

Hybrid Choice Models and accounting for the endogeneity of indicator variables: a Monte Carlo investigation

Wiktor Budziński,

Mikołaj Czajkowski



**Fundacja na rzecz
Nauki Polskiej**

► Why Hybrid Choice Models?

- Allow for inclusion of ‘soft’ variables such as perceptions and attitudes into the choice model using latent variables framework
- Direct incorporation of indicator variables into choice model may lead to biased estimates due to endogeneity and measurement problems
 - *“To what extent do you agree with the statement that the results of the survey will influence future policy?”*
(from 1 - ‘definitely disagree’ to 5 - ‘definitely agree’)
- More ‘behavioral’ approach for explaining preference heterogeneity

- ▶ Hybrid Choice models (HCM) usually consist of three parts:
 - ▶ Choice equations (utility):

$$V_{ijt} = \boldsymbol{\beta}_i' \mathbf{X}_{ijt} + e_{ijt}$$

$$\boldsymbol{\beta}_i = \boldsymbol{\Lambda} \mathbf{L} \mathbf{V}_i + \boldsymbol{\Omega} \mathbf{S} \mathbf{D}_i + \boldsymbol{\beta}_i^*$$

- ▶ Structural equations:

$$\mathbf{L} \mathbf{V}_i = \boldsymbol{\Psi}' \mathbf{X}_i^{str} + \boldsymbol{\xi}_i$$

- ▶ Measurement equations

$$\mathbf{I}_i = \boldsymbol{\Gamma} \mathbf{L} \mathbf{V}_i + \boldsymbol{\Phi} \mathbf{X}_i^{Mea} + \boldsymbol{\eta}_i$$

- ▶ Reasons for endogeneity (Chorus and Kroesen, 2014):
 - ▶ missing variables which influence both latent variable and choices of individuals
 - ▶ learning effects
 - ▶ individuals tend to align their attitudes with their actual choices in order to seem consistent
- ▶ Daly et al. (2011) states: *“The advantages of the latent variable framework over deterministic attitude incorporation are clear; the model is not affected by endogeneity bias [...]”*
- ▶ Similar statements in Hess and Stathopoulos (2013), Hess, Shires and Jopson (2013), Kløjgaard and Hess (2014) and Bello and Abdulai (2015)

- ▶ Two types of indicator variables endogeneity:
 - ▶ LV-endogeneity
 - ▶ Latent variable is endogenous in itself
 - ▶ Correlated error terms in choice model and structural equations
 - ▶ M-endogeneity
 - ▶ Indicator variables are endogenous, but latent variable is not
 - ▶ Correlated error terms in choice model and measurement equations
- ▶ Simulation with 1'000 individuals, 6 choice tasks per person, 3 alternatives per choice task (including the Status Quo)
- ▶ 1000 repetitions

► Data generating process:

LV-endogeneity

M-endogeneity

Utility function

$$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$$

$$\beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13} X_i^{Miss}$$

$$\beta_{2i} = \alpha_{21} + \alpha_{22} LV_i$$

$$\beta_{3i} = \alpha_{31} + \alpha_{32} LV_i$$

Indicator variables
(measurement
component)

$$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1}$$

$$I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2}$$

$$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1} + \alpha_{44} X_i^{Miss}$$

$$I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2} + \alpha_{54} X_i^{Miss}$$

Latent variables
(structural component)

$$LV_i^* = \alpha_{61} X_i^{SD} + \xi_i + \alpha_{62} X_i^{Miss}$$

$$LV_i^* = \alpha_{61} X_i^{SD} + \xi_i$$

► Estimated models:

- Base models allow to check whether simulation works properly, and the extend of measurement error:

	Model type	Measurement error	Endogeneity	Description
Model 1	Hybrid MNL	No	No	No missing variables
Model 2	MNL	Yes	No	No missing variables, indicator variables entering directly

- ▶ Next we analyze the extend of error arising due to:
 - ▶ Endogeneity and measurement bias jointly
 - ▶ Endogeneity bias and ignoring the preference heterogeneity
 - ▶ Endogeneity bias

	Model type	Measurement error	Endogeneity	Description
Model 3	MXL	Yes	Yes	missing, random, indicator variables entering directly
Model 4	Hybrid MNL	Controlled	Yes	missing
Model 5	Hybrid MXL	Controlled	Yes	missing, random

- ▶ These are models which are most likely to be used by researchers

- ▶ Lastly, we control for endogeneity using two different methods:
 - ▶ Directly modeling the correlation between latent factor and random parameters
 - ▶ Incorporating additional latent variable to account for correlation between error terms

