eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

V Workshop on Discrete Choice Modelling

Dealing with endogeneity in Discrete Choice Models

David Hoyos

Warsaw, 5 october 2016

Introduction

- Endogeneity refers to the existence of correlation between the deterministic part and the error term of a model.
- In the presence of endogeneity, parameter estimates are inconsistent thus invalidating any statistical inference performed.
- Although we know that this issue may be unavoidable in many practical cases, the severe consequences of its existence may deserve specific attention, especially in regards to welfare measures.
- The main source of endogeneity in environmental valuation may be found in the omission of contextual conditions in the choice situations.

Correcting for endogeneity in discrete choice models

Endogeneity in DCM may be addressed using the following methods:

1. Proxy variable approach (e.g. Wardman and Whelan, 2011; Tirachini et al., 2013).
2. Control Function method (Petrin and Train, 2002).
3. Maximum Likelihood approach to CF method for simultaneous estimation (e.g. Walker et al., 2007).
4. Latent variables approach (e.g. Hoyos et al., 2015).
5. Multiple Indicator Solution (MIS), as proposed by Guevara and Polanco (2015).

Basic model

The structural equation for the choice model is given by the random utility theory, which is used to link the deterministic model with a statistical model of human behavior. Under this framework, the utility of alternative i for respondent n is given by:

$$\begin{aligned} U^*_{in} &= X_{in}\beta_X + \beta_q q^*_{in} + e^*_{in} = X_{in}\beta_X + \epsilon^*_{in} && (1) \\ y_{in} &= 1[U^*_{in} \geq U^*_{jn}; \forall j \in C_n] && (2) \end{aligned}$$

Since q^* is latent, the actual error term of the choice model will be ϵ^*, thus endogeneity will exist if at least one of the elements in X depends on q^*.

Multiple Indicator Solution (I)

Let us assume that, instead of the latent variable q^*, the researcher observes two indicators:

$$\begin{aligned} q_{1in} &= \alpha_1^0 + \alpha_1 q_{in}^* + e_{q1in}^*, \\ q_{2in} &= \alpha_2^0 + \alpha_2 q_{in}^* + e_{q2in}^* \end{aligned}$$

And, let us assume that the latent variable q^* is generated by the following structural equation, where H is a matrix of exogeneous variables, with its first column equal to one, θ a vector of coefficients, and ω^* an exogenous error term:

$$q_{in}^* = H_{in}\theta + \omega_{in}^*$$

Multiple Indicator Solution (II)

Then, if q^*_{in} is replaced by q_{1in} in the utility, the new error term v^*_{in} is not correlated with any x and, by construction, the indicator q_1 is the only endogenous variable:

$$\begin{aligned} U^*_{in} &= X_{in}\beta_X + \frac{\beta_q}{\alpha_1}(q_{1in} - \alpha^0_1 - e^*_{q1in}) + e^*_{in}, \\ U^*_{in} &= X_{in}\beta_X + \gamma_1 q_{1in} + \left(-\frac{\beta_q \alpha^0_1}{\alpha_1} - \frac{\beta_q e^*_{q1in}}{\alpha_1}\right) + e^*_{in}, \\ U^*_{in} &= X_{in}\beta_X + \gamma_1 q_{1in} + v^*_{in}. \end{aligned}$$

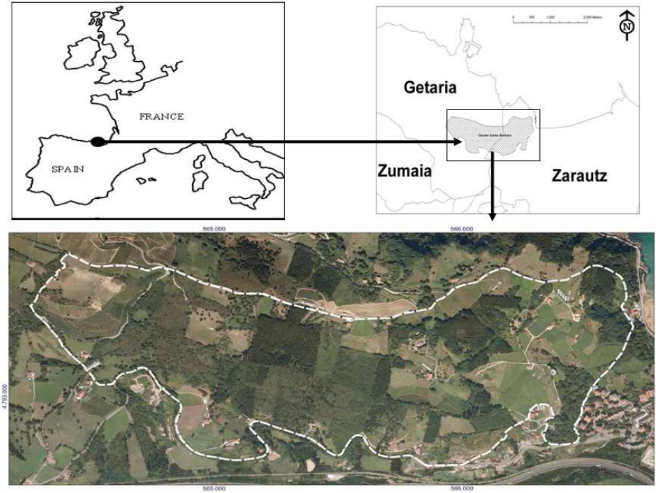
The inclusion of one of the indicators in the utility does not solve the endogeneity problem, it only changes the source of it into q_1. But now, we have a proper instrument for the only endogenous variable q_1, i.e. q_2.

