Hybrid Choice model

Mathematically, two components of hybrid choice model, SEM (Equations 2, 3 and 4) and DCM (Equations 1 and 5), includes separate equations for representing structural and measurement relationship between exogenous and endogenous variables respectively:

$$U_n = B\boldsymbol{x_n} + L\boldsymbol{x_n^*} + \varepsilon_n \quad (1)$$

$$\boldsymbol{x_n^*} = A\boldsymbol{x_n} + \gamma_n \quad (2)$$

$$\boldsymbol{i_{nr}^*} = D\boldsymbol{x_n^*} + \eta_n \quad (3)$$

$$i_{nr} = \begin{cases} 1 \quad if\, \boldsymbol{i_{nr}^*} \le \tau_1 \\ 2 \quad if\ \tau_1 < \boldsymbol{i_{nr}^*} \le \tau_2 \\ \dots \\ j\ if \boldsymbol{i_{nr}^*} > \tau_{j-1} \end{cases} \quad (4)$$

$$y_n = \begin{cases} o = 1 \quad if U_n \le \mu_1 \\ o = 2 \quad if \mu_1 < U_n \le \mu_2 \\ \dots \\ o = O\ if U_n > \mu_{o-1} \end{cases} \quad (5)$$

Equation 1 represents structural equations for DCM framework where *U* represents utility for each individual *n* ($n \in N$) explained by vector $\boldsymbol{x_n}$ ($K \times 1$) consisting of *K* observable explanatory variables, vector $\boldsymbol{x_n^*}(M \times 1)$ consisting of *M* unobserved latent variables identified from likert scale variables and error terms ε_n, assumed to be independently and identically distributed (i.i.d.) logistically distributed with Σ_ε as the covariance matrix. *B* and *L* are the matrices with coefficients of explanatory variables ($1 \times K$) and latent variables ($1 \times M$).

Equation 2 represents the structural equation for SEM framework to calculate unobserved latent variable $\boldsymbol{x_n^*}$ described by explanatory variables $\boldsymbol{x_n}$ ($K \times 1$) with their coefficient matrix *A* ($M \times K$), reflecting the effect of $\boldsymbol{x_n}$over latent variables. γ_n is the vector ($M \times 1$) of error terms assumed to be i.i.d. normally distributed with φ as the covariance matrix. Many terms in $\boldsymbol{x_n}$ may be zero depending upon their association with latent variables.

Equation 3 represents the measurement equation for the SEM framework based on a vector of random variable $\boldsymbol{i_{nr}^*}$ ($R \times 1$) assumed to be normally distributed and discrete in nature (Likert scale with J levels) for each indicator ($r \in R$) and individual n. The indicators are based on the vector of latent variables, $\boldsymbol{x_n^*}(M \times 1)$, estimated from equation 2 and matrix *D* ($R \times M$) capturing the effect of the latent variables on indicators. η_n is the vector ($R \times 1$) of error terms assumed to be i.i.d. normally distributed with ψ as the covariance matrix. Some terms in $\boldsymbol{x_n^*}$ may be zero depending upon the association of latent variables with the indicators. This association is identified using EFA assuming the cut-off value of 0.4. In Equation 4, the random variable $\boldsymbol{i_{nr}^*}$ is measured based on the observed vector of indicators and certain thresholds τ_{j-1} based on ordinal probit kernel where ($j \in J$). All the error terms ($\varepsilon_n, \gamma_n\ and\ \eta_n$) are assumed to be mutually independent. In this study, survey questions utilized 7- level Likert scale.

Equation 5 represents measurement equation for DCM framework, based on ordinal logit kernel, as the dependent variable, y, is categorical with three ordered categories (*O*) and measured from utility *U*, calculated in Equation 1, and certain thresholds μ_{O-1}.

1.1 Estimation using Maximum simulated likelihood.

Hybrid choice models can be estimated in two steps, i.e., sequentially and simultaneously. In sequential estimation, SEM framework is estimated first, which enables the flexibility of embedding the estimated latent variables into the DCM framework and then DCM is estimated traditionally, maximizing the likelihood function conditional on explanatory and latent variables. In simultaneous estimation, both SEM and DCM modeling frameworks are estimated together where the likelihood function is conditional on the explanatory, latent and indicator variables (estimating all the four equations 1 to 5 jointly)

The first term of integrand represents structural equation of DCM, second term represents the measurement equation of SEM and the third term represents the structural equation of latent variable, and the joint probability of both equations is integrated over a vector of the latent construct x^* as the latent variables follow this distribution:

$$\mathcal{L}(y_n|\boldsymbol{x_n}, \boldsymbol{x_n^*}; B, L, \Sigma_\varepsilon, D, A, \psi, \varphi) = \int_{x^*} f_y(y_n|\boldsymbol{x_n}, \boldsymbol{x_n^*}; B, L, \Sigma_\varepsilon) f_{i^*}(\boldsymbol{i_{nr}^*}|\boldsymbol{x_n^*}; D, \psi)\, f_{x^*}(\boldsymbol{x_n^*}|\boldsymbol{x_n}; A, \varphi)\mathrm{d}x^* \quad (6)$$

The first term of integrand represents structural equation of DCM, second term represents the measurement equation of SEM and the third term represents the structural equation of latent variable, and the joint probability of all equations equations is integrated over a vector of the latent construct x^* as the latent variables follow this distribution:

Density function f_y is estimated **as ordinal logit kernel** based on Equation 5. The integral in Equation 6 can be evaluated using the Monte Carlo simulation method with 150 Halton draws, and then the resulting likelihood was estimated using maximum simulated likelihood (MSL).