# Addressing Endogeneity in Residential Location Models

by

Cristian Angelo Guevara

Ingeniero Civil,
Magíster en Ciencias de la Ingenieria Mencion Transporte (2000)
Universidad de Chile, Santiago, Chile

Submitted to the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degree of

Master of Science in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2005





Signature of the Author_____
Department of Civil and Environmental Engineering
May 6, 2005

Certified by_____
Moshe E. Ben-Akiva
Edmund K. Turner Professor of Civil and Environmental Engineering Thesis Supervisor

Accepted by_____
Andrew Whittle
Chairman, Departamental Committee for Graduate Students

# Addressing Endogeneity in Residential Location Models

by

Cristian A. Guevara

Submitted to the Department of Civil and Environmental Engineering
on May 6, 2005, in partial fulfillment of the requirements
for the degree of Master of Science in Transportation

Abstract

Some empirical residential location choice models have reported dwelling-unit price estimated parameters that are small, not statistically significant, or even positive. This would imply that households are non-sensitive to changes in dwelling unit prices or location taxes, which is not only against intuition, but also makes the models useless for policy analysis.

One explanation for this result is price endogeneity, which means that the price is correlated with the error term in the econometric model. This problem is caused either by the simultaneous determination of the supply and the demand for dwelling units in aggregated models, or by omitted attributes that are correlated with the price, in the disaggregated ones. The treatment of endogeneity in discrete choice models is an area of ongoing research in econometrics. Therefore, methods to treat this problem began to be proposed only in the last decade, and have not been thoroughly analyzed for residential location models.

This thesis evaluated the available methods to treat endogeneity in discrete choice models. Each method was tested in terms of its applicability and robustness in a residential location choice framework, using a set of Monte Carlo experiments. The results showed that the control-function method (Petrin and Train, 2004) is the most promising one to address endogeneity in this framework because it is the best to handle individual level endogeneity and it is tractable with available estimation software.

Finally, the application of the control-function method to an example based on real data from Santiago de Chile showed not only that the problem of price endogeneity does exist in residential location choice models, but also that the control-function method gives a satisfactory answer to the problem. Further venues of research are discussed at the end of the thesis, in particular, the usage of non-parametric methods to improve the estimation results of the control-function method.

Thesis Supervisor: Moshe E. Ben-Akiva

Title: Edmund K. Turner Professor of Civil and Environmental Engineering

# Acknowledgements

I wish to thank my wife Erika: for all the love, the patience, the support, the joy, and for going to the other side of the planet, with a single piece of luggage and many uncertainties, just to share this challenge with me. My sons Diego and Bruno: for the joy, the hope, and the constantly renewed definitions of life that they give me every day.

My mother Amparo: for being a limitless source of encouragement (not impartial but persistent enough to make me believe half of it), inspiration, and example. My sisters Gabriela and Monica: for their constant support and encouragement.

Professor Moshe Ben-Akiva: for the invaluable advice that put me to this interesting and challenging research track, the careful review of the drafts of this thesis and for forcing me to make myself clear. Professor Nigel Wilson: for his always friendly advice, academic encouragement and time to hear me in the most difficult moments. Professor Joe Sussman: for his advice and financial support. Professor Sergio Jara-Diaz: for his personal advice, encouragement and support. Professor Francisco Martinez: for his academic advice in urban economics and land-use modeling. Professors Marcela Munizaga and Juan de Dios Ortuzar: for their support in my application to MIT. Alan Thomas and Esteban Godoy: for their help in obtaining and processing the data.

The Integrated Program on Urban, Regional and Global Air Pollution: for the financial support provided with funds of the Fideicomiso Ambiental del Valle de Mexico and the Mario Molina Center for Strategic Studies in Energy and the Environment. Fulbright-Chile and MIDEPLAN: for the fellowships that partially helped me in funding my studies. Aldo Signiorelli, Henry Malbran and Guillermo Diaz: for the additional economic support from Sectra and the Transportation Ministry of Chile.

Janine, Travis, Jeff, Bernardo and Alvaro: for making the sometimes overwhelming study days at MIT a very nice experience.

And finally, Patricia, Marco, Roberto, Cristina, Felipe, Barbara, Francisco and all the chilenos_en_boston who helped us in our Boston Challenge.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Motivation

The development of models to understand the behavior of an urban system as a whole is crucial, not only for the evaluation and exploration of policies to manage the level of pollution and congestion in the cities, but also for solving problems such as the efficient provision of: utilities, transportation infrastructure and schools.

A general framework to model the decisions relevant to travel demand and their interactions with the urban system can be stated in three interrelated stages, corresponding to the *urban development* or land-use, the *travel demand* and the *transportation system performance* (Figure 1-1). Depending on the level of detail and the scale of the analysis, this framework can be used to model aggregated flows of trips by geographical zones or disaggregated household and individual decisions that lead to a demand for travel, including mobility and lifestyle, activity/travel scheduling, and their revisions (Ben-Akiva et al., 1996).

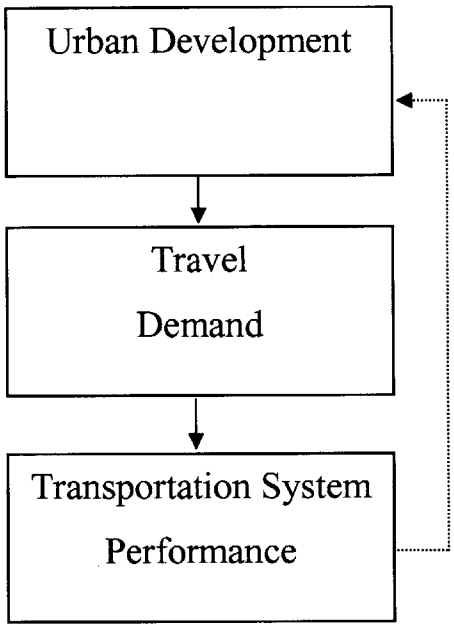



**Figure 1-1 Urban System Modeling Framework**

The key task in capturing the long-term behavior of an urban system is to forecast the changes in land-use patterns. Thus, understanding the behavior of the land market is crucial in the development of policy tools for shaping the urban system or managing the externalities associated with it. This observation is particularly relevant for cities of developing countries, such as Mexico or Chile, where dramatic changes in density and income composition of suburban areas is expected due to the simultaneous increase of motorization, income and highway infrastructure, just as it was observed in United States cities after World War II (Weisbord et al., 1980).

The land market itself is a very complex system comprised of many subsystems. There is an industrial and a residential land market. Each market has unique dynamics resulting from the interaction between agents offering and demanding industrial land or dwelling units. This research only focuses on the residential market due to its primary effect on the urban system through the determination of the core part of the origin-destination trip matrix. Furthermore, for the sake of simplicity, it will be assumed that the supply of dwelling units is exogenous, and thus, this thesis will be focused on residential location choice models.

## 1.2 Statement of the Problem

Within this framework, an econometric issue of relevance arises. Consider a situation where households choose to buy or rent dwelling units of a specific type in certain locations. If this problem is modeled using an aggregated framework, the number of dwelling units sold will necessarily depend upon the price and other characteristics of the dwelling unit. At the same time, the price of the dwelling unit will depend upon the demand for the dwelling unit. The simultaneous determination of these variables causes the independent variables of the model to be correlated with the error term (non-orthogonality), breaking then a crucial assumption made in the estimation of the model, which makes the estimated parameters inconsistent and biased.

Even in a discrete choice framework, where the simultaneous determination effect can be argued not to occur, a different but very plausible problem can arise. If a quality attribute, that is relevant for the decision maker, is not observed by the researcher and it is correlated with the price of the dwelling unit, the price will be correlated with the error term causing, again, a non-orthogonality problem.

Consider for example the residential location choice decision problem described in Figure 1-2. In this case, the decision maker can perceive all the attributes of the alternatives in her choice set, including the non-measurable one that is explained through the photographs. However the researcher can observe only the number of bedrooms, the size, the price and whether utilities are provided or not for each alternative. In the viewpoint of the researcher, individuals who decide to live in apartment *A* would be *not very smart* because they reject such a bargain as *B* or, more formally, just not sensitive to price.

However, what it is really occurring is that a relevant quality attribute, which is correlated with the price, has being omitted. It will be shown in this thesis that this problem implies that an upward bias for the price parameter (a negative parameter that becomes more positive) will usually occur in residential location models.

A) 1 Bedroom; 70 $m^2$; with utilities; \$2000 B) 1 Bedroom; 70 $m^2$; with utilities; \$100

**Figure 1-2: Example of an Omitted Quality Attribute**

Many empirical residential location choice models (Bhat and Guo, 2004; Sermons and Koppelman, 2001; Levine, 1998; Waddell, 1992, 1996; Quigley 1976) have reported non-significant, small or even positive dwelling unit price parameters. The hypothesis of this thesis is that this is caused by endogeneity due to the existence of some omitted attributes that are correlated with the dwelling unit's price. This hypothesis has been considered only in very few studies related to residential location (Bayer et al., 2004 and Ferreira, 2004). Furthermore, endogeneity in discrete choice models is an area of current development in econometrics, which makes the study of its effects in residential location modeling a challenging and interesting issue to address.

## 1.3 Thesis Objectives and Outline

The objectives of this thesis are the following:

1. To develop a conceptual framework for the treatment of endogeneity in residential location choice models.
2. To test the available methods to address endogeneity in discrete choice models of residential location, using a set of Monte Carlo experiments.
3. To verify the relevance of the endogeneity problem and the quality of the proposed correction methods, by estimating a model of residential location based on real data.

The structure of this thesis is as follows. The next chapter presents a critical analysis of the state of the art in residential location modeling in light of the problem of price endogeneity. The third chapter analyzes different issues related to the treatment of endogeneity in discrete choice models of residential location. In Chapter 4, a comparison of the different available methods to address endogeneity in discrete choice models of residential location is made through a set of Monte Carlo experiments. The fifth chapter is an application of the most suitable method to address endogeneity in residential location, as found in Chapter 4, to real data from Santiago de Chile. Finally, Chapter 6 presents the general conclusions and recommendations for further research.

# Chapter 2

# Residential Location Modeling

This chapter reviews the state of the art in residential location modeling in light of the price endogeneity problem. This task is accomplished in four steps. First, the chapter reviews theoretical considerations in modeling the residential location market. Second, the chapter reviews specification issues of relevance in the development of residential location choice models, including the selection of explanatory variables, sample and choice set definition, the level of aggregation of the alternatives and econometric issues related to the error structure that can be expected in modeling this market. Afterwards, a brief discussion about the relationship between land-use and transportation is presented. Finally a set of recent residential location models where price endogeneity appears to be an issue is presented.

# 2.1 Theoretical Considerations in Modeling the Residential Location Market

## 2.1.1 Microeconomic Framework

The general microeconomic framework to model residential location is based on the assumption that each household as a whole maximizes the combined *utility* of its members. This utility depends on the time and goods associated with the activities that each household's member performs, including transportation, work and other derived activities. It is also assumed that, conditional on a specific location *i*, households face constraints associated with: the available time that their members have to perform the activities, their budgets and the feasibility of developing each set of activities and their associated goods consumption.

The solution of this problem, conditional on each residential location *i*, corresponds to the set of activities that each member of each household perform and the associated time assignments and goods consumption. If these optimal values are replaced in the objective function of the utility maximization problem, it will be obtained what is known as the *conditional indirect utility* function.

Finally, it is assumed that households choose the residential location that has the larger conditional indirect utility function plus a random component, in what is known as the Random Utility Model (Ben-Akiva and Lerman, 1985).

## 2.1.2 Optimality of Current Location

The basic assumption required for modeling residential location choice using a utility maximization framework is that households are located in places that maximize their utilities given their budgets and other constraints. However, in reality, many households are located in non-optimal places due to the diverse costs (pecuniary, social or even emotional) associated with moving.

One way of addressing this problem is to model first the decision of whether to move, in order to consider the effect of inertia in the model. One example of this approach is Weisbord et al. (1980) where the authors used a Nested Logit model. However, to build a model like that, detailed information about the previous locations of the households is required, data that are not usually available, or at least, not with the required detail.

Another way to model the decision to move by any household is to add to the utility of alternatives new locations an additional cost associated with moving that is calculated based on external data and assumptions. However, the difficulty in obtaining adequate data to apply that approach makes its application infeasible.

Finally, other issue that would be relevant in addressing household residential location optimality is to consider a model to predict when a specific household decides to change from the rental to the ownership market, because each ownership status carries very different moving costs.

### 2.1.3 Market Size

Within this setting, the size of the market (that is, the number of households that want to move and the dwelling units that are available) will be a function of the characteristics of the system and also a function of time. This is because, even considering that the supply of new dwelling units is exogenous, any change in the characteristics of the system would affect the perception held by specific households about their present residential location, relative to the other alternatives. Sometimes, the effects of these changes will be strong enough to make the households to decide moving and thus to enter into the group of households seeking residential locations (the demand) and, at the same time, add their current dwelling units into the group of offered ones (the supply).

These changes can be external to the household, such as the provision of new transportation infrastructure or changes in school quality, crime or pollution rates in specific areas; or internal to the household such as changes in income, the births of children or problems with the neighbors.

### 2.1.4 Dwelling Unit's Price

Modeling dwelling unit's price in residential market is a challenging task because it does not only depend upon dwelling unit's attributes but also upon the differences between the demand and the supply which are caused both by the difficulties in adjusting the supply to the demand in the short run and because each good in this market has almost unique characteristics. Examples of different approaches to this problem can be found in Martinez and Henriquez (2005), Bayer et al. (2004) and Waddell (1996).

Martinez and Henriquez (2005) proposed a theoretically derived expression for the endogenous determination of dwelling unit's price or rent. This expression (2-1) is the expected value of the highest bid offered by the households for a specific dwelling unit and is derived under the assumption that bids are random variables that depend upon dwelling unit's characteristics. In their model, the bids consist of a systematic component $B$ and an error term that is assumed to be *identically and independently distributed* (*iid*) Extreme Value (0, $\mu$). Additionally, $\gamma$ is Euler's constant, $r_{vi}$ is the price of a dwelling unit of type $v$ located in $i$ and $N_g$ is the number of households of type $g$.

**(2-1)**
$$r_{vi} = \frac{1}{\mu} \ln\left( \sum_g N_g \exp\left(\mu B_{gvi}\right) \right) + \frac{\gamma}{\mu}$$

Conversely, Bayer et al. (2004) considered that dwelling unit's prices are endogenously determined as the prices that clear the market in an iterative process. This means that prices are adjusted up to the point where each dwelling unit is assigned to exactly one household. The authors affirm, based on the study of Berry (1994), that this procedure has a unique solution for the prices.

In reality, is not necessary that the each dwelling unit becomes chosen or that all households are able to find an affordable housing solution in the equilibrium. However, modeling a situation like that requires detailed information about the stock availability and household's budget, which is not usually available.

Finally, Waddell (1996) and Waddell and Borning (2004) in their URBANSIM model, used a micro-simulation approach to forecast the behavior of the land market. This model is implemented as a set of interacting model components that represent the

major actors and choices in the urban system. Within this framework, the author considered a hedonic price model that simulates the land prices for each modeling zone (cell), output that is used as an input by other model components. No theoretical considerations for the equilibrium in the system are stated by the authors.

### 2.1.5 Dwelling Unit's Supply

To model the long term equilibrium of the residential location market is necessary to have a model of dwelling units supply. This is because real estate developers should respond to price variations in the market, building more dwelling units where they expect to get greater benefits, changing with their decisions the system equilibrium by pulling down the higher dwelling units' prices and then affecting the demand structure in the whole system.

In this sense, the model of Martinez and Henriquez (2005) represents an interesting theoretical framework where real estates dwelling unit's supply are treated as random variables under a profit maximization framework that is consistent with the dwelling unit's demand model. On the other hand, the URBANSIM model of Waddell and Borning (2004) offers an attractive micro-simulation approach for this problem.

### 2.1.6 Workplace and Residential Location

A final question to address in modeling the equilibrium of the residential location market is the relationship between workplace location and residential location decisions.

It can be argued that the choice of work location is made conditional on the residential location, or visa versa. Alternatively, it can be argued that residential location decision is made in a longer time frame and thus, what households consider in their decision of where to live is the probability of having access to good jobs in each specific location.

If workplace determines residential location, the model must consider the distance between the present workplace and each residential location available. Other alternative

is to consider a more elaborated measure of the generalized cost, such as the expected maximum utility of the available modes, which depends on travel time and costs.

If, on the contrary, the residential location determines the workplace location, the residential location model should include a work accessibility measure, instead of the actual distance to workplace. Even though, if in this case the distance to workplace is included instead of the accessibility, this variable should be significant anyway, because actual work location can be seen as a proxy for work accessibility, because it is more probable to find a job in places where more jobs are offered. However, the meaning of the estimated parameters in that case will not be clear.

Now, if it is assumed that workplace location determines residential location and that residential location determines workplace location simultaneously, it would be necessary to address an endogeneity problem arising from this simultaneous determination, equivalent to the one described in this thesis for price in residential location modeling. The study of this effect is left for further research.

Finally, if residential location and workplace location occur in different time frames, what should be included in the model is the accessibility to a variety of workplaces instead of the actual distance to workplace.

## 2.2 Model Specification Issues

In this section is presented a set of specification issues that have to be considered in residential location modeling. These issues are related to the definition of the sample, the level of aggregation of the alternatives, the definition of the choice set, the error specification, and the type of explanatory variables that should be considered in residential location modeling.

### 2.2.1 Sample Definition

One alternative proposed in the literature to achieve the requirement of considering only optimally located households in the sample, is to include in the sample only households that have moved recently because it can be argued that these are the only ones that, for sure, are located in an optimum place (Weisbord et al., 1980). Another way of achieving this objective could be to consider only renters in the sample, under the assumption that they have lower moving cots associated. Against both the recently-moved and the renter assumptions can be argued that the modeling sample in these cases will not be representative of the whole population and that the selection criteria would imply such a reduction in the sample size that other estimation issues can arise.

Other issue to address in the sample definition is about the possibility that not every household chooses its location optimally on the basis of the attributes of the dwelling units. This is especially relevant in developing countries, where very-low income households live where they live because this dwelling unit is provided for free, or at a very non-market low price, by a relative (in what is known as "Allegados" or "Arrimados") or, in some cases, the place where they live was the result of the collective appropriation of a portion of land by the force (what is known as a "Toma" or "Ocupas"). In both cases, it is arguable that the households will be indeed located at their optimal locations but, because they do not have real feasible alternatives to their actual location, they cannot be modeled under the utility maximization framework used in this thesis.

Beyond this, if it would be possible to conduct a data collection activity specific for this research, without having to rely only on existing mobility surveys or census data, the ideal procedure would be to survey households that have recently changed their location. The survey would include questions about the alternatives that were considered in the process of choosing a residential location and their attributes.

### 2.2.2 Level of Aggregation of the Alternatives

A key consideration in modeling residential location choice is the definition of the alternatives. The usual approach is to consider administrative zones as the source to measure the spatial factors and as the alternatives themselves. This method follows Lerman (1975) who, arguing that the data is usually not available at an individual level, proposed a Multinomial Logit (MNL) choice model (2-2) where alternatives are groups of housing units instead of individual units. Within this framework, the author considered also a size variable (Ben-Akiva and Lerman, 1985) to account for the aggregation of alternatives.

Two problems arise with the aggregated approach described above. The first is that the definition of the *neighborhood* or the area of influence of the shared attributes is left as purely exogenous and subject to administrative definitions that are not necessarily aligned with the research interests. The second problem is that, by aggregating, the rich variability that can be found among dwelling units in the aggregated areas is lost.

