Chapter 16

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

Cristian Angelo Guevara and Moshe Ben-Akiva

Abstract

Endogeneity or nonorthogonality 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. When endogeneity occurs at the level of each observation, the principal technique used to treat for it is the control-function method, where a function that accounts for the endogenous part of the error term is constructed and is then included as an additional variable in the choice model. Alternatively, the latent-variable method can also address endogeneity. In this case, the omitted quality attribute is considered as a latent variable and modeled as a function of observed variables and/or measured through indicators. The link between the control-function and the latent-variable methods in the correction for endogeneity has not been established in previous work. This paper analyzes the similarities and differences between a set of variations of both methods, establishes the formal link between them in the correction for endogeneity, and illustrates their properties using a Monte Carlo experiment. The paper concludes with suggestions for future lines of research in this area.

16.1. Introduction

Demand models allow us to understand and forecast the behavior of individuals (or agents). This requires a range of assumptions regarding the behavior of the

Choice Modelling: The State-of-the-Art and the State-of-Practice

ISBN: 978-1-84950-772-1

individuals as a function of available information and about the statistical properties of the information itself. In order to have consistent estimators of model parameters, one critical assumption is that the observed model variables are uncorrelated to unobserved ones. The violation of this assumption is defined as endogeneity.

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

The next section describes the problem of endogeneity in discrete choice models. Then the basics of the control-function and the latent-variable methods are surveyed. Afterwards, the properties of both methods are contrasted; the equivalences and dissimilarities are studied; and a formal link between both is established. In Section 16.5, proposed formulations are illustrated and compared using synthetic data. The final section summarizes the principal findings, draws conclusions and suggests future lines of research in this area.

16.2. The Problem: Endogeneity in Discrete Choice Models

When modeling the behavioral response to a certain choice, if all the attributes that are relevant for the individuals are observed, the model estimators will be consistent; that is, they will be as close to the true model parameters as needed, with probability equal to one, if the sample size is large enough. In turn, if some relevant attributes are not observed, the estimators will be consistent if and only if those unobserved attributes are not correlated with the observed ones.

Consider, for example, modeling the choice made by individual n among combinations i of car makes and models. Utility U_{in} perceived by the individual is linear in the following attributes: price, size, fuel efficiency, safety features and whether the car is red or not (color). There is also an additive error term e_{in}, which is independent across alternatives and individuals.

$$U_{in} = ASC_i + \theta_p \text{price}_{in} + \theta_s \text{size}_{in} + \theta_e \text{efficiency}_{in} + \theta_s \text{safety}_{in} + \underbrace{\theta_c \text{color}_{in} + e_{in}}_{\varepsilon_{in}} \tag{16.1}$$

If the price, size, efficiency and safety can be perfectly measured, but the car's color is omitted, the model's error will be ε_{in} instead of e_{in}, as shown in Eq. (16.1). The omission of this variable will not compromise the consistency of the estimators, if and only if car color does not affect the observed attributes (price, size, efficiency and safety). Nonetheless, since the variance of ε_{in} will be larger than the variance of e_{in}, the scale of the estimated model will be smaller than the true scale.

Instead, if car prices are dependent on color, and color is omitted from the model, the crucial exogeneity assumption will be broken. For instance, if red cars become

more popular, retailers will adjust the price of red cars upwards to maximize profit. An external analyst will then observe that, for seemingly equal cars, which only differ in price (and unobserved color), some buyers choose the more expensive alternative. The analyst will then erroneously conclude that θ_p is smaller in absolute value than what it really is or even that it is positive, that would make the modeling effort completely worthless.

The price variable is frequently at the core of the endogeneity problem in a demand function because of the omission of correlated quality attributes. Formally, if it is assumed that the error term has zero mean, price is said to be endogenous if $E(p'\varepsilon) \neq 0$, where $E(\cdot)$ corresponds to the expected value, and p and ε are vectors where the variable price and the error term are stacked correspondingly across alternatives i and individuals n. Beyond the omission of attributes, endogeneity in discrete choice models may also be caused by errors in variables (Walker, Li, Srinivansan, & Bolduc, 2008), simultaneous determination or sample selection bias (Vella, 1992; Eklöf & Karlsson, 1997; Mabit & Fosgerau, 2009).

16.3. The Methods Under Study

16.3.1. The Control-Function Method

The control-function method is a procedure used to address endogeneity in econometric models. For a complete description of this method in the case of discrete choice models see Train (2009). This method is especially suitable when the endogeneity occurs at the level of each observation. This is the case, for example, of residential location choice models.

The 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 that accounts for the non-zero expected value of the error term, conditional on the observed attributes. The endogeneity problem is then solved by adding this control function as an explanatory variable in the choice utility.

For example, consider that only one observed attribute p is correlated with the error term of the utility ε and that a proper set of instrumental variables Z is available. To be a proper instrument, the elements in Z must be correlated with p, but at the same time, should not be correlated with ε. Consider now the ordinary least squares (OLS) regression of p as a function of Z, which will be henceforth labeled as the ''price equation''. Since the fitted errors of this OLS regression are orthogonal to the Z by construction (Greene, 2003), and since Z is not endogenous, it follows that the fitted errors of the price equation will capture all of p that is correlated with ε. Thus, if the fitted errors are included as auxiliary variables in the utility function, they will solve the endogeneity problem.

