Ann Reg Sci (2008) 42:341–356
DOI 10.1007/s00168-007-0154-2

ORIGINAL PAPER

Exploring the connections among job accessibility, employment, income, and auto ownership using structural equation modeling

Shengyi Gao · Patricia L. Mokhtarian · Robert A. Johnston

Received: 17 November 2006 / Accepted: 6 July 2007 / Published online: 18 August 2007

Abstract Using structural equation modeling, this study empirically examines the 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, CA. 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 are largely consistent with theory and/or with empirical observations across a variety of geographic contexts. It is demonstrated that structural equation modeling is a powerful tool for capturing the endogeneity among job accessibility, employment, income, and auto ownership.

JEL Classification C310 · R400 · R200

1 Introduction

In regional transportation and land use planning, job accessibility is often used to evaluate the ability of workers living in a zone to access jobs in all zones. It is

S. Gao (✉) · R. A. Johnston
Department of Environmental Science and Policy, University of California,
One Shields Ave., Davis, CA 95616, USA
e-mail: sgao@ucdavis.edu

R. A. Johnston
e-mail: rajohnston@ucdavis.edu

P. L. Mokhtarian
Department of Civil and Environmental Engineering, University of California,
One Shields Ave., Davis, CA 95616, USA
e-mail: plmokhtarian@ucdavis.edu

mathematically written as the sum across zones of the product of a travel friction factor (Kockelman 1997; Levinson 1998; Raphael 1998) from the given zone to each other zone, and the number of jobs in other zones.

A factor that is seldom emphasized in discussing job accessibility is location. Location in a travel model usually refers to a traffic analysis zone (TAZ), which has an explicit geographic boundary. In urban economics, by contrast, location is interpreted as a consumption bundle, including 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 fundamentally 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 result in changes in the others. Econometrically, this would lead to the circumstance that the error term in the equation for one variable will not be independent from the observed explanatory variables in the same equation—in violation of the assumptions required for ordinary least squares (OLS) regression. If so, an appropriate approach to describe the interdependence, or endogeneity, among the variables is to develop a simultaneous equation system instead of a single equation (Ihlanfeldt and Sjoquist 1990).

Unfortunately, in policy analyses, professional practice, and empirical studies, the intervening interactions are often treated as unidirectional and thus are simulated with classic linear regression models. 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 zero) is violated, and the 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).

It is very likely that the improper choice of model in this context will lead to improper interpretation of statistical inferences and impacts of policy implementation. 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 of the impact of job accessibility on employment, and attributed the insignificant impact of job accessibility on employment to this neglect and the consequent use of 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 1990). However, they only empirically demonstrated the first approach.

In this study, we propose a structural equation model (SEM) that depicts relationships among job accessibility, employment, income, and auto ownership. With census

tract data, we empirically implement the SEM and demonstrate that the SEM is a proper approach to explore these relationships.[1]

2 Literature review

In the 1950s, sociologists noticed a rapid decentralization of jobs in many American cities, especially those cities noted for blue collar industries, and started to explore the impacts of job decentralization on blacks who could not move to the suburbs with the jobs, due to residential discrimination. Employing a linear regression model, Kain (1968) first discussed the relationship between job decentralization and high unemployment rates among blacks living in central cities. He concluded that residential segregation plus the barrier of transportation to work led to a high unemployment rate among the blacks living in the central cities. Kain's findings were called the spatial mismatch hypothesis (SMH) and were empirically tested with aggregate and disaggregate data (Cooke 1997; Ellwood 1986; Holzer et al. 1994; Hughes and Madden 1991; Ihlanfeldt and Sjoquist 1990; Immergluck 1998; McLafferty and Preston 1996; Ong and Miller 2003; Raphael 1998). These studies produced inconsistent conclusions on the correlations between job accessibility and unemployment rates among blacks living in central cities. The inconsistencies might be attributed to three flaws in the empirical studies. First, many empirical studies did not take into account the endogeneity of residence, namely the correlation between employment location choice and residential location choice, and used the OLS procedure (Cooke 1997; Pastor and Adams 1996). Theoretically, if endogeneity exists, the premise of applying OLS procedures does not hold (Finkel 1995; Kline 2005). The estimates of the coefficients in OLS models will not be reliable, due to endogeneity bias. Secondly, some studies used youth who lived with their parents as their sample, to avoid endogeneity bias in linear regression models. In this case, the exogeneity of the residential location of the youth held only if the parents' residential location choice was exogenous, which is conceptually debatable and empirically hard to prove. Furthermore, anyone who did not live with their parents was excluded from the sample, which necessarily limited the generalizability of the results. Thirdly, none of these studies took into account the possible interactions between job accessibility and auto ownership and also the interactions between employment rate and auto ownership.