An alternative approach consists in considering each dwelling unit as an alternative, recovering the neighborhood information on administrative zones (as in Bayer et al. 2004) or using some kind of Geographic Information System (GIS) to recover neighborhood information searching for the average attributes within some general or specific dwelling unit area. Working on this idea, Guo (2004) found that neighborhood socioeconomic attributes are important in a small contiguous area of the dwelling unit and land-use attributes are relevant in a larger radius.

## 2.2.3 Choice Set definition

The definition of the choice set in residential location modeling is not trivial because the huge number of alternatives to consider (potentially, each dwelling unit in the city) creates a computational burden of considerable proportions. The usual approach to deal with this problem is to use a randomly chosen sample of the available alternatives. McFadden (1978) proved that the estimated parameters using this procedure are consistent for a MNL model. The problem with the use of the MNL model in residential location modeling is that it can usually be rejected because, for example, nearby areas share unobserved attributes that make them correlated.

Another approach is to consider only a reduced set of feasible alternatives, where each available alternative can be selected using an *ordered availability* assumption (Ben-Akiva and Boccara, 1995). In the case of residential location modeling, this means that the probability that each alternative belongs to the choice set is determined as a function of the difference between some socio-demographic characteristic of the household at their current location and those of each considered alternative. The idea behind this is that alternatives that are too different to the current location are not really considered by the households when deciding where to move.

Finally, about which dwelling units to consider as available in the choice set, an additional argument in favor of considering only households that have recently changed their location is that, if it is considered a model where each dwelling unit is an alternative, it can be argued that the only available dwelling units correspond precisely to the ones that have been chosen by a household recently. The other dwelling units that should be considered as being available in the model are the ones that are unoccupied. However, this last piece of information is usually difficult to collect.

## 2.2.4 Error Specification

The assumptions that are made on the error structure of the discrete choice model imply a trade-off between tractability and accuracy. The simplest assumption to make is to accept that utility errors $\varepsilon_i^n$ are *iid* Extreme Value $(0,\mu)$, which leads to a very

attractive closed form (Ben-Akiva and Lerman, 1985) for the choice probability in what is known as the Multinomial Logit (MNL) model

$$\textbf{(2-2)} \qquad P_n(i) = \Pr\left[\underbrace{V_i^n + \varepsilon_i^n}_{U_i^n} \geq Max_{j \in C_n}\left\{\underbrace{V_j^n + \varepsilon_j^n}_{U_j^n}\right\}\right] = \frac{e^{\mu V_i^n}}{\sum_{j \in C_n} e^{\mu V_j^n}},$$

where $P_n(i)$ is the probability that individual $n$ chooses alternative $i$, $V_i^n$ is the deterministic part of the utility $U_i^n$ that individual $n$ retrieves from alternative $i$ and $C_n$ is the set of alternatives available to individual $n$. The parameter $\mu$ is the scale parameter of the Extreme Value distribution, which is inversely proportional to the standard deviation of the error and has to be normalized, usually to one, because it is non-identifiable.

The basic problem with the MNL model is that it is not able to reproduce the non-uniform substitution rates that can be expected between alternatives. This problem can be expressed in what is known as the *Independence from Irrelevant Alternatives* (IIA) property. For model (2-2), the ratio of the probabilities of two alternatives is independent of all other alternatives.

This means that, for example, if a person chooses between hotels for vacation, and the five-stars hotels double their previous prices, the change in the probabilities of all other types of hotels (including the half-star ones) will be the same, instead of, as it is intuitively expected, have a stronger effect in the ones with better rating. It has to be noted however that, the IIA property no longer holds in aggregated predictions for heterogeneous population, reducing its impact.

Many ways of avoiding the IIA property have been proposed in the literature. First, this can be solved by considering that the errors are Multivariate Normal, which leads to the Probit model, where the correlation matrix is not forced to be diagonal. The problem however is that the Probit model does not have a closed form and requires numerical integration, which increases the computational burden, making the problem almost intractable (even for today's computers) for a model with over six alternatives.

One closed form model that belongs to the *Generalized Extreme Value* (GEV) models and allows for correlation between groups of alternatives is the *Nested Logit* (NL) model (Ben-Akiva and Lerman, 1985). In this model, correlated alternatives are grouped in

nests allowing different scale factors for each group. The probability that an individual $n$ chooses a specific alternative $i$ will depend on the product of the probability of choosing the nest $k$, where $i$ belongs, and the probability of choosing the alternative conditional on the nest being chosen.

The expression for this probability is (2-3), where $\mu$ and $\lambda$ are scale parameters of the errors of the root and the nest respectively. These parameters are not jointly identifiable, problem that can be solved by fixing one of them, usually $\mu$, to be equal to one.

**(2-3)**
$$P_n(i) = P_n(k)P_n(i/k) = \frac{\exp(\mu V_k^n)}{\sum_{l \in Nests(n)} \exp(\mu V_l^n)} \frac{\exp(\lambda V_i^n)}{\sum_{j \in C_{n/k}} \exp(\lambda V_j^n)}$$

The utility of each nest is calculated as the *Expected Maximum Utility* (EMU) or inclusive value of the alternatives contained in the nest.

**(2-4)**
$$V_k^n = \frac{1}{\lambda} \ln\left( \sum_{j \in A(n/k)} \exp(\lambda V_j^n) \right) + \frac{\gamma}{\lambda}$$

One problem that makes this model unsuitable for residential location choice is the fact that it is restricted to have each alternative in only one nest. This is a problem because correlation should exist between each pair of adjacent zones or areas and not in separated groups. This means that what is required is to have each zone belonging to a different nest with each of its adjacent zones.

A variation of the NL model that allows for such correlation is the Cross Nested Logit model (Vovsha, 1997) where each alternative is allowed to belong to more than one nest considering a specific weight for each one.

Other method proposed in the literature to deal with the IIA property is the Logit Kernel or Mixed Logit model (Train, 2003). This method in addition to the Extreme Value error of the MNL model also includes other error that introduces correlation among the alternatives. The choice probabilities are obtained by numerical or Monte Carlo integration.

This model can take into account, for example, the specification of flexible disturbances or random parameters. For example, if the subset of alternatives $S$ are correlated, then it would be possible to specify an additional error $e$ with distribution $f(e)$

in the utility of the alternatives that belong to $S$ and calculate $P_n(i)$ , the probability that an individual $n$ chooses alternative $i$ , as in (2-5),where $\delta_i^S$ is a dummy that is equal to one if $i$ belongs to $S$ and zero otherwise. The most efficient way found so far to calculate this integral numerically is the *Halton Draws* procedure (Train, 2000).

**(2-5)**
$$P_n(i) = \int_{-\infty}^{+\infty} \frac{\exp\left(V_i^n + \delta_i^S e\right)}{\sum_{j \in A(n)} \exp\left(V_j^n + \delta_j^S e\right)} f(e) de$$

An example where the error specification issues in residential location modeling are considered through closed forms and flexible error structures is the work of Bhat and Guo (2004). The authors developed a model that they called a Mixed Spatially Correlated Logit (MSCL) model for location choices, that is basically a Cross Nested Logit where each zone belongs to a different nest with each of its adjacent zones and, additionally, some parameters of the model are considered random in the estimation using the Logit Kernel approach.

The motivation of the authors behind this approach is to add some flexibility to the disturbances terms of the RUM without increasing too much the dimension of the integration problem in the Logit Kernel problem. This is done by representing all this interaction using only one estimable parameter that captures the correlation effect between contiguous spatial units.

The authors applied their model to a case study with 236 observations, comparing a MNL model versus their MSCL model. Despite the fact that some improvements are observed, it can't be asserted from the paper if these improvements came from the fact that some parameters in the MSCL model were considered random and/or because of the inclusion of the spatial correction effect.

Moreover, the central motivation in developing the MSCL, that is, to have a computationally cheaper alternative to Logit Kernel to address zonal correlation, was not tested, so no conclusions can be obtained about its real advantages.

## 2.2.5 Relevant Attributes and Characteristics in Residential Location Choice Decision

In this section, a general review of the type of location attributes and household characteristics that have been found in the literature to be relevant in residential location choice is presented.

The relevant attributes in residential location choice models can be divided in two classes, the ones that belong to the specific dwelling unit and others related to the environment (neighborhood) where it is located. The effect of each of these variables will depend on the socio-economic characteristics of the household involved. The results reported in the literature for these parameters so far are diverse and sometimes contradictory or against intuition (Guo, 2004).

Examples of attributes that belong to the specific dwelling unit are the commuting time associated with this specific dwelling unit, its price, size, age, number of rooms, architectural style, and if it is a single house, a duplex, a condominium or an apartment. All these attributes, but the price, share the feature of been completely exogenous to the residential location process.

The attributes related to the environment (neighborhood) where the dwelling unit is located can also be divided in two classes. The first class of neighborhood attributes is formed by socio-racial-economic attributes such as, for example, race and ethnicity, income, education, average age and family status. These neighborhood attributes are grouped in a different class because it has been empirically found that households tend to cluster themselves in groups that share similar attributes. Thus, they are expected to enter the utility function as the difference between household value for the correspondent characteristics and neighborhood attribute average.

Explanations for this phenomenon go in the line of easiness for establishing social networks, the type of educational experiences that children can have as the result of their interaction with the neighbors (Bayer, et al, 2004), segregation in the real estate market (especially by racial origin) or self characteristics validation. These neighborhood attributes share also the feature of being endogenous to the residential location process.

This means that they are, at the same time, the result of the residential location process and a variable considered in the choice model.

The second class of attributes related to dwelling unit's environment corresponds to the ones that are not valued in terms of their difference with household characteristics. Examples of these types of attributes are accessibility and attractiveness, residential density, school quality, safety, street cleanliness or land-uses expressed as the percentage of parks, industrial or trash dump areas. All these neighborhood attributes, except perhaps the residential density, share also the feature of been exogenous, unless a very long-term analysis is considered.

Finally, household characteristics are not only relevant in the location choice process in terms of their difference with neighborhood attributes. It is also important to consider the differences in the perception of the attributes that different household type or members should have. For example, it can be expected that low-income households perceive housing price as more important than high-income households. It have also been found that women tend to have more sensitivity to commuting time relative to males what have been explained as a result of the dual role of women who have to be near enough home to take care of children and other households tasks (Sermons and Koppelman 2001). Also, it can be expected for example that families with children would have a greater preference for houses instead of apartments. Thus, it would be necessary to consider in modeling residential location parameters differentiated by household socio-economic characteristics such as the ones described here.

# 2.3 Land Use and Transportation

## 2.3.1 About the Magnitude of the Link between Land-use and Transportation

Pickrell (1999) show that between 1900 and 1960, the densities of new residential developments in a sample of 10 United States (US) large urban areas, have been constantly declining up to be one fifth of the original ones. This can be argued to be the effect of the development of transportation technologies in the form of private automobile, motorbus, transit and highways, together with the rise in family income.

However, the author claims that nowadays the impact of transportation on land-use in US cities should be considerably less important. This is because the marginal effect of new transportation improvements should be smaller due to the characteristics of the technological development curve and because new transportation investments would probably be more or less redundant to already existing ones, reducing then their net impact.

Other fact that reduces the flexibility in residential location nowadays is the grater share of multi-workers households, which makes a residential move substantially more difficult.

In the case of developing countries such a Mexico or Chile, it can be expected that the phenomenon observed in US cities in the first half of the twentieth century, that is, a simultaneous increase in income, motorization and a substantial increase in highway infrastructure, would have the mid-term effect of increasing the sprawl of the city, moving higher income households to the lower density suburbs. However, the second argument about the multi-workers household composition is equally valid for developing countries' cities so, it can be expected that the net effect would be lower than the effect in the US.

Pickrell (1999) also discusses the feasibility of developing a computational model to forecast the long term behavior of the entire system of land-use and transportation. The author states that experience have shown that the great amount of information needed and

the simplifying assumptions used in the current models have made them of little accuracy or robustness in the prediction of the behavior of the system so far.

### 2.3.2 Measuring the Relationship between Land-use and Transportation

When deciding where to live, households as a whole have to make a trade-off between dwelling unit's amenities and travel cost to actual and potential activities developed by each of its' members.

Because individuals has more than one transportation mode alternative to reach their destinations is necessary to consider, for each origin and destination, the EMU that the individual can obtain by using any of the available modes, defined as the Generalized Cost (GC). The GC that household $n$ that is located at $I$ perceives by traveling to place $J$ has a nice closed form (2-6) if it is assumed that the error terms of the utility function are *iid* Extreme Value (0,$\lambda$) (Ben-Akiva and Lerman, 1985), where $V_k^{I,J,n}$ is the utility of mode $k$ that is available to travel from $I$ to $J$ for household $n$.

**(2-6)**
$$GC_n(I,J)=\frac{1}{\lambda}\ln\left(\sum_{k\in C_n^{IJ}}\exp\left(\lambda V_k^{I,J,n}\right)\right)+\frac{\gamma}{\lambda}$$

In the same way, in the case where *accessibility* (ACC) to different activities is what matters, this value can be calculated as the EMU of visiting, from location $I$, each and every location $J$ to develop a specific activity, expression that again has a closed form (2-7) if it is assumed that errors are *iid* Extreme Value (0,$\mu$).

**(2-7)**
$$ACC_n(I)=\frac{1}{\mu}\ln\left(\sum_{J\in Dest}\exp\left(\mu V^{I,J,n}\right)\right)+\frac{\gamma}{\mu}$$

Martinez (1995) used the same approach to also define the concept of *attractiveness* (ATT), which is the expected maximum utility of being visited from other locations as follows, if the errors are assumed to be *iid* Extreme Value (0,$\eta$).

**(2-8)**
$$ATT_n(J)=\frac{1}{\eta}\ln\left(\sum_{I\in Orig}\exp\left(\eta V^{I,J,n}\right)\right)+\frac{\gamma}{\eta}$$

More elaborated expressions for measures of the relationship between transportation and land-use can be drawn by considering all the omitted attributes and perceptions of the households by latent variables (Walker, 2001). However, this approach requires more data and has not yet been applied to build satisfactorily models of residential choice.

# 2.4 Price Endogeneity in Econometric Models of Residential Location

In order to show the potential relevance of the issues addressed in this thesis, in this section is reviewed as set of studies in residential location where price endogeneity appears to be a relevant issue that was not considered. Afterwards are presented a residential location model where price endogeneity was treated but possibly ineffectively and one case where almost the same considerations taken into account in this thesis were used.

One work where endogeneity appears to be an issue and was not treated[1] is Sermons and Koppelman (2001). In this work, the authors modeled couples commute behavior in residential location and found a small (but significant) housing cost parameter, which could be the result of the price endogeneity problem studied in this thesis.

Another source of endogeneity in this model could be simultaneous determination of the work location of each member of the modeled couple, what could cause the commuting distance parameters to be inconsistently estimated. In this case, it would be useful to apply some variation of the methods developed by Blundell and Powel (2004) to correct for male-female working salary endogeneity.

Housing price was not significant in the model developed by Bhat and Guo (2004). They also found that school quality was not significant in the residential location process of single worker households, attributing the unintuitive result to the lack of detail in how the urban attributes were measured. A similar result is found by Waddell (1996) for single worker households and by Quigley (1976) and Levine (1998) for high income households. Waddell (1992) also found a positive, but small, elasticity of price for white workers.

It is highly unlikely that these results represent real behavioral characteristics of single workers, high income or white households. The real problem behind these findings should be either data problems or the price endogeneity bias analyzed in this thesis.

---

[1] Finding publications where counterintuitive results are reported is difficult. The credit for finding almost all the works cited in this section belongs to the deep literature review made by Guo (2004).

Unsuccessful attempts were done to obtain the databases used in the cited studies, and thus, it was not possible to test these hypotheses directly.

For the best of the knowledge of the author of this thesis only two works have considered the treatment of price endogeneity in discrete choice models or residential location. The first one is Bayer et al. (2004). In their research about racial segregation in the housing market, the authors did correct for price endogeneity using the method proposed by Berry, Levinsohn and Pakes (BLP, 1995), which is explained latter in this thesis. The author found considerably more negative parameters for the dwelling unit's price when comparing their model to a non-corrected one.

However, the big problem with this study is that the authors considered one alternative specific constant for each and every household, which implies that the estimated parameters are inconsistent due to the *incidental parameters* problem (Wooldridge, 2002). The problem is that, as the sample size increases, the size of each "market" stays fix (in one) making impossible to estimate their fix-effect parameter consistently. Furthermore, the inconsistent estimation of the fix-effect parameter contaminates the estimation of the other parameters in the model, which makes all estimated parameters inconsistent.

The other example where price endogeneity is treated in residential location modeling is Ferreira (2004), who was a research assistant in the Bayer et al.(2004) study. The author analyzed taxes incentives in residential location market. In his model the author considered a correction for price endogeneity based in the control-function method proposed by Petrin and Train (2004). The only argument against the procedure applied is that in this case the author used, as instruments for the price, the attributes of nearby dwelling units (following the idea of Bresnaham,1997) instead of the prices of nearby dwelling units themselves, as proposed by Hausman (1997), which will be claimed to be a better approach in section 3.1.3. The author found considerably more negative price parameters when the endogeneity correction was applied.

# Chapter 3

# Endogeneity in Econometric Models

The objective of this chapter is to describe all the relevant issues in understanding the causes, consequences, and treatment of endogeneity in econometric models in light of residential location modeling.

With this purpose, will be first presented a review of what is endogeneity, when it is likely to occur, and how it can be solved in the case of linear econometric models. Using this as a background, are then analyzed in detail the available methods to treat endogeneity in discrete choice models. Finally a review of relevant applications of these methods in the literature is presented and discussed.

# 3.1 Endogeneity in Linear Models

## 3.1.1 When Endogeneity is Likely to Occur in Linear Models

Consider the linear regression model (3-1), where $Y$ is a vector defined as the dependent variable to be explained by the independent variables contained in the matrix $X$, $\beta$ is a vector of parameters, and $\varepsilon$ is a vector of errors or values that makes expression (3-1) an identity.

**(3-1)** $$Y = X\beta + \varepsilon$$

The estimator of the vector of parameters $\beta$ that minimizes the square of the modeling error (Greene, 2003), which is commonly identified as the ordinary least squares (OLS) estimator, is expression (3-2).

**(3-2)** $$Min_{\beta}\varepsilon^T\varepsilon = Min_{\beta}(Y - X\beta)^T(Y - X\beta) \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T Y$$

A desired small sample property for the estimator $\hat{\beta}$ is to be unbiased. This means that, at least on *average*, $\hat{\beta}$ should be equal to the true parameter $\beta$ or, formally, that its expected value, conditional on $X$, should be equal to $\beta$. In (3-3) the conditions required for $\hat{\beta}$ to be unbiased are analyzed.

**(3-3)**
$$\begin{aligned} E(\hat{\beta}/X) &= (X^T X)^{-1} X^T E[Y/X] = (X^T X)^{-1} X^T (X\beta + E[\varepsilon/X]) \\ &= \beta + \underbrace{(X^T X)^{-1} X^T E[\varepsilon/X]}_{BIAS} \end{aligned}$$

Thus, a necessary condition for $\hat{\beta}$ to be unbiased is to have $\varepsilon$ independent of $X$ or, equivalently, the expectation of $\varepsilon$ given $X$ equals zero ($E[\varepsilon/X] = 0$). Having $E[\varepsilon/X] \neq 0$ is commonly called non-orthogonality and, for the purpose of this thesis, it will be named also as endogeneity.

## Omitted Attributes

One example of when endogeneity occurs is the case when the researcher omits an attribute that is relevant in the true model. To demonstrate this, assume that in the true model an additional variable $U$ is relevant, so that the model can be expressed as shown in (3-4).

