MIT Portugal

Transportation Systems

Working Paper Series

Addressing Endogeneity in Discrete Choice Models: Assessing Control-Function and Latent-Variable Methods

Paper# TSI-SOTUR-09-03

June 2009

Cristian Angelo Guevara & Moshe Ben-Akiva
Massachusetts Institute of Technology

Addressing Endogeneity in Discrete Choice Models: Assessing Control-Function and Latent-Variable Methods

MIT Portugal Program
Transportation Systems Focus Area

Research Domain:
Transportation Systems Integration
Research Project:
Strategic Options for Integrating Transportation Innovations and Urban Revitalization
(SOTUR)

Paper#:
TSI-SOTUR-09-03

June 2009

Cristian Angelo Guevara
Research Assistant
ITS Lab, Massachusetts Institute of Technology
Department of Civil & Environmental Engineering
77 Massachusetts Ave., 1-249
Cambridge, MA 02139
crguevar@mit.edu

Moshe Ben-Akiva
Edmund K. Turner Professor of Civil and Environmental Engineering
Massachusetts Institute of Technology
Department of Civil & Environmental Engineering
77 Massachusetts Ave., 1-181
Cambridge, MA 02139
mba@mit.edu

This publication was made possible by the generous support of the Government of Portugal through the Portuguese Foundation for International Cooperation in Science, Technology and Higher Education and was undertaken in the MIT-Portugal Program.

ABSTRACT

Endogenity or non-orthogonality in discrete choice models occurs when the systematic part of the utility is correlated with the error term. Under this misspecification, the model's estimators are inconsistent. This problem is virtually unavoidable, for example, in discrete choice models of residential choice where endogeneity occurs at the level of each observation mainly because of the omission of attributes. In such a case, the principal technique to treat for endogeneity is the *control-function* method. This method consists in the construction of a function that accounts for the endogenous part of the error term which is then included as an additional variable in the model. Alternatively, the *latent-variable* method can also be viewed as a procedure to address the endogeneity problem in discrete choice models. In this case, the omitted quality attribute which is causing the endogeneity can be considered as a latent-variable and modeled, in a *structural equation*, as a function of observed variables and potentially enhanced through *indicators.* The main objective of this paper is to analyze similarities and differences among control-function and latent-variable techniques and the exploration of ways by means of which both methods would enhance each other in addressing endogeneity in discrete choice models. This objective is achieved by analyzing the properties of both methods and by testing their performance in the correction of the endogeniety problem in a Monte Carlo experiment. The paper concludes with the analysis of potential future lines of research in this area.

1 INTRODUCTION

The main goal of developing demand models is to forecast users' behavior using available information which is usually very limited. To achieve such a goal, a range of assumptions are needed. First, regarding the behavior of the individual as a function of that limited information and, second, regarding the statistical properties of the information itself. When these assumptions do not hold, forecasting capabilities of the model are invalid or, at least, challenged. One critical assumption to attain consistent estimators of model parameters is known as exogeneity. It assumes that observed model variables are uncorrelated to non-observed ones.

The analysis of methods to correct for endogeneity in models of discrete choice is an area of current development in econometrics (Louviere et al, 2005). One of those techniques corresponds to the control-function method which is particularly suitable when endogeneity occurs at the level of each observation. The purpose of this paper is to explore possible enhancements of the two stage control-function method to correct for endogeneity in discrete choice models that was applied by Guevara and Ben-akiva (2006), in the light of the latent variables approach.

This paper continues as follows. The next section describes the problem of endogeneity in discrete choice models and in the subsequent the basics of the control-function and the latent variables methods are revised. Then, both methods' properties are contrasted and potential equivalences and dissimilarities are studied. In section 5, proposed formulations are tested using an omitted attribute endogenous model of discrete choice constructed with synthetic data. The final section summarizes the principal findings, draws conclusions and reviews potential enhancements of this research.

2 THE PROBLEM: ENDOGENEITY IN DISCRETE CHOICE MODELS

Consider a group of individuals who face the selection of one alternative among a choice set. Assume that each individual methodically chooses the alternative from which a larger amount of *utility* can be retrieved. Consider that the *utility* of each alternative depends on its attributes, which are conditional on the individual's characteristics plus a specific error term.

Consider now an analyst who wants to depict individuals' behavior given some limited information. If the sample used for the estimation is infinitely large and all the attributes are observed by the analyst, the true model coefficients would be obtained. However, if some attributes are not observed the true model coefficients would be obtained if and only if those attributes are not correlated with the observed ones.