Finding appropriate instrumental variables is cumbersome. First the instruments need be sufficiently correlated with the endogenous variable, but the actual degree of

correlation required is difficult to determine. Second, the instruments need to be uncorrelated with the error term of the utility, which is unobserved. Formal tests for the validity of instruments exist for linear models. Guevara and Ben-Akiva (2008) developed and applied an adaptation of one of these tests for logit models.

16.3.2. The Latent-Variable Method

The latent-variable method is a technique used to account for unobserved or latent variables in econometric models. The basic idea of the method is to explicitly include the latent variables in the model specification, and then to integrate them out in the calculation of the likelihood of each observation. The problem is that the distribution of the latent variables is unknown. However, this distribution can be depicted under some conditions by accounting for the causality of each latent variable with other latent and measurable variables.

For example, in the case of a choice model, the random utility of each alternative in the choice set is latent since it cannot be measured. Instead, we observe the choices made by the individuals, choices that depend on the utilities of the different alternatives available to them. In this case, the observed choices are said to be indicators explained by the utilities (latent variables) through a measurement equation (the choice behavior).

As an alternative to measurement equations or in addition to them, the distribution of the latent variables can be also depicted by structural equations. In a choice model, the specification of the utility function is an example of a structural equation, where the utility (a latent variable) is written as a function of measurable attributes of each alternative and, potentially, also other latent variables such as unobserved quality attributes.

The latent-variable method can be estimated either sequentially or simultaneously. For a complete description of this method in discrete choice models see Walker and Ben-Akiva (2002).

16.4. Combining Control-Function and Latent-Variable Methods to Correct for Endogeneity

16.4.1. Issues to be Addressed

The latent-variable and the control-function methods were originally conceived for different purposes. Unlike the latent-variable method, the control-function method was specifically created to address endogeneity. The control-function method focuses on the statistical properties of the variables, while the latent-variable approach is primarily behaviorally based. The purpose of this section is to analyze similarities and differences between both methods that may be of relevance in the correction of endogeneity.

There are two issues to be addressed. First, the control-function method is estimated in two stages, whereas the latent-variable method is generally estimated simultaneously. This is important because the number of stages may impact the efficiency of the estimators. Section 16.4.2 proposes an approach to estimate the control-function method in one-stage based on the full information maximum likelihood (FIML) framework (Greene, 2003).

Second, it is not clear which is the counterpart, if any, for the components of the latent-variable method (e.g., the structural equations, the measurement equations and the latent variables) in the control-function approach. To achieve this goal, Section 16.4.3 discusses the conditions under which the latent-variable approach may be directly used to correct for endogeneity. Then, Sections 16.4.4 and 16.4.5 propose two ways to combine the latent-variable and the control-function methods to correct for endogeneity in discrete choice models.

16.4.2. Simultaneous Estimation of the Control-Function Method

The simultaneous estimation of the control-function method can be solved using an FIML approach, where the likelihood of both the choice model and the price equation are maximized simultaneously. Since data and parameters are shared by both models, this simultaneous procedure should increase efficiency. However, this potential increase in efficiency comes with the cost of making stronger assumptions about the joint distribution of the error terms in both models.

To explain the procedure proposed in this section, we first present the two-stage control-function method applied by Guevara and Ben-Akiva (2006). Consider that an individual n perceives a certain utility U_{in} for alternative i, which is a linear function of a set of attributes X_{in} and the price p_{in}. θ and θ_p are parameters and ε_{in} is an error term. In this model, utility U_{in} is a latent variable. Instead, we can observe the choice y_{in}, variable which is equal to one if the alternative i is chosen and zero otherwise. Assuming that individuals choose the alternative with the largest utility within the choice set C_n, the choice model can be formulated as it is shown in Eq. (16.2).

$$\begin{aligned} U_{in} &= \theta_p p_{in} + X'_{in}\theta + \varepsilon_{in} \\ y_{in} &= 1[U_{in} = \max_{j \in C_n} U_{jn}] \end{aligned} \tag{16.2}$$

If the error term ε is the distributed extreme value (0, μ), the resulting choice model is the logit (Ben-Akiva & Lerman, 1985). The likelihood of an observation ($L_n^{\text{Choice}}*$) is equal to Eq. (16.3), where i corresponds to the chosen alternative.

$$L_n^{\text{Choice}}* = \frac{e^{\mu(\theta_p p_{in} + X'_{in}\theta)}}{\sum_{j \in C_n} e^{\mu(\theta_p p_{jn} + X'_{jn}\theta)}} \tag{16.3}$$

The estimation of the model parameters by the maximization of Eq. (16.3), allows us to retrieve only $\mu\theta$ and $\mu\theta_p$, but not μ, θ or θ_p separately. Therefore, some

normalization is required for identification. This is usually done by setting the scale coefficient μ equal to one. Under this normalization, the scale μ disappears from Eq. (16.3).

