Online supplement to

"Pooled Versus Private Ride-hailing: A Joint Revealed and Stated Preference Analysis Recognizing Psycho-Social Factors"

By Shuqing Kang, Aupal Mondal, Aarti C. Bhat, and Chandra R. Bhat (corresponding author)

Table 1. Loading of Latent Constructs on Indicators (MEM)

Indicators	Tech-savviness		Sharing Propensity		GLP	
	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat
I like to be among the first to have the latest technology.	0.537	9.059				
Learning how to use new technologies is often frustrating for me. (inverse scale)	0.513	8.804				
Having internet connectivity everywhere I go is important to me.	0.356	6.864				
I like trying things that are new and different.	0.383	6.757				
I feel uncomfortable around people I do not know. (inverse scale)			0.349	8.632		
Traveling with a driver I don't know makes me feel uncomfortable. (inverse scale)			1.793	8.612		
For shared ride-hailing (e.g., uberPOOL, Lyft Share), traveling with unfamiliar passengers makes me uncomfortable. (inverse scale)			1.528	10.667		
Sharing my personal information or location via internet-enabled devices concerns me a lot. (inverse scale)			0.226	5.399		
I am concerned that my travel logs and personal information stored in AVs could be leaked. (inverse scale)			0.246	6.128		
The government should raise the gas tax to help reduce the negative impacts of transportation on the environment.					0.554	10.671
I am committed to an environmentally-friendly lifestyle.					0.938	9.917
I am committed to using a less polluting means of transportation (e.g., walking, biking, and public transit) as much as possible.					1.302	8.031

Table 2 ATE Table for Pooled RH -- Shopping Purpose

Variable	Base Level	Treatment Level	% Contribution by mediation through						Overall ATE
			RH familiarity direct effect	RH familiarity sharing propensity increase	Tech-savviness decrease	Sharing propensity increase	GLP increase	Pooled RH choice direct effect	
Pooled RH interest for the shopping purpose									
Socio-demographic									
Gender	Female	Male	0	45	-34	19	-2	0	0.019
Age	18-24	55+	-80	0	19	0	-1	0	-0.213
Race/Ethnicity	Other races	Non-Hispanic/Non-Latino White	-37	-14	0	-4	0	-45	-0.086
Education	High school or less	Graduate degree	61	0	0	0	3	36	0.129
Employment	Unemployed	Employed	0	71	0	29	0	0	0.020
Tenure	Own or other	Rent	100	0	0	0	0	0	0.112
Household income	< $150,000	≥ $150,000	65	0	-30	0	-5	0	0.026
Built environment									
Living environment	Urban/suburban	Rural	-100	0	0	0	0	0	-0.084
Transit access	Transit access	No transit access	-100	0	0	0	0	0	-0.067
Population density	Low	High	0	0	0	0	0	100	0.040
Trip level attributes									
Travel time	Current time	Decrease by 5 mins	-	-	-	-	-	100	0.026
Travel cost	Current cost	Decrease by $1	-	-	-	-	-	100	0.017
Additional passenger	Current scenario	1 additional passenger	-	-	-	-	-	-100	-0.032

Table 3 ATE Table for Pooled RH -- Leisure Purpose

Variable	Base Level	Treatment Level	% Contribution by mediation through						Overall ATE
			RH familiarity direct effect	RH familiarity sharing propensity increase	Tech-savviness decrease	Sharing propensity increase	GLP increase	Pooled RH choice direct effect	
Pooled RH interest for the leisure purpose									
Socio-demographic									
Gender	Female	Male	0	36	-50	10	-4	0	-0.006
Age	18-24	55+	-72	0	26	0	-2	0	-0.171
Race/Ethnicity	Other races	Non-Hispanic/Non-Latino White	-33	-13	0	-3	0	-51	-0.080
Education	High school or less	Graduate degree	55	0	0	0	6	39	0.125
Employment Status	Unemployed	Employed	0	80	0	20	0	0	0.015
Tenure type	Own or other	Rent	100	0	0	0	0	0	0.096
Income	< $150,000	≥ $150,000	52	0	-40	0	-8	0	0.003
Built environment									
Living environment	Urban/suburban	Rural	-100	0	0	0	0	0	-0.109
Transit access	Transit access	No transit access	-100	0	0	0	0	0	-0.058
Population density	Low	High	0	0	0	0	0	100	0.045
Trip level attributes									
Travel time	Current time	Decrease by 5 mins	-	-	-	-	-	100	0.021
Travel cost	Current cost	Decrease by $1	-	-	-	-	-	100	0.023
Additional passenger	Current scenario	1 additional passenger	-	-	-	-	-	-100	-0.031

Mathematical formulation of the GHDM for the current study

Since the main outcome variables are all binary models, they can be modeled as ordinal variables as well (with 0 and 1 as the ordered levels). Given all the indicators are ordinal in nature, the GHDM model is formulated with only ordinal outcomes.