For all these reasons, the findings in the literature are subject to challenge in terms of model framework and specialized samples, regardless of whether the results are supportive to the SMH. Thus, if the first approach suggested by Ihlanfeldt and Sjoquist (1990) does not work well due to biased samples, the only solution is to build a

[1] The difficulties of inferring causality from cross-sectional data, especially aggregate data, are well-known. However, in practice, aggregate cross-sectional data are widely used to explore the causal relationships between a dependent variable and explanatory variables, even with single-equation models. In fact, as we argue herein, it is for single-equation models that the use of cross-sectional data to infer causality is most questionable. When multi-equation SEMs are used to capture feedback loops and indirect relationships, cross-sectional data can provide an appropriate model for a system in dynamic equilibrium. Thus, although no modeling approach is perfect, we are consistent with the mainstream of SEM practice when we interpret our results in terms of causal effects.

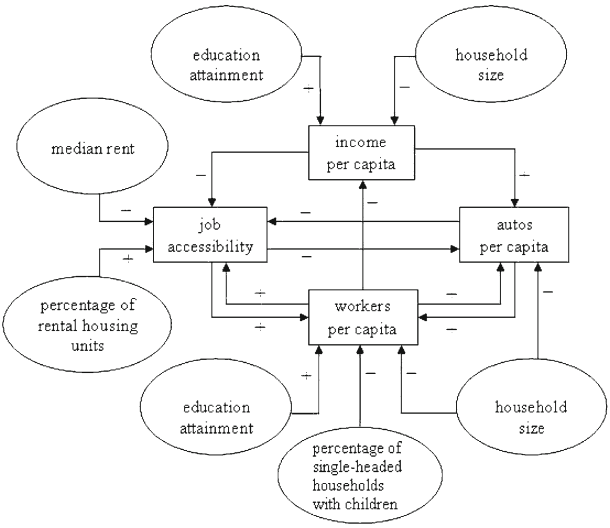

Fig. 1 Conceptual structural equation model

simultaneous equation system to properly reflect the endogeneity of all variables, using a generalizable sample.

3 Conceptual framework

From the perspective of economic theory, job accessibility, employment, and auto ownership are interdependent. Figure 1 illustrates the conceptual model from the perspective of the aggregate data available to us. A single arrow represents the direction of a direct effect from one variable to another variable, while reciprocal arrows represent possible interactions between two endogenous variables. The + and − signs represent the expected nature of the impacts. Land use patterns determine locations of jobs and residences, and as shown in the figure, we postulate that location-based characteristics such as employment and rent levels accordingly affect job accessibility. Income and auto ownership also affect residential location choice and job accessibility (Holzer 1991; Schimek 1996). Job accessibility, education attainment, and auto ownership are expected to have positive impacts on employment (Cervero et al. 2000; Kasarda and Ting 1996; Kockelman 1997; Ong and Blumenberg 1998), 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 (Dargay 2001; Kockelman 1997). 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 (because the endogenous variable of one equation is an explanatory variables in others) the explanatory variables are not independent from the error terms in an equation.

