th>Primary independent variables</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kain, 1968</td>
<td>Chicago, 1956 and Detroit, 1952 AREA Traffic Study Survey; zone level</td>
<td>Pooled data for blacks</td>
<td>OLS regression</td>
<td>Share of black workers in each zone</td>
<td>Distance to major ghetto; distance to nearest ghetto; black share of workers residing in a zone</td>
<td>Share of black workers in each zone positively associates with racial composition, and negatively associates with distances between work and residence.</td>
</tr>
<tr>
<td>Mooney, 1969</td>
<td>Census 1960; SMSA level</td>
<td>The blacks in the poorest tract; 25 SMSAs</td>
<td>OLS regression</td>
<td>Employment-population ratio</td>
<td>Ratio of the jobs of a sector in central city to the jobs of that sector in the SMSA</td>
<td>Employment-population ratio increases with the job shares in central cities of all occupations.</td>
</tr>
<tr>
<td>Greytak, 1974</td>
<td>Journey-to-work survey data; city level</td>
<td>1965 interviews in the cities across U.S.</td>
<td>OLS regression</td>
<td>Work trips</td>
<td>Time cost of the journey to work; work trip origination dummy, 1 if from central; race dummy; mode dummy</td>
<td>Nonwhites have higher time cost; number of work trips negatively associate with work trip origination from central city; time cost of nonwhite work trips increases with city size, while decreases for whites.</td>
</tr>
<tr>
<td>Masters, 1974</td>
<td>Census 1960; SMSA level</td>
<td>Large SMSAs</td>
<td>OLS regression</td>
<td>Ratio of non-white to white median income</td>
<td>Residential segregation indices</td>
<td>None of the indices have significant impacts on the ratio of non-white to white median income.</td>
</tr>
<tr>
<td>Ellwood, 1986</td>
<td>Census 1970; tract level</td>
<td>Black youth and white youth out of school; Chicago;</td>
<td>OLS regression</td>
<td>Employment rate</td>
<td>Five measures of job accessibility</td>
<td>No job accessibility measures have significant impacts on the employment rate of white and black youth.</td>
</tr>
</tbody>
</table>
Table 2.1 Summary of literature on spatial mismatch (Continued)

<table>
<thead>
<tr>
<th>Source</th>
<th>Data Source and Methodology</th>
<th>Sample</th>
<th>Analysis Type</th>
<th>Employment Variables</th>
<th>Job commuter Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farley, 1987</td>
<td>Census 1980 and Census 1977 for four industries; SMSA level</td>
<td>Black and Hispanic males over 16</td>
<td>OLS regression</td>
<td>Unemployment rate; Job share of four industries; Black share of population; education inequality</td>
<td>Black male unemployment rate tends to be high relative to that of white males where the metropolitan area has high percentage of blacks, decentralized jobs, centralized black population and a high level of racial inequality in education attainment.</td>
</tr>
<tr>
<td>Ihlanfeldt and Sjoquist, 1990a</td>
<td>1980 Public Use Microdata Samples; disaggregate</td>
<td>Philadelphia SMSA</td>
<td>OLS regression; logistic regression</td>
<td>Partial derivatives of employment probability (ln(P_j/P_i))</td>
<td>Mean travel time of youth; mean travel time of all low wage workers Proximity to jobs increases the likelihood of employment for both black and white youth. Proximity to jobs is attributed to the residential segregation.</td>
</tr>
<tr>
<td>Ihlanfeldt and Sjoquist, 1990b</td>
<td>1980 Public Use Microdata Samples; disaggregate</td>
<td>Chicago SMSAs, youth who live with parents</td>
<td>OLS regression</td>
<td>Partial derivatives of employment probability (ln(P_j/P_i))</td>
<td>Family income categories; residential location dummies Residential location has significant impacts on white youth and black youth employment. Racial differential is attributed to residential segregation.</td>
</tr>
<tr>
<td>Ihlanfeldt and Sjoquist, 1991</td>
<td>1980 Public Use Microdata Samples; SMSA level</td>
<td>23 SMSAs</td>
<td>Logistic regression</td>
<td>Employment probability</td>
<td>Commute time; income categories Job accessibility increases the likelihood of employment for white youth and black youth.</td>
</tr>
<tr>
<td>Taylor and Ong, 1995</td>
<td>American Housing Survey, 1977/1978-1985; zone level</td>
<td>10 metropolitan areas</td>
<td>OLS regression</td>
<td>Commute distance; commute time</td>
<td>Age; travel mode; education; race; residential mixed area; minority area The travel patterns of white and minority workers diverge over time. The commute distance of non-moving workers in minority areas decreases. The findings do not support SMH.</td>
</tr>
<tr>
<td>McLafferty and Preston, 1996</td>
<td>1980 and 1990 Public Use Microdata Samples; disaggregate</td>
<td>New York CMSA; persons with a job</td>
<td>OLS regression</td>
<td>Mean commute time</td>
<td>Race; gender; travel mode; mean weekly income Black men and women in central city have longer commutes than white men and women. Spatial mismatch does not exist in suburban areas, and does not exist in Latino areas.</td>
</tr>
</tbody>
</table>
Table 2.1 Summary of literature on spatial mismatch (Continued)

<table>
<thead>
<tr>
<th>Study</th>
<th>Years</th>
<th>Sample Characteristics</th>
<th>Methodological Approach</th>
<th>Variables</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holloway, 1996</td>
<td>1980 and 1990</td>
<td>Public Use Microdata Samples; MSA level</td>
<td>Logistic regression</td>
<td>Employment probability</td>
<td>The impacts of job accessibility decline between 1980 and 1990. The black teenagers living in job-accessible areas have no advantages in 1990 over those living in job-inaccessible areas.</td>
</tr>
<tr>
<td>Wyly, 1996</td>
<td>1980 and 1990</td>
<td>Public Use Microdata Samples</td>
<td>Structural equation modeling</td>
<td>Endogenous variable: hourly wages; working hours; commute distance</td>
<td>No evidence of increasing commute distances among blacks is found.</td>
</tr>
<tr>
<td>Kasarda and Ting, 1996</td>
<td>1980 and 1990</td>
<td>Public Use Microdata Samples; Housing Summary Tape File 3A; the Urban Underclass Database; city level</td>
<td>Structural equation modeling</td>
<td>Endogenous variable: % female jobless; % female in poverty; % female on welfare; % male jobless; % male in poverty; % male on welfare; skill mismatch; spatial mismatch</td>
<td>Skill mismatch and spatial mismatch are positively associated with percent jobless for male and female. Residential segmentation index is positively associated with spatial mismatch. The percent loss of low-skilled industry jobs has no effects on spatial mismatch. The percent low-skilled industry jobs decreases spatial mismatch.</td>
</tr>
<tr>
<td>Raphael, 1998</td>
<td>Census tract 1980 and 1990; tract level</td>
<td>Bay Area, California; youth</td>
<td>OLS regression</td>
<td>Employment rate</td>
<td>Job growth other than job levels has significant impacts on employment rate. Black youth have higher access to job levels and lower access to job growth than white youth.</td>
</tr>
<tr>
<td>Ong and Miller, 2003</td>
<td>Census tract 2000; tract level</td>
<td>Los Angeles</td>
<td>Two-stage regression</td>
<td>Employment rate</td>
<td>Job accessibility has no impacts on employment rate for males and females.</td>
</tr>
</tbody>
</table>
especially the latter (for example, Cooke, 1997; Pastor and Adams, 1996).

The endogeneity of residence can be simply described as an endogenous relationship between employment and job accessibility. On the one hand, in the SMH, job accessibility is thought to be a key factor that directly affects employment, especially for blacks. On the other hand, in residence location choice, employment-related variables such as income, attributes toward commuting, and congestion, etc. play an important role. Since residential location is a major determinant of job accessibility (see Equation (2.3)), employment-related variables also affect job accessibility. The relationship between job accessibility and employment rate can be written as:

Employment rate = \( f_1 \) (job accessibility, \( X_1, \delta_1 \))  
(2.1)

Job accessibility = \( f_2 \) (employment rate, \( X_2, \delta_2 \))  
(2.2)

where \( X_1, X_2 \) are vectors of explanatory variables, and \( \delta_1 \) and \( \delta_2 \) are error terms.

In Equation (2.1), \( \delta_1 \) is a component of employment rate, which is causally related to job accessibility in Equation (2.2). Similarly, \( \delta_2 \) is a component of job accessibility in Equation (2.2), which is a cause of employment rate in Equation (2.1). Thus, job accessibility is correlated with \( \delta_1 \) because of its relationship with employment rate (which is a function of \( \delta_1 \)) in Equation (2.2), and employment rate is correlated with \( \delta_2 \) because of its relationship with job accessibility (which is a function of \( \delta_2 \)). The endogeneity of residence violates a key assumption of OLS regression, namely that included explanatory variables are independent from excluded variables. It will unavoidably result in biased estimates in a single regression equation (Finkel, 1995; Kline, 1998). Whether the coefficient of job accessibility is significant in Equation (2.1) depends on the magnitude of the direct reciprocal impacts or indirect impacts of employment or both on job
accessibility. Conceptually, regardless of whether the coefficient of job accessibility is statistically significant, the estimates are not reliable due to the endogeneity bias. Logistic regression models are subject to the same problem. Therefore, the findings in the literature that are based on OLS models without considering endogeneity bias are subject to challenge, regardless of whether the results support the SMH or not.

As suggested by Ihlanfeldt and Sjoquist (1990b), two approaches can be adopted to improve the reliability of the estimates. The first approach is to sample people whose residence is strictly exogenous to their employment. Youth who live with their parents have been thought to be a sample that meets this condition. This approach was very popular in the 1990s, and studies using this approach provide results supportive to the SMH. However, as Ihlanfeldt and Sjoquist (1998) argue in their later literature review, this approach has at least three shortcomings. First, the exogeneity of residential location of the youth holds only if the parents’ residential location choice is exogenous, which is hard to prove. Second, youth who do not live with their parents are excluded from the sampling, which results in biased samples. Third, therefore, the findings cannot be generalized even to all youth, let alone to adults as well.

The second approach is to design a simultaneous equation system that can include both employment and job accessibility as endogenous variables. Ihlanfeldt and Sjoquist (1998) do not discuss the conceptual structure of this system due to the unavailability of data for an empirical test. To our best knowledge, only Kasada and Ting (1996) and Wyly (1996) used a method similar to this system, which will be further discussed below.

Using full-information maximum-likelihood (FIML)-based structural equation modeling, Kasarda and Ting (1996) study the feedback relationships among
unemployment rates, welfare enrollment rates, and poverty rates across cities. In their conceptual model, skill mismatch, which is measured by the percent of city jobs held by persons with more than a high school education versus percent of out-of-school adult residents with less than a high school education, and spatial mismatch, which is measured by the average number of minutes for the one-way work commute of city residents, do not affect each other. Neither of them increases the joblessness for males or females.

The percent of persons who have a high school education or less is assumed to increase skill mismatch and spatial mismatch. The percent of low-skilled industry jobs in 1980 is assumed to decrease skill mismatch and spatial mismatch. The percent loss of low-skilled industry jobs between 1980 and 1990 is assumed to increase skill mismatch and spatial mismatch. A higher residential segregation index in 1990 is assumed to increase spatial mismatch. Higher joblessness is assumed to increase the poverty rate. A higher poverty rate is assumed to increase the welfare rate. The welfare rate may increase or reduce the unemployment rate. Their data are aggregated at the city level, with a sample of the 100 largest cities in the U.S. They find that the percentage of females on welfare has no significant impact on female unemployment; the male welfare rate increases the male unemployment rate; the percentage of persons who have a high school education or less has no significant impact on spatial mismatch; the percent loss of low-skilled industry jobs between 1980 and 1990 has no impact on spatial mismatch; and all other impacts are significant and the directions are the same as expected. It is noted that the percentage of persons who have a high school education or less and the percent loss of low-skilled industry jobs between 1980 and 1990 indirectly increase percent male
joblessness and percent female joblessness by increasing skill mismatch rather than
spatial mismatch.

Wyly (1996) used structural equation modeling to explore the impacts of
occupational segmentation, industry segmentation, worked hours, and hourly wages on
commute distance for black males, black females, white males and white females. He did
not find any statistically significant differences between blacks and whites in commute
distance after controlling for the other variables, and found that commute distance is
determined by hourly wages and worked hours.

The foci of Kasarda and Ting (1996) and Wyly (1996) are not the causal relationship
between employment and job accessibility. Therefore, it is not surprising that their results
do not provide strong support to the SMH. However, their modeling methods shed light
on how to integrate employment and job accessibility into a system as two endogenous
variables.

Multi-stage regression is often used to address an endogeneity bias. Ong and Miller
(2003) use a two-stage regression model to study the impacts of job accessibility and
transportation access on employment ratio and unemployment rate in Los Angeles. They
handle the endogeneity of car ownership in the first stage by introducing an instrumental
variable for predicting car demand, which is designated as a linear function of car
insurance index, population density and bus availability. In the second stage regression,
however, job accessibility is still treated as exogenous to the employment ratio. Therefore,
their estimates are subject to the same endogeneity problems as those in OLS models.
Even if they had used instrumental variables for both job accessibility and car ownership,
their model only tells us how the variance in employment rate is explained by job
accessibility and car ownership, and misses how the variances in job accessibility and car ownership are affected by employment rate. Therefore, multi-stage regression is not a good method in this context, where bi-directional causality is possible.

Structural equation modeling is an *a priori* simultaneous equation system. This system consists of endogenous and exogenous variables. The exogenous variables impose their impacts on endogenous variables, but will not be affected by endogenous variables. The endogenous variables are connected to each other through unidirectional or bi-directional causality. Since this system can solve the endogeneity problem, it does not require the specialized samples (e.g. youth living with parents) found in some past empirical studies. The ability of the system to incorporate endogeneity into one covariance matrix and to provide a $\chi^2$ test for models and a t-test for each coefficient is especially attractive. Therefore, structural equation modeling is the modeling approach that is suggested by Ihlanfeldt and Sjoquist (1990b). In this paper, I will introduce two conceptual models which have different foci and explore different issues of the SMH.

### 2.3. CONCEPTUAL MODELS OF THE SMH

When we test the spatial mismatch hypothesis, we want to know not only how the employment ratio (or employment rate) is affected by job accessibility and other variables, but also how the employment ratio affects job accessibility and other variables in turn. Therefore, it is of great importance to build a correct conceptual structure to guide data collection and empirical test of the relationships among the variables.
2.3.1. Cross-Sectional Structural Equation Model of the SMH

Figure 2.1 illustrates a conceptual model of how the endogenous variables (within the dotted rectangle) affect each other and how they are affected by exogenous variables. The relationships are based on urban economic theories and results of empirical studies, and the variables are expected to be measured at an aggregate (e.g., census tract) level. The arrow represents the direction of the influence, and the sign +/- represents the expected sign of the coefficients. Four endogenous variables are included in the model: employment ratio, auto ownership, income, and job accessibility. They influence each other unidirectionally or bi-directionally. Socio-economic variables impose effects on the endogenous variables, but are not affected by them. Due to the existence of reciprocal interactions and loops, this model is a nonrecursive model. The endogenous variables are summarized in Table 2.2.

Figure 2.1 Cross-sectional conceptual model of the SMH
Table 2.2 Endogenous variables

<table>
<thead>
<tr>
<th>Endogenous variables</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment ratio</td>
<td>Employment ratio = (The number of employed residents in a tract)/(The number of people aged 16 to 64 in the same tract)</td>
</tr>
<tr>
<td>Income</td>
<td>Median household income</td>
</tr>
<tr>
<td>Auto ownership</td>
<td>Average number of vehicles per household</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>Equation (2.3)</td>
</tr>
</tbody>
</table>

2.3.1.1 Direct Impacts of Endogenous Variables

Impacts of Job Accessibility. Job accessibility is a core variable in spatial mismatch hypothesis testing. According to the spatial mismatch hypothesis, it is expected that job accessibility will have significant, positive impacts on employment ratios of blacks and whites. The employment ratio of blacks is expected to be more sensitive to job accessibility than that of whites because blacks are subject to housing segregation and employers’ possible discrimination. Therefore, the magnitude of impacts of job accessibility on blacks and whites should be different, and the impact differential should reflect the difference between blacks and whites in accessing jobs. In the literature on SMH, the lack of a unitary definition of job accessibility is thought to be a main reason for the inconsistencies in empirical studies (Ihlanfeldt and Sjoquist, 1998). In this paper, I choose the gravity model structure (Equation (2.3)), which is widely used in travel demand forecasting practice. This structure captures all job opportunities within the planning region and the interactions, which decay nonlinearly with the increase of
distance, between census tracts in the labor market. It is more reasonable than the summation of jobs within a fixed distance or travel time (Ellwood, 1986; Immergluck, 1998; Ong and Miller, 2003). Thus, the job accessibility of census tract i, $A_i$, is measured as follows:

$$A_i = \sum_{j=1}^{N} \text{Job}_j \exp (-\gamma d_{ij})$$  \hspace{1cm} (2.3)

where $i = 1, 2, \ldots N$, $j = 1, 2, \ldots N$; $N$ is the total number of census tracts in the research area; $\text{Job}_j$ is the total jobs in census tract $j$; $\gamma$ is a constant; $d_{ij}$ is distance from census tract $i$ to census tract $j$ which is measured by travel time by automobile on the highway network at a.m. peak period.

The measurement of job accessibility always contains a geographic location factor. In Equation (2.3), the location factor is represented by census tract $i$ where people reside. When people choose a neighborhood in which to live, what they choose is not just a house or location, but a consumption bundle. In the three components of job accessibility, jobs and travel time are exogenous while the residence is determined by income and other variables, and thus is endogenous in the structural equation model. Any variable which has direct impacts on residence will have direct impacts on job accessibility.

Job accessibility is assumed to negatively affect auto ownership. First, people or households may intentionally choose to live in neighborhoods with high job accessibility by non-auto modes to lower their dependency on the auto mode, given there is no racial discrimination in the housing market and people can choose to live where they want. Second, autos have no substantial marginal benefits for people living in neighborhoods with high job accessibility by non-auto modes. From the perspective of economics, the high fixed and variable costs make autos a luxury good for people with low incomes, and
lower their desires to own autos. Kockelman (1997) finds that job accessibility, which is measured by total jobs within 30 minutes’ auto travel, is negatively associated with auto ownership.

The impact of job accessibility on income is assumed to be only indirect, through the employment ratio.

**Impacts of Auto Ownership.** It has been found that auto ownership increases employment. Raphael and Rice (2002) find that there is a large difference between the employment rates of youth with and without private autos. Ong and Blumenberg (1998) and Cervero et al. (2000) find that owning a car significantly increases the probability of welfare participants being employed.

