A correction for endogeneity in choice models with psychological constructs

C. Angelo Guevara[1]
Aurélie Glerum[2]
Anna Fernández Antolín[2]
Michel Bierlaire[2]

3rd Symposium of the European Association for Research in Transportation
September 10th-12th 2014

1. Introduction

This research aims at integrating two methods mitigating parameter biases resulting from the omission of unobserved factors of decisions in discrete choice models. The first method is to estimate an integrated choice and latent variable model (Walker and Ben-Akiva, 2002, Walker, 2001), where a latent factor captures an unobserved qualitative attribute (e.g. the comfort of a transport mode). Omitting this attribute can lead to an inconsistent estimate of the parameters relative to economic factors (e.g. the travel time), and consequently to inconsistent willingness to pay indicators. The second method uses the multiple indicator solution (MIS) to introduce a factor of correction in the choice model, in order to obtain consistent estimates of the economic factors (Guevara and Polanco, 2013).

In Section 2, we present a comparative analysis between the use of integrated choice and latent variable models and the use of the MIS method to account for endogeneity biases. To do so, we use a transportation mode choice case study in Switzerland (Glerum et al., 2014) and show how the omission of comfort in public transportation biases the estimation of value of time (VOT) indicators. In Section 3 we present the theoretical framework to integrate psychological variables and endogeneity correction in choice models. In the presentation we will present an application of this framework on the mode choice case study and present Monte Carlo experiments to validate the correctness of willingness to pay estimates (such as VOT) obtained with this new method.

[1] UNIVERSIDAD DE LOS ANDES, Facultad de Ingeniería y Ciencias Aplicadas, caguevara@miuandes.cl
[2] ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE (EPFL), School of Architecture, Civil and Environmental Engineering (ENAC), Transport and Mobility Laboratory (TRANSP-OR), {aurelie.glerum, anna.fernandezantolin@epfl.ch, michel.bierlaire}@epfl.ch

2. A comparison between ICLV models and models using the MIS method

2.1 Theoretical framework

ICLV model framework

Let us consider an integrated choice and latent variable model where the choice of an alternative i depends on an economic factor t_{in} which is correlated with an unobserved attribute ξ_{in}, and on a set of other explanatory factors x_{in}. The utility of this alternative is specified as follows:

$$U_{in} = ASC_i + \beta_x x_{in} + \beta_t t_{in} + \beta_\xi \xi_{in} + e_{in} \quad (1)$$

where ASC_i, β_x, β_t and β_ξ are parameters to estimate and e_{in} is a random error term. Since t_{in} is correlated with ξ_{in} the above model is endogeneous. 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}. The structural equation of the latent variable model is given as follows:

$$\xi_{in} = \eta_0 + \eta s_n + e_\xi \quad (2)$$

where η_0, η are (vectors of) parameters to estimate, s_n is a vector of socio-economic characteristics of the respondent n, and e_ξ is an error term.

The measurement model specifies the following k measurement equations:

$$I_{kin} = \alpha_{k0} + \alpha_{k\xi}\, \xi_{in} + e_{I_{kin}} \quad (3)$$

where α_0 and $\alpha_{k\xi}$ are parameters to estimate, $e_{I_{kin}}$ is a random error term.

The ICLV method has the drawback that it does not fully capture endogeneity bias. For example, if we now assume that variable ξ_{in} represents comfort in a mode i instead of the mode's perception of comfort, Equation (2) could also depend on attributes of the alternative. Therefore it is unclear that e_ξ is uncorrelated with the other explanatory variables in the utility (as it should be to obtain consistent parameters). The method proposed in Section 3. proposes to address this issue.

MIS method

Instead of using the ICLV method to account for the omission of ξ_{in}, let us now consider a model with the same formulation of utility as in Equation (1).

Assuming that we have two indicators I_{1in} and I_{2in} which are related to the omitted variable ξ_{in} by the following relations:

$$I_{1in} = \alpha_0 + \alpha_\xi \xi_{in} + e_{I_{1in}} \tag{4}$$
$$I_{2in} = \delta_0 + \delta_\xi \xi_{in} + e_{I_{2in}} \tag{5}$$

Given Equation (4), we can replace ξ_{in} by $\frac{I_{1in}}{\alpha_\xi} - \frac{\alpha_0}{\alpha_\xi} - \frac{e_{I_{1in}}}{\alpha_\xi}$ in Equation (1), which becomes:

$$U_{in} = ASC_i - \theta_\xi \alpha_0 + \beta_x x_{in} + \beta_t t_{in} + \theta_\xi I_{1in} - \theta_\xi e_{I_{1in}} + e_{in} \tag{6}$$

where we have defined the following relation: $\theta_\xi = \frac{\beta_\xi}{\alpha_\xi}$. 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, 2010) and use I_{2in} as an instrument for I_{1in}. Since I_{2in} is correlated with I_{1in} by Equations (4) and (5) but uncorrelated with e_{I1in}, we can define the following relations:

$$I_{1in} = \gamma I_{2in} + \delta_{in} \tag{7}$$
$$e_{I_{1in}} = \beta_\delta \delta_{in} + \nu_{in} \tag{8}$$

