Transportation Research Part A 82 (2015) 240–254

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Critical assessment of five methods to correct for endogeneity in discrete-choice models

C. Angelo Guevara

Universidad de los Andes, Chile

ARTICLE INFO

Article history:
Received 2 July 2015
Received in revised form 1 October 2015
Accepted 16 October 2015
Available online 3 November 2015

Keywords:
Multiple indicator solution
Proxies
Control-function
Maximum-likelihood
Latent-variables
Monte Carlo

ABSTRACT

Endogeneity often arises in discrete-choice models, precluding the consistent estimation of the model parameters, but it is habitually neglected in practical applications. The purpose of this article is to contribute in closing that gap by assessing five methods to address endogeneity in this context: the use of Proxys (PR); the two steps Control-Function (CF) method; the simultaneous estimation of the CF method via Maximum-Likelihood (ML); the Multiple Indicator Solution (MIS); and the integration of Latent-Variables (LV). The assessment is first made qualitatively, in terms of the formulation, normalization and data needs of each method. Then, the evaluation is made quantitatively, by means of a Monte Carlo experiment to study the finite sample properties under a unified data generation process, and to analyze the impact of common flaws. The methods studied differ notably in the range of problems that they can address; their underlying assumptions; the difficulty of gathering proper auxiliary variables needed to apply them; and their practicality, both in terms of the need for coding and their computational burden. The analysis developed in this article shows that PR is formally inappropriate for many cases, but it is easy to apply, and often corrects in the right direction. CF is also easy to apply with canned software, but requires instrumental variables which may be hard to collect in various contexts. Since CF is estimated in two stages, it may also compromise efficiency and difficult the estimation of standard errors. ML guarantees efficiency and direct estimation of the standard errors, but at the cost of larger computational burden required for the estimation of a multifold integral, with potential difficulties in identification, and retaining the difficulty of gathering proper instrumental variables. The MIS method appears relatively easy to apply and requiring indicators that may be easier to obtain in various cases. Finally, the LV approach appears as the more versatile method, but at a high cost in computational burden, problems of identification and limitations in the capability of writing proper structural equations for the latent variable.

1. Introduction: Causes, impact and cures for endogeneity

Endogeneity occurs when some explanatory variables are correlated with the error term of an econometric model due to, among other things, omitted attributes, measurement or specification errors, simultaneous determination or self-selection. This issue is almost unavoidable in various practical cases, and it results in inconsistent estimators of the parameters, invalidating any type of analysis performed with the model. Forecasting or behavioral assessment for policy design may be seriously misled if they are based on models based in inconsistent estimators (see e.g. Guevara and Thomas, 2007).

E-mail address: caguevara@miuandes.cl

http://dx.doi.org/10.1016/j.tra.2015.10.005
0965-8564/

For example, in a mode choice model between public and private urban transportation, it is likely that the perceived level of discomfort (due to e.g. crowding) will grow with travel time. Since the perceived level of discomfort is relevant for the decision maker, but very difficult to measure by the researcher, its omission will cause endogeneity. This omission will make the model useless for assessing policies that enhance comfort, such as providing air conditioning, or redesigning the vehicle's layout. Besides, this omission will result in poor forecasting capabilities and in an overestimation of the value of travel time savings, which will be confounded with the improvements in comfort. Various researchers have reported results that seem to confirm the existence of endogeneity due to the omission of comfort in public transportation models (see e.g. Wardman and Whelan, 2011 and Tirachini et al., 2013).

Similar problems can be found in models of residential location. When choosing a residence, individuals take into consideration a large list of dwelling attributes, such as location, price, size, painting, layout, natural illumination, orientation or neighborhood attributes. However, the researcher is likely to be able to measure, at best, the dwelling price, size and location, omitting many relevant attributes. Those omitted attributes will be positively correlated with dwelling's price because of market forces. The better the dwelling, the larger will be number of households that would be willing to bid for it, increasing its market price. Therefore, a residential location model that neglects the impact of those omitted attributes will underestimate the impact of price in the choice process. The prevalence of this type of endogeneity has been evidenced by the results obtained by different researchers who have reported estimated coefficients of housing price that are not significant, or even positive, when endogeneity was not addressed (Guevara and Ben-Akiva, 2006, 2012; Guevara, 2005, 2010; Bhat and Guo, 2004; Sermonss and Koppelman, 2001; Waddell, 1992; Quigley, 1976).