Table 1 Descriptive statistics of the variables ($N = 266$)

Variable	Minimum	Maximum	Mean	Standard deviation
Percentage of rental housing units	1.59	99.24	40.31	20.93
Median rent	378.00	1,461.00	747.80	173.64
Percentage of single-headed households with children	1.00	28.00	11.44	5.16
Education attainment	0.01	0.54	0.16	0.11
Household size	1.30	4.34	2.72	0.52
Job accessibility	17,663.00	71,263.00	41,486.76	9,250.32
Workers per capita	0.23	0.74	0.45	0.08
Income per capita	6,754.00	49,729.00	21,615.80	8,174.85
Autos per capita	0.33	0.85	0.63	0.11

In the literature, the computation of the variables from aggregate data takes either a per-household (e.g., Schimek 1996) or per-capita (e.g., Kockelman 1997) form. 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, and household income also tends to be smaller in central-city census tracts. Thus, 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, we choose to compute the variables on a 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.

4 Data

This analysis is done at the census tract level in Sacramento County, CA. 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). We use the year 2000 census tract data to extract socio-economic variables, and the year 2000 travel demand forecasting model of the Sacramento Area Council of Governments (SACOG) to obtain the number of jobs for each tract and travel time (A.M. peak period) between TAZs.

The descriptive statistics of the variables are shown in Table 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 census tract data. *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 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 (for details, see Gao 2006).

5 Results and discussion

5.1 Conformance to multivariate normality

In maximum likelihood estimation of the SEM, 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) have demonstrated that excessive non-normality inflates the χ^2-statistic 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, non-normality 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. Multivariate normality can be tested in many ways (Bollen 1989; D'Agostino 1986; Mardia 1970), but Mardia's coefficients of skewness and kurtosis are used most often in structural equation modeling software. In AMOS 5, with which we estimated our hypothesized model, 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 non-normality generally report only univariate non-normality instead of multivariate non-normality (Andreassen et al. 2006; Hu and Bentler 1995; West et al. 1995; Yuan et al. 2005) and do not recommend a cutoff for the multivariate normality measure.

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. To minimize the risk of inflating the significance tests in this case due to a relatively large multivariate kurtosis, we 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. The larger the distance is, the larger the contribution an observation is making to the departure from multivariate normality (Bollen 1989; Mardia 1970). Removing the observations with biggest Mahalanobis distance will reduce the multivariate kurtosis and thus the critical ratio. We 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 observations 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.

We note that, for our sample, every time outliers are removed, the χ^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 χ^2-statistic is the product of the sample size minus one ($N - 1$) and the minimized discrepancy function ($F_{\min}$), a larger χ^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, we find that these observations are census tracts of great interest in policy analyses. These observations have much larger values of *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 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 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. Bagley and Mokhtarian (2000) 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, we 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

Table 2 Normality evaluation of the variables in the final model ($N = 266$)

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

original data. The results[2] reported in this paper are based on those 266 observations. The results of the multivariate normality evaluation are shown in Table 2.

5.2 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. 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. If they exist, and the conceptual model is disentangled as four linear regression models, the parameter estimates of the OLS models 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

[2] Under the specification of the conceptual model, we 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). For details, see Gao (2006).

estimates were unbiased and consistent, the significance tests on those parameter estimates would be unreliable in the presence of correlated error terms.

5.3 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), we report the model fit indices from several different index families. The model fit indices of the hypothesized model[3] are compared with the indices of the independence and saturated models, which are two opposite extreme cases. 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. All the indices in Table 3 suggest a good fit of the hypothesized model.

5.4 Direct and total effects in the SEM

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 (Hayduk 1987). Table 4 shows the significant standardized direct effects of the final model (the insignificant coefficients in the initial model are treated as zero 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 by *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 close to commercial or industrial land uses have more multi-family land use designations. Therefore, we 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 we do not have enough information about the rental housing market in the study area and do not know the other determinants of rent in the census data, such as square feet per housing unit, school quality, crime rate, and livability of the neighborhood, we do not know whether the rent per square foot in places with higher *job accessibility* is truly lower than that in places with lower *job accessibility*. This needs further study in the future.

