

Endogeneity and Sampling of Alternatives in Spatial Choice Models

by
Cristian Angelo Guevara-Cue

Ingeniero Civil, Magíster, Universidad de Chile
(2000)

Master of Science, Massachusetts Institute of Technology
(2005)

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

Doctor of Philosophy


at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY


September 2010

© 2010 Massachusetts Institute of Technology. All Rights Reserved

Author


Cristian Angelo Guevara-Cue
Department of Civil and Environmental Engineering
June 23rd, 2010

Certified by


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

Accepted by


Daniele Veneziano
Chairman, Departmental Committee for Graduate Students

Endogeneity and Sampling of Alternatives in Spatial Choice Models

by

Cristian Angelo Guevara-Cue

Submitted to the Department of Civil and Environmental Engineering
on June 23rd, 2010, in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in the field of Transportation

Abstract

Addressing the problem of omitted attributes and employing a sampling of alternatives strategy, are two key requirements of practical spatial choice models. The omission of attributes causes endogeneity when the unobserved variables are correlated with the measured variables, precluding the consistent estimation of the model parameters. The consistent estimation while sampling alternatives in non-Logit models has been an open problem for three decades. This dissertation is concerned with both the endogeneity and the sampling of alternatives in non-Logit models, two problems that have hindered the development of suitable modeling tools for urban policy analysis, but have been neglected in spatial choice modeling.

For the problem of endogeneity, this research applies, enhances, adapts, and develops efficient and tractable methods to correct and test for it in models of residential location choice, and also develops novel methods to validate the success of the correction. For the problem of sampling of alternatives in non-Logit models, this study develops and demonstrates a novel method to achieve consistency, relative efficiency, and asymptotic normality when the underlying model belongs to the Multivariate Extreme Value class. This development allows for the estimation of spatial choice models with more realistic error structures. Monte Carlo experiments and real data from Lisbon, Portugal, are employed to illustrate the significant benefits of these novel methods in correcting for endogeneity and addressing sampling of alternatives in non-Logit models, with specific reference to urban policy analysis.

Thesis Supervisor: Moshe E. Ben-Akiva

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

Thesis Committee

Moshe E. Ben Akiva (Chair)
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology

Steven R. Lerman
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology

P. Christopher Zegras
Department of Urban Studies and Planning
Massachusetts Institute of Technology

Denis E. Bolduc
Département d'économique
Université Laval

Stephane Hess
Institute for Transport Studies
University of Leeds

Paul Waddell
Department of City and Regional Planning
University of California, Berkeley

Acknowledgments

I am in debt to several people who provided assistance to me in the course of this research. Professor Moshe Ben-Akiva, my thesis supervisor, gave me valuable advice on various aspects throughout the research and shared his knowledge, experience and novel ideas with me. The members of my committee, Professors Steven Lerman, Christopher Zegras, Denis Bolduc, Stephane Hess and Paul Waddell, provided many important ideas and suggestions. Professors Kenneth Train and Joan Walker gave me helpful comments and advice. Tina Xue and the instructors from MIT's Writing Center helped me editing the thesis. The assistance I received from Luis Martinez and Weifeng Li, in obtaining, understanding and processing the data, was also of great value. Finally, the submission of the final version of this thesis would be literally impossible without the help from Travis Dunn.

The funding for this research came in part from the generous support of the Portuguese Government through the Portuguese Foundation for International Cooperation in Science, Technology and Higher Education, undertaken by the MIT-Portugal Program. Additional funding for this research came from the Martin Family Society of Fellows for Sustainability, from CONICYT (Comisión Nacional de Investigación Científica y Tecnológica, Gobierno de Chile), and from Universidad de los Andes, Chile.

I would also like to acknowledge the intangible help from the various friends I had the luck to meet in these two very intense years at MIT; the unfailing support of my mother and sisters, Amparo, Monica and Gabriela; as well as the constant cheer and love I got from my three beautiful kids, Diego, Bruno and Matilde, who are always my priority.

Finally, I wish to acknowledge the invaluable support of my wife, Erika, to whom this thesis is dedicated.

Table of Contents

| | |
|---|-----------|
| CHAPTER 1 INTRODUCTION..... | 11 |
| 1.1 MOTIVATION | 11 |
| 1.2 OBJECTIVES AND METHODOLOGY | 13 |
| 1.3 MODELING FRAMEWORK..... | 13 |
| 1.4 ENDOGENEITY | 15 |
| 1.5 SAMPLING OF ALTERNATIVES IN MEV MODELS | 17 |
| 1.6 CONTRIBUTIONS | 18 |
| 1.7 STRUCTURE OF THE THESIS | 19 |
| CHAPTER 2 ENDOGENEITY IN SPATIAL CHOICE MODELS..... | 20 |
| 2.1 OVERVIEW..... | 20 |
| 2.2 THEORETICAL CONSIDERATIONS | 21 |
| 2.2.1 <i>Causes of Endogeneity in Spatial Choice Models</i> | 21 |
| 2.2.2 <i>Methods to Correct for Endogeneity in Discrete Choice Models</i> | 22 |
| 2.2.3 <i>The Control-Function Method</i> | 24 |
| 2.2.4 <i>Change of Scale with the Control-function Method</i> | 27 |
| 2.2.5 <i>Simulation and Forecasting with the 2SCF Method</i> | 30 |
| 2.2.6 <i>Comparison between 2SCF and 2SIV Methods</i> | 32 |
| 2.2.7 <i>Efficiency and Calculation of Standard Errors with the 2SCF Method</i> | 34 |
| 2.2.8 <i>Testing for Endogeneity</i> | 35 |
| 2.3 MONTE CARLO EXPERIMENT..... | 36 |
| 2.3.1 <i>Model Setting</i> | 36 |
| 2.3.2 <i>Estimation with 2SCF and 2SIV Methods</i> | 36 |
| 2.3.3 <i>Forecasting with 2SCF and 2SIV Methods</i> | 39 |
| 2.4 APPLICATION TO REAL DATA | 41 |
| 2.4.1 <i>Overview</i> | 41 |
| 2.4.2 <i>Construction of the Database for Estimation</i> | 42 |
| 2.4.3 <i>Instrumental Variables</i> | 48 |
| 2.4.4 <i>Estimation Using the 2SCF Method</i> | 53 |
| 2.4.5 <i>Correction of Standard Errors</i> | 56 |
| 2.4.6 <i>Forecasting</i> | 57 |

| | |
|---|------------|
| 2.5 CONCLUSION | 58 |
| CHAPTER 3 EFFICIENCY AND TRACTABILITY IN THE CORRECTION FOR ENDOGENEITY USING LATENT-VARIABLE AND CONTROL-FUNCTION METHODS | 59 |
| 3.1 OVERVIEW..... | 59 |
| 3.2 THE LATENT-VARIABLE METHOD IN THE CORRECTION FOR ENDOGENEITY | 60 |
| 3.3 THE CONTROL-FUNCTION METHOD IN A MAXIMUM-LIKELIHOOD FRAMEWORK | 65 |
| 3.4 THE LINK BETWEEN LATENT-VARIABLE AND CONTROL-FUNCTION METHODS..... | 67 |
| 3.5 ASSUMPTIONS TO ACHIEVE EFFICIENCY AND TRACTABILITY..... | 69 |
| 3.6 MONTE CARLO EXPERIMENT..... | 71 |
| 3.7 APPLICATION TO REAL DATA | 73 |
| 3.8 CONCLUSION | 75 |
| CHAPTER 4 TESTING FOR THE VALIDITY OF INSTRUMENTAL VARIABLES IN DISCRETE CHOICE MODELS..... | 77 |
| 4.1 OVERVIEW..... | 77 |
| 4.2 VALIDATION OF INSTRUMENTS USING OVER-IDENTIFYING RESTRICTIONS | 78 |
| 4.2.1 <i>The Sargan Test in Linear Models</i> | 78 |
| 4.2.2 <i>The Amemiya-Lee-Newey Test in Discrete Choice models</i> | 82 |
| 4.3 TWO NOVEL TESTS FOR DISCRETE CHOICE MODELS | 85 |
| 4.3.1 <i>A Regression-based Test for Logit Models</i> | 85 |
| 4.3.2 <i>A Direct test for Discrete Choice Models</i> | 90 |
| 4.4 MONTE CARLO EXPERIMENT..... | 94 |
| 4.5 APPLICATION TO REAL DATA..... | 98 |
| 4.6 CONCLUSION | 102 |
| CHAPTER 5 SAMPLING OF ALTERNATIVES IN MULTIVARIATE EXTREME VALUE MODELS | 103 |
| 5.1 OVERVIEW..... | 103 |
| 5.2 ESTIMATION AND SAMPLING OF ALTERNATIVES IN LOGIT MODELS..... | 104 |
| 5.3 A NOVEL METHOD FOR MEV MODELS..... | 108 |
| 5.4 FORMULATION OF THE METHOD FOR NESTED LOGIT | 117 |
| 5.5 FORMULATION OF THE METHOD FOR CROSS-NESTED LOGIT | 121 |

| | |
|---|------------|
| 5.6 MONTE CARLO EXPERIMENT..... | 122 |
| 5.6.1 <i>Model Setting</i> | 122 |
| 5.6.2 <i>Assessment of the Methods with and without Re-sampling</i> | 126 |
| 5.6.3 <i>Expansion in Practice when Re-sampling is not Possible</i> | 130 |
| 5.6.4 <i>Additional Experiments</i> | 133 |
| 5.7 APPLICATION TO REAL DATA..... | 137 |
| 5.8 CONCLUSION..... | 141 |
| CHAPTER 6 CONCLUSION..... | 142 |
| 6.1 SUMMARY..... | 142 |
| 6.2 OVERALL CONCLUSION..... | 143 |
| 6.3 METHODOLOGICAL RECOMMENDATIONS..... | 143 |
| 6.4 EXTENSIONS..... | 145 |
| REFERENCES..... | 147 |

List of Tables

| | |
|---|-----|
| Table 2-1 Monte Carlo Experiment: Model Estimation with 2SCF and 2SIV..... | 37 |
| Table 2-2 Monte Carlo Experiment: Forecasting with Endogeneity Correction..... | 40 |
| Table 2-3 Summary of Lisbon's Residential Location Choice Database for Estimation..... | 47 |
| Table 2-4 Correlation Matrix of Dwelling Price and Instrumental Variables..... | 52 |
| Table 2-5 Lisbon's Logit Model: First Stage of 2SCF..... | 53 |
| Table 2-6 Lisbon's Logit Model: With and without Correction for Endogeneity..... | 55 |
| Table 2-7 Lisbon's Logit Model: Correction of 2SCF's Standard Errors by Bootstrapping..... | 56 |
| Table 2-8 Lisbon's Logit Model: Forecasting with and without Endogeneity Correction..... | 57 |
| Table 3-1 Monte Carlo Experiment: 2SCF and Maximum-likelihood Methods..... | 72 |
| Table 3-2 Lisbon's Logit Model: 2SCF and Maximum-likelihood Methods..... | 74 |
| Table 4-1 Monte Carlo Experiment: Performance of Tests for the Validity of Instruments..... | 96 |
| Table 4-2 Lisbon's Logit Model: Auxiliary Choice Model for Amemiya-Lee-Newey Test..... | 99 |
| Table 4-3 Lisbon's Logit Model: Auxiliary Regression for Regression-based Test..... | 100 |
| Table 4-4 Lisbon's Logit Model: Auxiliary Choice Model for Direct Test..... | 101 |
| Table 5-1 Monte Carlo Experiment: Sampling in MEV with and without Re-Sampling..... | 127 |
| Table 5-2 Monte Carlo Experiment: Different Estimators of Choice Probabilities..... | 131 |
| Table 5-3 Monte Carlo Experiment: Additional Experiments on Sampling in MEV..... | 134 |
| Table 5-4 Monte Carlo Experiment: Sampling in MEV. 1,000,005 Alternatives..... | 136 |
| Table 5-5 Lisbon's Nested Logit Model: Sampling 5 + 5 Alternatives..... | 139 |
| Table 5-6 Lisbon's Nested Logit Model: Sampling 500 + 500 Alternatives..... | 140 |

List of Figures

| | |
|--|-----|
| Figure 2-1 SOTUR Observations in Lisbon Metropolitan Area (LMA)..... | 43 |
| Figure 2-2 SOTUR (■) and Imokapa (★) Observations in Lisbon, Odivelas and Amadora | 44 |
| Figure 2-3 Matching of Dwellings from SOTUR (■) into IMOKAPA (★) | 46 |
| Figure 2-4 Dwelling Price and Instrumental Variables | 52 |
| Figure 4-1 Over-identification Allows Testing for the Validity of Instruments | 81 |
| Figure 5-1 Monte Carlo Experiment: Nesting Structure. 1,005 Alternatives | 122 |
| Figure 5-2 Monte Carlo Experiment: Estimators as \tilde{J}_2 Increases. <i>Expanded True Prob.</i> | 129 |
| Figure 5-3 Monte Carlo Experiment: Nesting Structure. 1,00,005 Alternatives | 135 |
| Figure 5-4 Lisbon's Nested Logit Model: Nesting Structure | 138 |

Chapter 1

Introduction

1.1 Motivation

A model is a simplified representation of a complex phenomenon. Models of urban systems are important decision support tools for policy analysis. Limitations in computational and methodological tractability have led to the formulation of models that consider the behavior of aggregates of agents. These models neglect to consider the interactions within the different decision levels and time scales involved in urban systems. These simplifications have significantly reduced the ability to perform adequate policy analysis (Ben-Akiva, 1973; Kitamura et al., 1996; Bowman, 1998; Badoe and Miller, 2000) and have consequently limited the ability to control traffic congestion, air pollution, noise and other externalities that jeopardize urban sustainability.

Any model will be only as valid as the behavioral assumptions on which it is based. Therefore, models of urban systems will be ultimately wrong if they neglect the fact that the behavior of the system is the end result of the choices made by millions of heterogeneous agents, with varying levels of information, unique motivations, and at distinct time and space scales. Consequently, it has become the common goal of various research teams around the world to work toward the development of microscopic integrated models of urban systems (Miller et al., 2004; Strauch et al., 2005; Waddell et al., 2008; Almeida et al., 2009).

The development of trustworthy and practical microscopic integrated models of urban systems is still a challenge. Current models are plagued by shortcomings such as the lack of a practical framework that can represent agent behavior (Ben-Akiva, 2010); the estimation, simulation and integration of different modeling components (Antonioni et al., 2008); and the collection, processing, and integration of data (Chen et al., 2009). In terms of the estimation and application, there are two important modeling drawbacks that are shared by several components of microscopic integrated models of urban systems. Microscopic spatial choice modeling requires a detailed representation of numerous quasi-unique alternatives. This would be impossible to implement in practice and results in the omission of certain attributes of the alternatives, and in that only a subset of the true choice-set can be considered by the researcher.

The need for omitting attributes and sampling of alternatives is common in different spatial choice models that are embedded into microscopic integrated urban models. These simplifications are required, for example, in models of residential or job location choice, where the number of dwellings or workplaces in the choice-set may be extremely large and varied. These simplifications are also required in route-choice models, where there may be many different routes linking two places. Equivalently, these simplifications are also necessary in activity-based models because the number of potential combinations of activities, schedules, duration, and participation choices may be enormous and heterogeneous.

The omission of attributes results in inconsistent estimators when the omitted attributes are correlated with the observed ones. This problem is known as endogeneity and it has been systematically ignored by the literature on transportation and spatial choice modeling. Besides, the problem of obtaining consistent estimators of the model parameters when only a sample of the true choice-set is available has been resolved only for Logit, a model type that is unrealistic for several spatial choice models. This research focuses on addressing the issues of endogeneity in discrete choice models and sampling of alternatives in Multivariate Extreme Value models, a family of closed-form choice models that includes Logit among other models that allow for more realistic error structures for spatial choice modeling.

1.2 Objectives and Methodology

This research focuses on addressing endogeneity and sampling of alternatives in Multivariate Extreme Value (MEV) models, two major model estimation drawbacks shared by several spatial choice models embedded into microscopic integrated urban models.

In terms of the motivation, the framework for analysis and the examples used, this thesis is concerned with the estimation of models of residential location choice, a case where the modeling drawbacks under study have special relevance. However, the methodological advances resulting from this research will be generally applicable to a vast range of choice models, including other spatial choice models such as route choice, activities scheduling and firm and job location.

The research methodology used in this study was threefold. In the first stage, I drew from different fields in order to enhance, adapt or develop potential solutions for the modeling drawbacks being studied. The proposed methods were developed while keeping in mind that they must be computationally tractable, theoretically based and behaviorally consistent with the problem of residential location choice. In the second stage, these advancements were assessed and enhanced using Monte Carlo experimentation. The performance of the proposed methods under diverse circumstances was compared. Finally, the methods under development were applied to a case study using real data on residential location choice from the city of Lisbon, Portugal. All Monte Carlo and real data experiments developed in this thesis were generated and estimated using the open-source software R (R Development Core Team, 2008).

1.3 Modeling Framework

This thesis is concerned with a modeling framework where there are agents (n) that choose an alternative (i) among a set of elements or choice-set C_n (typically households choosing among potential residences). Besides the agents making choices, the framework is completed by a researcher who wants to model the agents' behavior in order to develop policy analysis.

Households (n) are assumed to behave rationally. Households perceive certain utility from the combination of activities their members are involved in. When choosing among a set of potential dwellings (i), households evaluate the maximum level of utility (U_{in}) that they may achieve, conditional on the selection of each alternative. Then, households choose the alternative that allows them obtaining the largest level of utility.

Utility functions (U_{in}) are indirect in nature because they depend on the attributes of alternative i (typically dwelling price p_{in} and some other attributes x_{in} and q_{in}) and the characteristics of household n (typically income).

Utilities are considered to be random variables. Utilities are assumed to be compounded by a systematic part V_{in} and a random part ε_{in} . The systematic part is assumed to depend linearly on the dwelling's attributes (potentially interacted with household characteristics) with coefficients β^* . The random part consists of an error term or discrepancy (ε_{in}), which is a random variable.

$$U_{in} = V_{in} + \varepsilon_{in} = \beta_p^* p_{in} + \beta_x^* x_{in} + \beta_q^* q_{in} + \varepsilon_{in}$$

The researcher can observe the dwelling's attributes and the choices made by a total of N households, but not the utilities, which are latent. Assuming a certain distribution of the error terms (ε_{in}), the researcher can formulate the following choice probability model for alternative i :

$$P_n(i) = P(U_{in} \geq U_{jn} \quad \forall j \in C_n).$$

When the researcher observes the true choices, precisely measures all attributes (p_{in} , x_{in} and q_{in}) for the full choice-set C_n , and uses the correct distribution for the error term (ε_{in}), the researcher will be able to retrieve consistent estimators for the model parameters. This means that estimators $\hat{\beta}$ will be as close to β^* as desired (if N is large enough). This also implies that the choice probability model will be a reliable representation of household behavior, and would allow for the effective policy analysis. The main purpose of this thesis is to determine the impact and to investigate solutions for cases where certain attributes (like q_{in}) are not measured by the researcher, and when only a subset D_n of the true choice-set C_n is observed.

1.4 Endogeneity

Endogeneity is an inevitable problem for all spatial choice models. In the case of residential location choice, endogeneity usually occurs when a researcher who wants to model household behavior cannot account for all the attributes that may influence a household's final residential location choice. Since dwelling attributes are likely to be correlated with price, a model that accounts for price but omits other relevant attributes will suffer from endogeneity: the error term of the model will be correlated with the observed price. The result of this misspecification is that the model will fail to account for the correct impact of price in the choice process because the effect of price will be confounded with the impact of the omitted attributes.

Consider, for example, the case of seemingly equal apartments that differ only in two attributes: their price and their location within the building. An apartment that is in the corner of the building usually has a better view and better lighting. The preference for these attributes triggers a larger demand for corner apartments in the market, and a consequent increase in their price. Household's choices are then based on the trade-off between apartments' price and location within the building. If the researcher's model omits apartment's location, choices toward the more expensive apartments will be then misinterpreted as the result of an unrealistically small deterrence to price.

Endogeneity might significantly impact the suitability of models of urban systems as reliable tools for policy analysis. For example, consider that the policy under study is the distribution of a subsidy to urban residents geared toward encouraging households to reside in the city center. In this case the underestimation of the deterrence to price caused by endogeneity will result in an overestimation of the subsidy required and in a misleading picture of the effects of the policy. A policy maker deluded by this misspecified model may end up trashing the subsidy policy because it may seem too expensive to implement (as informed by the spurious model); or the policy maker may end up ignoring the model completely, only to apply subsidies at a level that seems intuitively reasonable. In both cases, the modeling effort is almost useless.

Different methods to treat for endogeneity in discrete choice models have been developed. One of them is known as the control-function method (Heckman, 1978,

Hausman, 1978). This technique corrects for endogeneity even when it occurs at the level of each alternative, making it more practical for residential location choice modeling when compared to the method proposed by Berry et. al (1995), which can only correct for endogeneity when it occurs at the level of markets or large sets of alternatives. The control-function method can be applied to Logit and non-Logit models, such as the Nested Logit or the Probit. In Chapter 2, I study the problem of endogeneity in models of residential location choice and analyze a two-stage version of the control-function method. First, I use Monte Carlo experimentation to study some theoretical issues about the application of the control-function method. Then, I deploy all the practical considerations involved in applying the method to estimate a model of residential location choice for Lisbon, Portugal.

One alternative to the control-function method is to consider the omitted attributes as latent variables (Walker and Ben-Akiva, 2002). In Chapter 3, I show that the simultaneous estimation of the control-function method in a full-information-maximum-likelihood framework (Train, 2009; Newey, 1987; Rivers and Vuong, 1988; Villas-Boas and Winner, 1999; Park and Gupta, 2009) is fully equivalent to the latent-variable approach. This method can be applied to Logit and non-Logit models, such as the Nested Logit or the Probit. Chapter 3 also shows how, given certain assumptions, the maximum-likelihood estimator can be reduced to a tractable form that avoids multidimensional integration. This avoidance is important because the large number of alternatives in residential location choice models makes integration impracticable. I also show that under these conditions, both the two-stage and the tractable maximum-likelihood estimator can efficiently estimate model parameters; however, only the standard errors of the latter do not need to be corrected by bootstrapping (Petrin and Train, 2002) or other techniques such as the delta-method (Karaca-Mandic and Train, 2003). The properties of the different estimators are studied using both Monte Carlo experimentation and real data.

Much like the other methods used to correct for endogeneity, the control-function method relies on the availability of valid instrumental variables. The instruments need to comply with two conflicting properties. They need to be correlated with the endogenous variable (the price) and, at the same time, to be uncorrelated with the unobserved

attributes that cause endogeneity. Whether or not the instrumental variables correlate with the endogenous variable is trivial to verify because the endogenous variable is observable. In turn, it is more difficult to verify that the instruments are uncorrelated with the omitted attributes because the omitted attributes are unobservable.

In Chapter 4, I review the state-of-the-art in testing for the validity of instruments, which can be summarized by the Sargan (1958) test for linear models and the Amemiya-Lee-Newey (Lee, 1992) test for discrete choice models. Then, I develop two novel tests for discrete choice models. The first test, termed Regression-based, was developed by adapting Sargan's test into the Logit framework. The second test, termed Direct, was constructed from a different framework, is much easier to implement using commercial software, and is applicable for Logit and non-Logit models. Monte Carlo experimentation on a binary Logit case showed that these two novel tests are statistically more powerful than the Amemiya-Lee-Newey test. The tests were also applied for the validation of the instruments used in the residential location choice model for Lisbon.

1.5 Sampling of Alternatives in MEV Models

The number of alternatives in spatial choice models is usually huge. Collection, processing and estimation costs for such big databases render the use of the full choice-set for modeling impractical. McFadden (1978) showed that the consistent estimation of Logit models using only a sample of the alternatives is possible by adjusting the likelihood function based on the sampling protocol. However, the Logit assumption is difficult to sustain in spatial choice models since the alternatives are expected to be correlated according to proximity or to be nested according to different decision levels.

Ignoring a non-Logit structure in spatial choice modeling may significantly impact the quality of spatial choice models. For example, if the underlying model is a Nested Logit with nests defined by geographical areas, a location subsidy will trigger more intra-area than inter-area household relocation. This effect would be impossible to capture with a Logit model, resulting in misleading guidance for urban policy analysis.

Few significant extensions of McFadden's consistency result to non-Logit models have been made. Some researchers have studied the problem of choice-based samples in

non-Logit models, which are cases where the complete choice-set is available but the observations are sampled conditional on the choices (Manski and Lerman, 1977; Manski and McFadden, 1981; Cosslett, 1981; Imbens and Lancaster, 1994; Garrow et al., 2005; Bielaire et al., 2009). Other advances have been made in the empirical study of the impact of sampling of alternatives in Logit Mixture models (McConnel and Tseng, 2000; Nerella and Bhat, 2004; Chen et al., 2005). Finally, for the case of the Nested Logit, the problem of sampling of alternatives has been largely ignored and erroneously assumed to be solvable by the application of the sampling correction derived by McFadden (1978) for the Logit model (Berkovec and Rust, 1985; Train et al., 1987; Hansen, 1987; Rivera and Tiglaio, 2005).

Building on an idea originated by Ben-Akiva (2009), in Chapter 5, I present a method that allows for the consistent estimation of model parameters for models belonging to the Multivariate Extreme Value (MEV) class, when only a sample of the true choice-set is observed. The MEV model class is a family of models that allows for different correlation structures among alternatives. The method is deployed in detail for the Nested and Cross-Nested Logit models, the principal members of the MEV class. I illustrate the properties of the method and the impact of the misspecification using Monte Carlo experimentation and real residential location data from the city of Lisbon. In the Lisbon case study, I combine the tools developed to address sampling of alternatives in MEV models with those to correct for endogeneity deployed in the previous chapters.

1.6 Contributions

Regarding the problem of endogeneity, I applied, enhanced, and developed methods to test and to correct for endogeneity in models of residential location choice, as well as methods to validate and apply such models in simulation. To achieve these goals, I synthesized the latest research in this topic and developed one of the first comprehensive applications to address this problem for residential location choice modeling.

I also studied some methodological issues that have been debated in the literature, and developed maximum-likelihood estimators that are consistent, efficient, and are tractable in problems with large choice-sets, such as residential location choice models. I also

developed two tractable tests for the validity of instrumental variables in discrete choice models that showed better power properties than an existing test in a set of binary Logit Monte Carlo experiments. In addition, I identified the link between the latent-variable and control-function methods in the correction for endogeneity in spatial choice models, and discussed the potential benefits that this link may allow.

Regarding the problem of sampling of alternatives, the main contributions of this doctoral dissertation are in the development and demonstration of a method for achieving consistency, relative efficiency, and asymptotic normality when the underlying model is MEV. This novel method is the first significant extension of McFadden's work on sampling of alternatives for Logit models in 30 years. It will make feasible the implementation of more realistic error structures in future applications on microscopic modeling and render the development of better tools for policy analysis.

1.7 Structure of the Thesis

This introductory chapter is followed by the four methodological chapters described before. Chapter 2 is concerned with endogeneity in spatial choice models and the application of a two-stage version of the control-function method to correct for endogeneity in residential location choice. Chapter 3 studies the link between the latent-variable and the control-function methods in the quest for efficiency and tractability in the correction for endogeneity. Chapter 4 is concerned with the development of tests for the validity of instruments in discrete choice models and their application to residential location choice models. In Chapter 5, I develop and assess a novel method to address the problem of sampling of alternatives in MEV models. Chapter 6 presents a summary of the methodological findings resulting from this thesis, analyses their impacts and limitations, derives modeling recommendations, and suggests further directions of research in this area. This is finally followed by the list of bibliographic references used in this study.

Chapter 2

Endogeneity in Spatial Choice Models

2.1 Overview

An econometric model is said to suffer from endogeneity when the systematic part of the utility is correlated with the error term. This problem is common in spatial choice models in general and in residential location choice models in particular. Endogeneity is a critical modeling failure that leads to the inconsistent estimation of model parameters. Intuitively, if a variable is endogenous, changes in the error term will be misinterpreted as resulting from changes of the endogenous variable, making impossible the consistent estimation of the model parameters.

In this chapter, I discuss the correction for endogeneity in residential location choice models using a two-stage version of the control-function method, the most suitable tool to address endogeneity in this framework. This chapter is divided into three parts. The first section presents a critical review of the theoretical aspects involved in the correction for endogeneity in residential location choice models. Then, I use Monte Carlo experimentation to study the properties of the different procedures deployed in the first section. Finally, I develop a comprehensive application of the formulation, estimation and correction for endogeneity in a discrete choice model of residential location for the city of Lisbon, Portugal.

2.2 Theoretical Considerations

2.2.1 Causes of Endogeneity in Spatial Choice Models

There are generally three causes of endogeneity. One cause is errors in variables. If a variable is measured wrong, that error will be propagated to the model's unobserved part, which will then be correlated with the wrongly measured variable, causing endogeneity. Errors in variables are unavoidable in models of residential location choice, just as they are inevitable in any econometric model. This source of endogeneity needs to be controlled by measuring the variables of the model as precisely as possible.

A second situation that may lead to endogeneity is known as simultaneous determination. This type of endogeneity can be observed, for example, in the joint determination of location and modal choices. People who are transit-oriented would more likely choose to live in dwellings that have better accessibility to transit and will consequently have relatively better travel times, compared to other people in the city. Since being transit-oriented means also having a relatively more positive error term in the mode choice model, this implies that travel time by transit will be correlated with the modal error, causing endogeneity.

In the case of residential location choice, endogeneity from simultaneous determination may be expected at an aggregated level because the aggregated demand for dwellings depends on their price and, their price depends on the demand for them. However, if the demand and supply are treated at a microscopic scale, this source of endogeneity might not be significant because the price of each dwelling is not likely to be determined by the choice made by any particular household. Moreover, the effect of all agents on dwelling price would become apparent only in the medium term, mitigating any potential endogeneity effect from this source in residential location choice models.

A third cause of endogeneity is the omission of variables that are relevant in the model and are correlated with some observed attributes. This source of endogeneity is unavoidable and significant in microscopic models of residential location choice. Therefore, it is the main motivation for this chapter. The large number and variety of the attributes that are relevant in location choice decisions makes it difficult to model this

phenomenon since it becomes impossible to measure or even to fully identify all of them. This omission becomes a problem when those attributes, which become part of the error, are correlated with the observed model variables.

Consider, for example, the case of two seemingly equal houses that differ only in that one has been recently renovated and consequently has a higher price. If the data on the renovation of the house is not available, the observation of the choice of the house with the higher price will lead to the erroneous conclusion that the sensitivity to price is smaller than it really is.

Numerous empirical applications in residential location choice modeling have shown estimated coefficients of dwelling price that are non-significant or even positive when endogeneity is not taken into account (Guevara and Ben-Akiva, 2006; Guevara, 2005; Bhat and Guo, 2004; Sermonss and Koppelman, 2001; Levine, 1998; Waddell, 1992; Quigley, 1976). This reinforces the idea that endogeneity is a prevalent problem in the field.

2.2.2 Methods to Correct for Endogeneity in Discrete Choice Models

Two main methods have been proposed to correct for endogeneity in discrete choice models when the endogenous variable is continuous. When endogeneity occurs at the level of a market or a group compounded by a sufficiently large set of observations, the problem can be solved by applying the BLP method proposed by Berry et al. (1995). This method consists of the estimation of an Alternative Specific Constant (ASC) for each market in order to account for the endogeneity problem.

Berry et al. (1995) apply their method in the choice of automobile models, a case where the price is expected to be endogenous by market. The problem is solved by calculating ASCs by markets that are geographically defined. Given the large number of ASCs required by this method, the estimation is performed iteratively using a contraction. In the second stage, the ASCs are regressed as a linear function of model variables.

If endogeneity is expected in the second stage of the BLP method, it can be addressed using the two-stage least-squares (2SLS) method for linear models (see, e.g., Greene, 2003). The first stage of the 2SLS method corresponds to an auxiliary regression of the

endogenous variable on instrumental variables. The instruments are variables that have to be correlated with the endogenous variable, but uncorrelated with the error term of the model. Then, the original model is estimated replacing the endogenous variable by the fitted values obtained from the auxiliary regression. The 2SLS method is described with further detail in Section 4.2.1.

The BLP method cannot be applied to correct for endogeneity in residential location choice models because endogeneity is expected to occur at the level of each alternative, caused by the omission of attributes that are specific to each dwelling. Therefore, the BLP method would entail, in residential location choice modeling, the estimation of ASCs for each alternative in the choice-set. This is generally impossible or, at least, would lead to over-fitting or incidental-parameter problems (Wooldridge, 2002). This seems to be a methodological problem in the work of Bayer et al. (2004), the only application of the BLP method in residential location choice, to the best of my knowledge.

Examples of applications of the BLP approach in transportation are Train and Winston (2007), who used the method to address price endogeneity at the consumer-level in vehicle choice modeling, and Walker et al. (2010), who used the method to address endogeneity in a model of peer group behavior.

The second method to treat for endogeneity in discrete choice models when the endogenous variable is continuous is known as the control-function method. This method is similar to the 2SLS method in that it relies on an auxiliary regression of the endogenous variable onto instruments. However, in the control-function method, instead of substituting the endogenous variable with the fitted counterpart obtained from the auxiliary regression, the endogenous variable is maintained in the model and the residuals of the auxiliary regression are used as additional variables. This method can handle endogeneity at the level of each alternative, and is then suitable for the problem of residential location choice modeling. Examples of previous applications of the control-function method in residential location choice are Guevara (2005), Guevara and Ben-Akiva (2006), and Ferreira (2010).

Other methods to correct for endogeneity in discrete choice models when the endogenous variable is continuous are the two-stage instrumental-variables (2SIV)

method, which is discussed in Section 2.2.6; the latent-variable method, which is presented in Section 3.2; and a method developed by Amemiya (1978), which is discussed in Section 4.2.2. All of these alternative methods are either outperformed by or grounded in the control-function method, which is described in detail in the next section.

Finally, it should be remarked that the methods studied in this thesis to address endogeneity are concerned with discrete choice models where the endogenous variables are continuous. When the endogenous variables are discrete, literature indicates (Wooldridge, 2002; Evans and Schwab, 1995) that the problem can only be solved by using maximum-likelihood methods, an approach that might become impractical in spatial choice models and is left for future research.

2.2.3 The Control-Function Method

The original idea of the control-function method comes from Hausman (1978) and Heckman (1978). In order to define the method and show how and why it effectively corrects for endogeneity, consider the behavioral model described in Eq. (2-1), where a group of N households (n) face the selection of a dwelling i among the J dwellings in the choice-set C_n .

$$\begin{aligned} U_{in} &= \beta_p p_{in} + \beta_x x_{in} + \varepsilon_{in} = \beta_p p_{in} + \beta_x x_{in} + \xi_{in} + e_{in} \quad n = 1, \dots, N; i \in C_n \\ p_{in} &= \alpha_z z_{in} + \delta_{in} \\ y_{in} &= 1[U_{in} = \max_{j \in C_n} \{U_{jn}\}] \end{aligned} \quad (2-1)$$

Household n perceives a certain utility U_{in} from dwelling i . The utility depends linearly on price p_{in} , an attribute x_{in} , and a zero mean error term ε_{in} , which can be decomposed into two parts ξ_{in} and e_{in} that also have zero mean. U_{in} is a latent variable. The researcher observes variables x_{in} , z_{in} , p_{in} and the choice y_{in} , which takes value 1 if the alternative i has the largest utility among the alternatives in choice-set C_n , and zero otherwise. The price p_{in} is determined as a linear function of variable z_{in} and a zero mean error δ_{in} , expression that is termed the price equation. For notational purposes, it will be considered from this point that U , p , x , ε , ξ , e , z , δ and y correspond to vectors compounded by the respective variables stacked by alternatives i and households n . This notation is maintained in the rest of the thesis.

Variables x and z are exogenous, meaning that they are uncorrelated with all error terms ε , ξ , e , and δ of the model. Variable x is said to be a control because it appears in the specification of the utility function. Variable z is said to be an instrument for price, because it does not appear in the utility function and is correlated with price. The error term e is uncorrelated with the observed variables p , x and z , and with the error term δ .

Endogeneity problems arise when δ is correlated with ξ . In this case, p will be correlated with ξ and the standard estimation methods will fail to retrieve consistent estimators of model parameters. This problem may occur, for example, if ξ contains relevant dwelling attributes that are correlated with p , but cannot be measured by the researcher.

The control-function method consists of the construction of an auxiliary variable, which when added to the systematic part of the utility function, the remaining error of the model will no longer be correlated with observed variables. To construct this auxiliary variable, note first that it is always possible to write ξ as the sum of its conditional expectation, given δ , and an error term v , such that

$$\xi_{in} = E(\xi_{in} | \delta_{in}) + v_{in}.$$

Then, the error term v will be orthogonal to δ by construction and therefore uncorrelated with it. Assuming then that ξ and δ are jointly Normal, we have

$$\xi_{in} = \beta_{\delta} \delta_{in} + v_{in},$$

where v will be independent of δ and will follow a Normal distribution with zero mean and a fixed variance σ_v^2 (Wooldridge, 2002).

The next step is to show that z is uncorrelated with v . To show why, note first that since z is a valid instrument, it must be uncorrelated with δ and ξ . Then, since δ and ξ have zero mean, the fact that they are uncorrelated with z implies that $E(\xi'z) = E(\delta'z) = 0$. Replacing these conditions into $\xi = \beta_{\delta} \delta + v$, it follows that

$$\begin{aligned} \xi &= \beta_{\delta} \delta + v \\ E(z' \xi) &= \beta_{\delta} E(z' \delta) + E(z' v) = 0 + E(z' v) = 0, \end{aligned}$$

where, given that v has zero mean, this implies that z is uncorrelated with v .

The final step is to show that v is uncorrelated with p . This can be achieved by noting that

$$p = \alpha_z z + \delta$$

$$E(v' p) = \alpha_z E(v' z) + E(v' \delta) = 0 + 0 = 0$$

Therefore, the endogeneity problem can be solved if this decomposed $\xi = \beta_\delta \delta + v$ is replaced in the utility function. Indeed, assuming (for the moment) that δ is observed, the remaining error $v + e$ in Eq. (2-2) will not be correlated with the observed attributes of the model: p , x and δ .

$$U_{in} = \beta_p p_{in} + \beta_x x_{in} + \varepsilon_{in} = \beta_p p_{in} + \beta_x x_{in} + \beta_\delta \delta_{in} + v_{in} + e_{in}$$

$$p_{in} = \alpha_z z_{in} + \delta_{in}$$

$$y_{in} = 1[U_{in} = \max_{j \in C_n} \{U_{jn}\}] \quad (2-2)$$

The practical problem that δ is not observed can be addressed in different ways. Chapter 3 analyzes the implementation of the model described in Eq. (2-2) under the maximum-likelihood and the latent-variable frameworks. Alternatively, this problem can be addressed by recalling that, since δ and z are uncorrelated, δ can be consistently estimated by using an ordinary-least-squares (OLS) regression of p on z . Therefore, if the consistent estimator of δ is inserted into the choice model, the consistency of the estimators of the model parameters would be guaranteed by the Slutsky theorem (Ben-Akiva and Lerman, 1985).

Formally, the following procedure, which is termed in this thesis the two-stage control-function (2SCF) method, can be devised to solve the endogeneity problem in the discrete choice model described in Eq. (2-1):

Stage 1: Estimate $\hat{\delta}$ by ordinary-least-squares (OLS).

$$p_{in} = \alpha_z z_{in} + \delta_{in} \xrightarrow{OLS} \hat{\alpha}_z \Rightarrow \hat{\delta}_{in} = p_{in} - \hat{p}_{in} = p_{in} - \hat{\alpha}_z z_{in}$$

Stage 2: Estimate the choice model using $\hat{\delta}$ as an additional variable.

$$U_{in} = \beta_p p_{in} + \beta_x x_{in} + \beta_\delta \hat{\delta}_{in} + \underbrace{\tilde{v}_{in} + e_{in}}_{\tilde{e}_{in}}$$

If it is additionally assumed that $\tilde{v} + e$ from Stage 2 follows, or can be approximated using an Extreme Value distribution, the model becomes a Logit, making the 2SCF easy

to estimate with commercial software using maximum-likelihood methods. This assumption might seem difficult to sustain at first. If e is distributed Extreme Value, there is no parametric distribution of \tilde{v} that would result in that $\tilde{v} + e$ is distributed Extreme Value. However if the sample is large enough, it is first possible to claim the Law of Large Numbers to say that $\tilde{v} + e$ will be normally distributed. The argument is completed using the results from Lee (1982) and Ruud (1983), which state that the approximation of a Normal by an Extreme Value distribution causes only negligible discrepancies.