Auto ownership is expected to have a negative impact on job accessibility. Urban economic theory predicts that people with high auto ownership will more likely live farther from jobs to gain higher residential amenities (Holzer, 1991). People with lower ownership will have to choose to reside in neighborhoods close to jobs or with good transit service. The empirical evidence supports the theories (Schimek, 1996).

Again, auto ownership is expected to affect income only indirectly, through employment ratio.

**Impacts of Employment Ratio.** The outcomes of the labor market can be represented by several different variables. Employment ratio, which in this paper is the ratio of total employed population to total population aged 16 to 64 in a census tract, is most often used in empirical studies. Rather than a single dependent variable as in OLS models, in structural equation modeling, employment ratio is set as an endogenous variable that imposes impacts on other endogenous variables, and in turn is affected by
other endogenous variables. It is common sense that labor income is a consequence of employment and is the main source of income for the majority of households. Employment ratio has positive and direct impacts on income, but will not be determined by income. The impact is unidirectional.

Employment ratio is assumed to positively affect auto ownership directly besides its indirect impacts through income, because employment prompts the acquisition of an auto almost independently of income – i.e. more as a fixed cost of employment than one that varies continuously with income.

The direct impact of employment ratio on job accessibility is designed to capture the impact of employment on voluntary residence self-selection which is separate from residence selection based on income. For example, employment-related variables such as a dislike of commuting and/or congestion may motivate some people to choose to live in neighborhoods close to their jobs. Under this circumstance, job accessibility is directly affected by employment status with a sign opposite to that based on income. This impact is assumed to exist (Ihlanfeldt and Sjoquist, 1989, 1992; Kain, 1992), but to our best knowledge has never been empirically tested in the context of the SMH. The positive sign on the link represents the assumption that employment ratio is positively associated with job accessibility, i.e. that tracts with higher employment ratios have higher job accessibility.

**Impacts of Income.** There is little doubt that income is the most important determinant in auto ownership (Dargay, 2001; Golob and Wissen, 1989; Kockelman, 1997). The owner of an automobile has to pay the fixed cost (purchase, registration, insurance) and variable costs for gasoline and maintenance fees. These expenses result in
lower auto ownership among the poor (Dargay, 2001; Ong and Blumenberg, 1998). A difference in auto ownership is expected to exist between white households and black households due to the differences in household income.

Income is assumed to have negative impacts on job accessibility. According to urban economics, people with higher incomes have flatter bid-rent curves and are more likely to make tradeoffs between housing amenities and commute costs. People with lower incomes have steeper bid-rent curves, and are more likely to live closer to jobs, and usually thus have higher job accessibility.

From Figure 2.1 and the discussion above, we can see that it is quite possible that the interaction between job accessibility and employment ratio is the case discussed in Equations (2.1) and (2.2), i.e. job accessibility and employment ratio are mutually cause and effect. In the SEM, combining exogenous variables and other endogenous variables, the interactions between the two variables are distinguished as cause and effect and are estimated simultaneously. The estimates, which I will further discuss below, are therefore not subject to simultaneity bias as they are in an OLS regression. Whether or not the effects are significant is determined by the hypothesis tests on the estimates of the parameters. As in an OLS regression, a failure to reject the null hypothesis on an estimate in the SEM implies that the hypothesized effect is not statistically significant. Since the simultaneity bias which makes a hypothesis test of an OLS estimate invalid is excluded, the hypothesis tests are valid. Besides the interactions between job accessibility and employment ratio, the interactions between job accessibility and auto ownership, and between auto ownership and employment ratio are subject to simultaneity bias as well in an OLS regression. Like the mutual cause and effect between job accessibility and
employment ratio, these two interactions are split as two direct effects and are estimated simultaneously in the SEM. Therefore, the estimates of these direct effects are not subject to simultaneity bias in the SEM.

2.3.1.2 Indirect Impacts of Endogenous Variables

The endogenous variables affect each other not only directly, but also indirectly through feedback loops. The indirect impacts are very important in the spatial mismatch context because they are a part of the endogeneity. Through structural equation modeling, the magnitude of indirect impacts can be quantified and separated from the direct impacts. Combined with the direct impacts, they will quantitatively address the endogeneity. In this model, I assume the following indirect impacts exist:

- Employment ratio → Income → Auto ownership
- Employment ratio → Income → Auto ownership → Job accessibility
- Employment ratio → Income → Job accessibility
- Employment ratio → Income → Job accessibility → Auto ownership
- Employment ratio → job accessibility → Auto ownership
- Employment ratio → Auto ownership → Job accessibility
- Job accessibility → Employment ratio → Income
- Job accessibility → Employment ratio → Income → Auto ownership
- Job accessibility → Employment ratio → Auto ownership
- Job accessibility → Auto ownership → Employment ratio → Income
- Job accessibility → Auto ownership → Employment ratio
- Auto ownership → Job accessibility → Employment ratio
• Auto ownership → Job accessibility → Employment ratio → Income
• Auto ownership → Employment ratio → Job accessibility
• Auto ownership → Employment ratio → Income
• Auto ownership → Employment ratio → Income → Job accessibility

2.3.1.3 Impacts of Exogenous Variables

Socio-economic variables are widely used in travel behavior and spatial mismatch modeling (Cervero et al., 2000; Ihlanfeldt and Sjoquist, 1990a, 1991; Kockelman, 1997; McLafferty et al., 1992; Ong and Blumenberg, 1998; Ong and Miller, 2003; Preston et al., 1998; Rachael, 1998; Schimek, 1996;). In the SEM context, they are exogenous variables and affect endogenous variables. Each endogenous variable may be affected by different exogenous variables. An exogenous variable may affect different endogenous variables in different ways. For example, household characteristics affect income, auto ownership, or job accessibility separately or all of them simultaneously.

Besides the conceptual considerations of the impacts of exogenous variables, including exogenous variables in the structural equation models is technically indispensable, to meet the conditions for model identification. The possible variables and their definitions are listed in Table 2.3.

Impacts of Competing Laborers. It seems to be intuitive that competing laborers will impose a direct negative impact on the employment ratio (Raphael, 1998). Job accessibility represents job opportunities, i.e., how many jobs are available to a census tract, while competing laborers represent the total number of laborers who compete for those jobs. Theoretically, all laborers within the research area, and even outside the
research area, may compete for the jobs accessible to a specific census tract. However, the competition for jobs will become weaker with the increase of distance. To be practical for the SMH testing, it is suitable to take into account the competing laborers within the research boundary. Thus, each census tract will have a measured number of competing laborers. The calculation method will be the same as Equation (2.3) except for

Table 2.3 Exogenous variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Potentially affected endogenous variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competing laborers</td>
<td>Total persons age 16-64, calculated by Equation (2.3), replacing jobs with laborers</td>
<td>Employment ratio; income</td>
</tr>
<tr>
<td>Education</td>
<td>Percentage of persons with college or higher degree</td>
<td>Employment ratio; income</td>
</tr>
<tr>
<td>Household size</td>
<td>Average household size</td>
<td>Auto ownership; income</td>
</tr>
<tr>
<td>Car insurance premium</td>
<td>Annual car insurance premium</td>
<td>Auto ownership</td>
</tr>
<tr>
<td>Transit availability*</td>
<td>A product of bus stops, buses per hour per stop, bus capacity, and service factor</td>
<td>Auto ownership; employment ratio</td>
</tr>
<tr>
<td>Rental housing percentage</td>
<td>Rental housing units divided by total housing units</td>
<td>Job accessibility</td>
</tr>
<tr>
<td>Median rent</td>
<td>Median rent in a census tract</td>
<td>Job accessibility</td>
</tr>
<tr>
<td>Single headed households</td>
<td>Percentage of single headed households with children</td>
<td>Employment ratio</td>
</tr>
</tbody>
</table>

* Light rail is included.

the replacement of Jobs with Laborers which is the total number of persons who are 16-to-64 years old in each tract. Besides affecting employment ratio, competing laborers will also negatively affect income. From the definition of competing laborers, it is easy to see that this variable will not directly affect auto ownership, or job accessibility.
Impacts of Education. Education represents job skills, and thus positively affects employment ratio and income (Blumenberg and Shiki, 2003; Ihlanfeldt and Sjoquist, 1989, 1990a, 1990b; Massey and Mullan, 1984). In the conceptual model, this variable is measured by the percentage of persons aged 16-64 with a college or higher degree. Some empirical studies (Bhat and Koppelman, 1993; Meurs, 1993; Raphael and Rice, 2002) include an education variable in car ownership logit and OLS models. I argue that education associates with car ownership through its impacts on income, rather than imposing direct impacts. I have not found any empirical studies on the causality between education and job accessibility, and assume that any such relationship is also indirect, through income.

Impacts of Household Size. Empirical studies (Meurs, 1993; Kockelman, 1996) have found that household size has significant positive impacts on auto ownership: the larger households will more likely have more vehicles. However, if auto ownership is represented by average number of vehicles per person, the impact of household size will be negative, reflecting the greater opportunities for intra-household sharing of vehicles in larger households, as well as the increased presence of non-drivers (children and elders) in the denominator. The impacts of household size on income are similar to those on auto ownership: positive if income is measured per household; negative if measured per capita. Household size itself doesn’t directly affect job accessibility or employment ratio.

Impacts of Car Insurance Premium. Ong and Miller (2003) found that car insurance premiums had significant negative impacts on auto ownership in Los Angeles. It might also be true in other metropolitan areas, although the size of the geographic area of the analysis may affect the variation of auto insurance premiums.
**Impacts of Transit Availability.** Transit is a substitute mode for private autos. The central city generally has better transit service than the suburban areas. Ong and Miller (2003) found that transit availability (which was a product of buses per hour, seats per bus and load factor) had significant negative impacts on auto ownership in Los Angeles. In this paper, I define transit availability as a product of transit (bus and light rail) stops, transit vehicles per hour, vehicle capacity and a service factor. The service factor is used to adjust the area with walk access in a census tract, and is the percentage of the transit-served area in a census tract. The absence of transit service from the neighborhoods dominated by black residents in the suburban areas is thought by Kain (1968) to be a main reason of low employment ratios among blacks. Therefore, transit availability is assumed to have positive impacts on employment ratio.

**Impacts of Rental Housing Percentage.** In conventional land use zoning, rental housing is generally apart from single family residential areas and closer to employment establishments. Renters have lower moving costs than home owners, and may respond to job changes by relocation. Rental housing percentage is thus expected to positively affect job accessibility. Rental housing percentage may correlate with employment ratio, income, and auto ownership because neighborhoods with high rental housing percentages may attract lower income households. However, it should not be thought of as a direct determinant of employment ratio, income, or auto ownership.

**Impacts of Median Rent.** The median rent is mainly determined by the characteristics of rental units and neighborhood environment. It is expected to affect people’s residential location choices and thus job accessibility. Units with higher rent are usually in neighborhoods which are more attractive and farther from employment centers.
Therefore, the sign of the impact on job accessibility is assumed to be negative. Since both rent and expenditures on an automobile are positively correlated with income, they have a spurious positive correlation with each other. Obviously, rent level is not a determinant of income or employment ratio, although it may be correlated with each of them.

**Impacts of Single-headed Households.** Some empirical studies (Blumenberg and Shiki, 2003; Blumenberg and Ong, 1998) find that single-headed households with children have higher unemployment rates than non-single headed households with children. In this case, as an exogenous variable, single-headed households with children (measured as the percentage of single-headed households with children in a census tract) is expected to have a negative impact on employment ratio, reflecting the potential impact on employment of having to provide child care and meet other domestic responsibilities alone. It is assumed not to directly affect auto ownership, job accessibility, or income.

2.3.1.4 *Empirical Test of the Conceptual Model*

This conceptual model describes a structural equation system (Figure 2.1). It is good for describing the causality between endogenous variables, and between exogenous variables and endogenous variables for whites or blacks. This conceptual model can be estimated with cross-sectional data at the aggregate level for whites only and blacks only separately, using commercial structural equation modeling software such as LISREL, EQS or AMOS. With regard to an empirical test, a major concern is how to represent jobs in job accessibility. The ideal measure of jobs is thought to be job openings (Ihlanfeldt,
and Sjoquist, 1998), but this is difficult to measure. Some empirical studies show that total job growth (Raphael, 1998) and total jobs (Sawicki and Moody, 2000, Shen, 2001) are reasonable alternatives. In travel demand forecast modeling, total jobs are widely used to calculate job accessibility. Generally speaking, the number of jobs in a metropolitan travel model comes from the Census Transportation Planning Package (CTPP) that organizes the employment data by “place of work” and by industry. An alternative employment data source is InfoUSA employment survey data. It is a nearly 100% sample survey database with detailed spatial information and industry information by the North American Industry Classification System (NAICS). It provides a researcher with an opportunity to classify jobs by job skills. Furthermore, this database is available by year, and thus can be used to calculate job growth by firms and by NAICS category.

A separate estimation of the model for whites or blacks will only give the direct impacts and indirect impacts, not the difference between whites and blacks. The $\chi^2$ test in SEM makes the comparison of differences between the structural coefficients for blacks and whites possible. In AMOS, the separate SEMs for blacks and whites can be estimated at one time. The t-test for each pair of parameters will show whether an individual coefficient for blacks is significantly different from that for whites. The final log-likelihood of this segmented model can be compared to that for the pooled model, and a Chi-square test conducted to test whether the relationships collectively differ by segment. In particular, a finding that the direct, or especially the total, impact of job accessibility on employment ratio was significantly larger for whites than for blacks would be considered evidence supporting the SMH.
2.3.2 Unconditional Change-Score Model of the SMH

The unconditional change-score model was originally used in multiple regression analyses with panel data (Finkel, 1995). It assumes that the relationships between dependent variables and independent variables remain stable over time. In other words, the coefficients of the independent variables at time t have the same sign and magnitude as those at time t - 1 (Finkel, 1995). This model can test some assumptions that cannot be tested by the cross-sectional model.

First, it measures the impacts of changes in the independent variables on the change of dependent variable(s) directly. In an unconditional change-score model, the change of a variable is the difference in its values measured at two time points in the same panel. In a cross-sectional model, the changes either are not measurable or are measured as inter-unit changes due to the limitations of cross-sectional data. The determinants of change in the dependent variable derived from a cross-sectional model are weaker in addressing the causality of the change in the dependent variable.

Second, the impacts of the independent variables with constant effects, regardless of whether they are observed, are cancelled out. In other words, some variables have constant impacts on the dependent variable over time and do not explain any changes in the dependent variable.

Third, an unconditional change-score model does not have to take into account cross impacts of independent variables at time t - 1 on the dependent variables at time t. Although an unconditional change-score model is generally applied in multiple regression analysis, its premise is also applicable to structural equation modeling. Figure 2.2 is constructed according to this idea.
In an unconditional change-score SEM, all the relationships between endogenous variables and between endogenous variables and exogenous variables can be examined in the form of their changes. The relationships between job accessibility and employment ratio will be paid extra attention. In this model, the change in job accessibility is interpreted as net growth of job opportunities. If job growth, not job level, affects employment ratio, then the growth of job accessibility will have positive impacts on the change in the employment ratio. Combined with the results from the cross-sectional SEM, it becomes possible to determine which measure of job accessibility has a more significant impact on employment, job level or job growth.
2.4. CONCLUSIONS

Classic multiple regression models and logistic regression models are widely used in social studies and provide reliable estimates when the assumptions of the models are not violated. According to economic theories and some empirical studies, we can conclude that the assumptions for applying conventional multiple regression models and logistic regression models are violated in the spatial mismatch context due to the endogeneity of residence and auto ownership. Ignoring the endogeneity in empirical studies makes the estimates in these models unreliable. Structural equation modeling is superior to the conventional single equation modeling methods in terms of its ability to solve the endogeneity bias problem through integrating job accessibility, employment ratio, and auto ownership into a simultaneous equation system, and sampling bias by using all cases rather than a selective subset. Furthermore, SEM has the ability to quantitatively distinguish direct effects and indirect effects, and statistically show the differences of the effects across ethnic groups of interest.

The cross-sectional model tests the impacts of the endogenous and exogenous variables, and the endogeneity of job accessibility and auto ownership. The unconditional change-score model tests the impacts of job growth. They test two different dimensions of the SMH. The data for testing the models are available for most cities and urban regions. Therefore, rather than being just conceptual models, they are testable.
References:


CHAPTER THREE

EXPLORING THE CAUSAL CONNECTIONS AMONG JOB ACCESSIBILITY, EMPLOYMENT, INCOME, AND AUTO OWNERSHIP USING STRUCTURAL EQUATION MODELING

Abstract

Using structural equation modeling, this study empirically examines the causal connections between job accessibility, workers per capita, income per capita, and autos per capita at the aggregate level with year 2000 census tract data in Sacramento County, California. Under the specification of the conceptual model, the model implied covariance matrix exhibits a reasonably good fit to the observed covariance matrix. The direct and total effects show that job accessibility has a negative effect on autos per capita, autos per capita has a positive effect on workers per capita and income per capita, workers per capita has a positive effect on income per capita and autos per capita, and education attainment has a positive effect on workers per capita, income per capita and autos per capita. Job accessibility has a negative total effect on workers per capita, income per capita and autos per capita. These results are largely consistent with theory and/or with empirical observations across a variety of geographic contexts. They suggest that structural equation modeling is a powerful tool for capturing the endogeneity among
job accessibility, employment, income and auto ownership, and has other advantages over linear regression in this context.

Key words: structural equation modeling, endogeneity, job accessibility, spatial mismatch
3.1. INTRODUCTION

Job access or job accessibility, which is usually defined as the product of number of jobs and a travel friction factor (Kockelman, 1997; Levinson, 1998; Raphael, 1998), is paid great attention in regional transportation and land use planning. From the perspective of transport, improving job accessibility can be achieved by improving the transportation infrastructure, thereby decreasing the friction of travel. For example, increasing the transit service area will increase job accessibility by transit, and increasing highway capacity will increase job accessibility by personal vehicle. From the perspective of land use, an increase in job accessibility can be achieved by implementing certain land use policies. For example, commercial/residential mixed use zoning may substantially increase the job accessibility in some traffic analysis zones (TAZs). Obviously, the two approaches are not mutually exclusive.

Besides the number of jobs and the travel friction factor, the third element associated with job accessibility is location. Location in a travel model usually refers to a traffic analysis zone (TAZ), which has an explicit geographic boundary. In real estate economics, location is interpreted as a consumption bundle including the neighborhood characteristics (physical environment, school quality, crime rate, zoning, access to services and jobs, etc.). To a household, choosing a residential location is choosing a consumption bundle. On the one hand, the willingness to pay for the bundle is subject to constraints on income, transportation availability, and commute distance allocation between the household members if a household has more than one worker. In this respect, the job accessibility of a household’s residential location is causally affected by these constraints. On the other hand, when a household makes a work location choice or makes
a decision about auto ownership, job accessibility will be a determinant in those decisions. In other words, job accessibility, employment, and auto ownership are interdependent, and a change in one of them will cause changes in others. From the perspective of econometrics, it is quite possible that the error term in the equation for one variable will not be independent from the observed values of the other variables. If so, a correct approach to describe the interdependence among the variables is to develop a simultaneous equation system instead of a single equation (Ihliefeldt and Sjoquist, 1990b).