Consider the case of an individual $q \in \{1, 2, ..., Q\}$. Let $l \in \{1, 2, ..., L\}$ be the index of the latent constructs and let z_{ql}^* be the value of the latent variable *l* for the individual *q*. z_{ql}^* is expressed as a function of its explanatory variables as,

$$z_{ql}^* = \boldsymbol{w}_{ql}^{\mathrm{T}} \boldsymbol{\alpha} + \eta_{ql}, \tag{1}$$

where $\boldsymbol{w}_{ql}$ $(D \times 1)$ is a column vector of the explanatory variables of latent variable *l* and $\boldsymbol{\alpha}$ $(D \times 1)$ is a vector of its coefficients. η_{ql} is the unexplained error term and is assumed to follow a standard normal distribution. Equation (1) can be expressed in the matrix form as,

$$\boldsymbol{z}_q^* = \boldsymbol{w}_q \boldsymbol{\alpha} + \boldsymbol{\eta}_q, \tag{2}$$

where $\boldsymbol{z}_q^*$ $(L \times 1)$ is a column vector of all the latent variables, $\boldsymbol{w}_q$ $(L \times D)$ is a matrix formed by vertically stacking the vectors $(\boldsymbol{w}_{q1}^{\mathrm{T}}, \boldsymbol{w}_{q2}^{\mathrm{T}}, ..., \boldsymbol{w}_{qL}^{\mathrm{T}})$ and $\boldsymbol{\eta}_q$ $(D \times 1)$ is formed by vertically stacking $(\eta_{q1}, \eta_{q2}, ..., \eta_{qL})$. $\boldsymbol{\eta}_q$ follows a multivariate normal distribution centered at the origin and having a correlation matrix of $\boldsymbol{\Gamma}$ $(L \times L)$, i.e., $\boldsymbol{\eta}_q \sim MVN_L(\boldsymbol{0}_L, \boldsymbol{\Gamma})$, where $\boldsymbol{0}_L$ is a vector of zeros. The variance of all the elements in $\boldsymbol{\eta}_q$ is fixed as unity because it is not possible to uniquely identify a scale for the latent variables. Equation (2) constitutes the SEM component of the framework.

Let $j \in \{1, 2, ..., J\}$ denote the index of the outcome variables (including the indicator variables). Let y_{qj}^* be the underlying continuous measure associated with the outcome variable y_{qj}. Then,

$$y_{qj} = k \text{ if } t_{jk} < y_{qj}^* \le t_{j(k+1)}, \tag{3}$$

where $k \in \{1, 2, ..., K_j\}$ denotes the ordinal category assumed by y_{qj} and t_{jk} denotes the lower boundary of the k^{th} discrete interval of the continous measure associated with the j^{th} outcome. $t_{jk} < t_{j(k+1)}$ for all *j* and all *k*. Since y_j^* may take any value in $(-\infty, \infty)$, we fix the value of $t_{j1} = -\infty$ and $t_{j(K_j+1)} = \infty$ for all *j*. Since the location of the thresholds on the real-line is not uniquely identifiable, we also set $t_{j2} = 0$. y_j^* is expressed as a function of its explanatory variables as,

$$y_{qj}^* = \boldsymbol{x}_{qj}^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{z}_q^{*\mathrm{T}} \boldsymbol{d}_j + \xi_{qj}, \tag{4}$$

where $\boldsymbol{x}_{qj}$ $(E \times 1)$ is a vector of size of explanatory variables for the continuous measure y_{qj}^*. $\boldsymbol{\beta}$ $(E \times 1)$ is a column vector of the coefficients associated with $\boldsymbol{x}_{qj}$ and $\boldsymbol{d}_j$ $(L \times 1)$ is the vector of coefficients of the latent variables for outcome *j*. ξ_{qj} is a stochastic error term that captures the effect of unobserved variables on y^*_{qj}. ξ_{qj} is assumed to follow a standard normal distribution. Jointly, the continuous measures of the *J* outcome variables may be expressed as,

$$\boldsymbol{y}^*_q = \boldsymbol{x}_q\boldsymbol{\beta} + \boldsymbol{d}\boldsymbol{z}^*_q + \boldsymbol{\xi}_q, \tag{5}$$