Consider now that price is endogenous because it is correlated with some unobserved variable that is relevant to the choice process. If the coefficients are estimated by maximizing the likelihood function shown in Eq. (16.3), the estimators will not be consistent. However if, as in Eq. (16.4), p_{in} can be written as a linear function of exogenous instruments Z_{in} and an error term v_{in}, the endogeneity problem can be solved if v_{in} is uncorrelated with Z_{in}, X_{in} and ε_{in}.

$$p_{in} = Z'_{in}\beta + v_{in} \tag{16.4}$$

The first stage of the control-function method consists of the estimation of the price equation using OLS. These estimators are then used to calculate the fitted errors of the price equation $\hat{v}$, which are then used as auxiliary variables that enter the utility function, in the second stage.

$$\begin{aligned} p_{in} &= Z'_{in}\beta + v_{in} \xrightarrow{\text{OLS}} \hat{v}_{in} \\ U_{in} &= \theta_p p_{in} + X'_{in}\theta + \theta_v \hat{v}_{in} + e_{in} \\ y_{in} &= 1[U_{in} = \max_{j \in C_n} U_{jn}] \end{aligned} \tag{16.5}$$

One way to transform the two-stage control-function method into a one-stage procedure, results by making an additional assumption on the distribution of the error term of the price equation. If the error v_{in} in Eq. (16.4) is distributed normal $(0, \sigma_v^2 I)$, the likelihood of the price equation $L_n^{p-\text{eq}}*$ for individual n will correspond to Eq. (16.6).

$$L_n^{p-\text{eq}}* = \prod_{j \in C_n} \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-(1/2\sigma_v^2)(p_{jn} - Z'_{jn}\beta)^2} \tag{16.6}$$

This means that if the price equation were estimated using maximum likelihood, the result would be exactly the same as if it were estimated using OLS. This can be noted in the fact that if we take the log of Eq. (16.6), the result is the negative of the sum of the squared errors plus multiplicative and additive scalars, which play no role in the likelihood maximization process.

It follows directly that if the errors in the price equation are normally distributed, the control-function method can be estimated simultaneously by considering the product of the likelihood of the price equation shown in Eq. (16.6) and the likelihood of the choice model shown in Eq. (16.3) as the objective function to be maximized. As it is shown in Eq. (16.7), in this case the error of the price equation should be included as an additional variable in the utility function.

$$L_n^{\text{FIML}}* = \frac{e^{\theta_p p_{in} + X'_{in}\theta + \theta_v (p_{in} - Z'_{in}\beta)}}{\sum_{j \in C_n} e^{\theta_p p_{jn} + X'_{jn}\theta + \theta_v (p_{jn} - Z'_{jn}\beta)}} \prod_{j \in C_n} \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-(1/2\sigma_v^2)(p_{jn} - Z'_{jn}\beta)^2} \tag{16.7}$$

This one-stage procedure can be seen as what is called FIML in econometrics literature since, instead of maximizing the likelihood of only one problem conditional on the estimated values of the previous stage, the likelihood to be maximized in this case simultaneously considers the information of both models.

16.4.3. Using the Latent-Variable Method to Correct for Endogeneity

In this section, we study how the latent-variable method can be used to correct for endogeneity in discrete choice models. For expositional purposes, we consider a concrete example: the choice of commuting mode. However, the statements can be straightforwardly extended to any type of discrete choice model.

Comfort, safety and reliability are quality attributes, which are difficult to measure, have an effect on mode choice, and are correlated with cost and travel time. If these omitted quality attributes are independent across individuals, the endogeneity problem caused by their omission can be solved by including alternative (modal) specific constants. However, comfort, safety and reliability might depend, in general, on the distance traveled or on the origin and destination pair of each trip, that would make the use of alternative (modal) specific constants insufficient.

The latent-variable approach can be used to correct for endogeneity in this case. This is achieved by including explicitly the omitted quality attribute as a latent variable q_{in} in the utility function. Then, the problem is how to depict the distribution of q_{in}. As discussed before, this can be done using structural and measurement equations.

Consider for example the case of comfort. Even though it is the difference in comfort among modes that affects the choice, comfort cannot be measured by itself. However, we know that comfort can be explained as a mixture of observable variables O_{in} such as passenger density, the age and model of the vehicle, the travel time (which is in X_{in}), the price or even other variables such as safety or status, which may be also latent. We can therefore postulate a structural equation where comfort is in the left hand side and O_{in} is in the right hand side, which includes an additive error as well.

The distribution of the latent variable may also be depicted if individuals provide information or indicators I_{in} on, for example, their declared degree of satisfaction with each mode or their qualitative appreciation of the comfort, safety and reliability experienced in each mode. In this case, a measurement equation can be formulated where the indicator I_{in} is in the left hand side explained by the latent variable q_{in} and a set of variables M_{in}, which may include the individual's characteristics, some components of X_{in}, the price and/or other latent variables as well.