Unfortunately, in policy analyses, professional practice, and empirical studies, the intervening interactions are often treated as unidirectionally causal and thus are simulated with classic linear regression models. Kockelman (1997) studied the impact of job accessibility on auto ownership in the San Francisco Bay Area and pointed out that job accessibility negatively affected auto ownership (number of autos in a household divided by the number of members in the household that are five years of age or older). Using regression, Schimek (1996) found a positive impact of the number of workers in a household on auto ownership. Raphael (1998) analyzed the impact of job accessibility on employment among youth who live with their parents, and pointed out that employment was significantly affected by job accessibility. Applying regression analysis to 2000 census tract data in Los Angeles, Ong and Miller (2003) found that job accessibility was a determinant of employment ratio for neither whites nor African-Americans.

According to statistical theory, if the assumption of independence of explanatory variables from error terms (i.e. that the covariance of explanatory variables with errors is 0) is violated, and the ordinary least squares (OLS) procedure is applied to estimate the
parameters of a simultaneous equation system, the estimators will be biased, inconsistent, and inefficient; furthermore, hypothesis tests on parameters will be invalid (Ramanathan, 2002). The numerical value of the cumulative bias will be contingent upon the magnitude and sign of the interdependence between the endogenous variables (Mayston, 2005).

For this reason, Ihlanfeldt and Sjoquist (1998) criticized the neglect of endogeneity (or simultaneity, i.e. the correlations of the explanatory variables with the errors) in many empirical studies regarding the impact of job accessibility on employment, and attributed the insignificant impact of job accessibility on employment to the neglect of endogeneity and incorrect modeling methods. They proposed two approaches: using a sample that is less subject to endogeneity (e.g. youth who live with their parents), and incorporating endogeneity into a system of simultaneous equations (Ihlanfeldt and Sjoquist, 1990b). Obviously, the second approach is more appealing than the first one because it is not subject to limitations in sampling and findings are more generalizable.

Based on theoretical analysis, Gao et al. (2006) proposed a structural equation model (SEM) that incorporates not only endogeneity but more generally, relationships among job accessibility, employment, income, and auto ownership. In this study, I empirically implement the SEM proposed by Gao et al. (2006), and compare the results with those obtained from linear regressions.

3.2. CONCEPTUAL MODEL

From the perspective of economic theory, the relationships among job accessibility, employment, and auto ownership are interdependent, and may be mutually causal. Land use patterns determine locations of jobs and residences and accordingly affect job
accessibility. Income, auto ownership, employment and rent levels affect residential location choice and job accessibility. Job accessibility, education attainment and auto ownership are expected to have positive impacts on employment, while living in a single-headed household with children is hypothesized to have negative impacts on employment due to heavy family responsibilities. It is plausible to assume that auto ownership is a function of job accessibility, employment, income, and household size. Income is a function of employment, education attainment, and household size. From the perspective of econometrics, job accessibility, employment, income, and auto ownership constitute the endogenous variables of an equation system, in which the explanatory variables are
not independent from the error terms in an equation. In the literature, the computation of the variables takes the form of either per-household (e.g. Schimek, 1996) or per-capita (e.g. Kockelman, 1997). In our context, however, expressing variables on a per-household basis confounds the effect of the variable itself with the effect of household size, which varies by census tract. For example, households tend to be smaller in census tracts close to the central business district (CBD) than in those farther away. Similarly, household income also tends to be smaller in central-city census tracts. The increase in household income from inner city to outer city is at least partially attributable to there being more workers in a household. To avoid the possible overestimation of structural coefficients between the endogenous variables due to the impact of household size as a common factor of the three variables at the household level, I choose to compute the variables on per-capita basis. Household size is taken as exogenous and a determinant of employment, income and auto ownership. In this way, the direct effects of household size on these variables can be distinguished from the direct effects of other determinants. However, this leads to two competing empirical models when the conceptual model is estimated, which will be further discussed below. Figure 3.1 illustrates the conceptual model. A single arrow represents a cause-effect relation from one variable to another variable. Reciprocal arrows represent the possible interactions between two endogenous variables. The + and – signs represent the expected directions of impacts.

To compare the results of the SEM with those of linear regression models, each of the four endogenous variables in the SEM is written as a linear function of the variables which are hypothesized to have a direct effect on it:

\[
job\ accessibility = \beta_{10} + \beta_{11} \ast percentage\ of\ rental\ housing\ units + \beta_{12} \ast median\ rent
\]
\[ + \beta_{13} * \text{workers per capita} + \beta_{14} * \text{income per capita} \]

\[ + \beta_{15} * \text{autos per capita} + \epsilon_{\text{job}} \]

\[ \text{workers per capita} = \beta_{20} + \beta_{21} * \text{education attainment} + \beta_{22} * \text{household size} \]

\[ + \beta_{23} * \text{percentage of single-headed households} \]

\[ + \beta_{24} * \text{job accessibility} + \beta_{25} * \text{autos per capita} + \epsilon_{\text{workers}} \]

\[ \text{income per capita} = \beta_{30} + \beta_{31} * \text{education attainment} + \beta_{32} * \text{household size} \]

\[ + \beta_{33} * \text{workers per capita} + \epsilon_{\text{income}} \]

\[ \text{autos per capita} = \beta_{40} + \beta_{41} * \text{household size} + \beta_{42} * \text{job accessibility} \]

\[ + \beta_{43} * \text{workers per capita} + \beta_{44} * \text{income per capita} + \epsilon_{\text{autos}} \]

where the \( \epsilon \)s are the error terms for each equation.

### 3.3. DATA

I take Sacramento County, California as the example to test the conceptual model. There are seven cities in Sacramento County. Except for two small cities (small in geographic size and population), the other five cities are spatially contiguous. The City of Sacramento is the central city and traditional employment center. The other cities are relatively new and tend to absorb decentralized newer jobs. The population distribution and job distribution in the study area have the characteristics of a typical urban county, and therefore I think it is reasonable to choose this area for an empirical study.

This analysis is done at the census tract level, with each census tract taken as an observation. There were 279 census tracts in Sacramento County in the year 2000. Three tracts are excluded from the sample because they constitute military bases and a state prison. Thus, the sample size for the initial analysis is 276, which is reasonably large for
structural equation modeling (Kline, 2005). I use census tract 2000 data purchased from GeoLytics, Inc. to extract socio-economic variables, and the year 2000 travel demand forecasting model used by the Sacramento Area Council of Governments (SACOG) to obtain the number of jobs for each tract, and travel time (A.M peak period) between traffic analysis zones (TAZs).

The descriptive statistics of the variables are shown in Table 3.1 for the final sample of 266 tracts (the deletion of 10 tracts from the initial 276 is discussed below). Median rent (median asking rent per month) and income per capita (dollars per year) are imported directly from the data CD. Percentage of rental housing units is defined as the percentage of renter-occupied housing units out of the total occupied housing units. Education attainment is calculated by dividing total persons who have a bachelor’s or higher degree by total population in a census tract, which can be interpreted as persons

Table 3.1 Descriptive statistics of the variables (N = 266)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of rental housing units</td>
<td>1.59</td>
<td>99.24</td>
<td>40.31</td>
<td>20.93</td>
</tr>
<tr>
<td>Median rent</td>
<td>378.00</td>
<td>1461.00</td>
<td>747.80</td>
<td>173.64</td>
</tr>
<tr>
<td>Percentage of single-headed households with children</td>
<td>1.00</td>
<td>28.00</td>
<td>11.44</td>
<td>5.16</td>
</tr>
<tr>
<td>Education attainment</td>
<td>0.01</td>
<td>0.54</td>
<td>0.16</td>
<td>0.11</td>
</tr>
<tr>
<td>Household size</td>
<td>1.30</td>
<td>4.34</td>
<td>2.72</td>
<td>0.52</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>17663.00</td>
<td>71263.00</td>
<td>41486.76</td>
<td>9250.32</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>0.23</td>
<td>0.74</td>
<td>0.45</td>
<td>0.08</td>
</tr>
<tr>
<td>Income per capita</td>
<td>6754.00</td>
<td>49729.00</td>
<td>21615.80</td>
<td>8174.85</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>0.33</td>
<td>0.85</td>
<td>0.63</td>
<td>0.11</td>
</tr>
</tbody>
</table>
with a college degree per capita. *Workers per capita* is computed by dividing the total number of employed persons by total population. *Household size* is calculated by dividing the total population by total number of households. *Autos per capita* is computed by dividing the total number of vehicles by total population. *Percentage of single-headed households with children* is defined as the percentage of single-headed households with children out of the total number of households.

*Job accessibility* at the census tract level is approximated based on the job accessibility at the TAZ level in SACOG’s travel model. Following Kockelman’s (1997) method, I assume that households in a census tract have no travel friction to access the jobs in the same tract, but do have travel friction to access jobs in other tracts. Thus, the job accessibility to a census tract is the sum of jobs in that tract, and jobs in other tracts discounted by the travel friction between them. Mathematically, it is written as below:

\[
\text{CTJobAccessibility}_m = \sum_{i=1}^{n_m} \text{Job}_i + \left( \sum_{i=1}^{n_m} \sum_{j\in\text{CTm}, TAZ\neq CTm}^{N-n_m} \text{Job}_i \* (1/\text{TravelTime}_{ij}) \right) / n_m \tag{3.5}
\]

where \(\text{CTJobAccessibility}_m\) is the job accessibility in census tract \(m\) (CT\(_m\)); \(i = 1, 2, \ldots n_m\), which is the number of TAZs contained in census tract \(m\); \(\text{Job}_i\) is the total jobs in TAZ \(i\); and \(\text{TravelTime}_{ij}\) is the travel time from TAZ \(i\) to TAZ \(j\) (where its inverse is the travel friction).

Thus, in the right side of Equation (3.5), the first part represents the total job count of all TAZs contained in census tract \(m\). The second part approximates the average job accessibility of a census tract to all other census tracts (actually, TAZs) in the county.
1/n_m in the second part is used to penalize the census tracts which consist of two or more TAZs so that the job accessibility in a census tract which consists of one TAZ is comparable to that in a census tract which consists of more than one TAZ. Otherwise, the job accessibility in the census tracts containing two or more TAZs will be considerably overestimated. That is important in this study because in Sacramento County, 109 out of 276 census tracts consist of only one TAZ; 149 census tracts consist of 2 to 5 TAZs; and 18 census tracts consist of 6 to 21 TAZs.

3.4. COMPETING SEM MODELS

Under the specification of the conceptual model, I obtained two equally good competing models in terms of significance tests on parameter estimates and goodness-of-fit of the model. Controlling for education attainment, percentage of single-headed households and autos per capita, job accessibility is a determinant of workers per capita in the first competing model while household size is a determinant of workers per capita in the second one, and all other relationships are the same. Including job accessibility and household size in the equation for workers per capita makes the direct effects of both variables insignificant (note that in each model, the competing variables are only significant at the 0.1 level). From the perspective of policy analysis, the first model implies that access to employment opportunities increases the odds of employment while the second model implies that a larger household size will increase the odds of employment. The first model has stronger theoretical grounds and is more useful for policy analysis. The SEM results presented in this paper are based on the first model.
(although the second one comes into play later). The direct effects and total effects of the second model are reported as appendices.

3.5. CONFORMANCE TO MULTIVARIATE NORMALITY

In maximum likelihood estimation of the SEM, which is used in this study, the multivariate normality of all variables is a key assumption. Simulation and empirical studies (Andreassen et al., 2006; Hu and Bentler, 1995; West et al., 1995; Yuan et al., 2005) have demonstrated that excessive nonnormality inflates the $\chi^2$-statistic and Root Mean Square Error of Approximation (RMSEA), and deflates some model fit indices like the Normed Fit Index (NFI) and the Comparative Fit Index (CFI), therefore leading to more rejections of the hypothesized model than are warranted. Furthermore, nonnormality leads to underestimation of standard errors of the parameter estimates and thus inflated t-statistics, which leads to more rejections of the null hypothesis in tests on parameters than it should. Some insignificant coefficients under multivariate normality will erroneously appear to be significant due to the inflation of the t-statistics under multivariate nonnormality. In contrast, the effects of slight nonnormality are negligible (Lei and Lomax, 2005). Multivariate normality can be tested in many ways (Bollen, 1989; D’Agostino, 1989; Mardia, 1970), but Mardia’s coefficients of skewness and kurtosis are used most often in structural equation modeling software. In PRELIS, a companion software package to LISREL, multivariate normality is measured by Mardia’s PK, an omnibus multivariate normality measure based on skewness and kurtosis. If the Mardia’s PK is smaller than 3, a sample is considered to be multivariate normally distributed (Siekpe, 2005). In AMOS 5, multivariate normality is measured by Mardia’s coefficient
of multivariate kurtosis, which is asymptotically distributed as \( N(0, 1) \). Therefore, a sample is considered to be multivariate normally distributed at the 0.05 level of significance if the critical ratio of Mardia’s coefficient of multivariate kurtosis is smaller than 1.96 (Mardia, 1970). Unfortunately, the simulation and empirical studies on nonnormality generally report only univariate nonnormality instead of multivariate nonnormality (Andreassen et al., 2006; Hu and Bentler, 1995; Lei and Lomax, 2005; West et al., 1995; Yuan et al., 2005). Therefore, in empirical studies (Andreassen et al., 2006), univariate normality is often used as the basis for assessing normality. A single variable is generally considered to be moderately skewed if the absolute value of its skewness index is smaller than 2 and the absolute value of kurtosis is smaller than 7, and excessively skewed if the absolute value of its skewness index is more than 3 and the kurtosis is around 21 (Curran et al., 1996). However, since univariate normality for each variable is only a necessary but not sufficient condition for multivariate normality, a small univariate kurtosis for each variable does not automatically preclude having large multivariate kurtosis. In contrast to the univariate case, cutoffs for slight, moderate and excessive multivariate nonnormality have not been established in the literature, to our best knowledge.

In this study, the skewness index of the variables varies between -1.03 and 0.97, and the kurtosis index varies between -0.40 and 4.16. Thus, the univariate distributions are considered to depart from normality only slightly (Curran et al., 1996; Lei and Lomax, 2005). However, compared with the critical value for a normal distribution (critical ratio = 1.96 at \( \alpha = 0.05 \)), the multivariate kurtosis (92.49) and critical ratio (54.60) are relatively large. Transformation (log, square root, Box-Cox) of the variables substantially
improves neither their univariate normality indices nor the multivariate kurtosis. Taking into account the desirability of having easily interpretable coefficients, I decided to use the raw data instead of the transformed data.

To minimize the risk of inflating the significance test in this case due to a relatively large multivariate kurtosis, I run the model with the original sample to obtain, for each observation, the Mahalanobis distance, which represents the distance of the vector of an observation from the vector of sample means for all variables. Since Mardia’s multivariate kurtosis is defined as the difference between the sample average of the Mahalanobis distance raised to the 4th power and the expected value of the distance to the 4th power (Bollen, 1989; Mardia, 1970), the larger the distance is, the larger the contribution an observation is making to the departure from multivariate normality. Removing the outliers will reduce the multivariate kurtosis and thus the critical ratio. I remove five outliers at a time and observe the changes to the critical ratio and goodness-of-fit. After 10 outliers are removed, the critical ratio becomes 10.73, and the multivariate kurtosis is 16.36. Only after 58 outliers are removed does the critical ratio fall below the 1.96 threshold (1.92), by which point the coefficients of one exogenous and two endogenous variables in the model have become insignificantly different from 0 at the 0.1 level of significance.

I note that, for our sample, every time outliers are removed, the $\chi^2$-statistic of the model on the reduced sample, including the sample with the best multivariate normality (N = 218), becomes larger than for the model on the original sample (N = 276). Given that the $\chi^2$-statistic is the product of the sample size minus one (N-1) and the minimized discrepancy function ($F_{\text{min}}$), a larger $\chi^2$-statistic with a smaller sample means an increase
of F_{min}, i.e., a greater discrepancy between the sample covariance matrix and the one implied by the SEM, i.e., a worse fitting model. Further, after examining the attributes of the excluded observations, I find that these observations are census tracts of great interest in policy analyses. These tracts are “outliers” on job accessibility and workers per capita, which are two key endogenous variables. The means of job accessibility and workers per capita for the reduced sample are close to those of the original sample, but the variances of the two variables are substantially smaller. Therefore, it is not surprising that the direct effects of job accessibility on workers per capita, and workers per capita on autos per capita are not significantly different from 0 when the sample size is reduced to 218.

As mentioned above, the sample consists of census tracts and thus is not a random sample. The “outliers” represent the consequences of land use and some policy factors. For example, downtown Sacramento is the main job center and hence has the highest job accessibility; subsidies for rental housing for low-income households lead to extremely low rent in some census tracts. Removing observations from the sample implies the loss of influence of the land use patterns, transportation network, and other policies. From this perspective, it is undesirable to remove any observation. On the other hand, it is also undesirable to have false inferences about causality between the variables due to the inflation of t-statistics. Therefore, some compromise is appropriate between the need to take full advantage of what the original data can tell us and the need for statistical confidence in what the data do tell us. Andreassen et al. (2006) discussed the effects of nonnormality on parameter estimates, the \( \chi^2 \)-statistic, the RMSEA and t-statistics. But they only reported univariate skewness and kurtoses and did not address what cutoffs for
Table 3.2 Normality evaluation of the variables in the final model (N = 266)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Skewness</th>
<th>Critical ratio of skewness</th>
<th>Kurtosis</th>
<th>Critical ratio of kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of rental housing units</td>
<td>0.52</td>
<td>3.44</td>
<td>-0.16</td>
<td>-0.52</td>
</tr>
<tr>
<td>Median rent</td>
<td>1.07</td>
<td>7.15</td>
<td>1.61</td>
<td>5.35</td>
</tr>
<tr>
<td>Education attainment</td>
<td>0.93</td>
<td>6.18</td>
<td>0.30</td>
<td>1.00</td>
</tr>
<tr>
<td>Household size</td>
<td>0.15</td>
<td>0.99</td>
<td>0.16</td>
<td>0.52</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>0.25</td>
<td>1.64</td>
<td>0.33</td>
<td>1.10</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>-0.36</td>
<td>-2.39</td>
<td>0.26</td>
<td>0.85</td>
</tr>
<tr>
<td>Income per capita</td>
<td>0.75</td>
<td>4.99</td>
<td>0.34</td>
<td>1.12</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>-0.42</td>
<td>-2.82</td>
<td>-0.46</td>
<td>-1.52</td>
</tr>
<tr>
<td>Multivariate</td>
<td></td>
<td>16.36</td>
<td></td>
<td>10.73</td>
</tr>
</tbody>
</table>

Multivariate kurtosis and critical ratio would yield a reasonably accurate estimation. Due to differences in the number of variables and the range of univariate kurtosis, their normality assessment is not very helpful for us to determine our multivariate kurtosis cutoffs in this study. Bagley and Mokhtarian (2002) discussed the tradeoff between the sample size and conformance to multivariate normality. In their case, the removal of 100 out of 615 observations led to a reduction of the critical ratio (72.28) to the desirable level (1.96) while the outcomes were not substantially affected. Their finding suggests that even when the multivariate distribution substantially departs from normal, the significance tests for the parameter estimates may still be robust. Therefore, I think that the sample retaining 266 observations (i.e. having removed the 10 most egregious outliers, dropping the multivariate critical ratio from 54.59 to 10.73) should produce a reasonably good estimation while keeping as much information as possible from the original data.
The results reported in this paper are based on those 266 observations. The results of the multivariate normality evaluation are shown in Table 3.2.