This fact can be shown with the following example. Consider that we are interested in modeling the choice of a car make and model. Potential buyer *n* considers in his or her selection of make and model *i* the following variables: price, size, fuel efficiency, safety features and whether the car is red or not (color). Consider that the marginal indirect utility perceived by the individual from each of these attributes corresponds to θ_k and that the utility U_{ni} is completed by some attributes that are specific to each alternative and measured by the alternative-specific-constant ASC_i, and an error term ε_{ni} which is specific to each individual an independent across alternatives.

$$U_{ni} = ASC_i + \theta_p price_{ni} + \theta_s size_{ni} + \theta_e efficiency_{ni} + \theta_s safety_{ni} + \underbrace{\theta_c color_{ni} + \varepsilon_{ni}}_{\delta_{ni}} \qquad (1)$$

Consider now that the analyst can perfectly measure price, size, efficiency and safety, but forgets to register the car's color. In that case, the model's error will be δ_{ni} instead of ε_{ni}, as shown in (1). The analyst's omission of this variable will not compromise the consistency of the estimators, if and only if car's color does not determine the observed attributes (price, size, efficiency and safety). Even though, the scale of the model would be affected by this omission, since the variance of δ will be larger than the variance of ε, and therefore the scale will be smaller.

In turn, if retailers vary car's prices depending on their color, the crucial exogeneity assumption will be broken. Consider that, for instance, red cars suddenly become more popular and retailers quickly adjust their prices upwards to maximize profit. In that case, the analyst will observe that, for seemingly equal cars, which only differ in observed price (and unobserved color), some buyers chose the more expensive alternative. Therefore the analyst will conclude that θ_p is smaller than it really is or even that it is positive, what is against common sense and would make the model worthless.

This model misspecification is called endogenity. The price variable frequently is at the heart of the endogeneity problem in a demand function. For example, in residential choice modeling the quasi-uniqueness of each dwelling unit makes unavoidable the omission of relevant quality attributes which will necessarily be correlated with the market price of the dwelling unit (Guevara and Ben-Akiva 2006). Beyond the omission of attributes, endogeneity in discrete choice models may also be caused by errors in variables (Walker et al, 2008), simultaneous determination, or sample selection bias (Vella, 1992; Eklöf and Karlsson, 1997; Mabit and Fosgerau, 2009).

3 THE METHODS UNDER STUDY

3.1 The Control-Function Method

The control-function method can be thought of as a two stage procedure to address endogeneity in econometric models. For a complete description, the reader is referred to Train (2009). This method is especially suitable for discrete models of residential location choice in which the endogeneity is expected to occur at the level of the individual dwelling unit, because of the omission of quality attributes.

Theoretical basis of the method is described in Heckman (1978), Hausman (1978), Petrin and Train (2005) and Blundell and Powell (2004). The basic idea is to construct a variable or control-function, which would account for the part of the expected value of the error term, conditional on the observed attributes, which is not zero. Thus, if this control-function is added as an explanatory variable in the econometric model, the endogeneity problem would be solved.

To illustrate the intuition behind the construction of the control-function, assume, without loss of generality, that only one observed variable *P* is correlated with the error term. Consider also a proper *instrumental variable Z*. This instrumental variable has to be correlated with the endogenous variable but, at the same time not correlated with the error term of the model. Under those assumptions, the control-function will simply correspond to the fitted error of the ordinary least squares (OLS) regression of *P* as a function of *Z*. This happens because the OLS estimator is a projection of the left hand side variable onto the space spanned by the right hand side variable and the fitted errors are, therefore, orthogonal to the instruments (Greene, 2003). Then, since the instrument is not correlated with the original error term, the fitted error of the price equation captures the part of *P* which is correlated with the error in the original model, and therefore, serves as a control for it.

3.2 The Latent Variables Approach

A complete review of the latent variables approach for models of discrete choice can be found in Walker and Ben-Akiva (2002). The basic idea in this case is that, together within the choice model, some latent or unobservable variables may play a relevant role in the choice behavior. These latent variables can be either determined though *structural equations* as a function of observed variables, or accounted through *measurement equations* within which some observable quantity or indicator is assumed to be a function of the latent variables.

For example, in the case of a mode choice model, a latent variable would correspond to the unobserved quality of the mode. Then, structural equations can be stated within which this quality attribute can be written as a function of, for example, the number of passengers per unit of space and the existence of air-conditioning. Additionally, if the passengers are asked to evaluate their appreciation of the mode's quality on a scale from 1 to 10, this survey question could be used as an indicator in a measurement equation of the true unobserved quality attribute.

The latent variable approach can be estimated either sequentially or simultaneously. In case the simultaneous estimation is considered, the joint likelihood of the model should be accounted for, and the latent variable should be integrated out, making some assumption on its distribution.

4 COMBINING CONTROL-FUNCTION AND LATENT VARIABLES TO CORRECT FOR ENDOGENITY

4.1 Aspects to be addressed

The latent variables and the control-function methods were originally conceived with different purposes. The former was specifically created to address endogeneity which is not the case of the latent-variable method. However, some equivalences and differences among both methods can be clearly identified.

The first issue is related to the fact that the control-function is estimated in two stages, whereas the latent-variables method is estimated, in general, simultaneously. In the next section, inspired by the latent variables approach, different alternatives are explored to adapt the control-function method so that it could be estimated simultaneously.

