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

Wiktor Budziński,

Mikołaj Czajkowski

- 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 heterogeneity

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

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

$$\boldsymbol{\beta}_i = \boldsymbol{\Lambda}\mathbf{LV}_i + \boldsymbol{\Omega}\mathbf{SD}_i + \boldsymbol{\beta}_i^*$$

  - Structural equations:

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

  - Measurement equations

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

- 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
  - ME-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)
- 100 repetitions

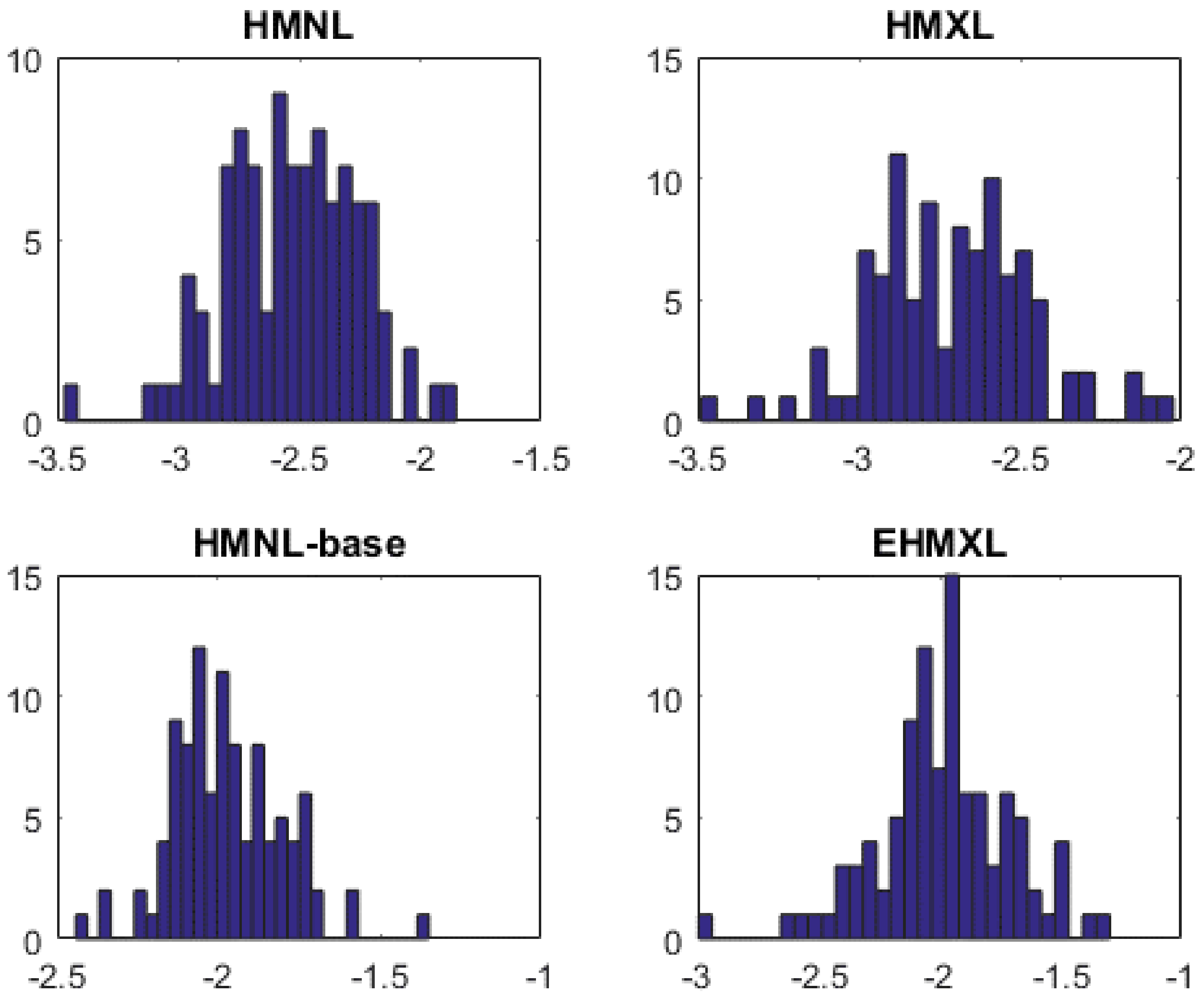
Data generating process:

	LV-endogeneity	ME-endogeneity
Utility function	$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$ $\beta_{1i} = -4 - 2LV_i^{norm} - 2X_i^{Miss}$ $\beta_{2i} = 5 + 2LV_i^{norm}$ $\beta_{3i} = -3 + 1LV_i^{norm}$	$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$ $\beta_{1i} = -4 - 2LV_i^{norm} - 2X_i^{Miss}$ $\beta_{2i} = 5 + 2LV_i^{norm}$ $\beta_{3i} = -3 + 1LV_i^{norm}$
Structural equations	$LV_i = -2X_i^{SD} + 1X_i^{Miss} + \xi_i$	$LV_i = -2X_i^{SD} + \xi_i$
Measurement equations	$I_{i1} = -1 + 1LV_i^{norm} + 0.5\eta_{i1}$ $I_{i2} = 1 - 1LV_i^{norm} + 0.5\eta_{i2}$	$I_{i1} = -1 + 1LV_i^{norm} + 1.5X_i^{Miss} + 0.5\eta_{i1}$ $I_{i2} = 1 - 1LV_i^{norm} - 0.5X_i^{Miss} + 0.5\eta_{i2}$

- Estimated models:

HMNL-base	The same specification as in DGP	No missing variables
MNL-base	Including indicator variables directly into the choice model	No missing variables
HMNL	The same specification as in DGP	X_i^{Miss} is missing
HMXL	The same specification as in DGP + random parameter for SQ	X_i^{Miss} is missing
EHMXL	The same specification as in DGP + random parameter for SQ + correlation between random parameter and ξ_i	X_i^{Miss} is missing
HMNL2	The same specification as in DGP + second LV in both measurement equations	X_i^{Miss} is missing
MXL	Including indicator variables directly into the choice model + random parameter for SQ	X_i^{Miss} is missing

Variable	True value	HMNL-base	MNL-base	HMNL	HMXL	EHMXL	MXL
Utility function							
SQ	-4	-3.9949* [-4.5009 -3.6244]	-2.9439 [-3.2650 -2.6970]	-3.4352* [-4.0052 -2.9512]	-4.0072* [-4.5648 -3.4461]	-3.9974* [-4.6198 -3.4770]	-4.0674* [-4.6970 -3.5655]
Quality	5	5.0086* [4.7234 5.3378]	4.4587 [4.2042 4.7083]	4.8979* [4.6343 5.2492]	4.9892* [4.7224 5.3371]	4.9990* [4.7320 5.3502]	4.7593* [4.5070 5.0488]
Cost	-3	-3.0028* [-3.2166 -2.8111]	-2.661 [-2.8241 -2.4999]	-2.8864* [-3.0828 -2.7077]	-3.0001* [-3.1969 -2.8215]	-3.0003* [-3.2093 -2.8171]	-2.8945* [-3.0963 -2.7162]
SQ x Miss (or RP)	-2	-2.0183* [-2.4011 -1.7051]	-1.7025* [-2.0087 -1.3931]	-	2.2335* [1.8243 2.6402]	2.0182* [1.5191 2.4058]	2.6393 [2.3189 2.9859]
SQ x LV (or I_1)	-2	-1.9716* [-2.3439 -1.5921]	-0.606 [-0.8884 -0.3111]	-2.5318 [-3.0434 -2.0386]	-2.7113 [-3.2260 -2.1443]	-1.9840* [-2.5466 -1.4656]	-1.3291 [-1.8569 -0.8997]
Quality x LV (or I_1)	2	2.0108* [1.7407 2.3351]	0.8114 [0.5385 1.0257]	2.1267* [1.8867 2.4722]	1.9849* [1.7443 2.3580]	2.0038* [1.7651 2.3573]	0.803 [0.5394 1.0548]
Cost x LV (or I_1)	1	1.0031* [0.8201 1.2025]	0.3549 [0.1610 0.5847]	0.8707* [0.7097 1.0437]	1.0053* [0.8192 1.1917]	1.0021* [0.8083 1.1980]	0.4351 [0.2278 0.6603]
SQ x I_2		-	0.6212 [0.3426 1.0147]	-	-	-	1.3251 [0.8985 1.8941]
Quality x I_2		-	-0.7745 [-1.0423 -0.4355]	-	-	-	-0.7731 [-1.0654 -0.4604]
Cost x I_2		-	-0.3424 [-0.5750 -0.1700]	-	-	-	-0.4228 [-0.6611 -0.1959]