### 3.6. CORRELATIONS OF THE ERROR TERMS

In the conceptual model, four endogenous variables are connected through several reciprocal loops. This structure implicitly suggests possible correlations among the error terms of the endogenous variables. It is logical to include the correlations between the error terms of the endogenous variables in the covariance matrix. The correlations between the error terms for *income per capita* and *job accessibility*, and *income per capita* and *autos per capita*, are demonstrated to be insignificant in the initial model and constrained to be zero in the final model. In the final model, the correlations between the error terms for *job accessibility* and *workers per capita*, *workers per capita* and *income per capita*, *workers per capita* and *autos per capita*, and *job accessibility* and *autos per capita* are -0.14, -0.87, -0.25 and 0.71, respectively and are all significant at the 0.01 level.

Allowing the correlations between the error terms of the endogenous variables is imperative because if they exist, the OLS parameter estimates of Equations (3.1), (3.2), (3.3) and (3.4) are inefficient (Greene, 1997). In other words, the standard errors of the parameter estimates will tend to be inflated. Thus, even if endogeneity bias were not a problem, and the OLS parameter estimates were unbiased and consistent, the significance tests on those parameter estimates would be unreliable in the presence of correlated error terms.
3.7. GOODNESS OF FIT OF THE SEM

In contrast to a linear regression model, the SEM does not have a unique goodness of fit measure that is widely accepted. Following the principles suggested by Bollen and Long (1993), Hoyle and Panter (1995) and Shah and Goldstein (2006), I report the model fit indices from several different index families. In Table 3.3, the saturated model is the model in which no constraints are placed on the population moments, and which fits the data perfectly. The independence or null model is the model which assumes that there are no correlations at all between the observed variables. The hypothesized model (called the default model in AMOS 5) is the final model for which I report our results. Therefore, the model fit index of the hypothesized model is between the two extremes. The closer to the saturated model and the farther from the independence model the fit index of the hypothesized model is, the better the hypothesized model.

According to this principle, we can see that all the fit indices in Table 3.3 suggest a good fit of the hypothesized model. Among these indices, the Expected Cross-validation Index (ECVI) and Akaike Information Criterion (AIC) are less often reported in the literature. ECVI measures a single sample approximation of the cross-validation coefficients, i.e. the generalizability of a solution obtained in one sample to an independent sample from the same population (Browne and Cudeck, 1989; Shah and Goldstein, 2006). The value of the ECVI of the hypothesized model is very close to that of the saturated model, suggesting a good generalizability of this model. The AIC adjusts the $\chi^2$-statistic by penalizing model complexity, and favors explaining the covariance with as few parameters as possible. In our case, the AIC shows that our parsimonious model explains the covariance relatively successfully.
Table 3.3 Selected model fit indices

<table>
<thead>
<tr>
<th>Model fit index</th>
<th>Independence model</th>
<th>Hypothesized model</th>
<th>Saturated model</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>2105.00</td>
<td>81.64</td>
<td>0.00</td>
<td>Measuring the discrepancy between the sample and population covariance matrices; the smaller, the better; sample size dependent</td>
</tr>
<tr>
<td>Degrees of freedom (d.f.)</td>
<td>36.00</td>
<td>6.00</td>
<td>0.00</td>
<td>Assessing the proportion of the variability in the sample covariance matrix explained by the model; GFI &gt; 0.9 suggests a good fit.</td>
</tr>
<tr>
<td>Goodness of fit index (GFI)</td>
<td>0.36</td>
<td>0.94</td>
<td>1.00</td>
<td>Not parsimony adjusted; normed; NFI &gt; 0.9 suggests a good fit.</td>
</tr>
<tr>
<td>Normed fit index (NFI)</td>
<td>0.00</td>
<td>0.96</td>
<td>1.00</td>
<td>Assessing the improvement of the hypothesized model over the independence model; IFI &gt; 0.9 suggests a good fit.</td>
</tr>
<tr>
<td>Incremental fit index (IFI)</td>
<td>0.00</td>
<td>0.96</td>
<td>1.00</td>
<td>Assuming noncentral chi-square distribution; assessing the improvement of the hypothesized model relative to the independence model. 0.90 or higher suggests a good fit.</td>
</tr>
<tr>
<td>Comparative fit index (CFI)</td>
<td>0.00</td>
<td>0.96</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Akaike information criterion (AIC)</td>
<td>2121.00</td>
<td>141.64</td>
<td>72.00</td>
<td>Parsimony adjusted; the closer to the value of saturated model, the better the hypothesized model.</td>
</tr>
<tr>
<td>Expected cross-validation index (ECVI)</td>
<td>8.00</td>
<td>0.53</td>
<td>0.27</td>
<td>Assessing the generalizability of a solution obtained in a sample to an independent sample from the same population. The smaller the value, the better the hypothesized model.</td>
</tr>
</tbody>
</table>

3.8. DIRECT EFFECTS IN THE SEM

AMOS 5 reports the direct effects and total effects in both unstandardized form and standardized form. An unstandardized coefficient retains scaling information and is interpreted as the number of units change in the “effect” variable in response to a unit
change in the “cause” variable when all other “cause” variables are held constant. A standardized coefficient is a transformation of the corresponding unstandardized coefficient from which the scaling information is removed. As a consequence of this transformation, it is interpreted as the number of standard deviations change of the “effect” variable in response to a unit standard deviation change of the “cause” variable when all other “cause” variables are held constant (Hoyle, 1995). I prefer to report the standardized direct, indirect and total effects in this study for three reasons. Firstly, the purpose of this study is more to understand the causal relations among the variables, and to demonstrate a method that incorporates endogeneity in a complex context than to predict the quantitative change of one variable upon the changes of other variables. Secondly, a wrong choice of the functional form may change not only the parameter estimates, t-statistics and model goodness of fit, but also the relative importance of the variables in a model. I want to test both aspects and using standardized effects makes it easier to assess the latter in this study. Thirdly, using standardized effects will make it easier for other researchers to compare our findings with theirs.

The direct effect of a variable is its structural coefficient, and is interpreted as the initial response (i.e. without taking into account any feedback effect through the loops) of the “effect” variable to the change in a “cause” variable. Figure 3.2 and Table 3.4 show the significant direct effects of the final model (the insignificant coefficients in the initial model are treated as 0 and are not included in the final model). As predicted in the conceptual model, *job accessibility* is significantly affected by *percentage of rental housing units* positively, and *median rent* negatively. This implies that the census tracts
with more renter occupied housing units (which are generally multi-family housing units) tend to have higher job accessibility. According to the Sacramento County general land use plan, the areas within or close to commercial or industrial land uses have more multi-family land use designations. Therefore, I would like to interpret the impact of rental housing units on job accessibility as partly a consequence of the land use policies in metropolitan planning. The negative correlation between median rent and job accessibility seems to be contradictory to classic location theory, which states that the higher the access to the urban center, the higher the rent is. Since I do not have enough
information about the rental housing market in the study area and do not control the
determinants of the rent in census data, such as square feet per housing unit, school
quality, crime rate, and livability of the neighborhood, I do not know whether the rent per
square foot in the place with higher job accessibility is truly lower than that in places with
lower job accessibility. This needs further study in the future.

In the conceptual model, we expect that the census tracts with higher income per
capita will have lower job accessibility because the richer neighborhoods are usually
farther from jobs. This assumption is not supported by the data in Sacramento County. In
Sacramento County, downtown Sacramento and Rancho Cordova are the two most
important employment centers. Areas closer to these two centers have higher job
accessibility. Thus, the suburban areas in this county, whose land use is dominated by
residential instead of employment centers (commercial or industrial), do not have high
job accessibility. By cross-tabulating the observations upon medians of job accessibility,
median rent, percentage of rental housing units and income per capita and mapping the
observations in a Geographic Information System (GIS), I find that after controlling for
median rent and percentage of rental housing units, 67 of 118 observations have higher
job accessibility and higher income per capita while the remaining 51 have higher job
accessibility and lower income per capita. Spatially, those census tracts with high income
per capita are very close to downtown Sacramento. Therefore, I think the positive
influence of income per capita on job accessibility correctly represents what is observed.

Why is the sign positive? Since I do not have individual socio-economic and
attitudinal data for each household, I cannot give an explicit explanation of this result.
Besides the possible self-selection in residence, another possible reason relates to the
Table 3.4 Standardized direct effects of the SEM

<table>
<thead>
<tr>
<th></th>
<th>Percentage of rental housing units</th>
<th>Median rent</th>
<th>Education attainment</th>
<th>Household size</th>
<th>Job accessibility</th>
<th>Workers per capita</th>
<th>Income per capita</th>
<th>Autos per capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job accessibility</td>
<td>0.64***</td>
<td>-0.21***</td>
<td>--</td>
<td>--</td>
<td>-</td>
<td>0.38***</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>--</td>
<td>--</td>
<td>0.29***</td>
<td>-</td>
<td>0.05*</td>
<td>--</td>
<td>0.56***</td>
<td>--</td>
</tr>
<tr>
<td>Income per capita</td>
<td>--</td>
<td>--</td>
<td>0.40***</td>
<td>-</td>
<td>--</td>
<td>0.76***</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>-0.60***</td>
<td>-0.80***</td>
<td>0.17***</td>
<td>0.37***</td>
<td>--</td>
</tr>
</tbody>
</table>

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

*: hypothesized but not statistically significant and therefore constrained to be 0 in final model
**: no direct effect hypothesized in conceptual model
distribution of one-person households. I note that the distribution of the proportion of
one-person households in the census tracts is similar to that of job accessibility. A higher
proportion of one-person households leads to a larger income per capita. In addition,
housing amenities, such as square feet per housing unit, school quality and parks, are
much less important to one-person households than to larger ones. High access to jobs, as
well as shopping and entertainment sites, seems to be more important to those one-person
households.

As predicted in the conceptual model, education attainment, job accessibility and
autos per capita positively affect workers per capita. Education attainment indicates job
skills. Higher job skills increase the opportunities to be hired. Autos per capita represents
mobility. Higher autos per capita can be interpreted as fewer constraints on
transportation in accessing or retaining jobs far away from the residence. The direct effect
of job accessibility on workers per capita is a key structural path. Its positive sign implies
that higher job accessibility will lead to more workers, which is highly desirable from a
policy analysis standpoint. It is noted that the direct effect of job accessibility on workers
per capita is significant at the 0.1 level. As discussed in the section on conformance to
multivariate normality, the magnitude of this parameter estimate and its significance are
sensitive to the sample distribution. When the sample has a multivariate normal
distribution (N = 218), it is not significant even at the 0.1 level. When the sample has a
relatively large multivariate kurtosis (N = 276), the coefficient is significant at the 0.01
level. In this respect, we should be cautious when we interpret the causal relationship
between these two variables. Compared with education attainment and autos per capita,
the positive effect of job accessibility on workers per capita is minor.
The influences on autos per capita are quite straightforward. The census tracts having higher income per capita have higher autos per capita, and the census tracts having higher workers per capita have higher autos per capita since having more workers implies a higher need for autos for commuting, all else being equal. The negative impact of job accessibility on autos per capita implies that in tracts having high job accessibility, households tend to make use of alternative modes of transportation and have lower dependence on personal vehicles.

In the 2000 census tract data, household income includes incomes from all sources. Due to lack of information on income from non-work sources, I cannot split income into work and non-work sources. I only use two variables to explain the variance of income in this study. The positive sign and large magnitude of workers per capita suggest that a job is the main source of income to the majority of households. Education attainment significantly increases income per capita.

The direct effects of workers per capita and autos per capita on job accessibility, percentage of single headed households on workers per capita and household size on income per capita have the same signs as predicted, but are not significant. Therefore, these four effects are constrained to zero in the final SEM. Thus, the direct reciprocal interaction between job accessibility and autos per capita, and job accessibility and workers per capita in the conceptual model are not supported by the data in Sacramento County. Autos per capita and workers per capita do not affect job accessibility directly in this case.
3.9. INDIRECT EFFECTS AND TOTAL EFFECTS IN THE SEM

In a recursive path model (a model without a feedback loop between any two variables), the indirect effect of a variable in the SEM refers to the changes in an “effect” variable in response to a unit change of a “cause” variable through one or more intervening variables, while direct effects are taken as constant (Hayduk, 1987). When coefficients are standardized, an indirect effect can be expressed as the product of all the direct effects involved in the path. If there is more than one indirect path, the indirect effect of a variable will be the sum of the indirect effects on all paths from the “cause” variable to the “effect” variable. In a nonrecursive SEM (a model with at least one feedback loop between two endogenous variables), the indirect effect becomes much more complicated because any change in one variable will cause changes in the variables in the feedback loop until equilibrium is reached. If equilibrium cannot be reached\(^1\), the hypothesized model will be empirically unidentifiable.

In the final model of this study, all four endogenous variables are connected through loops, and each endogenous variable has at least one indirect path leading to it. For example, household size directly affects autos per capita, but its effect is passed over to job accessibility through autos per capita and income per capita. The change in job accessibility caused by the indirect effect of household size in turn affects autos per capita. For this reason, besides the direct effects on an endogenous variable, an exogenous variable will affect other endogenous variables indirectly through the indirect paths. The same is true for any endogenous variable. For example, income per capita

\(^1\) The stable state of a nonrecursive model system is measured by the System Stability Index in AMOS 5. The widely accepted cutoff for the stability index is 1 (Kline, 2005). If the index is 1 or smaller, the system is thought to be stable. If a system is not stable, the estimates are thought to be unreliable. For this study, the stability index is 0.46.
does not affect workers per capita directly, but indirectly through three paths (i.e. income per capita → job accessibility → workers per capita, income per capita → autos per capita → workers per capita, and income per capita → job accessibility → autos per capita → workers per capita). AMOS 5 outputs the total indirect effects instead of a decomposition of indirect effects for each path, which makes it difficult to assess the role of the intervening variables in each indirect path. As can be seen clearly from Table 3.5, the indirect effects do not necessarily have the same sign as direct effects and are not necessarily smaller than direct effects in magnitude.

The total effect of a “cause” variable on an “effect” variable is the sum of its direct effect and indirect effect, and hence its sign and magnitude are determined by the signs and relative magnitudes of the direct and indirect effects. When it has the same sign as the direct effect, an indirect effect will strengthen the direct impacts of a “cause” variable on an “effect” variable (i.e. the magnitude of the total effect is larger than that of the direct effect or the indirect effect.). From Tables 3.4, 3.5, and 3.6, we can see that 8 out of the 12 total effects whose direct effects are significant have larger magnitudes (absolute values) than the corresponding direct effects due to the synergism of indirect effects, while 4 out of those 12 total effects are the net outcome of opposing direct and indirect effects. It is quite reasonable that workers per capita, income per capita and autos per capita have positive total effects on all endogenous variables.

It is noted that the magnitude of the negative indirect effect of job accessibility on workers per capita is far larger than its corresponding positive direct effect, leading to a negative sign on the total effect. As noted earlier, the direct effect should be interpreted with caution, since it has a lower significance and may be less robust in the presence of
non-normality than the other effects in the model. However, even if the direct effect were altogether negligible, the negative indirect (and hence total) effect is an important result deserving further discussion. Studies of spatial mismatch often suggest that increased employment among economically-disadvantaged groups could be achieved by bringing jobs and workers closer together (Kain, 1968; Raphael, 1998), i.e. by increasing job accessibility. This model suggests that such an approach will have an effect exactly opposite to the intended: i.e. it will reduce the number of workers rather than increase it. Why the apparently counter-intuitive effect? Figure 3.2 shows that the sizable negative indirect effect of job accessibility on workers per capita occurs through its negative effect on autos per capita (which has a positive direct effect on workers per capita). In other words, the higher the job accessibility, the lower the autos per capita, and the lower the autos per capita, the lower the workers per capita. According to our model, job skills (i.e. education attainment) are more important to employment (workers per capita) than is the direct positive impact of distance to work (job accessibility). If the subsidized housing only concentrates in some specific neighborhoods, and the transportation policy focuses only on transit, such policies might in the longer run be depriving disadvantaged households of the superior accessibility to the larger number and variety of more distant jobs that an automobile makes possible. This interpretation is supported by Shen and Sanchez’s (2005) finding that welfare recipients’ odds of employment were substantially improved by increasing car ownership compared to changing residential location.

Besides having direct effects on their corresponding endogenous variables, the exogenous variables have indirect and total effects on other endogenous variables. 

*Percentage of rental housing units* has negative total effects on *workers per capita*, 

income per capita and autos per capita, and affects autos per capita more than workers per capita and income per capita. This implies that census tracts with higher percentages of rental housing units will tend to have lower autos per capita. Education attainment has the biggest positive total effect on income per capita and smallest positive total effect on autos per capita. Household size has a larger negative total effect on autos per capita than on any of the other three endogenous variables. The negative sign is expected since larger households achieve economics of scale with respect to auto ownership due to a greater ability to share. Two single-person households are more likely to own two autos between them than is one two-person household. Thus, the smaller households are on average, the greater auto ownership per capita tends to be.