**(3-4)** $$Y = X\beta + U\alpha + \varepsilon\,;\, E(\varepsilon / XU) = 0\,;\, [XU] \text{full rank.}$$

Now, assume that the matrix of variables $U$ is unobserved by the researcher but that the true model is still (3-4). Then, if the expectation of the OLS estimator of the parameter $\beta$ is calculated without considering the variable $U$, expression (3-5) will hold.

**(3-5)** $$\begin{aligned} E(\hat{\beta}/\tilde{X}) &= (X^T X)^{-1} X^T E[Y/XU] = (X^T X)^{-1} X^T (X\beta + U\alpha + E[\varepsilon/XU]) \\ &= (X^T X)^{-1}[X^T X\beta + X^T U\alpha] \\ &= \beta + \underbrace{(X^T X)^{-1} X^T U\alpha}_{BIAS} \end{aligned}$$

Hence, $\hat{\beta}_{OLS}$ will be unbiased if and only if the observed variables $X$ are orthogonal with the unobserved ones $U$ or, if the vector of parameters $\alpha$ is zero. Otherwise, $\hat{\beta}_{OLS}$ will be upward biased if $X$ and $U$ are positively (negatively) correlated, and $\alpha$ is positive (negative) and downward biased otherwise.

## Errors in Variables

Another example of when endogeneity can occur is the case when the independent variables are measured with error. Consider the true model (3-6), where a single variable $x_t$* determines the dependent variable $y_t$, and where, instead of $x_t$* , $x_t$ is observed, which is a *noisy* version of $x_t$* as is shown in (3-7).

**(3-6)** $$y_t = \alpha + x_t{*}\beta + \varepsilon_{1t} \qquad E[\varepsilon_{1t} / x_t{*}] = 0$$

**(3-7)** $$x_t = x_t{*} + \varepsilon_{2t}$$

**(3-8)** $$y_t = \alpha + (x_t - \varepsilon_{2t})\beta + \varepsilon_{1t} = \alpha + x_t\beta + \underbrace{(\varepsilon_{1t} - \varepsilon_{2t}\beta)}_{\varepsilon_t}$$

In this case, as is shown in expression (3-8), the independent variable $x_t$ will be correlated with $\varepsilon_{2t}$ because of (3-7) and then with the error term $\varepsilon_t$ as long as $\varepsilon_{2t}\beta$ is different from zero, that is, if errors in variables exist and $\beta$ is different from zero. If this problem exists, $\hat{\beta}_{OLS}$ will be biased toward zero in what is known as the "attenuation bias" or the "iron law of econometrics" (Greene, 2003).

## Simultaneous Determination

Another example where the non-orthogonality problem arises is the case where the dependent and at least one of the independent variables are simultaneously or endogenously determined through different equations. This case can commonly occur in a model of market equilibrium between demand and supply where the selling price $p_t$ determines the quantity produced $q_t$ in the supply equation and the quantity produced determines the price in the demand equation, which depends also on another variable $I_t$ as is shown in expression (3-9).

**(3-9)**

$$\begin{aligned} &\textit{Supply} && q_t = \beta_{11} p_t + \varepsilon_{1t} \\ &\textit{Demand} && p_t = \beta_{21} q_t + \beta_{22} I_t + \varepsilon_{2t} \end{aligned}$$

The fact that the dependent variable is correlated with the error term, both in the supply and the demand equations, follows directly from an argument equivalent to the problem of errors in variables shown in (3-6)-(3-8) if one equation is replaced on the other, taking the dependent variable of each one as the true one.

As in the errors in variables case, if each equation of model (3-9) is estimated independently, a bias towards zero will occur. Instead, if the simultaneity is taken into account for both equations, it will be possible to estimate the parameter of the price in the supply equation consistently, but not to identify the parameters of the demand one.

To understand why the identification and consistency results described above occur, recall first that what can be observed are equilibrium points of the price and the quantity traded in the market (Figure 3-1). In the case described in (3-9) the supply equation is

invariant and the different equilibrium points are explained by a set of demand curves that is **not unique** because of variations of $I_t$.

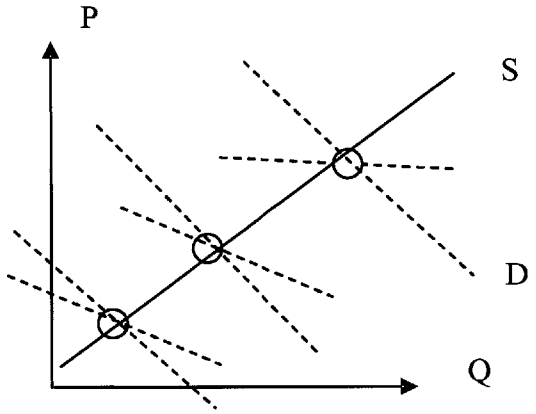



**Figure 3-1 Identification of Simultaneous Equations**

Then, the demand equation remained unidentified but, at the same time, its variation justified in (3-9) by the factor $I_t$, allows to draw or identify the supply equation.

This identification result can be shown through the derivation of the *reduced form* (Greene, 2003) of problem (3-9) by replacing one equation on the other and solving all the dependent variables as a function of the exogenous variable $I_t$.

**(3-10)**

$$q_t = \alpha_{11} I_t + \xi_{1t} = \frac{\beta_{11}\beta_{22}}{1-\beta_{11}\beta_{21}} I_t + \frac{\beta_{11}\varepsilon_{2t} + \varepsilon_{1t}}{1-\beta_{11}\beta_{21}} \quad \textit{Supply}$$

$$p_t = \alpha_{21} I_t + \xi_{1t} = \frac{\beta_{22}}{1-\beta_{11}\beta_{21}} I_t + \frac{\beta_{21}\varepsilon_{1t} + \varepsilon_{2t}}{1-\beta_{11}\beta_{21}} \quad \textit{Demand}$$

Thus, if equations (3-10) are estimated by OLS, the parameter of the supply equation in (3-9) can be identified as $\beta_{11} = \alpha_{11}/\alpha_{21}$. However, it is not possible to derive demand parameter $\beta_{21}$ as a function of $\alpha_{11}$ and $\alpha_{21}$. It can be shown that, under some assumptions, this procedure to identify the supply equation is fully equivalent to one where the price is regressed (OLS) on $I_t$, and then estimated prices of this model are used, instead of the actual prices, to estimate the supply equation. This last method is called *Instrumental Variables* (IV) and is described in more detail in the following section.

## 3.1.2 Instrumental Variables, the Method to Treat Endogeneity in Linear Models

The method to address the problem of non-orthogonality or endogeneity in linear models is called the *instrumental variables* (IV) procedure (Greene, 2003). It corresponds to project the dependent variable that is correlated with the error term onto a space that is orthogonal to the error's space, which is defined by another variable called instrument. Then, the estimated (instrumented) dependent variable will not suffer the non-orthogonality problem and can be used to estimate the model consistently.

Consider again the model (3-1) where the endogeneity problem exists. Consider also a matrix of variables $Z$ with $rank(Z) = L > rank(X) = K$; and a matrix $W_{(TxK)} = Z_{(TxL)}\hat{A}_{(LxK)}$, where $\hat{A}$ is a function of the data $X$. Thus, the following estimator of $\beta$ can be defined.

**(3-11)**
$$\hat{\beta}_{IV} = \left(W^T X\right)^{-1} W^T Y$$

The conditions under which the estimator (3-11) is unbiased are analyzed in (3-12).

**(3-12)**
$$\begin{aligned} E\left[\hat{\beta}_{IV}/X\right] &= E\left[\left(W^T X\right)^{-1} W^T Y/X\right] = E\left[\left(W^T X\right)^{-1} W^T (X\beta+\varepsilon)/X\right] \\ &= \beta + \left(W^T X\right)^{-1} E\left[W^T \varepsilon/X\right] = \beta + \left(W^T X\right)^{-1} E\left[\left(Z\hat{A}\right)^T \varepsilon/X\right] \\ &= \beta + \underbrace{\left(\hat{A}^T Z^T X\right)^{-1} \hat{A}^T E\left[Z^T \varepsilon/X\right]}_{BIAS} \end{aligned}$$

Thus, in order to have an unbiased[2] estimator of $\beta$ using (3-11) it would be required that $Z$ were, at the same time, orthogonal to the error term $E\left[Z^T \varepsilon\right] = 0$ and correlated with $X$ $E\left[Z^T X\right] \neq 0$.

It can be shown that the value of $\hat{A}$ that minimizes the variance of $\hat{\beta}_{IV}$ corresponds to $\hat{A} = \hat{V}^{-1} Z^T X$, where $\hat{V}$ is some consistent estimator of the variance of $Z^T \varepsilon$. In the case where $\varepsilon$ is spherical, that is $\text{var}(\varepsilon) = \sigma^2 I$, the estimator of the inverse of the variance

---

[2] To have an equivalent consistency result for large sample, the following weaker conditions are needed: $p\lim Z^T \varepsilon/T = 0$ *and* $p\lim Z^T X/T \neq 0$.

will be $\hat{V}^{-1} = \frac{1}{\sigma^2}\left(Z^T Z\right)^{-1}$ and then the optimal instrumental variables estimator will be (3-13).

$$\textbf{(3-13)} \qquad \hat{\beta}_{OIV} = \left(X^T Z\left(Z^T Z\right)^{-1} Z^T X\right)^{-1} X^T Z\left(Z^T Z\right)^{-1} Z^T Y \quad if \quad \text{var}(\varepsilon) = \sigma^2 I$$

Thus, under sphericality, the optimal instrumental variables estimator corresponds to do OLS of $X$ on $Z$ and then use the estimated $X$s to run OLS of $Y$ as is shown in (3-14). This estimator is also known as *two stages least squares* (2SLS).

$$\begin{aligned}
&X = Z\lambda + \delta \\
&\Rightarrow \hat{\lambda}_{STEP_I} = \left(Z^T Z\right)^{-1} Z^T X \quad \Rightarrow \quad \hat{X}_{STEP_I} = Z\left(Z^T Z\right)^{-1} Z^T X \\
&Y = \hat{X}\chi + \xi \\
&\Rightarrow \hat{\beta}_{2SLS} = \left(\hat{X}^T \hat{X}\right)^{-1} \hat{X}^T Y \\
&\Rightarrow \hat{\beta}_{2SLS} = \left(\left(Z\left(Z^T Z\right)^{-1} Z^T X\right)^T Z\left(Z^T Z\right)^{-1} Z^T X\right)^{-1} \left(Z\left(Z^T Z\right)^{-1} Z^T X\right)^T Y \\
&\Rightarrow \hat{\beta}_{2SLS} = \left(X^T Z\left(Z^T Z\right)^{-1} Z^T X\right)^{-1} X^T Z\left(Z^T Z\right)^{-1} Z^T Y = \hat{\beta}_{OIV}
\end{aligned} \qquad \textbf{(3-14)}$$

Hausman (1978) noted that and equivalent result is obtained if, instead of using the instrumented variables in the second stage, the model is ran with the original variables, but adding the fitted errors of the first stage as additional variables. This follows directly by recalling that, by construction, $X = \hat{X} + \hat{\delta}$ and applying this to the second stage model of the 2SLS procedure.

$$\textbf{(3-15)} \qquad Y = \hat{X}\chi + \xi \quad = \left(X - \hat{\delta}\right)\chi + \xi = X\chi - \hat{\delta}\chi + \xi$$

Thus, to have a completely equivalent result to what is obtained with the 2SLS procedure by adding the fitted errors as an additional variable, it should be considered that the parameter of the variable $X$ and the parameter of the fitted error $\hat{\delta}$ are equal in absolute value but with different sign.

As it will be noted later, this result is the basis for the method to correct for endogeneity in discrete choice models known as the control-function, the one that is claimed to be the most suitable to treat this problem in residential location modeling.

With this background, it is possible now to give an alternative explanation for the identification of the supply equation in (3-9) based on Hausman (1983). In that case the exogenous variable $I_t$ was used as an instrument for the price $p_t$ in the supply equation. This was possible because $I_t$ is correlated with $p_t$ (as long as $\beta_{22}$ is different to zero) and it is assumed to be uncorrelated with the error term of the supply curve.

Finally, a comment has to be made about what is called the *weak instruments* problem. Recall that the unbiased (or consistent) estimation of the parameters under endogeneity using the instrumental variables method relies in the assumptions that the instruments are correlated with the endogenous variables and, at the same time, they are uncorrelated with the error term. Sometimes it could be easy to have variables that are not correlated with the error term (such as, rainfall variation in China with residential location in Cambridge, MA), but having also this additional variables correlated with the endogenous one would be sometimes difficult.

Hahn and Hausman (2003) showed that in the presence of weak instruments the bias of the 2SLS estimator could be even greater than the one of the OLS estimator, that is, the cure is worst than the illness. Thus, the definition of appropriate instruments for modeling residential location is an important issue that will have to be addressed.

### 3.1.3 About Finding Appropriate Instruments

The selection of an appropriate set of instruments both in linear and non linear residential location modeling is a relevant and controversial issue itself. An example of this controversy can be seen in the series of papers between Hausman (1997) and Bresnaham (1997) where the authors discussed about the usage of instruments bases on the prices or on the attribute of the other markets, alternatives that are explained in the following paragraphs.

Assuming that the unobserved attributes affecting prices are zone (market) specific, one alternative to define instruments for dwelling unit price is to argue that observed average prices of the same product (dwelling unit type in the residential location market) in other zones are correlated with analysis-zone's product price and uncorrected with its

respective unobservable attributes. Thus, combinations of observed prices from other zones can be used as instruments (Nevo, 2001 and Hausman, 1997).

Another alternative is to use, as BLP (1995) did it, attributes of dwelling units from other zones as instruments. Yet, it is not clear why these attributes from other zones should be correlated with the attributes of the analysis-zone's dwelling unit, so the use of this approach will possibly lead to a weak instruments problem (Hahn and Hausman, 2003)

## 3.2 Endogeneity in Discrete Choice Models

In discrete choice models of residential location endogeneity is expected to occur as the result of the omission of attributes that are relevant to the decision maker but not observed by the researcher. Think for example in a group of houses with identical architecture but where one of them posses an especially beautiful and unrepeatable view of a nearby pond. Because of this unique positive attribute the price of that specific house would be larger, effect that cannot be explained by the researcher in his model if he is unable to obtain this very specific information.

The unexplained (for the researcher) fact that some households do choose expensive dwelling units, will be interpreted by the model as a low sensitivity to price, which leads to upward biased price parameters.

Formally, the problem of endogeneity can be stated as follows. Consider the utility that household $n$ obtains from dwelling unit $j$ decomposed into an observed $V(.)$ and an unobserved part $e_{jn}$

**(3-16)** $$U_{jn} = V(p_{jn}, x_{jn}, s_n) + e_{jn},$$

were $s_n$ denotes the observed characteristics of household $n$, $p_{jn}$ and $x_{jn}$ are the price and observed attributes of dwelling unit $j$ perceived by household $n$. Problems arises when $e_{jn}$ is correlated with $p_{jn}$ or $x_{jn}$ because discrete choice models assume that this correlation is zero or constant, which excludes the possibility of correlation with price (Train, 2003).

It was shown in (3-5) for the linear model that if the correlation between the observed and unobserved attributes has the same (different) sign of the parameter of the unobserved attribute, the estimated parameters of the observed attributes will be upward (downward) biased. The hypothesis is that the same result is valid for discrete choice models, but a formal proof of it is left for further research.

That would imply that an upward bias for the price parameter (a negative parameter that becomes more positive) will usually occur in the case of residential location. This follows from the fact that if the omitted quality attribute is measured in a positive way (for example, as more clean or less crime), it will be positively correlated with the price,

and its parameter will also be positive. Conversely, if the omitted quality attribute is measured in a negative way, it will be negatively correlated with the price, and its parameter will also be negative.

### 3.2.1 Methods to Treat Endogeneity in Discrete Choice Models

The treatment of endogeneity (or non-orthogonality) in non-linear models of discrete choice, such as the ones needed in residential location modeling, cannot be pursued by directly using instrumental variables, the standard method for linear models. As it was mentioned before, this is an ongoing area of research in econometrics where, at least, four proposed methods can be identified in the literature. In this section, these methods are described and analyzed and in Chapter 4 they are evaluated under the framework of residential location modeling through four Monte Carlo experiments.

#### BLP method

The first method to deal with endogeneity in discrete choice models was proposed by BLP (1995). The authors developed and applied an approach using product-market fix effects, which provides consistent estimation under endogeneity or omitted product attributes, with the assumption that endogeneity occurs at the market level.

Within this setting, Petrin and Train (2004) formulated this method by considering that the utility $U_{jn}$ that individual $n$, who is located in market $m$, perceives from alternative $j$, is decomposed into a part that is the same for all customers in a market, $\delta_{jm}$, plus an observed $V_{jnm}$ and an unobserved error part $\varepsilon_{jn}$ that is assumed to be *iid* Extreme Value.

(3-17) $$U_{jn} = \underbrace{\alpha p_{jm} + h(x_{jm}) + \xi_{jm}}_{\delta_{jm}} + V_{jnm}(p_{jm}, x_{jm}, s_n) + \varepsilon_{jn}$$

In this model it is assumed that prices $p_{jm}$ and other attributes $x_{jm}$ are the same within each market. The fix effect term $\delta_{jm}$ incorporates the average value of the omitted attributes $\xi_{jm}$ along with the other components of the utility that vary between markets

and that are assumed to depend linearly upon price and also other attributes through the parametric function *h(.)*.

As stated before, the assumption is that the endogeneity problem occurs at the market level, that is, that $p_{jm}$ is correlated only with $\xi_{jm}$ and not with $\varepsilon_{jn}$. Under this assumption, the procedure to estimate the model consists of three steps.

The **first step** is the estimation of a discrete choice model with a full set of *Alternative Specific Constants* (ASC) for every market. These ASC's represent the fix effects, $\delta_{mj}$, and absorb the endogeneity problem making all the estimated parameters consistent. The original method proposed by BLP (1995) considered an estimation using the *Generalized Method of Moments* (GMM) but Petrin and Train (2004) used *Maximum Likelihood* (ML), which is the optimal GMM if the error distribution is known. Both BLP (1995) and Petrin and Train (2004) considered $\varepsilon_{nj}$ to be *iid* Extreme Value.

The estimation of the model described above would imply a serious computational burden because of the considerable number of ASC's that have to be estimated, the product of the number of markets and the number of alternatives minus one (for identification).

However, BLP (1995) provided a method to solve this problem. Based on Berry (1994), the authors demonstrated that the $\delta_{jm}$ can be calculated iteratively as a contraction using expression (3-18), where $S_{jm}$ is the sample share of product $j$ in market $m$ and $F_{jm}$ is its forecasted share.

**(3-18)**
$$\delta_{jm}^{t+1} = \delta_{jm}^{t} + \ln\left(S_{jm}\right) - \ln\left(F_{jm}^{t}\right)$$

This iterative process is performed for each trial value of the other parameters of the model within the optimization process associated with the GMM or the ML estimation methods.

**Steps two and three** correspond to the application of 2SLS to the expression of the fix effect portion of (3-17) that are divided in two for the sake of clarity. Thus, step two is OLS regression (3-19) of the observed prices on exogenous instruments (or a function of them) and then step three corresponds to use the fitted values of prices to estimate the

model (3-20), using the previously estimated fix effects parameters as the dependent variable.

**(3-19)** $$p_{jm} = Z_m + \mu_{jm} \quad \Rightarrow \hat{p}_{jm} = Z_m \left(Z_m^T Z_m\right)^{-1} Z_m^T p_{jm}$$

**(3-20)** $$\hat{\delta}_{jm} = \alpha \hat{p}_{jm} + h\left(x_{jm}\right) + \xi_{jm}$$

As can be seen, the goal of BLP (1995) is to move the endogeneity problem from the non-linear model to a linear setting were the standard procedures of instrumental variables can be applied.