The application of the 2SCF method to some cases that are not covered by Eq. (2-1) implies small variations. First, when the model has various continuous endogenous variables, the only difference is that an auxiliary variable $\hat{\delta}_k$ has to be estimated for each endogenous variable k in Stage 1. Then, in Stage 2, each $\hat{\delta}_k$ has to be added to the systematic part of the utility. Instrumental variables can be shared among endogenous variables in the first stage of the method. However, to obtain identification, it is indispensable to have at least as many different instrumental variables as there are endogenous variables in the model. Second, when the exogenous variable x forms part also of the price equation, x should be included in the right hand side of the first stage of the 2SCF method. Otherwise, the residual $\hat{\delta}$ would be correlated with x in the second stage of the 2SCF, affecting the estimation of its coefficient. Finally, when the error term δ does not have mean zero, the method can be applied by including an intercept in the first stage of the 2SCF method.

2.2.4 Change of Scale with the Control-function Method

The correction for endogeneity using the control-function method produces consistent estimators of the model parameters but only up to a certain scale. That is, the ratios between the estimators are consistent estimators of the ratios of the parameters of the true model, but the actual estimators of the model parameters are inconsistent. This is also true, in general, for BLP, 2SIV and Amemiya's methods to correct for endogeneity in discrete choice models.

The change of scale in the control-function method results from the fact that the error term with the control-function correction in Eq. (2-2) is $v + e$, whereas the error term of

the original model shown in Eq. (2-1) was only e . Therefore, if the variance of v is not null, the control-function correction will trigger a change of scale in the estimated parameters. This effect is analogous to that of the omission of an orthogonal attribute in discrete choice models. An orthogonal attribute is one that truly and importantly belongs to the systematic part of the utility, but is uncorrelated with other observed attributes. The problem of the change of scale due to the omission of an orthogonal variable was originally studied by Yatchew and Griliches (1985) for the Probit model. Cramer (2007) extended this analysis to the binary Logit model. Here, I use their framework to determine the change of scale caused by the application of the 2SCF method in correcting for endogeneity in Logit models.

Consider the true model shown in Eq. (2-1) where ξ is observed, and assume that the error e is distributed Extreme Value $(0, \mu_e)$. As with any Logit model, the scale is not identifiable and normalization is required. The usual normalization is to set $\mu_e = 1$. This is equivalent to normalizing the variance of the differences of e across alternatives to be equal to $\sigma_e^2 = \pi^2/3$.

Consider now the model corrected for endogeneity using the control-function method described in Eq. (2-2). The usual normalization $\mu_{v+e} = 1$ would imply that $\sigma_{v+e}^2 = \pi^2/3$. However this normalization is incompatible with that assumed for the model in Eq. (2-1). To determine the correct normalization, consider first the ratio between the scales of the two models. Since v and e are uncorrelated by construction, this ratio will depend only on the variances of v and e as follows:

$$\frac{\mu_{v+e}}{\mu_e} = \frac{\sigma_e}{\sigma_{v+e}} = \frac{\sigma_e}{\sqrt{\sigma_v^2 + \sigma_e^2 + 2\text{cov}(v,e)}} = \frac{\sigma_e}{\sqrt{\sigma_v^2 + \sigma_e^2}} = \frac{1}{\sqrt{1 + \frac{\sigma_v^2}{\sigma_e^2}}}.$$

Then, if the normalization of the model in Eq. (2-1) $\sigma_e^2 = \pi^2/3$ is to be maintained, the compatible scale of the model shown in Eq. (2-2) should be

$$\mu_{v+e} = 1/\sqrt{1 + 3\sigma_v^2/\pi^2}. \quad (2-3)$$

This change of scale is unknown to the researcher in a practical application because the variance of v is not identifiable. This raises the natural question of what is the cost of

the omission of v in the estimation of the control-function method. It turns out that the cost of this omission is negligible. First, it is usually the ratio between the coefficients what is relevant, not their actual values, and the ratios are indeed obtained consistently with the change of scale that results from the application of the 2SCF method. Second, beyond the ratios, the other thing that is important is the effect in forecasting.

The first insight into the issue of forecasting comes from Wooldrige (2002). He proved, for binary Probit, that the omission of an attribute that is uncorrelated with other observed variables will not change the expected value of the derivative of the choice probability. There is no equivalent analytical result for Logit, but Cramer (2007), for binary Logit, and Daly (2008), for multinomial Logit, used Monte Carlo experimentation to show that the sample average of the derivative of the choice probability, which they termed the Average Sample Effect (ASE), differs insignificantly between the full model and a model that omits a variable that is uncorrelated with other observed variables.

Cramer's and Daly's results can be directly extended to the case of the change of scale caused by the application of the 2SCF method because the error term v acts as an omitted orthogonal attribute in Eq. (2-2). Assume that e and $e+v$ are distributed (or can be approximated) using an Extreme Value distribution. Term:

- $\hat{P}_n(i)$ the choice probability of alternative i calculated using estimators $\hat{\beta}$ from the model shown in Eq. (2-1), including the variable ζ in the utility, and
- $\hat{\hat{P}}_n(i)$ the choice probability calculated using estimators $\hat{\hat{\beta}}$ of the model shown in Eq. (2-2), omitting variable v .

Then, the extension of Cramer's and Daly's results to the analysis of the impact of the application of the 2SCF method in the ASE of price, for alternative i , in a Logit model, can be summarized as follows:

$$\overline{ASE}_p(i) = \frac{1}{N} \sum_{n=1}^N \frac{\partial \hat{P}_n(i)}{\partial p_{in}} = \frac{1}{N} \sum_{n=1}^N (1 - \hat{P}_n(i)) \hat{P}_n(i) \hat{\beta}_p \approx \frac{1}{N} \sum_{n=1}^N (1 - \hat{\hat{P}}_n(i)) \hat{\hat{P}}_n(i) \hat{\hat{\beta}}_p.$$

In summary, the application of the 2SCF differs from the true model in the omission of the error term v shown in Eq. (2-2). This omission causes a change in the scale of the estimators obtained using the 2SCF. However, all the meaningful properties of the model

remain the same as those of the true model. In Section 2.3 I use Monte Carlo experimentation to provide empirical evidence of the validity of this assertion.

2.2.5 Simulation and Forecasting with the 2SCF Method

Simulation and forecasting requires the calculation of the fitted probabilities outside the sample used for estimation. Using a weak Law of Large Numbers, Wooldridge (2002) shows that the expected value of the simulated choice probability of Probit can be consistently estimated using the residuals $\hat{\delta}$ (from the first stage of the 2SCF) as additional variables. That result can be extended to Logit or other MEV models using the same Law of Large Numbers and accepting that a Normal distribution can be approximated using an Extreme Value distribution. Eq. (2-4) shows the expression of the simulated probabilities that would have to be used in the case of the Logit model, where the $\hat{\beta}$'s are the estimators obtained by the application of the 2SCF method and the superscript 1 is used to highlight the attributes that vary in the forecasting phase.

$$\hat{P}^1(i) = \frac{1}{N} \sum_{n=1}^N \hat{P}_n^1(i) = \frac{1}{N} \sum_{n=1}^N \frac{e^{\hat{\beta}_p p_{in}^1 + \hat{\beta}_x x_{in}^1 + \hat{\beta}_\delta \hat{\delta}_{in}}}{\sum_{j \in C_n} e^{\hat{\beta}_p p_{jn}^1 + \hat{\beta}_x x_{jn}^1 + \hat{\beta}_\delta \hat{\delta}_{jn}}} \quad (2-4)$$

This estimator of the choice probabilities may be impractical in some cases because the data used to estimate the model might not be available for simulation, making the use of the residuals in simulating phase impossible. This occurs, for example, in microscopic integrated models of the urban system such as UrbanSim (Waddell et al., 2008), where the choice models are estimated using real data on households n and dwellings i , but are applied to synthetic populations \tilde{n} and \tilde{i} .

Wooldridge (2002) proposed a different estimator of the choice probabilities that seems to overcome the limitations that arise in forecasting with synthetic populations. The idea is to avoid the need for calculating $\hat{\delta}$ for the synthetic populations, addressing the change of scale caused by its omission. Wooldridge presents the correction required for the case of Probit. The equivalent correction for Logit can be applied, following the same derivation used before to arrive at Eq. (2-3), by dividing the estimators with the factor

$$\sqrt{1+3\hat{\beta}_\delta^2\hat{\sigma}_\delta^2/\pi^2},$$

where $\hat{\sigma}_\delta^2$ is the sample variance of the residuals of the first stage of the 2SCF. This estimator of the choice probabilities is shown in Eq. (2-5).

$$\hat{P}^1(\tilde{i}) = \frac{1}{\tilde{N}} \sum_{\tilde{n}=1}^{\tilde{N}} \frac{e^{\frac{\hat{\beta}_p}{\sqrt{1+3\hat{\beta}_\delta^2\hat{\sigma}_\delta^2/\pi^2}} p_{i\tilde{n}}^1 + \frac{\hat{\beta}_x}{\sqrt{1+3\hat{\beta}_\delta^2\hat{\sigma}_\delta^2/\pi^2}} x_{i\tilde{n}}^1}}{\sum_{j \in C_{\tilde{n}}} e^{\frac{\hat{\beta}_p}{\sqrt{1+3\hat{\beta}_\delta^2\hat{\sigma}_\delta^2/\pi^2}} p_{j\tilde{n}}^1 + \frac{\hat{\beta}_x}{\sqrt{1+3\hat{\beta}_\delta^2\hat{\sigma}_\delta^2/\pi^2}} x_{j\tilde{n}}^1}} \quad (2-5)$$

However, this estimator of the choice probabilities is inconsistent. The problem is that Eq. (2-5) neglects the fact that δ is correlated with p when the model suffers from endogeneity. Then, even after the correction of the scale, the aggregate price elasticities of Eq. (2-5) will be different from those of the true model. I will explore the effect of this problem later in Section 2.3 using Monte Carlo experimentation.

Instead of using Eq. (2-5) for the case of synthetic populations, one alternative is to construct a control-function for each synthetic dwelling \tilde{i} and household \tilde{n} using the following expression:

$$\hat{\delta}_{i\tilde{n}}^0 = p_{i\tilde{n}}^0 - \alpha_z z_{i\tilde{n}}^0,$$

where the superscript zero indicates that the synthetic data used in the calculation of $\hat{\delta}$ should come from the base year.

If the dwellings available for estimation in the first stage of the 2SCF are a random sample from the population, this expression can be calculated using the estimators $\hat{\alpha}_z$ of the first stage of the 2SCF. Otherwise, the coefficients α_z could be calculated by re-estimating the first stage of the 2SCF using the attributes of synthetic dwellings \tilde{i} and the characteristics of synthetic households \tilde{n} . In both cases, $\hat{\delta}_{i\tilde{n}}^0$ has to be included then as an auxiliary variable in the utility, as shown in Eq. (2-6).

$$\hat{P}^1(\tilde{i}) = \frac{1}{\tilde{N}} \sum_{\tilde{n}=1}^{\tilde{N}} \frac{e^{\hat{\beta}_p p_{i\tilde{n}}^1 + \hat{\beta}_x x_{i\tilde{n}}^1 + \hat{\beta}_\delta \hat{\delta}_{i\tilde{n}}^0}}{\sum_{j \in C_{\tilde{n}}} e^{\hat{\beta}_p p_{j\tilde{n}}^1 + \hat{\beta}_x x_{j\tilde{n}}^1 + \hat{\beta}_\delta \hat{\delta}_{j\tilde{n}}^0}} \quad (2-6)$$

The application of this simulator may still be cumbersome because it requires the criteria used to build the instruments with the real data to be valid for the synthetic population. If the synthetic prices are reliable but the validity of the criteria used to build the instruments is uncertain or difficult to implement for the synthetic data, it would still be possible to generate a consistent estimator of the simulated probabilities by using the Logit Mixture model shown in Eq. (2-7), where $f(\delta|p)$ is the conditional distribution of δ given p .

$$\hat{P}^l(\tilde{i}) = \frac{1}{\tilde{N}} \sum_{\tilde{n}=1}^{\tilde{N}} \int \cdots \int \frac{e^{\hat{\beta}_p p_{\tilde{i}\tilde{n}}^1 + \hat{\beta}_x x_{\tilde{i}\tilde{n}}^1 + \hat{\beta}_\delta \delta_{\tilde{i}\tilde{n}}}}{\sum_{j \in C_{\tilde{n}}} e^{\hat{\beta}_p p_{\tilde{i}j}^1 + \hat{\beta}_x x_{\tilde{i}j}^1 + \hat{\beta}_\delta \delta_{\tilde{i}j}}} f(\delta|p) d\delta \quad (2-7)$$

In a practical application, the multifold integral shown in Eq. (2-7) can be calculated using Monte Carlo integration, where $f(\delta|p)$ can be inferred from the sample (provided it is random) by estimating the auxiliary regression

$$\hat{\delta}_{\tilde{i}\tilde{n}} = \gamma_0 + \gamma_p p_{\tilde{i}\tilde{n}}^0 + \lambda_{\tilde{i}\tilde{n}},$$

where the superscript 0 indicates that this model is estimated using data from the base year.

Then, for each synthetic dwelling \tilde{i} and household \tilde{n} , several draws r of δ should be obtained using the expression

$$\delta_{\tilde{i}\tilde{n}r} = \hat{\gamma}_0 + \hat{\gamma}_p p_{\tilde{i}\tilde{n}}^0 + \varepsilon_{\tilde{i}\tilde{n}r},$$

where $p_{\tilde{i}\tilde{n}}^0$ is the price of the synthetic dwelling in the estimation year, $\hat{\gamma}$ are the estimators of the auxiliary regression for $\hat{\delta}$, and $\varepsilon_{\tilde{i}\tilde{n}r}$ is a random draw distributed Normal $(0, \hat{\sigma}_\lambda^2)$, where $\hat{\sigma}_\lambda^2$ is the sample variance of the residual λ of the auxiliary regression. Then, the choice probability for each household is obtained by averaging across draws. Finally, the probability of each synthetic dwelling shown in Eq. (2-7) is obtained by averaging across synthetic households.

2.2.6 Comparison between 2SCF and 2SIV Methods

The great similarity between the 2SCF and the 2SLS method used in linear models raises the question of why (instead of replacing the residuals as additional variables) it would be incorrect to substitute the endogenous price with the fitted price and then re-estimate the

model. I will term this alternative method as the two-stage instrumental-variables (2SIV) method.

Formally, if the price p is replaced by \hat{p} in the utility function,

$$U_{in} = \beta_p p_{in} + \beta_x x_{in} + \xi_{in} + e_{in}$$

$$U_{in} = \beta_p \hat{p}_{in} + \beta_x x_{in} + \underbrace{(\beta_p + \beta_\delta) \hat{\delta}_{in}}_{\psi} + v_{in} + e_{in},$$

the remaining error of the model ψ will be compounded by v , e and $\hat{\delta}$. Note that all the terms in ψ are uncorrelated, by construction, to the observed variables of this auxiliary model: \hat{p} and x . This fact implies that 2SIV will result in consistent estimators of the model coefficients.

The fact that 2SIV is consistent has been rarely stated in the literature and caused some confusion among practitioners. Newey (1985a) gives a formal demonstration of this finding for a case equivalent to the one studied in this thesis. Finally, it should be noted that, as with the 2SCF, consistency is attained only up to a scale since the variance of ψ is different from the variance of $\xi + e$, what causes a change of scale that is unknown to the researcher.

Making assumptions about the distribution of ψ is complicated, but not more than with the 2SCF. If e follows an Extreme Value distribution, there is no parametric distribution of v or $\hat{\delta}$ that would make ψ follow any known distribution. However, if the sample is large enough, which is where the consistency results are relevant, those assumptions become plausible because the Law of Large Numbers can be claimed to affirm that ψ follows a Normal distribution.

However, there is an important difference between the 2SIV and 2SCF that finally tips the balance in favor of the latter in the correction for endogeneity in discrete choice models. The problem is that it is not clear how to forecast using the 2SIV method. An intuitive way to forecast would be to replace the new values of p into a model with the 2SIV estimators $\hat{\beta}$, as shown in Eq. (2-8). However, such a procedure would leave a term that depends on δ in the unobserved part of the model. Since δ is correlated with p , the estimators of the simulated probabilities will be inconsistent, for the same reason that the estimators of the simulated probabilities of the model shown in Eq. (2-5) were

inconsistent. The Monte Carlo experiments performed later in Section 2.3 give some empirical evidence to support this claim.

$$\hat{p}^1(i) = \frac{1}{N} \sum_{n=1}^N \frac{e^{\hat{\beta}_p p_{in}^1 + \hat{\beta}_x x_{in}^1}}{\sum_{j \in C_n} e^{\hat{\beta}_p p_{jn}^1 + \hat{\beta}_x x_{jn}^1}} \quad (2-8)$$

2.2.7 Efficiency and Calculation of Standard Errors with the 2SCF Method

The estimation of the 2SCF in two stages has two negative consequences. The first is that the estimators of this model are, in general, inefficient. Chapter 3 analyses the conditions required to achieve efficiency in this case. The second consequence of using two stages is that the standard errors cannot be calculated from the inverse of the Fisher-information-matrix. This prevents the direct application of hypothesis testing. The need for correcting the standard errors comes from the fact that the second stage of the method treats the residuals of the first stage as if they were error free, which they are not. This correction is not trivial and may easily overcome the simplicity attained from the estimation in two stages.

There are at least three alternatives for addressing this problem. Karaca-Mandic and Train (2003) derived a correction by calculating the asymptotic variance-covariance matrix of the 2SCF using the delta-method (Wooldridge, 2002) to account for the effect of both stages in the likelihood function. Another way to address this correction is to use non-parametric methods. The best alternative, in this case, is to bootstrap the observations of the first stage. According to Karaca-Mandic and Train (2003), the empirical results of their method are equivalent to those attained with bootstrapping. The third alternative is to estimate the model using maximum-likelihood, while simultaneously taking into account both stages of the 2SCF. In Chapter 3 I develop a maximum-likelihood estimator that is tractable (under mild conditions) and efficient in the correction for endogeneity in problems of residential location choice. This estimator also allows for the calculation of the standard errors directly from the inverse of the Fisher-information-matrix.

2.2.8 Testing for Endogeneity

Rivers and Vuong (1988) and Wooldridge (2002) noted that the 2SCF provides a practical way to test for the presence of endogeneity. Under the null hypothesis, where the model does not suffer from endogeneity, the coefficient of the residuals included in the second stage of the 2SCF is equal to zero, and the standard errors calculated from the inverse of the Fisher-information-matrix are correct. This implies that it is possible to test for endogeneity directly from the output of the 2SCF using a Quasi-t test, a Likelihood-ratio test or a LaGrange-multiplier test for the null hypothesis that the residuals are exogenous.

Formally, the Quasi-t test version of a test for endogeneity of price in the example examined throughout the chapter can be implemented in the following four stages:

Stage 1: Estimate $\hat{\delta}$ by ordinary-least-squares (OLS).

$$p_{in} = \alpha_z z_{in} + \delta_{in} \xrightarrow{OLS} \hat{\alpha}_z \Rightarrow \hat{\delta}_{in} = p_{in} - \hat{p}_{in} = p_{in} - \hat{\alpha}_z z_{in}$$

Stage 2: Estimate the choice model by maximum-likelihood (ML) using $\hat{\delta}$ as an additional variable.

$$U_{in} = \beta_p p_{in} + \beta_x x_{in} + \beta_\delta \hat{\delta}_{in} + \tilde{v}_{in} + e_{in} \xrightarrow{ML} \hat{\beta}_\delta$$

Stage 3: Estimate the variance-covariance matrix using the inverse of the Fisher-information-matrix.

$$\hat{\Sigma}_\beta = -E \left(\frac{\partial \ln P_n(\hat{\beta})}{\partial \beta \partial \beta'} \right)^{-1} \Rightarrow \hat{\sigma}_{\beta_\delta}$$

Stage 4: Calculate the Quasi-t test, which follows a Student distribution with $N-1$ degrees of freedom.

$$t = \frac{\hat{\beta}_\delta}{\hat{\sigma}_{\beta_\delta}} \sim t_{N-1}$$

When testing for the endogeneity of diverse variables the procedure is equivalent. The only difference is that the final stages are replaced by those required for the calculation of a Likelihood-ratio or a LaGrange-multiplier test.

2.3 Monte Carlo Experiment

2.3.1 Model Setting

In this section I develop a Monte Carlo experiment to analyze the impact of endogeneity in discrete choice models and to assess the effectiveness of 2SCF and 2SIV in estimation and forecasting. The true model considered in this experiment is a binary Logit with a latent utility that depends linearly on four attributes x_1 , x_2 , p and ξ , and an error term e independent and identically distributed (*iid*) Extreme Value (0,1). The coefficients of each attribute are shown in Eq. (2-9).

$$U_{in} = -2p_{in} + 1x_{1in} + 1x_{2in} + 1\xi_{in} + e_{in} \quad (2-9)$$

Variable p (price) is defined as a function of ξ , an instrument z , and an error term $\tilde{\delta}$ *iid* Uniform (-1,1), with the coefficients shown in Eq. (2-10). Variables x_1 , x_2 , ξ and z were generated as *iid* Uniform (-3,3). The synthetic database consists of 2,000 observations and was generated 100 times.

$$p_{in} = 5 + 0.5\xi_{in} + 0.5z_{in} + \tilde{\delta}_{in} \quad (2-10)$$

Note that by virtue of Eq. (2-10) variables p and ξ are correlated. Therefore, if ξ is omitted in the specification of the utility function, the choice model will suffer from endogeneity. In turn, since x_1 and x_2 are not correlated with other variables, the model will not suffer from endogeneity if x_1 or x_2 are omitted. Note also that z is, by construction, a valid instrument. From Eq. (2-10) z is correlated with p and independent of e .

2.3.2 Estimation with 2SCF and 2SIV Methods

To assess the impact of endogeneity in the estimation of the model parameters and to evaluate the performance of the 2SCF and 2SIV methods studied to address it, five models were estimated for each repetition of the Monte Carlo experiment: the true model, a model where x_1 is omitted, a model where ξ is omitted, and two models where ξ is omitted but the problem is addressed using the 2SCF and the 2SIV methods.

For each model, the average, bias, mean squared error (MSE) and the t-test against the true values of the estimators of the model parameters are reported in Table 2-1. The

use of repetitions avoids the risk of dealing with a singular case that may bias the analysis and avoids the need for correcting the standard errors required in the application two-stage procedures.

Table 2-1 Monte Carlo Experiment: Model Estimation with 2SCF and 2SIV

| | Metric | $\hat{\beta}_p$ | $\hat{\beta}_{x_1}$ | $\hat{\beta}_{x_2}$ | $\hat{\beta}_\xi$ | $\hat{\beta}_\delta$ | $\hat{\beta}_\rho$ | $\hat{\beta}_p / \hat{\beta}_{x_2}$ |
|----------------|----------------|-----------------|---------------------|---------------------|-------------------|----------------------|--------------------|-------------------------------------|
| True Model | Average | -1.990 | 0.9960 | 0.9949 | 0.9957 | | | -1.980 |
| | Bias | 0.009561 | -0.004032 | -0.005127 | -0.004288 | | | 0.02022 |
| | MSE | 0.008985 | 0.003247 | 0.002755 | 0.002990 | | | 0.2148 |
| | t-test true | 0.1014 | -0.07094 | -0.09814 | -0.07867 | | | 0.04366 |
| Omitting x_1 | Average | -1.122 | | 0.5627 | 0.5641 | | | -1.998 |
| | Bias | 0.8778 | | -0.4373 | -0.4359 | | | 0.002259 |
| | MSE | 0.7742 | | 0.1923 | 0.1913 | | | 0.2550 |
| | t-test true | 14.53 | | -13.61 | -12.03 | | | 0.004473 |
| Omitting ξ | Average | -0.7994 | 0.6675 | 0.6689 | | | | -1.212 |
| | Bias | 1.201 | -0.3325 | -0.3311 | | | | 0.7881 |
| | MSE | 1.443 | 0.1119 | 0.1108 | | | | 0.7276 |
| | t-test true | 26.80 | -8.873 | -9.359 | | | | 2.415 |
| 2SCF | Average | -1.563 | 0.7813 | 0.7825 | | 1.078 | | -1.992 |
| | Bias | 0.4372 | -0.2187 | -0.2175 | | | | 0.008215 |
| | MSE | 0.1983 | 0.04955 | 0.04884 | | | | 0.2581 |
| | t-test true(*) | 5.161 | -5.277 | -5.531 | | 13.09(*) | | 0.01617 |
| 2SIV | Average | | 0.7208 | 0.7192 | | | -1.440 | -1.980 |
| | Bias | | -0.2792 | -0.2808 | | | 0.5598 | 0.01956 |
| | MSE | | 0.07924 | 0.08017 | | | 0.3189 | 0.2872 |
| | t-test true | | -7.788 | -7.713 | | | 7.512 | 0.03652 |

100 Repetitions. $N=2,000$. $J=2$. (*) t-test against zero for $\hat{\beta}_\delta$

The first row below the labels in Table 2-1 shows the estimators obtained from the true model. In this case all estimators of the model parameters are statistically equal (with 95% confidence) to their true values. The second row shows the estimators of the model that omits variable x_1 . This model does not suffer from endogeneity because x_1 is not correlated with other variables. The estimators in this case are consistent, but only up to a scale. It should be noted that the ratio between the coefficients of p and x_2 is statistically equal (with 95% confidence) to its true value (-2). In turn, each coefficient is significantly different from its respective true value. This is explained by the change of scale caused by the addition of the variance of x_1 to the error of the model. The change of scale

observed in Table 2-1 is of approximately 0.56, a value that can be approximately calculated by substituting, in Eq. (2-3), the variance of x_1 by σ_v^2 . Finally, the omission of variable x_1 reduced the efficiency of the estimators. This can be noted in the increase of the MSE of the estimator of the ratio between the coefficients of p and x_2 for this model, when compared to the respective MSE of the true model.

The third row in Table 2-1 shows the estimators that are obtained when ξ is omitted. This model suffers from endogeneity because ξ is correlated with p . In this case the estimators are different from those of the true model, but not only up to a scale. The ratios between coefficients are also affected. Since p and ξ are positively correlated, the omission of ξ causes a positive bias in the coefficient of p . Consequently, the ratio between the coefficients of p and x_2 is approximately -1.2 instead of -2, as it was in the true model. Intuitively, the problem is that positive shocks of ξ on the utility are confounded as the results of shocks of p , causing a positive bias in the estimator of the coefficient of p .

Consider now the case of the model that omits ξ , but is corrected using the 2SCF method. Note first that the estimator of the auxiliary variable is statistically different (with 95% confidence) from zero. This correctly confirms that endogeneity was present in the model without the correction. Second, although the model coefficients are not numerically equal to those of the true model, the ratios between them are the same. Particularly, the ratio between the coefficient of p and x_2 is again statistically equal (with 95% confidence) to -2. The change of scale between the estimators in this case is approximately 0.78, shift that can be calculated by considering the variance of v in Eq. (2-3). Lastly, similarly to what occurred with the omission of x_1 , although the correction for endogeneity resulted in consistent estimators up to a scale, the fact that the term v was omitted caused a reduction in efficiency. This can be noted in the increase of the MSE of the estimator of the ratio between the coefficients of p and x_2 for this model, when compared to the respective MSE of the true model.

Finally, consider the model that corrects for endogeneity using the 2SIV procedure, which is shown in the last row of Table 2-1. Equal to what occurred with 2SCF, although the scale of the model is different to that of the true model, the ratios between the coefficients are statistically equal (with 95% confidence) to that of the true model. This

confirms that 2SIV succeeds in correcting for endogeneity, as it was originally shown by Newey (1985a). Additionally, it can also be noted that the MSE of this model is larger than that of the true model, which results from the fact that this model is less efficient.

2.3.3 Forecasting with 2SCF and 2SIV Methods

The next step in the analysis of this Monte Carlo experiment is to show how the different models behave in the forecasting or simulation phase. To do so, I first use the estimators of the different models to calculate the ASE of price and the Aggregated Direct Elasticity (ADE) of price (Ben-Akiva and Lerman, 1985). The expressions for ASE and ADE of price, for a given alternative i , are the following:

$$\begin{aligned}\overline{\text{ASE}}_p(i) &= \frac{1}{N} \sum_{n=1}^N (1 - P_n(i)) P_n(i) \beta_p \\ \overline{\text{ADE}}_p(i) &= \frac{\beta_p}{\sum_{n=1}^N P_n(i)} \sum_{n=1}^N (1 - P_n(i)) P_n(i) p_{in}\end{aligned}\quad (2-11)$$

The experiment was repeated 100 times. Table 2-2 reports the average and standard errors of the ASE and ADE for $i=1$ across the repetitions. Additionally, I simulated the effect of increasing the price of alternative 1 by 50% for all n 's and calculated the average probability of choosing alternative 1 across the 2,000 observations, before ($\hat{P}^0(i)$) and after ($\hat{P}^1(i)$) the price shift.

The true model works as the benchmark. Table 2-2 shows that, in this case, the 50% increase in the price of alternative 1 resulted in a reduction of its choice probability from approximately 50% to 19%, a 31% reduction. Additionally, the ASE is approximately -0.16% and the ADE is approximately -1.6% in this case.

The results of the model where variable x_1 is omitted are concordant with the conclusions attained by Cramer (2007) and Daly (2008) about omitted orthogonal attributes in Logit models. Although this model resulted in an important change of scale, as it was noted in Table 2-1, the forecasting probabilities of the model, as well as the ASE and the ADE, are statistically equal (with 95% confidence) to those of the true model.

Instead, the results are very different when variable ξ is omitted. In this case there is an underestimation of approximately 10% of the change in the probability of choosing alternative 1 when its price is raised by 50%. The ASE and ADE are also significantly affected.

Table 2-2 Monte Carlo Experiment: Forecasting with Endogeneity Correction

| Model | ASE _p (1) | ADE _p (1) | $\hat{P}^0(1)$ | $\hat{P}^1(1)$ |
|--|------------------------|----------------------|----------------------|----------------------|
| True Model | -0.1610 (0.00470) | -1.608 (0.05922) | 0.5009 (0.00896) | 0.1850 (0.008616) |
| Omitting x_l | -0.1603 (0.00496) | -1.600 (0.05868) | 0.5013 (0.007275) | 0.1871 (0.008294) |
| Omitting ξ | -0.09632 (0.004146) | -0.962 (0.04520) | 0.5010 (0.007406) | 0.2865 (0.01007) |
| 2SCF Adding $\hat{\delta}$ | -0.1610 (0.006719) | -1.608 (0.07726) | 0.5013 (0.008182) | 0.1852 (0.01076) |
| 2SCF Scale Adjustment | -0.1363 (0.004187) | -1.362 (0.05051) | 0.5012 (0.008292) | 0.2260 (0.009275) |
| 2SCF Logit Mixture | -0.1612 (0.004393) | -1.613 (0.07539) | 0.5013 (0.007791) | 0.1844 (0.01058) |
| 2SIV | -0.1384 (0.004731) | -1.382 (0.05562) | 0.5013 (0.008451) | 0.2232 (0.009781) |

Standard errors in parenthesis. 100 Repetitions. $N=2,000$. $J=2$.

Consider now the models corrected for endogeneity caused by the omission of ξ . In the case of 2SCF, three alternatives to doing forecasting were analyzed. Table 2-2 shows that when the $\hat{\delta}$ used for estimation is also included as auxiliary variable during forecasting (Eq. 2-4), the results of the simulation of the 2SCF are indistinguishable from those of the true model. In turn, when $\hat{\delta}$ is not included in forecasting, but the scale is adjusted (Eq. 2-5), as it was suggested by Wooldridge (2002), there is a significant bias in the forecast. In this case the effect of the price shift in the choice probabilities is underestimated by approximately 4% and the elasticity is consequently underestimated by approximately 0.2%. Instead, when the Logit Mixture method described in Eq. (2-7) is used for forecasting, the results for the simulated probabilities are again statistically equal (with 95% confidence) to those obtained with the true model. The same occurs with the ASE and the ADE.

Finally, consider the results from the application of the 2SIV method in correcting for endogeneity (Eq. 2-8), which are shown in the last row of Table 2-2. It can be seen that the simulated probabilities, the ASE, and ADE are significantly different from those of the true model. For this example the bias results in a significant underestimation of the shift in the choice probability due to the change in prices. This means that, even though both 2SIV and 2SCF achieve the consistent estimation of the model coefficients up to a scale, their performance in the forecasting phase shows the latter to be more effective for use in models of discrete choice.

In summary, this Monte Carlo experiment showed first that the omission of an orthogonal attribute causes a change of scale in the estimated coefficients but it does not impact the ratio between the coefficients or the forecasting properties of the model. This same result also holds for the application of the 2SCF method in correcting for endogeneity. It was also shown that the 2SIV method results in the consistent estimation of the model coefficients up to a scale, but that the forecasting properties of the model are significantly worse. Finally it was shown that the best alternative for forecasting with the 2SCF method is to include the residuals estimated in the first stage into the utility. In cases where the residuals are unavailable, they can be calculated from respective instruments using the estimators of the first stage of the 2SCF, or simulated using the expression shown in Eq. (2-7). Instead, the alternative of simply adjusting the scale when the residuals are omitted in forecasting was shown to have poor simulation properties.

2.4 Application to Real Data

2.4.1 Overview

In this section, I use a case study based on real data to investigate and demonstrate the properties of the 2SCF method in correcting for endogeneity in discrete choice models of residential location. I begin by describing the construction of the database used for estimation. Then I describe the logic used in the construction of the instrumental variables and show and analyze the results of the application of the 2SCF and its effects in forecasting.

The case study is situated in the Portuguese municipalities of Lisbon, Odivelas and Amadora, which are located at the center of the Lisbon Metropolitan Area (LMA). The LMA is an urban system that includes the Portuguese capital city (municipality) of Lisbon and 17 surrounding municipalities. LMA covers approximately 3,000 km² and has 2,5 million inhabitants, of which approximately 20% live in Lisbon's municipality.

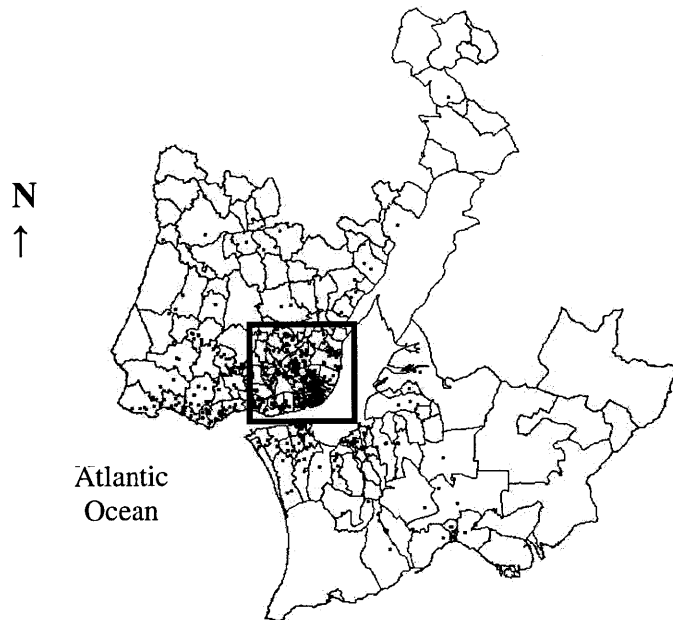
A comprehensive representation of household's location choice behavior might involve modeling their long-term plans regarding lifestyle, career and ownership choices, and account for the interactions among all household's members. The scope used in this research is simpler since the main purpose is to account particularly for endogeneity. It is assumed that the decision-maker is a household with certain characteristics and that it chooses among a set of available dwellings with certain attributes. It is also assumed that the underlying choice model is Logit. In Chapter 5 I analyze the impact of relaxing the Logit assumption in this context.

2.4.2 Construction of the Database for Estimation

The data to estimate the model was constructed using the combination of two sources. The first source was a small convenience online survey (SOTUR) conducted in 2009 by Martinez et al. (2010). This survey collected information on residential location, choice preferences, attitudes and household characteristics from 750 households across the entire LMA. Of the 750 observations from the SOTUR survey, only 342 are potentially useful for the estimation of the residential location choice model. Almost 50% of the observations were eliminated because they did not include information on household income or dwelling price. The rest were excluded because they corresponded to dwellings that were not traded in the open market, including cases where the dwellings were inherited, provided by an institution, or borrowed and/or rented under special conditions through friends or relatives.

Household's characteristics collected using the SOTUR survey include household size, level of education, monthly income (by ranges), and work location of the head-of-the-household. Dwelling attributes in the survey include age and price (both by ranges), area, number of bedrooms and location.

Figure 2-1 shows the LMA with zonal divisions at the level of the Freguesia, which are aggregations of census blocks. Each black square in Figure 2-1 represents the location of one of the 750 households interviewed using the SOTUR survey. The frame in the center corresponds approximately to the municipalities of Lisbon, Odivelas and Amadora, and is shown with more detail later in Figure 2-2.



**Figure 2-1 SOTUR Observations in Lisbon Metropolitan Area (LMA)
Zoning by Freguesia**

The information from the SOTUR survey can be used as the source for the characteristics of the households and their revealed choice but not as the source for the non-chosen alternatives. The reason is that the survey is not a random sample of the available dwellings in the market. Instead, the survey can be seen as a probability sample that was developed using a sampling protocol based on the choice probability. If the choice probabilities were known, it would be possible to draw non-chosen alternatives from the same survey and achieve consistent estimation of the model parameters by applying the sampling correction method proposed by McFadden (1978), which is described, in another context, in Section 5.2. However, the choice probabilities are unknown beforehand; thus eliminating this method as a viable option.

One way to avoid this limitation is to gather the attributes of the non-chosen alternatives from an independent source. The source used in this application is a snapshot of the dwellings that were advertised for sale in February 2007 within the municipalities of Lisbon, Odivelas and Amadora. The data was collected by Imokapa (www.imokapa.com) and is reported in detail by Martinez and Viegas (2009). The data contains attributes from 12,358 dwellings, including type, area, age, location and respective asking price. Over 70% of the observations belong to the Lisbon municipality.

Figure 2-2 corresponds approximately to the frame shown in the center of Figure 2-1. It shows the contrast between the observations from the SOTUR survey (black squares) and the data from Imokapa (grey stars) within the LMA sector covered by Imokapa.



Figure 2-2 SOTUR (■) and Imokapa (★) Observations in Lisbon, Odivelas and Amadora

Although the combined use of the two sources of data overcomes the problem of the unknown probabilities in the sampling protocol, it causes a different problem at the same time. Given that the two databases are from different years, cover different areas of the city and have different stratifications of dwelling attributes, their combination requires matching the observations of the SOTUR survey onto those of the Imokapa database.

The problem of multivariate matching has been extensively studied in diverse literature. Although various methods have been proposed, there is no consensus on which one is the most appropriate to address the matching problem since each procedure depends importantly on a set of usually unverifiable assumptions (Sekhon, 2010).

In this application, I address the matching problem using the nearest-neighbor approach (see, e.g., Duda et al., 2001), which can be stated as follows. First, ranges of acceptable discrepancies between the two databases are defined for each variable. This is needed because a perfect match between the two databases is almost impossible given that each dwelling is a quasi-unique combination of diverse attributes. Then, for each observation in the SOTUR survey and its respective range of variables, all dwellings from the Imokapa database falling into that range are identified. If no dwellings from Imokapa fall into the respective range of variables of the SOTUR observation, that record from the SOTUR survey has to be discarded. If several dwellings from Imokapa fall into the range, the match is defined for the nearest-neighbor, using some measure of distance.

The variables used in the matching process were four: the price, the age, the location, and the area of the dwelling. The discrepancies used for the first two matching variables (dwelling price and age) were defined as the ranges of those variables in the SOTUR survey, with the respective adjustments for inflation and for the year the data was collected.

The discrepancies allowed for dwelling location and dwelling area were determined by trading off the number of observations discarded and the stability of the estimators of the model coefficients obtained using the resulting database. In the case of dwelling location, the matching was enforced only at the level of the Freguesia, and in the case of dwelling area, discrepancies of up to 25 square meters were allowed.