Multiple Indicator Solution (III)

Then, under certain conditions, consistent estimation of β_X would be obtained by applying the following two-stage procedure:

1. Obtaining the residuals $\hat{\delta}_q$ from the OLS regression of q_{1in} on X_{in}, q_{2in}.
2. Estimating the choice model considering the following variables in the systematic part of the utility function: $X_{in}, q_{1in}, \hat{\delta}_q$.

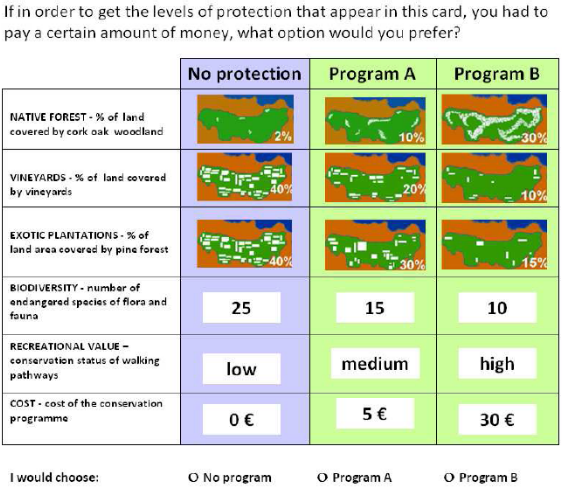
Garate-Santa Barbara N2000 site



Attributes and levels considered

Attributes	**Levels**					
Native Forest (AUT)	2 %	10 %	20 %	30 %		
Vineyard (VIN)	40 %	30 %	20 %	10 %		
Exotic tree plantations (FOR)	40 %	30 %	25 %	15 %		
Biodiversity (BIO)	25	25	10	5		
Recreation (REC)	Low	Medium	High	Very high		
Cost, in euros (COST)	0	5	10	30	50	100

Example of a choice card



If in order to get the levels of protection that appear in this card, you had to pay a certain amount of money, what option would you prefer?

	No protection	Program A	Program B
NATIVE FOREST - % of land covered by cork oak woodland	2%	10%	30%
VINEYARDS - % of land covered by vineyards	40%	20%	10%
EXOTIC PLANTATIONS - % of land area covered by pine forest	40%	30%	15%
BIODIVERSITY - number of endangered species of flora and fauna	25	15	10
RECREATIONAL VALUE – conservation status of walking pathways	low	medium	high
COST - cost of the conservation programme	0 €	5 €	30 €

I would choose: ○ No program ○ Program A ○ Program B

Model specification (MIS1)

Choice model:

$$U^*_{in} = ASC_i + \beta_1 AUT_{in} + \beta_2 VIN_{in} + \beta_3 FOR_{in} + \beta_4 BIO_{in} + \beta_5 REC_{in} + \\ + \beta_6 COST_{in} + \epsilon^*_{in}$$

$$\pi_{n,c_s} = \frac{exp(\mu_{0s} + \lambda'_s SD_n + \lambda_q q^{Alloc*}_{in})}{\sum_{s=1}^{C} exp(\mu_{0s} + \lambda'_s SD_n + \lambda_q q^{Alloc*}_{in})}$$

Auxiliary regression (for allocation probabilities only):

$$IND1^{qAlloc*}_{in} = \alpha_0 + \alpha_1 IND2^{qAlloc*}_{in} + \alpha_2 IND3^{qAlloc*}_{in} + ... + \alpha_K SD_{Kin} + \delta_{1in}$$