The latent-variable model considering the structural and the measurement equations can be expressed as it is shown in Eq. (16.8), where φ, α and α_q are parameters and ω_{in} and γ_{in} are, respectively, the error terms of the structural and the

measurement equations.

$$\begin{aligned} U_{in} &= \theta_p p_{in} + X'_{in}\theta + \theta_q q_{in} + e_{in} \\ y_{in} &= 1[U_{in} = \max_{j \in C_n} U_{jn}] \\ q_{in} &= O'_{in}\varphi + \omega_{in} \\ I_{in} &= M'_{in}\alpha + \alpha_q q_{in} + \gamma_{in} \end{aligned} \tag{16.8}$$

The likelihood of the choice model is calculated by integrating out the latent variable q_{in}, conditional on the structural and the measurement equations. The maximization of this likelihood will result in a consistent estimation of the parameters under some conditions on the error terms. First, to avoid endogeneity in the measurement and structural equations, it is required that $E(q'\gamma) = 0$, $E(M'_k\gamma) = 0$ and $E(O'_l\omega) = 0$, where k and l correspond to each component of M_{in} and O_{in}, respectively, stacked across alternatives and individuals. To avoid endogeneity due to simultaneous determination between the structural equation and the measurement equation, it is required that $E(\omega'\gamma) = 0$. Finally, to avoid endogeneity due to simultaneous determination between the structural equation and the utility, it is required that $E(\omega' e) = 0$.

Note that it is not necessary to impose the condition $E(e'\gamma) = 0$; 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 the latent variable in the measurement equation is in the right hand side.

In the next section we study the correspondence between this latent-variable formulation and the control-function method. The crucial issue is the identification of the link between elements q_{in}, I_{in}, M_{in} and O_{in} with the instrumental variables and other components of the control-function method. As it is explained in Sections 16.4.4 and 16.4.5 this correspondence can be either stated as a two-stage or as a one-stage procedure.

16.4.4. Two-Stage Latent-Variable/Control-Function Method

The link between the latent-variable and the control-function methods can be established as a two-stage procedure in the following way. Recall that in the control-function method the fitted errors of the price equation $\hat{v}_{in}$ are used directly to replace the omitted quality attribute q_{in}, although we know that there is a discrepancy between q_{in} and $\hat{v}_{in}$. Calling this discrepancy ω_{in} we can always write Eq. (16.9), where φ_0 is a constant to ensure that ω_{in} has zero mean.

$$q_{in} = \varphi_0 + \varphi_v \hat{v}_{in} + \omega_{in} \tag{16.9}$$

It can be shown (Train, 2009) that the assumptions required in the derivation of the control-function method imply that ω_{in} is not correlated with $\hat{v}_{in}$, X_{in} or e_{in}. It follows that Eq. (16.9) can be used as a structural equation where q_{in} is latent.

The resulting two-stage latent-variable/control-function model can be formulated as follows.

$$\begin{aligned} U_{in} &= \theta_p p_{in} + X'_{in}\theta + \theta_q q_{in} + e_{in} \\ y_{in} &= 1[U_{in} = \max_{j \in C_n} U_{jn}] \\ p_{in} &= Z'_{in}\beta + v_{in} \xrightarrow{OLS} \hat{v}_{in} \\ q_{in} &= \varphi_0 + \varphi_v \hat{v}_{in} + \omega_{in} \end{aligned} \tag{16.10}$$

The likelihood of each observation is calculated by integrating out the latent variable q_{in}. Making the change of variables implied by Eq. (16.9), and assuming that ω_{in} is distributed normal $(0, \sigma_\omega^2 \mathrm{I})$, the likelihood of each observation corresponds to Eq. (16.11).

$$L_n^{LV-2\text{Stage}} * = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \frac{e^{\theta_p p_{in} X'_{in}\theta + \theta_q(\varphi_0 + \varphi_v \hat{v}_{in} + \omega_{in})}}{\sum_{j \in C_n} e^{\theta_p p_{jn} + X'_{jn}\theta + \theta_q(\varphi_0 + \varphi_v \hat{v}_{jn} + \omega_{jn})}} \prod_{j \in C_n} \frac{1}{\sqrt{2\pi\sigma_\omega^2}} e^{-(1/2)(\omega_j^2/\sigma_\omega^2)} d\omega \tag{16.11}$$

Note that Eq. (16.11) may also be interpreted independently of the latent-variable approach. If an error term ω_{in} is added to $\hat{v}_{in}$ in model (16.5) and if this error term is then integrated out, the resulting model will be a logit mixture model equal to expression (16.11).

This formulation can be seen as a conceptual improvement to the two stage control-function method described in Eq. (16.7) since it addresses the fact that the omitted attribute does not correspond exactly to $\hat{v}_{in}$. On the other hand, compared to Eq. (16.7), this formulation relies on a stronger assumption on the joint distribution of the error term and involves the resolution of a multifold integral in which the number of dimensions is equal to the number of alternatives in the choice set. This may be impractical, for example, in models of residential choice where the number of alternatives is huge.