Finally, for cases where one SOTUR dwelling was assigned to more than one Imokapa dwelling, the approach used was to assign the match to the Imokapa dwelling that was geographically closer to the SOTUR observation under analysis.

The sole application of the matching criteria by geographic area reduced the number of observations from 342 to 178. The application of the other matching criterion resulted in a subset of only 66 valid observations from the SOTUR survey being matched into the Imokapa database.

The geographical component of this matching process is summarized in Figure 2-3. The black squares correspond to the SOTUR dwellings, and the grey stars correspond to the Imokapa dwellings that were matched. It can be noted that for some observations the geographical matching was almost perfect, whereas in other cases, the fact that location was only enforced at the level of the Freguesia, resulted in some non-negligible differences.

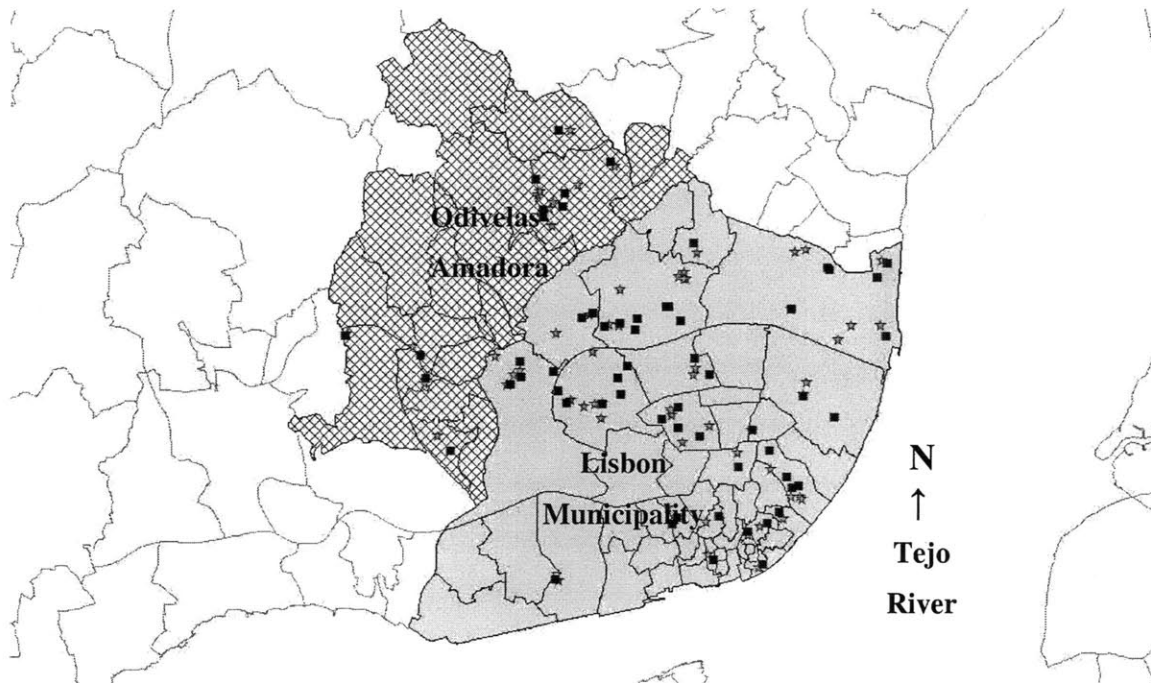


Figure 2-3 Matching of Dwellings from SOTUR (■) into IMOKAPA (★)

The application of the criteria to construct valid instrumental variables, which is described later in Section 2.4.3, further reduced the database available for estimation of the residential location choice models. The final database is compounded of 11,501 alternatives, from which only 63 correspond to chosen dwellings. The main descriptive statistics of the database are shown in Table 2-3.

Regarding dwelling attributes, Table 2-3 shows that dwellings from the Lisbon municipality tend to be more expensive and older than those from Odivelas and Amadora, although the differences are not statistically significant (with 95% confidence). Also, the dwellings from both regions have approximately equal area. Finally, dwellings from Lisbon are significantly closer, in average, to the workplace of the head-of-the-

households of the sample. Table 2-3 also shows the distribution of household location, classified by income. It should be noted that 51 out of 63 households reside in Lisbon municipality and that the larger share of households in the sample have an income that is between 2,000 and 5,000 Euros per month (€/M).

Table 2-3 Summary of Lisbon's Residential Location Choice Database for Estimation

| Municipality | Average Dwelling Attributes (Standard Deviation) | | | | Total Dwellings Available | Household Location | | | |
|-------------------------------------|---|----------------------------------|---------------------------|------------------|---------------------------------|---------------------------|------------------------------------|---------------------------|------|
| | Price 100,000 [€] | Distance to Workplace [Km] | Area [m ²] | Age [Years] | | Income <2,000 [€/M] | Income 2,000- 5,000 [€/M] | Income >5,000 [€/M] | Tot. |
| Lisbon | 2.356 (1.354) | 4.508 (2.389) | 99.30 (41.77) | 39.93 (36.21) | 8,018 | 16 | 28 | 7 | 51 |
| Odivelas and Amadora | 1.680 (0.8365) | 10.581 (1.253) | 98.44 (32.59) | 32.17 (31.68) | 3,483 | 5 | 7 | 0 | 12 |
| Total | 2.151 (1.260) | 6.347 (3.499) | 99.01 (39.22) | 37.58 (35.08) | 11,501 | 21 | 35 | 7 | 63 |

€/M: Euros per month

A final remark on the limitation of this database has to be acknowledged. First of all, although the number of alternatives in the choice set is very large and is a good representation of the housing market in the modeling area, it is not a cadastre and it does not correspond to the alternatives really faced by the households in the sample. Second, the number of observations available for estimation is small, and they were collected in a convenience online survey. These facts make the models that can be estimated from this database, susceptible to important biases.

In consequence, the models estimated using this database should be seen as preliminary in nature, as a proof of concept of the methodologies addressed in this research in addition to the empirical evidence gathered from Monte Carlo experimentation. Nevertheless, even with the small sample size and other limitations of the database, it worth noting that the models estimated using this database did provide significant evidence for most of the issues studied in this research, and shed important light about the practical issues associated with them.

2.4.3 Instrumental Variables

The quasi-uniqueness of dwelling-units and the limited capacity of the researcher in accounting for all the dwelling attributes shall cause price endogeneity in residential location choice modeling. To test and correct for endogeneity it is necessary to gather instruments, auxiliary variables that have to be relevant (correlated with dwelling price) and valid (uncorrelated with the omitted attributes). The instrumental variables proposed for this case study were constructed from the prices of other dwellings with similar observed attributes (other than price) and locating within certain vicinity. I begin by stating the logic used to sustain that such instrumental variables are valid and relevant, and then deploy the practical implementation of this logic for Lisbon's case study.

The first assumption required to sustain the validity of prices of other dwellings as instrumental variables rests in considering that endogeneity, caused by the simultaneous determination of dwelling price and household choice, is not a significant issue in microscopic modeling. As stated before in Section 2.2.1, this statement is supported by the fact that the behavior of a single household does not impact the price of any specific dwelling, contrasting with the impact that aggregated demand has on aggregated supply in the housing market. Under this assumption, the error term ε_{in} of alternative i will not be correlated (because of simultaneous determination) with the price p_{jn} of alternative j . This implies that the price of a dwelling j located nearby dwelling i can be used to construct valid instruments for the price of i .

The relevance of prices of other dwellings as instrumental variables is sustained by the existence of spatial autocorrelation, or what is known as the "first law of geography: everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). To formalize the argument, consider that the set $V(i)$ contains all dwellings j that are typologically and geographically near i . Then, if the price of the dwelling i is correlated with the price of the dwellings in $V(i)$, the following hedonic price equation can be formulated:

$$p_{in} = \rho_{p_{ij}} p_{jn} + \xi_{in} + v_{p_{in}} \quad \rho_{p_{ij}} \begin{cases} \neq 0 & j \in V(i) \\ = 0 & o/w \end{cases},$$

where v_p is an independent error term and ξ represents the omitted attributes that cause endogeneity in the model shown in Eq. (2-1). The coefficient ρ_p needs to be significantly different from zero to guarantee the relevance of the instruments. This need implies that the elements in $V(i)$ should be typologically and geographically near enough to dwelling i . Otherwise, the model may suffer from the weak instruments problem (see, e.g., Hahn and Hausman, 2002, and Stock et al., 2002). This issue is discussed further in Section 2.4.4.

It is worth noting what occurs when spatial autocorrelation impacts not only the prices of dwelling-units but also the error terms of the model. The reason is that it can be hypothesized that, when using prices of other dwellings as instrumental variables, the effects of spatial autocorrelation and endogeneity may be confounded.

Consider first that the independent error term e of the choice model shown in Eq. (2-1) is also spatially autocorrelated, such that

$$e_{in} = \rho_{e_{ij}} e_{jn} + v_{e_{in}} \quad \rho_{e_{ij}} \begin{cases} \neq 0 & j \in V(i) \\ = 0 & o/w \end{cases},$$

where v_e is an independent error term. In this case, spatial autocorrelation of e needs to be addressed, for example, using a Logit Mixture model, but this effect will not be confounded or affected by the endogeneity problem. The prices of nearby dwellings can still be used as instruments because p_{jn} will still be independent of $\xi_{in} + v_{p_{in}}$.

In turn, if the error term ξ , which represents the omitted attributes that cause endogeneity, is spatially autocorrelated, such that

$$\xi_{in} = \rho_{\xi_{ij}} \xi_{jn} + v_{\xi_{in}} \quad \rho_{\xi_{ij}} \begin{cases} \neq 0 & j \in V(i) \\ = 0 & o/w \end{cases},$$

the prices of nearby dwellings could not be used as instruments. The orthogonality between p_{jn} and ξ_{in} would be broken and then the estimators of the first stage of the 2SCF would be inconsistent. This problem can be shown by noting that

$$\begin{aligned} p_{in} &= \rho_{p_{ij}} p_{jn} + \xi_{in} + v_{p_{in}} \\ p_{in} &= \rho_{p_{ij}} p_{jn} + \rho_{\xi_{ij}} \xi_{jn} + v_{\xi_{ij}} + v_{p_{in}} \end{aligned}$$

The model fails because p_{jn} and ξ_{jn} are correlated in Eq. (2-1).

The problem that arises when ξ is spatially auto-correlated, falls into what Manski (1993) termed “reflection bias”. In practice, if two dwellings are too near, they can share some attributes that are omitted by the researcher, such as being close to a gas station or another firm that causes some type of externality. In that case the price of one dwelling cannot be used as an instrument for the price of the other since both prices might be correlated with the same omitted attribute and, therefore, with the same error term. One way of avoiding the reflection bias in spatial choice models of residential location is then to exclude from the set of potential instruments the dwellings that are too close to the dwelling for which instruments are sought.

The application of the 2SCF method under the effect of the reflection bias might be misleading for the researcher. Although the estimators obtained from that procedure will be inconsistent, the coefficient of the auxiliary variable $\hat{\delta}$ in the second stage of the 2SCF is likely to be statistically significant (with 95% confidence) because it will capture part of the spatial autocorrelation of the model. The significance of $\hat{\delta}$ may mislead the researcher, who may interpret the significance of the residuals as resulting from a successful correction for endogeneity. In that sense, the tests for the validity of instruments, studied in Chapter 4, become a critical tool to assess correctly the overall validity of the model.

In summary, a suitable logic to construct valid and relevant instruments for dwelling price is to use the prices of dwellings that are typologically and geographically near to the dwelling for which instruments are sought (to avoid the weak instruments problem), but are, at the same time, beyond certain threshold (to avoid the reflection bias problem). Formally, defining $V(i)$ as the set of all the dwellings that are geographically and typologically near enough to dwelling i , and terming $v(i)$ the subset of $V(i)$ containing the dwellings that are geographically closer to i , instruments z_i can be selected as the prices of any dwelling j in the set $V(i) \setminus v(i)$

$$z_i = p_j \quad j \in V(i) \setminus v(i).$$

The practical implementation of the logic to gather instruments for the residential location choice model for Lisbon has diverse components. First, to avoid the reflection bias, the instruments were gathered from the prices of dwellings located beyond 500

meters from the dwelling for which instruments are sought. Provided enough data were available, the suitability of this 500 meters threshold could be formally validated using the techniques deployed in Chapter 4 to test for the validity of instruments. In this application, decreasing the threshold reduced only slightly the significance of the null hypothesis that the instruments were valid. Even though, I decided to maintain the 500 meters limit because 5 blocks appears as a conservative and qualitatively reasonable limit beyond which unobservable local effects may become insignificant.

To guarantee the relevance of the instruments; that is, to guarantee their correlation with price, dwellings located beyond 5,000 meters from the dwelling for which instruments are sought, were excluded from $V(i)$. The 5,000 meters external limit was determined by trading off the number of alternatives left in the model and the adjustment of the first stage of the control-function method. The trade-off arises in this case because, on the one hand, the tighter the external limit becomes, the more likely it may be necessary to discard some alternatives because it may not be possible to find appropriate instruments complying with the defined limit. On the other hand, the more relaxed the external limit becomes, the lower the correlation between the instruments and the endogenous variable may become, leading potentially to a weak instruments problem.

Besides the need for having instrumental variables that are correlated with the endogenous variable and uncorrelated with the error term, in Chapter 4 is shown that testing for the validity of instruments becomes possible only when there are more instruments than endogenous variables, and when those instruments are not highly correlated among them. Therefore, two instruments (z_1 and z_2) were built for each observation in the Imokapa database. The first instrument z_1 was constructed as the average price of dwellings located within 500 and 2,500 meters from the dwelling for which instruments are sought, and which area and age differed less than 10% from it. Equivalently, the second instrument z_2 was constructed as the average price of dwellings located within 2,500 and 5,000 meters from the dwelling for which instruments are sought, and which area and age differed more than 10% but less than 40% from it. This setting for the instruments was determined by trading off a high correlation of z_1 and z_2 with the endogenous price, and a low correlation among the instruments.

Table 2-4 shows the variance-covariance matrix of dwelling price and their respective instruments z_1 and z_2 . Complementing Table 2-4, Figure 2-4 shows a plot of dwelling price against their respective instruments. It can first be noted that both instruments are relevant; that is, both are significantly correlated with the endogenous variable. Whether or not this correlation is large enough to avoid the weak instruments problem is an issue that will be discussed later in Section 2.4.4. Additionally, Table 2-4 shows that the correlation between z_1 and z_2 is approximately 82%. Although this is a relatively high correlation, it is still in a range where the power of the tests for the validity of instruments were not severely impacted, as it is later shown in the Monte Carlo experiments deployed in Chapter 4.

Table 2-4 Correlation Matrix of Dwelling Price and Instrumental Variables

| <i>Corr</i> | <i>Price</i> | z_1 | z_2 |
|--------------|--------------|--------|--------|
| <i>Price</i> | 1.000 | 0.8127 | 0.7443 |
| z_1 | 0.8127 | 1.000 | 0.8238 |
| z_2 | 0.7443 | 0.8238 | 1.000 |

Table 2-4 also shows that the fact the z_1 was built from dwellings that were typologically and geographically closer to the dwelling for which instruments are sought, makes z_1 more correlated with the endogenous variable, compared to z_2 . This consequently results in a larger slope in the plots of dwelling price against z_1 , than against z_2 , as shown in Figure 2-4.

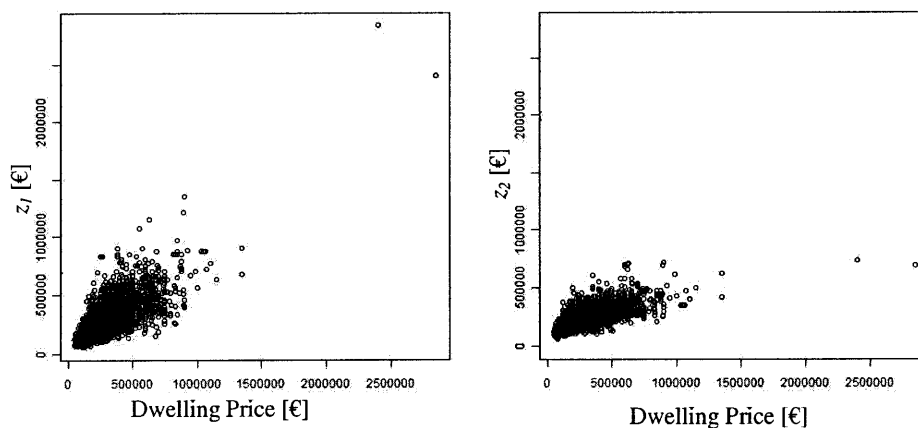


Figure 2-4 Dwelling Price and Instrumental Variables

Finally, it should be remarked that the thresholds defined to construct the instrumental variables are not hard constraints. A slight modification of the thresholds will qualitatively not impact the estimates of the model parameters, which are therefore robust to marginal changes in the implementation of the criterion proposed to construct the instrumental variables.

2.4.4 Estimation Using the 2SCF Method

In this section I present the estimation results for the residential location choice model of Lisbon. The specification considered is a Logit model where the systematic utility is linear for the following variables: dwelling price in 100,000 Euros (€), the distance from the dwelling to the workplace of the head-of-the-household in kilometers (Km), the log of dwelling area in square meters (m²), and the log of dwelling age in years (+1). Dwelling price was interacted with household income, which was stratified in three levels defined by the thresholds of 2,000 and 5,000 €/Month. The data consists of 63 observations, each one with the same choice set of 11,501 available dwellings. The estimators of the models, with and without the correction for endogeneity using the 2SCF method, are shown in Table 2-6.

The first stage in the application of the 2SCF method corresponds to the regression of the endogenous variable (price) on the instruments (z_1 and z_2). The results of this auxiliary regression are shown in Table 2-5.

Table 2-5 Lisbon's Logit Model: First Stage of 2SCF

| Variables | $\hat{\alpha}$ | s.e |
|-----------------|----------------|---------|
| Intercept | -3.023E+04 | 2450 |
| z_1 | 0.6995 | 0.01052 |
| z_2 | 0.483 | 0.01935 |
| R^2 | 0.6779 | |
| Adjusted R^2 | 0.6778 | |
| Sample Size N | 11,501 | |
| F | 1.210E+04 | |

The adjustment of the regression of the first stage of the 2SCF is highly relevant. If the instruments are exogenous but are not correlated enough with the endogenous

variable, the correction for endogeneity may worsen the model. This is known as the weak instruments problem, an issue that has been intensively studied for linear models. Hahn and Hausman (2002) showed that, for linear models, the strength of the instruments should be assured if the R^2 of the auxiliary regression is larger than 0.4. Also for linear models, Stock et al. (2002) suggested a threshold defined by an F test of more than 20 for each endogenous variable, to assure the strength of the instruments.

To the best of my knowledge, there is no systematic study of the weak instruments problems in a discrete choice framework. However, all Monte Carlo experiments estimated in this research confirmed that the thresholds established by Hahn and Hausman (2002) and Stock et al. (2002) were also appropriate for Logit models. Given that Table 2-5 shows that both the F and the R^2 criteria are surpassed in this case, we can affirm that there is evidence that the Lisbon's model does not suffer from the weak instruments problem.

The second stage of the 2SCF correction corresponds to the estimation of a residential location choice model that includes the residuals $\hat{\delta}$ as additional variables in the systematic utility. The results of this estimation are shown in the third column of Table 2-6 as a benchmark. The results of the model without the correction for endogeneity are shown in the second column of Table 2-6.

First of all, it should be noted that in Table 2-6, the signs of the coefficients of the model with and without the correction for endogeneity are as expected. The coefficient of dwelling area ($\hat{\beta}_5$) is positive, meaning that households prefer larger dwellings. The contrary occurs with dwelling price ($\hat{\beta}_1$), age ($\hat{\beta}_6$), and distance to workplace of the head-of-the-household ($\hat{\beta}_4$), which are perceived negatively. Also, the impact of dwelling price decreases with household income since $\hat{\beta}_2, \hat{\beta}_3 > 0$.

For the model without the correction for endogeneity, reported in the second column of Table 2-6, the coefficient of price for the wealthiest households (income over 5,000 €/month) is negative ($\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = -0.4270$) but small and with low statistical significance (t-test = -1.063). Arguably, this results from the omission of attributes that are correlated with dwelling price, causing endogeneity.

Table 2-6 Lisbon's Logit Model: With and without Correction for Endogeneity

| Variables | Without Endogeneity Correction | | With Endogeneity Correction | |
|---|--------------------------------|---------|-----------------------------|---------|
| | $\hat{\beta}$ | s.e | $\hat{\beta}$ | s.e |
| 1. Dwelling price (in 100,000 €) | -2.008 | 0.5150 | -2.811 | 0.6344 |
| 2. Dwelling price * 1[Income > 2,000 €/M] | 0.8136 | 0.5340 | 0.8542 | 0.5485 |
| 3. Dwelling price * 1[Income > 5,000 €/M] | 0.7674 | 0.4668 | 0.8089 | 0.4779 |
| 4. Distance to Workplace (in Km) | -0.2203 | 0.05064 | -0.2565 | 0.05335 |
| 5. Log [Dwelling Area (in m ²)] | 1.019 | 0.4982 | 2.232 | 0.7326 |
| 6. Log [Dwelling Age (in years) +1] | -0.3508 | 0.1076 | -0.4607 | 0.1192 |
| 7. $\hat{\delta}$ Control-function Auxiliary Variable | | | 1.054 | 0.4600 |
| Log likelihood at Convergence $L(\hat{\beta})$ | -563.00 | | -560.05 | |
| Log likelihood at Zero $L(0)$ | -589.06 | | -589.06 | |
| Adjusted ρ^2 | 0.05443 | | 0.06113 | |
| Sample Size N | 63 | | 63 | |
| Choice-set Size J | 11,501 | | 11,501 | |

Logit Model combining Imokapa database and SOTUR survey for Lisbon, Odivelas and Amadora
 Model estimated using the 2SCF method. Standard errors calculated by bootstrapping. €/M: Euros per month

Consider the model with the correction for endogeneity reported in the third column of Table 2-6. First of all, it should be noted that the coefficient of the auxiliary variable $\hat{\delta}$ is statistically significant (t-test=2.291). This confirms that endogeneity was present in the model before the correction. Additionally, the correction for endogeneity significantly changed the estimated coefficients. The coefficient of price for the wealthiest households is now more negative ($\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = -1.148$) and statistically different from zero (t-test=-2.172). Other model coefficients were also affected by the correction of price endogeneity, particularly the coefficient of dwelling area. This is because dwelling area and price are highly correlated (correlation = 0.7013) compared to other attributes, and then the impact of price endogeneity is significantly transferred to the coefficient of dwelling area. In general, the correction for price endogeneity resulted in a model that is more sensitive, not only to changes in price, but also to changes in area, age, and distance to workplace.

2.4.5 Correction of Standard Errors

The standard errors of the 2SCF method reported in Table 2-6 have already been corrected for the use of residuals from the first stage as if they were error free. The correction was performed by bootstrapping.

Bootstrapping in this case corresponds to the following four step procedure: 1) Estimate the 2SCF. 2) Collect 100 samples, with replacement, of the data used in the first stage of the 2SCF. Each sample should have the same size as that of the original data and is used to repeatedly estimate the coefficients of the first stage of the 2SCF. 3) Use the set of coefficients estimated in Step 2 to calculate a respective set of residuals for each observation of the choice model and use them to repeatedly estimate the second stage of the 2SCF. 4) Calculate the sample variance of the set of estimators for the choice model estimated in Step 3 and add it to the variance of each estimator obtained in Step 1.

Table 2-7 Lisbon's Logit Model: Correction of 2SCF's Standard Errors by Bootstrapping

| Variables | With Endogeneity Correction | | |
|---|--------------------------------|------------------|--------------------|
| | $\hat{\beta}$ | s.e Bootstrap | s.e Uncorrected |
| 1. Dwelling price (in 100,000 €) | -2.811 | 0.6344 | 0.6339 |
| 2. Dwelling price * 1[Income > 2,000 €/M] | 0.8542 | 0.5485 | 0.5485 |
| 3. Dwelling price * 1[Income > 5,000 €/M] | 0.8089 | 0.4779 | 0.4779 |
| 4. Distance to Workplace (in Km) | -0.2565 | 0.05335 | 0.05335 |
| 5. Log [Dwelling Area (in m ²)] | 2.232 | 0.7326 | 0.7322 |
| 6. Log [Dwelling Age (in years) +1] | -0.4607 | 0.1192 | 0.1191 |
| 7. δ Control-function Auxiliary Variable | 1.054 | 0.4600 | 0.4594 |
| Log likelihood at Convergence $L(\hat{\beta})$ | -560.05 | | |
| Log likelihood at Zero $L(0)$ | -589.06 | | |
| Adjusted ρ^2 | 0.06113 | | |
| Sample Size N | 63 | | |
| Choice-set Size J | 11,501 | | |

Logit Model combining Imokapa database and SOTUR survey for Lisbon, Odivelas and Amadora
Model estimated using the 2SCF method. €/M: Euros per month.

The impact of this correction is reported in Table 2-7. The variance added to the estimators using this procedure was minimal. It was always below the fourth decimal point for all coefficients. The standard errors that were mostly affected were those of the

coefficients of price and of the residuals of the first stage of the 2SCF. This minimal effect arguably results from the large sample size (11,501) of the first stage in this application.

2.4.6 Forecasting

The impact and importance of the correction for endogeneity in this experiment is not fully measured until its effects on forecasting are accounted for. This can be done by calculating the ASE and the ADE with and without the correction for endogeneity using the expression shown before in Eq. (2-11). Table 2-8 shows these statistics for all the variables of the model. The dwelling used as reference for these calculations corresponds to the dwelling chosen by the first household in the sample. In all cases, the calculations were made by including $\hat{\delta}$ in the utility.

Table 2-8 shows that, when looking at either ASE or ADE, the sensitivity of the model was significantly increased by the correction for price endogeneity. In addition, the correction also affected the sensitivity of other dwelling attributes. This demonstrates the importance of correcting for endogeneity on policy analysis. It shows that the misspecified model will significantly underestimate the impact, not only of a pricing policy, but also the impact of policies that may affect other attributes of dwelling-units.

Table 2-8 Lisbon’s Logit Model: Forecasting with and without Endogeneity Correction

| <i>Measure</i> | | Without Endogeneity Correction | With Endogeneity Correction |
|----------------|--------------------------------------|---|--|
| <i>ASE(I)</i> | Price (in 100,000 €) | -2.500E-04 | -4.292E-04 |
| | Distance to Workplace (in Km) | -2.743E-05 | -3.915E-05 |
| | Log[Area (in m²)] | 1.269E-04 | 3.407E-04 |
| | Log[Age (in years)+1] | -4.368E-05 | -7.033E-05 |
| <i>ADE(I)</i> | Price (in 100,000 €) | -3.813 | -5.340 |
| | Distance to Workplace (in Km) | -0.6944 | -0.8083 |
| | Log[Area (in m²)] | 4.475 | 9.803 |
| | Log[Age (in years)+1] | -0.3853 | -0.5059 |

2.5 Conclusion

In this chapter, I critically reviewed recent advances in correcting for endogeneity in discrete choice models using a two-stage version of the control-function method, analyzed some issues using Monte Carlo experimentation, and applied these results to a residential location choice model for the city of Lisbon.

The first issue analyzed is related with the change of scale derived from the application of the control-function method. Extending a result from Cramer (2007) and Daly (2008), I used Monte Carlo experimentation to show that the change of scale produced with the control-function method is harmless since it does neither affect the forecasting probabilities nor the ratio of the estimators.

Second, I studied the use of the control-function method in the forecasting or simulation phase, showing that just correcting the scale, as it was suggested by Wooldridge (2002), may lead to a significant bias. I also proposed an alternative method to do forecasting that may be useful in the microscopic simulation of urban systems.

Third, following a result from Newey (1985a), I showed that both the control-function method and the 2SIV method result in consistent estimates up to a scale, but the latter results in a bias when used in the forecasting phase. This fact tips the balance toward the use of the control-function method in the correction for endogeneity in discrete choice models.

Finally, the application to real data from the city of Lisbon gives further empirical evidence that the endogeneity problem is unavoidable in residential location choice modeling. This application also serves as a detailed account of the methodological steps that have to be followed in order to correct for endogeneity in this framework, particularly regarding the construction of valid instrumental variables.

Chapter 3

Efficiency and Tractability in the Correction for Endogeneity Using Latent-variable and Control-function Methods

3.1 Overview

The control-function method is the most suitable tool to address endogeneity in spatial choice models, when this misspecification occurs at the level of each alternative. Chapter 2 examined a two-stage version of the method (2SCF), which achieves consistency but not, necessarily, efficiency and also requires a complicated correction of the standard errors for statistical testing. The goal of this chapter is to develop an estimator that can overcome these limitations without compromising practicality in spatial choice models. Throughout the chapter, I use residential location choice as an example, but the results are generally extendable to a much broader range of spatial and discrete choice models.

I begin by exploring the properties of the latent-variable method, a procedure that can also be used to address endogeneity and is typically estimated efficiently using the maximum-likelihood method. Afterwards, I analyze the control-function method within

the maximum-likelihood framework and then establish its formal link with the latent-variable method. I use this common framework to propose an estimator that achieves consistency and efficiency. This estimator also avoids, under mild conditions, the need for integration over alternatives (a problem that becomes impractical in spatial choice models, where the choice-sets are huge). I finish by illustrating the properties of the estimator using Monte Carlo experimentation and real data.

3.2 The Latent-variable Method in the Correction for Endogeneity

The latent-variable method is a technique used to account for latent variables or unobserved constructs in econometric models (Walker and Ben-Akiva, 2002). The basic idea of the method is to explicitly include the latent variable in the model specification, and to integrate it out in the calculation of the likelihood of each observation. This integration requires knowledge of the distribution of the latent variable, which is obviously unknown. The problem is solved by inferring the distribution from structural and measurement equations. In a structural equation, the latent variable is written as a function of other observed or latent variables. In turn, in a measurement equation, there is some indicator or measured variable that can be written as a function of latent and observed variables.

The random utility model is an example of the latent-variable concept. In this framework, a decision-maker (the household) chooses among a set of alternatives (the dwellings) by comparing the utility attained from them. The researcher, who wants to model the behavior of the household, cannot observe these utilities. She can only observe the choice and a fraction of the utility, known as its systematic part, which is a function of observed attributes. In this case the random utility is the latent variable. The choice is an indicator determined by the choice behavior, which then corresponds to the measurement equation. Finally, the specification of the systematic and random parts of the utility corresponds to the structural equation of the random utility model in the latent-variable framework.

The latent-variable method has been widely used in discrete choice models applied to transportation, and have experienced increasing popularity after the work of Walker (2001). The main application of the latent-variable approach in the transportation framework is in modeling the problem of latent classes. Examples of this type of applications are Kamakura and Russell (1989), Chintagunta et al. (1991), Gopinath (1995), Greene and Hensher (2003), Lee et al. (2003) and Walker and Lee (2007).

In this section, I study how the latent-variable method can be used to correct for endogeneity in models of residential location choice. The objective is to show later, in section 3.4, how this framework is linked to the control-function method and the role of this connection in the efficient estimation of models to correct for endogeneity in discrete choice modeling.

Consider the problem represented by Eq. (3-1), where a household n chooses among a set of dwellings i that belong to the choice-set C_n . The choice corresponds to variable y_{in} , which takes value 1 if alternative i has the largest random utility U_{in} among the elements in the choice-set, and zero otherwise. The systematic part of the utility depends linearly on dwelling price p , and on other attributes represented by x and q . The random utility is completed by an unobserved part represented by the error term e , which has a multivariate probability density function $f_e(\cdot)$ that depends on a set of parameters Ω_e .

$$\begin{aligned} U_{in} &= \beta_p p_{in} + \beta_x x_{in} + \beta_q q_{in} + e_{in} \\ y_{in} &= 1[U_{in} = \max_{j \in C_n} \{U_{jn}\}] \end{aligned} \quad (3-1)$$

Dwelling attributes x and q are generally correlated with dwelling price. Therefore, if, for example, q cannot be measured by the researcher, endogeneity will arise. Under the latent-variable framework, this problem can be addressed by explicitly considering q as a latent variable.

The distribution of q can be inferred using structural and measurement equations. A structural equation in residential location choice modeling requires finding an observable variable h such that q can be written as a function of h , as shown in Eq. (3-2)

$$q_{in} = \lambda_h h_{in} + \omega_{in}, \quad (3-2)$$

where the relationship is assumed to be linear, λ_h is a coefficient, and ω is an error term distributed $f_\omega(\Omega_\omega)$.

As with any system of equations, Eq. (3-1) and Eq. (3-2) have to fulfill a series of conditions to be consistently estimatable. To avoid endogeneity in the utility function, e has to be uncorrelated with p , x , and q . Equivalently, ω has to be uncorrelated with h ; but it also has to be uncorrelated with e in order to avoid the simultaneous determination between Eq. (3-1) and Eq. (3-2).

Finding a suitable variable to play the role of h in Eq. (3-2) in residential location choice modeling may be difficult in practice. One possibility would be to use a representative attribute, such as the number of months since the last time the dwelling was painted or the pipes were replaced. However, even if it were possible to collect such specific information, it is not clear why it would account for the whole set of omitted attributes in such a way that ω would be uncorrelated with h and with e . The problem is that omitted dwelling attributes are likely to be correlated among themselves since they probably share common causes, such as a careful/careless landlord. Therefore, if only a representative attribute is included in h , other omitted attributes will become part of ω , potentially causing endogeneity. In Section 3.4, I show how the difficulty of finding a suitable variable h , from generally available data, can be addressed using the control-function approach.

A measurement equation in the residential location choice problem can be written if some measure or indicator that depends on the omitted attribute q is available. Consider, for example, that the researcher has information on dwelling's rotation rate (r): the average number of households that occupied each dwelling per year. It can be hypothesized that dwellings with better q might have smaller rotation rates. This would allow us to write the measurement equation shown in Eq. (3-3), where the relationship is assumed to be linear with coefficients θ , and an error term $\eta \sim f_{\eta}(\Omega_{\eta})$. The inclusion in Eq. (3-3) of x and p , in addition to q , accounts for the fact that those factors may also play a role in the determination of the rotation rate (r).

$$r_{in} = \theta_p p_{in} + \theta_x x_{in} + \theta_q q_{in} + \eta_{in} \quad (3-3)$$

Equivalent to what occurred with the structural equation, the consistent estimation of Eq. (3-3) requires η to be uncorrelated with ω , q , x , and p . However, it is not necessary in this case to assume that η is uncorrelated with e in order to avoid simultaneous

determination between Eq. (3-3) and Eq. (3-1). That is, the error term of the choice model may also explain the realization of the indicators without compromising the consistency of the whole model. This type of correlation will not generate endogeneity due to simultaneous determination because in the structural equation (the utility function) and in the measurement equation (Eq. (3-3)), the latent variable q is on the right-hand side.

The parameters $\beta, \theta, \Omega_\eta, \lambda, \Omega_\omega$ of the model defined by Eq. (3-1), (3-2) and (3-3) can be consistently and efficiently estimated by maximizing its likelihood. To write this likelihood it is necessary to make some assumptions about the distribution of the error terms. Assuming first that the observations are independent, it is possible to write the likelihood of each observation (n) separately. Then, if it is assumed for simplicity that e is *iid* Extreme Value ($0, \mu_e = 1$) and that η and e are independent, the likelihood of observation (n) can be written as shown in Eq. (3-4).

$$L_n^* = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \frac{e^{\mu_e(\beta_p p_{in} + \beta_x x_{in} + \beta_q q_{in})}}{\sum_{j \in C_n} e^{\mu_e(\beta_p p_{jn} + \beta_x x_{jn} + \beta_q q_{jn})}} f_r(r | p, x, q; \theta, \Omega_\eta) f_q(q | h; \lambda, \Omega_\omega) dq \quad (3-4)$$

The estimation of this model also requires assuming a particular distribution for η and ω , the error terms of the structural and measurement equations, respectively. Consider, for example, that η and ω are *iid* Normal with mean zero and variances σ_η^2 and σ_ω^2 , respectively. Note that this is equivalent to saying that the errors are homoscedastic and non-autocorrelated. Making the appropriate change of variables between q and ω , the likelihood in Eq. (3-4) becomes what is shown in Eq. (3-5).

$$L_n^* = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \frac{e^{\mu_e(\beta_p p_{in} + \beta_x x_{in} + \beta_q(\lambda_h h_i + \omega_i))}}{\sum_{j \in C_n} e^{\mu_e(\beta_p p_{jn} + \beta_x x_{jn} + \beta_q(\lambda_h h_j + \omega_j))}} \cdots \quad (3-5)$$

$$\cdots \prod_{j \in C_n} \frac{1}{\sqrt{2\pi\sigma_\eta^2}} \exp\left[-\frac{(r_j - \theta_p p_j - \theta_x x_j - \theta_q(\lambda_h h_i + \omega_j))^2}{2\sigma_\eta^2}\right] \frac{1}{\sqrt{2\pi\sigma_\omega^2}} \exp\left[-\frac{\omega_j^2}{2\sigma_\omega^2}\right] d\omega$$

The application of this model to residential location choice is still impractical because the likelihood requires the calculation of a multifold integral in the number of alternatives, which is usually huge. In this sense, it is important to note what occurs when the measurement equation is not available or when it is just ignored. In such cases, the likelihood reduces to Eq. (3-6).

$$L_n^* = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \frac{e^{\mu_e(\beta_p p_m + \beta_x x_m + \beta_q(\lambda_n h_i + \omega_n))}}{\sum_{j \in C_n} e^{\mu_e(\beta_p p_m + \beta_x x_m + \beta_q(\lambda_n h_j + \omega_j))}} \prod_{j \in C_n} \frac{1}{\sqrt{2\pi\sigma_\omega^2}} \exp\left(-\frac{\omega_j^2}{\sigma_\omega^2}\right) d\omega \quad (3-6)$$

Under this setting the error term ω is not identifiable because it is confounded with the error e . Note however that this happens only when the structural equation shown in Eq. (3-2) is linear. One possibility for the estimation of this model would be to normalize the variance $\sigma_\omega^2 = 1$ and maximize Eq. (3-6). However, this assumption does not improve the properties of the model when compared with a more practical option where the whole error $\omega + e$ is assumed to be distributed Extreme Value $(0, \mu_{\omega+e})$. In the latter case the model reduces to a Logit, obviating the need for integration across alternatives, as shown in Eq. (3-7).

$$L_n^* = \frac{e^{\mu_{\omega+e}(\beta_p p_m + \beta_x x_m + \beta_h h_m)}}{\sum_{j \in C_n} e^{\mu_{\omega+e}(\beta_p p_m + \beta_x x_m + \beta_h h_m)}} \quad (3-7)$$

Assuming that $\omega + e$ is distributed Extreme Value $(0, \mu_{\omega+e})$ may seem problematic since the sum of a normally distributed ω and variable e , which is Extreme Value, has an unknown distribution. However, it can be argued that if the sample is large enough, any Law of Large Numbers would make it possible to claim that $\omega + e$ follows a Normal distribution. Then, using the results by Lee (1982) and Ruud (1983), showing that the approximation of a Normal by an Extreme Value distribution causes negligible discrepancies, the resulting model becomes a Logit, as shown in Eq. (3-7).

The estimation of the model shown in Eq. (3-7) has some peculiarities compared to the model shown in Eq. (3-5). First, the omission of ω in this Logit model affects the scale $\mu_{\omega+e}$ of the estimators, which is then different from the scale of the original model $\mu_e = 1$. Additionally, the omission of the measurement equation (Eq. (3-3)) results in that only the product $\beta_h = \beta_q \lambda_n$ would be identifiable in this case (not β_q or λ_n separately) and that the efficiency of the estimators will be reduced.

In summary, the latent-variable approach can be used to address endogeneity due to the omission of attributes in residential location choice models. For this purpose it is necessary to obtain a variable h to construct appropriate structural equations. If the

measurement equation is ignored, the model reduces to a Logit under some mild conditions. If the measurement equation is also available, the efficiency of the estimators would be increased, more parameters of the model would be identified, but the solution of the problem would require integration over all the alternatives, a problem that may prove impractical in models of residential location choice.