3.10. THE SEM VERSUS LINEAR REGRESSION MODELS

In the context of this study, a main concern is the endogeneity of several important variables, which makes the parameter estimates in linear regression models biased, inconsistent and inefficient. Theoretically, a SEM, in which all the coefficients are estimated simultaneously, has the ability to disentangle the complicated correlations among the endogenous variables and error terms. Based upon the discussions in previous sections, we can conclude that the hypothesized model, in which the endogeneity is explicitly accounted for, generates plausible results for direct effects and total effects, and the existence of endogeneity is empirically demonstrated. Therefore, it is of interest to compare the results obtained from the SEM and linear regression methods: i.e. how misleading would the results be if the four equations of the SEM were naively estimated one at a time using OLS?
Table 3.5 Standardized indirect effects of the SEM

<table>
<thead>
<tr>
<th></th>
<th>Percentage of rental housing units</th>
<th>Median rent</th>
<th>Education attainment</th>
<th>Household size</th>
<th>Job accessibility</th>
<th>Workers per capita</th>
<th>Income per capita</th>
<th>Autos per capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job accessibility</td>
<td>-0.09</td>
<td>0.03</td>
<td>0.26</td>
<td>-0.12</td>
<td>-0.14</td>
<td>0.34</td>
<td>0.02</td>
<td>0.19</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>-0.30</td>
<td>0.10</td>
<td>0.07</td>
<td>-0.39</td>
<td>-0.51</td>
<td>0.16</td>
<td>0.06</td>
<td>0.09</td>
</tr>
<tr>
<td>Income per capita</td>
<td>-0.23</td>
<td>0.08</td>
<td>0.27</td>
<td>-0.30</td>
<td>-0.36</td>
<td>0.12</td>
<td>0.05</td>
<td>0.50</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>-0.58</td>
<td>0.19</td>
<td>0.10</td>
<td>-0.08</td>
<td>-0.10</td>
<td>0.08</td>
<td>-0.30</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Table 3.6 Standardized total effects of the SEM

<table>
<thead>
<tr>
<th></th>
<th>Percentage of rental housing units</th>
<th>Median rent</th>
<th>Education attainment</th>
<th>Household size</th>
<th>Job accessibility</th>
<th>Workers per capita</th>
<th>Income per capita</th>
<th>Autos per capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job accessibility</td>
<td>0.55a</td>
<td>-0.18a</td>
<td>0.26</td>
<td>-0.12</td>
<td>-0.14</td>
<td>0.34</td>
<td>0.40b</td>
<td>0.19</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>-0.30</td>
<td>0.10</td>
<td>0.36b</td>
<td>-0.39</td>
<td>-0.46b</td>
<td>0.16</td>
<td>0.06</td>
<td>0.65b</td>
</tr>
<tr>
<td>Income per capita</td>
<td>-0.27</td>
<td>0.08</td>
<td>0.67b</td>
<td>-0.30</td>
<td>-0.36</td>
<td>0.88b</td>
<td>0.05</td>
<td>0.50</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>-0.58</td>
<td>0.19</td>
<td>0.10</td>
<td>-0.69b</td>
<td>-0.90b</td>
<td>0.25b</td>
<td>0.07a</td>
<td>0.14</td>
</tr>
</tbody>
</table>

a: opposing direct and indirect effects
b: synergistic direct and indirect effects
Table 3.7 Standardized coefficients of the linear regression models

<table>
<thead>
<tr>
<th></th>
<th>Percentage of rental housing units</th>
<th>Median rent</th>
<th>Education attainment</th>
<th>Household size</th>
<th>Job accessibility</th>
<th>Workers per capita</th>
<th>Income per capita</th>
<th>Autos per capita</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job accessibility</td>
<td>0.44***</td>
<td>-0.34***</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>0.14*</td>
<td>0.47***</td>
<td>-0.34***</td>
<td>0.50</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>--</td>
<td>--</td>
<td>0.19***</td>
<td>-0.26***</td>
<td>-</td>
<td>--</td>
<td>0.50***</td>
<td>--</td>
<td>0.67</td>
</tr>
<tr>
<td>Income per capita</td>
<td>--</td>
<td>--</td>
<td>0.86***</td>
<td>-</td>
<td>--</td>
<td>0.10***</td>
<td>--</td>
<td>--</td>
<td>0.86</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>-0.24***</td>
<td>-0.32***</td>
<td>0.30***</td>
<td>0.47***</td>
<td>--</td>
<td>0.81</td>
</tr>
</tbody>
</table>

*: p < 0.1, **: p < 0.05, *** p < 0.01
-: hypothesized but not statistically significant and therefore constrained to be 0 in final model
--: not hypothesized in the conceptual model
Table 3.8 Percentage difference in magnitude between the coefficients of the OLS models and the direct effects of the SEM (base)

<table>
<thead>
<tr>
<th></th>
<th>Percentage of rental housing units</th>
<th>Median rent</th>
<th>Education attainment</th>
<th>Household size</th>
<th>Job accessibility</th>
<th>Workers per capita</th>
<th>Income per capita</th>
<th>Autos per capita</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Job accessibility</strong></td>
<td>-30.93</td>
<td>60.38</td>
<td></td>
<td></td>
<td>c</td>
<td>22.40</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td><strong>Workers per capita</strong></td>
<td></td>
<td>-33.80</td>
<td>c</td>
<td>d</td>
<td></td>
<td>-11.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Income per capita</strong></td>
<td></td>
<td>115.54</td>
<td></td>
<td></td>
<td></td>
<td>-86.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Autos per capita</strong></td>
<td></td>
<td>-60.13</td>
<td>-60.20</td>
<td></td>
<td>77.51</td>
<td>28.42</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

c: hypothesized but not statistically significant and therefore constrained to be 0 in final SEM
d: hypothesized but not statistically significant and therefore constrained to be 0 in final linear regression model

Table 3.9 Percentage difference in magnitude between the coefficients of the OLS models and the total effects of the SEM (base)

<table>
<thead>
<tr>
<th></th>
<th>Percentage of rental housing units</th>
<th>Median rent</th>
<th>Education attainment</th>
<th>Household size</th>
<th>Job accessibility</th>
<th>Workers per capita</th>
<th>Income per capita</th>
<th>Autos per capita</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Job accessibility</strong></td>
<td>-20.15</td>
<td>85.79</td>
<td></td>
<td></td>
<td>-58.70</td>
<td>17.21</td>
<td>-278.95</td>
<td></td>
</tr>
<tr>
<td><strong>Workers per capita</strong></td>
<td>-46.48</td>
<td>-33.50</td>
<td>d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-22.96</td>
</tr>
<tr>
<td><strong>Income per capita</strong></td>
<td>28.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-88.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Autos per capita</strong></td>
<td>-64.96</td>
<td>-64.56</td>
<td></td>
<td></td>
<td>21.95</td>
<td>571.43</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
d: hypothesized but not statistically significant and therefore constrained to be 0 in final linear regression model
The results of the final single-equation linear regressions are shown in Table 3.7². Comparing Tables 3.4 and 3.7, we can see that the parameter estimates in the linear regressions have the same signs as those of the direct effects in the SEM (the direct effect of household size on workers per capita in the final SEM is obtained in the competing model shown in the appendix). Even the insignificant parameter estimates in the initial linear regression models have the same signs as their counterparts in the initial SEM. The coefficients significant in the SEM are also significant in the final linear regressions as well. So far, this is encouraging.

Further comparing Tables 3.4 and 3.7, we can find at least three differences between the results of the SEM and the linear regression models. First, parameter estimates have different magnitudes in the SEM and the linear regression models. Table 3.8, in which the parameter estimates of the direct effects in the SEM are used as the base, shows the percentage differences in absolute magnitudes of the parameter estimates in the two types of models. Five out of 11 coefficients in the linear regressions are larger than in the SEM (not counting the three which are zero in the SEM) while 6 out of 11 are smaller. Except for one variable, the magnitude of the regression coefficients differs from that of the SEM direct effects by more than 22% in one direction or the other. The average difference in magnitude (absolute value) is 53.39%, and the biggest difference is 115.54%.

It might be argued, however, that a more appropriate comparison of the OLS models is to the total effects of the SEM, since a given coefficient in a single regression model might be capturing indirect as well as direct effects. In Table 3.9, therefore, the base for

---

² The OLS models (3.1), (3.2), (3.3) and (3.4) were initially estimated with full specifications to identify the insignificant parameters. They were then re-run, excluding the insignificant parameters, to obtain the final models.
comparison is the total effects of the SEM. Due to the addition of the indirect effects, the pattern of the percentage differences between the total effects in the SEM and the coefficients in the linear regressions is different from that based on the direct effect. The absolute value of the difference in percentage points varies between 17.21 and 571.43, and the average difference is 123%. Second, not only that, the relative importance of some variables is changed. For example, the coefficient of *percentage of rental housing units* on *job accessibility* is 68.42% larger than that of *income per capita* in the SEM, while it is 6.38% smaller in the linear regression model. Similar examples include the coefficients of *education attainment* and *job accessibility* on *workers per capita* and the coefficients of *education attainment* and *workers per capita* on *income per capita*. Third, the significance tests for some parameter estimates are different. The direct effects of *autos per capita* and *workers per capita* on *job accessibility* in the SEM are statistically zero ($p > 0.1$) while they are not zero in the OLS model. In the SEM, either household size (competing model) or job accessibility, but not both, has a significant effect on *workers per capita*. In the OLS model, household size has a significant impact on *workers per capita* while *job accessibility* does not. Obviously, these differences will lead to different conclusions about the relative importance of these variables in policy analysis.

### 3.11. DISCUSSION AND CONCLUSIONS

The main purpose of this study is to demonstrate a generalizable structural equation model, both to portray the causal connections among job accessibility, employment, income and auto ownership, and to confirm a violation of the assumption of the independence of the included explanatory variables from the error terms when a linear
regression model is applied in this context. Following the requirements of the SEM for normality, model specification, identification and assessment of model fit, I estimate the direct, indirect and total effects of the SEM with aggregate data at the census tract level. The model fit indices show that the model-implied covariance matrix is reasonably close to the observed sample covariance matrix. Therefore, the hypothesized model cannot be rejected. The failure to reject the SEM implies that the four key variables are connected through feedback loops instead of unidirectional causal links as suggested in linear regression models (Equations (3.1), (3.2), (3.3) and (3.4)) and therefore endogeneity among the key variables is demonstrated. Thus, the linear regression models should be rejected. Econometrically, the endogeneity bias leads to three consequences. First, the coefficients are either overestimated or underestimated. From the analysis in the previous section, we can see that the biases are substantial for most coefficients. Second, the relative magnitudes of the coefficients are distorted due to the overestimation for some parameters and underestimation for the others. This distortion is also severe for several variables in this case. Third, the significance tests for the coefficients are not reliable. Furthermore, in the SEM, the discrepancy function between the model-implied covariance matrix and the sample covariance matrix is minimized using all the information such as the covariances between the endogenous variables, the correlations between the exogenous variables and the correlations between the error terms in the sample, while in linear regression, only a part of the information is used to estimate the coefficients. Therefore, the structural coefficients and indirect effects in the SEM contain richer information than do the coefficients in the linear regression models. Accordingly,
the results generated by the SEM are more trustworthy than those obtained by the linear regression models.

As far as the interpretation of the results and policy implications are concerned, the SEM has at least two advantages over linear regression models. First, the SEM explicitly shows the direction of the impact from one variable to another due to its unique covariance structure while the linear regression model does not have the ability to distinguish between effects in both directions. For example, in the SEM, although I assume a reciprocal loop between job accessibility and autos per capita, the model shows that the direction of the impact is from job accessibility to autos per capita instead of the opposite or both. In the linear regression Equations (3.2) and (3.4), the coefficients between job accessibility and autos per capita are both significant, which can be interpreted as a unidirectional effect when only one equation is involved or bidirectional when two equations are involved. In other words, the inferred direction of the impact is contingent upon which model is used, and the correct inference (based on the SEM) of job accessibility $\rightarrow$ autos per capita but not conversely, would only be made by chance, if Equation (3.4) happened to be the only one estimated. Obviously, this could cause confusion about which relationship(s) should be taken into account in policy analysis. Second, the SEM shows, besides the direct effect of one variable on another variable, the indirect effect and total effect of one variable on all variables. The indirect effects, especially the indirect effects in a nonrecursive model like the one in this study, greatly increase the complexity of the model, but provide a complete picture of the causal relations among the variables. According to these effects, it is easy to identify how and
how much the effects of one variable are passed on to other variables, and the relative magnitudes of the effects on all the endogenous variables.

In terms of the relationships among the endogenous variables, the context of this study is similar to that of the spatial mismatch hypothesis (Ihlanfeldt and Sjoquist, 1998; Kain, 1968), which states that residential segregation plus a transportation barrier makes it harder for African-Americans to access jobs in suburban areas than for whites. The method and findings of this study will shed some light on the dispute whether there is a causal relation between job accessibility and employment. As I have shown, endogeneity among job accessibility, employment and auto ownership does exist, and the endogeneity bias is substantial and should not be neglected. For this reason, the results produced by linear regression models should not be used as evidence to support whether job accessibility has an impact on employment. SEM is a proper approach to solve the endogeneity issue. We can expect that, under the hypothesized model, African-Americans and whites will have different direct and total effects. In addition, when SEM is employed, the samples need not be limited to youth who live with their parents, which has been used to minimize the impact of endogeneity bias in linear and logistic regression models.
References:


Gao, Shengyi, Robert A. Johnston, and Patricia Mokhtarian. 2006. A conceptual approach to testing the spatial mismatch hypothesis. Unpublished manuscript.


## APPENDIX

Table 3.10 Standardized direct effects of the competing SEM

<table>
<thead>
<tr>
<th></th>
<th>Percentage of rental housing units</th>
<th>Median rent</th>
<th>Education attainment</th>
<th>Household size</th>
<th>Job accessibility</th>
<th>Workers per capita</th>
<th>Income per capita</th>
<th>Autos per capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job accessibility</td>
<td>0.64***</td>
<td>-0.21***</td>
<td>--</td>
<td>--</td>
<td>-</td>
<td>0.38***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workers per capita</td>
<td>--</td>
<td>--</td>
<td>0.30***</td>
<td>0.04*</td>
<td>--</td>
<td>--</td>
<td>0.51***</td>
<td></td>
</tr>
<tr>
<td>Income per capita</td>
<td>--</td>
<td>--</td>
<td>0.40***</td>
<td>--</td>
<td>0.76***</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autos per capita</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>-0.60***</td>
<td>-0.80***</td>
<td>0.17***</td>
<td>0.37***</td>
<td></td>
</tr>
</tbody>
</table>

* p < 0.1, ** p < 0.05, *** p < 0.001
-: hypothesized but not statistically significant and therefore constrained to be 0 in final model
--: no direct effect in conceptual model

Table 3.11 Standardized total effects of the competing SEM

<table>
<thead>
<tr>
<th></th>
<th>Percentage of rental housing units</th>
<th>Median rent</th>
<th>Education attainment</th>
<th>Household size</th>
<th>Job accessibility</th>
<th>Workers per capita</th>
<th>Income per capita</th>
<th>Autos per capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job accessibility</td>
<td>0.55</td>
<td>-0.18</td>
<td>0.26</td>
<td>-0.12</td>
<td>-0.14</td>
<td>0.33</td>
<td>0.40</td>
<td>0.17</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>-0.30</td>
<td>0.10</td>
<td>0.36</td>
<td>-0.39</td>
<td>-0.46</td>
<td>0.12</td>
<td>0.03</td>
<td>0.58</td>
</tr>
<tr>
<td>Income per capita</td>
<td>-0.27</td>
<td>0.08</td>
<td>0.67</td>
<td>-0.30</td>
<td>-0.36</td>
<td>0.86</td>
<td>0.03</td>
<td>0.44</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>-0.58</td>
<td>0.19</td>
<td>0.10</td>
<td>-0.69</td>
<td>-0.90</td>
<td>0.24</td>
<td>0.06</td>
<td>0.12</td>
</tr>
</tbody>
</table>
CHAPTER FOUR
WHAT ARE THE TRADEOFFS IN deleting observations
TO OBTAIN A MULTIVARIATE NORMAL DISTRIBUTION? AN
EMPIRICAL STUDY

Abstract

In this study, I examine the generalizability of a conceptual model built on census
tract data, test the cutoffs suggested in the literature for reliable estimates and hypothesis
testing statistics at the census tract (CT) and block group (BG) levels, and evaluate the
efficacy of deleting observations as an approach to improving multivariate normality in
structural equation modeling. The results show that the conceptual model can be applied
to a BG sample. The parameter estimates and standard errors of the parameter estimates
at the BG level are less sensitive to changes in sample size and multivariate normality
than are those at the CT level. The cutoffs for nonnormality suggested in the literature do
not perform well in this study. I argue that pursuing a multivariate normal distribution by
deleting observations should be balanced against loss of model power in the
interpretation of the results.

Key words: multivariate nonnormality; structural equation modeling; maximum
likelihood estimation; outliers
4.1. INTRODUCTION

Structural equation modeling (SEM) has been theoretically and empirically demonstrated to be powerful in disentangling the complexity of causality among variables in social studies. As with other statistical methods, assuming conceptual plausibility, the inferences of causality in the SEM are based on hypothesis tests on the model and the parameter estimates. If the data meet all the assumptions required by an estimation method, the results are assumed to be reliable.

In SEM, one of the main concerns about the data is whether the sample has a multivariate normal distribution, because that determines what estimation method will be used and to what extent the estimates are reliable. Generally speaking, data in the real world do not have even univariate, let alone multivariate, normal distributions (Micceri, 1989). Therefore, it seems more desirable for a researcher to choose a non-normal theory-based estimation method at the conceptual level (Browne, 1982, 1984) if the multivariate normal distribution is not found at the data screening stage. However, the asymptotic distribution free (ADF) method, a non-normal theory-based estimation method, needs a large sample size. In practice, when the sample size is not big enough or the sample is not severely non-normally distributed, the ADF approach is found to be not necessarily superior to normal theory-based estimation methods in terms of minimizing estimation biases (Chou et al., 1991; Muthén and Kaplan, 1992). For these reasons, the normal theory-based estimation methods are more popular than ADF in empirical studies.

Thus, the empirical studies are at risk of violating one of the key assumptions in SEM when a normal theory-based estimation method is applied to a sample with a multivariate nonnormal distribution. To minimize the risk, researchers may take two
approaches (Yuan et al., 2000). To those researchers focusing on the methodology of estimation, the hypothesis test statistics for the parameter estimates and the model can be improved by incorporating the effects of nonnormality into the estimators of the statistics (Browne, 1984; Dolan and Molenaar, 1991; Henly, 1993; Hu et al., 1992; Satorra and Bentler, 1988, 1994; Yuan and Bentler, 1997; Yuan et al., 2000). When this is done, as long as the nonnormality is not extreme, the hypothesis tests are still robust. However, the requisite improvements in methodology may not be available in all SEM software or are not well documented in some SEM software (for example, AMOS does not tell the user which estimator is used for the standard errors of a parameter estimate in its documentation). Therefore, to researchers focusing on applications of SEM, improving the normality of the data appears more practical. Transforming the raw data (Andreassen, 2006; Bollen, 1989) and deleting outliers (Bagley and Mokhtarian, 2000), which contribute greatly to the departure from the normal distribution and distort the covariance matrix (Bollen, 1987; Yuan and Bentler, 2001), are often recommended or used in empirical studies.