16.4.5. One-Stage Latent-Variable/Control-Function Method

To achieve the simultaneous estimation of the control-function method within the latent-variable framework, we propose a model that directly uses the information of the instrumental variables instead of information of the fitted errors of the price equation. This can be achieved by replacing Eq. (16.4) into Eq. (16.9) to obtain Eq. (16.12).

$$q_{in} = \varphi_0 + \varphi_v(p_{in} - Z'_{in}\beta) + \omega_{in} \tag{16.12}$$

The calculation of the likelihood of the model considering this replacement is not straightforward. Note first that the likelihood in Eq. (16.11) is implicitly written conditional on $\hat{v}_{in}$. Therefore, if the actual $v_{in} = p_{in} - Z'_{in}\beta$ is used instead of $\hat{v}_{in}$, the

likelihood of each observation should also include the likelihood of v_{in}. If it is assumed that ω_{in} is distributed normal (0, σ_ω^2I), and making the change of variables implied by Eq. (16.9), the likelihood of each observation corresponds to the following expression.

$$L_n^{LV-1\text{stage}} * = \prod_{j \in C_n} \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-(1/2)((p_{jn} - Z'_{jn}\beta)^2/\sigma_v^2)} \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \frac{e^{\theta_p p_{in} + X'_{in}\theta + \theta_q(\varphi_0 + \varphi_v(p_{in} - Z'_{in}\beta) + \omega_{in})}}{\sum_{j \in C_n} e^{\theta_p p_{jn} + X'_{jn}\theta + \theta_q(\varphi_0 + \varphi_v(p_{jn} - Z'_{jn}\beta) + \omega_{jn})}} \prod_{j \in C_n} \frac{1}{\sqrt{2\pi\sigma_\omega^2}} e^{-(1/2)(\omega_j^2/\sigma_\omega^2)} d\omega \tag{16.13}$$

Note that Eq. (16.13) can be also interpreted as a logit mixture extension of Eq. (16.7). Viewed in this way, it can be shown that Eq. (16.13) is equivalent to the formulations used by Villas-Boas and Winer (1999) and Park and Gupta (2009) to perform a simultaneous estimation of the control-function method or what Train (2009) denominates as maximum likelihood methods.

Like the one-stage estimator described in Section 16.4.2, this one-stage latent variable method will be more efficient than the two-stage method if the assumptions about the joint distribution of e_{in}, ω_{in} and v_{in} are true. However, it may be less efficient if these assumptions fail. Additionally, unlike the two-stage method, the estimation of this one-stage procedure involves the calculation of a multifold integral, which may require numerical approximations that may end up overshadowing any potential improvement in efficiency.

16.5. Monte Carlo Experiment

In this section we create synthetic data that have endogeneity. We then implement a set of variations of the proposed methods to correct for endogeneity and analyze their results. The aim is to illustrate the properties of the different methods and to remark some practical issues associated with their estimation and normalization. Results regarding, for example, relative efficiency are not at all conclusive, since this is just one experiment. The analysis of different specifications of synthetic and/or real data as well, may provide a conclusive answer to those issues in future research.

16.5.1. Experimental Design

The experiment considers 2000 (N) synthetic individuals who choose between two alternatives. Each individual (n) maximizes his/her utility (U_{in}), which is assumed to be a linear function of the attributes a_{in}, b_{in}, c_{in}, a quality attribute q_{in}, the price p_{in} of each available alternative (i) and an error term (e_{in}). The coefficients of each attribute

in the utility are shown in Eq. (16.14).

$$U_{in} = 1a_{in} + 1b_{in} + 1c_{in} + 2q_{in} - 1p_{in} + e_{in} \tag{16.14}$$

The error term is distributed iid extreme value (0, 1) which implies a logit form for the probability that individual n chooses alternative i. Additionally, price is determined by the price equation shown in Eq. (16.15), which is linear in the attributes c_{in}, q_{in}, z_{1in}, z_{2in}, and an error term δ_{in} that is distributed normal (0, 1).

$$p_{in} = 1c_{in} + 1z_{1in} + 1z_{2in} + \underbrace{1q_{in} + \delta_{in}}_{v_{in}} \tag{16.15}$$

Variables a_{in}, b_{in}, c_{in}, z_{1in}, z_{2in} and q_{in} were considered iid uniform (1, 10) for each individual and alternative. Variable p_{in} was generated using Eq. (16.15) as a function of c_{in}, q_{in} and the exogenous instruments z_{1in} and z_{2in}. Table 16.1 summarizes the synthetic data considered in this experiment.

Within this setting, variables c_{in} and q_{in} are correlated with price p_{in} but neither with a_{in} nor with b_{in}. Therefore if the variable q_{in} is omitted, price will be correlated with the error term of the utility, which, in this case, would be equal to $\varepsilon_{in} = 10q_{in} + e_{in}$. Equivalently, the error of the price equation would become $v_{in} = 1q_{in} + \delta_{in}$, whose variance would therefore be equal to $\sigma_v^2 = 7.75$. At the same time, variables z_{1in} and z_{2in} are, by construction, proper instruments for price since they are correlated with price, but not with the error terms ε_{in} and v_{in}.

16.5.2. Models Estimated

Using the synthetic data, seven models were estimated using the open-source software *R* (R Development Core Team, 2008). The first model is a logit model that includes all variables, and acts as a benchmark. In subsequent models variable q_{in} is omitted causing endogeneity. In Models III–VII different variations of the proposed methods to correct for endogeneity are applied. All results are reported in Table 16.2.

Table 16.1: Summary statistics of synthetic data ($N = 2000$).