In Sacramento County, although Rancho Cordova is a thriving suburban job center, the downtown area is the most important employment center in terms of existing

[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 we report our results.

Table 3 Selected model fit indices

Model fit index	Independence model	Hypothesized model	Saturated model	Note
χ^2	2,105.00	81.64	0.00	Measuring the discrepancy between the sample and population covariance matrices; the smaller, the better; sample size dependent
Degrees of freedom (df)	36.00	6.00	0.00	
Goodness of fit index (GFI)	0.36	0.94	1.00	Assessing the proportion of the variability in the sample covariance matrix explained by the model; GFI > 0.9 suggests a good fit
Normed fit index (NFI)	0.00	0.96	1.00	Not parsimony adjusted; normed; NFI > 0.9 suggests a good fit
Incremental fit index (IFI)	0.00	0.96	1.00	Assessing the improvement of the hypothesized model over the independence model; IFI > 0.9 suggests a good fit
Comparative fit index (CFI)	0.00	0.96	1.00	Assuming non-central chi-square distribution; assessing the improvement of the hypothesized model relative to the independence model. About 0.90 or higher suggests a good fit
Akaike information criterion (AIC)	2,121.00	141.64	72.00	Parsimony adjusted; the closer to the value of saturated model, the better the hypothesized model
Expected cross-validation index (ECVI)	8.00	0.53	0.27	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

jobs. Therefore, 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 typically in suburban areas while the downtown area has the highest *job accessibility*. The actual result is opposite to our assumption: higher per-capita incomes are associated with higher *job accessibility* in the model. By mapping the observations in a Geographic Information System (GIS), we find that those census tracts with higher

Table 4 Standardized direct and total effects of the SEM

	Percentage of rental housing units	Median rent	Education attainment	Household size	Job accessibility	Workers per capita	Income per capita	Autos per capita
Direct effect								
Job accessibility	0.64***	−0.21***	–	–		–	0.38***	–
Workers per capita	–	–	0.29***	–	0.05*		–	0.56***
Income per capita	–	–	0.40***	–	–	0.76***		–
Autos per capita	–	–	–	−0.60***	−0.80***	0.17***	0.37***	
Total effect								
Job accessibility	0.55[a]	−0.18[a]	0.26	−0.12	−0.14	0.34	0.40[b]	0.19
Workers per capita	−0.30	0.10	0.36[b]	−0.39	−0.46[a]	0.16	0.06	0.65[b]
Income per capita	−0.27	0.08	0.67[b]	−0.30	−0.36	0.88[b]	0.05	0.50
Autos per capita	−0.58	0.19	0.10	−0.69[b]	−0.90[b]	0.25[b]	0.07[a]	0.14

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

-: hypothesized but not statistically significant and therefore constrained to be 0 in final model

–: no direct effect hypothesized in conceptual model

[a] Opposing direct and indirect effects

[b] Synergistic direct and indirect effects

income per capita tend to be very close to downtown Sacramento or Rancho Cordova, where *job accessibility* is high. In other words, a high proportion of richer suburban areas have high *job accessibility* instead of low *job accessibility* as anticipated. Therefore, we think the positive influence of *income per capita* on *job accessibility* correctly represents what is observed.

Why does this happen? Since we do not have individual socio-economic and attitudinal data for each household, we cannot give an explicit explanation of this result. Besides the possible self-selection in residence, another possible reason relates to the distribution of one-person households. We 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 lead to a larger *income per capita*. But 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 (higher income per capita) 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 income from all sources. Due to lack of information on income from non-work sources, we cannot split income into work and non-work sources. We 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* and *workers per capita* have the same signs as predicted, but are not significant. Therefore, these five 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 (however, they have sizable indirect effects through *income per capita*, as noted below).

The indirect effect is the effect that a variable exerts on another variable through one or more endogenous variables. Depending on the sign, the indirect effect of one variable on another variable may strengthen or offset its corresponding direct effect. The sum of the direct and indirect effects of a variable is the total effect. Comparing the direct and total effects in Table 4, we can see that 8 out of the 12 total effects whose direct effects are significant have larger magnitudes 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? Because a 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.