	Model type	Measurement error	Endogeneity	Description
Model 6	Hybrid MXL	Controlled	Controlled	missing, random, correlation between and allowed
Model 7	Hybrid MNL	Controlled	Controlled	missing, additional LV in model specification

► LV endogeneity, base models

Variable	Parameter	True value of the parameter	Model 1	Model 2
<i>SQ</i> (constant)	α_{11}	-4.0000	-4.0570**	-3.0065
<i>SQ</i> (std. dev.)		2.0000	-	-
<i>SQ</i> (<i>LV</i>)	α_{12}	-2.0000	-2.0207**	-
<i>SQ</i> (<i>LV</i> ₂)		-2.0000	-	-
<i>SQ</i> (X^{Miss})	α_{13}	-2.0000	-2.0348**	-1.7178
<i>SQ</i> ($\cdot I_1$)		-2.0000	-	-0.7393
<i>SQ</i> ($\cdot I_2$)		2.0000	-	0.6912
<i>Quality</i> (constant)	α_{21}	5.0000	5.0353**	4.4806
<i>Quality</i> (<i>LV</i>)	α_{22}	2.0000	2.0051**	-
<i>Quality</i> ($\cdot I_1$)		2.0000	-	0.8744
<i>Quality</i> ($\cdot I_2$)		-2.0000	-	-0.8981
<i>Cost</i> (constant)	α_{31}	-3.0000	-3.0193**	-2.6749
<i>Cost</i> (<i>LV</i>)	α_{32}	1.0000	1.0267**	-
<i>Cost</i> ($\cdot I_1$)		1.0000	-	0.4073
<i>Cost</i> ($\cdot I_2$)		-1.0000	-	-0.3847

► LV endogeneity, usually used models

Variable	Parameter	True value of the parameter	Model 3	Model 4	Model 5
<i>SQ</i> (constant)	α_{11}	-4.0000	-4.1020**	-3.4864	-4.0366**
<i>SQ</i> (std. dev.)		2.0000	2.6399	-	2.2190
<i>SQ</i> (<i>LV</i>)	α_{12}	-2.0000	-	-2.577	-2.7313
<i>SQ</i> (<i>LV</i> ₂)		-2.0000	-	-	-
<i>SQ</i> (X^{Miss})	α_{13}	-2.0000	-	-	-
<i>SQ</i> ($\cdot I_1$)		-2.0000	-1.5334	-	-
<i>SQ</i> ($\cdot I_2$)		2.0000	0.0002	-	-
<i>Quality</i> (constant)	α_{21}	5.0000	4.7775**	4.9257**	5.0186***
<i>Quality</i> (<i>LV</i>)	α_{22}	2.0000	-	2.1344*	1.9910**
<i>Quality</i> ($\cdot I_1$)		2.0000	0.8687	-	-
<i>Quality</i> ($\cdot I_2$)		-2.0000	0.0006	-	-
<i>Cost</i> (constant)	α_{31}	-3.0000	-2.9073**	-2.9018**	-3.0144**
<i>Cost</i> (<i>LV</i>)	α_{32}	1.0000	-	0.8877	1.0228**
<i>Cost</i> ($\cdot I_1$)		1.0000	0.4960	-	-
<i>Cost</i> ($\cdot I_2$)		-1.0000	0.0110	-	-

► LV endogeneity, correcting for endogeneity

Variable	Parameter	True value of the parameter	Model 6	Model 7
<i>SQ</i> (constant)	α_{11}	-4.0000	-4.0386**	-4.0423**
<i>SQ</i> (std. dev.)		2.0000	1.9975**	-
<i>SQ</i> (<i>LV</i>)	α_{12}	-2.0000	-1.9934**	-1.8143
<i>SQ</i> (<i>LV</i> ₂)		-2.0000	-	-2.8433
<i>SQ</i> (X^{Miss})	α_{13}	-2.0000	-	-
<i>SQ</i> ($\cdot I_1$)		-2.0000	-	-
<i>SQ</i> ($\cdot I_2$)		2.0000	-	-
<i>Quality</i> (constant)	α_{21}	5.0000	5.0228***	5.0366**
<i>Quality</i> (<i>LV</i>)	α_{22}	2.0000	2.0098**	1.8125
<i>Quality</i> ($\cdot I_1$)		2.0000	-	-
<i>Quality</i> ($\cdot I_2$)		-2.0000	-	-
<i>Cost</i> (constant)	α_{31}	-3.0000	-3.0139***	-3.0220**
<i>Cost</i> (<i>LV</i>)	α_{32}	1.0000	1.0206**	0.9344*
<i>Cost</i> ($\cdot I_1$)		1.0000	-	-
<i>Cost</i> ($\cdot I_2$)		-1.0000	-	-