Variable	Mean	Standard error	Correlation						
			a	*b*	*c*	*q*	*p*	z_1	z_2
a	5.5	2.7	1.0	−0.0046	0.012	−0.017	−0.025	−0.017	−0.034
b	5.6	2.6	−0.0046	1.0	−0.014	−0.019	−0.012	−0.00011	−0.00030
c	5.6	2.6	0.012	−0.014	1.0	0.0060	0.48	0.0020	−0.015
q	5.4	2.7	−0.017	−0.019	0.0060	1.0	0.49	0.031	−0.033
p	22	5.3	−0.025	−0.012	0.48	0.49	1.0	0.52	0.48
z_1	5.5	2.6	−0.017	−0.00011	0.0020	0.031	0.52	1.0	0.029
z_2	5.4	2.6	−0.034	−0.00030	−0.015	−0.033	0.48	0.029	1.0

Table 16.2: Monte Carlo experiment different model estimators to address endogeneity.

Coeff.	True values	Model I: all variables included	Model II: q is excluded	Model III: two-stage control function	Model IV: simultaneous control function	Model V: price equation in utility	Model VI: two-stage control function/latent variable	Model VII: one-stage control function/latent variable
Choice model								
ASC	0.00	0.0681	0.00307	−0.0469	−0.0453	−0.0431	−0.0468	−0.0457
		(0.103)	(0.0541)	(0.0781)	(0.0782)	(0.0782)	(0.0781)	(0.0788)
θ_a	1.00	0.953	0.262	0.549	0.549	0.550	0.549	0.553
		(0.0618)	(0.0167)	(0.0322)	(0.0323)	(0.0323)	(0.0322)	(0.0448)
θ_b	1.00	0.979	0.280	0.575	0.574	0.574	0.575	0.578
		(0.0612)	(0.0172)	(0.0325)	(0.0325)	(0.0325)	(0.0325)	(0.0454)
θ_c	1.00	0.929	0.0787	0.554	0.554	−0.448	0.554	0.558
		(0.0641)	(0.0170)	(0.0364)	(0.0418)	(0.0350)	(0.0364)	(0.0520)
θ_q	2.00	1.90						
		(0.116)						
θ_p	−1.00	−0.954	−0.105	−0.560	−0.560	0.417	−0.560	−0.564
		(0.0574)	(0.00886)	(0.0289)	(0.0312)	(0.0278)	(0.0289)	(0.0439)
θ_{z_1}						−0.979		
						(0.0538)		
θ_{z_2}						−0.990		
						(0.0531)		
θ_v				0.977	0.978		1.00	1.00
				(0.0509)	(0.0522)		Fixed	Fixed
φ_v							0.977	0.984
							(0.0509)	(0.0753)
σ_ω							0.0000105	0.200
							(0.902)	(0.865)

Price equation								
Intercept	0.00			5.26	5.26		5.26	5.26
				(0.166)	(0.166)		(0.166)	(0.166)
β_{z_1}	1.00			1.02	1.01		1.02	1.01
				(0.0171)	(0.0153)		(0.0171)	(0.0153)
β_{z_2}	1.00			0.997	1.00		0.997	1.00
				(0.0168)	(0.0151)		(0.0168)	(0.0151)
β_c	1.00			1.03	1.03		1.03	1.03
				(0.0171)	(0.0171)		(0.0171)	(0.0171)
β_q	1.00							
σ_v	2.78				2.80			2.80
					(N/A)			(N/A)
N	2000	2000	2000	2000	2000	2000	2000	2000
L(0)	−1386.29	−1386.29	−1386.29	−1386.29	−1386.29	−1386.29	−1386.29	−1386.29
$L(\hat{\theta})$		−307.67	−1027.90	−526.33	−525.99	−525.78	−526.33	−526.02
$\hat{\theta}_a/\hat{\theta}_p$	−1.00	−0.999	−2.50	−0.981	−0.981	1.32	−0.981	−0.981
$\hat{\theta}_a/\hat{\theta}_c$	1.00	1.03	3.33	0.991	0.991	−1.23	0.991	0.991

Standard errors in parentheses.
Likelihood of the choice model $L(\hat{\theta})$.
(N/A): Standard error of $\hat{\sigma}_v^2$ cannot be retrieved since $\hat{\sigma}_v^2$ was estimated iteratively.

16.5.2.1. Model I: all variables included The first model corresponds to a logit model in which all the variables that are present in the true model shown in Eq. (16.14) are included. The estimates of this model are shown in the third column of Table 16.2, where it can be seen that all estimated coefficients are statistically equal to the true ones.

16.5.2.2. Model II: *q* is excluded The second model in Table 16.2 corresponds to the estimation of a logit model in which variable q was omitted from the utility specification. Since variable q is correlated with the price by construction, this model suffers from endogeneity. As expected, results show that the estimator of the price coefficient is positively biased in this case. Since the scale of the different models differ, the correct way to check the bias of the price coefficient is by comparing it with the estimated coefficient of variables a or b since those variables are independent, by construction, to all other variables and to the error term. These ratios are at the bottom of Table 16.2, where it can be noted that the coefficient of p is 2.5 times smaller (in absolute value) than the coefficient of a, instead of being equal, as they are in Eq. (16.14). The coefficient of c is also pushed down because it is correlated with price. Additionally, note that the log likelihood of the choice model $L(\hat{\theta})$ is substantially more negative than that of Model I.

16.5.2.3. Model III: two-stage control function The next model corresponds to the application of the two-stage control-function correction, as used in Guevara and Ben-Akiva (2006), over the model that excludes q. The estimators of this model are shown in the fifth column of Table 16.2.