3.3 The Control-function Method in a Maximum-likelihood Framework

The efficient estimation of the control-function method can be achieved by estimating it using the maximum-likelihood method, because the estimators will attain the Cramer-Rao lower bound (Ben-Akiva and Lerman, 1985). The first step toward this goal is to write the likelihood of the model described in Eq. (2-2) from Chapter 2.

$$\begin{aligned}
 U_{in} &= \beta_p p_{in} + \beta_x x_{in} + \overbrace{\beta_\delta \delta_{in} + v_{in}}^{\xi_{in}} + e_{in} \\
 p_{in} &= \alpha_z z_{in} + \delta_{in} \\
 y_{in} &= 1[U_{in} = \max_{j \in C_n} \{U_{jn}\}]
 \end{aligned}$$

Assuming independence between observations n , the likelihood for each observation can be written separately. Given that δ and v are independent, if it is assumed that $\delta \sim f_\delta(\Omega_\delta)$, $v \sim f_v(\Omega_v)$, and that e is distributed *iid* Extreme Value (0, $\mu_e = 1$), the likelihood of observation n can be written as the Logit Mixture model shown in Eq. (3-8).

$$L_n^* = f_\delta(\delta | \Omega_\delta) \int_{-\infty}^{+\infty} \frac{e^{\mu_e(\beta_p p_{in} + \beta_x x_{in} + \beta_\delta \delta_{in} + v_{in})}}{\sum_{j \in C_n} e^{\mu_e(\beta_p p_{jn} + \beta_x x_{jn} + \beta_\delta \delta_{jn} + v_{jn})}} f_v(v | \Omega_v) dv \quad (3-8)$$

Note that in this case the likelihood of v and the likelihood of δ need to be considered across all alternatives in choice-set C_n , but the latter does not need to be inside the integral because it is independent of v and e , and is fully determined by p and z , which are observed. This formulation is equivalent to that used by Villas-Boas and Winner (1999) and Park and Gupta (2009) to perform a simultaneous estimation of the control-function method, and to what Train (2009) terms maximum-likelihood methods.

It is interesting to compare the 2SCF method described in Chapter 2 with its maximum-likelihood counterpart presented here. The 2SCF method can be seen as a limited-information maximum-likelihood (LIML) version of the full-information maximum-likelihood (FIML) model represented by Eq. (3-8). This has both advantages and disadvantages. Beyond the simplification of the estimation procedures, the 2SCF method has the advantage of being more robust for misspecifications of the error structure. This is because the conditional distribution of the error of the second stage, given the residuals of the first stage, is compatible with various joint distributions of ζ and δ (Wooldridge, 2002). On the other hand, the 2SCF procedure has the disadvantages of not being necessarily efficient, and that the standard errors cannot be calculated from the inverse of the Fisher-information-matrix.

The need for integration over ν across alternatives in the choice-set in Eq. (3-8) may be problematic in models of residential location choice because the choice-set can be huge. Following the same argumentation used for the latent-variable models, if ν has the same variance σ_ν^2 across alternatives, ν will not be identifiable from e . Then, if the sample is large enough, it can be assumed that $\nu+e$ is distributed Normal. The normality assumption can equivalently result from assuming that e is distributed Normal. This is because ν was already Normal, and the sum of two normally distributed random variables is also normally distributed. Finally, based on the results by Lee (1982) and Ruud (1983), the distribution of $\nu+e$ can be safely approximated by an Extreme Value $(0, \mu_{\nu+e})$ distribution, avoiding the need for integration in this problem.

The formulation that results by assuming that $\nu+e$ are distributed, or can be approximated by, a Logit model is shown in Eq. (3-9). This formulation is termed the tractable maximum-likelihood estimator of the control-function method.

$$L_n^* = f_\delta(\delta | 0, \Omega_\delta) \frac{e^{\mu_{\nu+e}(\beta_p p_{in} + \beta_x x_{in} + \beta_\delta \delta_{in})}}{\sum_{j \in C_n} e^{\mu_{\nu+e}(\beta_p p_{jn} + \beta_x x_{jn} + \beta_\delta \delta_{jn})}} \quad (3-9)$$

The estimation of the model parameters by maximizing the likelihood shown in Eq. (3-9) requires making specific assumptions about the distribution of δ . For example, if it

is assumed that δ is *iid* Normal with variance σ_δ^2 , the likelihood can be rewritten as shown in Eq. (3-10).

$$L_n^* = \frac{e^{\mu_{v+\epsilon}(\beta_p p_{jn} + \beta_x x_{jn} + \beta_\delta(p_{jn} - \alpha_z z_{jn}))}}{\sum_{j \in C_n} e^{\mu_{v+\epsilon}(\beta_p p_{jn} + \beta_x x_{jn} + \beta_\delta(p_{jn} - \alpha_z z_{jn}))}} \prod_{j \in C_n} \frac{1}{\sqrt{2\pi\sigma_\delta^2}} \exp\left[-\frac{(p_{jn} - \alpha_z z_{jn})^2}{2\sigma_\delta^2}\right] \quad (3-10)$$

It is interesting to note what occurs when taking the log of Eq. (3-10), a monotonic transformation of the objective function of the maximum-likelihood problem that does not affect optimization results. In this case, the objective function to be maximized corresponds almost exactly to the sum of the objective functions of the first and second stages of the 2SCF procedure. The only changes are that the first component is weighted by a term that depends on the inverse of twice the variance of δ and that there is an additive constant term that also depends on the variance of δ .

Finally, regarding the efficiency of the 2SCF procedure as compared to the tractable maximum-likelihood estimator, the latter attains the Cramer-Rao lower bound and is therefore efficient. However, this does not necessarily mean that the 2SCF method is inefficient. If the error terms δ and ξ are *iid* (homoscedastic and non-autocorrelated) 2SCF will be efficient. This result was noted by Rivers and Vuong (1988) and is equivalent to what occurs in linear models between 2SLS and 3SLS methods (see, e.g., Greene, 2003). There is however one important difference between 2SCF and the maximum-likelihood approach. Although the estimated coefficients of the 2SCF will be consistent and efficient if the errors are homoscedastic and non-autocorrelated, the estimators of the standard errors (calculated using the inverse of the Fisher-information-matrix) will be inconsistent, precluding the direct application of hypothesis testing, unless they are corrected. This correction can be done using, for example, non-parametric methods such as bootstrapping.

3.4 The Link between Latent-variable and Control-function Methods

The latent-variable and the control-function methods are conceived from fairly different perspectives. The latent-variable method has a broad range of applications and is based

on accounting for the causality among observed and latent variables resulting from the behavior of the agents involved in the phenomena under study. In contrast, the control-function method is intended specifically for the correction for endogeneity and is mainly based on the statistical properties of the variables. Despite of the different origins and objectives, Guevara and Ben-Akiva (2010) noted that there is a link between the two approaches. In this section I analyze the connection between both methods and highlight the impact of the identification of this link on the efficient correction for endogeneity in residential location choice models.

The main issue in linking the latent-variable and the control-function methods is to identify the roles played by the different components of each method on its counterpart. This link becomes immediately apparent by comparing the likelihood functions shown in Eq. (3-7) and Eq. (3-9). It should be noted that the residuals of the first stage of the control-function method $\delta = p - \alpha_z z$ can play the role of variable h , the independent variable required in the specification the structural equation in the latent-variable approach shown in Eq. (3-2). Note that since δ is not deterministic, the likelihood of the model has to be multiplied by the likelihood of δ in Eq. (3-9). Therefore, it can be affirmed that the control-function framework allows for the construction, from valid instrumental variables, of variables that can play the role of h for the implementation of the structural equation in the latent-variable framework.

An alternative way to identify the link between the control-function and the latent-variable approaches is to note that the implementation of the former conveys the decomposition of the error term ε of the model into an endogenous part ξ and an exogenous part e . As it is shown in Eq. (2-2), ξ is then decomposed in two parts, where the first depends on δ and the second is an exogenous error v . Then, interpreting ξ as q , it follows directly that the expression

$$\xi = \beta_\delta \delta + v$$

constitutes a structural equation for ξ where δ plays the role of h in Eq. (3-2).

The link and synergy between the control-function and latent-variable approaches in modeling residential location choice is clear. If the researcher has information about an indicator such as the rotation rate of the dwellings, it would be possible to use the

control-function approach to build suitable structural equations, and apply the latent-variable framework to use the information from the indicator by means of a measurement equation. This will increase the efficiency of the estimators and allow for the identification of more model parameters. This subsequently achieves a more realistic representation of the behavior of the agents in the system. The cost, however, is that the model needs then to be integrated across alternatives, a calculation that may become impractical with large choice-sets.

The full assessment of the value of the identification of the link between the control-function and latent-variable approaches shall be addressed by the estimations of models with real data. This task is left for future research.

3.5 Assumptions to Achieve Efficiency and Tractability

The likelihoods used in the estimation of the control-function or the latent-variable methods will result in consistent and efficient estimators of the model parameters if Eq. (3-9) and Eq. (3-7) represent the true likelihood of the model, or are acceptable approximations of it in the sense established by White (1982). Therefore, it is important to study the mildness of the assumptions involved in the derivation of Eq. (3-9) (which are extendable to those of Eq. (3-7)), the impact of their failure, and possible strategies to address it.

First, the assumption on the homoscedasticity and non-autocorrelation for v results from the joint normality assumption between ξ and δ used in the derivation of the control-function method as it was described in Chapter 2. A failure would occur if the variance of the omitted attributes ξ depends on the instruments z or on the alternatives j . There is no a priori ground to suggest that this failure might occur, but if it did, it could be resolved using the Logit Mixture model described in Eq. (3-11). The cost is that the model would need to be integrated across alternatives, which would make this approach generally intractable in spatial choice modeling, unless some simplifying assumptions were considered for the structure of Ω_v . Feasible alternatives might include block homoscedasticity and/or autoregressive processes of degree 1 (see, e.g., Greene, 2003).

$$L_n^* = f_\delta(\delta | \Omega_\delta) \int \cdots \int \frac{e^{\mu_e(\beta_p p_{in} + \beta_x x_{in} + \beta_\delta \delta_{in} + v_{in})}}{\sum_{j \in C_n} e^{\mu_e(\beta_p p_{jn} + \beta_x x_{jn} + \beta_\delta \delta_{jn} + v_{jn})}} f_v(v | \delta, \Omega_v) dv \quad (3-11)$$

There is also no a priori ground to expect the failure of the assumption on the homoscedasticity and non-autocorrelation for δ required to arrive at Eq. (3-10). This failure might occur if the variance of δ is different across alternatives and observations depending, for example, on the instruments or on the alternatives. However, if a failure does occur, the cost of addressing it would not compromise the tractability of the model because it would not involve integration over the alternatives in the choice-set. This problem can be handled using any Feasible Generalized Least Squares (FGLS) procedure (see, e.g., Greene, 2003) in the specification of f_δ in Eq. (3-9). One alternative would be to obtain a consistent estimator $\hat{\Omega}_\delta$ of the variance Ω_δ and then use it in the specification of Eq. (3-9). Alternatively, if the specification of Ω_δ is simple enough, it would be possible to estimate it within the same maximum-likelihood problem. I will use the second approach for the estimation with real data, later in this chapter.

As discussed before, to reach the Logit closed form shown in Eq. (3-9) it was necessary to assume that $v+e$ followed an Extreme Value distribution. This assumption ultimately depends on the assumption that a Normal distribution can be acceptably approximated using an Extreme Value distribution. Concordant with the results by Lee (1982) and Ruud (1983), the Monte Carlo experiments shown in the next section demonstrate that this approximation is reasonably robust.

As a final point on efficiency, it should be noted that Newey (1987) studied a method developed by Amemiya (1978) to correct for endogeneity in discrete choice models. I describe this estimator later in Section 4.2.2. Newey (1987) showed that Amemiya's estimator is at least as efficient as the 2SCF, and globally efficient under some circumstances. This estimator is much more complicated to calculate than the 2SCF because it involves the estimation of various auxiliary models, including a minimum chi-squared procedure devised specially by Amemiya (1978).

In summary, when δ and v are homoscedastic and non-autocorrelated, the tractable maximum-likelihood method deployed in Eq. (3-10) would be preferred because it will

be easier to estimate, globally efficient and would allow for the direct calculation of the standard errors. In cases where there are suspicions that δ and ν have non-spherical structures (and writing the log-likelihood is impractical), Amemiya's method would be preferred because it is practical and will be at least as efficient as the 2SCF method. However, equivalent to what occurred with the 2SIV method, it is not clear how to forecast with the estimators obtained with Amemiya's method. Therefore if the model will be used for simulation (and writing the likelihood is impractical) the 2SCF should still be preferred.

3.6 Monte Carlo Experiment

In this section I revisit the binary choice Monte Carlo experiment developed in Chapter 2, defined by Eq. (2-9) and Eq. (2-10). The models studied in this case are the true model; the model that suffers from endogeneity resulting from the omission of ζ ; the model that corrects for endogeneity using the 2SCF method; and the tractable maximum-likelihood method deployed in Eq. (3-9). The last method considers that δ and ν are homoscedastic and non-autocorrelated, so that Eq. (3-10) is valid. A total of 100 repetitions of the data were generated. Table 3-1 shows the average, the bias, the mean squared error (MSE) and the t-test against the true value of ratio of the estimators of the coefficients of p and x_j . The results are classified by sample sizes N and by the diverse models estimated.

Table 3-1 shows that the ratio between the estimators of the coefficients of p and x_1 in the true model is almost identical and statistically equal (with 95% confidence) to its true value: -2. In turn, when ζ is omitted, endogeneity causes a significant positive bias of the estimator for this ratio, as in Chapter 2. The most relevant result reported in Table 3-1 corresponds to the comparison between 2SCF and the tractable maximum-likelihood estimator. Interestingly, the MSE for both methods are virtually identical. This is because the true model is *iid* across observations and alternatives. Therefore, Eq. (3-10) is valid, and the model falls under the case where both the 2SCF and the tractable maximum-likelihood estimators are efficient.

Another difference between the 2SCF and the maximum-likelihood methods resides in the calculation of the standard errors. The former requires bootstrapping and the latter

can be achieved by inverting the Fisher-information-matrix. It was found for these experiments that the impact of the correction required for the 2SCF depended on the sample size of each problem, ranging from changes in the third decimal of the standard error for $N=150$, to changes in the fifth decimal for $N=2,000$. It can be affirmed that, for computational time, when the maximum-likelihood method is compared with the 2SCF (with bootstrap), the former method outperforms the latter because bootstrapping requires the repetitive estimation of several models.

Table 3-1 Monte Carlo Experiment: 2SCF and Maximum-likelihood Methods

| $\hat{\beta}_p / \hat{\beta}_{x_i}$ | Metric | $N=150$ | $N=500$ | $N=1,000$ | $N=2,000$ |
|--|-------------|----------|-----------|-----------|-----------|
| True Model | Average | -2.011 | -2.002 | -2.012 | -2.003 |
| | Bias | -0.01073 | -0.001686 | -0.01247 | -0.002897 |
| | MSE | 0.1253 | 0.02829 | 0.01261 | 0.006819 |
| | t-test true | -0.03031 | -0.01002 | -0.1118 | -0.03510 |
| Omitting ξ | Average | -1.211 | -1.208 | -1.212 | -1.197 |
| | Bias | 0.7890 | 0.7919 | 0.7882 | 0.8035 |
| | MSE | 0.7071 | 0.6488 | 0.6282 | 0.6495 |
| | t-test true | 2.713 | 5.380 | 9.408 | 12.75 |
| 2SCF | Average | -2.042 | -1.994 | -2.006 | -1.999 |
| | Bias | -0.04193 | 0.006188 | -0.006141 | 0.0006357 |
| | MSE | 0.1884 | 0.04251 | 0.01882 | 0.01053 |
| | t-test true | -0.09706 | 0.03003 | -0.04481 | 0.006195 |
| Maximum-likelihood Homoscedastic non-Autoc. | Average | -2.042 | -1.994 | -2.006 | -1.999 |
| | Bias | -0.04200 | 0.006181 | -0.006175 | 0.0006229 |
| | MSE | 0.1885 | 0.04252 | 0.01882 | 0.01054 |
| | t-test true | -0.09721 | 0.02999 | -0.04506 | 0.006068 |

100 Repetitions. $J=2$

In summary, these experiments show that when δ and ν are *iid*, the 2SCF method is as efficient as the tractable maximum-likelihood estimator described in Eq. (3-10). The latter however, besides achieving consistency and efficiency, also allows direct hypothesis testing using the standard errors calculated from the inverse of the Fisher-information-matrix. This final fact implies that the tractable maximum-likelihood estimator also outperforms the 2SCF in terms of computational cost.

3.7 Application to Real Data

The final section of this chapter focuses on the application of the tractable maximum-likelihood method, which is described in Eq. (3-9), in the residential location choice model for Lisbon. Three models are shown in Table 3-2. The first corresponds to the 2SCF method estimated in Chapter 2. The second corresponds to the maximum-likelihood estimator described in Eq. (3-10), where it is assumed that both δ and ν are homoscedastic and non-autocorrelated. The last model corresponds to the maximum-likelihood method described in Eq. (3-9) where δ and ν are assumed to be non-autocorrelated, but only ν is assumed to be homoscedastic. The heteroscedasticity of δ in the third model reported in Table 3-2 was addressed by estimating two variances, $\sigma_{\delta_1}^2$ for the dwellings in the Lisbon municipality, and $\sigma_{\delta_2}^2$ for the dwellings located in the municipalities of Odivelas and Amadora. For both maximum-likelihood models, the variances were estimated in one stage within the optimization procedure.

Table 3-2 shows that the estimators of the choice model coefficients are statistically equal (with 95% confidence) among the three models. To make this comparison appropriate, the standard errors of the 2SCF method include the correction calculated by bootstrapping. It should be noted that, besides the standard deviations (σ_{δ} , σ_{δ_1} and σ_{δ_2}), the only notable differences among the estimated models occur with the coefficient of the intercept of the price equation and with the value of the likelihoods. In the former, the difference comes from a change of units. In the first stage of the 2SCF estimated in Chapter 2, the prices were considered in Euros and in the maximum-likelihood estimation, all prices were considered in hundreds of thousands of Euros. After adjusting for this change of units, the intercept is also similar for the three methods, and almost identical for the first two. Equivalently, the difference among the likelihoods of the 2SCF and the maximum-likelihood models, results from the likelihood reported for the 2SCF method is only that of the choice model and, in the maximum-likelihood models, the likelihood reported is the joint likelihood of the price equation and the choice model.

Table 3-2 Lisbon's Logit Model: 2SCF and Maximum-likelihood Methods

| Variables | 2SCF | MaxLik Homoscedastic non-Autoc. | MaxLik δ Heteroscedastic non-Autoc. |
|--|-----------------------|---------------------------------------|--|
| 1. Dwelling price (in 100,000 €) | -2.811 (0.6344) | -2.812 (0.6340) | -2.818 (0.6356) |
| 2. Dwelling price * 1[Income > 2,000 €/M] | 0.8542 (0.5485) | 0.8543 (0.5485) | 0.8533 (0.5485) |
| 3. Dwelling price * 1[Income > 5,000 €/M] | 0.8089 (0.4779) | 0.8087 (0.4780) | 0.8085 (0.4782) |
| 4. Distance to work (in Km) | -0.2565 (0.0534) | -0.2565 (0.05336) | -0.2565 (0.05335) |
| 5. Log [Dwelling Area (in m ²)] | 2.232 (0.7326) | 2.232 (0.7323) | 2.233 (0.7325) |
| 6. Log [Dwelling Age (in years) +1] | -0.4607 (0.1192) | -0.4607 (0.1191) | -0.4609 (0.1191) |
| 7. δ | 1.054 (0.4600) | 1.055 (0.4595) | 1.062 (0.4625) |
| α_0 Intercept Price Equation | -3.023.E+04 (2450) | -0.3024 (0.02448) | -0.3159 (0.02391) |
| α_{z1} Instrument z_1 | 0.6995 (0.01052) | 0.6994 (0.01051) | 0.6937 (0.01035) |
| α_{z2} Instrument z_2 | 0.4830 (0.01935) | 0.4827 (0.01933) | 0.4818 (0.01873) |
| σ_δ | | 0.7150 (0.004714) | |
| σ_{δ_1} Lisbon | | | 0.7700 (0.006132) |
| σ_{δ_2} Odivelas and Amadora | | | 0.5707 (0.006970) |
| Adjusted R^2 | 0.6779 | | |
| Log likelihood at Convergence $L(\hat{\beta})$ | -563.00 | -13,020.67 | -12,830.26 |
| Log likelihood at Zero $L(\beta=0; \sigma=1)$ | -589.06 | -46,892.31 | -46,892.31 |
| Adjusted ρ^2 | 0.05443 | 0.7226 | 0.7266 |
| Sample Size Choice Model N | 63 | 63 | 63 |
| Choice-Set Size J /Sample Size First Stage | 11,501 | 11,501 | 11,501 |

Standard errors in parenthesis. €/M: Euros per month.

Finally, the virtual equality among the estimators of the 2SCF and both maximum-likelihood estimators is a sign that the specification of the model is correct. Formally, comparing the first 7 coefficients of the two maximum-likelihood models using a

Hausman's (1978) test, the null hypothesis that the three sets of estimators are statistically equal (with 95% confidence) is not rejected. This implies that both models satisfactorily corrected for endogeneity and resulted in consistent estimators of the model parameters. The differences among the estimators are due only to the increase in efficiency attained with consideration of a more general variance-covariance matrix.

Note also that the log-likelihood of the third model is substantially more positive than the likelihood of the model where only one standard deviation term is considered. Evaluated through a Likelihood-ratio test, this loosely rejects the null hypothesis that the standard deviation of Lisbon and Odivelas-Ámadora are the same. This implies that the third model produced a significant increase in efficiency and should then be preferred.

3.8 Conclusion

In this chapter, I explored the possibility of addressing endogeneity in residential location choice models by combining the control-function and the latent-variable frameworks. I showed that the control-function method allows for the construction of structural equations that can be used to implement the latent-variable method. The full value of the identification of this link remains to be identified in future research using appropriate real data.

I also showed that when there are no measurement equations, the latent-variable and the control-function methods become the same maximum-likelihood model. Also, under mild conditions, the estimation of the common maximum-likelihood model avoids the calculation of a multifold integral, a problem that becomes impractical in residential location choice models.

Additionally, I pointed out, following Rivers and Vuong (1988), that 2SCF will achieve efficiency if the error terms of the model are homoscedastic and non-autocorrelated. However, even in that case, the standard errors of the estimators cannot be calculated directly from the inverse of the Fisher-information-matrix, as they do when the maximum-likelihood approach is used.

I also showed that if the error of the first stage of the 2SCF is heteroscedastic and autocorrelated, the problem can be solved under the maximum-likelihood framework

through the estimation of a simile of the Feasible Generalized Least Squares method in linear models. This method can be implemented in two stages or simultaneously, depending on the complexity of the structure of the variance-covariance matrix.

Chapter 4

Testing for the Validity of Instrumental

Variables in Discrete Choice Models

4.1 Overview

The crucial assumption required for the correction for endogeneity using the control-function or any other method, is the availability of suitable instrumental variables. Instruments have to be relevant (correlated with the endogenous variable) and also valid (uncorrelated with the error term of the model). The second requirement is particularly difficult to test because the error term is not observed.

For linear models, Sargan (1958) noted that if the model is over-identified (if there are more instruments than endogenous variables) the residuals of the instrumental-variables (IV) regression can be used to test for instruments exogeneity. For discrete choice models, Lee (1992) noted that an estimator developed by Amemiya (1978), and studied by Newey (1987), can play the role of the Sargan test in the validation of instruments in this context.

In this chapter I present the details of the Sargan test for linear models and then those of the Amemiya-Lee-Newey test for discrete choice models. Next, I develop a novel Regression-based test for the validity of instruments in Logit models using the concept of

generalized residuals developed by Cox and Snell (1968), and the asymptotic results from an omitted-attributes test developed by McFadden (1987). Additionally I propose a Direct test for the validity of instruments that is applicable to various types of discrete choice models and has some practical advantages. Finally, I analyze the performance of the proposed tests using Monte Carlo experimentation and real data on residential location choice from Lisbon, Portugal.

4.2 Validation of Instruments Using Over-identifying Restrictions

4.2.1 The Sargan Test in Linear Models

Verifying that the instruments are not correlated with the error term of the model is cumbersome, and may seem impossible, because the error term is unobserved. However, Sargan (1958), and later Basman (1960), noted that testing in linear models is feasible when the model is over-identified.

To describe the Sargan test, reconsider the problem formulated in previous chapters but transformed (only for this section) into a linear model where variable y is continuous, as follows:

$$y_i = \beta_0 + \beta_p p_i + \beta_x x_i + \xi_i + e_i \quad i = 1, \dots, N$$

$$y_i = \beta_0 + \beta_p p_i + \beta_x x_i + \varepsilon_i$$

$$p_i = \alpha_0 + \alpha_z z_i + \delta_i,$$

where z is a valid instrument and $\text{corr}(\delta, \xi) \neq 0$.

As before, δ is correlated with ξ , so p is correlated with ξ and therefore, the omission of ξ will cause endogeneity. Variables x , e and z are independent among them and independent of all other variables and error terms in the model. Under this setting z is a good instrument because it is correlated with p and independent of e and ξ .

Since this model is linear, endogeneity can be solved either using the control-function or the two-stage least-squares (2SLS) methods. Consider the 2SLS procedure. The first stage of 2SLS corresponds to the OLS regression of the endogenous variable p on the

instrument z . The estimators of this model $\hat{\alpha}$ are then used to calculate the fitted values $\hat{p} = \hat{\alpha}_0 + \hat{\alpha}_z z$ and the residuals $\hat{\delta} = p - (\hat{\alpha}_0 + \hat{\alpha}_z z)$. The fitted values and the residuals are orthogonal by construction (see, e.g., Greene, 2003). Also, note that \hat{p} is linear in z . That is, \hat{p} is in the plane that is spanned by z and therefore, like z , \hat{p} is also uncorrelated with ξ and e .

The second stage of the 2SLS procedure corresponds to the replacement of p (in the model where ξ is omitted) with the fitted value \hat{p} . As shown in Eq. (4-1), the error term of this auxiliary regression $\tilde{\varepsilon} = \beta_p \hat{\delta} + \xi + e$ will not be correlated with the observed variables \hat{p} and x . Therefore the estimators of this 2SLS regression will be consistent.

$$\begin{aligned} y_i &= \beta_0 + \beta_p (\hat{p}_i + \hat{\delta}_i) + \beta_x x_i + \varepsilon_i \\ y_i &= \beta_0 + \beta_p \hat{p}_i + \beta_x x_i + \underbrace{\beta_p \hat{\delta}_i + \xi_i + e_i}_{\tilde{\varepsilon}_i} \xrightarrow{OLS} \hat{\beta} \end{aligned} \quad (4-1)$$

Since $\hat{\beta}_0$, $\hat{\beta}_p$ and $\hat{\beta}_x$ are consistent estimators of β_0 , β_p and β_x , a consistent estimator of the error ε can be obtained by replacing $\hat{\beta}_0, \hat{\beta}_p, \hat{\beta}_x$ in the model where ξ is omitted to obtain $\hat{\varepsilon} = y - (\hat{\beta}_0 + \hat{\beta}_p p + \hat{\beta}_x x)$. If the instrument z is valid, then z should be uncorrelated with ε and also with its consistent estimator $\hat{\varepsilon}$. Since $\hat{\varepsilon}$ is observed, it is tempting to just calculate the correlation between $\hat{\varepsilon}$ and z to test for the validity of z . However, in this case $\hat{\varepsilon}$ is orthogonal to z by construction and testing is therefore impossible because, even if z is invalid, it will be uncorrelated with $\hat{\varepsilon}$.

To show that $\hat{\varepsilon}$ is orthogonal to z by construction note first that

$$\hat{\varepsilon} = y - (\hat{\beta}_0 + \hat{\beta}_p p + \hat{\beta}_x x) = y - (\hat{\beta}_0 + \hat{\beta}_p \hat{p} + \hat{\beta}_x x + \hat{\beta}_p \hat{\delta}) = \hat{\varepsilon} - \hat{\beta}_p \hat{\delta}.$$

Then, since $\hat{\delta}$ is orthogonal to z because it is the residual from the regression of p on z , $\hat{\varepsilon}$ will be orthogonal to z if $\hat{\varepsilon}$ is orthogonal to z . Note then that $\hat{\varepsilon}$ is orthogonal to \hat{p} because it is the residual of the 2SLS regression shown in Eq. (4-1). Then, decomposing \hat{p} into $\hat{p} = \hat{\alpha}_0 + \hat{\alpha}_z z$, note that

$$\hat{\varepsilon}' \hat{p} = \hat{\alpha}_0 \sum_{i=1}^N \hat{\varepsilon}_i + \hat{\alpha}_z \hat{\varepsilon}' z = \hat{\alpha}_z \hat{\varepsilon}' z = 0,$$

where $\sum_{i=1}^N \hat{\varepsilon}_i = 0$

because the model has an intercept. This means that $\hat{\varepsilon}$ is always orthogonal to z , no matter how correlated ε and z might be in reality. Testing for the validity of z is therefore impossible.

To apply the Sargan test it is necessary to make a small shift to the model described above in order to avoid the problem of having $\hat{\varepsilon}$ be orthogonal to z by construction. Consider now that the price is a linear function of two instruments, z_1 and z_2 .

$$p_i = \alpha_0 + \alpha_{z_1} z_{1i} + \alpha_{z_2} z_{2i} + \delta_i$$

In this case, the model is said to be over-identified since it has more instruments than endogenous variables. Under this setting, \hat{p} will be a particular linear combination of z_1 and z_2 ,

$$\hat{p} = \hat{\alpha}_0 + \hat{\alpha}_{z_1} z_1 + \hat{\alpha}_{z_2} z_2$$

and $\hat{\varepsilon}$ will be orthogonal, by construction, to \hat{p} , but neither (necessarily) to z_1 , nor to z_2 .

To show why, note that the orthogonality between $\hat{\varepsilon}$ and \hat{p} implies that

$$\hat{\varepsilon}' \hat{p} = \hat{\alpha}_0 \sum_{i=1}^N \hat{\varepsilon}_i + \hat{\alpha}_{z_1} \hat{\varepsilon}' z_1 + \hat{\alpha}_{z_2} \hat{\varepsilon}' z_2 = \hat{\alpha}_{z_1} \hat{\varepsilon}' z_1 + \hat{\alpha}_{z_2} \hat{\varepsilon}' z_2 = 0$$

but neither that $\hat{\alpha}_{z_1} \hat{\varepsilon}' z_1 = 0$ nor that $\hat{\alpha}_{z_2} \hat{\varepsilon}' z_2 = 0$.

This result implies that, z_1 and z_2 would only be uncorrelated with $\hat{\varepsilon}$ by chance, only if they are indeed good instruments. Then, it is possible to use $\hat{\varepsilon}$ to test the validity of instruments z_1 and z_2 because, if z_1 and z_2 are good instruments, $\hat{\varepsilon}$ will be a consistent estimator of ε and, at the same time, $\hat{\varepsilon}$ will not be orthogonal to z_1 and z_2 by construction.

Figure 4-1 shows this result graphically. The figure represents a case where only three observations are available, which makes it possible to draw the vectors in a 3-dimensional space. Vectors z_1 and z_2 are in the plane \mathbf{Z} . \hat{p} , the OLS estimator of the regression of p on z_1 and z_2 , is also in the plane \mathbf{Z} . Variable x is in a plane that is not orthogonal to the plane \mathbf{Z} . $\hat{\varepsilon}$, the residual of the second stage of the 2SLS method, is orthogonal, by construction, to \hat{p} and x . However, note that the angles between $\hat{\varepsilon}$ and

z_1 , and between $\hat{\varepsilon}$ and z_2 , are far from being a right angle. The only cases where $\hat{\varepsilon}$ and z_1 or z_2 would be orthogonal are when the model is just identified ($z_1 = z_2$) or when the instruments are valid. In this 3-dimensional example, the latter option can occur only when x is in the plane spanned by z_1 and z_2 .

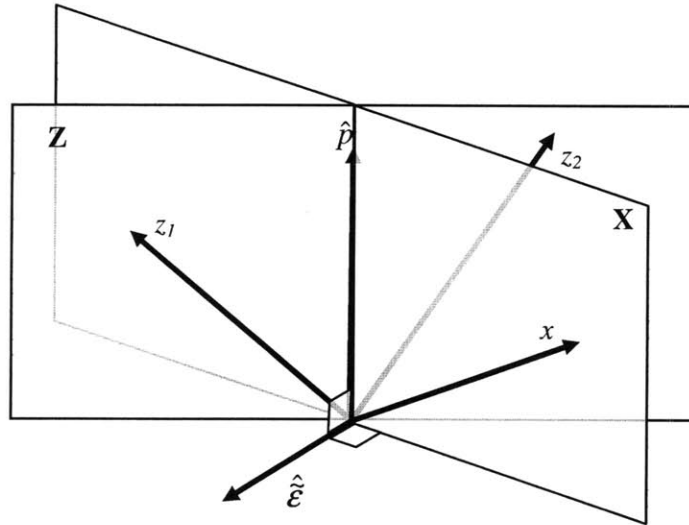


Figure 4-1 Over-identification Allows Testing for the Validity of Instruments

The statistic of the Sargan test is constructed by estimating an OLS regression of the residual $\hat{\varepsilon}$ of the 2SLS procedure on all the exogenous variables of the model, which for this example are z_1 , z_2 and x .

$$\hat{\varepsilon}_i = \theta_0 + \theta_x x_i + \theta_{z_1} z_{1i} + \theta_{z_2} z_{2i} + \psi_i \quad (4-2)$$

The Sargan test is calculated as a LaGrange-multiplier test, where the null hypothesis is that all θ 's in Eq. (4-2), except for the intercept, are equal to zero. This corresponds to the S statistic shown in Eq. (4-3), where R^2 is the unadjusted coefficient of determination of the regression shown in Eq. (4-2), and N is the number of observations. The S statistic is distributed χ_{df}^2 with a number of degrees of freedom (df) equal to the degree of over-identification of the problem (the number of additional instruments available), which is equal to 1 in this example.

$$S = NR^2 \sim \chi_{df}^2 \quad (4-3)$$

If the test is rejected (S is larger than the critical value for a certain level of significance) this is evidence that the specification of the model is incorrect and/or that at

least one of the instruments is invalid. The test gives no information on which one might be the invalid instrument. If the test is accepted (S is small) it is evidence that both instruments are suitable and that there are no other model specification issues. However, as with any statistical test, it could also be that the instruments are really not valid and the test just has low power.

Regarding the power of the Sargan test, Newey (1985b) shows that over-identification tests are inconsistent. These tests are blind to certain alternate hypothesis, meaning that, in certain cases, the power of the tests is never equal to one, even when the sample size goes to infinity. To account for this fact, over-identification tests are sometime stated under the assumption that, at least, a subset of the instruments are exogenous (Stock, 2001), a condition that cannot be tested. This consideration seems to discourage the use of methods based on instrumental variables because they are grounded in an unverifiable assumption. However, De Blander (2008) shows that the alternate hypotheses for which over-identification tests are blind is very peculiar. If and only if the instruments appear in ζ in the same linear combination that they appear in the price equation, over-identification tests will not be able to detect the endogeneity of the instruments. The assumption that this particular event does not occur seems easier to defend than to attack. This fact gives a reasonable sustain for the usage of tests for over-identifying restrictions for the validity of instruments.

4.2.2 The Amemiya-Lee-Newey Test in Discrete Choice models

Amemiya (1978) proposed a two-stage minimum-chi squared estimator for the simultaneous equations Probit model that Newey (1987) proved to be efficient compared to other two-stage procedures. Later, Lee (1992) noted that Amemiya's estimator can also be used to test for the validity of instruments. The test can be extended to other discrete choice models.

To describe Amemiya's estimator under the setting used in this thesis, reconsider the discrete choice problem stated in previous chapters where household n chooses an alternative i among those in the choice-set C_n . Households make their choices based on a latent utility U_{in} that depends on the price p , a control x and an error ε . The price p

depends on two instruments z_1 and z_2 , and the error term δ . The model suffers from endogeneity because δ is correlated with ε .

$$\begin{aligned} U_{in} &= \beta_p p_{in} + \beta_x x_{in} + \varepsilon_{in} \\ p_{in} &= \alpha_{z_1} z_{1in} + \alpha_{z_2} z_{2in} + \delta_{in} \\ y_{in} &= 1[U_{in} = \max_{j \in C_n} \{U_{jn}\}] \end{aligned}$$

The first step in the calculation of Amemiya's estimator is to replace the equation of price, into the structural equation of the utility function. By this, the following equation for the utility is obtained:

$$\begin{aligned} U_{in} &= \beta_p (\alpha_{z_1} z_{1in} + \alpha_{z_2} z_{2in} + \delta_{in}) + \beta_x x_{in} + \varepsilon_{in} \\ U_{in} &= \underbrace{\beta_p \alpha_{z_1}}_{\pi_1} z_{1in} + \underbrace{\beta_p \alpha_{z_2}}_{\pi_2} z_{2in} + \underbrace{\beta_x}_{\pi_3} x_{in} + \underbrace{\varepsilon_{in} + \beta_p \delta_{in}}_{\tilde{\varepsilon}_{in}}. \end{aligned}$$

This equation is termed a reduced-form equation for the utility, where the right hand side is compounded only by exogenous variables: instruments (z_1 and z_2) and controls (x). Note that the price equation is then also a reduced-form equation.

By means of this transformation, the model no longer suffers from endogeneity since neither z_1 , z_2 nor x are correlated with ε or δ . Then the estimation of this model would result in consistent estimators $\hat{\pi}$ of π_1 , π_2 and π_3 . Note that consistency is only up to a scale because the variance of $\tilde{\varepsilon}_{in}$ is different from the variance of ε_{in} in the true model.

The researcher is however interested in gathering consistent estimators of the parameters of the structural equation of the utility, which in this example correspond to β_p and β_x . β_x can be retrieved directly from the estimator of π_3 . In turn, a consistent estimator for β_p can be obtained by means of a two-stage procedure.

Note first that it is possible to obtain consistent estimators $\hat{\alpha}$ of α_{z_1} and α_{z_2} , by regressing p on z_1 and z_2 . Then, using the estimators $\hat{\pi}$ and $\hat{\alpha}$, the following set of equations can be constructed:

$$\begin{aligned} \hat{\pi}_1 &= \beta_p \hat{\alpha}_{z_1} \\ \hat{\pi}_2 &= \beta_p \hat{\alpha}_{z_2}. \end{aligned}$$

These two equations can be seen as observations from the following auxiliary model

$$\hat{\pi} = \beta_p \hat{\alpha} + \gamma,$$

where γ is an error term, and β_p is the only coefficient to be estimated. Note that the auxiliary model for this example has one estimatable coefficient and only two observations because there is only one endogenous variable and two instruments. Each additional endogenous variable would result in an additional estimatable coefficient, and each additional instrument would result in an additional observation.

To estimate this auxiliary model, Amemiya (1978) proposed the following minimum chi-squared estimator

$$\min_{\beta_p} (\hat{\pi} - \beta_p \hat{\alpha})' \hat{W}^{-1} (\hat{\pi} - \beta_p \hat{\alpha})$$

Amemiya (1978) proved that this estimator is consistent. Newey (1987) proved that if \hat{W} is a consistent estimator of the variance-covariance matrix of $(\hat{\pi} - \beta_p \hat{\alpha})$, Amemiya's estimator will also be, at least, as efficient any two-stage estimator, such as 2SCF or 2SIV.

The calculation of \hat{W} is cumbersome for three reasons. First, the matrix needs to be invertible. If it is not invertible, it would be possible to use the pseudo-inverse (Rao and Mitra, 1971). Second, the calculation requires a consistent estimator of β_p . This estimator can be obtained from a preliminary estimation using $\hat{W} = I$, or from the two-stage methods used to address endogeneity described in previous chapters. Third, to achieve efficiency, the calculation of \hat{W} requires a consistent estimator of the joint asymptotic variance-covariance matrix of $\hat{\pi}$ and $\hat{\alpha}$. One possible simplifying assumption for this last requirement would be to consider the result from Hausman (1978) in order to state that

$$\text{Var}(\hat{\pi} - \beta_p \hat{\alpha}) \approx \text{Var}(\hat{\pi}) - \hat{\beta}_p^2 \text{Var}(\hat{\alpha}),$$

where the variance-covariance matrices of $\hat{\pi}$ and $\hat{\alpha}$ can be retrieved from the estimators of the previous stages of Amemiya's method, and $\hat{\beta}_p$ is a consistent estimator of β_p . I use this transformation in the Monte Carlo experiments and in the application with real data, later in this chapter.