The second concern is to study how the components of the latent variables approach (the structural equations, the measurement equations and the latent variables themselves) find their respective counterpart, if any, in the control-function approach. Conceptually, the control-function method focuses on the statistical properties of the variables while the latent-variable approach is primarily behaviorally based. Moreover, despite their potential similarity, it is not clear the relationship between the statistical properties of the instrumental variables and, for example, those of the right hand side variables of the measurement equations.

4.2 Simultaneous Estimation of the Control-Function Method

The first aspect to address in the liaison between the control-function method and the latent variables approach is related to the simultaneous estimation of the former. This issue can be addressed using a Full Information Maximum Likelihood (FIML) approach where the likelihood of both the choice model and the equation used to build the control-function, are estimated simultaneously. Since data is shared by both models, this simultaneous procedure will necessarily increase efficiency although it is not clear up to what extent. However, this potential increase in efficiency is not free since the estimation using FIML implies considering assumptions about the joint distribution of the errors in both models. This approach to the control-function method has been previously applied, with variations, by Villas-Boas and Winner (1999) and Park and Gupta (2009).

To explain the procedure proposed in this section, we present first the two stage control-function method as it was applied by Guevara and Ben-Akiva (2006). Consider that an individual *n* perceives a certain utility U_{nj} from an alternative *j* that is a linear function of a set of attributes X_{nj} and the price p_{nj}, a vector of parameters θ, a parameter θ_p and an error term ε_{nj} , as it is shown in (2).

$$U_{nj} = \theta_p p_{nj} + X_{nj}^{'}\theta + \varepsilon_{nj} \tag{2}$$

Assuming that the error term ε is distributed Extreme Value 1 (0,μ) the choice model resulting is the Logit model (Ben-Akiva and Lerman, 1985) and the likelihood of an observation (L_n) corresponds to expression (3), where C_n is the choice set of individual *n* and *i* corresponds to the chosen alternative in for that individual.

$$L_n^{Choice} = \frac{e^{\mu\left(\theta_p p_{ni} + X_{ni}^{'}\theta\right)}}{\sum_{j\in C_n} e^{\mu\left(\theta_p p_{nj} + X_{nj}^{'}\theta\right)}} \tag{3}$$

From the estimation of (3) only $\mu\theta$ and $\mu\theta_p$ can be retrieved but not μ neither θ nor θ_p separately. Therefore, normalization is required to attain identification. This is usually done by setting the scale coefficient to be equal to one. Under this normalization the scale μ "disappears" from expression (3).

Consider now that price is endogenous because it is correlated with some variable which is relevant to the choice process. As explained before, if the likelihood function (3) is maximized, the estimated coefficients obtained by such procedure would not be consistent. However, consider that p_{nj} can also be written as a function of exogenous instruments Z_{nj}, a vector of parameters β, and an error term v as it is shown in (4). We will call this expression the price equation model.

$$p_{nj} = Z'_{nj}\beta + v_{nj} \tag{4}$$

If the instruments are appropriate, the fitted errors of the price equation will account for the part of the price which is correlated with the error term ε in equation (2). This can be shown by noting that $E(\varepsilon_{nj} \mid p_{nj}) = E(\varepsilon_{nj} \mid Z'_{nj}\beta + v_{nj})$ and, if the instruments are not correlated with ε, it follows directly that $E(\varepsilon_{nj} \mid p_{nj}) = E(\varepsilon_{ni} \mid v_{nj})$. Then, if ε and v are assumed to be jointly normal, $E(\varepsilon_{nj} \mid p_{nj}) = \theta_v v_{nj}$, term which will therefore account for the conditional mean of the error which is not equal to zero and, therefore, corrects for the endogeneity problem if it is included in the utility.

Therefore, the first stage of the traditional control-function correction corresponds to the estimation of the price equation using Ordinary Least Squares (OLS) to obtain the fitted errors $\hat{v}$, which are then used as an auxiliary variable of the utility function in the second stage

$$\begin{aligned} p_{nj} &= Z'_{nj}\beta + v_{nj} \xrightarrow{OLS} \hat{v}_{nj} \\ U_{nj} &= \theta_p p_{nj} + X'_{nj}\theta + \theta_v \hat{v}_{nj} + e_{nj} \end{aligned} \tag{5}$$

The enhancement of the two stage control-function method into a one stage procedure follows directly by recalling that, if it is assumed that the error v in (4) is distributed Normal(0, $\sigma^2_v I$), the likelihood of the price equation will correspond to the following expression.

$$L_n^{p-eq.} = \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{1}{2\sigma_v{}^2}\sum\limits_{j\in C_n}\left(p_{nj} - Z'_{nj}\beta\right)^2} \tag{6}$$

Therefore, if the price equation were to be estimated by maximum likelihood, the result would be exactly the same as if it were estimated using OLS since, if we take the logarithm of expression (6), what lasts is precisely the sum of squared residuals plus a multiplicative and an additive constant.