Transforming (e.g. square root, logarithm, Box-Cox) raw data reduces the multivariate skewness and kurtosis of all variables collectively by reducing the univariate skewness and kurtosis of each individual variable. According to our experience, when the univariate nonnormality is severe, transformation will substantially reduce the univariate skewness and kurtosis. When univariate nonnormality is moderate or slight, transformation has only a minor effect. However, slight nonnormality of the individual variables may still lead to a large multivariate skewness or kurtosis. Therefore, transformation alone is unlikely to lead to a multivariate normal distribution. In addition,
transformation implies that the relationships of a variable with other variables are assumed to be curvilinear instead of linear. This assumption may not be true. It is not uncommon for a researcher to find that the hypothesized SEM becomes worse in terms of model fit indices, or even worse, that the model is empirically unidentifiable after some variables are transformed. In this case, even if transformation improves normality \textit{per se}, it is not helpful for improving the model as a whole. Therefore, the role of transformation needs to be assessed on a case-by-case basis.

In contrast to transformation, deleting outliers focuses on lowering \textit{multivariate} skewness and kurtosis. In AMOS, multivariate normality is measured by Mardia’s multivariate kurtosis. The outliers are indicated by their Mahalanobis distances, which represent the squared distance, in standard units, of the vector of an observation from the vector of sample means for all variables. The larger the distance is, the larger the contribution an observation is making to Mardia’s multivariate kurtosis and hence to the departure from multivariate normality. Deleting an outlier will decrease Mardia’s multivariate kurtosis. Outliers can be deleted until the multivariate kurtosis index reaches the desired level. The advantage of deleting outliers is that it retains the assumption of linear relationships among the variables. The disadvantage of deleting outliers is obvious: it means loss of observations, and hence information and model power. At the extreme, the results could be a Procrustean model trimmed to fit allegedly “typical” cases, but which ignores the departures from “typical” that characterize real empirical data. Therefore, deleting outliers until multivariate normality is reached may also be undesirable.
Even if choosing transformation to improve multivariate normality, a researcher is unlikely to obtain a sample with a multivariate normal distribution without also deleting observations. He will often estimate the parameters and assess the model with maximum likelihood under some degree of multivariate nonnormality. Therefore, what is important is not whether a sample has a multivariate normal distribution, but to what extent the hypothesis tests on the parameter estimates and the model are reliable when the data are multivariate nonnormal. In other words, we want to know what level of multivariate nonnormality is acceptable for a reasonably accurate (i.e. less biased) estimation of parameters, standard errors and chi-square statistic. If this level of multivariate normality is reached by deleting outliers, to what extent is model power lost due to the loss of information in the deleted observations?

In this study, I try to make use of a unique property of census data to accomplish three goals: (1) to test the generalizability of the model I built with a census tract (CT) sample in our previous study (Gao et al., 2006b, i.e. Chapter Three), (2) to test the cutoffs suggested in the literature for reliable estimates and test statistics, at the CT and the block group (BG) levels, and (3) to evaluate the efficacy of deleting observations as an approach to improving multivariate normality.

4.2. UNIVARIATE NORMALITY VERSUS MULTIVARIATE NORMALITY

---

3 Census tract and census block group data are most often used in social studies. Spatially, a census tract may have more than one census block group. For a given study area, a sample consisting of census block groups has more observations than a sample consisting of census tracts. Since a census block group spans a smaller space than a census tract, the households in a census block group are more homogeneous than those in a census tract. Thus, the two samples will have the same sample means. But the census block group sample will have larger variation and accordingly standard deviations.
For a sample, the univariate normal distribution of a variable is assessed by its measures of skewness ($b_1$) and kurtosis ($b_2$), which are written as (Bollen, 1989, p.420):

$$b_1 = \frac{(m_4)}{(m_2)^2}$$

(4.1)

$$b_2 = \frac{m_4}{(m_2)^2}$$

(4.2)

$$m_r = \frac{\sum (X - \bar{X})^r}{N}$$

(4.3)

where $m_r$ is the $r$th sample moment.

$b_1$ describes how much a distribution departs from symmetry and $b_2$ describes the peakedness of a distribution. For a univariate normal distribution, $b_1$ is equal to 0 and $b_2$ is equal to 3 (in AMOS 5, the measure of kurtosis is equal to $(b_2 - 3)$, thus the kurtosis of a normal distribution is 0. For convenience of discussion, I will follow the rule in AMOS 5). Univariate nonnormality simply means that at least one of the measures of skewness and kurtosis is not 0.

A multivariate normal distribution implies that each variable in a sample has a univariate normal distribution and each pair of variables has a bivariate normal distribution (Hayduk, 1987). Multivariate normality can be measured in many ways, but Mardia’s (1970) coefficients of multivariate skewness and kurtosis or an omnibus measure based on both coefficients (for example, Mardia’s PK in PRELIS) are most
commonly used in structural equation modeling software. In AMOS 5, only Mardia’s coefficient of multivariate kurtosis is available and is computed as:

Mardia’s multivariate kurtosis

\[
\begin{align*}
\text{Mardia's multivariate kurtosis} &= \frac{1}{N} \sum_{i=1}^{N} \left[ (x_i - \bar{x})\hat{S}^{-1}(x_i - \bar{x}) \right]^2 - \frac{p(p+2)(N-1)}{N + 1} \\
&= \frac{1}{N} \sum_{i=1}^{N} \left[ (x_i - \bar{x})\hat{S}^{-1}(x_i - \bar{x}) \right]^2 - \frac{p(p+2)(N-1)}{N + 1} \\
\end{align*}
\]

(4.4)

where \(X_i\) is the \(i\)-th observation on the \(p\) observed variables, \(\bar{X}\) is the vector of the means, \(\hat{S}^{-1}\) is the unbiased estimate of the population covariance matrix, \(p\) is the number of observed variables, and \(N\) is the number of observations. The first part is usually written as \(b_{2,p}\). The second part is the expected value of \(b_{2,p}\), and is usually written as \(\beta_{2,p}\).

Associated with Mardia’s coefficient are its standard error and critical ratio. They are calculated as:

standard error of multivariate kurtosis = \(\sqrt{\frac{8p(p+2)}{N}}\)  

(4.5)

critical ratio (c.r.)

\[
= \text{Mardia’s multivariate kurtosis} / \text{standard error of multivariate kurtosis}
\]

(4.6)

The critical ratio of Mardia’s multivariate kurtosis is asymptotically distributed as \(N(0, 1)\). Thus, a sample can be considered to be multivariate normally distributed at the 0.05 level of significance when the critical ratio is smaller than 1.96, indicating that the coefficient of multivariate kurtosis is not significantly different from zero (Mardia, 1970).
From the above equations and definitions, it can be seen that univariate normality describes the distribution of only one variable in the sample while multivariate normality describes the joint distribution of all variables in the sample. A univariate normal distribution for each variable does not guarantee a multivariate normal distribution for the group of them. Specifically, the univariate normal distribution of each variable is a necessary, but not sufficient, condition for having a multivariate normal distribution (West et al., 1995). However, univariate nonnormal distributions for each variable will generally result in a multivariate nonnormal distribution.

What if a sample with a moderate size does not have a multivariate normal distribution and the ML estimation is used? It has been shown that as long as the nonnormality is not extreme, the ML parameter estimates are still unbiased and consistent, but are inefficient (Bollen, 1989; Browne, 1982, 1984; Yuan et al., 2005). Nonnormality leads to an overestimation of the chi-square statistic (indicating the degree of discrepancy between the model-implied and sample-derived covariance matrices), potentially leading to false rejection of the model as whole, and the underestimation of standard errors of parameter estimates, leading to inflated t-statistics and hence possibly erroneous attributions of significance of specific relationships in the model (Muthén and Kaplan, 1985). These theoretical predictions about the effects of nonnormality on parameter estimates, the chi-square statistic, and the standard errors of the parameter estimates are supported by Monte Carlo simulation studies, in which the values of the skewness and kurtosis can be well controlled at desired levels. Thus, the marginal differences between the samples having normal and nonnormal distributions can be attributed to nonnormality.
Muthén and Kaplan (1985) designed five combinations of multivariate skewness and kurtosis (see footnote 1 for details) to study the effects of nonnormality on chi-square statistics and parameter estimates with a sample N = 1000. The chi-square statistic and model rejection frequencies for the model in Case 4 (multivariate skewness = 15.421, multivariate kurtosis = 21.408⁴) were 84% and 67% higher, respectively, than those in Case 1 (multivariate skewness = 0.082, multivariate kurtosis = 0). The effects of nonnormality on the chi-square statistic and model rejection frequencies in Case 2 (multivariate skewness = 1.796, multivariate kurtosis = 0.624), Case 3 (multivariate skewness = 5.625, multivariate kurtosis = 6.648) and Case 5 (multivariate skewness = 0.131, multivariate kurtosis = 13.92) were negligible. In all cases, the differences between the parameter estimates and the actual parameters were no more than 4.2% and were negligible. They concluded (p.187) that “if most variables have univariate skewnesses and kurtoses in the range -1.0 to +1.0, not much distortion is to be expected” and that “this is largely independent of number of variables and number of categories”, and that “when most skewnesses and/or kurtoses are larger in absolute value than 2.0, and correlations are large (say 0.5 and higher), distortions of ML and GLS chi-squares and standard errors are very likely”. Hallow (1985) tested the impacts of nonnormality which was measured by univariate skewness (-1.25 < skewness < 2.0) and kurtosis (-1.0 < kurtosis < 8.0), and Mardia’s kurtosis (-4.9 < Mardia’s kurtosis < 49.1). The results showed that, compared with the parameter estimates of the base condition (multivariate

⁴ The original paper actually used multivariate relative kurtosis, defined as the sample multivariate kurtosis \(\beta_2^p\) divided by the expected value of the sample multivariate kurtosis \(\beta_2^p\), which was 0.989, 1.026, 1.277, 1.892 and 1.580 in Cases 1, 2, 3, 4 and 5, respectively. The number of variables \(p\) in the model was 4. We recalculated the relative kurtosis as Mardia’s multivariate kurtosis based on \(p = 4\) and \(N = 1000\) to be consistent with the multivariate kurtosis used in the present paper. The univariate skewness and kurtosis were the same for each variable in a case. The univariate skewness/kurtosis for the five cases were 0.000/0.000, -0.742/-0.334, -1.217/1.615, -2.028/2.898 and 0.000/2.785, respectively.
normal distribution), the parameter estimates were still unbiased. The chi-square statistics were not significantly inflated by nonnormality (all 12 nonnormal conditions). However, at least one standard error of the parameter estimate under each nonnormal condition was found to be negatively or positively biased.

Using samples drawn from a non-normal population with $\beta_{1,6}$ (multivariate skewness) = 0 and $\beta_{2,6}$ (multivariate kurtosis) = 63.9, and a normal population with $\beta_{1,6}$ = 0 and $\beta_{2,6}$ = 48, Henly (1993) studied the effects of sample size, distribution, and nonnormality on the chi-square statistic. She found that 1) for the cases whose sample size was smaller than 300, even if the sample was multivariate normally distributed, the parameter estimates and standard errors were subject to biases due to the small sample size; 2) when a sample was multivariate normally distributed and its size was bigger than 300, the parameter estimates and standard errors were unbiased; and 3) the sample sizes should be at least 600 to obtain unbiased parameter estimates for samples with a multivariate nonnormal distribution; and 4) for samples with a multivariate nonnormal distribution, regardless of the sample size, the model rejection frequencies were substantially higher than for the corresponding samples with a normal distribution, and the standard error estimates obtained from maximum likelihood estimation appeared not to be useful.

Curran et al. (1996) compared the percentage bias of chi-square statistics and percentages of rejecting models under moderate nonnormality (skewness = 2, kurtosis = 7) and extreme nonnormality (skewness = 3, kurtosis = 21), with those under a normal distribution (skewness = 0, kurtosis = 0), and found that the chi-square statistics were positively biased by nonnormality. Lei and Lomax (2005) tested the impacts of
nonnormality with six combinations of univariate skewness (0 < skewness < 1.74) and kurtosis (0 < kurtosis < 3.8). They found that, compared with a normal distribution, nonnormality did not have significant impacts on the means of standard errors for the parameter estimates and on the parameter estimates, but had a significant impact on the chi-square statistics. They also found that either skewness or kurtosis led to a significant change in the chi-square statistic.

In a broad sense, the findings on the consequences of violating the normality assumption in these studies are consistent. A larger sample helps obtain more accurate estimates for parameters, standard errors, and the chi-square statistic. The estimates of the chi-square statistic and standard errors are more sensitive to nonnormality than are the parameter estimates. What is not clear enough is how to set up cutoffs for slight, moderate, and extreme nonnormality, based on the measures of multivariate skewness or kurtosis, or both, because the assumption in ML estimation is multivariate normality, not univariate normality. Univariate skewness and kurtosis can only roughly measure multivariate normality. In addition, focusing on univariate nonnormality makes it extremely difficult to compare the marginal contribution of nonnormality when two samples have different univariate skewness and kurtosis for each variable. Muthén and Kaplan’s (1985) recommendations are actually based on univariate skewness and kurtosis and do not well explain the results obtained by Curran et al. (1996) and Lei and Lomax (2005). Furthermore, I think more evidence is needed to exclude the possible effects of the number of variables on the assessment of multivariate normality of a sample. This can be done only if the normality is measured by multivariate skewness and/or kurtosis.
4.3. CONCEPTUAL MODEL AND DATA

For the purpose of testing the generalizability of the conceptual model and the impacts of sample size and nonnormality on model fit indices, parameter estimates, and standard errors of the parameter estimates, I use year 2000 census tract (CT) and census block group (BG) data. The study area is Sacramento County, California. There are 279 census tracts and 792 census blocks in this county. Three cases are deleted from the census tract sample because these tracts are prisons and military bases. For the same reason, five census block groups are deleted from the BG sample.

Figure 4.1 illustrates the conceptual model (for detailed information, see Gao et al., 2006a, 2006b, i.e. Chapters Two and Three). The variables are computed in the same way at both geographic levels. Median rent (median asking rent per month) and income per capita (per year) are imported directly from the data CD I purchased from Geolytics, Inc. Percentage of rental housing units is defined as the percentage of renter-occupied housing units out of the total occupied housing units. Education attainment is calculated by dividing total persons who have a bachelor’s or higher degree by the total population in a census tract. Workers per capita is computed by dividing the total number of employed persons by total population. Household size is calculated by dividing the total population by the total number of households. Autos per capita is computed by dividing the total number of vehicles by the total population. Percentage of single-headed households with children (Percentage of single-headed households in Figure 4.1) is defined as the percentage of single-headed households with children out of the total number of households. Job accessibility is calculated as:
Figure 4.1 Conceptual causal relationships among job accessibility, employment, income and auto ownership

\[
CT\text{JobAccessibility}_m = \sum_{i=1}^{n_m} Job_i + \left( \sum_{i=1}^{n_m} \sum_{j=1}^{N-n_m} Job_i \cdot \{(1/TravelTime_{ij})\} / n_m \right)
\]

(4.7)

where \(CT\text{JobAccessibility}_m\) is the job accessibility in census tract \(m\); \(n_m\) is the number of traffic analysis zones (TAZs) contained in census tract \(m\) (\(TAZ_i \subset CT_m\)); \(Job_i\) is the number of jobs in TAZ \(i\), and \(TravelTime_{ij}\) is the travel time from TAZ \(i\) to \(j\) in the three-hour AM peak period on the highway network (for a detailed explanation, see Gao et al., 2006b).
Table 4.1 Descriptive statistics of the variables at the CT level (N = 276)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of rental housing units</td>
<td>1.59</td>
<td>100.00</td>
<td>41.12</td>
<td>22.27</td>
</tr>
<tr>
<td>Median rent</td>
<td>0.29</td>
<td>1.46</td>
<td>0.75</td>
<td>0.18</td>
</tr>
<tr>
<td>Percentage of single-headed households with children</td>
<td>1.00</td>
<td>28.00</td>
<td>11.26</td>
<td>5.30</td>
</tr>
<tr>
<td>Education attainment</td>
<td>0.01</td>
<td>0.54</td>
<td>0.17</td>
<td>0.11</td>
</tr>
<tr>
<td>Household size</td>
<td>1.25</td>
<td>5.98</td>
<td>2.72</td>
<td>0.60</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>1.43</td>
<td>7.55</td>
<td>4.18</td>
<td>0.99</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>0.03</td>
<td>0.74</td>
<td>0.46</td>
<td>0.09</td>
</tr>
<tr>
<td>Income per capita</td>
<td>0.68</td>
<td>5.09</td>
<td>2.18</td>
<td>0.85</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>0.04</td>
<td>0.85</td>
<td>0.63</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Table 4.2 Descriptive statistics of the variables at the BG level (N = 787)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of rental housing units</td>
<td>0.00</td>
<td>100.00</td>
<td>40.58</td>
<td>25.50</td>
</tr>
<tr>
<td>Median rent</td>
<td>0.00</td>
<td>1.88</td>
<td>0.75</td>
<td>0.23</td>
</tr>
<tr>
<td>Percentage of single-headed households with children</td>
<td>0.00</td>
<td>52.77</td>
<td>16.62</td>
<td>9.18</td>
</tr>
<tr>
<td>Education attainment</td>
<td>0.00</td>
<td>0.65</td>
<td>0.17</td>
<td>0.13</td>
</tr>
<tr>
<td>Household size</td>
<td>1.18</td>
<td>9.40</td>
<td>2.71</td>
<td>0.70</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>1.47</td>
<td>7.28</td>
<td>4.19</td>
<td>0.88</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>0.00</td>
<td>0.82</td>
<td>0.45</td>
<td>0.11</td>
</tr>
<tr>
<td>Income per capita</td>
<td>0.45</td>
<td>7.20</td>
<td>2.19</td>
<td>1.01</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>0.02</td>
<td>1.13</td>
<td>0.64</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Tables 4.1 and 4.2 show the descriptive statistics of the variables (to place the magnitudes of *job accessibility*, *income per capita* and *median rent* roughly in the same range, the raw values of these three variables are divided by 10000, 10000 and 1000, respectively). From Tables 4.1 and 4.2, we can see that each variable has nearly the same mean at the CT and BG level. Logically enough, the variation of all the variables except for *job accessibility* is larger at the BG level. Because 718 of 787 BGs consist of only one TAZ (compared to only 109 out of 276 CTs), *job accessibility* at the BG level is less aggregated than at the CT level and hence more desirable.

