International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 14, Number 12 (2021), pp. 1202-1205
© International Research Publication House. http://www.irphouse.com

Multiple indicator solution: an alternative approach to correct for Endogeneity in discrete choice models

Thomas Edison Guerrero Barbosa[1], Sir-Alexci Suarez Castrillon[2] and Isbelia Karina Rincon Parada[3]

[1] *Engineering Faculty, CERG, University Francisco of Paula Santander Ocaña, Colombia*

[2, 3] *Engineering Faculty, GRUCITE, University Francisco of Paula Santander Ocaña, Colombia*

Abstract

Endogenous models lead to wrong decision making; therefore, the correction of this problem in econometric modelling should be a common task. In this paper, the Multiple Indicator Solution method is used to check the recovery of parameters in discrete-choice models. Besides, it corrects the endogeneity problem. For this, Monte Carlo simulation is used to test our working hypothesis. Our findings indicate that the Monte Carlo simulation is a proper tool when it is impossible to have real data. On the other hand, it is found that the method is a useful approach to correct for endogeneity in discrete choice models.

Keywords: Endogeneity, Multiple Indicator Solution (MIS), Discrete Choice Models (DCM), Monte Carlo simulation.

I. INTRODUCTION

Endogeneity is a problem in modelling econometric that yields estimation inconsistent of the model parameters [1].This anomaly can affect any type of models. In linear models has have significance findings to address it [2] but in another type of models (for example, Discrete Choice Models - DCM) the results are scarce. Usually, the endogeneity arises due to omitted attributes, measurement or specification error, simultaneous estimation and/or self-selection [3]. If the endogeneity is not corrected, any conclusion or analysis coming from the model will be wrong.

Multiple Indicator Solution (MIS) method is a proper approach in linear models to reach the consistent parameters [4] and, in this way, to correct for endogeneity. More recently, Guevara and Polanco [5] implemented a novel approach of the MIS method but adjusted to correct for endogeneity in DCM. This paper used this last version of the MIS method and test recovery of model coefficients using the Monte Carlo simulation from a code in R software [6].

To use the MIS method, the modeller needs of the indicators. They are questions or sentences graded by the respondents through a survey. The modeller uses the grades to measure respondents' attitudes and/or perceptions about their decision-making [7]. For each endogenous variable, at least two indicators are needed. Indicators are typically collected using Likert scales [8]. In theory, the collection of indicators using surveys to apply the MIS method is more straightforward than collecting instrumental variables [9], [10] (also known as instruments). In practice, this fact is an advantage of the MIS method.

The rest of the paper is organised as follows. Section 2 describe the Monte Carlo simulation to recovery the parameters in endogenous DCM using the MIS method. The results section describes the main findings coming from the simulation. In the final section, we conclude.

II. A MONTE CARLO SIMULATION TO RECOVER THE PARAMETERS USING THE MIS METHOD

We designed a Monte Carlo experiment to recover the DCM parameters with binary choice using the MIS method. The exciting aspect of Monte Carlo experiments is that the modeller controls all simulation conditions. This fact implies that the modeller defines key aspect such as the population parameters, the sampling distribution of the explanatory variables, the number of replications, and sample size. For the explanatory purposes, let consider the DCM represented by the utility function (U_{in}) for the alternative *i* and individual *n* as follows:

$$U_{in} = ASC_i + \beta_t t_{in} + \beta_c c_{in} + \beta_q q_{in} + \epsilon_{in} \qquad (1)$$

Where ASC_i is the alternative specific constant,β_t, β_c and β_q are parameters of the model, whereas ϵ_{in} is an exogenous error term. The distribution of the error term defines the type of DCM. For example, if ϵ_{in} distributes Gumbel (also called Extreme Value Type I), the popular Multinomial Logit (MNL) model is obtained [11]. Given the aim of our experiment, it will be assumed that the term $\beta_q q_{in}$ is omitted by the modeler. Therefore, the new specification of the DCM as shown in (2):

$$U_{in} = ASC_i + \beta_t t_{in} + \beta_c c_{in} + \varepsilon_{in} \qquad (2)$$