As a result, if it is assumed that the errors in the price equation are normally distributed, in order to achieve the simultaneous estimation of the control-function method it would only be needed to consider as the objective function to be maximized, the product of the likelihood of the price equation and the likelihood of the choice model where in the second, the error of the price equation is considered as an additional variable in the utility function, as it is shown in (7). This procedure is called Full Information Maximum Likelihood (FIML) in econometrics literature (Greene, 2003).

$$L_n^{FIML.} = \frac{e^{\theta_p p_{ni} + X_{ni}^{'}\theta + \theta_\nu \left(p_{ni} - Z_{ni}^{'}\beta\right)}}{\sum_{j \in C_n} e^{\theta_p p_{nj} + X_{nj}^{'}\theta + \theta_\nu \left(p_{nj} - Z_{nj}^{'}\beta\right)}} \prod_{j \in C_n} \frac{1}{\sqrt{2\pi\sigma_\nu^2}} e^{-\frac{1}{2\sigma_\nu{}^2} \sum_{j \in C_n} \left(p_{nj} - Z_{nj}^{'}\beta\right)^2} \tag{7}$$

The FIML estimation of the control-function method can be directly obtained by maximizing expression (7). However, a procedure that may facilitate the performance of the estimation is to consider an iterative process in which, for a given iteration *k*, problem (7) is solved conditional on a given variance of $\hat{\sigma}_{\nu\_k}^2$ and then its value is calculated in the next iteration (until convergence) as it is shown in expression (8) where *N* is the sample size and *J* is the size of the choice set, which is assumed to be equal across the sample. The extension to the case of different choice set sizes is obvious.

$$\hat{\sigma}_{\nu\_k+1}^2 = \frac{1}{JN} \sum_{n,i \in C_n} \hat{\nu}\left(\hat{\beta}_k\right)^2 \tag{8}$$

4.3 Using the Latent-Variables Method to Correct for Endogeneity

Two Stages

One way of using the latent variables method to address endogeneity would be to maintain the two stages procedure of the control-function method but with a shift. The problem of endogeneity comes from the fact that some quality attribute *q* was omitted in the specification of the choice model utility. Therefore, instead of using directly the fitted error of the price equation as the omitted attribute, one can consider a structural equation where the omitted quality attribute is a latent variable written, in an structural equation, as the sum of this fitted error and an additional error term, as it is shown in expression (9).

$$q_{nj} = \hat{\nu}_{nj} + \gamma_{nj} \tag{9}$$

The same result may be attained if the fitted errors are used instead as indicators of the latent variable in a measurement equation. This can be easily noted by reversing expression (9). In both cases the choice utility is specified as including the omitted quality attribute *q*, as it is shown in expression (10).

$$U_{ni} = \theta_p p_{nj} + X_{nj}^{'}\theta + \theta_q q_{nj} + e_{nj} \tag{10}$$

This approach can be seen as an improvement to the two stage control-function method since it addresses the fact that the omitted attribute does not corresponds exactly to the fitted error of the first stage. However, it is arguably not ideal since it still relies on an independent OLS estimation to obtain the fitted errors, losing the potential gain in efficiency that may be achieved from a joint estimation.

If it is assumed that γ in (9) is distributed N(0,σ^2_γI), the likelihood of each observation in this case corresponds to expression (11).

$$L_n^{LV-2Stage} = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\frac{e^{\theta_p p_{ni} X'_{ni}\theta+\theta_q\left(\hat{v}_{ni}+\gamma_{ni}\right)}}{\sum_{j\in C_n} e^{\theta_p p_{ni}+X'_{nj}\theta+\theta_q\left(\hat{v}_{ni}+\gamma_{nj}\right)}}\prod_{j\in C_n}\frac{1}{\sqrt{2\pi\sigma_\gamma^2}}e^{-\frac{1}{2}\frac{\gamma_j^2}{\sigma_\gamma^2}}\,d\gamma \tag{11}$$

Just as it occurred with FIML, one disadvantage of this approach is that it relies on an assumption about the whole distribution of the error term and not just about its expected value. This could be an important issue when the true distribution differs significantly from the distribution assumed in the latent variables model.

It is worth noting also that the estimation of model (11) involves solving a multifold integral in which the number of dimensions is equal to the number of alternatives in the choice set. Since the number of alternatives in, for example, residential location, may be huge, the solving algorithm will necessarily involves Monte Carlo integration with potentially important costs in accuracy. Therefore, even though this approach supposes a theoretical improvement from the two stage control-function estimation since it recognizes that the omitted attribute is a latent variable, the computational burden involved in its application may gloom any improvement in practice.