### 4.4. ASSESSMENT OF MULTIVARIATE NORMALITY

In AMOS 5, univariate and multivariate normalities are evaluated in one step. Tables 4.3 and 4.4 show the results of the assessment of the univariate normality for each variable and the multivariate normality for the two samples at the CT and BG levels. At the CT level, the univariate skewness varies between -1.04 and 0.97, and the univariate kurtosis varies between -0.12 and 4.16. According to the cutoffs suggested by Muthén and Kaplan (1985) and Curran et al. (1996), this sample is considered to have a slightly nonnormal distribution, which will not significantly distort the ML estimates. However, the multivariate kurtosis (multivariate kurtosis = 92.94, critical ratio = 54.87) is much larger than those in the literature (Hallow, 1985; Henly, 1993; Muthén and Kaplan, 1985). This indicates a severe multivariate nonnormality of the sample, which will severely distort the ML estimates. At the BG level, the univariate skewness and kurtosis have larger ranges. *Median rent* and *income per capita* are slightly nonnormal while *household size* is severely nonnormal in terms of skewness. The kurtosis of *household
Table 4.3 Assessment of normality at the CT level (N = 276)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Skewness</th>
<th>Critical ratio of skewness</th>
<th>Kurtosis</th>
<th>Critical ratio of kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of rental housing units</td>
<td>0.60</td>
<td>4.04</td>
<td>-0.12</td>
<td>-0.40</td>
</tr>
<tr>
<td>Median rent</td>
<td>0.97</td>
<td>6.60</td>
<td>1.59</td>
<td>5.40</td>
</tr>
<tr>
<td>Percentage of single-headed households with children</td>
<td>0.51</td>
<td>3.47</td>
<td>-0.07</td>
<td>-0.24</td>
</tr>
<tr>
<td>Education attainment</td>
<td>0.89</td>
<td>6.01</td>
<td>0.13</td>
<td>0.44</td>
</tr>
<tr>
<td>Household size</td>
<td>0.96</td>
<td>6.48</td>
<td>4.16</td>
<td>14.10</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>0.38</td>
<td>2.55</td>
<td>0.65</td>
<td>2.21</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>-0.62</td>
<td>-4.22</td>
<td>1.70</td>
<td>5.77</td>
</tr>
<tr>
<td>Income per capita</td>
<td>0.80</td>
<td>5.41</td>
<td>0.50</td>
<td>1.71</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>-1.04</td>
<td>-7.02</td>
<td>2.22</td>
<td>7.51</td>
</tr>
<tr>
<td>Multivariate</td>
<td></td>
<td></td>
<td>92.94</td>
<td>54.87</td>
</tr>
</tbody>
</table>

Table 4.4 Assessment of normality at the BG level (N = 787)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Skewness</th>
<th>Critical ratio of skewness</th>
<th>Kurtosis</th>
<th>Critical ratio of kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of rental housing units</td>
<td>0.54</td>
<td>6.13</td>
<td>-0.62</td>
<td>-3.57</td>
</tr>
<tr>
<td>Median rent</td>
<td>1.10</td>
<td>12.55</td>
<td>3.47</td>
<td>19.87</td>
</tr>
<tr>
<td>Percentage of single-headed households with children</td>
<td>0.82</td>
<td>9.44</td>
<td>1.49</td>
<td>8.51</td>
</tr>
<tr>
<td>Education attainment</td>
<td>0.93</td>
<td>10.68</td>
<td>0.31</td>
<td>1.77</td>
</tr>
<tr>
<td>Household size</td>
<td>2.23</td>
<td>25.56</td>
<td>15.93</td>
<td>91.23</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>-0.11</td>
<td>-1.21</td>
<td>0.44</td>
<td>2.54</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>-0.24</td>
<td>-2.79</td>
<td>0.57</td>
<td>3.28</td>
</tr>
<tr>
<td>Income per capita</td>
<td>1.25</td>
<td>14.29</td>
<td>2.56</td>
<td>14.65</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>-0.62</td>
<td>-7.08</td>
<td>0.69</td>
<td>3.92</td>
</tr>
<tr>
<td>Multivariate</td>
<td></td>
<td></td>
<td>101.61</td>
<td>101.29</td>
</tr>
</tbody>
</table>
### Table 4.5 Assessment of normality at the CT level (N = 218)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Skewness</th>
<th>Critical ratio of skewness</th>
<th>Kurtosis</th>
<th>Critical ratio of kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of rental housing units</td>
<td>3.93</td>
<td>86.52</td>
<td>0.24</td>
<td>1.46</td>
<td>-0.64</td>
<td>-1.92</td>
</tr>
<tr>
<td>Median rent</td>
<td>0.49</td>
<td>1.21</td>
<td>0.71</td>
<td>4.27</td>
<td>0.14</td>
<td>0.43</td>
</tr>
<tr>
<td>Percentage of single-headed households with children</td>
<td>0.04</td>
<td>0.26</td>
<td>0.33</td>
<td>1.97</td>
<td>-0.49</td>
<td>-1.46</td>
</tr>
<tr>
<td>Education attainment</td>
<td>0.01</td>
<td>0.44</td>
<td>0.96</td>
<td>5.81</td>
<td>0.31</td>
<td>0.94</td>
</tr>
<tr>
<td>Household size</td>
<td>1.81</td>
<td>4.09</td>
<td>0.29</td>
<td>1.76</td>
<td>-0.35</td>
<td>-1.05</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>2.05</td>
<td>6.01</td>
<td>-0.07</td>
<td>-0.42</td>
<td>0.10</td>
<td>0.31</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>0.24</td>
<td>0.66</td>
<td>-0.45</td>
<td>-2.74</td>
<td>-0.24</td>
<td>-0.72</td>
</tr>
<tr>
<td>Income per capita</td>
<td>0.68</td>
<td>3.98</td>
<td>0.59</td>
<td>3.53</td>
<td>-0.21</td>
<td>-0.63</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>0.33</td>
<td>0.84</td>
<td>-0.40</td>
<td>-2.38</td>
<td>-0.44</td>
<td>-1.34</td>
</tr>
<tr>
<td>Multivariate</td>
<td></td>
<td></td>
<td></td>
<td>3.56</td>
<td>1.87</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4.6 Assessment of normality at the BG level (N = 650)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Skewness</th>
<th>Critical ratio of skewness</th>
<th>Kurtosis</th>
<th>Critical ratio of kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of rental housing units</td>
<td>0</td>
<td>100.00</td>
<td>0.47</td>
<td>4.89</td>
<td>-0.62</td>
<td>-3.24</td>
</tr>
<tr>
<td>Median rent</td>
<td>0.26</td>
<td>1.44</td>
<td>0.84</td>
<td>8.70</td>
<td>1.00</td>
<td>5.21</td>
</tr>
<tr>
<td>Percentage of single-headed households with children</td>
<td>0</td>
<td>33.66</td>
<td>0.50</td>
<td>5.18</td>
<td>-0.07</td>
<td>-0.38</td>
</tr>
<tr>
<td>Education attainment</td>
<td>0</td>
<td>0.50</td>
<td>0.83</td>
<td>8.64</td>
<td>-0.05</td>
<td>-0.25</td>
</tr>
<tr>
<td>Household size</td>
<td>1.43</td>
<td>4.28</td>
<td>0.20</td>
<td>2.04</td>
<td>-0.42</td>
<td>-2.16</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>1.78</td>
<td>6.05</td>
<td>-0.26</td>
<td>-2.74</td>
<td>0.16</td>
<td>0.81</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>0.17</td>
<td>0.74</td>
<td>-0.22</td>
<td>-2.25</td>
<td>-0.14</td>
<td>-0.74</td>
</tr>
<tr>
<td>Income per capita</td>
<td>0.56</td>
<td>4.31</td>
<td>0.47</td>
<td>4.92</td>
<td>-0.35</td>
<td>-1.83</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>0.20</td>
<td>0.99</td>
<td>-0.45</td>
<td>-4.68</td>
<td>-0.21</td>
<td>-1.07</td>
</tr>
<tr>
<td>Multivariate</td>
<td></td>
<td></td>
<td></td>
<td>2.17</td>
<td>1.97</td>
<td></td>
</tr>
</tbody>
</table>
size indicates a severe nonnormal distribution of this variable. The large multivariate kurtosis implies that the sample at the BG level has a severely multivariate nonnormal distribution.

To reach a multivariate normal distribution, I deleted five observations at a time, based on the Mahalanobis distance. After the first five observations were deleted, the multivariate kurtosis dropped sharply, from 92.94 to 24.82 at the CT level and from 101.61 to 31.99 at the BG level. The critical ratios fall below 1.96 after 58 observations are deleted at the CT level and 137 observations at the BG level. Thus, 21% and 17% of observations had to be deleted, respectively, in order to achieve the desired critical ratio. The univariate skewness and kurtosis, and multivariate kurtosis of the two reduced samples having a multivariate normal distribution, are shown in Tables 4.5 and 4.6. From the tables, it can be seen that, when the critical ratio of multivariate kurtosis is smaller than 1.96, the absolute values of univariate skewness and kurtosis for all variables are not necessarily close to 0, but are smaller than 1. In this respect, it appears that measuring only multivariate kurtosis may be good enough for the purpose of assessing multivariate normality.

It is noted that, among the deleted observations, only 3 in the CT sample and 6 in the BG sample can be considered to be outliers in terms of their Mahalanobis distances being much larger than those of the other observations. Deleting these outliers lowers not only univariate nonnormality but also multivariate nonnormality (by 71.6% at the CT level and by 86.4% at the BG level). Therefore, it is the outliers that lead to severe univariate and multivariate nonnormality. Except for these outliers, the marginal contributions of
the other deleted observations to the departure of the sample from normality are minor, and the cumulative contribution of these observations is relatively small.

4.5. TRADEOFF OF DELETING OBSERVATIONS: NORMALITY VS. POWER

The samples at both the CT and the BG levels are not random samples. The size is generally determined by the area of an empirical study (usually a metropolitan area or a county). The CTs or BGs within a study area will be taken as observations. Therefore, the attributes of the observations are fundamentally affected by the land use pattern of the study area. The spatial heterogeneity of land use determines the uniqueness of the observations. One observation is unlikely to be fully represented by another observation in terms of its attributes. Deleting an observation leads to a loss of the unique information in that observation. The influence of deleting an observation is not identical for different attributes. Therefore, the extent to which the loss can be compensated depends on the similarities between the observations. The cumulative effects of deleted observations on the model power and the interpretation of the structural coefficients may be significant and cannot be neglected. In other words, when I delete some observations to obtain higher multivariate normality, I take the loss of information in the deleted observations as a tradeoff.

The descriptive statistics of the two multivariate normally distributed samples are Tables 4.11 and 4.12 in Appendix. Comparing the means of the variables in Table 4.1 and Table 4.11, I find that the variation in the means of the variables is no more than ±9% while the standard deviations decrease by 12% to 23%. A similar comparison of Table
4.2 and Table 4.12 shows that, after 137 observations are deleted and multivariate normality is reached, the changes in the means of the variables are no more than ±6%. The changes in the standard deviations are between 9% and 23%. The average percent changes in the means and standard deviations of all the variables at the BG level are slightly smaller than those at the CT level. This suggests that the negative marginal and cumulative effects of deleting an observation are smaller at the BG level than at the CT level.

I further checked the spatial distribution of the deleted observations by mapping the census tracts and block groups in a geographic information system (GIS). Among the first 10 deleted observations at the CT level, seven were in, or next to, downtown Sacramento and had high job access to the traditional job centers, and two had high job access to newer job centers. Checking all 58 deleted observations at the CT level, I found that the deleted tracts covered almost all of downtown Sacramento and the other three important employment centers (Rancho Cordova, California State University at Sacramento, and Sacramento International Airport). From the standpoint of policy analysis, these observations are so important that none of the observations should be deleted. The conclusions which are based on a sample that is missing such important observations are meaningless and misleading.

The loss of information at the BG level is mitigated to some extent due to the smaller geographic size of the BGs. The 10 deleted tracts discussed above contain 27 BGs. Each of the tracts contains more than one BG. In the BG sample, only 4 out of those 27 BGs are identified as outliers in the first 10 deleted observations. In other words, a large portion of the information lost in the CT sample is kept in the BG sample.
Therefore, a BG sample is more desirable than a CT sample if only a few observations have to be deleted to lower the multivariate nonnormality.

4.6. THE PARAMETER ESTIMATES AND STANDARD ERRORS

By ML estimation, the hypothesized model is estimated at the CT and BG levels. From Table 4.7, we can see that all coefficients are significant at the 0.05 level when the CT sample size is 276 while three coefficients (i.e., the direct effects of percentage of single-headed households with children and job accessibility on workers per capita, and workers per capita on autos per capita) are insignificant at the 0.1 level when the CT sample size is reduced to 218. Besides these three parameter estimates, five parameter estimates are more than 20% larger in the reduced CT sample than in the original CT sample. Accordingly, the standard errors of these parameter estimates change with the change of sample distribution and sample size. Eight of 13 standard errors become larger while two become smaller, and only three are unchanged.

At the BG level (see Table 4.8), the direct effect of median rent on job accessibility is not significant at the 0.1 level in the sample of 787 observations while it is significant at the 0.05 level when the sample size is reduced to 650. The direct effects of job accessibility on workers per capita become insignificant ($p = 0.604$) when a multivariate normal distribution is reached ($N = 650$). Five of the parameter estimates have a variation in magnitude of more than 20%. After the sample size is reduced, four of 13 parameter estimates have larger standard errors, and two of 13 have smaller standard errors.

These results have several implications. First, the conceptual model can be estimated at both the CT and BG levels. In this respect, I may claim that the causal structure still
Table 4.7 Unstandardized parameter estimates and their standard errors at the CT level

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Job accessibility</td>
<td>Percentage of rental housing units</td>
<td>0.031</td>
<td>0.002</td>
<td>** ***</td>
<td>0.222</td>
<td>0.002</td>
<td>** ***</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>Median rent</td>
<td>-0.089</td>
<td>0.028</td>
<td>0.001</td>
<td>-0.176</td>
<td>0.030</td>
<td>** ***</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>Income per capita</td>
<td>0.422</td>
<td>0.060</td>
<td>** ***</td>
<td>0.356</td>
<td>0.068</td>
<td>** ***</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>Percentage of single-headed households with children</td>
<td>-0.088</td>
<td>0.037</td>
<td>0.019</td>
<td>0.058</td>
<td>0.056</td>
<td>0.296</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>Education attainment</td>
<td>0.172</td>
<td>0.048</td>
<td>** ***</td>
<td>0.181</td>
<td>0.048</td>
<td>** ***</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>Job accessibility</td>
<td>0.009</td>
<td>0.003</td>
<td>0.006</td>
<td>0.003</td>
<td>0.004</td>
<td>0.474</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>Autos per capita</td>
<td>0.443</td>
<td>0.050</td>
<td>** ***</td>
<td>0.536</td>
<td>0.050</td>
<td>** ***</td>
</tr>
<tr>
<td>Income per capita</td>
<td>Education attainment</td>
<td>3.492</td>
<td>0.562</td>
<td>** ***</td>
<td>3.427</td>
<td>0.441</td>
<td>** ***</td>
</tr>
<tr>
<td>Income per capita</td>
<td>Workers per capita</td>
<td>6.347</td>
<td>0.934</td>
<td>** ***</td>
<td>5.464</td>
<td>0.644</td>
<td>** ***</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>Household size</td>
<td>-0.084</td>
<td>0.009</td>
<td>** ***</td>
<td>-0.149</td>
<td>0.014</td>
<td>** ***</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>Job accessibility</td>
<td>-0.079</td>
<td>0.007</td>
<td>** ***</td>
<td>-0.109</td>
<td>0.012</td>
<td>** ***</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>Workers per capita</td>
<td>0.418</td>
<td>0.055</td>
<td>** ***</td>
<td>0.079</td>
<td>0.076</td>
<td>0.296</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>Income per capita</td>
<td>0.050</td>
<td>0.006</td>
<td>** ***</td>
<td>0.054</td>
<td>0.008</td>
<td>** ***</td>
</tr>
</tbody>
</table>

***: p value < 0.000
Table 4.8 Unstandardized parameter estimates and their standard errors at the BG level

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Explanatory variable</th>
<th>Estimate (N = 787)</th>
<th>Standard error (N = 787)</th>
<th>P value (N = 787)</th>
<th>Estimate (N = 650)</th>
<th>Standard error (N = 650)</th>
<th>P value (N = 650)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job accessibility</td>
<td>Percentage of rental housing units</td>
<td>0.022</td>
<td>0.001</td>
<td>***</td>
<td>0.021</td>
<td>0.001</td>
<td>***</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>Median rent</td>
<td>-0.140</td>
<td>0.087</td>
<td>0.107</td>
<td>-0.198</td>
<td>0.088</td>
<td>0.025</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>Income per capita</td>
<td>0.279</td>
<td>0.030</td>
<td>***</td>
<td>0.288</td>
<td>0.039</td>
<td>***</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>Percentage of single-headed households with children</td>
<td>-0.001</td>
<td>0.000</td>
<td>***</td>
<td>-0.001</td>
<td>0.000</td>
<td>***</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>Education attainment</td>
<td>0.216</td>
<td>0.035</td>
<td>***</td>
<td>0.273</td>
<td>0.033</td>
<td>***</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>Job accessibility</td>
<td>0.013</td>
<td>0.004</td>
<td>0.003</td>
<td>0.002</td>
<td>0.004</td>
<td>0.604</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>Autos per capita</td>
<td>0.333</td>
<td>0.036</td>
<td>***</td>
<td>0.334</td>
<td>0.031</td>
<td>***</td>
</tr>
<tr>
<td>Income per capita</td>
<td>Education attainment</td>
<td>3.094</td>
<td>0.478</td>
<td>***</td>
<td>1.691</td>
<td>0.451</td>
<td>***</td>
</tr>
<tr>
<td>Income per capita</td>
<td>Workers per capita</td>
<td>7.598</td>
<td>0.855</td>
<td>***</td>
<td>7.873</td>
<td>0.720</td>
<td>***</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>Household size</td>
<td>-0.074</td>
<td>0.006</td>
<td>***</td>
<td>-0.120</td>
<td>0.008</td>
<td>***</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>Job accessibility</td>
<td>-0.117</td>
<td>0.007</td>
<td>***</td>
<td>-0.133</td>
<td>0.010</td>
<td>***</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>Workers per capita</td>
<td>0.454</td>
<td>0.035</td>
<td>***</td>
<td>0.395</td>
<td>0.035</td>
<td>***</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>Income per capita</td>
<td>0.056</td>
<td>0.005</td>
<td>***</td>
<td>0.052</td>
<td>0.006</td>
<td>***</td>
</tr>
</tbody>
</table>

***: p value = 0.0
holds when the spatial resolution becomes finer and the variables are less aggregated. The failure of the significance test of the direct effects of median rent on job accessibility can be attributed to the higher nonnormality caused by the outliers at the BG level because the coefficient becomes significant after the first five outliers are deleted from the sample. As shown in Tables 4.1, 4.2, 4.3 and 4.4, the variables at the BG level have larger variations and more outliers (six) and depart from multivariate normality further than at the CT level. After the first five outliers are deleted (N = 782) and the multivariate kurtosis becomes comparable with that at the CT level (N = 276), the direct effect of median rent on job accessibility becomes moderately significant (p = 0.085). An analysis of the first five outliers at the BG level shows that they are aggregated with other BGs at the CT level and thus have less contribution to the departure from the multivariate normality than at the CT level. Second, in terms of explaining causal relationships among the variables, the model leads to the same conclusions both at the CT and BG levels after the extreme outliers are excluded. Third, a sample of higher spatial resolution and less aggregation is more desirable. Due to its smaller physical size, a BG has fewer households than a CT. The households within a BG are more homogeneous than those within a CT. The characteristics of a BG can be better represented by a variable in per capita form than those of a CT can. The more important point is that the changes of parameter estimates and standard errors of those parameter estimates are smaller at the BG level than at the CT level, due to the larger sample size. Fourth, regardless of whether taken as an explanatory or a predictive model, the model leads to different conclusions when multivariate normality is reached by deleting the observations with a larger Mahalanobis distance at both levels. The correlations become weaker after too many
observations are deleted. Therefore, a sample having a multivariate normal distribution obtained by deleting some observations does not necessarily produce the expected results. Fifth, the standard errors of the parameter estimates do not necessarily become smaller when the multivariate kurtosis becomes smaller by deleting the outliers, presumably because the smaller kurtosis is counteracted by the smaller sample size, which tends to increase standard errors, all else equal.