The utilization of Amemiya's procedure as an estimation method to correct for endogeneity in discrete choice models seems less attractive than those analyzed in previous chapters. Indeed, it is difficult to obtain an appropriate estimator \hat{W} such that the method would be efficient, particularly when few instruments are available.

Additionally, the estimation of the method cannot be performed with commercial software, the calculation of the correct standard errors is complex, and it is unclear how to do forecasting with Amemiya's method.

However, there is an important byproduct from Amemiya's procedure. Lee (1992) noted that the objective function of Amemiya's estimator can be used to construct a test of over-identifying restrictions. This test can be used to test for the validity of instrumental variables in discrete choice models. Intuitively, if the instruments are valid, the model will be consistent. How far the objective function is from zero will depend solely on the degree of over-identification of the model, the number of extra instruments available. In turn, if the instruments are invalid, the estimators will be inconsistent. How far the objective function of Amemiya's estimator is from zero will be affected by the inconsistency caused by the use of invalid instruments. Lee (1992) showed that Eq. (4-4), known as the Amemiya-Lee-Newey statistic, follows a chi-squared distribution with degrees of freedom (df) equal to the degrees of over-identification of the model, which is equal to 1 in this example.

$$ALN = N(\hat{\pi} - \hat{\beta}_p \hat{\alpha})' \hat{W}^{-1} (\hat{\pi} - \hat{\beta}_p \hat{\alpha}) \sim \chi_{df}^2 \quad (4-4)$$

4.3 Two Novel Tests for Discrete Choice Models

4.3.1 A Regression-based Test for Logit Models

In this section I develop a novel test for the validity of instruments that is applicable to Logit models and was originally sketched by Guevara and Ben-Akiva (2008). The test is an extension of the Sargan test that is grounded in an application of the concept of generalized residuals (Cox and Snell, 1968) applied to Logit Models by McFadden (1987).

The crucial step in the adaptation of the Sargan test for the Logit model lies in the identification of a Logit analogous for the residuals of the 2SLS regression in linear models. Cox and Snell (1968) were the first to define the residuals in a nonlinear framework. McFadden (1987) developed a series of Regression-based tests for Logit where he used the concept of generalized residuals and derived the transformations

required to mimic the asymptotic distribution of the Logit errors with linear regressions. These tests were constructed based on a LaGrange-multiplier test for omitted attributes in Logit models with linear utilities.

To describe McFadden's omitted attributes test, consider a problem where N households face the choice among J alternatives in a choice-set C_n . Each household n retrieves a random utility U_{in} from each alternative in their choice-set. The utilities depend linearly on the attributes x and z , and an error term ε that is distributed *iid* Extreme Value and is also uncorrelated with z and x .

$$U_{in} = \beta_x x_{in} + \beta_z z_{in} + \varepsilon_{in}$$

$$y_{in} = 1[U_{in} = \max_{j \in C_n} \{U_{jn}\}]$$

The researcher wonders if the attribute z is exogenous to the model; that is, if $\beta_z = 0$. If the dimension of z is one, the null hypothesis $H_0 : \beta_z = 0$ can be tested using a Quasi-t, a Likelihood-ratio test or a LaGrange-multiplier test. If the dimension of z is larger than one, only the last two tests are suitable.

McFadden (1987) used the following LaGrange-multiplier test for this problem, where β is the vector of model parameters, L is the log-likelihood of the model, $I(\beta)$ is the Fisher-information-matrix, and the degrees of freedom (df) are equal to the dimension of z .

$$\frac{\partial L}{\partial \beta}(\hat{\beta} | \hat{\beta}_z = 0) [I(\hat{\beta} | \hat{\beta}_z = 0)]^{-1} \frac{\partial L}{\partial \beta}(\hat{\beta} | \hat{\beta}_z = 0) \sim \chi_{df}^2$$

Extending a result obtained by Engle (1984) for binary Logit, McFadden (1987) showed that if the underlying model is Logit with linear utilities, this LaGrange-multiplier test would be asymptotically equivalent to a two-stage Regression-based test. To show this result, consider that the null hypothesis is true, $\beta_z = 0$. Then, the log-likelihood of the model can be written as

$$L_n = \sum_{j \in C_n} y_{jn} \beta_x x_{jn} - \ln \sum_{j \in C_n} \exp(\beta_x x_{jn})$$

Taking the derivatives of the log-likelihood with respect to β_x results in

$$\frac{\partial L_n}{\partial \beta_x} = \sum_{j \in C_n} y_{jn} x_{jn} - \frac{1}{\sum_{j \in C_n} \exp(\beta_x x_{jn})} \sum_{j \in C_n} \exp(\beta_x x_{jn}) x_{jn},$$

$$\frac{\partial L_n}{\partial \beta_x} = \sum_{j \in C_n} y_{jn} x_{jn} - \sum_{j \in C_n} P_n(j) x_{jn} = \sum_{j \in C_n} (y_{jn} x_{jn} - P_n(j) x_{jn})$$

where

$$P_n(j) = \frac{\exp(\beta_x x_{jn})}{\sum_{k \in C_n} \exp(\beta_x x_{kn})} \text{ is the choice probability of alternative } j.$$

Note that

$$\sum_{j \in C_n} \left(P_n(j) \sum_{k \in C_n} P_n(k) x_{kn} - y_{jn} \sum_{k \in C_n} P_n(k) x_{kn} \right) = 0 \text{ because } \sum_{j \in C_n} y_{jn} = \sum_{j \in C_n} P_n(j) = 1. \text{ Then}$$

$$\frac{\partial L_n}{\partial \beta_x} = \sum_{j \in C_n} \left(y_{jn} x_{jn} - P_n(j) x_{jn} - y_{jn} \sum_{k \in C_n} P_n(k) x_{kn} + P_n(j) \sum_{k \in C_n} P_n(k) x_{kn} \right)$$

$$\frac{\partial L_n}{\partial \beta_x} = \sum_{j \in C_n} \left[y_{jn} \left(x_{jn} - \sum_{k \in C_n} P_n(k) x_{kn} \right) - P_n(j) \left(x_{jn} - \sum_{k \in C_n} P_n(k) x_{kn} \right) \right]$$

$$\frac{\partial L_n}{\partial \beta_x} = \sum_{j \in C_n} (y_{jn} - P_n(j)) \left(x_{jn} - \sum_{k \in C_n} P_n(k) x_{kn} \right).$$

Note that this transformation of the derivative of the log-likelihood resulted in a condition that is equivalent to the orthogonality property of OLS estimates that establishes that the residuals of an OLS regression are orthogonal to the independent variables of the model (see, e.g., Greene, 2003). In this case, the term $(y_{jn} - P_{jn})$ corresponds to the crude residuals (termed by Cox and Snell, 1968), and the columns of the matrix of independent variables correspond to the following transformation of variable x :

$$x_{jn} - \sum_{k \in C_n} P_n(k) x_{kn}.$$

The derivative of the log-likelihood remains unchanged if it is multiplied and divided by the square root of the probability of choosing each alternative. McFadden (1987) shows that this transformation assures that the generalized residuals defined later will

have the same asymptotic properties as those of the Logit model. Under this transformation, the derivative of the log-likelihood becomes:

$$\frac{\partial L_n}{\partial \beta_x} = \sum_{j \in C_n} \underbrace{\frac{(y_{jn} - P_n(j))}{\sqrt{P_n(j)}}}_{\bar{\varepsilon}_j} \underbrace{\left(x_{jn} - \sum_{k \in C_n} P_n(k) x_{kn} \right)}_{\bar{x}_j} \sqrt{P_n(j)} = \sum_{j \in C_n} \bar{\varepsilon}_j \bar{x}_j. \quad (4-5)$$

The intuition of McFadden's test of omitted attributes is the following. If the model is estimated omitting z , the expression in Eq. (4-5) will be equal to zero for the estimated values of the model parameters. Term $\hat{P}_n(i)$ the fitted probabilities resulting from this model. Then, under the null hypothesis that $\beta_z = 0$, the expression shown in Eq. (4-6) should also be similar to zero.

$$\sum_{j \in C_n} \frac{(y_{jn} - \hat{P}_n(j))}{\sqrt{\hat{P}_n(j)}} \left(z_{jn} - \sum_{k \in C_n} \hat{P}_n(k) z_{kn} \right) \sqrt{\hat{P}_n(j)} = \sum_{j \in C_n} \bar{\varepsilon}_{jn} \bar{z}_{jn} \quad (4-6)$$

In other words, Eq. (4-6) indicates that $\bar{\varepsilon}$ should be almost orthogonal to \bar{z} if $\beta_z = 0$. Then, intuitively, the R^2 of a regression of $\bar{\varepsilon}$ onto \bar{x} and \bar{z} should be very small. McFadden (1987) formally proved that a test based on the R^2 of an OLS regression of $\bar{\varepsilon}$ onto \bar{x} and \bar{z} is asymptotically equal to the LaGrange-multiplier test for the omission of z . McFadden's test of omitted attributes can be stated as follows:

Stage 1) Estimate a Logit model considering only x and use the estimators of this model to calculate the fitted probabilities $\hat{P}_n(i)$ and the following auxiliary variables:

$$\begin{aligned} \bar{x}_{in} &= \left(x_{in} - \sum_{j \in C_n} x_{jn} \hat{P}_n(j) \right) \sqrt{\hat{P}_n(i)} & \bar{z}_{in} &= \left(z_{in} - \sum_{j \in C_n} z_{jn} \hat{P}_n(j) \right) \sqrt{\hat{P}_n(i)} \\ \bar{\varepsilon}_{in} &= \frac{(y_{in} - \hat{P}_n(i))}{\sqrt{\hat{P}_n(i)}}. \end{aligned}$$

Stage 2) Estimate an OLS regression $\bar{\varepsilon}_{in} = \theta_0 + \theta_x \bar{x}_{in} + \theta_z \bar{z}_{in}$. The statistic of the test is

$$M = \tilde{N} R^2 \sim \chi_{df}^2,$$

where \tilde{N} corresponds to the number of cases. If the number of alternatives J is the same for all households, $\tilde{N} = N(J - 1)$. McFadden (1987) showed that statistic M is distributed

χ^2_{df} , where the degrees of freedom df are equal to the number of omitted attributes being tested. df is equal to 1 in this example.

The usefulness of McFadden's omitted attribute test, in the quest for testing the validity of instruments in Logit models, is in that it formally establishes an expression for the generalized residuals of a Logit model and in that it makes the appropriate normalization that allows testing the properties of the model from those residuals.

Using McFadden's derivations it is possible to propose an analogy for the Sargan test that is applicable in Logit models. Consider that the true model can be defined as follows

$$\begin{aligned} U_{in} &= \beta_p p_{in} + \beta_x x_{in} + \varepsilon_{in} = \beta_p p_{in} + \beta_x x_{in} + \xi_{in} + e_{in} \\ p_{in} &= \alpha_{z_1} z_{1in} + \alpha_{z_2} z_{2in} + \delta_{in} \\ y_{in} &= 1[U_{in} = \max_{j \in C_n} \{U_{jn}\}] \end{aligned}$$

where ξ and δ are correlated causing endogeneity, and where z_1 and z_2 are valid instruments.

Under this setting, the following Regression-based test for the validity of instruments in Logit models can be proposed. This test was originally sketched by Guevara and Ben-Akiva (2008).

Stage 1) Estimate the price equation, using OLS to obtain the residuals $\hat{\delta}$

$$p_{in} = \alpha_0 + \alpha_{z_1} z_{1in} + \alpha_{z_2} z_{2in} + \delta_{in} \xrightarrow{OLS} \hat{\delta}_{in}$$

Stage 2) Estimate the choice model, including $\hat{\delta}$ as an additional variable

$$U_{in} = \beta_p p_{in} + \beta_x x_{in} + \beta_\delta \hat{\delta}_{in} + \tilde{e}_{in} \xrightarrow{ML} \hat{P}_n(i)$$

These first two stages correspond to the application of the 2SCF method defined in Chapter 2. The estimators of this model can be used to calculate the fitted probabilities $\hat{P}_n(i)$, which are then used to proceed to Stage 3.

Stage 3) Calculate the following auxiliary variables

$$\begin{aligned} \bar{x}_{in} &= \left(x_{in} - \sum_{j \in C_n} x_{jn} \hat{P}_n(j) \right) \sqrt{\hat{P}_n(i)} & \bar{\tilde{e}}_{in} &= \frac{(y_{in} - \hat{P}_n(i))}{\sqrt{\hat{P}_n(i)}} \\ \bar{z}_{1in} &= \left(z_{1in} - \sum_{j \in C_n} z_{1jn} \hat{P}_n(j) \right) \sqrt{\hat{P}_n(i)} & \bar{z}_{2in} &= \left(z_{2in} - \sum_{j \in C_n} z_{2jn} \hat{P}_n(j) \right) \sqrt{\hat{P}_n(i)}. \end{aligned}$$

Now that we have a simile for the residuals of the control-function regression, Stage 4 mimics the final stage of the Sargan test.

Stage 4) Regress the generalized residuals $\widehat{\bar{e}}_{in}$ by OLS on the transformed instruments $\widehat{\bar{z}}_{1in}$ and $\widehat{\bar{z}}_{2in}$ and the control $\widehat{\bar{x}}_{in}$.

$$\widehat{\bar{e}}_{in} = \theta_0 + \theta_x \widehat{\bar{x}}_{in} + \theta_{z_1} \widehat{\bar{z}}_{1in} + \theta_{z_2} \widehat{\bar{z}}_{2in} \quad (4-7)$$

Finally, calculate the unadjusted multiple correlation coefficient R^2 of this regression and calculate the statistic

$$S_{RB} = \tilde{N}R^2 \sim \chi_{df}^2,$$

where $\tilde{N} = N(J - 1)$ is the number of cases, and the degrees of freedom (df) corresponds to the degree of over-identification, which in this example is equal to 1.

The outcome of this test can be interpreted in the same way as the outcome for the Sargan test. If the null hypothesis is rejected, this means that at least one of the instruments is correlated with the error (is invalid) or that there is another model misspecification. If the null hypothesis is accepted, this is evidence that both instruments are valid.

4.3.2 A Direct test for Discrete Choice Models

The application of the Amemiya-Lee-Newey test and the Regression-based test may be cumbersome and vulnerable to data-processing errors because they involve the calculation of auxiliary variables and/or fitted probabilities, as well as the estimation of several auxiliary regressions. For this reason, I present an alternative test for the validity of instruments that only involves the estimation of discrete choice models with commercial computational packages. I term this the Direct test for the validity of instruments in discrete choice models.

To describe Direct the test, consider that there is a set of K instrumental variables to correct for price endogeneity in the choice model used as an example throughout the chapter. If all K instrumental variables are valid, the model estimated using the control-function correction will be consistent. Then, the subsequent inclusion of any instrument as an additional variable into the corrected model should produce a non-significant

increase in the log-likelihood. In turn, if the instruments are invalid, they will be correlated with the error term of the model, and the inclusion of any instrument as additional variables into the model corrected for endogeneity, should result in a significant increase in the log-likelihood. This suggests an alternative test for the validity of instruments.

Note that only $K-1$ out of all K instruments used in the construction of the residuals $\hat{\delta}$ can be included at the same time as additional variables into the model corrected for endogeneity. The problem is that $\hat{\delta}$ was constructed as a linear function of the endogenous variable (p) and all K instruments. Then, a model including p , $\hat{\delta}$ and all K instruments will be perfectly collinear, making the model non-estimatable. For the example used in this chapter, considering that z_1 and z_2 are used to construct $\hat{\delta}$, the Direct test would just correspond to the test for exogeneity of z_1 (or z_2).

The Direct test proposed in this section is, in some sense, similar to the Refutability test used by Card (1995). In the Refutability test, the validity of an instrument is tested by including it in a model that was corrected for endogeneity using an alternative instrument. In that case, the validity of one instrument is conditional on the validity of the other instrument. Instead, for the Direct proposed in this section, all instruments are used to correct for endogeneity and then, the alternate hypothesis is that, at least, one of the instruments is invalid. Also, equivalently to what pointed out by De Blander (2008) for linear models, the Direct test will have no power if the instrumental variables appear in the same linear combination in the price equation and in the utility function.

Two issues have to be remarked about the Direct test. The first is that the test will be valid only asymptotically. This is because the fact that $\hat{\delta}$ was built using z_1 will reduce the size of the test in finite samples. The second issue is that although the 2SCF results in consistent estimators of the model parameters, all statistical tests derived from it are invalid. This problem can be avoided by using the tractable maximum-likelihood estimator studied in Chapter 3. However, in practice, the impact of using the 2SCF in hypothesis testing is minimal, and its usage simplifies enormously the calculation of the statistics.

If the degrees of over-identification are only 1, the Direct test can be calculated as a Quasi-t test, as Lagrange-multiplier test or as a Likelihood-ratio test. In a general case, only the two last options are suitable. Interestingly, the LaGrange-multiplier and Likelihood-ratio versions of the Direct test consider a statistic that is distributed χ^2 , with degrees of freedom equal to the degrees of over-identification of the problem, the same distribution of the statistics of the Amemiya-Lee-Newey test and of the Regression-based test.

The LaGrange-multiplier version of the Direct test can be applied to any discrete choice model. In the particular case of Logit, the test can be calculated using the R^2 of the following auxiliary regression:

$$\widehat{e}_{in} = \theta_0 + \theta_x \widehat{x}_{in} + \theta_p \widehat{p}_{in} + \theta_\delta \widehat{\delta}_{in} + \theta_{z_1} \widehat{z}_{1in}.$$

The LaGrange-multiplier version of the Direct test can be alternatively implemented considering, instead of a linear regression, the estimation of a modified choice model in the final stage. This variation comes from an alternative implementation of McFadden's (1987) test for omitted attributes, used by Train et al. (1989) to test for non-IIA error structures. First, it is necessary to calculate a slightly different version of the auxiliary variable for the instrument

$$\overline{\widehat{z}}_{1in} = \left(z_{1in} - \sum_{j \in C_n} z_{1jn} \widehat{P}_j(j) \right).$$

Then, the choice model is re-estimated considering the following specification of the utility function:

$$V_{in} = \beta_p p_{in} + \beta_x x_{in} + \beta_\delta \widehat{\delta}_{in} + \beta_{\overline{z}_1} \overline{\widehat{z}}_{1in}.$$

Finally, the test is implemented as a Quasi-t test for the null hypothesis that $\beta_{\overline{z}_1} = 0$.

The Likelihood-ratio version of the Direct test can be applied to any discrete choice model and has the important advantage of requiring only an auxiliary estimation of the choice model and no need for additional transformations. Under these considerations, the following two-stage Direct test for the validity of instruments can be proposed:

Stage 1) Estimate the price equation using OLS to obtain the residuals $\widehat{\delta}$

$$p_{in} = \alpha_0 + \alpha_{z_1} z_{1in} + \alpha_{z_2} z_{2in} + \delta_{in} \xrightarrow{OLS} \hat{\delta}_{in}$$

Stage 2) Estimate the 2SCF model and retrieve the log-likelihood L_{2SCF} .

$$U_{in} = \beta_p p_{in} + \beta_x x_{in} + \beta_\delta \hat{\delta}_{in} + \tilde{e}_{in} \xrightarrow{ML} L_{2SCF}$$

Stage 3) Estimate the choice model including $\hat{\delta}$ and one of the instruments (for example z_1) as additional variables and retrieve the log-likelihood L_D .

$$U_{in} = \beta_p p_{in} + \beta_x x_{in} + \beta_\delta \hat{\delta}_{in} + \beta_{z_1} z_{1in} + \tilde{e}_{in} \xrightarrow{ML} L_D$$

The evaluation of the null hypothesis $H_0 : \hat{\beta}_{z_1} = 0$ can be done using a Likelihood-ratio test comparing the likelihood of the model estimated in Stage 3 with that estimated in Stage 2 as follows

$$S_{DIRECT} = -2(L_{2SCF} - L_D) \sim \chi_{df}^2,$$

where the degrees of freedom (df) in this case is equal to 1.

The Direct test is equivalent to the Amemiya-Lee-Newey and the Regression-based test for the validity of instruments for Logit models in that they require over-identification to be performed. Their outcomes can also be equivalently interpreted. If the null hypothesis $H_0 : \hat{\beta}_{z_1} = 0$ is rejected, then at least one of the instruments is not valid, although we cannot tell which one. If the null hypothesis is accepted, this is evidence that the instruments are appropriate.

In summary, three tests for the validity of instruments in discrete choice models have been identified: The Amemiya-Lee-Newey test, the Regression-based test, and the Direct test. Both the Amemiya-Lee-Newey and the Direct test can be applied to any discrete choice model. The Regression-based test can be applied only for Logit. On the other hand, Amemiya-Lee-Newey and the Regression-based test require the estimation of auxiliary regressions and the calculation of auxiliary variables, whereas the Direct test can be implemented with a single re-estimation of the choice model. This simplicity makes the Direct test extremely attractive for practitioners. In the next section I use Monte Carlo experimentation to compare the size and power properties of the three tests for the validity of instruments investigated in this chapter.

4.4 Monte Carlo Experiment

In this section, I perform a series of Monte Carlo experiments to demonstrate and to investigate the behavior of the tests for the validity of instruments in discrete choice models. For experimentation purposes, the true or underlying model used to develop the experiments is a binary Logit model where the utility of each alternative depends on its price p , a control x , and an error term ε , which is divided into two components, ξ and e . ξ represents an omitted attribute that is correlated with p , and e is an *iid* error distributed Extreme Value (0,1). The value of the model coefficients in the true model are shown in Eq. (4-8).

$$U_{in} = -1p_{in} + 1x_{in} + \underbrace{\xi_{in} + e_{in}}_{\varepsilon_{in}} \quad (4-8)$$

The price p was constructed as a function of ξ and two exogenous variables z_1 and z_2 , as shown in Eq. (4-9), where $\tilde{\delta}$ is an error *iid* Normal (0,1). Variables x , z_1 , z_2 , and ξ were constructed Uniform (-3,3). Under this setting, if ξ is omitted in the specification of the utility, the price will be correlated with the error term ε causing endogeneity. On the other hand z_1 and z_2 are valid instruments for p because they are correlated with it, and uncorrelated with ε .

$$p_{in} = 0.5\xi_{in} + 0.5z_{1in} + 0.5z_{2in} + \tilde{\delta}_{in} \quad (4-9)$$

Dahlberg et al. (2008), De Blander (2008), Newey (1985b), and others have shown that the Sargan test has low power properties in linear models. To analyze the power properties of the tests for the validity of instruments in discrete choice models, I build two invalid instruments: b_1 and b_2 and investigated the success of the test in detecting that the instruments are invalid.

Variables b_1 and b_2 are invalid instruments because they are correlated with ξ and therefore, with the error term ε of the model. Following the motivation shown in Figure 4-1, it can be expected that the power properties of the test can be reduced when the invalid instruments become highly correlated, or equivalently, as the angle between b_1 and b_2 shrinks. In this case, the residuals $\hat{\varepsilon}$ would become almost orthogonal to the instruments, by construction, yielding to false acceptances of the null hypothesis. To

evaluate this hypothesis, the invalid instruments were constructed as shown in Eq. (4-10), where $c \in (0,1)$ and ψ_{in} and $\tilde{\psi}_{in}$ were generated *iid* Normal (0,1).

$$\begin{aligned} b_{1in} &= 1\xi_{in} + 1p_{in} + \psi_{in} \\ b_{2in} &= c b_{1in} + (1-c)p_{in} + \tilde{\psi}_{in} \end{aligned} \quad (4-10)$$

Under this setting the correlation between b_1 and b_2 will increase with c . b_1 and b_2 will be correlated with ξ and p for all values of c , which will make them invalid instruments but also relevant in the price equation. This allows differentiating the problem that the instruments are correlated with the error term, from the problem that the instruments are weak.

The tests analyzed in these experiments were the Amemiya-Lee-Newey test, the Regression-based test and the Direct test. A total of 100 realizations of the data and different sample sizes N were used in the analysis. The performance of the tests was evaluated considering three situations: 1) two valid instruments (z_1 and z_2) are used to correct for endogeneity using the 2SCF method; 2) one valid (z_1) and one invalid (b_1) instrument are used in the correction for endogeneity; 3) two invalid instruments (b_1 and b_2) are used in the application of the 2SCF method. In the third experiment, the correlation among the invalid instruments was changed by varying the values of variable c .

Table 4-1 shows the number of times each test resulted in an acceptance at 5% significance, and the corresponding bias, mean squared error (MSE) and the t-test against the true value of the ratio between the estimators of the coefficient of p and of the coefficient of x , which is -1 in this experiment. As discussed in Chapter 2, to check the consistency of the estimators, it is necessary to look at the ratio of the coefficients and not at the coefficients themselves, because the estimators obtained with the control-function correction are only consistent up to a scale.

Consider the case where two valid instruments are used to correct for endogeneity. These results are reported in the first four rows (below the headings) of Table 4-1, for sample sizes N of 100, 500, 1,000 and 2,000, respectively. In this case the bias, the MSE and the value of the t-test of the ratio $\hat{\beta}_p / \hat{\beta}_x$ against its true value, are small for all the sample sizes analyzed. This means that when the two valid instruments are used, the

2SCF method satisfactorily addressed the endogeneity problem caused by the omission of attribute ζ . It would therefore be desirable to have the tests for the validity of instruments accept the null hypothesis that the instruments are not correlated with the model error. Table 4-1 shows that all tests have similar size. The empirical confidence is equally near to the nominal confidence (95%) for all sample sizes and for all the tests.

Table 4-1 Monte Carlo Experiment: Performance of Tests for the Validity of Instruments

| N | Acceptances out of 100 5% significance | | | Bias | MSE | t-test true |
|---|---|----------------------|--------|------------|----------|-------------|
| | Amemiya- Lee-Newey | Regression- based | Direct | | | |
| $\hat{\beta}_p / \hat{\beta}_x$ | | | | | | |
| 2 Valid Instruments | | | | | | |
| 100 | 92 | 91 | 91 | 0.1014 | 0.1185 | 0.3083 |
| 500 | 91 | 92 | 92 | 0.002114 | 0.02153 | 0.01441 |
| 1,000 | 92 | 93 | 94 | -0.002516 | 0.008195 | -0.02780 |
| 2,000 | 95 | 95 | 96 | 0.00009412 | 0.003647 | 0.001558 |
| 1 Valid and 1 Invalid Instrument | | | | | | |
| 100 | 18 | 12 | 11 | -0.7572 | 0.6070 | -4.130 |
| 500 | 0 | 0 | 0 | -0.7761 | 0.6105 | -8.565 |
| 2 Invalid Instruments c=0.1 Correlation $b_1, b_2=0.7718$ | | | | | | |
| 100 | 6 | 2 | 2 | -0.6464 | 0.4544 | -3.382 |
| 500 | 0 | 0 | 0 | -0.6703 | 0.4573 | -7.471 |
| 2 Invalid Instruments c=0.5 Correlation $b_1, b_2=0.9012$ | | | | | | |
| 100 | 39 | 30 | 27 | -0.7918 | 0.6605 | -4.320 |
| 500 | 0 | 0 | 0 | -0.8055 | 0.6563 | -9.255 |
| 2 Invalid Instruments c=0.9 Correlation $b_1, b_2=0.9489$ | | | | | | |
| 100 | 95 | 91 | 91 | -0.9280 | 0.8907 | -5.409 |
| 500 | 79 | 56 | 57 | -0.9389 | 0.8883 | -11.37 |
| 1,000 | 68 | 46 | 49 | -0.9392 | 0.8848 | -17.92 |
| 2,000 | 43 | 14 | 15 | -0.9406 | 0.8866 | -22.34 |
| 5,000 | 4 | 0 | 0 | -0.9449 | 0.8935 | -36.79 |

100 Repetitions. $J=2$

Consider the case where one valid and one invalid instrument are used. In this experiment, the bias, the MSE and the value of the t-test of the ratio $\hat{\beta}_p / \hat{\beta}_x$ against its true value, are large for all the sample sizes N analyzed. This means that, because one of the instruments was invalid, the 2SCF method did not solve the endogeneity problem caused by the omission of the attribute ζ . It would then be desirable to have the tests for the validity of instruments reject the null hypothesis. The results in Table 4-1 show that

for a sample size of 100 observations, the Amemiya-Lee-Newey test resulted in 18 false acceptances, the Regression-based resulted in 12 and the Direct test resulted in only 11. These results show that both the Regression-based and the Direct test have better power properties than the Amemiya-Lee-Newey test. For sample sizes of 500 and larger (not reported in Table 4-1), the number of false acceptances became zero for all three tests.

Consider the case where both instruments are invalid, that is, when both are correlated with the omitted attribute that causes endogeneity. Table 4-1 shows that for all the values of c analyzed, the bias, MSE, and the t-test of the ratio $\hat{\beta}_p / \hat{\beta}_x$ against its true value, are significantly large. In this case it would be desirable to have the tests reject the null hypothesis. Table 4-1 shows that when the correlation between the invalid instruments is 0.7718 and the sample size is 100, there are only 2 out of 100 false acceptances for the Regression-based and for the Direct tests. For the Amemiya-Lee-Newey test, the number of Type II errors increases up to 6, which further shows that this test has lower power. Again, for sample sizes of 500 and larger (not reported in Table 4-1), the number of false acceptances is zero for all three tests. Something similar occurs when the correlation between the invalid instruments increases to 0.9012. In this case, there are Type II errors only when the sample size is 100. The false acceptances are 27 for the Direct test, 30 for the Regression-based test and 39 for the Amemiya-Lee-Newey test. The picture is very different when the correlation jumps to 0.9489. In this case, there are false acceptances even when the sample size is as large as 5,000 observations. Interestingly, for all cases the power properties of the Amemiya-Lee-Newey test were always below those of the Regression-based and the Direct tests.

In summary, the Monte Carlo experiments showed that, for this setting, the Regression-based and Direct tests have similar size and power properties and that their power is superior to that of the Amemiya-Lee-Newey test. The Direct test showed to be a reliable tool for testing the validity of instruments in this framework. This is attractive for practitioners since the Direct test is easily calculable with commercial packages because it only involves the re-estimation of the choice model with an additional variable. Additionally, it became evident that the correlation between the instruments can severely affect the power of the tests, even for large sample sizes. To the best of my knowledge, this has not been noted before and raises a warning for the usual practice (see, e.g.,

Nichols, 2007) of attaining over-identification, to be able to test for the validity of instruments, by generating additional instruments as non-linear transformations of available instruments.

4.5 Application to Real Data

In this section I re-visit the residential location choice model of Lisbon estimated in Chapter 2. Although the process behind the construction of the instruments used for the correction for endogeneity and their effect on the estimates were theoretically sound, it is necessary to perform formal tests to verify their validity.

The tests for the validity of instruments rely on the over-identification of the model. The model estimated in Chapter 2 considered two instruments (the averages of two different sets of dwellings) to correct for one continuous endogenous variable (dwelling price). As it was noted in the Monte Carlo experiments, if the instruments are highly correlated, the power of the tests may be severely affected. In the case of this residential location choice model, the correlation between the instruments equaled to 0.8238, as shown in Table 2-4. This is below the empirical threshold of ~0.95 found in the Monte Carlo experiments and therefore gives some confidence in the power of the tests for the validity of instruments calculated for Lisbon's model.

I begin by calculating the Amemiya-Lee-Newey test. For this test, it is necessary to estimate two models: 1) the regression of the price on the instruments, and 2) the estimation of a choice model where the price is substituted by the instrumental variables. The former corresponds to the same model reported in Table 2-5. The implementation of the latter has one small shift compared to the models estimated in the Monte Carlo experiment. In the application with real data, the endogenous attribute (price) was interacted with a household characteristic (Income). Under this consideration, the specification of the price part of the utility of the auxiliary (reduced-form) choice model needs to be adjusted as shown in the following expression:

$$U_{in} = (\pi_{z_1} z_1 + \pi_{z_2} z_2) (1 + \pi_{2000} 1[Income > 2,000] + \pi_{5000} 1[Income > 5,000]) + \dots + \varepsilon_{in}.$$

The estimators of this model are shown in Table 4-2.

Amemiya's estimator is obtained by solving the following problem using the method described in Section 4.2.2

$$\hat{\pi}_{z_1} = \beta_p \hat{\alpha}_{z_1} + \gamma_1$$

$$\hat{\pi}_{z_2} = \beta_p \hat{\alpha}_{z_2} + \gamma_2,$$

where $\hat{\pi}$ correspond to the estimators of the instrumental variables in the auxiliary choice model reported in Table 4-2. $\hat{\alpha}$ are the estimators of the coefficients of the instrumental variables obtained in the first stage of the 2SCF method, which are reported in Table 2-5.

Table 4-2 Lisbon's Logit Model: Auxiliary Choice Model for Amemiya-Lee-Newey Test

| Variables | Reduced-Form Model | |
|--|--------------------|---------|
| | $\hat{\pi}$ | s.e |
| 1. z_1 | -1.759 | 0.5261 |
| 2. z_2 | -0.9197 | 0.7184 |
| 3. 1[Income > 2,000 €/M] | 0.7300 | 0.1918 |
| 4. 1[Income > 5,000 €/M] | 0.3496 | 0.2226 |
| 4. Distance to Workplace (in Km) | -0.2418 | 0.05279 |
| 5. Log [Dwelling Area (in m ²)] | 1.902 | 0.7263 |
| 6. Log [Dwelling Age (in years) +1] | -0.4291 | 0.1181 |
| Log likelihood at Convergence $L(\hat{\beta})$ | -566.66 | |
| Log likelihood at Zero $L(0)$ | -589.06 | |
| Adjusted ρ^2 | 0.04992 | |
| Sample Size N | 63 | |
| Choice-Set Size J | 11,501 | |

Logit Model combining Imokapa database and SOTUR survey for Lisbon, Odivelas and Amadora. €/M: Euros per month.

The statistic of the Amemiya-Lee-Newey test was calculated using the expression shown in Eq. (4-4). The value of the statistic is shown in Eq. (4-11), where it should be noted that it is far below the threshold to reject the null hypothesis that the instruments are valid.

$$ALN = 0.1162 < \chi_{1,95\%}^2 = 3.842 \quad (4-11)$$

The second test performed is the Regression-based test described in Section 4.3.1. First, using the estimates of the model corrected for endogeneity reported in Table 2-6, I

calculated the fitted probabilities $\hat{P}_n(i)$. Then, I calculated the auxiliary variables as shown below

$$\begin{aligned}\bar{\hat{x}}_{in} &= \left(x_{in} - \sum_{j \in C_n} x_{jn} \hat{P}_n(j) \right) \sqrt{\hat{P}_n(i)} & \bar{\hat{e}}_{in} &= (y_{in} - \hat{P}_n(i)) / \sqrt{\hat{P}_n(i)} \\ \bar{\hat{z}}_{1in} &= \left(z_{1in} - \sum_{j \in C_n} z_{1jn} \hat{P}_n(j) \right) \sqrt{\hat{P}_n(i)} & \bar{\hat{z}}_{2in} &= \left(z_{2in} - \sum_{j \in C_n} z_{2jn} \hat{P}_n(j) \right) \sqrt{\hat{P}_n(i)},\end{aligned}$$

where x corresponds to the explanatory variables of the residential location choice model estimated in Chapter 2, including the distance to workplace, the log of the area, and the log of the age (+1) of each dwelling.

The next step corresponds to the estimation of an OLS regression of the generalized residuals $\bar{\hat{e}}_{in}$ as a function of the other auxiliary variables $\bar{\hat{x}}$ and $\bar{\hat{z}}$. The results of this OLS regression are shown in Table 4-3. Note that the R^2 of this model is very small and that all variables are statistically equal to zero with 95% confidence. This is a first indication that the instruments are valid and therefore, that the correction for endogeneity was successful. The formal Regression-based test statistic is calculated using the R^2 from Table 4-3, as shown in Eq. (4-12). Note that the statistic is far below the critical value with 95% confidence for the chi-square distribution with one degree of freedom. This result confirms again that the instruments are valid.

Table 4-3 Lisbon’s Logit Model: Auxiliary Regression for Regression-based Test

| Variables | $\hat{\alpha}$ | s.e |
|--------------------------------------|----------------|----------|
| 1. Intercept | 1.363E-05 | 0.001252 |
| 2. $\bar{\hat{z}}_1$ | 0.03436 | 0.3830 |
| 3. $\bar{\hat{z}}_2$ | -0.07846 | 0.5454 |
| 4. $\bar{\hat{x}}_{dist_workplace}$ | 0.00017220 | 0.05555 |
| 5. $\bar{\hat{x}}_{\log(area)}$ | 0.02352 | 0.7491 |
| 6. $\bar{\hat{x}}_{\log(age+1)}$ | -0.004543 | 0.1247 |
| R^2 | 3.295e-08 | |
| Adjusted R^2 | -6.868e-06 | |
| Sample Size $N*J$ | 724,563 | |

Logit Model combining Imokapa database and SOTUR survey for Lisbon, Odivelas and Amadora.

$$S_{RB} = N(J-1)R^2 = 0.2387 < \chi_{1,95\%}^2 = 3.842 \quad (4-12)$$

The final test performed corresponds to the Direct test for the validity of instruments proposed in Section 4.3.2. This test is constructed from the estimation of a Logit model where the utility function includes not only residuals of the first stage of the 2SCF, but also one of the instrumental variables. The results of the estimation of this model are shown in Table 4-4. It should be noted that the coefficient of z_1 in Table 4-4 is not statistically significant (with 95% confidence), as evaluated by a Quasi-t test. This means that the null hypothesis that both instruments are valid is accepted. Equally, Eq. (4-13) shows the statistic of the Direct test calculated as a Likelihood-ratio test, where it can be noted that the outcome is the same, the null hypothesis that both instruments are valid is accepted.

Table 4-4 Lisbon's Logit Model: Auxiliary Choice Model for Direct Test

| Variables | Direct test | |
|--|---------------|---------|
| | z_1 | |
| | $\hat{\beta}$ | s.e |
| 1. Dwelling price (in 100,000 €) | -2.976 | 1.227 |
| 2. Dwelling price * 1[Income > 2,000 €/M] | 0.8533 | 0.5482 |
| 3. Dwelling price * 1[Income > 5,000 €/M] | 0.8093 | 0.4787 |
| 4. Distance to workplace (in Km) | -0.2562 | 0.05336 |
| 5. Log [Dwelling Area (in m ²)] | 2.255 | 0.7461 |
| 6. Log [Dwelling Age (in years) +1] | -0.4650 | 0.1219 |
| 7. δ | 1.215 | 1.122 |
| 8. z_1 | 0.1498 | 0.9566 |
| Log likelihood at Convergence $L(\hat{\beta})$ | -560.04 | |
| Log likelihood at Zero $L(0)$ | -589.06 | |
| Adjusted ρ^2 | 0.06285 | |
| Sample Size N | 63 | |
| Choice-Set Size J | 11,501 | |

Logit Model combining Imokapa database and SOTUR survey for Lisbon, Odivelas and Amadora. €/M: Euros per month

$$S_{DIRECT} = 0.02450 \ll \chi_{1,95\%}^2 = 3.842 \quad (4-13)$$

4.6 Conclusion

In this chapter, I summarized the state-of-the-art in testing for the validity of instruments in discrete choice models. Then I developed two novel tests for the validity of instruments in this framework. The first test was termed the Regression-based and is applicable only to Logit models. This test is an adaptation of the Sargan test for linear models that uses the asymptotic results derived by McFadden (1987) to construct a simile for the residuals in Logit models. The second test developed was termed the Direct test. This test is applicable to diverse choice models and can be easily applied using the outputs from commercial software.

Using Monte Carlo experimentation, I showed that the tests behave as expected and proved, for the binary Logit experiments analyzed, that the Regression-based and Direct tests have better power properties compared to the available Amemiya-Lee-Newey test. I also showed that, when the instruments are highly correlated, the power of the tests may be severely affected. Finally, the application to real data confirmed that the price of similar dwellings, within a certain vicinity, make appropriate instrumental variables for endogeneity in residential location choice modeling. In addition, this application showed that the tests under study were applicable, and performed adequately.

Chapter 5

Sampling of Alternatives in Multivariate

Extreme Value Models

5.1 Overview

The computational burden and the impossibility of identifying or measuring the attributes of a huge number of alternatives in spatial choice models, makes it necessary to only consider a subset of the choice-set in practical applications. McFadden (1978) demonstrated that if the model underlying the choice process is Logit, the problem of sampling of alternatives and estimation can be addressed by adding a corrective constant to the systematic utility of each alternative.