In this case, the variables of Eq. (16.15) used to build the control function were the two instruments z_{1in} and z_{2in}, variable c_{in} and an intercept to guarantee that $\hat{v}_{in}$ has zero expected mean. Either z_{1in} or z_{2in} could be excluded from the calculation of $\hat{v}_{in}$ but not the two at the same time. In turn, the inclusion of c_{in} in the calculation of $\hat{v}_{in}$ is crucial in this case. Although c_{in} is not correlated with ε_{in}, it is correlated with p_{in} and it is an attribute of U_{in}. Therefore, if c_{in} is excluded from the calculation of $\hat{v}_{in}$, and $\hat{v}_{in}$ is used as an auxiliary variable in the choice model, the remaining error e_{in} would be correlated with c_{in}, causing endogeneity again. In general, all model variables that are correlated with the endogenous one, but not with the model error, should be used as instruments in the price equation.

The estimators of the coefficients of the price equation are statistically equal to the true values. The only exception is the intercept, which is larger because of the omission of q_{in}. Regarding the choice model parameters, it can be noted the two-stage control-function method satisfactorily corrected the endogeneity problem since the sign of the coefficients is correct and their ratio, relative to variables a and b, are near to the true ones. Also, the log likelihood of the choice model of Model III is substantially more positive than that of Model II.

16.5.2.4. Model IV: simultaneous control function The next model corresponds to the FIML model described in expression (16.7), which is labeled here as the

simultaneous control-function model. The estimators of this model are shown in the sixth column of Table 16.2.

The estimation of this model has one particularity. Since the variance of the price equation σ_v^2 is a function of other parameters β_l in the model, the correspondence between β_l and σ_v^2 must be guaranteed by the inclusion of a constraint. Instead of including this constraint as part of the maximum likelihood estimation procedure, we consider it iteratively in the following way: (i) For a given iteration k, likelihood shown in Eq. (16.7) is maximized conditional on a given variance $\hat{\sigma}^2_{v\_k}$. (ii) Then, using Eq. (16.16) the resulting estimators are used to calculate the variance to be used in the next iteration. (iii) The process is repeated until convergence.

$$\hat{\sigma}^2_{v\_k+1} = \frac{1}{JN} \sum_{n,j \in C_n} \hat{v}^2_{jn}(\hat{\beta}_k) \tag{16.16}$$

In Eq. (16.16), 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 across individuals is obvious. Note that the use of this iterative procedure implies that the standard error of this estimator cannot be retrieved.

The estimation results of this model show that the endogeneity problem was solved since the sign of the coefficients of p and c are correct and their size is statistically equal, in absolute value, to that of a. A slight improvement can also be noted in the choice model likelihood.

Additionally, there is an increase in the standard errors of some coefficients in the choice model. This could be misinterpreted as a reduction in efficiency resulting from the simultaneous estimation. However, it should be noted that the estimators of the standard errors in the two-stage control function, as in all two-stage procedures, are biased. Therefore, the comparison is not really possible with a single experiment.

16.5.2.5. Model V: price equation in utility An easy (but incorrect) way to achieve simultaneity in the estimation of the control-function method corresponds to the direct replacement of the price equation in the utility as it is shown in Eq. (16.17).

$$\begin{aligned}
v_{in} &= p_{in} - Z'_{in}\beta \\
U_{in} &= \theta_p p_{in} + X'_{in}\theta + \theta_v v_{in} + e_{in} \\
U_{in} &= \theta_p p_{in} + X'_{in}\theta + \theta_v (p_{in} - Z'_{in}\beta) + e_{in} \\
U_{in} &= (\theta_p + \theta_v) p_{in} + X'_{in}\theta - Z'_{in}\beta\theta_v + e_{in} \\
U_{in} &= \tilde{\theta}_p p_{in} + X'_{in}\theta - Z'_{in}\tilde{\beta} + e_{in}
\end{aligned} \tag{16.17}$$

This procedure is equivalent to including the instruments directly in the utility function. The estimation results of this model are shown in the seventh column of Table 16.2. It can be noted that, although the likelihood was substantially improved, the estimators of p and c have the wrong sign.

The reason for these misleading results is that the true coefficients cannot be identified in Model V. Note that the coefficient $\tilde{\theta}_p$ in Eq. (16.17) corresponds to the sum of the true coefficient of price (θ_p) and the coefficient of the control-function (θ_v). Indeed, if the correct estimators of θ_p and θ_v (taken from the control-function method estimations of Models III or IV) are summed, the result is almost equal to the estimator of θ_p that is obtained in Model V.

16.5.2.6. Model VI: two-stage control function/latent variable This model corresponds to the two-stage latent-variable/control-function model described in Eq. (16.11). In this case the choice model includes a latent variable, whose distribution is then depicted using a structural equation in which the fitted errors of the price equation are used as explanatory variables.

The integration of the latent variable was performed using the Hermite procedure with 136 points (Judd, 1998). Some coefficients of Eq. (16.11) need to be normalized to achieve identification. First φ_0 is set to be equal to zero, since all deviations from the mean is already captured by the alternative specific constant of the choice model. Also, θ_v is set to be equal to one to allow the identification of φ_v and the variance of ω.