A third example, among many others, is interurban mode choice modeling. In this case, different services (bus, train or airplane) are likely to compete among them in terms of price and different dimensions of level of service. Thus, a premium alternative will be more highly priced but, in compensation, it may offer shorter travel or waiting time, fewer transfers, larger space between seats, better seats, food, entertainment, safety and security, a smaller carbon footprint, additional amenities, or just a more attentive and caring crew. The choice-maker will somehow take into consideration some of these amenities, but the researcher will usually be able to account only for price, travel and waiting time, and maybe the number of transfers. Therefore, a choice model omitting some of the various dimensions of the level of service will suffer of endogeneity. Evidence of this type of endogeneity in interurban mode choice, although treated at an aggregated level, can be found in Mumbower et al. (2014).

Various methods have been developed to obtain consistent estimators in spite of the presence of endogeneity. The methods depend on the particular model that is being analyzed. This article focuses on the problem of endogeneity in discrete-choice models. For the case of linear models, the reader is referred to the comprehensive book of Wooldridge (2010).

When endogeneity is present in a discrete-choice model, the methods to correct for it differ importantly on the assumptions considered. For the more general case, the problem could be addressed using nonparametric methods. The reader is referred to Matzkin (2007) for a detailed review and analysis of the conditions needed for achieving identification in nonparametric discrete-choice models. Further analysis of this topic can be found in Chesher (2003, 2005, 2010) and Chesher et al. (2013). In the latter paper, the authors study in particular the discrete-choice case and show that their results also apply in the presence of parametric restrictions. In spite of its generality, the complexity of the nonparametric approach seems to have precluded so far its application in practice.

When the structural equation of the latent utility of a discrete-choice model is linear, and the endogenous explanatory variable is discrete; the problem can only be formally solved using Full Information Maximum Likelihood (FIML). An example of such type of application in transportation is the work of Abay et al. (2013), analyzing the relation between injury's severity and seat belt use in two-vehicle crashes. For achieving identification in this case it is essential to be able to write the endogenous and the dependent variables as a function of exogenous variables that work as instruments.

The Latent-Variables (LV) approach (Walker and Ben-Akiva, 2002) can be classified among the FIML methods to address endogeneity. The idea of the method is to include the latent variable – which can be continuous or discrete – in the model, and to integrate it out to calculate the likelihood function. The conditional distribution of the latent variable, which is needed for the integration, is written using structural and measurement equations. The method requires writing the latent variable as a function of exogenous variables in the structural equation, what may prove challenging for various practical cases.

When the endogenous explanatory variable is continuous in a discrete-choice model with linear utility, there are some alternative methods to address endogeneity that are easier to apply than FIML or nonparametric methods. For example, if endogeneity occurs at the level of groups of observations, such as when automobile prices are determined in equilibriums that occur by regional markets, the problem can be solved using the BLP approach (Berry et al., 1995). This method consists in taking the endogeneity out of the choice model by including alternative specific constants for each alternative in each group or market. Then, in a second stage, the estimated constants are regressed on other explanatory variables, where remaining endogeneity can be resolved using any method for lineal models.

If endogeneity does not only occur at the level of groups of observations but at the level of each observation, the Control-Function (CF) method can be used instead. This method was originally proposed by Rivers and Vuong (1988) for binary Probit, as a generalization of a method proposed by Heckman (1978). Petrin and Train (2002) extended the CF method to Logit Mixture models. The CF function is conceptually similar to the Two Stages Least Squares (2SLS) method (see. e.g. Wooldridge, 2010) for linear models and, as it, the CF can be applied in two stages or simultaneously, case in which it is termed Maximum-Likelihood (ML) (Train, 2009). The CF method has been shown to be particularly suitable to address endogeneity for various types of discrete-choice models (Ferreira, 2010; Guevara and Ben-Akiva, 2006, 2012). The key aspect of the CF