The Logit model requires the assumption that the error terms of the random utilities are uncorrelated among alternatives. This assumption may be invalid for some spatial choice models. In residential location, the error terms may be correlated among dwellings located nearby. Equivalently, in route choice modeling, routes that share sets of common links may be perceived as more similar than other routes that are complete substitutes, breaking from the Logit assumption.

Building on an idea originated by Ben-Akiva (2009), in this chapter, I extend McFadden's results to the Multivariate Extreme Value (MEV) models, a class of closed-

form discrete choice models that allows for different degrees of correlation among alternatives. The chapter is structured as follows. The next section describes McFadden's results on sampling of alternatives in Logit models. Next, the proposed extension to MEV models is presented. The following sections describe the formulation of the proposed methodology to the Nested and the Cross-Nested Logit models, the main members of the MEV family. Then, the effects of the proposed methodology are analyzed using a Monte Carlo experiment and real data on residential location choice from Lisbon, Portugal. The final section summarizes the main conclusions, implications, and potential extensions of this research.

5.2 Estimation and Sampling of Alternatives in Logit Models

Consider the random utility U_{in} that a household n retrieves from alternative i , which can be written as the sum of a systematic part V and a random error term ε , as shown in Eq. (5-1)

$$U_{in} = V_{in} + \varepsilon_{in} = V(x_{in}, \beta^*) + \varepsilon_{in}, \quad (5-1)$$

where the systematic utility depends on variables x and parameters β^* .

Then, if ε is distributed *iid* Extreme Value $(0, \mu)$, the probability that n will choose alternative i will correspond to the Logit model shown in Eq. (5-2), where C_n is the choice-set of J_n elements from which household n chooses an alternative. The scale μ in Eq. (5-2) is not identifiable and usually normalized to equal 1.

$$P_n(i) = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}} \quad (5-2)$$

Consider that, of the true choice-set C_n , only a subset D_n with \tilde{J}_n elements is sampled by the researcher. For estimation purposes, D_n must include (and therefore depends on) the chosen alternative i . Otherwise, the quasi-log-likelihood of the model may become unbounded, making the estimation of the model parameters impossible. To understand why, consider the case of a utility function that is linear in at least one variable x , which can take positive and negative values. If for at least one of the N observations x takes a

positive value for the chosen alternative and, for all alternatives in D_n (which by chance does not include the chosen) x takes negative values, then the quasi-log-likelihood of the model will always increase with the coefficient of x . In other words, the problem will be unbounded.

The joint probability that household n will chose alternative i and that the researcher will construct the set D_n corresponds to $\pi(i, D_n)$. Using the Bayes theorem, this joint probability can be rewritten as shown in Eq. (5-3). $\pi(D_n | i)$ is the conditional probability of constructing the set D_n , given that alternative i was chosen. $\pi(i | D_n)$ is the conditional probability of choosing alternative i , given that the set D_n was constructed.

$$\pi(i, D_n) = \pi(D_n | i)P_n(i) = \pi(i | D_n)\pi(D_n) \quad (5-3)$$

Since the events of choosing each one of the alternatives in C_n are mutually exclusive and totally exhaustive, it is possible to use the Total Probability theorem (see, e.g., Bertsekas and Tsitsiklis, 2002) to write the probability $\pi(D_n)$ of constructing the set D_n as shown in Eq. (5-4), where the second equality holds because $\pi(D_n | j) = 0 \forall j \notin D_n$.

$$\pi(D_n) = \sum_{j \in C_n} \pi(D_n | j)P_n(j) = \sum_{j \in D_n} \pi(D_n | j)P_n(j) \quad (5-4)$$

Substituting Eq. (5-4) and the Logit choice probability $P_n(i)$ shown in Eq. (5-2) into Eq. (5-3), Eq. (5-5) is obtained by canceling and re-arranging terms.

$$\pi(i | D_n) = \frac{e^{V_{in} + \ln \pi(D_n | i)}}{\sum_{j \in D_n} e^{V_{jn} + \ln \pi(D_n | j)}} \quad (5-5)$$

The expression $\ln \pi(D_n | j)$ is termed the sampling correction.

Eq. (5-5) indicates that the conditional probability of choosing alternative i , given that a particular choice-set D_n was constructed, depends only on the alternatives in D_n . This results from the cancellation of the denominators when dividing the probabilities of two alternatives in the Logit model, which is known as the Independence of Irrelevant Alternatives (IIA) property. Note that although IIA is a convenient mathematical property, it results from the assumption that the error structure is *iid*, a statement that may be unrealistic in spatial choice models.

McFadden (1978) demonstrated that if $\pi(D_n | j) > 0$ and known for all j in D_n , and if the true model is Logit with choice-set C_n , it is possible to obtain consistent estimators of the model parameters β^* by maximizing the following quasi-log-likelihood function:

$$QL_{Logit,D} = \sum_{n=1}^N \ln \frac{e^{V(x_{in}, \beta) + \ln \pi(D_n | i, x_n)}}{\sum_{j \in D_n} e^{V(x_{jn}, \beta) + \ln \pi(D_n | j, x_n)}}. \quad (5-6)$$

To demonstrate McFadden's (1978) consistency result, assume that sets C and D do not vary across the sample. This assumption is not essential and can be easily generalized, but helps to reduce the notation considerably.

Then, note that maximizing Eq. (5-6) is the same as maximizing Eq. (5-6) times $1/N$, which is in turn a sample analog for the expected value $E(\cdot)$ of the log-likelihood of Eq. (5-5) over the population.

$$\frac{1}{N} \sum_{n=1}^N \ln \frac{e^{V(x_{in}, \beta) + \ln \pi(D_i, x_n)}}{\sum_{j \in D} e^{V(x_{jn}, \beta) + \ln \pi(D_j, x_n)}} \approx E \left(\ln \frac{e^{V(x_i, \beta) + \ln \pi(D_i, x)}}{\sum_{j \in D} e^{V(x_j, \beta) + \ln \pi(D_j, x)}} \right)$$

The expected value depends on the true parameters β^* , the sampling protocol used to draw D , and the density function of data $f(x)$ as follows:

$$E(\cdot) = \int \ln \left(\frac{e^{V(x_i, \beta) + \ln \pi(D_i, x)}}{\sum_{j \in D} e^{V(x_j, \beta) + \ln \pi(D_j, x)}} \right) f(i, D, x) di dD dx$$

$$E(\cdot) = \int \sum_{i \in C} \sum_{D \subseteq C} \ln \left(\frac{e^{V(x_i, \beta) + \ln \pi(D_i, x)}}{\sum_{j \in D} e^{V(x_j, \beta) + \ln \pi(D_j, x)}} \right) P(i | C, \beta^*, x) \pi(D | i, x) f(x) dx.$$

In re-arranging terms, recall that $\pi(D | j, x) = 0 \forall j \notin D$. Then, it is possible to obtain

$$E(\cdot) = \int \sum_{D \subseteq C} \left[\frac{\sum_{j \in D} e^{V(x_j, \beta^*) + \ln \pi(D_j, x)}}{\sum_{j \in C} e^{V(x_j, \beta^*)}} \sum_{i \in D} \ln \left(\frac{e^{V(x_i, \beta) + \ln \pi(D_i, x)}}{\sum_{j \in D} e^{V(x_j, \beta) + \ln \pi(D_j, x)}} \right) \frac{e^{V(x_i, \beta^*) + \ln \pi(D_i, x)}}{\sum_{j \in D} e^{V(x_j, \beta^*) + \ln \pi(D_j, x)}} \right] f(x) dx.$$

Note that the only part of $E(\cdot)$ depending on variables β (the arguments of the quasi-likelihood maximization problem) have the form of

$$\sum_{i \in D} \phi(\beta^*) \ln \phi(\beta), \text{ where } \sum_{i \in D} \phi(\beta) = 1.$$

This expression has a maximum at $\beta = \beta^*$ because

$$\begin{aligned} \frac{\partial}{\partial \beta} \left[\sum_{i \in D} \phi(\beta^*) \ln \phi(\beta) \right]_{\beta=\beta^*} &= \sum_{i \in D} \phi(\beta^*) \frac{1}{\phi(\beta^*)} \frac{\partial \phi(\beta)}{\partial \beta} \Big|_{\beta=\beta^*} \\ \frac{\partial}{\partial \beta} \left[\sum_{i \in D} \phi(\beta^*) \ln \phi(\beta) \right]_{\beta=\beta^*} &= \sum_{i \in D} \frac{\partial \phi(\beta)}{\partial \beta} \Big|_{\beta=\beta^*} = 0 \end{aligned}$$

where the last equality holds because

$$\sum_{i \in D} \phi(\beta) = 1.$$

Under normal regularity conditions, this maximum is unique and the maximum of Eq. (5-6) converges in probability to the maximum of the true likelihood. Therefore it yields consistent estimators of the model parameters (Newey and McFadden, 1986).

Eq. (5-6) can be simplified if the sampling protocol is such that the sampling correction $\ln \pi(D_n | i)$ is the same for all alternatives. Then, the correction term will cancel out in Eq. (5-6) and can be ignored. The effects of using other sampling protocols are studied by Manski and McFadden (1981), Ben-Akiva and Lerman (1985), Watanatada and Ben-Akiva (1979) and Frejinger et al. (2009).

Diverse applications of McFadden's results on sampling of alternatives for Logit models can be found in the literature. Some examples are Parsons and Kealy (1992) and Sermons and Koppelman (2001). In turn, the extension of McFadden's results to non-Logit models is a problem for which few little progress have been made in the last 30 years. Some advances have been done for choice-based samples; cases where the full choice-set is available to the researcher, but the observations are instead sampled depending on the choices. First, Manski and Lerman (1977) proposed a consistent but inefficient estimator for non-Logit models. This estimator was also used by Cosslett (1981) and by Imbens and Lancaster (1994). Later, Garrow et al. (2005) proposed an efficient estimator for a particular case of the Nested Logit model. Lastly, Bierlaire et al. (2008) proposed an alternative estimator that is applicable to MEV models with choice based samples and does not require knowledge of the sampling protocol.

Additionally, some analyses have been done regarding the impact of sampling of alternatives in Logit Mixture models. For example, McConnel and Tseng (2000), and Nerella and Bhat (2004), used Monte Carlo experimentation to study the problem of sampling of alternatives in random coefficients Logit models and found that sampling causes only small changes to parameter estimates. In turn, Chen et al. (2005) used Monte Carlo experimentation to show that, for Logit Mixture models that capture correlation among alternatives, the effects of sampling might be severe. Finally, Domanski (2009), citing an unpublished paper attributed to Haefen and Jacobsen, claims that the use of the expectation-maximization algorithm (Train, 2009) might result in the consistent estimation of model parameters while sampling of alternatives in random coefficients Logit Mixture model.

Regarding the problem of sampling of alternatives for the Nested Logit, several authors have directly applied McFadden's results for Logit without any modification. Examples of these type of applications include Berkovec and Rust (1985), Train et al. (1987), Hansen (1987), and Rivera and Tiglaio (2005). As it will be shown later, this approach may significantly impact the estimators of the model parameters. Finally, to the best of my knowledge, the only attempt to deal with the problem of sampling of alternatives in the Nested Logit model corresponds to the work of Lee and Wadell (2010). These authors use a method based on an idea originally suggested by Ben-Akiva (2009), which I further develop in the next section.

5.3 A Novel Method for MEV Models

In this section, I present a novel methodology to address the problem of sampling of alternatives and estimation for Multivariate Extreme Value (MEV) models, based on an idea originated by Ben-Akiva (2009).

The genesis of MEV models goes back to 1973, when Ben-Akiva proposed the Nested Logit model. Afterwards, McFadden (1978) showed that the Logit, the Nested Logit and other models belonged to a more general class of closed-form choice models that can handle diverse correlation structures among alternatives in the choice-set. McFadden originally denominated this class of models as Generalized Extreme Value

(GEV) models. Since the error terms for this class of models follow a MEV distribution, the models themselves are termed here as MEV.

The joint distribution of the error terms of the utilities in MEV models has the form

$$F(\boldsymbol{\varepsilon}_{1n}, \dots, \boldsymbol{\varepsilon}_{Jn}) = e^{-G(e^{-\varepsilon_{1n}}, \dots, e^{-\varepsilon_{Jn}}; \boldsymbol{\gamma})}, \quad (5-7)$$

where G is a generating function that is specific to each member of the MEV family, and $\boldsymbol{\gamma}$ is a set of distribution parameters. McFadden (1978) shows that if the generating function G complies with certain requirements the choice model implied by Eq. (5-7) will be consistent with the random utility maximization theory. Later, Ben-Akiva and Lerman (1985) show that the MEV choice probability can be written in a Logit form as shown in Eq. (5-8)

$$P_n(i) = \frac{e^{V(x_{in}, \beta) + \ln G_i(\langle e^{V_{in}} \rangle_{I \in C_n}; \boldsymbol{\gamma})}}{\sum_{j \in C_n} e^{V(x_{jn}, \beta) + \ln G_j(\langle e^{V_{jn}} \rangle_{I \in C_n}; \boldsymbol{\gamma})}}, \quad (5-8)$$

where $G_i(\langle e^{V_{in}} \rangle_{I \in C_n}; \boldsymbol{\gamma}) = \frac{\partial G(e^{V_{1n}}, \dots, e^{V_{Jn}}; \boldsymbol{\gamma})}{\partial e^{V_{in}}} \equiv G_{in}$.

Given the Logit form of the MEV model, it might look as if the problem of sampling of alternatives can be easily extended to MEV by following the same process of analysis deployed before for Logit, as shown in Eq. (5-3)-(5-5). That procedure results in the following expression for the conditional probability of choosing alternative i , given that set D_n was constructed:

$$\pi(i | D_n) = \frac{e^{V(x_{in}, \beta) + \ln G_i(\langle e^{V_{in}} \rangle_{I \in C_n}; \boldsymbol{\gamma}) + \ln \pi(D_n|i)}}{\sum_{j \in D_n} e^{V(x_{jn}, \beta) + \ln G_j(\langle e^{V_{jn}} \rangle_{I \in C_n}; \boldsymbol{\gamma}) + \ln \pi(D_n|j)}}.$$

Then, the same demonstration used by McFadden (1978) can be used to show that the maximization of the following quasi-log-likelihood function

$$QL_{MEV, D, C} = \sum_{n=1}^N \ln \pi(i | D_n) = \sum_{n=1}^N \ln \frac{e^{V(x_{in}, \beta) + \ln G_i(\langle e^{V_{in}} \rangle_{I \in C_n}; \boldsymbol{\gamma}) + \ln \pi(D_n|i)}}{\sum_{j \in D_n} e^{V(x_{jn}, \beta) + \ln G_j(\langle e^{V_{jn}} \rangle_{I \in C_n}; \boldsymbol{\gamma}) + \ln \pi(D_n|j)}} \quad (5-9)$$

leads to consistent estimators of the model parameters. However, it can be immediately noted that Eq. (5-9) is not practical. Even though the denominator of the choice probability depends only on D_n , the argument of the term $\ln G_{in}$ still depends on the full choice-set C_n . Ben-Akiva (2009) suggests that this problem might be solved if G_{in} is replaced by an estimator that depends only on the subset D_n .

In this chapter, I formalize the idea proposed by Ben-Akiva (2009), analyze the conditions required for its success, study the asymptotic properties of the estimators resulting from it, determine the correct expansion factors required in some relevant examples, and study the properties of the estimators using Monte Carlo experimentation and real data.

The results on consistency, asymptotic normality and efficiency can be summarized in the following theorem:

Theorem: Given N observations, a choice-set C_n of cardinality J_n , and a subset D_n of cardinality \tilde{J}_n . If

- a) $\pi(D_n | j) > 0 \quad \forall j \in D_n$ and $\pi(D_n | j) = 0 \quad \forall j \notin D_n$,
- b) the choice model is MEV and $G_{in} = \frac{\partial G(e^{V_{in}}, \dots, e^{V_{jn}}; \gamma)}{\partial e^{V_{in}}}$,
- c) $G_{in} = f(B_i(C_n))$ where f is continuous and twice-differentiable,
- d) $\hat{B}_i(D_n)$ is a consistent (in \tilde{J}_n) and unbiased estimator of $B_i(C_n)$, and
- e) $Var(\hat{B}_{in}) = K_n / \tilde{J}_n$ with K_n scalar;

then, the maximization of the quasi-log-likelihood function

$$QL_{MEV, D} = \sum_{n=1}^N \ln \hat{\pi}(i | D_n) = \sum_{n=1}^N \ln \frac{e^{V(x_n, \beta) + \ln f(\hat{B}_i(D_n)) + \ln \pi(D_n, i)}}{\sum_{j \in D_n} e^{V(x_n, \beta) + \ln f(\hat{B}_j(D_n)) + \ln \pi(D_n, j)}} \quad (5-10)$$

yields, under general regularity conditions, consistent estimators (in N) of the model parameters β^* , as \tilde{J}_n increases with N at any rate. If \tilde{J}_n grows faster than \sqrt{N} , the estimators of the model parameters will be consistent, asymptotically normal, and as efficient as the estimators obtained from the maximization of a quasi-log-likelihood

shown in Eq. (5-9). Finally, if J_n is finite and the protocol is sampling without replacement, \tilde{J}_n needs to increase only up to $\tilde{J}_n = J_n$ in order to achieve consistency and relative efficiency.

Proof. Given that \hat{B}_{in} is a consistent estimator of B_{in} , as \tilde{J}_n grows, the Slutsky theorem guarantees that $\ln f(\hat{B}_i(D_n))$ will also be a consistent estimator of $\ln G_{in}$, because the log and f are continuous. Equivalently, since $\pi(i|D_n)$ is continuous in $\ln G_{in}$, the Slutsky theorem guarantees that $\hat{\pi}(i|D_n)$ will be a consistent estimator of $\pi(i|D_n)$. Finally, McFadden's consistency results for Logit, shown in Eq. (5-6), guarantees that the maximization of the quasi-log-likelihood shown in Eq. (5-10) will result in the consistent estimation of the model parameters as N grows.

Note that the claim of McFadden's consistency result is established as N grows, but the consistency of \hat{B}_{in} , $\ln f(\hat{B}_i(D_n))$ and $\hat{\pi}(i|D_n)$ is established as \tilde{J}_n grows. To rely legitimately on the Slutsky theorem, it is indispensable to determine a concordance between \tilde{J}_n and N . This concordance can be established by analyzing the asymptotic properties of the estimators.

The asymptotic distribution of the estimators of the model parameters that result from the maximization of the quasi-log-likelihood shown in Eq (5-10) can be derived using the two-stage approach employed by Train (2009, section 10.5) to analyze the asymptotic properties of simulation-based estimators. In a first stage, I will analyze the asymptotic distribution of the sample average of the score, the gradient of the quasi-log-likelihood shown in Eq. (5-10). In a second stage I will use those results to derive the asymptotic distribution of the estimators of the model parameters.

Consider that the choice-sets C and D , of cardinalities J and \tilde{J} respectively, do not vary across observations, and that there is a single term $\ln G_n$ that needs to be approximated for each observation n . Then, instead of B_{in} , the term considered in this case should be B_n . These assumptions are not essential, and can be easily generalized, but

help in substantially reducing the notational burden. With the same purpose, I will refer to the whole set of model parameters β and μ , just as β .

Under this setting, I will term $\hat{g}(\beta)$ the sample average of the gradient of the quasi-log-likelihood evaluated using the estimator \hat{B}_n , as follows:

$$\hat{g}(\beta) = \frac{1}{N} \sum_{n=1}^N \hat{g}_n(\beta) = \frac{1}{N} \sum_{n=1}^N \frac{\partial \ln \hat{\pi}_n}{\partial \beta} = \frac{1}{N} \sum_{n=1}^N \frac{\partial}{\partial \beta} \ln \frac{e^{V(x_n, \beta) + \ln f(\hat{B}_n) + \ln \pi_n(D_i, \beta)}}{\sum_{j \in D} e^{V(x_j, \beta) + \ln f(\hat{B}_n) + \ln \pi_n(D_j, \beta)}}.$$

To study the asymptotic distribution of $\hat{g}(\beta)$ in the vicinity of the true values β^* , consider the following re-arrangement of terms

$$\hat{g}(\beta^*) = \underbrace{g(\beta^*)}_{A_1} + \underbrace{[E(\hat{g}(\beta^*)) - g(\beta^*)]}_{A_2} + \underbrace{[\hat{g}(\beta^*) - E(\hat{g}(\beta^*))]}_{A_3}.$$

The first term $A_1 = g(\beta^*)$ is the statistic that is being approximated by $\hat{g}(\beta^*)$, where

$$g(\beta) = \frac{1}{N} \sum_{n=1}^N \frac{\partial \ln \pi_n(\beta)}{\partial \beta} = \frac{1}{N} \sum_{n=1}^N \frac{\partial}{\partial \beta} \ln \frac{e^{V(x_n, \beta) + \ln G_n(C) + \ln \pi_n(D_i, \beta)}}{\sum_{j \in D} e^{V(x_j, \beta) + \ln G_n(C) + \ln \pi_n(D_j, \beta)}}.$$

The second term $A_2 = E(\hat{g}(\beta^*)) - g(\beta^*)$ corresponds to the bias of the estimator of $g(\beta^*)$. The third term A_3 corresponds to the noise of the approximation, which is the difference between a particular realization of $\hat{g}(\beta^*)$, and its expected value.

Consider the noise term A_3 , which can be rewritten as follows:

$$\begin{aligned} A_3 &= \hat{g}(\beta^*) - E(\hat{g}(\beta^*)) \\ A_3 &= \frac{1}{N} \sum_n [\hat{g}_n(\beta^*) - E(\hat{g}_n(\beta^*))] \\ A_3 &= \frac{1}{N} \sum_n d_n, \end{aligned}$$

where each d_n is the deviation of $\hat{g}(\beta^*)$ from its expectation for observation n . Note that each d_n depends on a particular draw of alternatives to construct the set D . This means that there is a distribution of values of d_n depending on all possible draws of alternatives in D . The distribution of d_n has zero mean because the expectation is subtracted in the creation of d_n . Also, note that the variance of d_n should decrease with the cardinality of D because $\hat{g}(\beta^*)$ should become closer to its expected value as \tilde{J} increases. To account

for this effect, the variance of d_n can be expressed as S_n/\tilde{J} , where S_n is the variance when $\tilde{J} = 1$. Then, relying on the generalized version of the central limit theorem (Train, 2009), the noise A_3 will have the following limiting distribution:

$$\sqrt{N}A_3 \xrightarrow{d} \text{Normal}(0, \mathbf{S}/\tilde{J}),$$

where \mathbf{S} is the population mean of S_n . Consequently, the asymptotic distribution of the noise A_3 will be

$$A_3 \overset{a}{\sim} \text{Normal}(0, \mathbf{S}/\tilde{J}N).$$

It is interesting to note what occurs with the noise A_3 when N increases but \tilde{J} is fixed. In this case, $\sqrt{N}A_3$ will have a limiting distribution, but will not vanish as N increases. In turn, the asymptotic variance of the noise A_3 will decrease as N increases, even if \tilde{J} is fixed. Note also that when the protocol is sampling without replacement and J is finite, \tilde{J} needs to increase only up to J , since from that point $\hat{g}(\beta) = E(\hat{g}(\beta)) = g(\beta)$.

Consider the bias term A_2 . This bias exists because the method described in Eq. (5-10) considers an unbiased estimator \hat{B}_n of B_n , but the calculation of $\hat{g}(\beta)$ involves a series of nonlinear transformations of \hat{B}_n . The bias can be studied by taking a second order Taylor's approximation of $\hat{g}(\beta)$ around $\hat{B}_n = B_n$. Noting that $\hat{g}_n(\beta, B_n) = g_n(\beta)$, it follows that

$$\hat{g}_n(\beta) = g_n(\beta) + \frac{\partial \hat{g}_n}{\partial \hat{B}_n} [\hat{B}_n(\beta) - B_n(\beta)] + \frac{1}{2} \frac{\partial^2 \hat{g}_n}{\partial \hat{B}_n^2} [\hat{B}_n(\beta) - B_n(\beta)]^2 + o_n.$$

Then, taking expectations (over possible realizations of the set D), recalling that \hat{B}_n is an unbiased estimator of B_n , and considering that the discrepancy o_n has zero mean, this Taylor's approximation can be rewritten as

$$E(\hat{g}_n(\beta)) - g_n(\beta) = \frac{1}{2} \frac{\partial^2 \hat{g}_n(\beta)}{\partial \hat{B}_n^2} \text{Var}(\hat{B}_n(\beta)).$$

Note that the $\text{Var}(\hat{B}_n(\beta))$ should decrease as \tilde{J} increases because then \hat{B}_n will become progressively closer to B_n . Assuming that this relationship can be captured by the expression $\text{Var}(\hat{B}_n(\beta)) = K_n/\tilde{J}$, where K_n is a scalar, the bias A_2 can be rewritten as

$$A_2 = E(\hat{g}(\beta)) - g(\beta) = \frac{1}{N} \sum_n E(\hat{g}_n(\beta)) - g_n(\beta)$$

$$A_2 = \frac{1}{N} \sum_n \frac{1}{2} \frac{\partial^2 \hat{g}_n(\beta)}{\partial \hat{B}_n^2} \frac{K_n}{\tilde{J}}$$

$$A_2 = \frac{Z}{\tilde{J}}$$

where Z is the sample average of $\frac{K_n}{2} \frac{\partial^2 \hat{g}_n}{\partial \hat{B}_n^2}$.

The bias A_2 will vanish as N increases, if and only if \tilde{J} increases also with N . Otherwise, $\hat{g}(\beta)$ will be an inconsistent estimator of $g(\beta)$. Instead, an even stronger assumption is required to achieve asymptotic normality. To understand why, consider the bias A_2 normalized for sample size N

$$\sqrt{N}A_2 = \frac{\sqrt{N}}{\tilde{J}} Z.$$

This term will vanish as N increases, if and only if \tilde{J} increases faster than \sqrt{N} . Otherwise, the estimator $\hat{g}(\beta)$ will have neither a limiting nor an asymptotic distribution.

Equivalent to what occurred with the noise A_3 , note that when the protocol is sampling without replacement and J is finite, \tilde{J} needs to increase only up to J , since from that point $E(\hat{g}(\beta)) = g(\beta)$ because any resorting of the alternatives in the choice-set C will have no impact on the choice probabilities.

In summary, it was shown that if \tilde{J} increases with N at any speed, $\hat{g}(\beta) \xrightarrow{p} g(\beta)$ and when \tilde{J} increases faster than \sqrt{N} , $\hat{g}(\beta)$ will be asymptotically Normal. Given that $\hat{g}(\beta) \xrightarrow{p} g(\beta)$, the limiting and asymptotic distributions of $\hat{g}(\beta)$ will be the same as those of $g(\beta)$.

To study the asymptotic properties of $g(\beta)$, label \mathbf{W} the population variance of $g_n(\beta^*)$. Then, assuming that $g(\beta)$ equals zero in the population, by the central limit theorem, the limiting distribution of $g(\beta)$ corresponds to

$$\sqrt{N}(g(\beta^*)-0) \xrightarrow{d} \text{Normal}(0, \mathbf{W}),$$

and the asymptotic distribution corresponds to

$$g(\beta^*) \stackrel{a}{\sim} \text{Normal}(0, \mathbf{W}/N).$$

It is then possible to combine the results for the components of $\hat{g}(\beta)$ in order to study the asymptotic distribution of the estimators $\hat{\beta}$ of the model parameters β . This can be achieved by taking a first-order Taylor's expansion of $\hat{g}(\hat{\beta})$ around the true values β^*

$$\hat{g}(\hat{\beta}) = \hat{g}(\beta^*) + \hat{R}[\hat{\beta} - \beta^*] + o_n,$$

where $\hat{R} = \partial \hat{g} / \partial \beta$ and the discrepancy o_n disappears asymptotically. Then, note that the estimators $\hat{\beta}$ of the model parameters β are defined by the condition $\hat{g}(\hat{\beta}) = 0$, because dividing Eq. (5-10) by N does not impact the solution of the problem. It follows that the limiting distribution of the estimators is

$$\sqrt{N}(\hat{\beta} - \beta^*) = \sqrt{N}(-\hat{R}^{-1})\hat{g}(\beta^*) = \sqrt{N}(-\hat{R}^{-1})(A_1 + A_2 + A_3). \quad (5-11)$$

As established before if \tilde{J} increases faster than \sqrt{N} the terms A_2 and A_3 will vanish. Under this condition, the term A_1 in Eq. (5-11) becomes asymptotically equal to $g(\beta)$, which has a limiting distribution of $\sqrt{N}(g(\beta^*)-0) \xrightarrow{d} \text{Normal}(0, \mathbf{W})$. Note that $\hat{R} \xrightarrow{p} \mathbf{R}$, where $\mathbf{R} = E(\hat{R})$. This implies that the limiting distribution of the estimators of the model parameters becomes

$$\sqrt{N}(\hat{\beta} - \beta^*) \xrightarrow{d} \text{Normal}(0, \mathbf{R}^{-1}\mathbf{W}\mathbf{R}^{-1}), \quad (5-12)$$

and their asymptotic distribution will be

$$\hat{\beta} \stackrel{a}{\sim} \text{Normal}(\beta^*, \mathbf{R}^{-1}\mathbf{W}\mathbf{R}^{-1}/N) = \text{Normal}(\beta^*, \mathbf{\Omega}/N), \quad (5-13)$$

where $\mathbf{\Omega} = \mathbf{R}^{-1}\mathbf{W}\mathbf{R}^{-1}$, $\mathbf{W} = \text{Var}\left(\frac{\partial \ln \pi_n(\beta^* | D)}{\partial \beta}\right)$ and $\mathbf{R} = E\left(\frac{\partial^2 \ln \pi_n(\beta^* | D)}{\partial \beta \partial \beta'}\right)$.

$\mathbf{\Omega}$ is usually defined as the “robust” or “sandwich” variance-covariance matrix of the estimators of the model parameters (Train, 2009). Berndt et al. (1974) proposed an estimator of $\mathbf{\Omega}$ that is known as the BHHH matrix and is used, for example, by the discrete-choice estimation software Biogeme (Bierlaire, 2003). To deploy the BHHH matrix for this case, note that \mathbf{R} is the Hessian of the model shown in Eq. (5-9). A consistent estimator of \mathbf{R} is its sample analog, which can be constructed from the Hessian of the quasi-log-likelihood shown in Eq. (5-10). Equivalently, the variance-covariance matrix of the score of the model shown in Eq. (5-10), evaluated at the estimated values $\hat{\mathbf{W}}(\hat{\boldsymbol{\beta}})$, is a consistent estimator of \mathbf{W} . Given that $\hat{\mathbf{g}}(\hat{\boldsymbol{\beta}}) = 0$, $\hat{\mathbf{W}}(\hat{\boldsymbol{\beta}})$ can be calculated as the outer product of the scores of the model shown in Eq. (5-10). In summary, the BHHH estimator for the variance-covariance matrix of the estimators of the model parameters resulting from the maximization of the quasi-log-likelihood function shown in Eq. (5-10), corresponds to the following expression:

$$\hat{\mathbf{\Omega}} = \left[\frac{\partial^2 \ln \hat{\pi}(\hat{\boldsymbol{\beta}} | D)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right]^{-1} \left[\sum_{n=1}^N \frac{\partial \ln \hat{\pi}_n(\hat{\boldsymbol{\beta}} | D)}{\partial \boldsymbol{\beta}} \frac{\partial \ln \hat{\pi}_n(\hat{\boldsymbol{\beta}} | D)}{\partial \boldsymbol{\beta}'} \right] \left[\frac{\partial^2 \ln \hat{\pi}(\hat{\boldsymbol{\beta}} | D)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right]^{-1}.$$

These results imply that the estimators obtained by the maximization of Eq. (5-10) will have the same variance-covariance matrix as the estimators that would be obtained by using Eq. (5-9); that is, if the full choice-set C is available for the calculation of the expansion of the term $\ln G_n$. Then, it can be affirmed that estimators obtained by maximizing Eq. (5-10) are efficient among all possible approximations of the model described in Eq. (5-9). **Q.E.D.**

It is interesting to note that the estimators obtained by maximizing Eq. (5-9) are not globally efficient because Eq. (5-9) is not the true log-likelihood and therefore the Crammer-Rao lower bound is not attained. This also implies that the estimators obtained by using McFadden’s (1978) method for Logit are also inefficient. McFadden (1978) did not study the asymptotic distribution of his estimators. However, following the same line of analysis deployed in this section, it can be shown that the asymptotic distribution of McFadden’s (1978) estimators will be equal to Eq. (5-13), using instead Eq. (5-6) to calculate the terms \mathbf{R} and \mathbf{W} .

Additionally, the fact that the estimators obtained with the method deployed in Eq. (5-10) will not be consistent unless \tilde{J} increases with N , implies that, in practice, we should test the stability of the estimators of the model parameters as a function of \tilde{J} . If the estimators for different values of \tilde{J} are statistically equal, we can be sure that the finite sample (of alternatives) bias is negligible. Otherwise, \tilde{J} should be increased until attaining stability. This is equivalent to the need for testing the stability of Logit Mixture's estimators as a function of the number of draws, in the simulated maximum-likelihood framework (Walker, 2001).

The practical implementation of the method to achieve consistency and asymptotic normality under sampling of alternatives in MEV models depends on the specific MEV model and the sampling protocol being considered. In the next two sections, I analyze this implementation in detail for the Nested and the Cross-Nested Logit models, respectively. Then, for illustrative purposes, in Section 5.6, I develop a Monte Carlo experiment where the performance of the method is analyzed under different circumstances. Finally, in Section 5.7, the methodology is applied to a Nested Logit of residential location choice that was estimated using real data from Lisbon, Portugal.

5.4 Formulation of the Method for Nested Logit

The Nested Logit model is a closed-form discrete choice model that allows for the correlation among random components of the utilities of alternatives that belong to mutually exclusive and totally exhaustive subsets (or nests) of the full choice-set. In this model, the marginal choice probabilities are written as the product of the conditional probability of choosing each alternative (given that the nest is chosen) and the marginal probability of choosing the nest. The utility of a nest is defined as the inclusive value or the expected maximum utility of choosing the alternatives that belong to that nest (Ben-Akiva and Lerman, 1985).

McFadden (1978) showed that the Nested Logit model can be alternatively formulated as a member of the MEV family. The generating function G for a Nested Logit model with M nests is

$$G\left(\left\langle e^{V_{in}} \right\rangle_{i \in C_n}; \gamma\right) = \sum_{m=1}^M \left(\sum_{i \in C_{m(i)n}} e^{\mu_m V_{in}} \right)^{\frac{\mu}{\mu_m}}, \quad (5-14)$$

where $m(i)$ is the nest to which i belongs, γ is the set of scales μ_m of the nests, and $C_{m(i)n}$ is the set of alternatives that belong to the nest $m(i)$. In this case, $\ln G_{in}$ corresponds to the expression shown in Eq. (5-15).

$$\ln G_{in} = \left(\frac{\mu}{\mu_{m(i)}} - 1 \right) \left(\ln \sum_{j \in C_{m(i)n}} e^{\mu_{m(i)} V_{jn}} \right) + \ln \mu + (\mu_{m(i)} - 1) V_{in} \quad (5-15)$$

Then, if a sample $D_{m(i)n}$, is drawn from the true choice-set $C_{m(i)n}$, the only term that would be affected (and therefore needs to be approximated) is the sum of the exponentials of the systematic utilities, the argument of the *logsum*. The sum of the exponentials will be denoted as

$$B_{in} = \sum_{j \in C_{m(i)n}} e^{\mu_{m(i)} V_{jn}}.$$

One way of approximating B_{in} is by constructing an expanded sum of the exponentials of the utilities of the alternatives in $D_{m(i)n}$. Then, the challenge would be to determine the expansion factors w_{jn} required to obtain an unbiased and consistent estimator of the sum of the exponentials.

To obtain an unbiased estimator, the expansion factors have to comply with the conditions shown in Eq. (5-16), where the first expectation is taken over all values of x , and the second expectation is taken over x and all potential sets $D_{m(i)n}$.

$$E(B_{in}) - E(\hat{B}_{in}) = 0 = E_x \left(\sum_{j \in C_{m(i)n}} e^{\mu_{m(i)} V_{jn}} \right) - E_{x,D} \left(\sum_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} V_{jn}} \right) \quad (5-16)$$

Note that each $e^{\mu_{m(i)} V_{jn}}$ can be seen as a random variable with mean $\eta_{m(i)n}$, the mean of the empirical distribution of $e^{\mu_{m(i)} V_{jn}}$. In this case the first component of Eq. (5-16) becomes

$$E(B_{in}) = E \left(\sum_{j \in C_{m(i)n}} e^{\mu_{m(i)} V_{jn}} \right) = J_{m(i)n} \eta_{m(i)n}.$$

The expansion factors w_{jn} required to obtain an unbiased estimator of B_{in} shall depend on the sampling protocol. For analytical purposes I will consider first that the protocol is sampling without replacement and then that it is sampling with replacement. Finally, I will show that the expansion factors w_{jn} required in both cases can be summarized in a single expression.

Consider first that the protocol is sampling **without** replacement by nest. Then, using the following indicator function

$$1_{j \in D_{m(i)n}} = \begin{cases} 1 & \text{if } j \in D_{m(i)n} \\ 0 & \text{o/w} \end{cases}$$

it is possible to rewrite $E(\hat{B}_{in})$ in Eq. (5-16) as follows:

$$E(\hat{B}_{in}) = E\left(\sum_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} Y_{jn}}\right) = E\left(\sum_{j \in C_{m(i)n}} 1_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} Y_{jn}}\right).$$

Then, by the Law of Total Expectations (also known as the Law of Iterated Expectations), which is equivalent to the total probability theorem used in Eq. (5-4),

$$\begin{aligned} E(\hat{B}_{in}) &= E\left(E\left(\sum_{j \in C_{m(i)n}} 1_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} Y_{jn}} \mid 1_{j \in D_{m(i)n}}\right)\right) \\ E(\hat{B}_{in}) &= E\left(\sum_{j \in C_{m(i)n}} 1_{j \in D_{m(i)n}} w_{jn} E\left(e^{\mu_{m(i)} Y_{jn}} \mid 1_{j \in D_{m(i)n}}\right)\right) = E\left(\sum_{j \in C_{m(i)n}} 1_{j \in D_{m(i)n}} w_{jn} \eta_{m(i)n}\right) \\ E(\hat{B}_{in}) &= \sum_{j \in C_{m(i)n}} E(1_{j \in D_{m(i)n}}) w_{jn} \eta_{m(i)n}, \end{aligned}$$

where $E(e^{\mu_{m(i)} Y_{jn}} \mid 1_{j \in D_{m(i)n}}) = \eta_{m(i)n}$ results from the fact that the distribution of $e^{\mu_{m(i)} Y_{jn}}$ determines the sampling of $D_{m(i)n}$, but the causality does not go in the other direction.

Given this result, one way for Eq. (5-16) to equal zero is by having

$$w_{jn} = 1/E(1_{j \in D_{m(i)n}}),$$

where $E(1_{j \in D_{m(i)n}})$ is the probability of drawing alternative j , because the protocol in this case is sampling without replacement.

