Capturing non-linearities between observed and unobserved variables: how to model comfort while correcting for endogeneity

Anna Fernández Antolín[1], C. Angelo Guevara-Cue[2] and Michel Bierlaire[1]

[1]Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne
[2]Universidad de los Andes, Chile

March 26, 2015

1 Introduction

Endogeneity is an issue that often arises in demand modeling. One of the assumptions to derive random utility models such as logit, probit, nested logit and cross nested logit is that the deterministic part of the utility function is independent from the unobserved factors. If this assumption is violated, it may result in inconsistent estimates of the parameters. This is what is known as endogeneity.

An example of this is found in transportation, when comfort is not included in the model. Typically, when modeling mode choice between public transportation and private modes we have an observed attribute (travel time, travel cost) that is correlated with an unobserved attribute (perception of comfort). When comfort is omitted we may obtain biased estimates for the parameters associated with time and/or cost. This can be seen intuitively as follows: if people are traveling at peak hours when public transportation is very congested, the disutility towards public transportation caused by discomfort is captured by the travel time parameter. It results in a downwards-estimated parameter for travel time, since it captures both the disutility towards public transportation caused by travel time and the disutility caused by discomfort. In a similar way, transportation systems that are more expensive because they are more comfortable – like traveling in the first class in a train – have an upwards estimated parameter related to cost. This parameter is capturing on the one hand the disutility for high prices, but on the other hand the fact that travelers are willing to pay higher prices to travel in a more comfortable way. It can even result in positive estimates for parameters related to cost. This results, of course, in wrong willing to pay estimates, which can have bad consequences.

However, endogeneity is rarely assessed and corrected for in practical applications. This is due to the fact that although several methods to correct for it exist, they are not

easy to apply since most of them need what is known as instruments. A review of the different methods can be found in Fernández-Antolín *et al.* (2014).

In terms of specification of the utility of public transportation, there is evidence in the literature that suggests to use the interaction between travel time and comfort. In other words, it is not the discomfort that plays a role in the utility function, but the discomfort multiplied by the travel time. Intuitively, given the same level of discomfort, it affects the traveler differently if he is traveling for 5 minutes or for 1 hour. This places a methodological challenge for the application of methods to correct for endogeneity, such as the MIS, because they consider the endogeneity is explained by an additive term and it is, at first, how the interaction effect should be accounted for with them.

2 Research question

The methods that we propose to correct for endogeneity are the multiple indicator solution (MIS) and the extended multiple indicator solution (EMIS). The MIS method was introduced by Wooldridge (2002) in the context of linear models and generalized by Guevara & Polanco (2013) for non linear models. An extension of the framework was presented at hEART in 2014 and case study of this method was presented in Fernández-Antolín *et al.* (2015) (forthcoming).

The research question that we want to address is the following: can the multiple indicator solution (MIS) and the extended multiple indicator solution (EMIS) capture non-linear effects between an observable and an unobservable attribute? From a theoretical point of view it is shown that the method works, under some assumptions.

Let us consider a utility function defined as follows:

$$U_{in} = ASC_i + \beta_t t_{in} + \beta_x x_{in} + \beta_\xi t_{in}\xi_{in} + e_{in}, \tag{1}$$

where ASC_i, β_x β_t and β_ξ are parameters to estimate, t_{in} is the travel time, x_{in} are the other variables entering the utility function and e_{in} is a random error term. We assume that t_{in} is correlated with ξ_{in}, so that the above model is endogenous. For instance, i could be a transport mode alternative, where the travel time t_{in} could be correlated with a variable representing the perception of comfort in a transport mode ξ_{in}. Note that we could also consider price to be endogenous instead of travel time. The derivation would remain the same.