One Stage

In an attempt to achieve the simultaneous estimation of the control-function model within the latent variable method, the following model can be proposed with the basic idea of using directly the information of the instrumental variables instead of the fitted errors of the price equation. This can be achieved by combining equations (9) and (4) in the following way.

$$\begin{aligned} q_{nj} &= v_{nj} + \gamma_{nj} \\ q_{nj} &= p_{nj} - Z'_{nj}\beta + \gamma_{nj} \end{aligned} \tag{12}$$

Again, if it is assumed that γ is distributed N(0,σ^2_γI), the likelihood of each observation corresponds to the following expression.

$$L_n^{LV-1\,stage} = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\frac{e^{\theta_p p_{ni}+X_{ni}\theta+\theta_q\left(p_{ni}-Z_{ni}\beta+\gamma_{ni}\right)}}{\sum_{j\in C_n} e^{\theta_p p_{ni}+X_{nj}\theta+\theta_q\left(p_{ni}-Z_{ni}\beta+\gamma_{nj}\right)}}\prod_{j\in C_n}\frac{1}{\sqrt{2\pi\sigma_\gamma^2}}e^{-\frac{1}{2}\frac{\gamma_j^2}{\sigma_\gamma^2}}\,d\gamma \tag{13}$$

In the following section the methods proposed are tested and compared in terms of their ability to correct for the endogeneity in a Monte Carlo experiment.

5 MONTE CARLO EXPERIMENT

To study the different alternatives proposed to enhance the correction for endogeneity, we create synthetic data in way that the omission of a quality attribute will necessarily cause endogeneity. The experiment considers 2000 (N) synthetic individuals who choose between three alternatives. Each individual (n) maximizes its utility (U_{ni}), which was assumed to be a linear function of the attributes (a, b, c, a quality attribute q and the price p) of each available alternative (i) and an error term (e_{ni}).

$$U_{ni} = 10a_{ni} + 10b_{ni} + 10c_{ni} + 10q_{ni} - 10p_{ni} + e_{ni} \tag{14}$$

The error term is constructed to be distributed *iid* Extreme Value (0,1) what implies a Logit form for the probability that individual n chooses alternative i. Additionally, price is determined by the price equation shown in (15), which is linear in the attributes c, q, z_1 , z_2, and an error term v_{ni} which was assumed to be distributed Normal (0,0.01).

$$p_{ni} = 0.5c_{ni} + 0.5q_{ni} + 0.5z_{1ni} + 0.5z_{2ni} + \nu_{ni} \tag{15}$$

Variables a, b, c and q were considered *iid* Uniform (1,2) for each individual and alternative. Instruments z_1 and z_2 were considered *iid* Uniform (0,1). Variable p was generated using eq. (15), as a function of c, d and the exogenous instruments z_1 and z_2. Within this setting, variables c and q are correlated with price p but not with either a or b. Therefore if, for example variable q is omitted, price will be correlated wit the error term which, in this case, would be equal to $\varepsilon_{ni} = 10q_{ni} + e_{ni}$. At the same time, variables z_1 and z_2 are, by construction, proper instruments for price since they are correlated with it, but not with the error term ε_{ni} . The following table summarizes the synthetic data considered in these experiments.

Table 1: Summary Statistics of Synthetic Data

			Correlation						
Variable	**Mean**	**Standard Error**	***a***	***B***	***c***	***q***	***p***	z_1	z_2
a	1.5	0.29	1.0						
b	1.5	0.29	0.00019	1.0					
c	1.5	0.29	0.018	0.010	1.0				
q	1.5	0.28	-0.0068	0.0081	-0.016	1.0			
p	2.0	0.29	-0.011	0.012	0.52	0.47	1.0		
z_1	0.50	0.29	-0.00033	0.024	0.026	-0.013	0.50	1.0	
z_2	0.50	0.29	-0.033	-0.021	0.025	-0.013	0.51	-0.014	1.0

Using these synthetic data, seven models were estimated. The first five were estimated using the open source software *R* (R Development Core Team, 2008). The first model (Model I) corresponds to a Logit model in which all the variables that are present in the true model are included. The estimates of this model are shown in the fourth column of Table 2, where it can be noted that the estimated coefficients are statistically equal to the true coefficients.

The second model (Model II) in Table 2 corresponds to the estimation of model in which variable *q* was omitted from the utility specification. Since variable *q* is correlated with the price by construction, this model suffers of endogeneity. As expected, the coefficient of price is positively biased. Since the scale of the different models is not necessarily the same, the correct way to check that the coefficient of price is biased, is by comparing it with the estimated coefficient of variables *a* or *b*, since those variables are independent by construction, to all other variables in the model and also to the error term. Subsequently, it can be noted in this case that the coefficient of *p* is 3 times smaller (in absolute value) than the coefficient of *a*. In the same way, variable *c* is also pushed down (~50%) because it is correlated with the price. Additionally, it can be noted that the log-likelihood of the choice model $L(\hat{\theta})$ is substantially smaller than the one of model I.

Table 2: Monte Carlo Experiment.
Performance of Different Model Estimators to Address Endogeneity.