For the simulation purposes, we must suppose that the explanatory variables t_{in} and q_{in} are correlated as shown in (3).

$$t_{in} = \alpha_i + \alpha_q q_{in} + \omega_{in} \qquad (3)$$

Note that in (2) $\varepsilon_{in} = \beta_q q_{in} + \epsilon_{in}$, therefore ε_{in} and t_{in} are correlated throughq_{in}. In this way, the endogeneity arises and t_{in} is considered endogenous. Now, let suppose that the variable q_{in} and error terms φ_{in} can explain two indicators (I_{1in} and I_{2in}), as shown in (4) and (5):

$$I_{1in} = \alpha_{\mathrm{I}_1} + \alpha_{1q} q_{in} + \varphi_{1in} \tag{4}$$

$$I_{2in} = \alpha_{\mathrm{I}_2} + \alpha_{2q} q_{in} + \varphi_{2in} \tag{5}$$

The two-stage MIS method to address the endogeneity in DCM [5] is used. In the first stage, shown in (6), the residuals (□$\hat{\delta}_{in}$) are obtained from the ordinal least squares (OLS) regression of the indicator (I_{1in}) on the variablesc_{in}, t_{in} andI_{2in}. It is also possible to estimate the residuals ($\hat{\delta}_{in}$) obtained from the OLS regression of the indicator (I_{2in}) on the variablesc_{in}, t_{in} andI_{1in}. In the second stage, shown in (7), we estimate the DCM considering the residuals ($\hat{\delta}_{in}$) and the indicators (I_{1in} orI_{2in}) as explanatory variables within the utility function.

$$\begin{aligned} I_{1in} = \alpha_1 + \alpha_t t_{in} + \alpha_c c_{in} + \alpha_{I_{2in}} I_{2in} + \delta_{in} \\ \xrightarrow{OLS} \hat{\delta}_{in} = I_{1in} - \hat{I}_{1in} \end{aligned} \tag{6}$$

$$\begin{aligned} U_{in} = \widehat{ASC}_i + \hat{\beta}_t t_{in} + \hat{\beta}_c c_{in} + \hat{\beta}_{I_{1in}} I_{1in} \\ + \hat{\beta}_{\hat{\delta}_{in}} \hat{\delta}_{in} + \hat{\varepsilon}_{in} \end{aligned} \tag{7}$$

The design of the generation data process included the choices simulation following the parameters specified in Table 1. For simulation purposes, the term ϵ_{in} follows an independent and homoscedastic (IID) Gumbel distribution, whereas the termsω_{in}, φ_{1in} and φ_{2in} distribute Normal (0, 1).

Table 1. Parameters of the Monte Carlo simulation

Parameter	**Value**
ASC_i	-1.0
β_t	-4.0
β_c	-2.0
β_q	-1.0
α_i	0.5
α_q	3.0
$\alpha_{I_1} = \alpha_{I_2}$	0.5
$\alpha_{1q} = \alpha_{2q}$	3.0

IV. RESULT

The estimated models were of binomial logit with linear utilities. The estimated models were the following:

1) *True model*: Containing all the explanatory variables considered in (1). This will be the benchmark model.
2) *Endogenous model*: It was estimated excluding q_{in} in (1). It is the model shown in (2).
3) *MIS-1 model*: This is the model corrected for endogeneity. Here, the variable $\hat{\delta}_{in}$ (coming from the first stage as explained above) and the indicator 1 (I_{in}^1) are included as variables within the utility function. The first and second stage of the MIS method are shown in (8) and (9), respectively.