► M endogeneity, base models

Variable	Parameter	True value of the parameter	Model 1	Model 2
<i>SQ</i> (constant)	α_{11}	-4.0000	-4.0156**	-2.9964
<i>SQ</i> (std. dev.)		2.0000	-	-
<i>SQ</i> (LV)	α_{12}	-2.0000	-1.9873**	-
<i>SQ</i> (LV_2)		-2.0000	-	-
<i>SQ</i> (X^{Miss})	α_{13}	-2.0000	-2.0090**	-0.3479
<i>SQ</i> ($\cdot I_1$)		-2.0000	-	-1.1991
<i>SQ</i> ($\cdot I_2$)		2.0000	-	0.7840
<i>Quality</i> (constant)	α_{21}	5.0000	4.9947***	4.4629
<i>Quality</i> (LV)	α_{22}	2.0000	2.0050**	-
<i>Quality</i> ($\cdot I_1$)		2.0000	-	1.4961
<i>Quality</i> ($\cdot I_2$)		-2.0000	-	-1.0484
<i>Cost</i> (constant)	α_{31}	-3.0000	-3.0040***	-2.6792
<i>Cost</i> (LV)	α_{32}	1.0000	1.0054**	-
<i>Cost</i> ($\cdot I_1$)		1.0000	-	0.6897
<i>Cost</i> ($\cdot I_2$)		-1.0000	-	-0.4089

► M endogeneity, usually used models

Variable	Parameter	True value of the parameter	Model 3	Model 4	Model 5
<i>SQ</i> (constant)	α_{11}	-4.0000	-4.0666**	-3.9291**	-3.9508**
<i>SQ</i> (std. dev.)		2.0000	1.9103*	-	0.3076
<i>SQ</i> (LV)	α_{12}	-2.0000	-	-2.5606	-2.5659
<i>SQ</i> (LV_2)		-2.0000	-	-	-
<i>SQ</i> (X^{Miss})	α_{13}	-2.0000	-	-	-
<i>SQ</i> ($\cdot I_1$)		-2.0000	-1.8777*	-	-
<i>SQ</i> ($\cdot I_2$)		2.0000	1.2132	-	-
<i>Quality</i> (constant)	α_{21}	5.0000	4.3937	4.6903*	4.6909*
<i>Quality</i> (LV)	α_{22}	2.0000	-	1.7203	1.7130
<i>Quality</i> ($\cdot I_1$)		2.0000	-0.4716	-	-
<i>Quality</i> ($\cdot I_2$)		-2.0000	-1.6357	-	-
<i>Cost</i> (constant)	α_{31}	-3.0000	-2.7522*	-2.8764**	-2.8788**
<i>Cost</i> (LV)	α_{32}	1.0000	-	0.8606	0.8639
<i>Cost</i> ($\cdot I_1$)		1.0000	-0.1664	-	-
<i>Cost</i> ($\cdot I_2$)		-1.0000	-0.8642	-	-

► M endogeneity, correcting for endogeneity

Variable	Parameter	True value of the parameter	Model 6	Model 7
<i>SQ</i> (constant)	α_{11}	-4.0000	-3.9427**	-4.0194**
<i>SQ</i> (std. dev.)		2.0000	0.9762	-
<i>SQ</i> (LV)	α_{12}	-2.0000	-2.9241	-1.9582**
<i>SQ</i> (LV_2)		-2.0000	-	-2.1173*
<i>SQ</i> (X^{Miss})	α_{13}	-2.0000	-	-
<i>SQ</i> ($\cdot I_1$)		-2.0000	-	-
<i>SQ</i> ($\cdot I_2$)		2.0000	-	-
<i>Quality</i> (constant)	α_{21}	5.0000	4.6312*	4.9841***
<i>Quality</i> (LV)	α_{22}	2.0000	1.6474	1.9898**
<i>Quality</i> ($\cdot I_1$)		2.0000	-	-
<i>Quality</i> ($\cdot I_2$)		-2.0000	-	-
<i>Cost</i> (constant)	α_{31}	-3.0000	-2.8530*	-3.0027***
<i>Cost</i> (LV)	α_{32}	1.0000	0.8385	1.0010**
<i>Cost</i> ($\cdot I_1$)		1.0000	-	-
<i>Cost</i> ($\cdot I_2$)		-1.0000	-	-

- ▶ Currently used Hybrid Choice models do not account for the endogeneity of indicator variables
- ▶ Measurement bias can be substantial
 - ▶ Even with continuous indicator variables
- ▶ Possible solutions
 - ▶ Allowing for correlation between error terms in structural equations and choice model may help
 - ▶ Additional Latent Variables to capture residual correlation
 - ▶ Identification may be impossible, particularly with the two-step estimation procedure
 - ▶ The former does not work with M endogeneity