Model	Coeff.	True Values	Model I Include all variables	Model II Ommit q	Model III 2 stages C-Funct	Model IV Simult C-Funct	Model V Price eq. in Utility	Model VI Lat Vars 2 Stages	Model VII Lat Vars 1 Stage
Choice Model	ASC$_1$	0.00	-0.363 (0.117)	-0.107 (0.0784)	-0.357 (0.116)	-0.358 (0.116)	-0.359 (0.116)	-0.217 (6.01)	-0.359 (0.120)
	ASC$_2$	0.00	-0.100 (0.114)	-0.0287 (0.0771)	-0.0698 (0.113)	-0.0690 (0.113)	-0.0670 (0.114)	-0.0577 (106)	-0.0670 (0.115)
	θ_a	10.0	10.6 (0.494)	4.98 (0.194)	10.4 (0.484)	10.4 (0.485)	10.45 (0.486)	6.41 (0.232)	10.4 (0.689)
	θ_b	10.0	10.4 (0.487)	4.85 (0.192)	10.2 (0.473)	10.2 (0.474)	10.2 (0.475)	6.28 (0.236)	10.2 (0.684)
	θ_c	10.0	10.4 (0.504)	3.21 (0.183)	10.1 (0.493)	10.2 (0.521)	-0.106 (0.272)	5.73 (0.195)	1.94 (0.0530)
	θ_q	10.0	10.6 (0.512)						
	θ_p	-10.0	-10.9 (0.545)	-1.62 (0.163)	-10.7 (0.539)	-10.7 (0.573)	10.1 (0.574)	-5.66 (0.164)	12.2 (0.762)
	θ_{z_1}						-10.2 (0.507)		
	θ_{z_2}						-10.6 (0.539)		
	θ_μ				20.8 (1.01)	20.9 (1.03)		10.9 (0.0824)	-2.10 (0.0660)
Price/Structural Equation	Intercept	0.00			0.763 (0.0107)	0.763 (0.0107)		0.763 (0.0107)	
	β_{z_1}	0.500			0.492 (0.00642)	0.492 (0.00591)		0.492 (0.00642)	4.87 (0.173)
	β_{z_2}	0.500			0.493 (0.00645)	0.507 (0.00593)		0.493 (0.00645)	5.07 (0.196)
	β_c	0.500			0.505 (0.00645)	0.492 (0.00641)		0.505 (0.00645)	0.977 (0.111)
	β_d	0.500							
N		2000	2000	2000	2000	2000	2000	2000	2000
$L(0)$		-2197.22	-2197.22	-2197.22	-2197.22	-2197.22	-2197.22	-2197.22	-2197.22
$L(\hat{\theta})$			-514.12	-1133.01	-521.96	-521.84	-521.73	-602.00	-521.73
$\hat{\theta}_a/\hat{\theta}_p$		-1.00	-0.977	-3.07	-0.973	-0.973	1.03	-1.13	0.854
$\hat{\theta}_a/\hat{\theta}_c$		1.00	1.02	1.55	1.03	1.03	-98.6	1.12	5.38

(*)Estimator Standard Error in Brackets

The next model (Model III) is reported in the sixth column of Table 2 and corresponds to the two stages control-function correction, as it was described in Guevara and Ben-Akiva (2006). In this case it can be noted that the price equation coefficients are statistically equal to the true values and the omission of q resulted in a non-zero intercept. Regarding the choice model, it can also be noted that the coefficient of a is again around its true value of 10. This occurs just because the error term in equation (15) was built to be very small and because all the true variables in (15), but the omitted quality attributes, were considered in the estimation of the price equation. In general, we may expect larger errors in the price equation and also that some of the true instruments may not be available. In that case, we might observe an increase in error of the choice model and therefore, a reduction in the scale (Guevara and Ben-Akiva, 2006).

More important than the adjustment in scale, it can be noted that the inclusion of the control-function as an additional variable satisfactorily corrected for the endogeneity problem since the absolute value of the ratio of the coefficients of, both variables p and c with the coefficient of a, are again around 1 and the coefficients have the correct signs. Equally relevant, the log-likelihood of the choice model is substantially larger than the model without this correction and almost equal to the one attained with the true model. Finally it is important to remark that, despite that variable c is not correlated with q, the omission of q affected it equally since c was correlated with p. In the same way, if the price equation used to build the control-function does not include c as an explanatory variable, the method will correct the bias only for p but not for c affecting the consistency of the estimators. Thus, the general advice is to use as instruments in the price equation all the model variables that are correlated with price but not with the model error.

The following model estimated corresponds to the FIML model described in expression (7), which is labeled here as the simultaneous Control-Function (Model IV). In this application, the weight between the choice model and the price equation likelihoods, that is, the inverse of the variance of the price equation was calculated iteratively using expression (8). The iterative method was preferred since it showed better performance and stability. The cost of this option however, is that we had no estimate of the standard deviation of this estimator.

First, it can be noted that the log-likelihood of the choice model in this case is almost the same as the one attained with the model where the instruments were just included as additional variables in the choice model. However, in this case, the endogeneity problem is correctly solved since the sign of the coefficients of p and c are correct and their size is statistically equal to that of a. The gain with the joint estimation was a slight improvement of the choice model likelihood which is accompanied with an increase in the standard errors of the some coefficients of the choice model. This could be misinterpreted as a potential reduction, instead of an increase in efficiency resulting from the simultaneous estimation. The truth however is that the estimators of the standard errors in the two stages control-function were an incorrect approximation, since they did not account for the whole variability of the two underlying models.