$$\begin{aligned} I_{in}^1 = \alpha_1 + \alpha_t t_{in} + \alpha_c c_{in} + \alpha_{I_{in}^2} I_{in}^2 + \delta_{in} \xrightarrow{OLS} \hat{\delta}_{in} \\ = I_{in}^1 - \hat{I}_{in}^1 \end{aligned} \tag{8}$$

$$V_{in} = \widehat{ASC}_i + \hat{\beta}_t t_{in} + \hat{\beta}_c c_{in} + \hat{\beta}_{I_{in}^1} I_{in}^1 + \hat{\beta}_{\hat{\delta}} \hat{\delta}_{in} \tag{9}$$

4) *MIS-2 model*: This is the model corrected for endogeneity. Here, the variable $\hat{\delta}_{in}$ (coming from the first stage as explained above) and the indicator 2 (I_{in}^1) are included as variables within the utility function. The first and second stage of the MIS method are shown in (10) and (11), respectively.

$$\begin{aligned} I_{in}^2 = \alpha_1 + \alpha_t t_{in} + \alpha_c c_{in} + \alpha_{I_{in}^1} I_{in}^1 + \delta_{in} \xrightarrow{OLS} \hat{\delta}_{in} \\ = I_{in}^2 - \hat{I}_{in}^2 \end{aligned} \tag{10}$$

$$V_{in} = \widehat{ASC}_i + \hat{\beta}_t t_{in} + \hat{\beta}_c c_{in} + \hat{\beta}_{I_{in}^2} I_{in}^2 + \hat{\beta}_{\hat{\delta}} \hat{\delta}_{in} \tag{11}$$

The correction of endogeneity in DCM produces a change in the scale of the estimators [12], therefore, the ratio of the estimators $\frac{\hat{\beta}_t}{\hat{\beta}_c}$ must be checked instead of the estimators themselves.

The results for 100 repetitions run in the simulation and a sample size of 10000 individuals are displayed in Table 2. Using the parameters shown in Table 1, the true population ratio is $\frac{\beta_t}{\beta_c} = \frac{-4}{-2} = 2$. From Table 2 can also be seen the mean of the ratios for each model, the t-test against the *true* ratios and the bias (in percentage).

Table 2. Statistics of the Monte Carlo simulation

Model	$\frac{\widehat{\boldsymbol{\beta}}_t}{\widehat{\boldsymbol{\beta}}_c}$	**True** $\frac{\boldsymbol{\beta}_t}{\boldsymbol{\beta}_c}$	**% Bias** $\frac{\widehat{\boldsymbol{\beta}}_t}{\widehat{\boldsymbol{\beta}}_c}$	**t-test**^{2}
True	1.999	2.0	0.0	0.15
Endogenous	2.164	2.0	8.2	28.86
MIS-1	2.003	2.0	0.1	0.23
MIS-2	2.011	2.0	0.5	0.98

As shown in Table 2, the *endogenous model* reaches the highest bias (8.2%). It is significant since the t-test is equal to 28.86. This fact tests the hypothesis of inconsistent parameter estimation from the *endogenous model*. The *endogenous model* was not possible to recover the parameters of the model. On the other hand, the *true model* shows a low bias. This value is not statistically different from zero (t-test 0.15). Therefore, it can be stated that the *true model* can recover the parameters correctly. Finally, we can conclude that the models corrected for endogeneity (MIS-1 and MIS-2) can recover the parameters appropriately. Although there is bias from both models, the t-test indicates that this bias is not statistically significant. This allows us to state that the MIS method is a proper approach to recover the parameters and correct models affected for endogeneity. There was no difference when I_{in}^1 was used in the OLS of I_{in}^2 or vice versa. This means that both indicators are sufficiently correlated, allowing an adequate correction for endogeneity.