The positive side of the application of the product-market fix effects approach in residential location modeling is that, as long as it can be accepted that endogeneity occurs in disjointed markets, no further assumptions about the error structure are need to be made.

However, in spite of the fact that some geographical market segmentation exists, it is not clear how to define the boundaries of the markets that share unobserved characteristics in the residential location market. It can be equivalently argued that all dwelling units belong to the same market (because people can freely move across the city) and also that each dwelling unit belongs to its own market in the sense of having unique unobserved attributes. It can also occur that the endogeneity problem occur in overlapped markets. In all this cases, the assumption required by the BLP method will not hold.

## Control-function method

The second available method is the control-function method (Heckman, 1978 and Hausman, 1978) applied in a discrete choice environment (Petrin and Train, 2004 and Blundell and Powell, 2004). The central idea behind this method is to make use of the information that observed prices have about omitted attributes.

Consider utility function (3-16) and a set of appropriate instruments $Z$ that are correlated with the price but not with the error term $e_{jn}$. The price can always be written as the sum its expectation conditional on the instruments $Z$ and an error term $\mu_{jn}$.

**(3-21)** $$p_{jn} = E\lfloor p_{jn}/Z \rfloor + \mu_{jn}$$

Then, if expression (3-21) is estimated by OLS, the fitted prices will not be correlated with the unobserved part of the utility (3-16), $e_{jn}$. This follows from the fact that by doing OLS the fitted prices correspond to the projection of actual prices onto the space formed by the instruments $Z$, which are, by assumption, orthogonal to $e_{jn}$. For the same reason, the fitted errors of (3-21) will be orthogonal to the fitted prices and, because they correspond to the difference between fitted and observed prices, they will contain the part of actual prices that is correlated with the error $e_{jn}$.

Therefore, if $\hat{\mu}$ are used to estimate the conditional expectation of $e_{jn}$ (3-22), the fitted residual of this model $\hat{\varepsilon}_{jn}$ will be orthogonal to $\hat{\mu}$ and, in consequence, uncorrelated to the prices.

**(3-22)** $$e_{jn} = E\lfloor e_{jn}/\hat{\mu} \rfloor + \varepsilon_{jn} = f_{jn}(\hat{\mu}) + \hat{\varepsilon}_{jn},$$

Thus, if the control-function $f_{jn}(\hat{\mu})$, which is the OLS estimation of the expectation of $e_{jn}$ conditional on $\hat{\mu}$, is considered as an additional linear variable in the utility function (3-16), the endogeneity problem would be solved.

Petrin and Train (2004) state that if it is assumed that the covariance matrices of $e$ and $\hat{\mu}$ are diagonal, the control-function that enters the utility is just proportional to the price residual of the respective alternative and individual, or the "own error"

**(3-23)** $$f_{jn}(\hat{\mu}) = E\lfloor e_{jn}/\hat{\mu} \rfloor = \lambda \hat{\mu}_{jn},$$

where the parameter $\lambda$ is function of the covariance of $e$ and $\hat{\mu}$ and the variance of $\hat{\mu}$.

Other interpretation of the control-function method goes in the line of being a generalization of the traditional instrumental variables method, as it was described in (3-15) and is discussed in the following sub-section.

The method is applied in two straight forward steps. The **first step** of the control-function method corresponds to the OLS regression (3-24) of the price on exogenous instruments and the calculation of the fitted error $\hat{\mu}_{jn}$. The difference of this case with the

BLP method is that the prices are allowed to vary by individual observations and not only by market.

**(3-24)** $$p_{jn} = Z_{jn} + \mu_{jn} \quad \Rightarrow \hat{\mu}_{jn} = \left[Z - Z\left(Z^T Z\right)^{-1} Z^T p\right]_{jn}$$

The **second step** corresponds to the estimation of a discrete choice model considering, as an additional variable, a function of the fitted errors.

**(3-25)** $$U_{jn} = V\left(p_{jn}, x_{jn}, s_{n},\right) + f\left(\hat{\mu}_{jn}\right) + \varepsilon_{jn}$$

An equivalent procedure could be pursued by estimating steps one and two simultaneously using the latent-variable approach (Walker, 2001). This would lead to an increase in efficiency. The study of this alternative method is left for further research.

The control-function method can be applied in cases where the fix effects approach is not feasible, for example, when price varies endogenously over every observation instead than over groups. This makes the control-function method more promising for residential location modeling, where unique unobserved attributes of each dwelling-unit are expected to produce endogeneity at an individual level.

However, to apply this approach is necessary to assume the error structure of the system of equations in order to determine the theoretically correct function of the residuals $f_{jn}(\hat{\mu})$ to include as an extra variable.

This problem could reduce dramatically the robustness of the model, but it can apparently be avoided by using non-parametric techniques to recover the error structure from the sample (Blundell and Powell, 2004). The study of these correction methods is analyzed in Chapter 4 obtaining however non-conclusive results.

## Traditional Instrumented Prices

The next method is a particular case of the control-function method. This corresponds to use instrumented prices to estimate de discrete choice model instead of adding the fitted errors as an additional variable.

Following what was shown in (3-15), this procedure can be viewed as a special case of the control-function method where two conditions have to hold. First, it is necessary to

assume that the variance covariance matrices of the utility error and the price equation are diagonal and thus by (3-23) the control-function is proportional to the own error of the price equation. Second it has to be assumed also that the parameters of the control-function and of the price are equal.

Other limitation of this method compared with the control-function one is that it would not allow the application of the non-parametric methods proposed by Blundell and Powel (2004) to avoid the misspecification problems associated with the assumption of a specific error structure.

## Unobservable Instruments

The fourth method is known as unobservable instruments and it was proposed by Matzkin (2004). This method is based on the inclusion of an extra endogenous variable in the model which is correlated with the original endogenous one (the dwelling unit price, in the residential location modeling case) only through exogenous perturbations.

Formally, this mean that if $X$ is the endogenous variable that has to be treated, what is required is to find an auxiliary variable $X^*$ such that $X^*$ is an "exogenous perturbation" of $X$ in the sense that, for some function $s$ and some unobservable variable $\eta$

**(3-26)** $$X = s(X^*, \eta),$$

where $\eta$ satisfies some type of independence with the error term of the original function, that is, it is exogenous.

If a variable like $X^*$ can be found or built, Matzkin (2004) demonstrated that consistent estimates are obtained if $X^*$ is included as an extra variable in the model. The author probed that this procedure can be applied even for non-linear functions using non-parametric and semi-parametric methods.

One practical application of this method is the study of Train and Winston (2004). The authors modeled the choice of automobile brands considering the retained (or re-selling) price of the automobile as an additional variable in the model arguing that, despite this additional variable is correlated with the omitted attributes, it is correlated with the price only through characteristics of the vehicle, which they claimed to be

exogenous. Thus, the retained price apparently complies with the characteristics required by the auxiliary variable in the Matzkin method.

However, it is not clear why the characteristics through which price and retained price are correlated should be exogenous because it is fairly possible that some of them would be also unobserved and then endogenous.

The principal problem in using this method in residential location modeling is that it is not yet clear which kind of variable should comply with this characteristics required by the Matzkin method. It can be argued that assessment price or characteristics of other dwelling units in the neighborhood are potential candidates as long as it can be accepted that these variables are exogenous perturbations of the price. It can be even considered that synthetic instruments for price could be built just by adding an exogenous perturbation to actual prices. The study of these alternatives is left for further research.

Finally, it is possible to think in making a link between the control-function method and the usage of Matzkin type instruments. By construction (3-24), the fitted errors of the price equation are correlated with the error term, that is, they are endogenous. At the same time, they are correlated with the price only through the instruments, which are exogenous by definition. Hence, it can be argued that the control-function method is actually a method to construct Matzkin auxiliary variables. A formal proof and further analysis of this relationship is left for future research.

### 3.2.2 Applications of the Correction Methods Found in the Literature

Beyond the seminal study of BLP (1995), the research of Petrin and Train (2004) discussed earlier and the studies in residential location of Bayer et al. (2004) and Ferreira (2004) analyzed in the previous chapter, a set of relevant studies where variations of the methods described in this chapter are presented in this section.

Trajtenberg (1989) studied the market of Computed Tomography Scanners using a Nested Logit approach with the aim of looking for a way to measure the benefits of these product innovations. As the firsts models developed by the author turned out to have unintuitive parameter values due to, possibly, endogeneity, the author used a correction

method based on the inclusion of the residuals from a price equation in the utility function.

However, the author did not also enter the actual price in this model, and thus, he didn't estimate the price coefficient along with the rest of the model estimates, but used outside information to built it.

Villas-Boas and Winer (1999) analyzed endogeneity in brand choice models of Yogurt and Ketchup using scanner data. The authors tested for the presence of price endogeneity by applying a variation of the control-function method that corresponds to consider specific error structures, first Extreme Value and then Normal, between the price equation error and the utility error.

In both cases the authors considered a linear function of the own error as the control-function concluding that, not accounting for these effects, can result in a substantial bias in parameters estimates.

Blundell and Powell (2004) developed and implemented semi-parametric methods for estimating binary choice models with continuous endogenous regressors. The authors used a control-function method to account for endogeneity in binary choice models. In cases where the correlation matrices between the price equation and the utility error is not diagonal, the authors showed that, under some assumptions, the control-function method can be adapted to work using semi-parametric methods.

The model was applied to investigate the importance of correcting for the endogeneity of partner's income in a labor market participation model for a sample of married British men. The paper's results show a strong effect of correcting for endogeneity in this example and indicate that adjusting for endogeneity without the non-parametric corrections proposed, can give a misleading picture of the impact in the participation of an exogenous change in partner's income.

# Chapter 4

# Evaluation of Methods to Address Endogeneity in Residential Location Models

This chapter presents a theoretical, qualitative and a quantitative comparison of different methods to treat endogeneity in discrete choice models found in the literature, in light of residential location models.

In the first section, the methods are compared theoretically and qualitatively in terms of their robustness to the most plausible error structures that can be expected in residential location models. With the same objective, in the second section the methods are empirically evaluated under different settings of the error structure and quality of instruments through four Monte Carlo experiments.

# 4.1 About the Error Structure in Residential Location Models

Because the goods in the residential market have almost unrepeatable characteristics, it can be argued that the endogeneity problem in this market occurs at the individual level caused by the omission of attributes that are specific to each alternative. However, it can also be expected that, because nearby areas share unobserved attributes, some kind of zonal or market endogeneity should also exist.

If the endogeneity problem does occur exclusively at an individual level and instruments can be built at that level also, the available methods to correct for endogeneity will then be the control-function method, the traditional instruments method and the Matzkin method. In this case the BLP method will inconsistently estimate the parameters of the model because it needs to consider one ASC for each dwelling unit.

Among the possible methods, the most suitable to address the problem is the control-function method because it can efficiently handle individual endogeneity as long as good instruments for dwelling unit's price can be obtained. An instrument for the price in this case can be constructed as the average of the price of other the dwelling units in the neighborhood. This follows from the fact that this variable will be correlated with the original dwelling unit price through unobserved-shared-neighborhood attributes and, at the same time, uncorrelated with the error term, because this variable does not share the unobserved attributes that are specific to the original dwelling unit.

The traditional instruments approach corresponds to a special case of the control-function method where, as it was stated in 3.1.2, it is not only that it's simplest form is considered but also that, within it, the parameter of the price and the parameter of the control-function are forced to be equal.

The Matzkin method has the problem that it would be difficult to identify what kind of auxiliary variables or "exogenous perturbations" can be used in the residential location market to apply this method. Also, the application of this method on its more sophisticated variations would be computationally more complex than the other approaches.

If the endogeneity problem does occur exclusively in a zonal or market basis, the method to apply is the BLP method. In that case no further assumptions on the error structure have to be done to estimate the parameters of the model consistently.

However, even if it is accepted the assumption that endogeneity occurs in a zonal basis, in the residential location market it can be expected that this effect occurs in overlapped or ambiguously defined geographical areas and not in disjointed areas. This precludes for the use of the BLP method because, in this case, the endogeneity problem will occur in a combination of zonal and individual levels or it will not be possible to identify the market boundaries to apply it.

Thus, this thesis proposes that, even though the control-function method requires specific assumptions about the error structure, this method is the most promising to address endogeneity in residential location choice models principally because it can easily handle individual endogeneity case when where its better competitor, the BLP method, is inconsistent.

An additional argument in its favor is that the computational burden of the control-function method is considerably less than the one of the BLP method, and it can be efficiently applied using available estimation software. This argument is valid not only for the simpler version of the control-function method, but also for more elaborated ones that consider the estimation using non-parametric of feasible generalized least squares methods that are already coded in conventional estimation software.

If the endogeneity problem does occur on a market-differentiated level and the control-function method is applied, Petrin and Train (2004) concluded that a misspecification $\eta_{jn}$ will occur, which corresponds to the difference between the zonal error component $\xi_{jm}$ and the control-function $f_{jn}(\mu,\lambda)$.

**(4-1)**
$$\eta_{jn} = \left(\xi_{jm} - f_{jn}(\mu,\lambda)\right)$$

To solve this misspecification problem, the authors first proposed to estimate the parameter of the control-function $\hat{\mu}$ as random. The second alternative proposed by Petrin and Train (2004) to handle this misspecification corresponds to the inclusion in the control-function not only the own fitted errors but also the sum of the errors of other

alternatives of the same type in the area, and the sum of the errors of other alternatives of other types in the area. Formally, the specification of the control-function in this case would correspond to (4-2), where *J(j)* is the set of alternatives of the same type of *j*.

**(4-2)** $$f_{jn}(\mu, \lambda_0, \lambda_1, \lambda_2) = \lambda_0 \mu_{jn} + \lambda_1 \left( \sum_{k \neq j, k \in J(j)} \mu_{kn} \right) + \lambda_2 \left( \sum_{k \notin J(j)} \mu_{kn} \right)$$

In the following sections the effectiveness of these correction methods is analyzed through four Monte Carlo experiments.

# 4.2 Comparison Using Monte Carlo Experiments

The objective of this section is to analyze the robustness of the different available methods to treat endogeneity in residential location models using a set of Monte Carlo experiments.

For each Monte Carlo experiment a synthetic population of 2000 households was generated, where each household faces a choice among three alternatives residential locations. The residential location choice behavior is assumed to be governed by a Random Utility Model (Ben-Akiva and Lerman, 1985), where each household ($n$) maximizes its *Utility* ($U_{in}$), which is assumed to be a linear function ($V_{in}$) of the attributes ($a$, $b$, $c$, $d$ and the price $p$) of each available dwelling unit alternative ($i$), with specific parameters and an error term ($\varepsilon_{in}$), as in (4-3).

**(4-3)** $$U_{in} = V_{in} + \varepsilon_{in} = 10 * a_{in} + 10 * b_{in} + 10 * c_{in} + 10 * d_{in} - 10 * p_{in} + \varepsilon_{in}$$

The error term of this utility function $\varepsilon_{in}$ was constructed to be *iid* (over alternatives and households) Extreme Value (0,1).

Additionally, it was assumed that dwelling units' prices are determined by the linear function (4-4) of attributes $c$, $d$ and $z$, with parameters $\alpha$, which are specific to each Monte Carlo experiment, and an error term $\delta_{in}$.

**(4-4)** $$p_{in} = \alpha_c * c_{in} + \alpha_d * d_{in} + \alpha_z * z_{in} + \delta_{in}$$

The goal underlying this thesis is to find out, for example, what occurs if the researcher does not observe an attribute that is correlated with the price, such as ***d***.

In this chapter are presented four Monte Carlo experiments. The first one is a case where endogeneity is simulated to occur at an individual level, and good instruments are available. The second experiment differs from the first one only in that the instruments are poorer. The third experiment considers that endogeneity occurs in perfectly disjointed markets and that good instruments are available. Finally, the last experiment is the same as the third one, but with bad instruments.

As it was discussed before, the control-function method should perform better in the first two cases and the BLP method should be the most suitable for the latter ones.

For each Monte Carlo experiment, three endogeneity non-corrected and four endogeneity corrected models were estimated. The non-corrected models correspond to: A) a model where all variables were included, B) a model where the variable excluded *a* is not correlated with the price and C) a model where the variable excluded *d* is correlated with the price. The corrected models estimated in each case correspond to: D) the application of the control-function method, E) the simple Instrumental Variable method, F) the usage of Matzkin type instruments and G) the BLP method.

### 4.2.1 Monte Carlo Experiment One: Individual Variation and Good Instruments

For the first Monte Carlo experiment variables *a, b, c, d* and the additional instrument *z*, were created using the random number generator macro contained in Microsoft's Excel software considering that the variables were *iid* uniform (0,1) for each household and alternative. Attribute *p* was generated, using the price equation (4-5), as a function of *c, d* and the exogenous instrument *z*. The error term $\delta_{in}$ was constructed to be *iid* Normal (0, 0.1)[3]. Within this setting, variables *c* and *d* are correlated with the *price p* but not *a* nor *b*, as follows.

**(4-5)** $$p_{in} = 0.5 * c_{in} + 0.5 * d_{in} + 0.5 * z_{in} + \delta_{in}$$

Finally, an additional variable *m*, that corresponds to the type of instruments suggested by Maztkin (2004), was constructed as a function of variables *d* and *z* as shown in expression (4-6) , where the error term $\xi_{in}$ is considered to be Normal (0,1).

**(4-6)** $$m_{in} = 0.5 * d_{in} + 0.5 * z_{in} + \xi_{in}$$

Variable $m_{in}$ in (4-6) complies with the simplest case requirements of the auxiliary variable proposed by Matzkin (2004) to include in the model, as it was explained in

(3-26). If (4-6) is replaced in (4-5) it can be noted that the endogenous variable $p_{in}$ can be written as a function of the auxiliary variable $m_{in}$ plus an exogenous perturbation $\eta_{in}$.

(4-7) $$p_{in} = m_{in} + 0.5 * c_{in} + \delta_{in} - \xi_{in} = m_{in} + \eta_{in}$$

Within this setting, seven models where estimated. Their parameters are shown by column in Table 4-1 and Table 4-3, together with the direct price elasticity for each of the three alternatives, $e_{11}$, $e_{22}$, $e_{33}$, evaluated at the sample mean of each attribute. The first column of these tables corresponds to the labels. The second one corresponds to the true values of the parameters to be estimated taken from (4-3).