Estimated choice model:

$$U^*_{in} = ASC_i + \beta_1 AUT_{in} + \beta_2 VIN_{in} + \beta_3 FOR_{in} + \beta_4 BIO_{in} + \beta_5 REC_{in} + \\ + \beta_6 COST_{in} + \epsilon^*_{in}$$

$$\pi_{n,c_s} = \frac{exp(\mu_{0s} + \lambda'_s SD_n + \lambda_{qs} IND1^{qAlloc*}_{in} + \lambda_{\delta s}\hat{\delta}_{1in})}{\sum_{s=1}^{C} exp(\mu_{0s} + \lambda'_s SD_n + \lambda_{qs} IND1^{qAlloc*}_{in} + \lambda_{\delta s}\hat{\delta}_{1in})}$$

Model specification (MIS2)

Choice model:

$$\begin{aligned} U_{in}^{*} = ASC_i + \beta_1 AUT_{in} + \beta_2 VIN_{in} + \beta_3 FOR_{in} + \beta_4 BIO_{in} + \beta_5 REC_{in} + \\ + \beta_6 COST_{in} + \beta_q q_{in}^{Util*} + \epsilon_{in}^{*} \end{aligned}$$

$$\pi_{n,c_s} = \frac{exp(\mu_{0s} + \lambda_s' SD_n + \lambda_q q_{in}^{Alloc*})}{\sum_{s=1}^{C} exp(\mu_{0s} + \lambda_s' SD_n + \lambda_q q_{in}^{Alloc*})}$$

Auxiliary regressions:

$$IND1_{in}^{qAlloc*} = \alpha_0 + \alpha_1 IND2_{in}^{qAlloc*} + \alpha_2 IND3_{in}^{qAlloc*} + \ldots + \alpha_K SD_{Kin} + \delta_{1in}$$

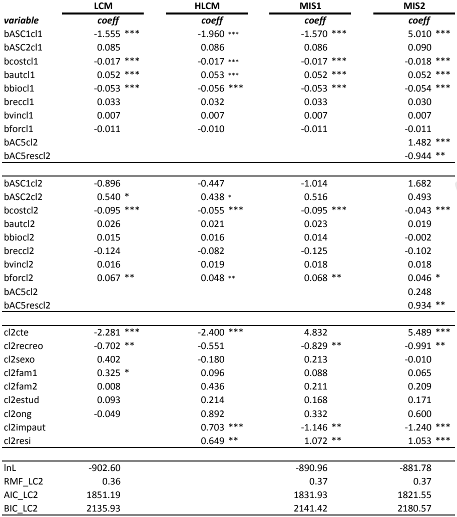
$$IND1_{in}^{qUtil*} = \gamma_0 + \gamma_1 IND2_{in}^{qUtil*} + \gamma_2 IND3_{in}^{qUtil*} + \ldots + \gamma_K SD_{Kin} + \delta_{2in}$$

Estimated choice model:

$$\begin{aligned} U_{in}^{*} = ASC_i + \beta_1 AUT_{in} + \beta_2 VIN_{in} + \beta_3 FOR_{in} + \beta_4 BIO_{in} + \beta_5 REC_{in} + \\ + \beta_6 COST_{in} + \beta_7 IND1_{in}^{qUtil*} + \beta_8 \hat{\delta}_{2in} + v_{in}^{*} \end{aligned}$$