The boxplot in Figure 1 shows the parameter ratios $\frac{\widehat{\boldsymbol{\beta}}_t}{\widehat{\boldsymbol{\beta}}_c}$ for the models estimated. Boxplots were introduced by Tukey [13] and allow getting an idea of the dispersion (accumulation) of the values drawn. The mean for the 100 replications is represented by as a black dot. The dashed line represents the true ratio value. As can be seen, the variance of the corrected ratios seems lower than those of the endogenous ratios. Note that the mean the parameter ratios $\frac{\widehat{\boldsymbol{\beta}}_t}{\widehat{\boldsymbol{\beta}}_c}$ drawn for the MIS-1 and MIS-2 models are close to the true ratio, therefore, it can be concluded that the MIS method is a proper approach to recover the parameters when the model suffers for endogeneity.

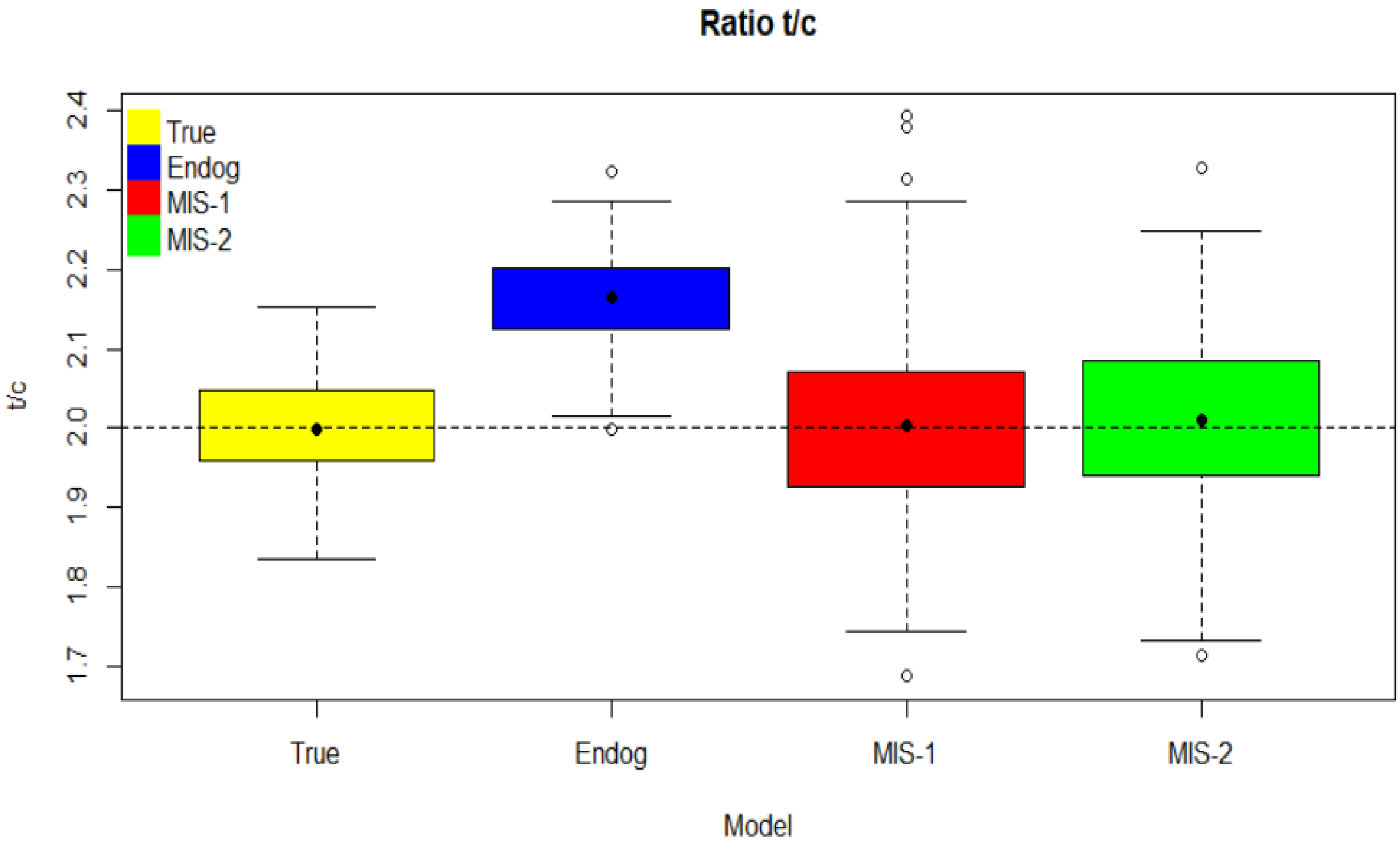

Fig. 1. Boxplots of parameter ratios for the models estimated.