**Table 4-1 Monte Carlo Experiment One. Models 1-A to 1-D**

| Variable | TRUE Values | MODEL 1-A Complete Model | | MODEL 1-B Omitting a non Correlated with p | | MODEL 1-C Omitting d Correlated with p | | MODEL 1-D Control Function Correction | |
|---|---|---|---|---|---|---|---|---|---|
| *ASC1* | | -0.0783 | (-0.674) | -0.0221 | (-0.312) | 0.0333 | (0.418 ) | 0.0225 | (0.251 ) |
| *ASC2* | | 0.0883 | (0.777 ) | -0.0324 | (-0.456) | 0.0420 | (0.533 ) | -0.0163 | (-0.183) |
| *a* | 10.0 | 10.5 | (21.4 ) | | | 5.14 | (25.5 ) | 6.45 | (24.7 ) |
| *b* | 10.0 | 10.6 | (21.4 ) | 4.04 | (25.2 ) | 5.25 | (25.7 ) | 6.52 | (24.7 ) |
| *c* | 10.0 | 10.4 | (20.1 ) | 4.14 | (21.4 ) | 3.02 | (16.3 ) | 6.28 | (21.3 ) |
| *d* | 10.0 | 10.4 | (20.6 ) | 4.32 | (21.9 ) | | | | |
| *p* | -10.0 | -10.6 | (-19.0) | -4.13 | (-18.0) | -1.00 | (-5.74) | -6.61 | (-17.6) |
| $\hat{\mu}$ | | | | | | | | 8.95 | (18.2 ) |
| *S. Size* | | 2000 | | 2000 | | 2000 | | 2000 | |
| *LL(0)* | | -2197.22 | | -2197.22 | | -2197.22 | | -2197.22 | |
| *Final LL* | | -517.93 | | -1355.03 | | -1098.94 | | -865.21 | |
| Adj $\rho^2$ | | 0.761 | | 0.383 | | 0.497 | | 0.606 | |
| $e_{11}$ | -11.5 | -12.7 | | -4.82 | | -1.16 | | -7.67 | |
| $e_{22}$ | -11.7 | -11.9 | | -4.82 | | -1.15 | | -7.71 | |
| $e_{33}$ | -11.6 | -11.8 | | -4.78 | | -1.14 | | -7.65 | |

t-test in bracklets

$e_{ii}$ direct elasticity alternative i

*LL* Log-likelihood

Adj. $\rho^2$ = 1 – (final *LL* - #Attributes)/*LL(0)*

$\hat{\mu}$ household-alternative own fitted error of the price equation

In the third column of Table 4-1 are presented the estimated parameters of the **Model 1-A**[4], which is the estimation of a MNL model[5] that includes all the attributes that are

[3] Such a small variance is just to be sure that the endogeneity problem within this setting is relevant. If the variance is large, then a bigger part of the unobserved part of $p$ would be explained by this *white noise* and not by the unobserved attribute $d$ and this will reduce or eliminate endogeneity problem.

[4] The label assigns a "1" for the Monte Carlo experiment and an "A" for the choice model considered.

[5] All MNL models were estimated using the software BIOGEME, Bierlaire *et al.* (2004).

relevant in the choice behavior, that is, *a*, *b*, *c*, *d* and *p*. Not surprisingly, for this model all estimated parameters and elasticities are statistically equal to the true ones.

In the fourth column of Table 4-1 are presented the estimated parameters of **Model 1-B,** where attribute *a* was omitted. Theoretically, because by construction attribute *a* is not correlated with any observed attribute (in particular with the price), this model should be consistent. However, the estimated parameters are smaller than the real ones. This is because of the change in the scale parameter of the model. The error term in this model is wider because of the omission of *a*. This implies a bigger variance and then a smaller scale parameter (Ben-Akiva and Lerman, 1985), which explains the fact that estimated parameters are smaller than the real ones.

However, it can be noted also that all parameters have approximately[6] the same absolute value, as occurred in the true model, fact that confirms the theoretical result about consistent estimates. The elasticities are also affected by the difference in the scale parameter of the model producing values that are smaller than the ones of the true model but, at least, the relative values of the elasticities of each alternative are similar as occur in the true model.

In the fifth column of Table 4-1 are shown the estimated parameters of **Model 1-C** where attribute *d*, which is correlated with the price because of expression (4-4), was omitted. In this case, the estimated parameters are quite different from the ones of the true model, beyond the scale parameter difference. Taking the parameters of *a* and *b* as the base, the parameter of price *p*, is **5 times smaller than it should** and the parameter of *c* is almost **two times smaller**. That is, as expected, under the omission of a relevant attribute that is correlated with the price, the price parameter is upward biased, making the models almost useless or at least non-trustable. Elasticities are also affected by the omission of *d* but now resulting in values even smaller than with the omission of *a*.

The following estimated model is **Model 1-D** where, again, attribute *d* is omitted. However, in this case the control-function method (Petrin and Train, 2004) was applied

[6] A Likelihood ratio test was performed rejecting the difference of the parameters with 95% of confidence.

to correct for endogeneity. Under the error structure assumed[7], the control-function method to correct for the omitted attributes is performed in two straightforward steps.

The first step is to regress (OLS)[8] dwelling-unit price $p$ on exogenous instruments, excluding $d$, because it is assumed to be not observed, and to calculate the fitted errors of this model as in (4-8).

$$p_{in} = \alpha_1 * c_i + \alpha_2 * z_i + \mu_{in} \quad \Rightarrow \quad \hat{\mu}_{in} = p_{in} - \hat{\alpha}_1 * c_i - \hat{\alpha}_2 * z_i \tag{4-8}$$

The estimated results of this model are reported in Table 4-2. The adjustment of only 57% is due to the omission of variable $d$, which explains, by construction (4-5), an important part of $p$. As expected, the estimated parameters of $c$ and $z$ are statistically equal (at a 95% of confidence interval) to the true ones.

**Table 4-2 Price Equation Model 1-D and 1-E**

| *Variables* | Parameters | |
|---|---|---|
| **Intercept** | 0.768 | (60.8 ) |
| **c** | 0.488 | (61.8 ) |
| **z** | 0.499 | (62.8 ) |
| **Adjusted $R^2$** | 0.569 | |
| ***S. Size*** | 6000 | |

t-test in brackets

The next step is the estimation of a MNL choice model, excluding $d$ as a variable, but using the fitted residuals from the price equation (4-8) as an extra variable in the utility function.

$$U_{in} = ASC_i + \beta_1 * a_i + \beta_2 * b_i + \beta_3 * c_i + \beta_4 * p_i + \phi * \hat{\mu}_{in} + e_{in} \tag{4-9}$$

The results of the application of this model are shown in the sixth column of Table 4-1 (Model 1-D). In this case, the parameters of the observed attributes $a$, $b$, $c$ and $p$, have approximately[9] the same absolute value as occurred in the true model, so it can be claimed that the inclusion of $\hat{u}_{in}$ as the extra variable in the choice model satisfactorily corrected the problem of the omission of $d$. Elasticities in this case are substantially

[7] It was assumed that the variance-covariance matrices of ε, μ and are diagonal. Petrin and Train (2004) affirm that, within this setting, the control-function is a constant multiplied by the own price residuals. However, a formal proof for this result is left for future research.

[8] OLS models were estimated using the respective routines included in the software MicroSoft EXCEL.

[9] A Likelihood ratio test was performed rejecting the difference of the parameters with 95% of confidence.

nearer to the true values compared with the ones of the Model 1-C without the control-function correction, and even better than the ones of the Model 1-B where *a* was omitted. This last statement could be explained by the fact that the inclusion of the control-function as an additional variable accounts for a part of the additional variability of the error term, moving the scale parameter up again and, with it, the estimated parameters and the elasticities.

The following estimated model is **Model 1-E** where, again, attribute *d* is omitted. In this case the traditional instrumental variables procedure was used to control for endogeneity. As stated before, Hausman (1978) noted that this method is equivalent to the control-function method (3-14), as long as the function of the fitted errors used is just the contemporary error and the parameters of the price and the control-function are forced to be equal. The procedure was developed in two steps.

The first step corresponded to regress (OLS) the dwelling-unit price *p* on exogenous instruments, excluding *d* and to calculate the fitted price of this model as in (4-10). The estimated results of this model are the ones presented in Table 4-2.

**(4-10)** $$p_{in} = \alpha_1 * c_i + \alpha_2 * z_i + \mu_{in} \quad \Rightarrow \quad \hat{p}_{in} = \hat{\alpha}_1 * c_i + \hat{\alpha}_2 * z_i$$

The second step corresponded to the estimation of a MNL choice model, excluding *d* as a variable, and using the fitted prices from the price equation (4-10) instead of actual prices as a variable in the utility function.

**(4-11)** $$U_{in} = ASC_i + \beta_1 * a_{in} + \beta_2 * b_{in} + \beta_3 * c_{in} + \varphi * \hat{p}_{in} + \mathrm{e}_{in}$$

The results of the estimation of the model (4-11) are shown in the third column of Table 4-3. As expected, the estimated parameters and elasticities of this model are very similar to the ones of the control-function method, Model 1-D. The differences arise, as was shown in (3-15), from the fact that in the control-function method the parameter of the fitted error is allowed to be different from the one of the price and in the instrumental variables method they are forced to be equal[10].

---

[10] A likelihood ratio test was performed between models 1-D and 1-E showing that, with 95% of confidence, the parameters are different. That is, the control-function method is statistically superior to the instrumental variables method. For this test was considered one degree of freedom, which is the additional parameter that is allowed to be different in the control-function method.

The following estimated model is **Model 1-F** where, again, attribute *d* is omitted. In this case the correction was made using the Matzkin type instrument *m*, constructed for this Monte Carlo experiment (4-6). This method was applied in one step by estimating a choice model (4-12) where *m* was included as an additional variable.

**(4-12)** $$U_{in} = ASC_i + \beta_1 * a_{in} + \beta_2 * b_{in} + \beta_3 * c_{in} + \beta_4 * p_{in} + \psi * m_{in} + e_{in}$$

The results of the application of this model are shown in the fourth column of Table 4-3. As expected, the estimated parameters and elasticities of this model are very similar to the ones of the control-function method.

The last estimated model is **Model 1-G**, which is the application of the product market fix effects (BLP) procedure considering (erroneously in this Monte Carlo experiment) that the data is ordered in 40 groups of fifty households each one. This procedure was developed in the following three steps.

The first step corresponded to the estimation of a choice model that excludes *d* but includes alternative specific constants ($\delta_{im(n)}$) for each alternative[11] *i* and for each one of the 40 markets *m(i)*.

**(4-13)** $$U_{in} = \delta_{im(n)} + \beta_1 * a_{in} + \beta_2 * b_{in} + \beta_3 * c_{in} + \beta_4 * p_{in} + e_{in}$$

The estimated results of the first step of the product market fix effect procedure are shown in fifth column of Table 4-3. It can be seen that, as in this Monte Carlo experiment endogeneity does not occur in differentiated markets (but individually), the $\delta_{im(n)}$ can't eliminate endogeneity and then the model is very poor, as bad as the model were *d* is omitted without any correction, Model 1-C, that is in the fifth column of Table 4-1.

In the second step, dwelling-unit price *p* was regressed (OLS) on exogenous instruments, excluding *d*. Then, using the estimated parameters of this model the fitted price was calculated as shown in (4-14).

**(4-14)** $$p_{in} = \alpha_1 * c_i + \alpha_2 * z_i + \mu_{in} \quad \Rightarrow \quad \hat{p}_{in} = \hat{\alpha}_1 * c_i + \hat{\alpha}_2 * z_i$$

[11] For identification, is necessary to fix the ASC of one of the alternatives in each market to be zero.

In the third step, the estimated alternative specific constants of model (4-13) were regressed (OLS) on instruments and fitted prices instead of actual prices (4-15). The bar over each attribute indicates that it is being used the market average (what can safely be done because the model is linear ) and the difference with respect to the base alternative of each observation.

**(4-15)** $$\hat{\bar{\delta}}_{im} = \gamma_1 * \bar{a}_{im} + \gamma_2 * \bar{b}_{im} + \gamma_3 * \bar{c}_{im} + \gamma_4 * \bar{\hat{p}}_{im} + \varsigma_{in}$$

The estimated results of the third step of the product-market fix effect procedure are shown in the sixth column of Table 4-3. In this case, endogeneity is treated in the model by using instrumented prices what causes the price parameter to become bigger. However, the poor adjustment of the model (adjusted $R^2$ of only 9%, which appears on the row that corresponds to the adjusted $\rho^2$ of the second column of Model 1-G in Table 4-3) indicates that the estimated parameters are not completely trustable. This can be the result of the limited sample size.

The net effect of the application of the product-market fix effects is the sum of the estimated parameters of each variable and is shown in the last column of Table 4-3. It can be noted that even thought the assumption of market differentiated endogeneity is not valid in this Monte Carlo experiment, the final parameters of the model are very similar in absolute value between them (as in the true model) and the elasticities are nearer to the real ones than in Model 1-C, but below the ones of Model 1-D. This can be the result of the loss of variability due to the aggregation of the variables in Model 1-G.

Finally, the computational cost of running the BLP method in this case was huge compared to the other ones. For this 2000 sample, all MNL models, but the last one, took around 10 seconds to run in a conventional PC (Pentium 4, 2.20 GHz, 224 MB of RAM). Instead, the product-market fix effects took around 15 minutes, that is, 90 times more. This is explained by the fact that 77 additional parameters have to be estimated, due to the inclusion of the ASC by market, increasing considerably the computational burden. Despite this number can sound big, note that the number of markets or zones was intentionally maintained small enough to make the model estimable with standard software leaving, on the other hand, only a reduced number (80) of observations to run

the third stage of the product-market fix effects procedure, which can certainly affect the estimation results.

**Table 4-3 Monte Carlo Experiment One. Models 1-E to 1-G**

| Variable | TRUE Values | MODEL 1-E IV Correction | | MODEL 1-F Matzkin Correction | | MODEL 1-G BLP MNL Model | | MODEL 1-G BLP ASC IV Model | | MODEL 1-G BLP FINAL |
|---|---|---|---|---|---|---|---|---|---|---|
| *ASC1* | | 0.0115 | (0.130 ) | 0.0786 | (0.891 ) | | | 0.109 | (1.71 ) | 0.109 |
| *ASC2* | | 0.0259 | (0.299 ) | 0.0904 | (1.05 ) | | | 0.109 | (1.71 ) | 0.109 |
| *a* | 10.0 | 6.21 | (24.9 ) | 6.13 | (25.0 ) | 5.39 | (24.9 ) | 1.35 | (1.27 ) | 6.74 |
| *b* | 10.0 | 6.24 | (25.0 ) | 6.29 | (25.3 ) | 5.53 | (25.3 ) | -0.29 | (-0.300) | 5.24 |
| *c* | 10.0 | 6.10 | (21.4 ) | 6.21 | (20.7 ) | 3.18 | (16.3 ) | 1.70 | (1.29 ) | 4.88 |
| *d* | 10.0 | | | | | | | | | |
| *p* | -10.0 | -6.37 | (-17.6) | -6.37 | (-16.7) | -1.04 | (-5.67) | -4.66 | (-3.01) | -5.70 |
| *m* | | | | 6.54 | (16.8 ) | | | | | |
| *S. Size* | | 2000 | | 2000 | | 2000 | | 80 | | 2000 |
| *LL(0)* | | -2197.22 | | -2197.22 | | -2197.22 | | | | |
| *Final LL* | | -907.38 | | -913.11 | | -1050.97 | | | | |
| Adj $\rho^2$ | | 0.584 | | 0.581 | | 0.483 | | 0.0903 | | |
| $e_{11}$ | -11.5 | -7.47 | | -7.36 | | -1.22 | | -5.30 | | -6.53 |
| $e_{22}$ | -11.7 | -7.48 | | -7.33 | | -1.21 | | -5.49 | | -6.65 |
| $e_{33}$ | -11.6 | -7.41 | | -7.26 | | -1.20 | | -5.44 | | -6.59 |

t-test in brackets

$e_{ii}$ direct elasticity alternative i

*LL* Log-likelihood

Adj. $\rho^2 = 1 - (\text{final } LL - \#\text{Attributes})/LL(0)$

*m* Matzkin type instrument

Adj. $\rho^2$ corresponds to Adj. $R^2$ for Model G

In models with real data many more markets can be expected, increasing the number of estimated parameters and then the computational burden. However, this can be addressed using the contraction (or *calibration*) procedure (3-18) described by BLP (1995), which is not allowed in the conventional package (BIOGEME) used in this study and thus, goes beyond the scope of the present thesis.

## 4.2.2 Monte Carlo Experiment Two: Individual Variation and Weak Instruments

The setting of this Monte Carlo experiment is the same as the last one in everything but the fact that the instrumental variable $z$ is now only slightly correlated with the price and with the Matzkin type instrument. Formally, in this model variables $p$ and $m$ were generated using the following expressions, and all other variables remained the same.

**(4-16)** $$p_{in} = 0.5 * c_{in} + 0.5 * d_{in} + 0.01 * z_{in} + \delta_{in},$$

(4-17) $$m_{in} = 0.5 * d_{in} + 0.01 * z_{in} + \xi_{in},$$

Thus, in this case the instrument $z$ is poorly correlated with $p$ and $m$, and also the endogeneity problem is more serious, because $d$, the omitted attribute, explains a bigger part of the price.

For this experiment, the same set of models was estimated and their results are shown in Table 4-4 and Table 4-5. As stated before, in this case the endogeneity problem is more serious than what happened in the Monte Carlo experiment one. This can be noted by looking at the estimated results for Model 2-C that are reported in the fifth column of Table 4-4, where now, when attribute $d$ is omitted, the estimated parameter of price is positive. In all cases, the proposed corrections restored, at least, the correct sign for the price parameter.

**Table 4-4 Monte Carlo Experiment Two. Models 2-A to 2-D**

| Variable | TRUE Values | MODEL 2-A Complete Model | | MODEL 2-B Omitting a non Correlated with p | | MODEL 2-C Omitting d Correlated with p | | MODEL 2-D Control Function Correction | |
|---|---|---|---|---|---|---|---|---|---|
| *ASC1* | 0.00 | -0.240 | (-2.103) | -0.0903 | (-1.313) | -0.0956 | (-1.10) | -0.0957 | (0.251 ) |
| *ASC2* | 0.00 | -0.0317 | (-0.285) | -0.0871 | (-1.275) | -0.0762 | (-0.885) | -0.0762 | (-0.183) |
| *a* | 10.0 | 10.4 | (22.3 ) | | | 6.45 | (25.5 ) | 6.45 | (24.7 ) |
| *b* | 10.0 | 10.4 | (22.2 ) | 3.88 | (25.5 ) | 6.39 | (25.3 ) | 6.39 | (24.7 ) |
| *c* | 10.0 | 10.6 | (18.9 ) | 4.22 | (18.0 ) | 2.15 | (10.5 ) | 3.29 | (21.3 ) |
| *d* | 10.0 | 10.6 | (19.0 ) | 4.29 | (17.9 ) | | | | |
| *p* | -10.0 | -11.0 | (-15.2) | -4.26 | (-11.9) | 2.01 | (7.98 ) | -0.312 | (18.2 ) |
| $\hat{\mu}$ | | | | | | | | 2.33 | (-17.6) |
| *S. Size* | | 2000 | | 2000 | | 2000 | | 2000 | |
| *LL(0)* | | -2197.22 | | -2197.22 | | -2197.22 | | -2197.22 | |
| *Final LL* | | -541.96 | | -1438.56 | | -915.79 | | -915.78 | |
| Adj $\rho^2$ | | 0.750 | | 0.343 | | 0.580 | | 0.606 | |
| $e_{11}$ | -9.87 | -11.7 | | -4.33 | | 2.07 | | -0.320 | |
| $e_{22}$ | -9.98 | -10.7 | | -4.29 | | 2.01 | | -0.311 | |
| $e_{33}$ | -9.90 | -10.6 | | -4.26 | | 1.99 | | -0.308 | |

t-test in bracklets

$e_{ii}$ direct elasticity alternative i

*LL* Log-likelihood

Adj. $\rho^2$ = 1 – (final *LL* - #Attributes)/*LL(0)*

$\hat{\mu}$ household-alternative own fitted error of the price equation

All proposed methods are less efficient in the correction of the problem under weak instruments but some differences can be noted. The most robust in this particular case is Model 2-G, the BLP method, and the worst one Model 2-D, the control-function. A case apart is Model 2-F the Matzkin method where parameters in this example are biased

upward instead of downward, what serves to show the misleading pictures that can be obtained if week instruments are used to apply the correction methods.