6 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, we 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 correlation as treated in many empirical studies and therefore endogeneity among the key variables is demonstrated. Thus, the linear regression models should be rejected in this context. 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 total 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 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. 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 non-recursive model like the one in this study, greatly increase the complexity of the model, but provide a complete picture of the 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. These two advantages will substantially lower the odds of an analyst misinterpreting the correlations among the 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 shed some light on the dispute whether there is a causal relation between job accessibility and employment. As we have shown, endogeneity among job accessibility, employment, and auto ownership does exist, and the endogeneity bias is substantial and should not be neglected.[4] 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.

There is nearly always a tradeoff between explanatory power and straightforwardness when choosing a model in a study. A more complex model generally brings in more explanatory power but needs more effort to estimate and interpret it. As demonstrated in this study, choosing a more complicated model is important not only for improving the accuracy of the parameter estimates, but also for exploring relationships among critical variables that cannot be explained by simple models. For these reasons, in this context, the benefits of the additional complexity of the structural equation model outweigh its costs.

Acknowledgments We would like to thank the University of California Transportation Center for funding this research. We would like also to thank the two anonymous peer reviewers of this paper for their valuable comments.

References

Andreassen TW, Lorentzen BG, Olsson UH (2006) The impact of non-normality and estimation methods in SEM on satisfaction research in marketing. Qual Quan 40:39–58

Bagley MN, Mokhtarian PL (2000) The impact of residential neighborhood type on travel behavior: a structural equations modeling approach. Ann Reg Sci 36:279–297

Bollen KA (1989) Structural equations with latent variables. Wiley, New York

Bollen KA, Long JS (1993) Introduction. In: Bollen KA, Long JS (eds) Testing structural equation models. SAGE Publications, Newbury Park, pp 1–15

[4] For the details of the comparison between the SEM and the linear regression models, see Gao (2006).