HMNL
10
5
0
-3.5
-3
-2.5
-2
-1.5
HMXL
15
10
5
0
-3.5
-3
-2.5
-2
HMNL-base
15
10
5
0
-2.5
-2
-1.5
-1
EHMXL
15
10
5
0
-3
-2.5
-2
-1.5
-1

Variable	True value	HMNL-base	MNL-base	HMNL	HMXL	EHMXL	HMNL2	MXL
Utility function								
SQ	-4	-3.9906* [-4.4414 -3.4806]	-2.9831 [-3.3499 -2.6748]	-3.9218* [-4.4117 -3.4665]	-3.9421* [-4.4120 -3.4749]	-3.9350* [-4.3918 -3.4995]	-4.0098* [-4.5693 -3.5050]	-4.0823* [-4.5824 -3.6172]
Quality	5	5.0169* [4.6879 5.2771]	4.4738 [4.2009 4.7062]	4.7127 [4.4286 4.9429]	4.7135 [4.4325 4.9391]	4.6508 [4.3286 4.8903]	5.0107* [4.6596 5.3122]	4.4063 [4.1470 4.6497]
Cost	-3	-3.0031* [-3.2032 -2.8169]	-2.6774 [-2.8503 -2.5095]	-2.8775* [-3.0397 -2.7008]	-2.8799* [-3.0383 -2.7010]	-2.8536* [-3.0257 -2.6736]	-3.0036* [-3.1903 -2.7977]	-2.7533 [-2.8974 -2.6019]
SQ x Miss (or RP/LV2)	-2	-2.0090* [-2.3235 -1.7493]	-0.3265 [-0.7444 0.0572]	-	0.8616 [0.4942 1.7094]	0.9866 [0.4703 1.6758]	-2.1307* [-2.4328 -1.7481]	1.9289* [1.6231 2.2177]
SQ x LV (or I_1)	-2	-2.0031* [-2.4012 -1.5482]	-0.6528 [-0.9526 -0.3617]	-2.578 [-3.0498 -2.0381]	-2.5824 [-3.0499 -2.0591]	-2.9621 [-3.5008 -2.3941]	-1.9657* [-2.4395 -1.5098]	-1.0035 [-1.2234 -0.8042]
Quality x LV (or I_1)	2	2.0058* [1.6895 2.2645]	0.8301 [0.5990 1.1070]	1.717 [1.4181 1.9407]	1.7105 [1.4175 1.9421]	1.6407 [1.3262 1.9075]	1.9972* [1.6629 2.2792]	-0.2308 [-0.3739 -0.0608]
Cost x LV (or I_1)	1	0.9839* [0.8243 1.1671]	0.3443 [0.1787 0.5617]	0.8438* [0.6379 1.0020]	0.8468* [0.6394 1.0105]	0.8228 [0.6216 0.9822]	0.9826* [0.7997 1.1662]	-0.0916 [-0.2074 0.0193]
SQ x I_2		-	0.6457 [0.2516 0.9509]	-	-	-	-	1.019 [0.6110 1.4040]
Quality x I_2		-	-0.8143 [-1.0563 -0.5328]	-	-	-	-	-1.2968 [-1.5695 -0.9740]
Cost x I_2		-	-0.3444 [-0.5482 -0.0897]	-	-	-	-	-0.6996 [-0.8788 -0.5243]

- Currently used Hybrid Choice models do not account for the endogeneity of indicator variables
- Measurement bias can be substantial
  - Even with continuous indicator variables
  - In some instances endogeneity bias can correct measurement bias
- 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