**Table 4-5 Monte Carlo Experiment Two. Models 2-E to 2-G**

| Variable | TRUE Values | MODEL 2-E IV Correction | | MODEL 2-F Matzkin Correction | | MODEL 2-G BLP MNL Model | | MODEL 2-G BLP ASC IV Model | | MODEL 2-G BLP FINAL |
|---|---|---|---|---|---|---|---|---|---|---|
| *ASC1* | | -0.0870 | (-1.02) | 0.129 | (0.755 ) | | | -0.0410 | (-0.530) | -0.0410 |
| *ASC2* | | -0.0494 | (-0.584) | -0.0597 | (-0.364) | | | -0.0410 | (-0.530) | -0.0410 |
| *a* | 10.0 | 6.27 | (25.7 ) | 24.2 | (15.3 ) | 6.81 | (24.9 ) | 2.38 | (2.18 ) | 9.19 |
| *b* | 10.0 | 6.19 | (25.5 ) | 23.9 | (15.4 ) | 6.77 | (25.3 ) | -0.218 | (-0.222) | 6.55 |
| *c* | 10.0 | 3.82 | (0.469 ) | 20.5 | (14.8 ) | 2.31 | (16.3 ) | -0.763 | (-0.567) | 1.55 |
| *d* | 10.0 | | | | | | | | | |
| *p* | -10.0 | -1.49 | (-0.0895) | -17.1 | (-13.6) | 2.13 | (-5.67) | -5.62 | (-0.883) | -3.49 |
| *m* | | | | 38.8 | (14.9 ) | | | | | |
| *S. Size* | | 2000 | | 2000 | | 2000 | | 80 | | 2000 |
| *LL(0)* | | -2197.22 | | -2197.22 | | -2197.22 | | | | |
| *Final LL* | | -949.64 | | -244.92 | | -871.38 | | | | |
| Adj $\rho^2$ | | 0.565 | | 0.885 | | 0.565 | | 0.0281 | | |
| $e_{11}$ | -9.87 | -1.54 | | -16.2 | | 2.15 | | -5.58 | | -3.49 |
| $e_{22}$ | -9.98 | -1.49 | | -17.5 | | 2.10 | | -5.85 | | -3.56 |
| $e_{33}$ | -9.90 | -1.48 | | -17.3 | | 2.08 | | -5.80 | | -3.53 |

t-test in bracklets

$e_{ii}$ direct elasticity alternative i

*LL* Log-likelihood

Adj. $\rho^2$ = 1 – (final *LL* - #Attributes)/*LL(0)*

*m* Matzkin type instrument

Adj. $\rho^2$ corresponds to Adj. $R^2$ for Model G

Finally, in Table 4-6 are presented the estimated results of the price equation used for Models 2-D and 2-E. It can be noted that, despite the general adjustment is acceptable, the correlation between the instrument *z* and the endogenous variable *p* is too small, what leads to the problems described before.

**Table 4-6 Price Equation Model 2-D and 2-E**

| *Variables* | Parameters | |
|---|---|---|
| **Intercept** | 0.768 | (60.8 ) |
| **c** | 0.488 | (61.8 ) |
| **z** | 0.00870 | (1.10 ) |
| **Adjusted $R^2$** | 0.389 | |
| ***S. Size*** | 6000 | |

t-test in brackets

The general conclusion is that it is extremely relevant to be sure to have good instruments to develop any of the analyzed correction methods. Also, apparently, the BLP method is more robust for this kind of problems.

## 4.2.3 Monte Carlo Experiment Three: Zonal Variation and Good Instruments

The setting of this Monte Carlo experiment is equal to the first one in terms of the true parameters considered but differs in the fact that variable *d* was defined as equal by markets or zones defined as the 40 clusters of 50 household considered in the definition of the BLP Model G. Variables *b*, *c* and *p* were construct *iid* uniform as half varying by zone and half individually. Variable *a* was constructed as varying only individually. All model parameters remained the same as in (4-3), (4-5) and (4-6). Under this setting the BLP method should perform better.

**Table 4-7 Monte Carlo Experiment Three. Models 3-A to 3-D**

| Variable | TRUE Values | MODEL 3-A Complete Model | | MODEL 3-B Omitting a non Correlated with p | | MODEL 3-C Omitting d Correlated with p | | MODEL 3-D Control Function Correction | |
|---|---|---|---|---|---|---|---|---|---|
| *ASC1* | | -0.173 | (-1.159) | -0.0711 | (-0.727) | 0.102 | (1.25 ) | 0.155 | (1.31 ) |
| *ASC2* | | 0.0151 | (0.110 ) | -0.157 | (-1.74) | 0.243 | (2.96 ) | 0.0819 | (0.712 ) |
| *a* | 10.0 | 9.49 | (17.8 ) | | | 3.32 | (19.7 ) | 6.45 | (19.5 ) |
| *b* | 10.0 | 9.66 | (18.6 ) | 4.02 | (24.0 ) | 3.21 | (24.5 ) | 6.57 | (21.3 ) |
| *c* | 10.0 | 9.56 | (17.7 ) | 3.84 | (20.5 ) | 0.845 | (9.64 ) | 5.32 | (19.5 ) |
| *d* | 10.0 | 9.31 | (18.4 ) | 3.97 | (23.1 ) | | | | |
| *p* | -10.0 | -9.09 | (-16.5) | -3.76 | (-16.8) | 1.90 | (20.2 ) | -3.76 | (-14.2) |
| $\hat{\mu}$ | | | | | | | | 9.68 | (20.6 ) |
| *S. Size* | | -377 | | 2000 | | 2000 | | 2000 | |
| *LL(0)* | | -2197.22 | | -2197.22 | | -2197.22 | | -2197.22 | |
| *Final LL* | | -386.24 | | -875.26 | | -1109.49 | | -555.15 | |
| Adj $\rho^2$ | | 0.825 | | 0.594 | | 0.492 | | 0.744 | |
| $e_{11}$ | -13.0 | -12.0 | | -4.30 | | 2.04 | | -4.46 | |
| $e_{22}$ | -6.53 | -5.92 | | -3.42 | | 1.68 | | -3.06 | |
| $e_{33}$ | -11.0 | -5.91 | | -3.80 | | 2.11 | | -3.98 | |

t-test in bracklets

$e_{ii}$ direct elasticity alternative i

*LL* Log-likelihood

Adj. $\rho^2 = 1 - (\text{final } LL - \#\text{Attributes})/LL(0)$

$\hat{\mu}$ household-alternative own fitted error of the price equation

In this case, the same seven models were estimated and their results are shown in Table 4-7 and Table 4-8 . As expected, the BLP method Model 3-G is the one that does it better in recovering the true parameters, at least in recovering the elasticities true values, but not as good as the control-function did it in Monte Carlo experiment one, Model 1-C. Moreover, note that he parameter of *b* and *c* in the last column of Table 4-8 are seriously

upward biased. This could be explained by the reduced number of observations (80) used to model the endogeneity problem by zone in this Monte Carlo experiment.

The control-function method, Model 3-C in Table 4-7, was not as good as the first stage of the BLP method, Model 3-G in Table 4-8, in solving the endogeneity bias because the price parameter of Model 1-C is almost half of the parameters *a* and *b*. However, the estimated parameters in this case have the correct sign and values that are more similar between each other, as occurred in the true model.

**Table 4-8 Monte Carlo Experiment Three. Models 3-E to 3-G**

| Variable | TRUE Values | MODEL 3-E IV Correction | | MODEL 3-F Matzkin Correction | | MODEL 3-G BLP MNL Model | | MODEL 3-G BLP ASC IV Model | | MODEL 3-G BLP FINAL |
|---|---|---|---|---|---|---|---|---|---|---|
| *ASC1* | | -0.114 | (-1.53) | 0.00967 | (0.113 ) | | | -0.440 | (-0.200) | -0.440 |
| *ASC2* | | 0.0730 | (1.00 ) | 0.208 | (2.45 ) | | | -0.440 | (-0.200) | -0.440 |
| *a* | 10.0 | 2.78 | (19.3 ) | 3.56 | (19.9 ) | 10.2 | (17.0 ) | | | 10.2 |
| *b* | 10.0 | 3.00 | (25.4 ) | 3.43 | (24.6 ) | 10.8 | (13.8 ) | 9.39 | (2.70 ) | 20.2 |
| *c* | 10.0 | 2.19 | (19.8 ) | 4.58 | (13.9 ) | 10.1 | (11.6 ) | 7.13 | (1.70 ) | 17.3 |
| *d* | 10.0 | | | | | | | | | |
| *p* | -10.0 | -1.74 | (-12.1) | -5.34 | (-8.85) | -9.06 | (-9.40) | -2.49 | (-0.500) | -11.6 |
| *m* | | | | 7.43 | (11.9 ) | | | | | |
| *S. Size* | | 2000 | | 2000 | | 2000 | | 80 | | 2000 |
| *LL(0)* | | -2197.22 | | -2197.22 | | -2197.22 | | | | |
| *Final LL* | | -1321.13 | | -1029.44 | | -359.95 | | | | |
| Adj $\rho^2$ | | 0.396 | | 0.528 | | 0.798 | | 0.0274 | | |
| $e_{11}$ | -13.0 | -1.79 | | -5.83 | | -8.03 | | -2.76 | | -11.9 |
| $e_{22}$ | -6.53 | -1.76 | | -4.66 | | -11.0 | | -2.84 | | -14.9 |
| $e_{33}$ | -11.0 | -1.77 | | -5.88 | | -8.83 | | -2.03 | | -8.63 |

t-test in bracklets

$e_{ii}$ direct elasticity alternative i

*LL* Log-likelihood

Adj. $\rho^2 = 1 - (\text{final } LL - \#\text{Attributes})/LL(0)$

*m* Matzkin type instrument

Adj. $\rho^2$ corresponds to Adj. $R^2$ for Model G

This fact, joined to the easy implementation of this method using standard software, shows that, even if the assumptions behind the application of the control-function method are not valid, this method can be used as a good first step to test for the presence of endogeneity.

Additionally, the price equation estimations for Models 3-D and 3-E are presented in Table 4-9, where it can be seen that, as expected, the estimated parameters are statistically equal to the true ones.

**Table 4-9 Price Equation Model 3-D and 3-E**

| *Variables* | Parameters | |
|---|---|---|
| **Intercept** | 0.434 | (23.2 ) |
| **c** | 0.508 | (43.4 ) |
| **z** | 0.569 | (49.8 ) |
| **Adjusted $R^2$** | 0.413 | |
| ***S. Size*** | 6000 | |

t-test in brackets

## 4.2.4 Monte Carlo Experiment Four: Zonal Variation and Weak Instruments

The only difference in the setting of this Monte Carlo experiment compared with the previous one is that the instrumental variable $z$ is assumed to be only slightly correlated with the price and with the Matzkin type instrument, just as it was for Monte Carlo experiment two. Also all model parameters remain the same as in (4-16), (4-17).

For this setting, the same set of models was estimate and their results are shown in Table 4-10 and Table 4-11. As occurred with Monte Carlo experiment two, the general conclusion is that it is extremely relevant to be sure to have good instruments to develop any of the analyzed correction methods.

In this case, the only procedure that solves the endogeneity problem, in the sense of, at least, recovering a negative price parameter is the first stage of the BLP method Model 4-G, where the inclusion of the zone-alternative specific constants addressed the endogeneity problem that in this case occurs by market. However, the other stages in the BLP method are so bad that the final model has again a positive parameter for the price.

A case apart is Model 4-F the Matzkin method where parameters in this example are biased upward instead of downward, what serves again to show the misleading pictures that can be obtained if week instruments are considered.

**Table 4-10 Monte Carlo Experiment Four. Models 4-A to 4-D**

| Variable | TRUE Values | MODEL 4-A Complete Model | | MODEL 4-B Omitting a non Correlated with p | | MODEL 4-C Omitting d Correlated with p | | MODEL 4-D Control Function Correction | |
|---|---|---|---|---|---|---|---|---|---|
| *ASC1* | | -0.206 | (-1.32) | -0.127 | (-1.29) | 0.248 | (2.07 ) | 0.286 | (2.36 ) |
| *ASC2* | | -0.122 | (-0.813) | -0.250 | (-2.65) | 0.0393 | (0.332 ) | 0.131 | (1.07 ) |
| *a* | 10.0 | 10.6 | (16.7 ) | | | 6.80 | (19.0 ) | 6.98 | (18.9 ) |
| *b* | 10.0 | 10.8 | (17.5 ) | 4.17 | (23.7 ) | 7.17 | (20.7 ) | 7.26 | (20.5 ) |
| *c* | 10.0 | 10.4 | (14.2 ) | 4.01 | (12.3 ) | 0.487 | (3.9 ) | -3.40 | (-4.06) |
| *d* | 10.0 | 10.2 | (14.1 ) | 4.15 | (12.5 ) | | | | |
| *p* | -10.0 | -9.41 | (-8.88) | -3.89 | (-6.45) | 6.74 | (20.8 ) | 14.6 | (8.39 ) |
| $\hat{\mu}$ | | | | | | | | -7.75 | (-4.68) |
| *S. Size* | | 2000 | | 2000 | | 2000 | | 2000 | |
| *LL(0)* | | -2197.22 | | -2197.22 | | -2197.22 | | -2197.22 | |
| *Final LL* | | -332.36 | | -845.23 | | -522.54 | | -511.14 | |
| Adj $\rho^2$ | | 0.846 | | 0.613 | | 0.759 | | 0.764 | |
| $e_{11}$ | -9.29 | -9.01 | | -3.22 | | 5.69 | | 12.4 | |
| $e_{22}$ | -4.58 | -4.31 | | -2.66 | | 4.05 | | 8.40 | |
| $e_{33}$ | -8.84 | -4.38 | | -2.98 | | 5.60 | | 12.5 | |

t-test in brackets

$e_{ii}$ direct elasticity alternative i

*LL* Log-likelihood

Adj. $\rho^2$ = 1 – (final *LL* - #Attributes)/*LL(0)*

$\hat{\mu}$ household-alternative own fitted error of the price equation

**Table 4-11 Monte Carlo Experiment Four. Models 4-E to 4-G**

| Variable | TRUE Values | MODEL 4-E IV Correction | | MODEL 4-F Matzkin Correction | | MODEL 4-G BLP MNL Model | | BLP ASC IV Model | | BLP FINAL |
|---|---|---|---|---|---|---|---|---|---|---|
| *ASC1* | | -0.113 | (-1.58) | 0.0267 | (0.145 ) | | | 1.81 | (0.700 ) | 1.81 |
| *ASC2* | | -0.0612 | (-0.833) | 0.0798 | (0.447 ) | | | 1.81 | (0.700 ) | 1.81 |
| *a* | 10.0 | 2.65 | (19.0 ) | 14.6 | (14.6 ) | 11.4 | (16.1 ) | | | 11.4 |
| *b* | 10.0 | 2.97 | (25.3 ) | 15.4 | (15.0 ) | 11.2 | (13.2 ) | 12.9 | (3.60 ) | 24.0 |
| *c* | 10.0 | -0.553 | (-1.15) | 18.9 | (13.7 ) | 10.3 | (8.40 ) | -5.59 | (-0.400) | 4.71 |
| *d* | 10.0 | | | | | | | | | |
| *p* | -10.0 | 3.89 | (4.04 ) | -21.8 | (-11.6) | -8.87 | (-4.50) | 22.4 | (0.90 ) | 13.5 |
| *m* | | | | 36.3 | (13.6 ) | | | | | |
| *S. Size* | | 2000 | | 2000 | | 2000 | | 80 | | 2000 |
| *LL(0)* | | -2197.22 | | -2197.22 | | -2197.22 | | | | |
| *Final LL* | | -1361.07 | | -237.30 | | -310.24 | | | | |
| Adj $\rho^2$ | | 0.378 | | 0.889 | | 0.821 | | 0.170 | | |
| $e_{11}$ | -9.3 | 3.25 | | -21.6 | | -5.36 | | 23.4 | | 13.6 |
| $e_{22}$ | -4.6 | 2.59 | | -7.61 | | -7.92 | | 1.82 | | 1.80 |
| $e_{33}$ | -8.8 | 3.01 | | -20.1 | | -7.04 | | 25.6 | | 15.3 |

t-test in brackets

$e_{ii}$ direct elasticity alternative i

*LL* Log-likelihood

Adj. $\rho^2$ = 1 – (final *LL* - #Attributes)/*LL(0)*

*m* Matzkin type instrument

Adj. $\rho^2$ corresponds to Adj. $R^2$ for Model G

Finally, in Table 4-12 are presented the estimated results of the price equation used for Models 4-D and 4-E. As with Monte Carlo experiment two, it can be noted that the

adjustment is acceptable, but the correlation between the instrument z and the endogenous variable $p$ is too small as it is in the true model (4-16).

**Table 4-12 Price Equation Model 4-D and 4-E**

| *Variables* | Parameters | |
|---|---|---|
| **Intercept** | 0.434 | (23.2 ) |
| **c** | 0.508 | (43.4 ) |
| **z** | 0.0786 | (6.88 ) |
| **Adjusted $R^2$** | 0.242 | |
| ***S. Size*** | 6000 | |

t-test in brackets

## 4.2.5 Non-Parametric Corrections

Following the idea proposed by Petrin and Train (2004), some corrections were explored over the Monte Carlo experiment three defined in 4.2.3 when, mistakenly, the control-function method is applied in a disjointed market endogeneity setting, case where the estimated parameters of the model will be biased.

Three different non-parametric models were estimated. The first corresponded to consider not only the contemporary fitted error $\hat{u}$ as an additional variable, but also the powers of it (up to nine) as a non-linear approximation of the true control-function. The results are shown in the fourth column of the Table 4-13, where it can be noted a slight improvement in the adjusted occurred, but not statistically significant at any confidence level, considering that in this case 8 additional variables were included. Furthermore, no significant changes are observed in the price parameters either.

The second method consisted in considering the parameter of the fitted error $\hat{u}$ as random. This last task was developed using the routine for estimating Mixed Logit models in BIOGEME, considering 500 Halton Draws (Train, 2000). The results are shown in the fifth column of Table 4-13, where it can be noted, again, that no significant improvements were achieved using this non-parametric correction.

The third and final non-parametric method to correct for endogeneity was, following Petrin and Train (2004), to include not only the contemporary fitted error $\hat{u}$ as an additional variable, but also the average fitted error of the other alternatives that belong to

the same market or zone. This variable was defined as $\hat{u}2$. The results are shown in the last column of Table 4-13, where it can be noted that, despite the significant improvement in the adjustment of the model, controlling for the change in the scale parameter, no significant changes in the size of the price parameter were achieved using this non-parametric correction.

The fact that in this experiment, by construction, the control-function method does have a specification problem makes rather surprising not to find significant improvements by the application of the no-parametric procedures described.