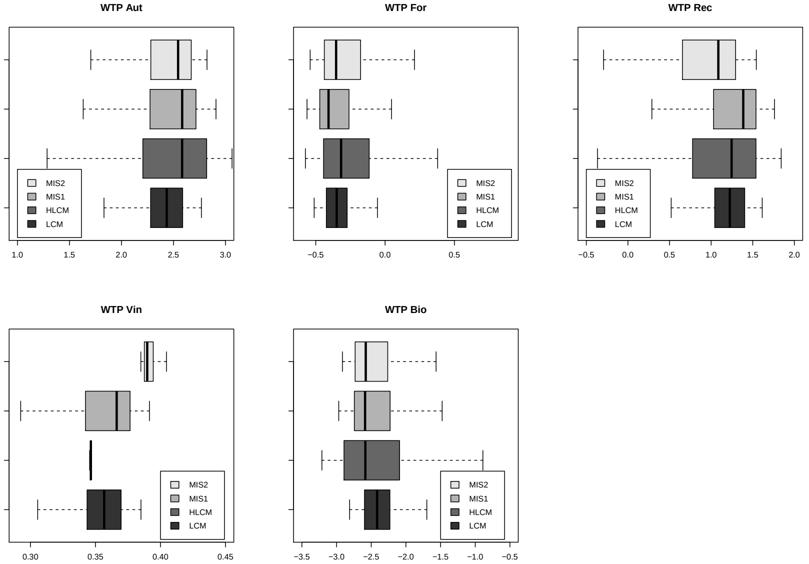
$$\pi_{n,c_s} = \frac{exp(\mu_{0s} + \lambda_s' SD_n + \lambda_{qs} IND1_{in}^{qAlloc*} + \lambda_{\delta s} \hat{\delta}_{1in})}{\sum_{s=1}^{C} exp(\mu_{0s} + \lambda_s' SD_n + \lambda_{qs} IND1_{in}^{qAlloc*} + \lambda_{\delta s} \hat{\delta}_{1in})}$$

Estimation results



	LCM	HLCM	MIS1	MIS2
variable	*coeff*	*coeff*	*coeff*	*coeff*
bASC1cl1	-1.555 ***	-1.960 ***	-1.570 ***	5.010 ***
bASC2cl1	0.085	0.086	0.086	0.090
bcostcl1	-0.017 ***	-0.017 ***	-0.017 ***	-0.018 ***
bautcl1	0.052 ***	0.053 ***	0.052 ***	0.052 ***
bbiocl1	-0.053 ***	-0.056 ***	-0.053 ***	-0.054 ***
breccl1	0.033	0.032	0.033	0.030
bvincl1	0.007	0.007	0.007	0.007
bforcl1	-0.011	-0.010	-0.011	-0.011
bAC5cl2				1.482 ***
bAC5rescl2				-0.944 **
bASC1cl2	-0.896	-0.447	-1.014	1.682
bASC2cl2	0.540 *	0.438 *	0.516	0.493
bcostcl2	-0.095 ***	-0.055 ***	-0.095 ***	-0.043 ***
bautcl2	0.026	0.021	0.023	0.019
bbiocl2	0.015	0.016	0.014	-0.002
breccl2	-0.124	-0.082	-0.125	-0.102
bvincl2	0.016	0.019	0.018	0.018
bforcl2	0.067 **	0.048 **	0.068 **	0.046 *
bAC5cl2				0.248
bAC5rescl2				0.934 **
cl2cte	-2.281 ***	-2.400 ***	4.832	5.489 ***
cl2recreo	-0.702 **	-0.551	-0.829 **	-0.991 **
cl2sexo	0.402	-0.180	0.213	-0.010
cl2fam1	0.325 *	0.096	0.088	0.065
cl2fam2	0.008	0.436	0.211	0.209
cl2estud	0.093	0.214	0.168	0.171
cl2ong	-0.049	0.892	0.332	0.600
cl2impaut		0.703 ***	-1.146 **	-1.240 ***
cl2resi		0.649 **	1.072 **	1.053 ***
lnL	-902.60		-890.96	-881.78
RMF_LC2	0.36		0.37	0.37
AIC_LC2	1851.19		1831.93	1821.55
BIC_LC2	2135.93		2141.42	2180.57

WTP measures



Questions?

Thank-you very much for your attention!

David Hoyos

University of the Basque Country (UPV/EHU)

Faculty of Economics and Business Administration
Department of Applied Economics III (Econometrics and Statistics)

Tel.: +34 94 601 7019
Email: david.hoyos@ehu.eus

http://www.ehu.eus/david.hoyos