Cervero R, Sandoval O, Landis J (2000) Transportation as a stimulus to welfare-to-work: private versus public mobility. J Plan Edu Res 22:50–63
Cooke TJ (1997) Geographic access to job opportunities and labor force participation among women and African Americans in the Greater Boston Metropolitan Area. Urban Geogr 18:213–217
Curran PJ, West SG, Finch JF (1996) The robustness of test statistics to non-normality and specification error in confirmatory factor analysis. Psych Meth 1:16–29
D'Agostino RB (1986) Tests for the normal distribution. In: D'Agostino RB, Stephens MA (eds) Goodness-of-fit techniques. Marcel Dekker, New York, pp 367–419
Dargay JM (2001) The effect of income on car ownership: evidence of asymmetry. Trans Res Part A 35:807–821
Ellwood DT (1986) The spatial mismatch hypothesis: are there teen-age jobs missing in the ghetto? In: Freeman RB, Holzer HJ (eds) The black youth employment crisis. University of Chicago Press, Chicago, pp 147–187
Greene, WH (1997) Econometric analysis, 3rd edn. Prentice Hall, Upper Saddle River. NJ
Gao SY (2006) Building causal connections among job accessibility, employment, income, and auto ownership using structural equation modeling: a case study in Sacramento County. PhD Dissertation, Transportation Technology and Policy Graduate Group, University of California, Davis, June. Available at http://www.ice.ucdavis.edu/files/ice/PhDdissertation.pdf. Accessed on 4 June 2007
Finkel SE (1995) Causal analysis with panel data. SAGE Publications, Thousand Oaks
Hayduk LA (1987) Structural equation modeling with LISREL: essentials and advances. The John Hopkins University Press, Baltimore
Holzer HJ (1991) The spatial mismatch hypothesis: what has the evidence shown? Urban Stud 28:105–122
Holzer HJ, Ihlanfeldt KR, Sjoquist DL (1994) Work, search, and travel among white and black youth. J Urban Econ 35:320–345
Hoyle RH, Panter AT (1995) Writing about structural equation models. In: Hoyle RH (ed) Structural equation modeling: concepts, issues and applications. SAGE Publications, Thousand Oaks, pp 159–176
Hu LT, Bentler PM (1995) Evaluating model fit. In: Hoyle RH (ed) Structural equation modeling: concepts, issues and applications. SAGE Publications, Thousand Oaks, pp 76–79
Hughes MA, Madden JF (1991) Residential segregation and the economic status of black workers: new evidence for an old debate. J Urban Econ 29:28–49
Ihlanfeldt KR, Sjoquist DL (1990) The effect of residential location on the probability of Black and White teen teenagers having a job. Rev Reg Stud 20:10–20
Ihlanfeldt KR, Sjoquist DL (1991) The effect of job access on black and white youth employment: a cross-sectional analysis. Urban Stud 28:255–265
Ihlanfeldt KR, Sjoquist DL (1998) The spatial mismatch hypothesis: a review of recent studies and their implications for welfare reform. Hous Pol Deb 9:849–892
Immergluck D (1998) Job proximity and the urban employment problem: do suitable nearby jobs improve neighborhood employment rates? Urban Stud 35:7–23
Kain J (1968) Housing segregation, negro employment, and metropolitan decentralization. Quar J Econ 88:513–519
Kasarda JD, Ting KF (1996) Joblessness and poverty in America's central cities: causes and policy prescriptions. Hous Pol Deb 7:387–419
Kline RB (2005) Principles and practice of structural equation modeling, 2nd edn. The Guilford Press, New York
Kockelman KM (1997) Travel behavior as function of accessibility, land use mixing, and land use balance: evidence from San Francisco Bay Area. Transp Res Rec 1607:116–125
Lei M, Lomax RG (2005) The effect of varying degrees of non-normality in structural equation modeling. Struct Equ Model 12:1–27
Levinson DM (1998) Accessibility and the journey to work. J Trans Geogr 6:11–21
Mardia KV (1970) Measures of multivariate skewness and kurtosis with applications. Biometrika 57:519–530
Mayston D (2005) Structural determinants of cumulative endogeneity bias. Discussion papers in economics. Available at http://www.york.ac.uk/depts/econ/dp/0511.dbf. Accessed on 30 November 2005
McLafferty S, Preston V (1996) Spatial mismatch and employment in a decade of restructuring. Prof Geogr 48:417–467
Ong PM, Blumenberg E (1998) Job access, commute and travel burden among welfare recipients. Urban Stud 35:77–93

Ong PM, Miller D (2003) Spatial and transportation mismatch in Los Angeles. Available at: http://www.uctc.net/papers/653.pdf. Accessed on 20 April 2004

Pastor MJ, Adams AR (1996) Keeping down with the joneses: neighbors, networks, and wages. Rev Reg Stud 26:115–145

Ramanathan R (2002) Introductory econometrics with applications, 5th edn. Harcourt Brace & Company, Orlando

Raphael S (1998) The spatial mismatch hypothesis and black youth joblessness: evidence from the San Francisco Bay Area. J Urban Econ 43:79–111

Shah R, Goldstein SM (2006) Use of structural equation modeling in operations management research: looking back and forward. J Oper Manag 24:148–169

Schimek P (1996) Household motor vehicle ownership and use: how much does residential density matter? Transp Res Rec 1552:120–125

Shen Q, Sanchez TW (2005) Residential location, transportation, and welfare-to-work in the United States: a case study of Milwaukee. Hous Pol Deb 16:393–431

Yuan KH, Bentler PM, Zhang W (2005) The effect of skewness and kurtosis on mean and covariance structure analysis: The univariate case and its multivariate implication. Socio Meth & Res 34:240–258

West SG, Finch JF, Curran PJ (1995) Structural equations models with non-normal variables. In: Hoyle RH (ed) Structural equation modeling: concepts, issues and applications. SAGE Publications, Thousand Oaks, pp 56–75