**Table 4-13 Non Parametric Corrections for Control-Function**

| Variable | TRUE Values | Control Function Correction | | Control Function $\hat{\mu}, \hat{\mu}^2, ..., \hat{\mu}^9$ | | Control Function Random $\hat{\mu}$ | | Control Function Random $\hat{\mu}, \hat{\mu}2$ | |
|---|---|---|---|---|---|---|---|---|---|
| *ASC1* | | 0.155 | (1.31 ) | 0.196 | (1.61 ) | 0.155 | (1.31 ) | 0.165 | (1.35 ) |
| *ASC2* | | 0.0819 | (0.712 ) | 0.184 | (1.45 ) | 0.0818 | (0.711 ) | 0.109 | (0.915 ) |
| *a* | 10.0 | 6.45 | (19.5 ) | 6.35 | (19.6 ) | 6.45 | (19.5 ) | 6.86 | (19.2 ) |
| *b* | 10.0 | 6.57 | (21.3 ) | 6.45 | (21.3 ) | 6.57 | (21.2 ) | 7.04 | (20.7 ) |
| *c* | 10.0 | 5.32 | (19.5 ) | 5.24 | (19.1 ) | 5.32 | (19.5 ) | 5.77 | (19.3 ) |
| *p* | -10.0 | -3.76 | (-14.2) | -3.56 | (-12.9) | -3.76 | (-14.2) | -4.14 | (-14.6) |
| $\hat{\mu}$ | | 9.68 | (20.6 ) | | | 9.68 | (20.6 ) | -0.757 | (-0.574) |
| *st dev* $\hat{\mu}$ | | | | | | 0.0685 | (0.204 ) | 0.0245 | (0.0698 ) |
| $\hat{\mu}2$ | | | | | | | | 11.3 | (8.03 ) |
| *st dev* $\hat{\mu}2$ | | | | | | | | 0.00197 | (0.00563 ) |
| *S. Size* | | 2000 | | 2000 | | 2000 | | 2000 | |
| *LL(0)* | | -2197.22 | | -2197.22 | | -2197.22 | | -2197.22 | |
| *Final LL* | | -555.15 | | -542.09 | | -555.13 | | -519.82 | |
| Adj $\rho^2$ | | 0.744 | | 0.746 | | 0.744 | | 0.759 | |
| $e_{11}$ | -13.0 | -4.46 | | -3.32 | | -4.46 | | -3.72 | |
| $e_{22}$ | -6.53 | -3.06 | | -3.82 | | -3.06 | | -4.64 | |
| $e_{33}$ | -11.0 | -3.98 | | -3.80 | | -3.98 | | -4.35 | |

t-test in bracklets

$e_{ii}$ direct elasticity alternative i

*LL* Log-likelihood

Adj. $\rho^2 = 1 - (\text{final } LL - \#\text{Attributes})/LL(0)$

$\hat{\mu}$ household-alternative own fitted error of the price equation

$\hat{\mu}2$ Average fitted error of other dwelling units in the same "Comuna"

This result can be attributed to the small size of the sample, or other issues, but it is not really clear yet where the problem comes from, making this an interesting source for future research.

An especially significant result that can be noted is the small significance of the variances of the random parameters of $\hat{u}$ and $\hat{u}2$. That is surprising because, following

Petrin and Train (2004), it was expected that this should be a non-parametric correction method particularly appropriate for cases where zonal endogeneity was erroneously treated using the control-function method (4-1), as precisely occur in this Monte Carlo experiment. This indicates that the study of such corrections needs further research attention.

# Chapter 5

# Application with Real Data from Santiago de Chile

In this chapter, the effect of price endogeneity in the residential location model is analyzed using an actual dataset that was collected in the Origin and Destination Survey in 2001 (Sectra, 2003) in the city of Santiago de Chile.

Two residential location models were estimated. The first one takes into account all the relevant issues in modeling this kind of system that were described in section 2.2 and that can be addressed with the available data, but without correcting for endogeneity at all. Then the control-function method, identified as the most promising method to treat endogeneity in residential location models, is applied in its simplest form and in one variation of it.

# 5.1 Data Description

In 2001, the Ministry of Planning and Cooperation of the government of Chile, adviced by Sectra (2003), designated the Pontificia Universidad Católica de Chile to develop the study "Actualización de Encuestas de Origen y Destino de Viajes, V Etapa" (EOD 2001), a mobility survey for the urban area of the city of Santiago de Chile.

The project considered the following field tasks: the household survey, the mode intercept survey, the complementary external cordon survey, flow counts in barrier and in a set of stations across the city, level of service measures and a fare survey. The information was collected in two periods of the year: a normal period and a summer period, on both work days and weekends.

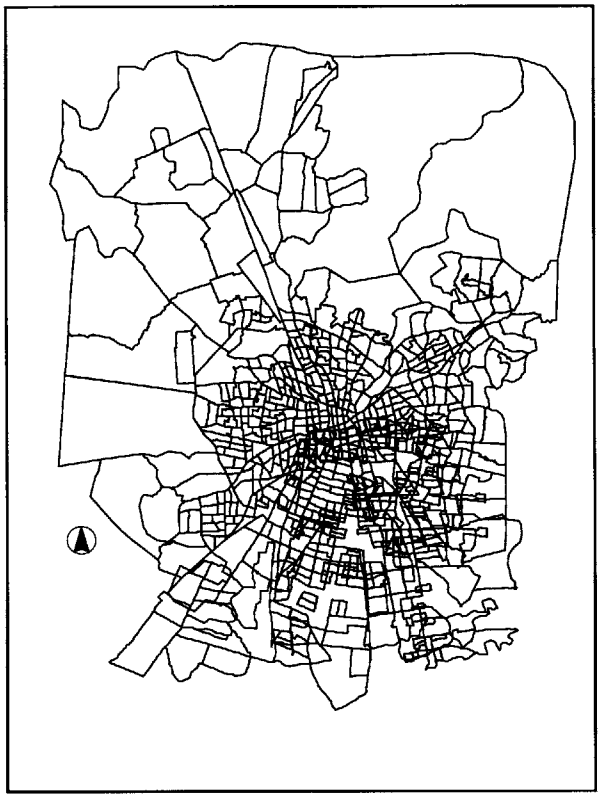
The study area corresponded to 38 "comunas" or municipalities, of which 32 belong to the "Provincia" of Santiago, and the other 6 to the adjacent "Provincias" of Maipo, Cordillera and Chacabuco. These municipalities can be grouped in 6 sectors: the "Centro," which is the Central Business District (CBD), the East and South-East where the wealthiest households live, especially in the former, and the North, East and South, as shown in Figure 5-1. In Appendix A is presented a detailed description of the socioeconomic characteristics of the population by Sector.

The population in the study area is over five million people on a surface of over 400 square kilometers. For this city, Sectra (2003) obtained 15,537 randomly selected household interviews that correspond to 59,763 persons and 153,413 trips. Other characteristics of the database are presented in Appendix A.


Source EOD 2001 Final Report Sectra (2003)

**Figure 5-1 Study Area EOD 2001 Santiago de Chile: Municipalities and Sectors**

Beyond the political divisions of the study area in 6 sectors and 38 municipalities, Sectra (2003) divided the area into 789 analysis zones as is shown in Figure 5-2.



Source EOD 2001 Final Report Sectra (2003)

**Figure 5-2 EOD 2001 Zones**

The other source of information used in this thesis was obtained from a database of land-use by municipality, registered in the *Servicio de Impuestos Internos* (SII), the Chilean equivalent to the Internal Revenue Service of the US. The information is divided into 19 different land-uses including housing, industry, education and commerce. For each land-use type in the database is reported the available the number of constructions and the built square meters.

# 5.2 Base Residential Location Model

## 5.2.1 Modeling Sample Definition

Using the provided data, the first step was the estimation of a residential location model to compare the effects of the application of the control-function method to correct for endogeneity.

To ensure that modeled households were located in places that maximize their utility, only renters were considered in the modeling sample. This assumes that renters face lower cost barriers in moving and then are more probably optimally located. A second reason to consider renters is that there is a better representation of the housing costs for them in this survey through rental cost, a value that has to be estimated in the case of home owners. This sample, defined as the *renters' sample*, was used for the stratification of the modeling households and consists of 1,228 observations.

Additionally, the effect of considering only *recently-moved* renters in the modeling sample was analyzed. Trading off between the resulting sample size and the probability of getting optimally located households, it was decided to define the *recently-moved* limit as below 2 years of residence in the current location because no significant differences in the results were found below this limit[12].

Another exclusion rule was to eliminate the households where the rent cost represented a high share of the reported income. This is under the assumption that, if the rent is too high (in some cases was even greater than the reported income), it could be a mistake in the reported rent or income or, if not, that the household is helped by some relatives to cover the rent, putting in doubt their free choice of the place to live relative to their income. The limit was established as one standard deviation to the right on the

[12] If the *recently-moved* limit is established below 6 months of residence, the result of the base model no longer shows the positive cost parameter for the higher income strata, as occurs with the 2 year limit case. However, in this case the sample size is too small (less than 200) and, even here, the change in the parameters with the control-function correction is also observed, that is, the price endogeneity problem is proved to exist, but it is not as bad as for the 2 years limit. This could be an indication that the inertia bias is positively correlated with the omitted attributes bias described in this thesis or that the small size of the sample causes other sources of problems.

observed sample, which corresponds to households where the ratio between the rent and the income is no larger than 0.5.

Finally, households where the income was not reported by the surveyed family but imputed by the Consultant who processed the survey, and households that were located in suburban "Comunas" for which no land-use information from the SII was available, were excluded also. After the application of all these filters, the *modeling sample* consists of 630 households.

## 5.2.2 Model Specification and Results

Using the modeling sample described above, a residential location model was constructed by defining for each observation 10 additional alternatives randomly selected from the rest of the sample. As stated before, following McFadden (1978), if it is assumed that the error is *iid* Extreme Value, consistent estimates of the model parameters can be obtained.

More elaborate error structures for these models are left for future research. Some alternatives could be to consider adjacent zones correlation, as was done by Bhat and Guo (2004), or nests with similar variables such as apartments or condominiums.

Some relevant information such as transportation levels of service or activity distributions through city to built accessibility measures was not available. These variables were approximated by the simple straight distance between the dwelling unit location and the workplace. Despite these limitations, the model satisfactorily accomplishes the objective of testing the proposed methods to correct for price endogeneity in residential location.

The following paragraphs describe the explanatory variables that were found, after a process of hypothesis and statistical testing, to be relevant in the utility function of a MNL model of residential location. Additionally an interpretation is given for the respective estimated parameters, which can be found in the first column of Table 5-2, at the end of this chapter.

## Apartment Dummy

This variable is a dummy for the dwelling unit type that takes the value 1 when it is an apartment. The additional classes in the sample are Condominium and House.

## Apartment Dummy for Large Households

This variable is the same as last one, but now multiplied by a dummy that takes value 1 for households (families) comprised by more than four members. Combined with the estimated parameter of the previous dummy reported in the second column of Table 5-2, these estimated parameters indicate that small families tend to favor, all other things equal, apartments instead of houses or condominiums. This effect is the inverse for large families, which can be explained (accepting that larger families are the ones that have children); by the extra value that having a house has in raising children.

## Big Apartment Buildings Dummy for High Income Households

This variable is a dummy that takes value 1 when the building where the apartment is located is of more than 4 floors. This dummy was found significant and positive only for households with monthly income over $650,000 Chilean pesos (around 1000 US Dollars), which is the highest 25% income group of the renters' sample. An interpretation of this could be that some correlation exists between the newest and the highest buildings, an effect that is stronger in the wealthiest neighborhoods.

## Condominium Dummy for High Income Households

This variable is a dummy for the dwelling unit type that takes the value 1 when it is a Condominium. This dummy was found significant and positive only for households with monthly income over $650,000 Chilean pesos. For other types of households this dummy was not significant, meaning that houses are perceived as equal to Condominiums. This can be explained because Condominiums do offer quality attributes that are different from those ones of single houses only for high income oriented types, where it can be found communitarian services such as golf fields, parks or swimming pools. Lower income oriented Condominiums are maybe just a different brand for a group of houses.

## Cost divided by Income

This variable is the ratio between the dwelling unit monthly rental cost and the household monthly income. As expected, the parameter is negative, meaning that, everything else equal, households prefer cheaper dwelling units.

## Cost divided by Income, Non-Low Income

This variable is the same as the last one, but now multiplied by a dummy that identifies the non-low income households, defined as having a monthly income over $230.000 Chilean pesos (around 400 US Dollars), which is the 25% lowest income group of the renters' sample. As expected, this parameter is positive, meaning that, as the household income increases, the marginal utility of income, which is the sum of this parameter and the previous one, decreases. Note that in this case, the total cost parameter is still negative, as is intuitively expected.

## Cost divided by Income, High Income

This variable is the same as the last one but in this case it was multiplied by a dummy that identifies the high income households, defined as the ones that have a monthly income over $650.000 Chilean pesos. As expected, this parameter is positive, meaning that, as the income increases, the marginal utility of income, which is the sum of this and the previous two parameters, decreases. However, **note that in this case, the total cost parameter is positive (+ 0.96)**, which is against intuition and theory because it would mean that, everything else equal, high income people prefer apartments with higher rents. This result makes the model useless for policy analysis and in a typical study it would be attributed to poor data variability or other issues and probably solved by not reporting a specific parameter for this income stratum, but a general one where this effect is lost.

## Absolute Difference between Household Income and Average Zonal Income

This variable measures the observed tendency of clustering in neighborhoods of similar socioeconomic characteristics, in this case income, by estimating the absolute

value of the difference between the household income and the average income observed in the zone of the dwelling unit alternative. As expected, the estimated parameter is negative, confirming the clustering hypothesis found also in other studies.

### Zonal Percentage of Head of the Household with High Education

This variable is the percentage of households that have a head of the household (role that is defined by the household itself when it is surveyed) with more than High School education in the zone where the dwelling unit belongs.

### Zonal Percentage of Head of the Household with High Education, Dummy High Education

This variable is the same as the last one but now multiplied by a that takes value one if the head of the household of the modeled household does have more than High School education. The estimated results of this variable combined with the parameter of the last one indicate that households prefer, other things equal, zones where the educational level is similar to the one of the household, what could be another expression of the clustering hypothesis.

### Distance to Work for the Head of the Household

This variable is the distance between each available housing alternative and the declared workplace of the head of the household of the modeled household. As expected, the estimated parameter of this variable is negative, meaning that households prefer residential locations that are nearer to the workplace of the head of the household.

### Distance to Work for the Head of the Household, Dummy One Worker

This variable is the same as the last one but now multiplied by a dummy that takes value 1 if the household has only one worker and zero otherwise. The estimated parameter is also negative, indicating that, when the head of the household is the only worker (the head of the household), her commuting time is more relevant in the

residential location decision than when the effect on a second worker has to be evaluated in the residential location decision.

## Percentage of Housing Square Meters by "Comuna"

This variable was included as a measure of how housing oriented an area is, with the idea that, as more housing in an area is present, more competition for similar attributes would exist, and thus a better differentiation and quality. The estimated parameter is positive as expected.

## Dummy West Area

This variable is a dummy indicating if the dwelling unit is located in the west area of the city, comprised by sectors North, East and South (Figure 5-1). The reason for this is that the other part of the city is the wealthiest one (see appendix A), not only correlated with a better level of public services such as sidewalks, parks, safety and security, but also charged with a positive bias in things like job seeking and even some level of racial differentiation. As expected, everything equal, households are more inclined not to live in the west area of the city, because the estimated parameter is negative.

# 5.3 Corrected Model Using Control-Function Method

The next step was the application of the control-function method to correct for price endogeneity as was described in the preceding chapters. The instruments for the price used in this case $\overline{p}_n$ were built, following a proposition of Hausman (1997), as the average price of the other dwelling units that are in the same "Comuna". Within this setting, the price equation model (5-1) was estimated doing OLS of the prices on the instrument and additional dummies indicating the type of dwelling unit under analysis.

**(5-1)** $$p_{in} = \alpha_0 + \alpha_1 * \overline{p}_n + \alpha_2 * Apartment + \alpha_3 * Condo + \delta_{in}$$

The objective of this price equation is to correct for the endogeneity problem and not to make a precise forecast of the dwelling unit's price. More elaborated forms of this equation, such as the one proposed by Martinez and Henriquez (2005) and that was shown in (2-1), are left for future research.

The estimated results of model (5-1) are shown in the Table 5-1 where it can be seen that the adjustment of the model is acceptably high (Hananh and Hausman, 2003), with an adjusted $R^2$ of 40%. It can be seen also that the instrument $\overline{p}_n$ is highly correlated with the dwelling unit price.

**Table 5-1 Price Equation Instrumental Variables OLS**

| *Variables* | Parameters | |
|---|---|---|
| **Intercept** | 7,633.4 | (1.7103 ) |
| $\overline{p}_n$ | 0.94077 | (26.777 ) |
| **Apartment Dummy** | -3,686.1 | (-0.75210) |
| **Condominium Dummy** | 50,899 | (2.4020 ) |
| ***Adjusted $R^2$*** | 0.39804 | |
| ***S. Size*** | 1228 | |

t-test in brackets

As it was explained before, the estimated parameters of model (5-1) are used to calculate a new variable $\hat{\mu}$, as is shown in expression (5-2), which is added to the residential location model as an additional variable. To be consistent with the formulation

used in the original model $\hat{\mu}$ was divided by the household income to enter the utility function.

**(5-2)** $$\hat{\mu} = p_{in} - \hat{\alpha}_0 + \hat{\alpha}_1 * \overline{p}_n + \hat{\alpha}_2 * Apartment + \hat{\alpha}_3 * Condo$$

The results of the estimation of the corrected residential location choice model are shown in the third column of Table 5-2, at the end of this chapter. The estimated parameters in this case are fairly similar to the ones of the base case. The only important difference, as expected, is related to the parameter of the rent, which is corrected downwards because of endogeneity. This correction is big enough to solve the problem of having a positive rent parameter for the higher income strata.

It can also be noted that the addition of the control-function variable did improve the statistical adjustment of the model, which can be noted by the significant increase in the likelihood. This can be formally tested by performing a likelihood radio test, and can be also verified by the fact that the estimated parameter of $\hat{\mu}$ is significantly different from zero at a 95% of confidence level.

The final step was the application of a non-parametric correction for the fact that, if the endogeneity problem does not occur at an individual level as it was implicitly assumed, the parameters of the model with the control-function correction are biased.

This objective was accomplished by considering not only the own fitted errors as additional variables, but also the average fitted errors of the other dwelling units that belong to the same "Comuna", variable that was defined as $\hat{\mu}2$. This procedure was used with the purpose of capturing the correlation effect between near dwelling units, following what was proposed by Petrin and Train, (2004) and that is described in (4-2).

Additionally, in this last procedure, the parameters of the control-function were allowed to be random in order to address, as was also proposed in Petrin and Train (2004), the misspecifications that arise because of not considering the zonal effect. This last task was developed using the routine for estimating Mixed Logit models in BIOGEME, considering 500 Halton Draws (Train, 2000)[13].

---

[13] More draws were used in this case because the standard deviation of $\hat{\mu}$, which is in any case statistically equal to zero, happens to be negative due to numerical problems if less draws are used.

The results are shown in the last column of Table 5-2, where it can be noted that the application of this procedure did not improve the statistical adjustment of the model, but slightly corrected the price parameters downward, that is, increasing the correction for the endogeneity problem.