Now, assume two indicators $t_{in}I_{1in}$ and $t_{in}I_{2in}$ which are related to the omitted variable $t_{in}\xi_{in}$ by the following relations:

$$\begin{aligned} t_{in}I_{1in} &= \alpha_0 + \alpha_\xi t_{in}\xi_{in} + e_{I_{1in}} \\ t_{in}I_{2in} &= \delta_0 + \delta_\xi t_{in}\xi_{in} + e_{I_{2in}} \end{aligned} \tag{2}$$

From equation (2) we obtain $\xi_{in} = \frac{t_{in}I_{1in} - \alpha_0 - e_{I_{1in}}}{\alpha_\xi t_{in}}$. By substituting this expression in equation (1) and denoting $\theta_\xi = \frac{\beta_\xi}{\alpha_\xi}$ we obtain

$$U_{in} = ASC_i + \beta_t t_{in} + \beta_x x_{in} + \theta_\xi t_{in}I_{1in} - \theta_\xi\alpha_0 - \theta_\xi e_{I_{1in}} + e_{in} \tag{3}$$

The above model is still endogeneous since I_{1in} is correlated with $e_{I_{1in}}$. We will hence apply the control function method (similarly as in Guevara-Cue (2010)) and use I_{2in} as an instrument for I_{1in}. Since I_{2in} is correlated with I_{1in} by equations (2) but uncorrelated with $e_{I_{1in}}$, we can define the following relations:

$$t_{in}I_{1in} = \gamma_0 + \gamma_1 t_{in}I_{2in} + \gamma_t t_{in} + \gamma_x x_{in} + \delta_{in} \tag{4}$$

$$e_{I_{1in}} = \beta_\delta \delta_{in} + \nu_{in} \tag{5}$$

where δ_{in} captures the part of $e_{I_{1in}}$ which is correlated with I_{1in} and ν_{in} is an error term. Given these equations, the utility function in equation (3) can be rewritten as follows:

$$U_{in} = (ASC_i - \theta_\xi \alpha_0) + \beta_t t_{in} + \beta_x x_{in} + \theta_\xi t_{in}I_{1in} - \theta_\xi \beta_\delta \delta_{in} - \theta_\xi \nu_{in} + e_{in} \tag{6}$$

And by denoting $\tilde{ASC}_i := ASC_i - \theta_\xi \alpha_0 \quad , \theta_\delta := -\theta_\xi \beta_\delta \quad$ and $\tilde{e}_{in} := -\theta_\xi \nu_{in} + e_{in}$ we obtain:

$$U_{in} = \tilde{ASC}_i + \beta_t t_{in} + \beta_x x_{in} + \theta_\xi t_{in}I_{1in} + \theta_\delta \delta_{in} + \tilde{e}_{in} \tag{7}$$

where there is no endogeneity anymore.

However, it should be noted that for the method to work, we need to assume that $t_{in}I_{in}$ is a proper indicator for $t_{in}\xi_{in}$, an assumption that may be questionable under some circumstances. In addition, even if the method properly corrects for endogeneity, there is the question of forecasting, in particular, regarding the question of whether or not the researcher would be able to distinguish the effect of changes in comfort, from changes in travel time. In this paper, we show from a theoretical point of view that this is indeed possible. The theoretical proof is complemented by Monte Carlo experiments, as well a comparison with the integrated choice model and latent variables (ICLV) (Walker (2001))approach.

3 Bibliography

References

Fernández-Antolín, Anna, Stathopoulos, Amanda, & Bierlaire, Michel. 2014. Exploratory Analysis of Endogeneity in Discrete Choice Models. *In: Proceedings of the 14th Swiss Transport Research Conference.*

Fernández-Antolín, Anna, Guevara-Cue, C. Angelo, & Bierlaire, Michel. 2015. Correcting for endogeneity using the MIS method: a case study with revealed preference data. *In: Proceedings of the 15th Swiss Transport Research Conference.*

Guevara, C. A., & Polanco, D. 2013. Correcting for endogeneity without instruments in discrete choice models: the multiple indicator solution. *In: Proceedings of the Third International Choice Modelling Conference.*

Guevara-Cue, Cristian Angelo. 2010. *Endogeneity and Sampling of Alternatives in Spatial Choice Models*. Thesis, Massachusetts Institute of Technology. Thesis (Ph. D.)–Massachusetts Institute of Technology, Dept. of Civil and Environmental Engineering, 2010.

Walker, Joan Leslie. 2001. *Extended discrete choice models: integrated framework, flexible error structures, and latent variables*. Ph.D. thesis, Massachusetts Institute of Technology.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.