The next model corresponds to an example of simultaneous estimation. It can be noted that if the control-function method conveys the addition of the fitted error as a variable in the choice model, the simultaneity could be achieved by just replacing the price equation directly in the utility function as it is shown in (16).

$$\left.\begin{aligned} U_{ni} &= X_{ni}^{T}\theta + \varepsilon_{ni} \\ p_{ni} &= Z_{ni}^{T}\beta + \nu_{ni} \end{aligned}\right\} U_{ni} = X_{ni}^{T}\theta + \theta_{\nu}\left(p_{ni} - Z_{ni}^{T}\beta\right) + \varepsilon_{ni}$$

$$\begin{aligned} U_{ni} &= \theta_{p} p_{ni} + \overline{X}_{ni}^{T}\theta + \theta_{\nu}\left(p_{ni} - Z_{ni}^{T}\beta\right) + \varepsilon_{ni} \\ U_{ni} &= \left(\theta_{p} + \theta_{\nu}\right) p_{ni} + \overline{X}_{ni}^{T}\theta - Z_{ni}^{T}\beta\theta_{\nu} + \varepsilon_{ni} \\ U_{ni} &= \widetilde{\theta}_{p} p_{ni} + \overline{X}_{ni}^{T}\theta - Z_{ni}^{T}\widetilde{\beta} + \varepsilon_{ni} \end{aligned} \qquad (16)$$

That is, apparently, the same result attained with the control-function model would be obtained if the instruments are just added as additional variables to the choice model. The results of this model are shown in the eighth column of Table 2 (Model V). First it can be noted that the log-likelihood of this model significantly larger than the one of model II, and even slightly superior to the one attained with the control-function correction. This occurs because the model is now estimated simultaneously. However, model estimators are substantially biased. Note that the price coefficient has the incorrect sign and that the ratio between the coefficient of *a* and *c* is now around 100!. It should be remarked that an analyst may be deceived if he or she is blindly studying the inclusion of additional variables to the model using, for example a Likelihood Ratio tests because he or she would end up confidently (but erroneously) including z_1 and z_2 as model variables.

The question of why, if the control-function model and the model including *z* use exactly the same information end up with very different results, rises naturally. The answer can be retrieved from equation (16). It can be noted that the true coefficients can not be identified. For example, the coefficient of price in this case will correspond to the sum of the true coefficient of price (θ_p) and the coefficient of the control-function (θ_v). It can actually be noted that if those coefficients are retrieved from the two stages control-function estimation, they sum up to approximately 10, the actual estimation result for the coefficient of price in the model in which *z* is included in the utility

The next step was to estimate the two latent variables models proposed in the previous section. These models were estimated using the software ICLV2 (Bolduc, 2007). To solve the integral required by the latent variables model, this software assumes a normal distribution of the error terms and solves the integrals using simulation.

The first latent variable model (Model VI) corresponds to the improved two stage version of the control-function where the choice model considers a latent variable that is a function of the fitted error of the price equation and an error term (11). The results of this model are reported in the ninth column of Table 2. It can immediately be noted that this procedure successfully corrected the endogeneity problem since the ratio between the absolute value of the parameters of *a*, *p* and *c* are again around 1 and have the correct sign. The correction, however, tend to be below the level of precision attained with FIML and the two stages control-function method. Regarding the log-likelihood of the choice model, it can be noted that it is substantially larger than the one of the endogenous model but not as good as the one of the two stage control-function. One possible explanation for these results is that this estimator conveyed the use of simulation to calculate the integral and, potentially, the error related with that procedure may surpass the potential improvements gained from the consideration of the latent variable.

An additional important comment regarding the latent variables model is that it was detected that the model was not robust to the starting point used for the estimation. In some cases the model was not estimable and for others it attained convergence to incorrect parameters. It can be speculated that this weakness is explained by simulation error.

Additionally, regarding the comparison between the two stage latent variables control-function, and the simple control-function method, it should be noted that the former relies more heavily on the normality distribution of the error terms since in that case, the whole distribution needs to be integrated out, whereas in the second, only the mean needs to be estimated. The variables in this case were built uniform instead of normal. Despite that the model was estimable, additional experiments (not reported here) for which the variables were normally distributed showed a slightly improved behavior. The true normality of the error terms in the latent variables approach should however play an important role in the estimation of models with real data.

The final column of Table 2 (Model VI) corresponds to the estimated parameters of the latent variables model described in equations (12). It can be seen that, despite that the log-likelihood of the choice model in this case is substantially better than the sequential latent variables model; it does not satisfactorily correct the endogeneity problem. Similar to what occurred with the model in which the instruments were included as explanatory variables in the choice model, the price coefficient in this case is positive, making the model useless.

One possible explanation for this final result is that despite model (13) is identified, the coefficient of β that maximizes the likelihood is not necessarily the same that solves the price equation and, therefore, will not make the correct projection needed to correct for endogeneity. Therefore, what may be needed in this case is to include also the likelihood of the price equation in expression (13). Note that this is just equivalent to formulate a mixture of the model considered to derive (7). The exploration of this extensions are beyond the possibilities of the current version of estimation software used for this research and therefore are left for further research.