The results of this model are reported in the eighth column of Table 16.2. This procedure successfully corrected the endogeneity problem since the parameters have the correct sign and the ratios between the absolute value of the parameters of a, p and c are near one. The standard errors in this model tend to be smaller than those of Model IV for all the coefficients. In contrast, the likelihood is slightly worse. Concordantly, the estimator of σ_ω is statistically equal to zero.

These results indicate that, for this particular experiment, the proposed method is outperformed by the two-stage and the simultaneous control-function as well. It can be speculated that the numerical approximations needed for the calculation of the integral required in this model, may play a role in this outcome.

16.5.2.7. Model VII: one-stage control function/latent variable The final method under study corresponds to the one-stage control-function/latent-variable method whose likelihood is described in Eq. (16.13). The estimation procedure in this case considers the same normalization used for Model VI and the iterative procedure described in Eq. (16.16).

The results are shown in the ninth column of Table 16.2. It can be noted that this method succeeded in the correction of endogeneity. In this case the likelihood of the choice model is slightly better than for Models VI and III, but worse than that of Model IV. Also, it can be noted a small improvement the efficiency of the estimators of the price equation, but a worse result in the case of choice model. Additionally, the estimator of σ_ω is more significant than that of Model VI, but still statistically equal to zero for any reasonable level of confidence.

It should be remarked that a formal comparison of the relative efficiency of the different models is not possible with a single experiment since the standard errors of

the two-stage procedures are biased. Further analysis on this issue is left for further research.

16.6. Conclusion

This paper explores different alternatives to address endogeneity in discrete choice models by combining the control-function and the latent-variable methods. It was first shown that both methods can be independently used to treat for endogeneity. In the case of the control-function, the availability of instrumental variables is necessary; whereas in the case of the latent-variable method, the availability of indicators and/or other observed variables that may explain the omitted quality attributes is required. The preference for one method over the other depends on the particular case and on the data available. For example, in mode choice models it is easier to have proper indicators than instruments to correct for endogeneity in travel time, whereas the contrary occurs in residential location choice when correcting for endogeneity of price. It was also shown how both methods may be seen as equivalent. The instruments used in the control-function method can be used to build the fitted errors of the price equation and then those fitted errors can be considered as explanatory variables in a structural equation in the latent-variable method.

The specifications proposed were estimated and assessed, based on their ability to correct for endogeneity in a Monte Carlo experiment. Normalization and other estimation procedures associated with each method were described. Of the five specifications analyzed, four succeeded in correcting the endogeneity. The only failure, as expected, corresponds to the case when the instruments are directly used as additional variables in the choice model.

The specifications analyzed involved estimation procedures performed in one and two stages. Although simultaneous estimators should increase efficiency, they also imply stronger distributional assumptions and a significant computational burden associated with the calculation of a multifold integral that depends on the number of alternatives. The Monte Carlo experiment performed in this research is inconclusive regarding these issues. Therefore, future research in this area should include the analysis of the relative performance of the different methods using real data and different specifications and repetitions of the synthetic data. Further analysis should also be performed regarding the study of methods to reduce the computational burden associated with the estimation of the integrals in the methods proposed.

Acknowledgments

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

Program. Research assistance by Kamil Sveda and editorial assistance by Tina Xue are greatly appreciated.

References

Ben-Akiva, M., & Lerman, S. (1985). *Discrete choice analysis: Theory and application to travel demand*. Cambridge, MA: The MIT Press.

Blundell, R., & Powell, J. (2004). Endogeneity in semi-parametric binary response models. *Review of Economic Studies*, *71*, 655–679.

Eklöf, J., & 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.

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

Guevara, C. A., & Ben-Akiva, M. (2006). Endogeneity in residential location choice models. *Transportation Research Record*, *1977*, 60–66.

Guevara, C. A., & Ben-Akiva, M. (2008). A Lagrange multiplier test for the validity of instruments in MNL models: An application to residential choice. European Transport Conference, Leeuwenhorst, The Netherlands.

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.

Judd, K. (1998). *Numerical methods in economics*. Cambridge, MA: MIT Press.

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*, *16*(3–4), 255–265.

Mabit, S., & Fosgerau, M. (2009). Mode choice endogeneity in value of travel time estimation. *Proceedings of International Choice Modelling Conference*, Leeds.

Park, S., & Gupta, S. (2009). A simulated maximum likelihood estimatorfor the random coefficient logit model using aggregate data. *Journal of Marketing Research*, *46*(4), 531–542.

Petrin, A., & 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, available at: http://www.R-project.org

Train, K. (2009). *Discrete choice methods with simulation* (2nd ed.). Cambridge, UK: Cambridge University Press.

Vella, F. (1992). Simple tests for sample selection bias in censored and discrete choice models. *Journal of Applied Econometrics*, *7*, 413–421.

Villas-Boas, M., & Winer, R. (1999). Endogeneity in brand choice models. *Management Science*, *45*, 1324–1338.

Walker, J., & Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical Social Sciences*, *43*(3), 303–343.

Walker, J., Li, J., Srinivansan, S., & Bolduc, D. (2008). Mode choice endogeneity in value of travel time estimation. *Proceedings of the Transportation Research Board Annual Meeting*.