**Table 5-2 Residential Location Models Using EOD 2001 Santiago de Chile**

| Variables | Base Residential Location Model | | Control Function Linear $\hat{\mu}$ | | Control Function Random $\hat{\mu}, \hat{\mu}2$ | |
|---|---|---|---|---|---|---|
| **Apartment D** | 0.140 | (1.16 ) | 0.130 | (1.08 ) | 0.125 | (1.03 ) |
| **Apartment D LF** | -0.840 | (-3.25) | -0.856 | (-3.28) | -0.857 | (-3.28) |
| **Apt. Floors > 4 D HI** | 0.263 | (1.18 ) | 0.278 | (1.25 ) | 0.280 | (1.26 ) |
| **Condominium D HI** | 1.11 | (2.09 ) | 1.27 | (2.39 ) | 1.28 | (2.41 ) |
| **Cost/Income** | -2.33 | (-5.91) | -4.86 | (-6.80) | -5.02 | (-5.44) |
| **Cost/Income D I > I1** | 1.06 | (1.75 ) | 1.70 | (2.59 ) | 1.67 | (2.53 ) |
| **Cost/Income D I > I2** | 2.23 | (2.65 ) | 1.97 | (2.42 ) | 1.96 | (2.39 ) |
| **Diff with Zonal Ave. Income** | -0.630 | (-6.83) | -0.459 | (-4.68) | -0.462 | (-4.68) |
| **% hHH with HE by zone** | -1.20 | (-3.35) | -0.908 | (-2.60) | -0.903 | (-2.58) |
| **% hHH with HE by zone D HE hHH** | 2.89 | (7.42 ) | 2.82 | (7.27 ) | 2.82 | (7.25 ) |
| **Distance to Work hHH** | -0.119 | (-9.14) | -0.125 | (-9.45) | -0.125 | (-9.42) |
| **Distance to Work D OW** | -0.0299 | (-1.55) | -0.0299 | (-1.54) | -0.0300 | (-1.54) |
| **% of housing SM by Comuna** | 1.40 | (4.88 ) | 1.45 | (5.10 ) | 1.46 | (5.11 ) |
| **West Area D** | -0.368 | (-1.90) | -0.429 | (-3.55) | -0.434 | (-3.55) |
| $\hat{\mu}$ | | | 2.50 | (4.38 ) | 2.77 | (2.60 ) |
| **Standard Deviation** $\hat{\mu}$ | | | | | 0.0000931 | (0.000131 ) |
| $\hat{\mu}2$ | | | | | 2.28 | (0.254 ) |
| **Standard Deviation** $\hat{\mu}2$ | | | | | 4.34 | (0.355 ) |
| **Sample Size** | 630 | | 630 | | 630 | |
| ***LL(0)*** | -1501.27 | | -1501.27 | | -1501.27 | |
| **Final *LL*** | -1222.80 | | -1210.54 | | -1210.49 | |
| **Adj** $\rho^2$ | 0.176 | | 0.184 | | 0.182 | |

D; Summy, 1 if valid, zero if not

LF: Large Family

HI: High Income the 25% high of the sample, over $650.000 (I2)by month

MI: Middle Income. Over 230000 (I1)and below 650000 (I2) by month

LI: Low Income the 25% low of the sample. Below $230000 (I1) by minth

SM: Square Meters

OW: One worker Family

HB: "House Boss", defined by the surveyed family

t-test in bracklets

*LL* Log-likelihood

Adj. $\rho^2$ = 1 – (final *LL* - #Attributes)/*LL(0)*

$\hat{\mu}$ household-alternative own fitted error of the price equation

$\hat{\mu}2$ Average fitted error of other dwelling units in the same "Comuna"

It can be noted also that, despite the parameter of $\hat{\mu}$ was allowed to be random, its empirical standard deviation is non-significantly (at any confidence level) different to

zero, what means that this parameter is not random. The parameter and the standard deviation of the additional element of the control-function ( $\hat{\mu}2$ ) are also non-significant, but not as small as the variance of the first one.

These results can be an indication that the zonal effect does not exists, and thus the simplest version of the control-function method is appropriate, or that the non-parametric corrections methods used are not appropriate to address the misspecification problem. The clarification of this question is left for future research.

# Chapter 6

# Synthesis, Conclusions and

# Recommendations for Further Research

## 6.1 Synthesis

This thesis searched for and tested available methods to treat endogeneity in discrete choice models of residential location. These tasks were developed by first reviewing the literature on residential location modeling and on endogeneity in linear and discrete choice models. Four methods to correct for endogeneity in discrete choice models were identified: the control-function method, the traditional instruments method, the Matzkin method and the BLP method.

Then, the methods were tested using four Monte Carlo experiments representing different error structures and quality of instruments. The results showed that the control-function method (Petrin and Train, 2004) is the most promising one to address endogeneity in residential location models.

Finally, the control-function method was satisfactorily applied to correct for endogeneity in a model of residential location choice that was estimated using real data from Santiago de Chile.

# 6.2 Conclusions

Three main conclusions were obtained from the development of this thesis. The first one is that, definitely, price endogeneity is a problem in discrete choice models of residential location. This follows not only from the previous studies cited in Section 2.4 where questionable results that can be attributed to price endogeneity were reported, but also from the estimation of the model of residential location of Santiago de Chile reported in Chapter 5.

As it was theoretically derived in 3.1.1, the endogeneity problem in the Santiago model produced an upward bias for the price parameter, strong enough to cause the price parameter for the wealthiest households to be positive. In a typical study, without being aware of the endogeneity effect, this result would be attributed to poor data variability or other issues and probably solved by not reporting a specific parameter for this income stratum, but a general one where this undesirable result is lost. Instead, in this thesis, that unintuitive result was satisfactorily solved with the application of the appropriate corrections.

The second conclusion is that the control-function method is the most suitable way to correct for endogeneity in discrete choice models of residential location. This follows from its applicability using standard estimation packages and from the fact that the better alternative to it, the BLP method, fails to recover consistent estimates when endogeneity occurs at an individual level, a characteristic that is very likely to occur in residential location modeling.

The final conclusion is about the characteristics of the instrumental variables in residential location modeling. First of all, by the application of the Monte Carlo experiments, this thesis found that, if the instruments are not correlated enough with the endogenous variables, the bias of the corrected model would be even greater than the one of the uncorrected model. This is concordant with the results shown by Hahn and Hausman (2003) and points out that finding appropriate instruments is crucial in the usage of the correction methods. Also, by the estimation of the model with real data in Chapter 5, it was found that an appropriate instrument for endogenous price of dwelling

units in discrete choice models of residential location is, following an idea of Hausman (1997), the average price of other dwelling units in the area.

## 6.3 Recommendations for Further Research

In the development of this thesis relevant research questions remained unanswered and thus, should be addressed in future studies. The first is the specification of the control-function that corresponds to the expected value of the utility errors conditional in the price equation errors. How to identify the cases when the conditions under which it can be expressed as a linear function of the own error hold and how to correct the model when these conditions do not hold, are not completely clear.

The necessity of research in this area is reinforced by the fact that, rather surprisingly, the third Monte Carlo experiment showed that no significant improvements to the control-function method were found when applying the non-parametric corrections suggested by Petrin and Train (2004), despite the fact that in this case, by construction, the control-function method does have a specification problem.

The second question that remained open is the derivation of a formal demonstration of the relationship between the instruments proposed by Matzkin(2004) and the control-function method. Apparently, the control-function method can be viewed as a way to built auxiliary variables of the type proposed by the Matzkin (2004).

Also about the Matzkin method, it would be interesting to explore different auxiliary variables to use in this method within the framework of residential location. Potential candidates are assessment price or characteristics of other dwelling units in the neighborhood. It can be even considered that synthetic instruments for price could be built just by adding an exogenous perturbation to actual prices.

Furthermore, it would be interesting to address in the future the apparent similarities between the theory behind the Matzkin (2004) and Walker (2001). It seems that Walker's method is more general than Matzkin's method because the former allows for the specification of structural equations, which would lead to a better representation of individual's behavior with available data.

Another area of research related with the application of the control-function method is the exploration of the possibility of enhancing the efficiency of the estimated parameters by performing the estimation of the choice and the price model simultaneously by using, for example, a latent variable approach (Walker, 2001).

Another open question identified is the study of the convenience of using theoretically-derived price equations, such as expression (2-1) which was proposed by Martinez and Henriquez (2005). Having better specified price equations should lead to more efficient corrections of the endogeneity problem.

Another interesting area of research not explored is the investigation of the presence and treatment of endogeneity between the workplace and the residential location decisions and between family members, as was done by Blundell and Powell (2004).

Finally, the development of a general equilibrium theory for urban systems, and the study of the conditions for which it holds, is an interesting task that would help, not only in the specification of better models of residential location, but also in understanding their impact and evaluating the real possibilities of implementing policies to shape the urban system. In this case, the work of Martinez and Henriquez (2005) is a good reference point that would be improved if it can be joined with the micro-simulation approach developed by Waddell and Borning (2004).

# Appendix A

# Description Santiago 2001 Mobility Survey

**Table A-1 Vehicle Possession in Santiago City**

| Vehicles | Households | Persons | Vehicles/ Household | Persons/ Household | Veh/ 1000 People |
|---|---|---|---|---|---|
| 855,057 | 1,513,938 | 5,772,617 | 0.56 | 3.81 | 148.1 |

Source: Sectra (2003)

**Table A-2 Households Distribution as a Function of Dwelling Unit Ownership**

| | Dwelling Unit Ownership | | | | | Total |
|---|---|---|---|---|---|---|
| | Owned | Rented | Institution | Relatives | N/A | |
| **Total** | 1,117,423 | 284,015 | 9,334 | 102,178 | 988 | 1,513,938 |
| **(%)** | 73.8 | 18.8 | 0.6 | 6.8 | 0.1 | 100 |

Source: Sectra (2003)

**Table A-3 Population Households and Motorization Rate**

| Sector | Vehicles | HH | People | Vehicles/ HH | People/ HH | Veh/ 1000 People |
|---|---|---|---|---|---|---|
| **North** | 90,982 | 201,466 | 822,763 | 4.08 | 0.45 | 110.58 |
| **West** | 151,935 | 314,262 | 1,240,814 | 3.95 | 0.48 | 122.45 |
| **East** | 287,902 | 258,805 | 876,462 | 3.39 | 1.11 | 328.48 |
| **"Centro"** | 28,427 | 78,936 | 230,674 | 2.92 | 0.36 | 123.23 |
| **South** | 121,665 | 334,586 | 1,317,826 | 3.94 | 0.36 | 92.32 |
| **S-East** | 174,145 | 325,883 | 1,284,079 | 3.94 | 0.53 | 135.62 |
| **Total** | 855,057 | 1,513,938 | 5,772,617 | 3.81 | 0.56 | 148.12 |

Source: Sectra (2003)

**Table A-4 Households Distribution by Income**

| Sector | Income (thousand of Chilean pesos. $ November 2001) | | | | | | | | | | | | Total | (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | > 5,000 | | 1,600-5,000 | | 450-1,600 | | 280-450 | | 150-280 | | < 150 | | | |
| | HH | % | HH | % | HH | % | HH | % | HH | % | HH | % | | |
| North | 97 | 2 | 2,323 | 3 | 44,904 | 10 | 60,145 | 16 | 58,203 | 16 | 35,794 | 15 | 201,466 | 13 |
| (%) | 0.0 | | 1.2 | | 22.3 | | 29.9 | | 28.9 | | 17.8 | | | |
| West | 0 | 0 | 2,101 | 3 | 81,807 | 18 | 90,203 | 24 | 86,460 | 23 | 53,691 | 23 | 314,262 | 20 |
| (%) | 0.0 | | 0.7 | | 26.0 | | 28.7 | | 27.5 | | 17.1 | | | |
| East | 4,026 | 90 | 65,021 | 81 | 126,78 | 29 | 31,601 | 8 | 19,599 | 5 | 11,775 | 5 | 258,805 | 17 |
| (%) | 1.6 | | 25.1 | | 49.0 | | 12.2 | | 7.6 | | 4.5 | | | |
| Centro | 0 | 0 | 1,754 | 2 | 31,272 | 7 | 23,938 | 6 | 14,121 | 4 | 7,851 | 3 | 78,936 | 5 |
| (%) | 0.0 | | 2.2 | | 39.6 | | 30.3 | | 17.9 | | 9.9 | | | |
| South | 245 | 5 | 1,720 | 2 | 64,505 | 15 | 83,421 | 22 | 108,939 | 29 | 75,757 | 32 | 334,586 | 22 |
| (%) | 0.1 | | 0.5 | | 19.3 | | 24.9 | | 32.6 | | 22.6 | | | |
| S – East | 134 | 3 | 7,804 | 10 | 95,457 | 22 | 88,930 | 24 | 85,512 | 22 | 48,047 | 21 | 325,883 | 22 |
| (%) | 0.0 | | 2.4 | | 29.3 | | 27.3 | | 26.2 | | 14.7 | | | |
| Total | 4,502 | | 80,724 | | 444,728 | | 378,236 | | 372,832 | | 232,915 | | 1,513,938 | |
| (%) | 0.3 | | 5.3 | | 29.4 | | 25.0 | | 24.6 | | 15.4 | | 100 | |

Source: Sectra (2003)

**Table A-5 Vehicle Possession by Income Level (thousand of Chilean Pesos 2001)**

| Income Level | Number of Vehicles | | | | Total |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 or more | |
| > 5,000 | 313 | 1,104 | 868 | 2,217 | 4,502 |
| 1,600-5,000 | 3,820 | 26,391 | 36,186 | 14,328 | 80,724 |
| 4,500-1,600 | 136,191 | 229,246 | 65,391 | 13,900 | 444,728 |
| 280-450 | 226,758 | 135,619 | 13,955 | 1,904 | 378,236 |
| 150-280 | 284,116 | 84,492 | 3,940 | 284 | 372,832 |
| < 150 | 207,626 | 22,973 | 2,226 | 90 | 232,915 |
| All | 858,824 | 499,825 | 122,566 | 32,723 | 1,513,938 |

Source: Sectra (2003)

**Table A-7 Vehicle Possession by Sector**

| Sector | Number of Vehicles | | | | Total |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 or more | |
| North | 124,271 | 66,384 | 9,184 | 1,628 | 201,466 |
| West | 187,824 | 106,143 | 16,142 | 4,154 | 314,262 |
| East | 73,556 | 109,707 | 55,722 | 19,820 | 258,805 |
| "Centro" | 55,218 | 20,031 | 2,942 | 745 | 78,936 |
| South | 232,621 | 85,908 | 13,936 | 2,121 | 334,586 |
| South-East | 185,335 | 111,651 | 24,641 | 4,256 | 325,883 |
| Total | 858,824 | 499,825 | 122,566 | 32,723 | 1,513,938 |

Source: Sectra (2003)

# References


Bayer, P., McMillan, R., and Reuben, K. (2004). "Residential Segregation in General Equilibrium." **Working Paper, Department of Economics Yale University**.

Ben-Akiva M.E. and Lerman, S. R. (1985). **Discrete Choice Analysis. : Theory and Application to Travel Demand,** The MIT press, Cambridge MA.

Ben-Akiva, M. and Boccara, B. (1995). "Discrete choice models with latent choice sets." **International Journal of Research in Marketing** 12: 9-24

Ben-Akiva, M. Bowman, J. and Gopinath, D. (1996). "Travel Demand Model System for the Information Era." **Transportation** 23: 241-266.

Berry, S. (1994). "Estimating Discrete Choice Models of Product Differentiation." **RAND Journal of Economics**, 25: 242-262.

Bhat, Ch. and Guo J. (2004). "A mixed spatially correlated logit model: formulation and application to residential choice modeling." **Transportation Research Part B.** 38. 147-168.

Bierlaire, M., Bolduc, D. and Godbout, M. (2004). "An introduction to BIOGEME (Version 1.0)," **URL:roso.epfl.ch/mbi/biogeme/doc/tutorial.pdf**

BLP (1995).Berry, S., Levinsohn, J and Pakes, A. "Automobile prices in market equilibrium." **Econometrica** 63: 841-889.

Blundell, R. and Powell J. (2004). "Endogeneity in semi-parametric binary response models." **Review Of Economic Studies** 71 (3): 655-679.

Bresnaham, T. (1997). "The apple-cinnamon Cheerios war: Valuing new goods, identifying market power, and economic measurement." **Working paper, Stanford University**.

Ferreira, F. (2004). "You can Take it with you: Transferability of Proposition 13 Tax Benefits, Residential Mobility, and Willingness to Pay for Housing Amenities." **Working Paper No. 72. Center for Labor Economics, University of California, Berkeley**.

Greene, W. (2003).**Econometric Analysis**. 5th edition, Prentice Hall.

Guo, J. (2004). "Addressing Spatial Complexities in Residential Location Choice Models." **PhD Dissertation**, University of Texas at Austin.

Hahn, J. and Hausman J. (2003), "Weak Instruments: Diagnosis and Cures in Empirical Econometrics." **The American Economic Review** 93 (2): 118-128 (8).

Hausman, J. (1997). "Valuation of new goods under imperfect competition," in R. Gordon and T. Bresnaham, eds. **The economics of new goods**, University of Chicago Press, Chicago.

Hausman, J. (1983). "Specification and Estimation of Simultaneous Equations Models." In Z. Griliches and M. Intriligartor (eds.), **Handbook of Econometrics.** 1: 391-448.

Hausman, J. (1978), "Specification Tests in Econometrics." **Econometrica** 46: 1251-1272

Heckman, J. (1978). "Dummy endogenous variables in a simultaneous equation system." **Econometrica** 46: 931-959

Lerman, S. (1975). "A disaggregate behavioral model for urban mobility decisions." **PhD Dissertation**, Massachusetts Institute of Technology.

Levine, J. (1998). "Rethinking Accessibility and Jobs-Housing balance." **Journal of American Planning Association**, 64(2): 133-149.

Martinez, F. (1996). "MUSSA: A Land-use Model for Santiago City." **Transportation Research Record 1552**: Transportation Planning and Land-use at State, Regional and Local Levels: 126-134.

Martinez, F. (1995). "Access: The transport-land use Economic Link." **Transportation Research B**. 29 (6): 457-470.

Martinez, F. and Henriquez, R. (2005). "The RB&SM: A Random Bidding and Supply Land-use Equilibrium Model." **Working paper Universidad de Chile**

Matzkin, R. (2004). "Unobservable instruments." **Working paper Northwestern University**.

McFadden, D.L. (1978). "Modeling the choice of residential location." In Karlqvist et. al. (eds), **Spatial Interaction Theory and Planning Models**, North-Holland, Amsterdam: 75-96.

Nevo, A (2001). "Measuring Market Power in the ready to eat Breakfast Cereal Industry." **Econometrica** 69: 307-342

Petrin, A. and Train, K. (2004). "Omitted product attributes in discrete choice models." **National Bureau of Economic Research**, Working paper.

Pickrell, D. (1999). "Transportation and Land-use." In, Gomez-Ibanes, Tye and Winston eds. **Essays in Transportation Economics and Policy.** Washington, USA.

Quigley, J. (1976). "Housing Demand in the short run: An analysis of politomous choice," **Regional Science and Urban Economics**, 15: 41-63.

Sectra (2003).**Santiago de Chile 2001 Mobility Survey: Final Report**. MIDEPLAN, Chile.

Sermonss, M and Koppelman, F. (2001). "Representing the differences between female and male commute behavior in residential location choice models." **Geography** 9: 101-110.

Train, K. (2003).**Discrete Choice Methods with Simulation**. Cambridge University Press.

Train, K. (2000). "Halton Sequences for Mixed Logit." **Working Paper No. E00-278, Department of Economics, University of California, Berkeley**.

Train, K., and Winston, C. (2004). "Vehicle Choice Behavior and the Declining Market Share of U.S. Automakers," **Working paper, Economics, University of California, Berkeley**.

Trajtenberg, M. (1989). "The welfare analysis of product Innovations, with an Application to Computed Tomography Scanners," **Journal of Political Economy**, 97: 444-479.

Varian, H. (1992) **Microeconomic Analysis. 3rd ed.** Norton eds. New York..

Vilas-Boas, J. and Winer, R. (1999). "Endogeneity in Brand Choice Models." **Management Science**, 45 (10): 1324-1338.

Vovsha, P. (1997). "The Cross-Nested Logit Model: Application to mode choice in the Tel-Aviv Metropolitan Area." **Transportation Research Record**: 1607, 6-15.

Waddell (1996). "Accesibility and Residential Location: The Interaction of Workplace, Residential Mobility,Tenure and Location Choices. **Lincoln Land Institute TRED Conference.**

Waddell (1992). "A Multinomial Logit Model of Race and Urban Structure." **Urban Geography**, 13(2): 127-141.

Waddell, P. and Borning, A. (2004). "A Case Study in Digital Government: Developing and Applying UrbanSim, a System for Simulating Urban Land-use, Transportation, and Environmental Impacts." **Social Science Computer Review,** 22 (1):37-51.

Walker, J. (2001). "Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables." **PhD Dissertation**, Massachusetts Institute of Technology.

Weisbord G., Ben-Akiva M. and Lerman M. (1980). "Tradeoffs in residential location decisions: Transportation versus other factors." **Transportation Policy and Decision-Making**, 1 (1): 13-26.

Wooldridge, J.W. (2002). **Econometric Analysis of Cross-Section and Panel Data**. MIT Press.