Consider now that the protocol is sampling **with** replacement by nest. Then it is necessary to define the set $\tilde{D}_{m(i)n}$ and the indicator function \tilde{n}_{jn} . The former is a set that

includes all the repetitions of the alternatives sampled, and the latter corresponds to the number of times alternative j is repeated in the set $\tilde{D}_{m(i)n}$. Then \hat{B}_{in} can be rewritten as follows

$$\hat{B}_{in} = \sum_{j \in \tilde{D}_{m(i)n}} \tilde{w}_{jn} e^{\mu_{m(i)} Y_{jn}} = \sum_{j \in C_{m(i)n}} \tilde{n}_{jn} \tilde{w}_{jn} e^{\mu_{m(i)} Y_{jn}}, \text{ and therefore}$$

$$E(\hat{B}_{in}) = \sum_{j \in C_{m(i)n}} \tilde{w}_{jn} \eta_{m(i)n} E(\tilde{n}_{jn}) \text{ and then } \tilde{w}_{jn} = 1/E(\tilde{n}_{jn}).$$

Finally, since

$$\hat{B}_{in} = \sum_{j \in \tilde{D}_{m(i)n}} \tilde{w}_{jn} e^{\mu_{m(i)} Y_{jn}} = \sum_{j \in D_{m(i)n}} \tilde{n}_{jn} \tilde{w}_{jn} e^{\mu_{m(i)} Y_{jn}} = \sum_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} Y_{jn}},$$

the expansion factors required to obtain an unbiased estimation of the sum of the exponentials, for the case of sampling with replacement, are equal to

$$w_{jn} = \tilde{n}_{jn} / E(\tilde{n}_{jn}).$$

The expansion factors required when the protocol is with or without replacement can be summarized in a single expression by noting that, when the protocol is sampling without replacement, $\tilde{n}_{jn} = 1$ if j is in $D_{m(i)n}$, and $E\left(1_{j \in \tilde{D}_{m(i)n}}\right)$ is also the expected number of times alternative j would be drawn to form the set $D_{m(i)n}$. Then, the general expression for the expansion factors required to obtain an unbiased estimator of B_{in} can be denoted as shown in Eq. (5-17).

$$w_{jn} = \frac{\tilde{n}_{jn}}{E(\tilde{n}_{jn})} \quad (5-17)$$

The next step is to prove that the expansion factors shown in Eq. (5-17) will lead to consistent estimators of B_{in} as $\tilde{J}_{m(i)n}$ increases. This results directly from any weak Law of Large Numbers. Actually, consistency would be granted even if no expansion factors were considered at all. As $\tilde{J}_{m(i)n}$ grows, even an estimator of B_{in} that only considers the simple sum of the exponentials of the alternatives in $D_{m(i)n}$ will eventually be as near to B_{in} as desired, as $\tilde{J}_{m(i)n}$ increases. The difference with the expansion factors shown in Eq. (5-17) is that the speed of convergence will be much faster, leading to better finite sample

properties. In addition, having an unbiased estimator is what allows for the derivation of the results on efficiency and asymptotic normality.

5.5 Formulation of the Method for Cross-Nested Logit

The Cross-Nested Logit model is a closed-form discrete choice model that allows for correlation among the random components of the utilities of all alternatives in the choice-set. Similar to the Nested Logit, the Cross-Nested Logit considers a set of nests m . However, in the Cross-Nested Logit model the nests are totally exhaustive but not mutually exclusive in the coverage of the alternatives in the choice-set. The correlation structure is defined by a non-negative weight α_{jm} representing the degree of belonging of alternative j to the nest m . Examples of applications of the Cross-Nested Logit model and variations of it are the works of Small (1987), Vovsha (1997), Vovsha and Bekhor (1998), Bierlaire (2001), and Papola (2004).

The Cross-Nested Logit model can be formulated as a member of the MEV family. In general, with M nests, the generating function G that results in the Cross-Nested Logit model is

$$G\left(\langle e^{V_{in}} \rangle_{i \in C_n}; \gamma\right) = \sum_{m=1}^M \left(\sum_{j \in C_n} \alpha_{jm} e^{\mu_m V_{in}} \right)^{\frac{\mu}{\mu_m}},$$

where m are the nests, γ corresponds to the set of scales μ_m of the nests, and the weights $\alpha_{jm} \geq 0$. Then, $\ln G_{in}$ corresponds to the following expression:

$$\ln G_{in} = \ln \sum_{m=1}^M \left(\mu \alpha_{im} e^{V_i(\mu_m-1)} \left(\sum_{j \in C_n} \alpha_{jm} e^{\mu_m V_j} \right)^{\frac{\mu-\mu_m}{\mu_m}} \right).$$

Just as it occurred with the Nested Logit, if a sample D_n is drawn from the true choice-set C_n , the only term affected will be the sum of the exponentials, which is now weighed by the terms α_{jm} . Then, consistency, relative efficiency, and asymptotic normality can be achieved for the Cross-Nested Logit while sampling of alternatives, using the following estimator:

$$\hat{B}_{in} = \sum_{j \in D_n} w_{jn} \alpha_{jm} e^{\mu_m V_j} \approx B_{in} = \sum_{j \in C_n} \alpha_{jm} e^{\mu_m V_j}.$$

The same derivation used in Eq. (5-16)-(5-17) can be used to show that the expansion factors w_{jn} required in this case are also those shown in Eq. (5-17).

5.6 Monte Carlo Experiment

5.6.1 Model Setting

The following Monte Carlo experiment was performed to analyze and illustrate the properties of the proposed method in achieving consistency in the case of sampling of alternatives in MEV models. The setting of this experiment is summarized in Figure 5-1. The true or underlying model is a Nested Logit with 1,005 alternatives, among which the first 5 belong to one nest ($J_1 = 5$) and the other 1,000 to a second nest ($J_2 = 1,000$). The systematic utilities V_{in} depends upon two variables, x_1 and x_2 , which were constructed *iid* Uniform (-1,1) for the $N=2,000$ observations. The true coefficients of the model are $\mu = 1, \mu_1 = 2, \mu_2 = 3, \beta_{x_1} = \beta_{x_2} = 1$.

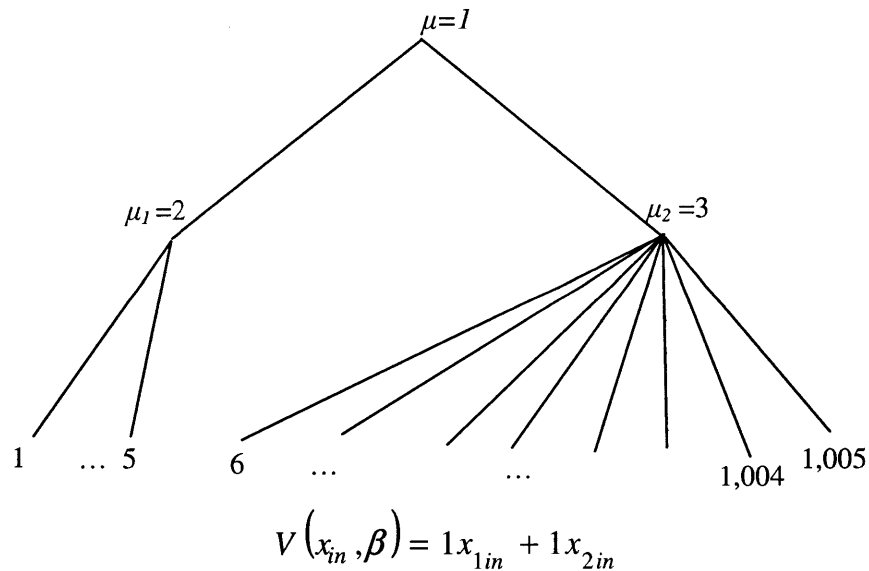


Figure 5-1 Monte Carlo Experiment: Nesting Structure. 1,005 Alternatives

$N=2,000 \quad J_1 = 5 \quad \tilde{J}_1 = 5; \quad J_2 = 1,000 \quad \tilde{J}_2 = 5 \text{ and } 500$

The methodology used to implement the Nested Logit model shown in Figure 5-1 for Monte Carlo experimentation differs from that used in the experiments deployed before. Previously, the chosen alternative for each observation was generated as that with the largest random utility. That methodology required the generation of the Extreme Value error term for each alternative and observation, a task that is easy to perform for the binary Logit. In turn, the generation of error terms from a 1,005-dimensional non-*iid* Multivariate Extreme Value distribution using Eq. (5-7) is much more complicated in terms of computational time and precision. Therefore, the approach used in this case is the following. First the choice probability was calculated replacing the true values of the parameters in Eq. (5-8); then, these choice probabilities were used to build a discrete cumulative density function by alternative; then, a random number Uniform (0,1) was generated for each observation; and finally, the chosen alternative was determined, from the random number, using the inverse of the cumulative density function.

The sampling protocol used to draw alternatives from the choice-set in this experiment was stratified importance sampling without replacement by nest. First, the chosen alternative for each observation was included. Then non-chosen alternatives were randomly sampled, without replacement by nest, to make a total of $\tilde{J}_1 = 5$ for the first nest, and $\tilde{J}_2 = 5$ and $\tilde{J}_3 = 500$ for the second nest.

Given this sampling protocol, the conditional probability of constructing a particular set D_n for observation n , given that alternative i was chosen, corresponds to

$$\pi_n(D | i) = \frac{\binom{J_{m(i)} - 1}{\tilde{J}_{m(i)} - 1}^{-1} \binom{J_{m' \neq m(i)}}{\tilde{J}_{m' \neq m(i)}}^{-1}}{\binom{J_{m(i)} - 1}{\tilde{J}_{m(i)} - 1}^{-1} \binom{J_{m' \neq m(i)}}{\tilde{J}_{m' \neq m(i)}}^{-1}},$$

where $m' \neq m(i)$ is the nest to which i does not belong and the expression on parenthesis correspond to the binomial coefficient.

It can be shown that

$$\frac{\binom{J_{m(i)} - 1}{\tilde{J}_{m(i)} - 1}}{\binom{J_{m(i)} - 1}{\tilde{J}_{m(i)} - 1}} = \frac{(J_{m(i)} - 1)!}{(\tilde{J}_{m(i)} - 1)! (J_{m(i)} - 1 - (\tilde{J}_{m(i)} - 1))!} = \frac{\tilde{J}_{m(i)}}{J_{m(i)}} \binom{J_{m(i)}}{\tilde{J}_{m(i)}},$$

and therefore, the conditional probability of constructing the set D_n , given that alternative i was chosen, corresponds to

$$\pi_n(D|i) = \frac{J_{m(i)}}{\tilde{J}_{m(i)}} \left[\begin{pmatrix} J_1 \\ \tilde{J}_1 \end{pmatrix}^{-1} \begin{pmatrix} J_2 \\ \tilde{J}_2 \end{pmatrix}^{-1} \right]. \quad (5-18)$$

Given that the second term in Eq. (5-18) does not vary across alternatives, it will cancel out when taking the log to calculate the sampling correction $\ln \pi(D_n|i)$. Then, the estimator of the conditional probability of choosing alternative i , given that the set D_n was constructed, will correspond to Eq. (5-19)

$$\hat{\pi}(i|D_n) = \frac{e^{V(x_{in}, \beta) + \ln f(\hat{B}_{in}(D_n)) + \ln \frac{J_{m(i)}}{\tilde{J}_{m(i)}}}}{\sum_{j \in D_n} e^{V(x_{jn}, \beta) + \ln f(\hat{B}_{jn}(D_n)) + \ln \frac{J_{m(j)}}{\tilde{J}_{m(j)}}}}, \quad (5-19)$$

$$\text{where } \ln f(\hat{B}_{in}(D_n)) = \left(\frac{\mu}{\mu_{m(i)}} - 1 \right) \left(\ln \sum_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} V_{jn}} \right) + \ln \mu + (\mu_{m(i)} - 1) V_{in}.$$

The final step corresponds to the specification of the expansion factors w_{jn} . This task is substantially different when the same set D_n used for the sampling correction is or is not used for the expansion of the sum of the exponentials.

Consider first that the set D_n is used also for the expansion of the sum of the exponentials. Then, given that the sampling protocol is without replacement, the numerator in Eq. (5-17) will equal 1. $E(\tilde{n}_{jn})$, the expected number of times alternative j might be sampled to construct the set D_n , remains to be calculated. Given that the protocol is without replacement, $E(\tilde{n}_{jn})$ corresponds to the probability of sampling alternative j .

$E(\tilde{n}_{jn})$ can be calculated using the Law of Total Expectations. The idea is to divide the space into mutually exclusive and totally exhaustive events with known probabilities of occurrence, and for which the conditional expectation of \tilde{n}_{jn} is also known. Consider the following events:

- A_1 : The chosen alternative is j
- A_2 : The chosen alternative is not j , but it is within those in the nest $m(j)$
- A_3 : The chosen alternative does not belong to the nest $m(j)$.

The events A_1 , A_2 and A_3 are totally exhaustive and mutually exclusive because only one alternative is chosen and the nests in the Nested Logit model are mutually exclusive and totally exhaustive. The probabilities of these three events depend on the choice probabilities:

$$\begin{aligned}
 P(A_1) &= P_n(j) && : \text{The probability of choosing alternative } j. \\
 P(A_2) &= \sum_{\substack{l \in C_{m(j)} \\ l \neq j}} P_n(l) && : \text{The probability of choosing other alternatives in } m(j), \text{ which} \\
 &&& \text{is equal to the sum of their choice probabilities.} \\
 P(A_3) &= 1 - \sum_{l \in C_{m(j)}} P_n(l) && : \text{The probability of choosing an alternative outside } m(j), \\
 &&& \text{which is equal to 1 minus the probability of the nest } m(j).
 \end{aligned}$$

The conditional expectations of \tilde{n}_{jn} given the events A_1 , A_2 and A_3 are also known:

$$\begin{aligned}
 E(\tilde{n}_{jn} | A_1) &= 1 && : \text{Because the chosen alternative is always sampled.} \\
 E(\tilde{n}_{jn} | A_2) &= \frac{\tilde{J}_{m(j)} - 1}{J_{m(j)} - 1} && : \text{Because if } j \text{ is not chosen, but the chosen alternative is in} \\
 &&& m(j), \text{ only } \tilde{J}_{m(j)} - 1 \text{ out of } J_{m(j)} - 1 \text{ alternatives remain to be} \\
 &&& \text{sampled from the nest } m(j). \\
 E(\tilde{n}_{jn} | A_3) &= \frac{\tilde{J}_{m(j)}}{J_{m(j)}} && : \text{Because if the chosen alternative is in not in } m(j), \tilde{J}_{m(j)} \text{ out} \\
 &&& \text{of } J_{m(j)} \text{ alternatives remain to be sampled from the nest } m(j).
 \end{aligned}$$

Then, by the Law of Total Expectations, the expected number of times alternative j might be drawn will correspond to

$$E(\tilde{n}_{jn}) = E(\tilde{n}_{jn} | A_1)P(A_1) + E(\tilde{n}_{jn} | A_2)P(A_2) + E(\tilde{n}_{jn} | A_3)P(A_3).$$

By replacing terms, Eq. (5-20) is finally obtained.

$$E(\tilde{n}_{jn}) = P_n(j) + \frac{\tilde{J}_{m(j)} - 1}{J_{m(j)} - 1} \sum_{\substack{l \in C_{m(j)} \\ l \neq j}} P_n(l) + \frac{\tilde{J}_{m(j)}}{J_{m(j)}} \left(1 - \sum_{l \in C_{m(j)}} P_n(l) \right) \quad (5-20)$$

The expression shown in Eq. (5-20) for the denominators of the expansion factors depends on the choice probabilities, which are unknown beforehand in an application with real data. In section 5.6.2, I analyze alternatives to achieve this goal in practice.

Consider now the case when a set D_n is used for the sampling correction $\ln \pi(D_n | i)$, and a different set \tilde{D}_n is drawn for the expansion of the sum of the exponentials. I term this alternative procedure re-sampling. In this case, the conditional probability of

choosing alternative i , given that the sets D_n and \tilde{D}_n were drawn, will correspond to Eq. (5-21).

$$\hat{\pi}(i | D_n, \tilde{D}_n) = \frac{e^{v(x_i, \beta) + \ln f(\hat{\beta}_m(\tilde{D}_n)) + \ln \frac{J_{m(i)}}{J_{m(i)}}}}{\sum_{j \in D_n} e^{v(x_j, \beta) + \ln f(\hat{\beta}_m(\tilde{D}_n)) + \ln \frac{J_{m(j)}}{J_{m(j)}}}} \quad (5-21)$$

As stated before, to formulate Eq. (5-3), the set D_n must include the chosen alternative. Otherwise, the quasi-log-likelihood of the model may become unbounded, making impossible the estimation of the model parameters. In turn, the set \tilde{D}_n used for the expansion of the sum of the exponentials in Eq. (5-21) does not need to include the chosen alternative, as long as D_n does it. This small difference is relevant because, if the sampling protocol used to build the set \tilde{D}_n does not require drawing the chosen alternative forcedly, there is no need for knowing the choice probabilities beforehand to calculate the expansion factors w_{jn} .

Then, the implementation of the expansion method in practice becomes considerably simpler. Consider for example that the sampling protocol used to build the set \tilde{D}_n was importance sampling without replacement by nest. Under this setting, the denominators of the expansion factors, the equivalent to Eq. (5-20), would simply be the ratio shown in Eq. (5-22), where $\tilde{J}_{m(j)}$ corresponds to the cardinality of \tilde{D}_n .

$$E(\tilde{n}_{jn}) = \frac{\tilde{J}_{m(j)}}{J_{m(j)}} \quad (5-22)$$

5.6.2 Assessment of the Methods with and without Re-sampling

Given this Monte Carlo experiment, the sampling protocol described and the expansion proposed, five models were estimated and the results are shown in Table 5-1. The first (*No Sampling* in Table 5-1) corresponds to the true model, where no sampling was applied. This model is estimated as a benchmark for the best possible estimators that could be expected for this particular experiment.

The second model (*Full InG_{in}* in Table 5-1) corresponds to the application of sampling of alternatives and the corresponding sampling correction, but using the full

choice-set to evaluate the term $\ln G_{in}$, as shown in Eq. (5-9). Even though this model is impractical because it requires knowledge of the full choice-set, it was estimated to show that Eq. (5-9) is correct, and to quantify and to differentiate the effects of sampling of alternatives, when having a reduced choice-set, from its effects in the approximation of $\ln G_{in}$.

Table 5-1 Monte Carlo Experiment: Sampling in MEV with and without Re-Sampling

| Experiments | | No Sampling | | Full $\ln G_{in}$ | | Unexpanded | | Expanded True Prob. | | Expanded Re-Sampling | |
|--|------------------|-------------|---------|-------------------|---------|------------|----------|---------------------|---------|----------------------|---------|
| | | est. | s.e | est. | s.e | est. | s.e | est. | s.e | est. | s.e |
| $\tilde{J}_1 = 5$ $\tilde{J}_2 = 5$ | β_{x_1} | 1.009 | 0.04681 | 0.9906 | 0.06112 | 2.570 | 0.1612 | 0.9102 | 0.06020 | 0.9301 | 0.06705 |
| | β_{x_2} | 1.062 | 0.04933 | 1.027 | 0.06253 | 2.630 | 0.1649 | 0.9276 | 0.06124 | 0.9558 | 0.06818 |
| | μ_1 | 2.055 | 0.2076 | 2.111 | 0.2289 | 0.2655 | 0.006477 | 2.211 | 0.2688 | 1.976 | 0.2913 |
| | μ_2 | 2.824 | 0.1125 | 2.881 | 0.1291 | 1.130 | 0.07562 | 3.313 | 0.1786 | 2.853 | 0.1567 |
| | $L(\hat{\beta})$ | -10,312.09 | | -1,942.70 | | -2,036.24 | | -1,968.59 | | -2,030.30 | |
| | $L(0)$ | -13,825.49 | | -4,605.17 | | -4,605.17 | | -4,605.17 | | -4,605.17 | |
| | $\bar{\rho}^2$ | 0.2544 | | 0.5790 | | 0.5587 | | 0.5734 | | 0.5583 | |
| $\tilde{J}_1 = 5$ $\tilde{J}_2 = 500$ | β_{x_1} | 1.009 | 0.04681 | 1.005 | 0.04678 | 0.7534 | 0.04708 | 1.005 | 0.04679 | 1.004 | 0.04679 |
| | β_{x_2} | 1.062 | 0.04933 | 1.055 | 0.04915 | 0.7913 | 0.04950 | 1.056 | 0.04918 | 1.055 | 0.04917 |
| | μ_1 | 2.055 | 0.2076 | 2.065 | 0.2088 | 2.730 | 0.3086 | 2.063 | 0.2088 | 2.065 | 0.2091 |
| | μ_2 | 2.824 | 0.1125 | 2.832 | 0.1130 | 3.785 | 0.2186 | 2.831 | 0.1131 | 2.834 | 0.1133 |
| | $L(\hat{\beta})$ | -10,312.09 | | -9,115.24 | | -9,117.40 | | -9,115.91 | | -9,115.37 | |
| | $L(0)$ | -13,825.49 | | -12,449.12 | | -12,449.12 | | -12,449.12 | | -12,449.12 | |
| | $\bar{\rho}^2$ | 0.2544 | | 0.2681 | | 0.2679 | | 0.2681 | | 0.2675 | |

$N=2,000$. $J_1 = 5$. $\tilde{J}_1 = \tilde{J}_1 = 5$; $J_2 = 1,000$ $\tilde{J}_2 = \tilde{J}_2 = 5$ and 500

The third model estimated (*Unexpanded* in Table 5-1) considers that a set D_n was sampled from the full choice-set C_n , that the corresponding sampling correction was applied, and that the same set D_n was used to construct the term $\ln G_{in}$, without any expansion term. This model acts as a benchmark because it corresponds to what has been used to date by the researchers to estimate Nested Logit models under sampling of alternatives (see, e.g., Berkovec and Rust, 1985; Train et al., 1987; Hansen, 1987; and Rivera and Tiglaio, 2005).

The fourth model estimated (*Expanded True Prob.* in Table 5-1) corresponds to the method proposed for cases where the same set D_n is used for the sampling correction and for the expansion of $\ln G_{in}$ using Eq. (5-20). The calculation of Eq. (5-20) involves knowledge of the choice probabilities, which are unknown beforehand in a real application. However, in this Monte Carlo experiment the true choice probabilities are available beforehand and are therefore used to show the performance of the method proposed for the expansion of the sum of the exponentials.

The last model estimated (*Expanded Re-sampling* in Table 5-1) corresponds to the method proposed for cases where a set D_n is used for the sampling correction, and a different set \tilde{D}_n , generated independently from the chosen alternative, is used for the expansion of $\ln G_{in}$ using Eq. (5-22). For fair comparison with other models, the number of alternatives considered in the set \tilde{D}_n is the same as that used for the set D_n ; that is, $\tilde{J}_n = J_n$.

The first result that should be noted in Table 5-1 is that, as expected, all estimated coefficients for the *No Sampling* and *Full* $\ln G_{in}$ models are statistically equal (with 95% confidence) to the true values. Regarding *Full* $\ln G_{in}$, note that, as the sample size increases, the standard error of the estimators is reduced as a result of the increment in the number of cases $N(\tilde{J} - 1)$. In other words, efficiency increased as more information became available.

Regarding the model *Unexpanded*, note that for $\tilde{J}_2 = 5$, the model estimates are very far from the true values. Remarkably, one of the scale coefficients is even below one, which makes this result inconsistent with utility maximization (Ben-Akiva and Lerman, 1985). The bias in this model is reduced substantially for $\tilde{J}_2 = 500$. This occurs because the *Unexpanded* formulation collapses to the true model as the sample size increases. However, even for the large \tilde{J}_2 , the estimators are still statistically different (with 95% confidence) from the true values.

In the case of the *Expanded True Prob.* method, all estimates in Table 5-1 are remarkably better than those of the *Unexpanded* model and statistically equal (with 95%

confidence) to the true ones with 95% confidence, even for $\tilde{J}_2 = 5$. For the bias, note that it is not negligible for $\tilde{J}_2 = 5$, but for $\tilde{J}_2 = 500$, it is significantly reduced.

Figure 5-2 shows the evolution of the estimators as \tilde{J}_2 is increased for the model *Expanded True Prob.* As \tilde{J}_2 approaches J_2 , the estimators of the model collapse to those of the *No Sampling* model. Remarkably, the estimators quickly stabilize for \tilde{J}_2 below 100 and are never far from the true values. As shown in Table 5-1, even for a sample size as small as $\tilde{J}_2 = 5$, all the estimators are statistically equal (with 95% confidence) to the true values.

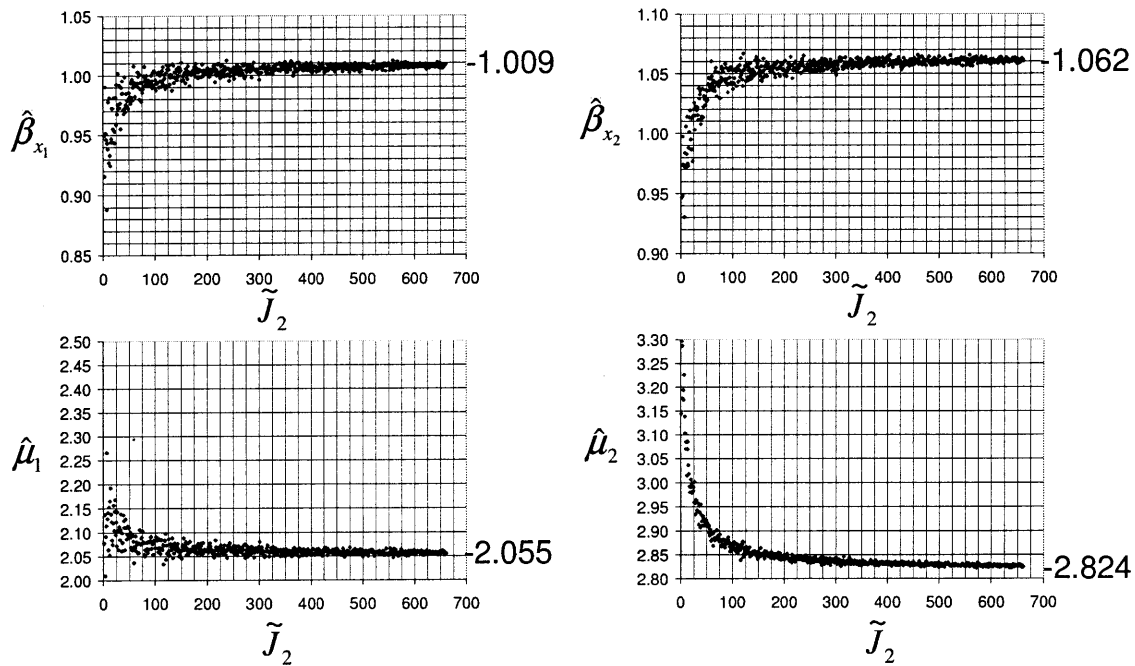


Figure 5-2 Monte Carlo Experiment: Estimators as \tilde{J}_2 Increases. *Expanded True Prob.*

Figure 5-2 is also useful for analyzing the small sample bias. First, note that the coefficient that has the poorer convergence behavior (larger variance and slope) is $\hat{\mu}_2$, the scale of the second nest. It can be hypothesized that this occurs because sampling is performed only from the second nest. Figure 5-2 also shows that both scales $\hat{\mu}_1$ and $\hat{\mu}_2$ are biased upward and the model coefficients $\hat{\beta}$ are biased downward. The experiments analyzed did not allow proposing hypotheses to explain this result. Further analysis of the

finite sample properties of this estimator, and potential ways to improve them, are left for future research.

Finally, the last column in Table 5-1 shows that the results for the *Expanded Re-Sampling* method are qualitatively equal to those of the *Expanded True Prob* method. This indicates that if re-sampling to perform the expansion is possible, it should be preferred because it avoids approximating the choice probabilities in order to perform the expansion. In the next section, I analyze the performance of different procedures that can be used in practice when re-sampling is not possible.

5.6.3 Expansion in Practice when Re-sampling is not Possible

When re-sampling is not possible the results of the method for sampling of alternatives in MEV shown in Table 5-1, require knowledge of the choice probabilities, which are not available in an application with real data. To avoid this problem, three methods used to approximate the choice probabilities are examined and the results are summarized in Table 5-2.

One alternative is to approximate the probability of the chosen alternative to equal 1, and the probability of the non-chosen alternatives to equal zero. This model is termed *Expanded All or Nothing* in Table 5-2. Replacing these assumptions in Eq. (5-20), the expansion factors used in this case will correspond to the following:

$$w_{jn} = 1 \quad \text{if } j \text{ is the chosen alternative}$$

$$w_{jn} = \frac{J_{m(j)} - 1}{\tilde{J}_{m(j)} - 1} \quad \text{if } j \text{ is not chosen, but another alternative in } m(j) \text{ is chosen}$$

$$w_{jn} = \frac{J_{m(j)}}{\tilde{J}_{m(j)}} \quad \text{if } j \text{ is not chosen, and no other alternative in } m(j) \text{ is chosen.}$$

The expansion factors that result in this case are equivalent to those used by Frejinger et al. (2009) to approximate the denominator of a Logit model with sampling of alternatives, and to those used by Lee and Waddell (2010) to expand a Nested Logit model under sampling of alternatives. That is, although it is not mentioned by those authors, they implicitly approximated the probability of the chosen alternative to 1, and the probability of the non-chosen alternatives to 0.

Table 5-2 Monte Carlo Experiment: Different Estimators of Choice Probabilities

| Experiments | | Expanded True Prob. | | Expanded All or Nothing | | Expanded Population Shares | | Expanded Iterative Prob. | |
|--|------------------|---------------------|---------|-------------------------|---------|----------------------------|---------|--------------------------|---------|
| | | est. | s.e | est. | s.e | est. | s.e | est. | s.e |
| $\tilde{J}_1 = 5$ $\tilde{J}_2 = 5$ | β_{x_1} | 0.9102 | 0.06020 | 0.7440 | 0.05335 | 1.133 | 0.06906 | 0.9444 | 0.06528 |
| | β_{x_2} | 0.9276 | 0.06124 | 0.7565 | 0.05417 | 1.158 | 0.07020 | 0.9630 | 0.06641 |
| | μ_1 | 2.211 | 0.2688 | 2.787 | 0.3327 | 1.685 | 0.2151 | 2.031 | 0.2734 |
| | μ_2 | 3.313 | 0.1786 | 4.328 | 0.2817 | 2.714 | 0.1251 | 3.210 | 0.1808 |
| | $L(\hat{\beta})$ | -1,968.59 | | -1,864.44 | | -1,982.65 | | -1,991.85 | |
| | $L(0)$ | -4,605.17 | | -4,605.17 | | -4,605.17 | | -4,605.17 | |
| | $\bar{\rho}^2$ | 0.5734 | | 0.5960 | | 0.5703 | | 0.568 | |
| $\tilde{J}_1 = 5$ $\tilde{J}_2 = 500$ | β_{x_1} | 1.005 | 0.04679 | 1.005 | 0.04673 | 1.007 | 0.04681 | 1.005 | 0.04679 |
| | β_{x_2} | 1.056 | 0.04918 | 1.055 | 0.04912 | 1.058 | 0.04920 | 1.056 | 0.04918 |
| | μ_1 | 2.063 | 0.2088 | 2.066 | 0.2088 | 2.059 | 0.2083 | 2.063 | 0.2088 |
| | μ_2 | 2.831 | 0.1131 | 2.833 | 0.1131 | 2.825 | 0.1125 | 2.831 | 0.1130 |
| | $L(\hat{\beta})$ | -9,115.91 | | -9,114.88 | | -9,115.92 | | -9,115.92 | |
| | $L(0)$ | -12,449.12 | | -12,449.12 | | -12,449.12 | | -12,449.12 | |
| | $\bar{\rho}^2$ | 0.2681 | | 0.2682 | | 0.2681 | | 0.2681 | |

$N=2,000$. $J_1 = 5$, $\tilde{J}_1 = 5$; $J_2 = 1,000$ $\tilde{J}_2 = 5$ and 500

A second possibility to approximate the choice probabilities needed for the calculation of the expansion factors is to use the population shares of each alternative. Although the true population shares are not available in a real application, good approximations of them are clearly plausible from different sources (Census data for spatial choice models or flow counts in route choice modeling). This method is termed *Expanded Population Shares* in Table 5-2. Replacing the population shares in Eq. (5-20), the expansion factors implied by this procedure are the following:

W_j = population share of alternative j

$$w_{jn} = \frac{1}{W_j + \frac{\tilde{J}_{m(j)} - 1}{J_{m(j)} - 1} \sum_{\substack{l \in C_{m(j)n} \\ l \neq j}} W_l + \frac{\tilde{J}_{m(j)}}{J_{m(j)}} \left(1 - \sum_{l \in C_{m(j)n}} W_l \right)} \quad \forall n = 1, \dots, N; \forall j \in C_n.$$

Finally, an iterative method can be proposed. This method starts with an estimation of the population shares of each alternative, and then estimates the choice probabilities for each observation, iteratively, until convergence. This method is termed *Expanded Iterative Prob.* in Table 5-2 and can be summarized as follows.

Step 0:

k=0

W_j = population share of alternative j

$$w_{jn}^k = \frac{1}{W_j + \frac{\tilde{J}_{m(j)} - 1}{J_{m(j)} - 1} \sum_{\substack{l \in C_{m(j)n} \\ l \neq j}} W_l + \left(1 - \sum_{l \in C_{m(j)n}} W_l\right) \frac{\tilde{J}_{m(j)}}{J_{m(j)}}} \quad \forall n = 1, \dots, N; j \in C_n$$

Step 1:

Estimate the model using w_{jn}^k to get $\hat{\beta}$ and $\hat{P}_n^k(j) = \frac{e^{\hat{v}(x_{jn}, \hat{\beta}) + \ln f(\hat{\beta}_{jn}(w^k))}}{\sum_{l \in D_n} w_{ln}^k e^{\hat{v}(x_{ln}, \hat{\beta}) + \ln f(\hat{\beta}_{jn}(w^k))}}$

Step 2:

$$w_{jn}^{k+1} = \frac{1}{\hat{P}_n^k(j) + \frac{\tilde{J}_{m(j)} - 1}{J_{m(j)} - 1} \sum_{\substack{l \in D_{m(j)n} \\ l \neq j}} w_{ln}^k \hat{P}_n^k(l) + \frac{\tilde{J}_{m(j)}}{J_{m(j)}} \left(1 - \sum_{l \in D_{m(j)n}} w_{ln}^k \hat{P}_n^k(l)\right)}$$

Step 3:

k=k+1

Go to step 1 until convergence.

Convergence can be stated in terms of the estimated parameters of the model, the expansion factors, or the choice probabilities. For the applications of the iterative procedure in this thesis, the following stopping criterion was used:

$$\max_{n,j} \left| \hat{P}_n^k(j) - \hat{P}_n^{k+1}(j) \right| \leq 1/(10J).$$

The three methods proposed to approximate the choice probability when re-sampling is not possible were used in the estimation of the problem of sampling of alternatives for the Nested Logit model described in Figure 5-1. Table 5-2 shows the results of the three methodologies, compared to the results obtained with the *Expanded True Prob.* method.

Consider the case of the *Expanded All or Nothing* and the *Expanded Population Shares* procedures. Table 5-2 shows that for $\tilde{J}_2 = 5$, the estimators of both methods are statistically different (with 95% confidence) to the true ones. Although, comparing these results with those of the *Unexpanded* method reported in Table 5-1, it should be noted that the new estimators have a smaller bias. For $\tilde{J}_2 = 500$, the *Expanded All or Nothing* and the *Expanded Population Shares* estimators are statistically equal (with 95% confidence) to those obtained by using the *Expanded True Prob.* method, and also statistically equal (with 95% confidence) to the true values.

Finally, for the *Expanded Iterative Prob.* method, Table 5-2 shows that for $\tilde{J}_2 = 5$ and $\tilde{J}_2 = 500$ the estimates are statistically equal (with 95% confidence) to those obtained using the *Expanded True Prob.* method, and also statistically equal (with 95% confidence) to the true values.

In conclusion, the Monte Carlo experiments showed that the sampling of alternatives causes a significant bias in the estimators of the model parameters when the choice model is Nested Logit. In addition, the proposed method for expanding the sum of the exponentials performed well, even for small sample sizes. In cases where it is possible to obtain an additional sample to expand the sum of the exponentials, the method proposed is easily applicable. When it is not possible to re-sample, the method requires knowledge of the choice probabilities in order to build the expansion factors. In this final case, an iterative procedure showed satisfactory results.

5.6.4 Additional Experiments

In this section, I present four additional experiments to illustrate the performance of the proposed method for addressing sampling of alternatives in MEV models, under different circumstances.

The first three experiments explore the effect of the distribution of the data. These experiments consider the same structure described in Figure 5-1. The only difference is that the distributions of attributes x_1 and x_2 vary across observations. Under this setting, the estimators of the model parameters were obtained for 30 repetitions using the *Expanded True Prob.* method and for different values of \tilde{J}_2 . Table 5-3 reports the bias,

mean squared error (MSE) and t-test against the true value of the scale of the second nest $\hat{\mu}_2$ for each experiment.

The first experiment is termed *Uniform Mixture*. For the first 1,000 observations, x_1 was drawn from an *iid* Uniform (-1,1) distribution and x_2 from an *iid* Uniform (-1.5,1.5) distribution. For the second half of the observations, x_1 was drawn from an *iid* Uniform (0,2) distribution and x_2 from an *iid* Uniform (-3,1) distribution. Table 5-3 shows that the sample size required to obtain an estimator of $\hat{\mu}_2$ statistically equal (with 95% confidence) to its true value is larger than 50 alternatives in this case. This value is larger than that obtained for the experiment reported in Table 5-1 and shows that the threshold required for attaining valid estimates of the model parameters depends on the data.

Table 5-3 Monte Carlo Experiment: Additional Experiments on Sampling in MEV

| $\hat{\mu}_2$ | Uniform Mixture | | | Varying \tilde{J}_2 | | | Normal Uniform | | |
|---------------|-----------------|---------|-------------|-----------------------|---------|-------------|----------------|---------|-------------|
| \tilde{J}_2 | Bias | MSE | t-test true | Bias | MSE | t-test true | Bias | MSE | t-test true |
| 10 | 0.6251 | 0.4341 | 3.004 | 0.5293 | 0.3146 | 2.850 | 1.305 | 1.831 | 3.660 |
| 25 | 0.5137 | 0.2932 | 3.005 | 0.3355 | 0.1378 | 2.113 | 0.9991 | 1.068 | 3.772 |
| 50 | 0.3031 | 0.1127 | 2.100 | 0.2141 | 0.06596 | 1.509 | 0.7401 | 0.5873 | 3.719 |
| 100 | 0.1709 | 0.04645 | 1.302 | 0.1355 | 0.03640 | 1.008 | 0.5010 | 0.2813 | 2.874 |
| 250 | 0.09168 | 0.02408 | 0.7324 | 0.07410 | 0.02166 | 0.5828 | 0.2557 | 0.08757 | 1.716 |
| 500 | 0.03459 | 0.01532 | 0.2911 | 0.05311 | 0.01906 | 0.4167 | 0.1092 | 0.02940 | 0.8265 |

$N=2,000$. $J_1 = 5$. $J_2 = 1,000$ $\tilde{J}_1 = 5$; Average and variance from 30 repetitions. *Expanded True Prob.*

The second experiment is termed *Varying \tilde{J}_2* . This experiment considers the same structure and distribution of the data used in the *Uniform Mixture* experiment. The only difference is that the number of drawn alternatives varies across individuals following a Discrete Uniform distribution with limits

$$\left[\left[\tilde{J}_2/2 \right], \left[2\tilde{J}_2 \right] \right].$$

Then, for example, for $\tilde{J}_2=10$ in Table 5-3, the number of alternatives considered for each of the 2,000 observations can be any integer between 5 and 20, with equal probability. The results of this experiment are shown in the second column of Table 5-3. Although this experiment is not directly comparable with the *Uniform Mixture* setting, it can be affirmed that the fact that, in both cases, sample sizes around 50 were large

enough to obtain an estimator of the scale of the second nest that was statistically equal (with 95% confidence) to its true value, is evidence that varying the sample size across observations causes only minor impacts in the estimation procedure.

The third experiment is termed *Normal Uniform*. In this case x_1 is *iid* Normal (0,1) for the first 1,000 observations and Normal (1,2) for the rest. In turn x_2 *iid* Uniform (1,3) for the first 1,000 observations and Uniform (0,4) for the rest. The results of this experiment are shown in the third column of Table 5-3. This experiment shows that the sample size required to attain estimators that are statistically equal (with 95% confidence) to the true values is now between 100 and 250. This result is further evidence that the performance of the method can be significantly affected by the distribution of the data.