4.7. DISCREPANCY FUNCTION AND CHI-SQUARE STATISTIC

In ML estimation, the chi-square statistic is the product of (N-1) and the discrepancy function (F_{ml}). Thus, it is very sensitive to sample size and cannot well reflect the effects of nonnormality on the discrepancy function when the change in nonnormality is accompanied by a change in sample size, as in this study. Monte Carlo simulation studies have demonstrated that, when sample size is controlled, multivariate nonnormality will lead to an inflation of the chi-square statistic (Curran et al., 1996; Lei and Lomax, 2006; Muthén and Kaplan, 1985). Figures 4.2 and 4.3 are two plots showing the discrepancy function for each sample. In Figure 4.2 (at the CT level), F_{ml} is the smallest at N = 276 and then becomes larger with the decrease in N. This is exactly the opposite from what is expected. At the BG level, the F_{ml} is largest at N = 787 and fluctuates slightly around 0.21 when the sample size becomes smaller. These plots suggest that deleting outliers improves multivariate normality but may or may not lead to a smaller discrepancy function and chi-square statistic (see Figures 4.4 and 4.5).
Figure 4.2 Discrepancy function $F_{ml}$ under different multivariate kurtoses at the CT level

Figure 4.3 Discrepancy function $F_{ml}$ under different multivariate kurtoses at the BG level
Figure 4.4 Chi-square statistic $(N-1)F_{ml}$ under different multivariate kurtoses at the CT level

Figure 4.5 Chi-square statistic $(N-1)F_{ml}$ under different multivariate kurtoses at the BG level
Comparing Figures 4.2 and 4.3, we can see that the maximum and minimum $F_{ml}$ at the BG level are substantially smaller than their counterparts at the CT level, and the range at the BG level is smaller as well. This suggests that when the sample covariance structures and multivariate normality are similar, a larger sample size is more helpful in minimizing the discrepancy function.

4.8. THE MODEL FIT INDICES

Following the principles suggested by Bollen and Long (1993), Hoyle and Panter (1995), and Shah and Goldstein (2006), I report the model fit indices from several different index families. From Tables 4.9 and 4.10, we can see that the changes in the sample size and multivariate normality do not have substantial effects on GFI, IFI and CFI at the CT and BG levels and do have substantial effects on the chi-square statistic, AIC and ECVI. At the CT level, as discussed in the section above, deleting 21% of the observations leads to a substantial increase of $F_{ml}$ (64%) and, accordingly, an increase of the chi-square statistic (29%) and failures of hypothesis tests on three parameter estimates. Furthermore, AIC and ECVI increases 15% and 46%, respectively. Therefore, deleting the observations at the CT level largely changes the covariance structure which is extremely undesirable.

At the BG level, deleting 17% of the observations leads to a decrease of 24% in the chi-square statistic and a decrease of 17% in AIC. Comparing the model fit indices across the CT and BG levels, we can see that the model at the BG level has a larger chi-square statistic and Akaike Information Criterion (AIC) due to a larger sample size, and a smaller expected cross-validation index (ECVI) than at the CT level. Since ECVI
Table 4.9 Assessment of model fit at the CT level

<table>
<thead>
<tr>
<th>Model fit index</th>
<th>Full sample (N = 276)</th>
<th>Reduced sample (N = 218)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Independence model</td>
<td>Hypothesized model</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>2218.75</td>
<td>76.68</td>
</tr>
<tr>
<td>Degrees of freedom (d.f.)</td>
<td>36.00</td>
<td>9.00</td>
</tr>
<tr>
<td>Goodness of fit index (GFI)</td>
<td>0.33</td>
<td>0.95</td>
</tr>
<tr>
<td>Normed fit index (NFI)</td>
<td>0.00</td>
<td>0.97</td>
</tr>
<tr>
<td>Incremental fit index (IFI)</td>
<td>0.00</td>
<td>0.97</td>
</tr>
<tr>
<td>Comparative fit index (CFI)</td>
<td>0.00</td>
<td>0.97</td>
</tr>
<tr>
<td>Akaike information criterion (AIC)</td>
<td>2236.75</td>
<td>148.68</td>
</tr>
<tr>
<td>Expected cross-validation index (ECVI)</td>
<td>8.13</td>
<td>0.54</td>
</tr>
</tbody>
</table>
Table 4.10 Assessment of model fit at the BG level

<table>
<thead>
<tr>
<th>Model fit index</th>
<th>Full sample (N = 787)</th>
<th>Reduced sample (N = 650)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Independence model</td>
<td>Hypothesized model</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>4103.26</td>
<td>178.59</td>
</tr>
<tr>
<td>Degrees of freedom (d.f.)</td>
<td>36.00</td>
<td>9.00</td>
</tr>
<tr>
<td>Goodness of fit index (GFI)</td>
<td>0.42</td>
<td>0.96</td>
</tr>
<tr>
<td>Normed fit index (NFI)</td>
<td>0.00</td>
<td>0.96</td>
</tr>
<tr>
<td>Incremental fit index (IFI)</td>
<td>0.00</td>
<td>0.96</td>
</tr>
<tr>
<td>Comparative fit index (CFI)</td>
<td>0.00</td>
<td>0.96</td>
</tr>
<tr>
<td>Akaike information criterion (AIC)</td>
<td>4121.26</td>
<td>250.59</td>
</tr>
<tr>
<td>Expected cross-validation index (ECVI)</td>
<td>5.24</td>
<td>0.32</td>
</tr>
</tbody>
</table>
indicates the generalizability of a hypothesized model (Shah and Goldstein, 2006), with smaller being better, this result suggests that it is better to estimate the model at the BG level than at the CT level. The goodness of fit index (GFI), normed fit index (NFI), incremental fit index (IFI) and comparative fit index (CFI) are almost the same at the CT and BG levels and seem not to be sensitive to the change of sample size and the degree of multivariate normality.

4.9. DISCUSSION AND CONCLUSIONS

In the literature, it is demonstrated that multivariate nonnormality will deflate standard errors of the parameter estimates and inflate the chi-square statistic (Muthén and Kaplan, 1985). The core of the question is not whether nonnormality will distort the estimates if maximum likelihood estimation is used, but at what level of nonnormality the ML estimates are still robust. Although the sample size and univariate skewness and kurtosis can be well controlled in Monte Carlo simulation studies and the marginal contributions of univariate nonnormality caused by skewness or kurtosis, or both, can be easily captured by calculating rejection frequencies using chi-square statistics and parameter estimates, the answers to the question are not satisfactory. The recommended cutoff (-1< univariate skewness or kurtosis < +1) for trustworthy estimates is not well supported by simulation studies (Curran et al., 1997; Hallow, 1985).

Our results show that, at this cutoff for univariate skewness and/or kurtosis, the multivariate kurtosis can be large and multivariate nonnormality can be extreme (critical ratio >> 1.96). Some outliers (three at the CT level) are included in the sample (N = 276). It is possible that the estimates of the standard errors and chi-square statistic are severely
biased. When a sample with a multivariate normal distribution (critical ratio < 1.96) is obtained by deleting some observations, the univariate skewnesses and kurtoses are within the range between -1 and +1. However, the reduced sample with a multivariate normal distribution does not include some observations containing important information about the covariance among the variables. The covariance structure of the reduced sample seems quite different from that of the original sample, although the sample means are similar. Therefore, I suggest that 1.96 should not be used as a steadfast threshold of the critical ratio for practical conformance to multivariate normality. In our CT sample, after the first five observations (i.e. all the true outliers) are deleted, the multivariate kurtosis drops to 24.8 and the critical ratio is 14.5. In the BG sample, after the first ten observations (all the true outliers) are deleted, the multivariate kurtosis drops to 25.4 and the critical ratio is 25.2. The parameter estimates and standard errors are very similar across the two samples. In Muthén and Kaplan’s (1985) study, even when the multivariate kurtosis was as high as 21.408, the biases for the estimates of parameters and standard errors of the parameter estimates were no more than 5%. This suggests that the biases for the parameter estimates and the standard errors of the parameter estimates in our results may be controlled at an acceptable level when the multivariate kurtosis is around 25. At this point, the loss of information due to deleting observations seems to be minimized while the results of the significance tests on the parameter estimates are what are desired from a standpoint of policy analysis.

It should be emphasized that deletion of the observations in this study is determined only by Mahalanobis distance and is stopped when the critical ratio of the Mardia’s coefficient of multivariate kurtosis is smaller than 1.96. The deleted observations can be
classified into two groups. The mean of the Mahalanobis distances of the observations in the first group is much larger than the mean of the distances of the observations in the reduced sample. The difference between each distance in the first group is substantial. Therefore, these observations may be considered as outliers. There are only a few observations (for example, 3 observations in the CT sample and 6 in the BG sample) in this group.

The outliers greatly increase the multivariate kurtosis and some univariate skewnesses and/or kurtoses, and distort the estimates for parameters, standard errors, and the chi-square statistic (Yuan et al., 2001). From the perspective of lowering the distortion of the estimates, these outliers should be either transformed or deleted. However, these outliers, at least in this study, are of interest in policy analysis. They indicate the consequences of some land use and transportation policies in practice. For example, mixed land uses largely increase job accessibility. Subsidized housing projects lead to concentration of low-income households and single-headed households with children. These outliers are, of course, a part of the data. Understanding the correlations among the attributes of the outliers is essential to policy analysis. Deleting the outliers means a loss of unique information. Therefore, even if these outliers may lead to inefficient estimates, they are desirable for policy analysis. The roles of the outliers should be reviewed carefully before they are deleted.

The mean of the Mahalanobis distances of the observations in the second group is substantially smaller than that in the first group, while still larger than that in the reduced sample. The differences among the distances in the second group are small. The larger distances simply mean that the observations are a little farther from the sample mean than
those in the reduced sample. It appears to be improper to consider these observations as outliers. Our suggestion is to keep these observations in the sample even if the critical ratio of the multivariate kurtosis is somewhat larger than desired as a result.

The results suggest, as well, that the conceptual model can be applied to a BG sample and performs better in terms of discrepancy function, chi-square, and standard errors, when some observations have to be deleted from the sample to meet the requirement of multivariate normality. A BG sample has more observations and contains richer spatial information than a CT sample. Therefore, regardless of whether used as explanatory or predictive, all else equal, a model at the BG level is more desirable than one at the CT level.
References:


Gao, Shengyi, Robert A. Johnston, and Patricia Mokhtarian. 2006a. A conceptual approach to testing the spatial mismatch hypothesis. Unpublished manuscript.


APPENDIX

Table 4.11 Descriptive statistics of the variables at the CT level (reduced sample, N = 218)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of rental housing units</td>
<td>3.93</td>
<td>86.52</td>
<td>38.9538</td>
<td>17.62</td>
</tr>
<tr>
<td>Median rent</td>
<td>0.50</td>
<td>1.21</td>
<td>0.75</td>
<td>0.14</td>
</tr>
<tr>
<td>Percentage of single-headed households with children</td>
<td>4.00</td>
<td>26.00</td>
<td>12.11</td>
<td>4.63</td>
</tr>
<tr>
<td>Education attainment</td>
<td>0.01</td>
<td>0.44</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>Household size</td>
<td>1.81</td>
<td>4.09</td>
<td>2.79</td>
<td>.46</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>2.05</td>
<td>6.01</td>
<td>4.05</td>
<td>0.76</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>0.24</td>
<td>0.66</td>
<td>0.45</td>
<td>0.08</td>
</tr>
<tr>
<td>Income per capita</td>
<td>0.68</td>
<td>3.98</td>
<td>2.05</td>
<td>0.72</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>0.33</td>
<td>0.84</td>
<td>0.63</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Table 4.12 Descriptive statistics of the variables at the BG level (reduced sample, N = 650)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of rental housing units</td>
<td>0</td>
<td>100.00</td>
<td>39.56</td>
<td>23.09</td>
</tr>
<tr>
<td>Median rent</td>
<td>0.26</td>
<td>1.44</td>
<td>0.74</td>
<td>0.19</td>
</tr>
<tr>
<td>Percentage of single-headed households with children</td>
<td>0</td>
<td>50.93</td>
<td>17.62</td>
<td>8.32</td>
</tr>
<tr>
<td>Education attainment</td>
<td>0.06</td>
<td>0.50</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td>Household size</td>
<td>1.43</td>
<td>4.28</td>
<td>2.73</td>
<td>0.54</td>
</tr>
<tr>
<td>Job accessibility</td>
<td>1.78</td>
<td>6.05</td>
<td>4.14</td>
<td>0.76</td>
</tr>
<tr>
<td>Workers per capita</td>
<td>0.17</td>
<td>0.74</td>
<td>0.45</td>
<td>0.10</td>
</tr>
<tr>
<td>Income per capita</td>
<td>0.56</td>
<td>4.31</td>
<td>2.07</td>
<td>0.79</td>
</tr>
<tr>
<td>Autos per capita</td>
<td>0.20</td>
<td>0.99</td>
<td>0.63</td>
<td>0.13</td>
</tr>
</tbody>
</table>
CHAPTER FIVE

CONCLUSIONS

In this dissertation, chapters Two, Three and Four focused on three inherently cohesive issues and collectively addressed the methodological problems in testing the causal relationships among job accessibility, employment, income, and auto ownership.

Chapter Two reviewed the literature on the spatial mismatch hypothesis and proposed two conceptual structural equation models to test the hypothesis. The first model was a cross-sectional model for testing the causal relationships among job accessibility, employment ratio and auto ownership. The core of this model is the assumption of reciprocal interaction among job accessibility, employment ratio and auto ownership, and the feedback loops through income. This model structure incorporated the interdependency among the four endogenous variables and the correlations among the error terms of the endogenous variables. Therefore, the estimates of the structural coefficients were not subject to endogeneity bias as they were in linear regression models, in which the coefficients between the variables were estimated separately. The second model is an unconditional change score model for testing the causal relations among the changes of the variables. In this model, each variable was computed as the differences between two points of time.

Chapter Three empirically demonstrated the cross-sectional model. The results of the SEM showed that (1) job accessibility had a positive direct effect on workers per capita and a negative direct effect on autos per capita; (2) workers per capita had a positive direct effect on autos per capita and income per capita; 3) income per capita had
a positive direct effect on job accessibility and autos per capita; (4) autos per capita had a positive direct effect on workers per capita; (5) education attainment had a positive direct effect on workers per capita and income per capita; (6) household size had a negative direct effect on autos per capita (and workers per capita in the competing model); (7) job accessibility was affected positively by percentage of rental housing units and negatively by median rent; (8) the error terms of job accessibility, workers per capita and autos per capita were significantly correlated; (9) workers per capita, income per capita, and autos per capita had positive total effects on all endogenous variables; (10) job accessibility had a negative total effect on all other endogenous variables; and (11) workers per capita and autos per capita did not have significant direct effects on job accessibility. The significant correlations among the error terms of job accessibility, workers per capita, and autos per capita suggest that the estimates in single-equation linear regression models are not efficient. In addition, when the reciprocal interactions between job accessibility and autos per capita, and job accessibility and workers per capita were estimated separately in the linear regression model, the coefficients on both directions were significant. When they were estimated simultaneously in the SEM, the direct effects of autos per capita and workers per capita on job accessibility were not significant. Since the simultaneous estimation used more information in the data than the estimations in the separate linear regression models, the results of the SEM were more trustworthy than those of the linear regression models. The failure to reject the reciprocal structural coefficients suggested the existence of endogeneity. Therefore, in this context, the linear regression models were subject to endogeneity bias, and hence not suitable.
Chapter Four examined the generalizability of the model developed in Chapter Three, and the effects of multivariate nonnormality on model goodness-of-fit, parameter estimates, and standard errors of the parameter estimates. When the CT sample was substituted with the BG sample, a structural coefficient (median rent on job accessibility) became insignificant due to high multivariate nonnormality. After the outliers were deleted from the sample and the multivariate kurtosis was lowered, the coefficient of median rent on job accessibility became significant. Therefore, the model structure developed on the CT sample could be applied to the BG sample. After multivariate normality was reached through deleting observations with bigger Mahalanobis distances, the sample moments (means and standard deviations) of the reduced samples became smaller than those of original samples. Three coefficients at the CT level and a coefficient at the BG level became insignificant. From a policy standpoint, the deleted observations contained important socio-economic information. Furthermore, a sample with smaller Mardia’s multivariate kurtosis did not show substantial improvements in discrepancy function and model fit indices at both the CT and BG level. Therefore, deletion of observations, if it has to be used, should be limited to true outliers instead of all observations with bigger Mahalanobis distances. I argued that the cutoff for normality should be multivariate-instead of univariate-based. When Mardia’s multivariate kurtoses of the CT and BG samples were around 25, our results were consistent with those in the literature and seemed to be trustworthy. For this reason, the tradeoffs between the loss of information and the desirability of the results in policy analysis seemed to favoring accepting that level of nonnormality.