IV. CONCLUSION

This paper shows the results from a Monte Carlo simulation experiment where the recovery of a discrete choice model's parameters is tested. The estimated models were one *true*, one *endogenous* and two models corrected using the MIS method. The recovery of the parameter ratio was compared, and the bias compared to the reference model (*true* model).

Our findings show that the MIS method is a proper approach to correct for endogeneity in DCM. Furthermore, through the simulation, it was possible to show the adverse effects of the inconsistent parameters. In this way, any analysis or conclusion coming from *endogenous models* will be wrong; note that the bias in the ratio of parameters from the *endogenous model* is statistically significant. Besides, if both indicators are highly correlated with the omitted variable, then they are able to capture the effect of the omitted variable within the econometric model.

REFERENCES

[1] W. H. Greene y J. A. Hernández Sánchez, *Análisis econométrico*. Madrid: Prentice-Hall, 1998.

[2] P. Ebbes, D. Papies, y H. van Heerde, «The Sense and Non-Sense of Holdout Sample Validation in the Presence of Endogeneity», *Mark. Sci.*, vol. 30, pp. 1115-1122, nov. 2011, doi: 10.1287/mksc.1110.0666.

[3] C. Guevara, «Critical assessment of five methods to correct for endogeneity in discrete-choice models», *Transp. Res. Part Policy Pract.*, vol. 82, n.º C, pp. 240-254, 2015.

[4] J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, 2.ª ed. Cambridge, MA, USA: MIT Press, 2010.

[5] C. A. Guevara y D. Polanco, «Correcting for endogeneity due to omitted attributes in discrete-choice models: the multiple indicator solution», *Transp. Transp. Sci.*, vol. 12, n.º 5, pp. 458-478, may 2016, doi: 10.1080/23249935.2016.1147504.

[6] «R: a language and environment for statistical computing». https://www.gbif.org/es/tool/81287/r-a-language-and-environment-for-statistical-computing (accedido dic. 05, 2021).

[7] H. Siringoringo, «Perceived usefulness, ease of use, and attitude towards online shopping usefulness towards online airlines ticket purchase.», *Procedia - Soc. Behav. Sci.*, vol. Vol. 81, pp. 212-216, jun. 2013.

[8] R. Likert, «A technique for the measurement of attitudes», *Arch. Psychol.*, vol. 22 140, pp. 55-55, 1932.

[9] J. J. Heckman, «Dummy Endogenous Variables in a Simultaneous Equation System», *Econometrica*, vol. 46, n.º 4, pp. 931-959, 1978, doi: 10.2307/1909757.

[10] D. Rivers y Q. H. Vuong, «Limited information estimators and exogeneity tests for simultaneous probit models», *J. Econom.*, vol. 39, n.º 3, pp. 347-366, 1988.

[11] T. A. Domencich y D. McFadden, «Urban travel demand - a behavioral analysis», Art. n.º Monograph, 1975, Accedido: dic. 05, 2021. [En línea]. Disponible en: https://trid.trb.org/view/48594

[12] C. Guevara y M. Ben-Akiva, «Change of Scale and Forecasting with the Control-Function Method in Logit Models», *Transp. Sci.*, vol. 46, pp. 425-437, ago. 2012, doi: 10.2307/23263552.

[13] J. W. Tukey, «Exploratory data analysis». 1977. Accedido: dic. 05, 2021. [En línea]. Disponible en: https://zbmath.org/?q=an%3A0409.62003