where $\boldsymbol{y}^*_q$ $(J \times 1)$ and $\boldsymbol{\xi}_q$ $(J \times 1)$ are the vectors formed by vertically stacking y^*_{qj} and ξ_{qj}, respectively, of the *J* dependent variables. $\boldsymbol{x}_q$ $(J \times E)$ is a matrix formed by vertically stacking the vectors $\left(\boldsymbol{x}^{\mathrm{T}}_{q1}, \boldsymbol{x}^{\mathrm{T}}_{q2}, ..., \boldsymbol{x}^{\mathrm{T}}_{qJ}\right)$ and $\boldsymbol{d}$ $(J \times L)$ is a matrix formed by vertically stacking $\left(\boldsymbol{d}^{\mathrm{T}}_1, \boldsymbol{d}^{\mathrm{T}}_2, ..., \boldsymbol{d}^{\mathrm{T}}_J\right)$. $\boldsymbol{\xi}_q$ follows a multivariate normal distribution centered at the origin with an identity matrix as the covariance matrix (independent error terms). $\boldsymbol{\xi}_q \sim MVN_J(\boldsymbol{0}_J, \mathbf{I}_J)$. We assume the terms in $\boldsymbol{\xi}_q$ to be independent because it is not possible to uniquely identify all the correlations between the elements in $\boldsymbol{\eta}_q$ and all the correlations between the elements in $\boldsymbol{\xi}_q$. Further, because of the ordinal nature of the outcome variables, the scale of $\boldsymbol{y}^*_q$ cannot be uniquely identified. Therefore, the variances of all elements in $\boldsymbol{\xi}_q$ is fixed to one. The reader is referred to Bhat (2015) for further nuances regarding the identification of coefficients in the GHDM framework.

Substituting Equation (2) in Equation (5), $\boldsymbol{y}^*_q$ can be expressed in the reduced form as

$$\boldsymbol{y}^*_q = \boldsymbol{x}_q\boldsymbol{\beta} + \boldsymbol{d}\left(\boldsymbol{w}_q\boldsymbol{\alpha} + \boldsymbol{\eta}_q\right) + \boldsymbol{\xi}_q, \tag{6}$$

$$\boldsymbol{y}^*_q = \boldsymbol{x}_q\boldsymbol{\beta} + \boldsymbol{d}\boldsymbol{w}_q\boldsymbol{\alpha} + \boldsymbol{d}\boldsymbol{\eta}_q + \boldsymbol{\xi}_q. \tag{7}$$

In the right side of Equation (7), $\boldsymbol{\eta}_q$ and ξ_q are random vectors that follow the multivariate normal distribution and the other variables are constants. Therefore, $\boldsymbol{y}^*_q$ also follows the multivariate normal distribution with a mean of $\boldsymbol{b} = \boldsymbol{x}_q\boldsymbol{\beta} + \boldsymbol{d}\boldsymbol{w}_q\boldsymbol{\alpha}$ (all the elements of $\boldsymbol{\eta}_q$ and $\boldsymbol{\xi}_q$ have a mean of zero) and a covariance matrix of $\boldsymbol{\Sigma} = \boldsymbol{d}\boldsymbol{\Gamma}\boldsymbol{d}^{\mathrm{T}} + \mathbf{I}_J$.

$$\boldsymbol{y}^*_q \sim MVN_J(\boldsymbol{b}, \boldsymbol{\Sigma}). \tag{8}$$

The parameters that are to be estimated are the elements of $\boldsymbol{\alpha}$, strictly upper triangular elements of $\boldsymbol{\Gamma}$, elements of $\boldsymbol{\beta}$, elements of $\boldsymbol{d}$ and t_{jk} for all *j* and $k \in \{3, 4, ..., K_j\}$. Let $\boldsymbol{\theta}$ be a vector of all the parameters that need to be estimated. The maximum likelihood approach can be used for estimating these parameters. The likelihood of the q^{th} observation will be,

$$L_q(\boldsymbol{\theta}) = \int_{v_1 = t_{1y_{q1}} - b_1}^{v_1 = t_{1(y_{q1}+1)} - b_1} \int_{v_2 = t_{2y_{q2}} - b_2}^{v_2 = t_{2(y_{q2}+1)} - b_2} \cdots \int_{v_J = t_{Jy_{qJ}} - b_J}^{v_J = t_{J(y_{qJ}+1)} - b_J} \phi_J(v_1, v_2, \ldots, v_J \mid \boldsymbol{\Sigma}) dv_1 dv_2 \ldots dv_J, \tag{9}$$

where, $\phi_J\left(v_1, v_2, \ldots, v_J \mid \boldsymbol{\Sigma}\right)$ denotes the probability density of a *J* dimensional multivariate normal distribution centered at the origin with a covariance matrix $\boldsymbol{\Sigma}$ at the point $(v_1, v_2, \ldots, v_J)$. Since a closed form expression does not exist for this integral and evaluation using simulation techniques can be time consuming, we used the One-variate Univariate Screening technique proposed by Bhat (2018) for approximating this integral. The estimation of parameters was carried out using the *maxlik* library in the GAUSS matrix programming language.

References

Bhat, C.R., 2018. New matrix-based methods for the analytic evaluation of the multivariate cumulative normal distribution function. *Transportation Research Part B*, 109, 238-256.