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 (6) can be rewritten as follows:

$$U_{in} = \widetilde{ASC_\iota} + \beta_x x_{in} + \beta_t t_{in} + \theta_\xi I_{1in} + \theta_\delta \delta_{in} + \widetilde{e_{\iota n}} \tag{9}$$

where we have defined new variables, that is, $\widetilde{ASC_\iota} \coloneqq ASC_i - \theta_\xi \alpha_0$ and $\theta_\delta \coloneqq -\theta_\xi \beta_\delta$. In the above equation, we have the following random error term $\widetilde{e_{\iota n}} \coloneqq -\theta_\xi \nu_{in} + e_{in}$.

The model with the MIS correction is estimated in two stages. First δ_{in} is obtained by taking the residual values of Equation (7). Second, all parameters of Equation (9) are estimated by maximum likelihood. Note that using the full information maximum likelihood would render a one-stage estimation possible.

2.2 Application to the omission of comfort in a mode choice model

In the previous subsection, we have presented two methods which can mitigate biases due to the omission of an unobserved variable in a logit model. In what follows we present an analysis of VOT indicators in a mode choice case study in Switzerland, and show how the above methods allow to obtain more realistic values of time than a logit model where the perception of comfort in public transportation is omitted.

The case study consists of an analysis of transportation mode choices of individuals in low-density areas of Switzerland (Glerum et al., 2014). The mode alternatives are public transportation (PT), private motorized modes (PMM) (car, motorbike, etc.) and soft modes (SM) (bike, walk).

Three models are estimated: (1) a logit model, including travel times, travel costs, socio-economic characteristics of the respondent, (2) an ICLV model, which has the same specification as the logit model, but it additionally includes a latent variable capturing the perception of comfort in public transportation and (3) a logit model including the MIS correction to remedy the omission of the variable capturing the perception of comfort in public transportation.

The values of time are reported in Table 1. For both the ICLV model and the logit model with MIS correction, the VOTs are lower than the values obtained in the simple logit model. It is also observed that the VOT for the logit model with the MIS correction is very close to a reference VOT for commuting trips in Switzerland, which is equal to 27.81 CHF/hour (Axhausen et al., 2008).

Model	**Value of time for PT [CHF/hour]**
Logit model	36.15
ICLV model	22.60
Logit model with MIS correction	28.73

Table 1: Values of time resulting from the estimation of the logit model, the ICLV model and the logit model with the MIS correction.

3. Proposition of extension

The methods described in Section 2.1 are designed to address different goals. On the one hand the purpose of the ICLV model is to assess the impact of a perceptional variable on choice. On the other hand the aim of the model with the MIS correction is to correct endogeneity biases. We propose here a more complete framework which achieves both goals.

The utility of Equation (9) is assumed to be the same, except that we now consider a random coefficient for θ_ξ. More precisely, this coefficient's mean is assumed to be a function of socio-economic characteristics of the respondent. We hence assume the following relation:

$$\theta_\xi \sim Dist(\eta_0 + \eta s_n, \sigma) \qquad (10)$$

where $Dist$ could typically be a normal distribution, and parameters η_0, η and σ should be estimated to assess the heterogeneity of the impact of the latent psychological variable on choice.

The above model hence combines a measurement of the effect of an unobserved variable and a correction for endogeneity. The presentation will include estimation results of the above model on the data from the transport mode choice case study, as well as a comparative analysis with the methods considered in Section 2.1. In addition, we will perform Monte Carlo experiments to show that estimates obtained with the new method are consistent.

4. Bibliography

Axhausen, K.W., Hess, S., König, A., Abay, G., Bates, J. J. and Bierlaire,M. (2008). Income and distance elasticities of values of travel time savings: New Swiss results, *Transport Policy* **15**(3): 173–185.

Glerum, A., Atasoy, B. and Bierlaire, M. (2014). Using semi-open questions to integrate perceptions in choice models, *Journal of Choice Modelling* **10**: 11–33.

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

Guevara, C. A. (2010). *Endogeneity and Sampling of Alternative in Spatial Choice Models,* PhD thesis, Massachusetts Institute of Technology.

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

Walker, J. L. (2001). *Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables,* PhD thesis, Massachusetts Institute of Technology.