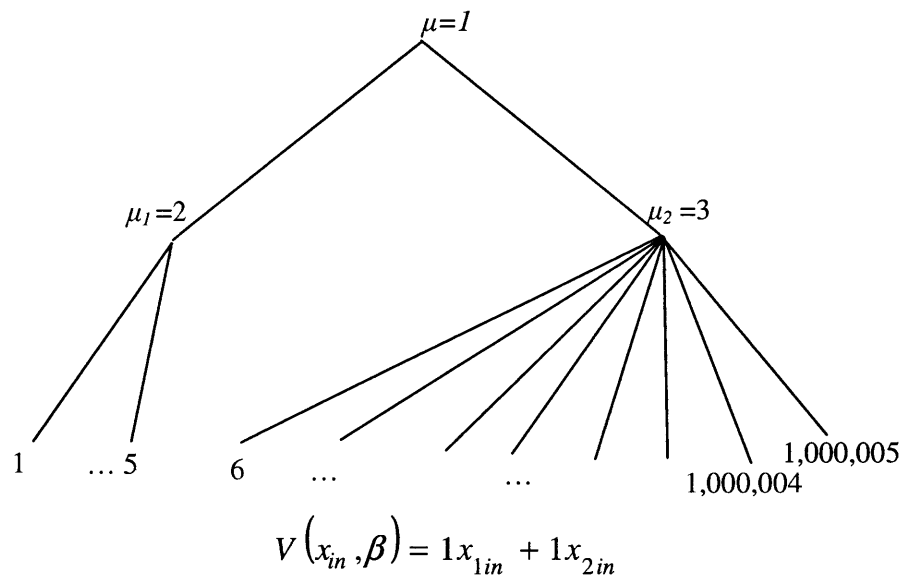


Figure 5-3 Monte Carlo Experiment: Nesting Structure. 1,00,005 Alternatives

The fourth experiment sheds light on whether or not the sample size required to attain a desirable bias can be stated as a percentage of the cardinality of the true choice-set. The experiment described in Figure 5-1 was modified only regarding the number of alternatives in the second nest, which is 1,000,000 in this case. The distribution of x_1 and x_2 are again *iid* Uniform (-1,1) for the $N=2,000$ observations. The model is described in Figure 5-3 and the results are reported in Table 5-4.

Table 5-4 Monte Carlo Experiment: Sampling in MEV. 1,000,005 Alternatives

| Experiments | | True Values | Unexpanded | | Expanded True Prob. | |
|--|------------------|-------------|------------|----------|---------------------|---------|
| | | | Est. | s.e | est. | s.e |
| $\tilde{J}_1 = 5$ $\tilde{J}_2 = 5$ | β_{x_1} | 1 | 2.947 | 0.3594 | 0.9403 | 0.07193 |
| | β_{x_2} | 1 | 2.820 | 0.3412 | 0.9118 | 0.06894 |
| | μ_1 | 2 | 0.1427 | 0.003651 | 1.877 | 0.5237 |
| | μ_2 | 3 | 1.073 | 0.1322 | 3.372 | 0.2203 |
| | $L(\hat{\beta})$ | | -1,348.82 | | -1,341.87 | |
| | $L(0)$ | | -3,670.51 | | -3,670.51 | |
| | $\bar{\rho}^2$ | | 0.6336 | | 0.6355 | |
| $\tilde{J}_1 = 5$ $\tilde{J}_2 = 500$ | β_{x_1} | 1 | 1.887 | 0.4404 | 1.014 | 0.05930 |
| | β_{x_2} | 1 | 1.784 | 0.4162 | 0.9629 | 0.05586 |
| | μ_1 | 2 | 0.1896 | 0.02426 | 1.836 | 0.455 |
| | μ_2 | 3 | 1.645 | 0.3837 | 3.054 | 0.162 |
| | $L(\hat{\beta})$ | | -9,253.96 | | -9,241.78 | |
| | $L(0)$ | | -12,710.46 | | -12,710.46 | |
| | $\bar{\rho}^2$ | | 0.2723 | | 0.2732 | |

$N=2,000$. $J_1 = 5$, $\tilde{J}_1 = \tilde{J}_1 = 5$; $J_2 = 1,000,000$ $\tilde{J}_2 = \tilde{J}_2 = 5$ and 500

In this case the true model is not estimatable with commercial software because the computational costs are too high. In turn, it is possible to simulate the choices by each observation, and then to sample a small number of alternatives from the true choice-set for subsequent estimation. Using this sampling procedure, samples of 5 and 500 alternatives were drawn from the second nest.

Table 5-4 contrasts the estimators that are obtained using the *Unexpanded* and *Expanded True Prob.* methods. Similar to what occurred in the experiments reported in Table 5-1, the estimators of the *Expanded True Prob.* method are also statistically equal (with 95% confidence) to their true values, even for a sample size as small as 5. However, comparing Table 5-1 with Table 5-4, it can be noted that the confidence is smaller in the case where the true choice-set has 1,000,005 alternatives.

Given that the quality of the estimators obtained with samples of 5 and 500 are qualitatively equal when the cardinality of the true choice-set is 1,005 or 1,000,005, it can

be affirmed that there is evidence that the sample size required to obtain acceptable estimators is independent of the true cardinality of the choice-set.

In summary, these additional experiments gave evidence that the sample size required to obtain good estimators while sampling alternatives in MEV models depends on the distribution of the data available and cannot be expressed as a percentage of the cardinality of the true choice-set. In general, an appropriate strategy to determine if the size of the sample of alternatives is large enough might be to test the stability of the estimators with different number of alternatives sampled.

5.7 Application to Real Data

The final step corresponds to the demonstration of the method proposed for sampling of alternatives and estimation in MEV models using real data. I revisited the residential location choice model for Lisbon, which was estimated in previous chapters as a Logit model. In this case, I considered a Nested Logit model, allowing for correlation between alternatives on a geographic base. The structure used is shown in Figure 5-4. I considered one nest for the 3,483 alternatives that belong to the Municipalities of Odivelas and Amadora, and the other 8,018 alternatives from the Municipality of Lisbon, were considered to belong to the root of this Nested Logit model.

The nesting structure used is simple principally because of the small number of observations available. More interesting structures, such as multilevel nests by Freguesia and municipalities, were impossible to estimate. However, the nesting structure considered does serve well its main purpose of demonstrating the methodology for sampling of alternatives and estimation developed in this chapter. Despite its simplicity, the nesting structure is concordant with what is observed in the city. The municipalities of Odivelas and Amadora are approximately what Rayle (2008) defined (using a factor-analysis approach) as the “Inner Periphery” of the central LMA, a sector that has marked differences with the Lisbon’s Municipality.

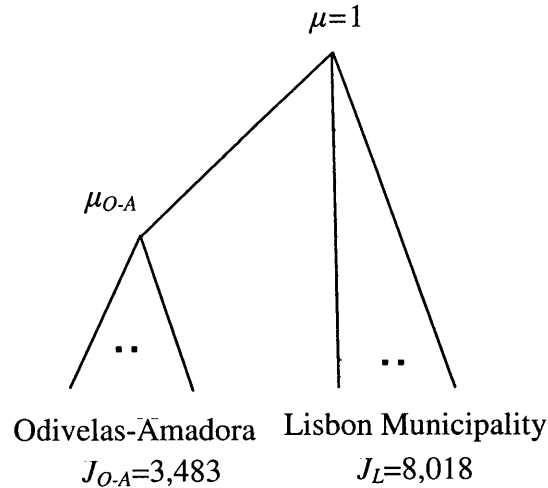


Figure 5-4 Lisbon’s Nested Logit Model: Nesting Structure

Under this setting, a Nested Logit was estimated with the assumption that the 11,501 alternatives corresponded to the true choice-set. This model considered the correction for endogeneity caused by the omission of attributes using the 2SCF function method. The results of this model are reported in the second column of Table 5-5 and are repeated in Table 5-6. Note that the estimators of the parameters of this Nested Logit model have the same sign and tend to be upward scaled, when compared with those obtained for the Logit model reported in Table 2-6. The main difference is in the estimator of the scale of the nest, which is statistically different (with 95% confidence) from 1 and therefore causes an important change in the elasticities of the model.

To demonstrate the method studied in this chapter, I performed two experiments where I sampled a set of alternatives in the choice-set and then re-estimated the model with and without the expansion of the sum of the exponentials proposed in this chapter. The sampling protocol used in the first experiment was the following. First, the chosen alternative was included. Then, alternatives were randomly drawn from the Odivelas-Amadora nest and from the root (Lisbon) up to make a total of 5 alternatives for each case.

The results of the model estimated using this sampling protocol, are shown in Table 5-5. In the third column are reported the estimators of the *Unexpanded* model where the

sampling correction was applied but the sum of the exponentials of the Odivelas-Amadora nest was calculated using only the 5 alternatives sampled. Note that several estimators are statistically different (with 95% confidence) from those of the original model. Remarkably, the estimator of the scale of the Odivelas-Amadora nest is highly positively biased. This means that the use of the *Unexpanded* model for simulation will cause an important overestimation of the substitution among dwellings in the Odivelas-Amadora nest. The fourth column of Table 5-5 reports the estimators of the model estimated using the *Expanded Iterative Prob.* method, where the sampling correction is applied and the sum of the exponentials is expanded using the iterative procedure described in Section 5.6.3. Equivalent to what occurred in the Monte Carlo experiments, the estimators are remarkably similar to those of the model without sampling and statistically equal (with 95% confidence) to them.

Table 5-5 Lisbon's Nested Logit Model: Sampling 5 + 5 Alternatives

| Variables | No Sampling | | Unexpanded | | Expanded Iterative Prob. | |
|--|---------------|---------|---------------|--------|--------------------------|--------|
| | $\hat{\beta}$ | s.e | $\hat{\beta}$ | s.e | $\hat{\beta}$ | s.e |
| 1. Dwelling price (in 100,000 €) | -4.393 | 0.7058 | -3.095 | 0.6498 | -5.374 | 0.8947 |
| 2. Dwelling price * 1[Income > 2,000 €/M] | 1.213 | 0.5769 | 1.291 | 0.4756 | 1.834 | 0.7048 |
| 3. Dwelling price * 1[Income > 5,000 €/M] | 0.9463 | 0.5284 | 0.5298 | 0.5364 | 0.7604 | 0.6779 |
| 4. Distance to Workplace (in Km) | -0.1774 | 0.0538 | -0.1617 | 0.0528 | -0.1732 | 0.0639 |
| 5. Log [Dwelling Area (in m ²)] | 4.217 | 0.7854 | 2.220 | 0.5530 | 4.454 | 1.0324 |
| 6. Log [Dwelling Age (in years) +1] | -0.6381 | 0.1158 | -0.4850 | 0.1180 | -0.7252 | 0.1604 |
| 7. $\hat{\delta}$ Control-function Aux. Var. | 1.987 | 0.4711 | 0.6193 | 0.3864 | 2.145 | 0.5763 |
| 8. μ_{O-A} Odivela-Amadora Nest | 1.329 | 0.09414 | 5.480 | 3.053 | 1.392 | 0.1266 |
| Log likelihood at Convergence $L(\hat{\beta}, \hat{\mu})$ | -547.89 | | -94.96 | | -93.53 | |
| Log likelihood at Zero $L(\hat{\beta} = 0, \hat{\mu} = 1)$ | -589.06 | | -134.13 | | -134.13 | |
| Adjusted ρ^2 | 0.08518 | | 0.3666 | | 0.3623 | |
| Sample Size N | 63 | | 63 | | 63 | |
| Choice-set Size J | 11,501 | | 10 | | 10 | |
| Estimation Time [seconds] | 363.0 | | 1.080 | | 10.65 | |

Nest Amadora and Odivelas. Root Lisbon municipality. Models include sampling correction. Models corrected for endogeneity with 2SCF. Sample 5 alts. from Odivelas-Amadora nest and 5 from Lisbon municipality. €/M: Euros per month.

The second experiment corresponded to the application of the same sampling protocol as before, but with alternatives that were sampled up to make a total of 500 for

the Odivelas-Amadora nest and 500 for the root (Lisbon). The results of the models estimated using this sampling protocol are shown in Table 5-6. Equivalent to what occurred with the Monte Carlo experiments, the estimators of the *Unexpanded* and of the *Expanded Iterative Prob.* models are similar to those of the model without sampling when \tilde{J} is large. All estimators are statistically equal (with 95% confidence) in both cases. The only significant difference is that the bias of the estimator of the scale of the Odivelas-Amadora's nest is smaller for the *Expanded Iterative Prob.* model.

Table 5-6 Lisbon's Nested Logit Model: Sampling 500 + 500 Alternatives

| Variables | No Sampling | | Unexpanded | | Expanded Iterative Prob. | |
|--|---------------|---------|---------------|---------|--------------------------|---------|
| | $\hat{\beta}$ | s.e | $\hat{\beta}$ | s.e | $\hat{\beta}$ | s.e |
| 1. Dwelling price (in 100,000 €) | -4.393 | 0.7058 | -4.349 | 0.6780 | -4.347 | 0.7054 |
| 2. Dwelling price * 1[Income > 2,000 €/M] | 1.213 | 0.5769 | 1.242 | 0.5649 | 1.184 | 0.5776 |
| 3. Dwelling price * 1[Income > 5,000 €/M] | 0.9463 | 0.5284 | 0.9566 | 0.5290 | 0.9923 | 0.5333 |
| 4. Distance to Workplace (in Km) | -0.1774 | 0.0538 | -0.1766 | 0.05288 | -0.1811 | 0.05380 |
| 5. Log [Dwelling Area (in m ²)] | 4.217 | 0.7854 | 4.177 | 0.7450 | 4.223 | 0.7902 |
| 6. Log [Dwelling Age (in years) +1] | -0.6381 | 0.1158 | -0.6362 | 0.1123 | -0.6321 | 0.1161 |
| 7. $\hat{\delta}$ Control-function Aux. Var. | 1.987 | 0.4711 | 1.908 | 0.4460 | 1.937 | 0.4683 |
| 8. μ_{O-A} Odivela-Amadora Nest | 1.329 | 0.09414 | 1.510 | 0.1618 | 1.326 | 0.09340 |
| Log likelihood at Convergence $L(\hat{\beta}, \hat{\mu})$ | -547.89 | | -382.38 | | 382.95 | |
| Log likelihood at Zero $L(\hat{\beta} = 0, \hat{\mu} = 1)$ | -589.06 | | -424.25 | | 424.25 | |
| Adjusted ρ^2 | 0.08518 | | 0.1223 | | 0.1162 | |
| Sample Size N | 63 | | 63 | | 63 | |
| Choice-set Size J | 11,501 | | 1,000 | | 1,000 | |
| Estimation Time [seconds] | 363.0 | | 55.27 | | 220.8 | |

Nest Amadora and Odivelas. Root Lisbon municipality. Models include sampling correction. Models corrected for endogeneity with 2SCF. Sample 500 alts. from Odivelas-Amadora nest and 500 from Lisbon municipality. €/M: Euros per month.

Finally, Table 5-5 and Table 5-6 report also the computational time used in the estimation of the different models. In the case where only 10 alternatives were sampled, the differences in computational costs were huge. The true model that considers the full choice-set of 11,501 alternatives took approximately 350 times more seconds to be estimated than the *Unexpanded* model, and approximately 35 times more than the *Expanded Iterative Prob.* method. The differences are reduced to 7 and 1.7 times respectively, when 1,000 alternatives are sampled. These differences in estimation time,

together with the evidence gathered from the Monte Carlo experiment with one million alternatives, reflect the significant gains that can be obtained with sampling. The methodological developments of this chapter will allow taking benefit of these gains in the implementation of spatial choice models with more realistic error structures.

5.8 Conclusion

Sampling of alternatives for non-Logit models is a problem that has been open for over 30 years, and that have hindered the development of suitable spatial choice models. This chapter proposes a novel method to address this issue for MEV models and illustrates its properties by means of a Monte Carlo experiment applied to the Nested Logit model, and a case study based on real data on residential location choice from Lisbon, Portugal.

The first interesting result is that the estimation in MEV models, under sampling of alternatives when the full $\ln G_{in}$ is considered, recovers true parameters, even if only a small number of alternatives is sampled. Second, the experiments show that when $\ln G_{in}$ is approximated by the proposed methodology, the results are always better than those obtained when ignoring the fact that only a subset of the true choice-set is available, and that the latter method performs poorly for small sample sizes.

When it is not possible to re-sample alternatives independently from the chosen one, in order to approximate the term $\ln G_{in}$, the proposed method involves knowledge of the choice probabilities. To avoid this inconvenience, three procedures were analyzed, among which the iterative procedure performed the best, and work reasonably well, even for small samples.

Finally, it should be noted that the proposed method is biased for a fixed sample size, and that the bias could be significant for a small \tilde{J} . This problem can be addressed by testing the stability of the estimators to different values of \tilde{J} . Future investigation regarding the small sample of alternatives bias shall involve the development methods to control or to quantify this bias.

Chapter 6

Conclusion

6.1 Summary

The purpose of this doctoral dissertation was to address endogeneity and sampling of alternatives in non-Logit models, two critical model estimation weaknesses that have been neglected in spatial choice models and have a significant impact in the development of suitable models of urban systems.

For endogeneity, I investigated diverse estimation and forecasting drawbacks that were debated or neglected in previous literature. First, I showed that the change of scale resulting from the use of the control-function method to correct for endogeneity does not impact the relevant properties of the model. Second, I studied the approach required for forecasting with models corrected for endogeneity, and devised a novel procedure to forecast with synthetic populations. Third, I studied the link between the latent-variable approach and the control-function method to correct for endogeneity, and developed a tractable maximum-likelihood estimator that achieves consistency, efficiency, and asymptotic normality, and allows for the direct calculation of the standard errors of the estimators of the model parameters. Finally, I identified a criterion to build instrumental variables to address price endogeneity in models of residential location choice, and tested its validity using two novel tests of over-identifying restrictions for discrete choice

models. These novel tests showed better power properties than the existing Amemiya-Lee-Newey test in a set of binary Logit Monte Carlo experiments.

For sampling of alternatives in non-Logit models, I studied the problem of obtaining consistent estimators when the underlying model belongs to the Multivariate Extreme Value class, a family of models that includes the Logit and other models that allow for more realistic substitution patterns among alternatives, such as the Nested Logit and the Cross-Nested Logit. For this problem, I implemented a method to achieve consistency, relative efficiency and asymptotic normality, building on an idea originated by Ben-Akiva (2009). I studied the performance of the method using both Monte Carlo experimentation and real data, and showed that it functioned remarkably well, even for small sample sizes.

6.2 Overall Conclusion

The main conclusion of this research is that the estimation and simulation of spatial choice models are significantly affected by the inevitable omission of attributes and by the need for sampling of alternatives. These issues can and should be addressed using the methods surveyed and developed in this research. Empirical evidence from Monte Carlo experimentation and real data was provided to show the impact of these drawbacks in models estimators, and to demonstrate how they may influence policy analysis. Empirical evidence also showed that the proposed methods for addressing these modeling drawbacks were successful and feasible with commercial software and generally available data.

6.3 Methodological Recommendations

Diverse methodological recommendations for future modeling efforts are derived from this research.

For endogeneity, the first recommendation is to test for it using any test for omitted attributes applied to the auxiliary variable used in the second stage of the 2SCF method. The second recommendation is a criterion for the construction of instruments to correct for endogeneity in residential location choice models. It was shown that prices of similar

dwelling within a certain vicinity made valid instruments in this framework. It is recommended that the dwellings used to construct the instruments should be selected among those outside a certain threshold (for example, 500 mts.) in order to avoid reflection bias, and within a certain boundary (for example, 5,000 mts. and differing less than 40% in area and age) to ensure their relevance.

The third recommendation regarding endogeneity is that, given its relative simplicity, the use of the Likelihood-ratio version of the Direct test for the validity of instruments is recommended. In addition, it was shown that the power of the tests for the validity of instruments might be severely affected if the instruments are highly correlated (above 0.95). This highlights the importance of avoiding the practice of generating instruments as nonlinear transformations of existing instruments in order to achieve over-identification.

The fourth recommendation regarding endogeneity is to use, when possible, the tractable maximum-likelihood estimator derived in Chapter 3. This estimator achieves consistency, efficiency and asymptotic normality in the correction for endogeneity, and permits the direct calculation of the standard errors of the model from the inverse of the Fisher-information-matrix. Otherwise, the two-stage estimator can be used to achieve efficiency (under some mild assumptions), but the calculation of the standard errors should be addressed using bootstrap or the delta-method.

The fifth recommendation regarding endogeneity is that, in cases where there is endogeneity and some indicator that theoretically depends on the omitted attributes is available, the use of the joint framework derived in Chapter 3 is recommended for modeling the latent-variable and control-function methods. This combined method would result in an increase of efficiency of the estimates, and in a more realistic representation of the behavior of the agents. The cost is that it might be necessary to evaluate a multifold integral in the number of alternatives, a procedure that may be impractical in spatial choice models.

For sampling of alternatives in MEV models it is recommended the use of the method that involves re-sampling (independent of the chosen alternative) in the generation of the expansion required to address this modeling drawback. When re-sampling is not possible, the iterative procedure described in Chapter 5 is preferred, as it shows substantially better

performance compared to other alternatives. Lastly, small sample bias should be addressed by testing the stability of the estimates attained with this method, as a function of the cardinality of the set used for the expansion.

6.4 Extensions

The developments and the analysis performed in this thesis have diverse limitations that future research shall address in different ways.

Regarding the methodologies developed to address endogeneity, it would be interesting to evaluate how the market clearing process in the housing market may affect the assumptions used in the implementation of the control-function method. It would also be interesting to explore the power properties of the tests for the validity of instruments under other circumstances, including diverse choice models, and real databases. The empirical study of the link between the control-function and the latent-variable methods should be useful in the assessment of the practical value of this joint approach. Another line of research in this area corresponds to the analysis of problems where the endogenous variable is discrete and not continuous as it was considered throughout this thesis. Finally, it would also be interesting to develop a systematic investigation of the problem of weak instruments in discrete choice models, extending existing research for linear models.

Regarding the method developed to address sampling of alternatives in MEV models, it would be interesting to explore the feasibility of controlling for the bias that is inevitably present in finite samples. Another interesting line of research is the extension of the approach used for solving the problem of sampling of alternatives in MEV models, into other non-Logit models, such as the Logit Mixture.

Additionally, it would be interesting to investigate the feasibility of applying the methods surveyed and developed in this thesis into larger databases and other spatial choice models, such as job and firm location, route choice or activity scheduling.

Finally, it would be interesting to assess the full impact of the methodological advances of this research in policy analysis. This might be achieved by applying these advancements in the framework of an operational microscopic integrated urban model

such as UrbanSim (Waddell et al., 2008). In particular, it would be interesting to investigate whether the integrated nature of the system will amplify or mitigate the effect of the corrections for endogeneity and sampling of alternatives in MEV models.

References

- Almeida, A., M. Ben-Akiva, F. Pereira, A. Ghauche, C. Guevara, S. Niza and C. Zegras (2009), "A Framework for Integrated Modeling of Urban Systems," Presented at the 45th ISOCARP Conference, Porto, Portugal.
- Amemiya, T. (1983), "A Comparison of the Amemiya GLS and the Lee-Maddala-Trost GZSLS in a Simultaneous-equations Tobit Model," **Journal of Econometrics**, 23, 295-300.
- Amemiya, T. (1979), "The Estimation of a Simultaneous Equation Tobit Model," **International Economic Review**, 20, 169-181.
- Amemiya, T. (1978), "The Estimation of a Simultaneous Equation Generalized Probit Model," **Econometrica**, 46, 1193-1205.
- Antoniou, C. (2008), **Online Calibration for Dynamic Traffic Assignment - Theory, Methods and Application**, VDM Verlag Dr. Müller Publishers, Saarbrücken, Germany.
- Badoe, D. and E. Miller (2000), "Transportation and Land-use Interaction: Empirical Findings in North America, and their Implications for Modeling," **Transportation Research Part D**, 5, 235-263.
- Basman, R. (1960), "On Finite Sample Distributions of Generalized Classical Linear Identifiability Test Statistics," **Journal of the American Statistical Association**, 5, 650-659.
- Bayer, P., R. McMillan and K. Rueben (2004), "Residential Segregation in General Equilibrium," Working Paper 885, Economic Growth Center, Yale University.
- Ben-Akiva, M. (2010), "Planning and Action in a Model of Choice," in Hess and Daly eds. **Choice Modelling: The State-of-the-Art and the State-of-Practice**, Emerald Publishing, Bingley, UK, 19-34.
- Ben-Akiva, M. (2009), "Sampling of Alternatives in Non-Logit Models," Unpublished Manuscript, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.

- Ben-Akiva, M. (1973), "Structure of Passenger Travel Demand Models," Ph.D. Thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.
- Ben-Akiva, M. and S. Lerman (1985), **Discrete Choice Analysis, Theory and Application to Travel Demand**, MIT Press, Cambridge, MA.
- Berkovec, J. and J. Rust (1985), "A Nested Logit Model of Automobile Holdings for One Vehicle Households," **Transportation Research B**, 19(4), 275-285.
- Berndt, E., H. Hall, R. Hall and J. Hausman (1974), "Estimation and Inference in Nonlinear Structural Models," **Annals of Economic and Social Measurement**, 3/4, 653-665.
- Berry, S., J. Levinsohn and A. Pakes (1995), "Automobile Prices in Market Equilibrium," **Econometrica**, 63(4), 841-90.
- Bertsekas, D. and J. Tsitsiklis (2002), **Introduction to Probability**, Athena Scientific Press, Belmont, MA.
- Bhat, C. and J. Guo (2004), "A Mixed Spatially Correlated Logit Model: Formulation and Application to Residential Choice Modeling," **Transportation Research Part B**, 38(2), 147-168.
- Bierlaire, M. (2003), "BIOGEME: A Free Package for the Estimation of Discrete Choice Models," **Proceedings of the 3rd Swiss Transportation Research Conference**, Ascona, Switzerland.
- Bielaire, M. (2001), "A Theoretical Analysis of the Cross-Nested Logit Model," **Annals of Operations Research**, 144(1), 287-300.
- Bierlaire, M., D. Bolduc and D. McFadden (2008), "The Estimation of Generalized Extreme Value Models from Choice-Based Samples," **Transportation Research Part B: Methodological**, 42(4), 381-394.
- Bowman, J. (1998), "The Day Activity Schedule Approach to Travel Demand Analysis," Ph.D. Thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.
- Card, D. (1995), "Using Geographic Variation in the College Proximity to Estimate the Return to Schooling," in Christophides, Garnt and Swidinsky eds. **Aspects of Labour Market Behaviour: Essays in Honor of John Vanderkamp**, University of Toronto Press, Toronto, Canada, 201-222.

- Chen, Y, L. Duann and W. Hu (2005), "The Estimation of Discrete Choice Models with Large Choice-Set," **Journal of the Eastern Asia Society for Transportation Studies**, 6, 1724-1739.
- Chen, J., J. Newman and M. Bierlaire (2009), "Modeling Route Choice Behavior from Smartphone GPS Data," presented at the 12th International Conference on Travel Behavior Research, Jaipur, India.
- Chintagunta, P., D. Jain and N. Vilcassim (1991), "Investigating Heterogeneity in Brand Preferences in Logit Models for Panel Data," **Journal of Marketing research**, 28, 417-528.
- Cox, D. and E. Snell (1968), "A General Definition of Residuals," **Journal of the Royal Statistical Society B**, 30, 248-275.
- Cramer, J. (2007), "Robustness of Logit Analysis: Unobserved Heterogeneity and Mis-specified Disturbances," **Oxford Bulletin of Economics and Statistics**, 69(4), 545-555.
- Cosslett, S. (1981) "Maximum Likelihood Estimator for Choice-Based Samples," **Econometrica**, 49(5), 1289-1316.
- Dahlberg, M., E. Mörk and P. Tovmo (2008), "Power Properties of the Sargan Test in the Presence of Measurement Errors in Dynamic Panels," **Applied Economics Letters, Taylor and Francis Journals**, 15(5), 349-353.
- Daly, A. (2008), "Elasticity, Model Scale and Error," presented at the European Transport Conference, Leeuwenhorst, The Netherlands.
- De Blander, R. (2008), "Which Null Hypothesis Do Overidentification Restrictions Actually Test?" **Economics Bulletin**, 3(76), 1-9.
- Domanski, A. (2009), "Estimating Mixed Logit Recreation Demand Models with Large Choice Sets," presented at the Agricultural and Applied Economics Association Annual Meeting, Milwaukee, WI.
- Duda, R., P. Hart and D. Stork (2001), **Pattern Classification**, Second Edition, John Wiley publishers, NY.
- Engle, R. (1984), "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics," in Griliches and Intrilligator eds. **Handbook of Econometrics**, II, Amsterdam, North Holland.
- Evans, W. and R. Schwab (1995), "Finishing High-school and Starting College: Do Catholic Schools Make a Difference?" **Quarterly Journal of Economics**, 110, 941-974.

- Ferreira, F. (2010), "You Can Take It with You: Proposition 13 Tax Benefits, Residential Mobility, and Willingness to Pay for Housing Amenities," **Journal of Public Economics**, 94 (9-10), 661-673.
- Frejinger, E., M. Bierlaire and M. Ben-Akiva (2009), "Sampling of Alternatives for Route Choice Modeling," **Transportation Research Part B: Methodological**, 43(10), 984-994.
- Garrow, L., F. Koppelman and L. Nelson (2005), "Efficient Estimation of Nested Logit Models using Choice-Based Samples," in Mahmassani ed. **Transportation and Traffic Theory Flow, Dynamics and Interactions: Proceedings of the 16th International Symposium on Transportation and Traffic Theory**, Oxford, UK, 525-544.
- Gopinath, D. (1995), "Modeling Heterogeneity in Discrete Choice Processes: Application to Travel Demand," Ph.D. Thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.
- Greene, W. (2003), **Econometric Analysis**, 5th Edition, Prentice Hall, New York.
- Greene, W. and D. Hensher (1991), "A Latent Class Model for Discrete Choice Analysis: Contrasts with Mixed Logit," **Transportation Research Part B: Methodological**, 37(8), 681-698.
- Guevara, C. (2005), "Addressing Endogeneity in Residential Location Models," M.Sc. Thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.
- Guevara, C. and M. Ben-Akiva (2010), "Addressing Endogeneity in Discrete Choice Models: Assessing Control-function and Latent-Variable Methods" in Hess and Daly eds. **Choice Modelling: The State-of-the-Art and the State-of-Practice**, Emerald Publishing, Bingley, UK, 353-370.
- Guevara, C. and M. Ben-Akiva (2008), "A Lagrange Multiplier Test for the Validity of Instruments in MNL Models: An Application to Residential Choice," presented at the European Transport Conference, Leeuwenhorst, The Netherlands.
- Guevara, C. and M. Ben-Akiva (2006), "Endogeneity in Residential Location Choice Models," **Transportation Research Record**, 1977, 60-66.
- Hahn, J. and J. Hausman (2002), "Weak Instruments: Diagnosis and Cures in Empirical Econometrics," **American Economic Review**, 93, 118-128.

- Hansen, E. (1987), "Industrial Location Choice in Sao Paulo Brazil, A Nested Logit Model," **Regional Science and Urban Economics**, 17, 89-108.
- 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.
- Imbens, G. and T. Lancaster (1994), "Combining Micro and Macro Data in Microeconomic Models," **Review of Economic Studies**, 61(4), 655-80.
- Kamakura, W. and J. Russell (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structuring," **Journal of Marketing Research**, 25, 379-390.
- Karaca-Mandic, P. and K. Train (2003), "Standard Error Correction in Two-stage Estimation with Nested Samples," **Econometrics Journal**, 6(2), 401-407.
- Kitamura R., E. Pas, C. Lula, T. Lawton and P. Benson (1996), "The Sequenced Activity Mobility Simulator (SAMS): An Integrated Approach to Modeling Transportation, Land Use and Air Quality," **Transportation**, 23(3), 267-291.
- Lee, L. (1992), "Amemiya's Generalized Least Squares and Tests of Overidentification in Simultaneous Equation Models with Qualitative or Limited Dependent Variables," **Econometric Reviews**, 11(3), 319-328.
- Lee, L. (1982), "Specification Error in Multinomial Logit Models," **Journal of Econometrics**, 20, 197-209.
- Lee, L. (1981), "Simultaneous Equations Models with Discrete and Censored Variables," in Manski and McFadden eds. **Structural Analysis of Discrete Data with Econometric Applications**, MIT Press, Cambridge, MA, 346-364.
- Lee, B.H. and P. Waddell (2010), "Residential Mobility and Location Choice: a Nested Logit Model with Sampling of Alternatives," **Transportation**, 37(4), 587-601.
- Lee, B.J., A. Fujiwara, J. Zhang and Y. Sugie (2003), "Analysis of Mode Choice Behaviors Based on Latent Class Models," presented at the 10th International Conference on Travel Behavior Research, Lucerne, Switzerland.
- Levine, J. (1998), "Rethinking Accessibility and Jobs-Housing Balance," **Journal of American Planning Association**, 64, 133-149.

- Manski, C. (1993), "Identification of Endogenous Social Effects: The Reflection Problem," **Review of Economic Studies**, 60(3), 531-42.
- Manski, C. and S. Lerman (1977), "The Estimation of Choice Probabilities from Choice Based Samples," **Econometrica**, 45(8), 1977-1988.
- Manski, C. and D. McFadden (1981), "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in Manski and McFadden eds. **Structural Analysis of Discrete Data with Econometric Applications**, MIT Press, Cambridge, MA, 2-50.
- Martinez, L. and J. Viegas (2009), "Effects of Transportation Accessibility on Residential Property Values: A Hedonic Price Model in the Lisbon Metropolitan Area," **Transportation Research Record**, 2115, 127-137.
- Martinez, L., J. Abreu and J. Viegas (2010), "Assessment of Residential Location Satisfaction in Lisbon Metropolitan Area," presented at the 89th Transportation Research Board Annual Meeting, Washington, DC.
- McConnel, K. and W. Tseng (2000), "Some Preliminary Evidence on Sampling of Alternatives with the Random Parameters Logit," **Marine Resource Economics**, 14, 317-332.
- McFadden, D. (1978), "Modeling the Choice of Residential Location," in Karlquist, Lundqvist, Snickers and Weibull eds. **Spatial Interaction Theory and Residential Location**, North Holland, Amsterdam, 75-96.
- McFadden, D. (1987), "Regression Based Specification Tests for the Multinomial Logit Model," **Journal of Econometrics**, 34, 63-82.
- Miller, E., J. Hunt, J. Abraham and P. Salvini (2004), "Microsimulating Urban Systems," **Computers, Environments and Urban Systems**, 28, 9-44.
- Nerella, S. and C. Bhat (2004), "A Numerical Analysis of the Effect of Sampling of Alternatives in Discrete Choice Models," **Transportation Research Record**, 1894, 11-19.
- Newey, W. (1987), "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," **Journal of Econometrics**, 36, 231-250.
- Newey, W. (1985a), "Semiparametric Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," **Annales de L'insee**, 59/60, 219-237.
- Newey, W. (1985b), "Generalized Method of Moments Specification Testing," **Journal of Econometrics**, 29, 229-256.

- Newey, W. and D. McFadden (1986), "Large Sample Estimation and Hypothesis Testing," in Engle and McFadden eds. **Handbook of Econometrics**, 4(36), 2111-2245.
- Nichols, A. (2007), "Causal Inference with Observational Data," **Stata Journal**, 7 (4), 507–541.
- Park, S. and S. Gupta (2009), "A Simulated Maximum Likelihood Estimator for the Random Coefficient Logit Model Using Aggregate Data," **Journal of Marketing Research**, 46(4), 531-542.
- Parsons, G. and M. Kealy (1992), "Randomly Drawn Opportunity Sets in a Random Utility Model of Lake Recreation," **Land Economics**, 68(1), 93-106.
- Petrin, A. and K. Train (2002), "Omitted Product Attributes in Discrete Choice Models," Working Paper, Department of Economics, University of California, Berkeley, CA.
- Papola, A. (2004), "Some Developments on the Cross-nested Logit Model," **Transportation Research Part B**, 38, 833–851.
- Quigley, J. (1976), "Housing Demand in the Short Run: An Analysis of Polytomous Choice," **Explorations in Economic Research**, 3, 76–102.
- R Development Core Team (2008), **R: A Language and Environment for Statistical Computing**, R Foundation for Statistical Computing, Vienna, Austria, URL <http://www.R-project.org>.
- Rao, C. and S. Mitra (1971), **Generalized Inverse of a Matrix and its Applications**, J. Wiley, New York, NY.
- Rayle, L. (2008), "Initial Insights from Spatial Analysis of the Lisbon Metropolitan Area," Working Paper, Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA.
- Rivera, M. and N. Tiglao (2005), "Modeling Residential Location Choice, Workplace Location Choice and Mode Choice of Two-Worker Households in Metro Manila," **Proceedings of the Eastern Asia Society for Transportation Studies**, 5, 1167 – 1178.
- Rivers, D. and Q. Vuong (1988), "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," **Journal of Econometrics**, 39, 347-366.
- Ruud, P. (1983), "Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Models," **Econometrica**, 51, 225–228.

- Sargan, J. (1958), "The Estimation of Economic Relationships Using Instrumental Variables," **Econometrica**, 26, 393-415.
- Sekhon, J. (2010), "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The matching Package for R," **Journal of Statistical Software**, forthcoming.
- Sermons, M. and F. Koppelman (2001), "Representing the Differences between Female and Male Commute Behavior in Residential Location Choice Models," **Geography**, 9, 101-110.
- Small, K. (1987), "A Discrete Choice Model for Ordered Alternatives," **Econometrica**, 55(2), 409-424.
- Stock, J. (2001), "Instrumental Variables in Statistics and Econometrics," in Smelser and Baltes eds. **International Encyclopaedia of the Behavioural Sciences**, Elsevier Publishing, New York, 7577-7582.
- Stock, J., J. Wright and M. Yogo (2002), "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," **American Statistical Association Journal of Business and Economic Statistics**, 20(4), 518-529.
- Strauch, D., R. Moeckel, M. Wegener, J. Gräfe, H. Muhlhans, G. Rindsfuser and K. Beckmann (2005), "Linking Transport and Land Use Planning: The Microscopic Dynamic Simulation Model ILUMASS," in Atkinson, Foody, Darby and Wu eds. **Geodynamics**, CRC Press, Boca Raton, Florida, 295-311.
- Tobler, W. (1970), "A Computer Movie Simulating Urban Growth in the Detroit Region," **Economic Geography**, 46(2), 234-240.
- Train, K. (2009), **Discrete Choice Methods with Simulation, 2nd Edition**, Cambridge University Press, New York, NY.
- Train, K., D. McFadden and M. Ben-Akiva (1987), "The Demand for Local Telephone Service: A fully Discrete Model of Residential Calling Patterns and Service Choice," **Rand Journal of Economics**, 18, 109-123.
- Train, K., M. Ben-Akiva and T. Atherton (1989), "Consumption Patterns and Self-selecting Tariffs," **The Review of Economics and Statistics**, 71(1), 62-73.
- Train, K. and C. Winston (2007), "Vehicle Choice Behavior and the Declining Market Share of US Automakers," **International Economic Review**, 48(4), 1469-1496.

- Villas-Boas, A. and R. Winer (1999), "Endogeneity in Brand Choice Models," **Management Science**, 45, 1324–1338.
- Vovsha, P. (1999), "Comparative Analysis of Different Spatial Interaction Models," presented at the 78th Annual Meeting of the Transportation Research Board, Washington, DC.
- Vovsha, P. and S. Bekhor (1998), "The Link-Nested Logit Model of Route Choice: Overcoming the Route Overlapping Problem," **Transportation Research Record**, 1645, 133–142.
- Waddell, P. (1992), "A Multinomial Logit Model of Race and Urban Structure," **Urban Geography**, 13, 127–141.
- Waddell, P., L. Wang and X. Liu (2008), "UrbanSim: An Evolving Planning Support System for Evolving Communities," in Brail eds. **Planning Support Systems for Cities and Regions**, Lincoln Institute for Land Policy, Cambridge, MA, 103-138.
- Walker, J. (2001), "Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables," Ph.D. Thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.
- Walker, J. and M. Ben-Akiva (2002), "Generalized Random Utility Model," **Mathematical Social Sciences**, 43(3), 303-343.
- Walker, J., E. Ehlers, I. Banerjee and E. Dugundji (2010), "Correcting for Endogeneity in Behavioral Choice Models with Social Influence Variables," Working Paper, University of California, Berkeley, CA.
- Walker, J. and J. Lee (2007), "Latent Lifestyle Preferences and Household Location Decisions," **Journal of Geographical Systems**, 9(1), 77-101.
- Watanatada, T. and M. Ben-Akiva (1979), "Forecasting Urban Travel Demand for Quick Policy Analysis with Disaggregate Choice Models: a Monte Carlo Simulation Approach," **Transportation Research**, 13(A), 241–248.
- White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," **Econometrica**, 50, 1–25.
- Wooldridge, J. (2002), **Econometric Analysis of Cross-Section and Panel Data**, MIT Press, Cambridge, MA.
- Yatchew, A. and Z. Griliches (1985), "Specification Error in Probit Models," **The Review of Economics and Statistics**, 67, 134–139.