6 CONCLUSION

This paper explored different alternatives to address endogeneity in discrete choice models combining the control-function and the latent variables methods. Its was found that the most appropriate way to combine both methods is to consider the omitted attribute as a latent variable which can be then written either as a function of the fitted errors of the price equation in a structural equation, or alternatively, as the right hand side of a measurement equation in which the fitted errors are used as indicators.

Under this framework, five alternative methods were analyzed by means of a Monte Carlo experiment. In this experiment, endogeneity was created, in a trinomial logit model, by the omission of an attribute of the systematic utility that was correlated with one of the remaining attributes.

The first method corresponded to the two stages control-function method as it was applied by Guevara and Ben-Akiva (2006). Using this method it was possible to correct for the endogeneity problem and recover also the scale. However it has to be pointed out that scale may not be recovered if, for example, only one of the instruments is used as an instrument. The second method corresponded to FIML or simultaneous control-function method, in which the likelihood of both the choice model and the price equation are maximized simultaneously. The results showed that the endogeneity problem can also solved in this case with also a relatively improvement in efficiency.

The third model corresponded to the substitution of the price equation in the choice utility. The experiment showed that, if endogeneity exists, when the instruments are directly included in the utility function the problem of endogeneity is not solved at all, but the likelihood of the model can be substantially improved what may deceive the analyst.

The fourth method corresponded to the two stages latent variable model in which a latent variable in the choice model utility was considered to be a function, through a structural equation, of the fitted error of the price equation. Despite this model corrected for the endogeneity problem, its performance was below the two stages control-function and the FIML, arguably, because the computational burden associated with the calculation of the multifold integral.

The last method considered corresponded to the estimation of the latent variables approach in a single stage where instead of considering the fitted errors from an OLS estimation of the price equation, its coefficients are estimated simultaneously as part of the structural equation. As with the case where the instruments were directly included in the utility, this model resulted in a significant improvement of the likelihood but, did not correct at all for the endogeneity problem.

Finally, some relevant future lines of research can be identified. The first extension should consider the analysis of the relative performance of the different methods under different simulated data including different sample sizes and number of alternatives. A second area of research corresponds to the analysis of methods to reduce the computation burden associated with the estimation of the integrals in the latent variables approach is also another potential line of research. Finally, the estimation of the one stage latent variable method presented in this paper where, additionally, the likelihood of the price equation is considered simultaneously, appears a reasonable extension which may potentially address the endogeneity problem.

7 REFERENCES

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

Blundell, R., and Powell, J. (2004). Endogeneity in Semi-Parametric Binary Response Models. *Review of Economic Studies*, Vol. 71, pp. 655-679.

Bolduc, D. (2007). The Integrated Choice and Latent Variable Model (ICLV2). Software manual version 2.2

Greene, W. (2003). *Econometric Analysis*. 5th Edition. New York: Prentice Hall.

Guevara, C. A. and Ben-Akiva, M. (2006). Endogeneity in Residential Location Choice Models. *Transportation Research Record* 1977, pp. 60-66.

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

Heckman, J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, Vol. 46, pp. 931-959.

Louviere, J., Train, K., Ben-Akiva, M., Bhat, C., Brownstone, D., Cameron, T., Carson, C., Deshazo, J., Fiebig, D., Greene, W., Hensher, D., Waldman, D. (2005). Recent Progress on Endogeneity in Choice Modeling. *Marketing Letters*. Vol. 16, No. 3-4, pp. 255-265.

Park, S. and Gupta, S. (2009). A Simulated Maximum Likelihood Estimator for the Random Coefficient Logit Model Using Aggregate Data. *Journal of Marketing Research* forthcoming.

Petrin, A., and Train, K. (2005). Omitted Product Attributes in Discrete Choice Models. Working Paper, National Bureau of Economic Research.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Walker, J. and Ben-Akiva, M. (2002). Generalized Random Utility Model. *Mathematical Social Sciences,* Vol. 43, No. 3, pp. 303-343.

Vella, F. (1992). Simple Tests for Sample Selection Bias in Censored and Discrete Choice Models. *Journal of Applied Econometrics*, Vol. 7, pp. 413–421.

Villas-Boas, M. and Winer, R. (1999). Endogeneity in Brand Choice Models. *Management Science*, Vol. 45, pp. 1324–1338.

Eklöf, J. and Karlsson, S. (1997). Testing and Correcting for Sample Selection Bias in Discrete Choice Contingent Valuation Studies. Working Paper No. 171. Stockholm School of Economics, Sweden.

Mabit, S. and Fosgerau, M. (2009). Mode Choice Endogeneity in Value of Travel Time Estimation. Proceedings of International Choice Conference, Leeds.

Walker, J., Li, J., Srinivansan, S., and Bolduc, D. (2008). Mode Choice Endogeneity in Value of Travel Time Estimation. Proceedings of the Transportation Research Board Annual Meeting.