Online Supplement to

"MaaS in Car-Dominated Cities:
Modeling the adoption, frequency, and characteristics of ride-hailing trips in Dallas, TX"

Patrícia S. Lavieri and Chandra R. Bhat (corresponding author)

An Overview of the Generalized Heterogeneous Data Model

Latent Variable Structural Equation Model

Let *l* be an index for latent variables (*l*=1,2,3,4). Consider the latent variable z_l^* and write it as a linear function of covariates:

$$z_l^* = \boldsymbol{\alpha}_l' \boldsymbol{w} + \eta_l, \quad (1)$$

where $\boldsymbol{w}$ is a $(\tilde{D} \times 1)$ vector of observed covariates (excluding a constant), $\boldsymbol{\alpha}_l$ is a corresponding $(\tilde{D} \times 1)$ vector of coefficients, and η_l is a random error term assumed to be standard normally distributed for identification purposes (See Bhat, 2015). Next, define the $(4 \times \tilde{D})$ matrix $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \boldsymbol{\alpha}_4)'$, and the (4×1) vectors $\boldsymbol{z}^* = (z_1^*, z_2^*, z_3^*, z_4^*)'$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3, \eta_4)'$. To accommodate interactions among the unobserved latent variables, we allow a MVN correlation structure for $\boldsymbol{\eta}$, that is $\boldsymbol{\eta} \sim MVN_4[\mathbf{0}_4, \boldsymbol{\Gamma}]$, whue $\boldsymbol{\Gamma}$ is (4×4) correlation matrix. A general covariance structure is adopted because in our study there are no conceptual reasons to establish causal relationships between the latent variables. In matrix form, we may write Equation (1) as:

$$\boldsymbol{z}^* = \boldsymbol{\alpha} \boldsymbol{w} + \boldsymbol{\eta}. \quad (2)$$

Latent Variable Measurement Equation Model Components

As mentioned earlier, we consider a combination of ordinal and nominal outcomes explained by a latent variable vector $\boldsymbol{z}^*$ and, when relevant, a set of other endogeneous and exogenous variables as well.

Consider *N* ordinal outcomes for the individual, and let *n* be the index for the ordinal outcomes $(n = 1, 2, ..., N)$, in our application *N*=13 for 906 individuals and *N*=12 for the remainder of the sample. Also, let J_n be the number of categories for the n^{th} ordinal outcome $(J_n \geq 2)$ and let the corresponding index be j_n $(j_n = 1, 2, ..., J_n)$. Let $\tilde{y}_n^*$ be the latent

underlying variable whose horizontal partitioning leads to the observed outcome for the n^{th} ordinal variable. Assume that the individual under consideration chooses the a_n^{th} ordinal category. Then, in the usual ordered response formulation, for the individual, we may write:

$$\tilde{y}_n^* = \tilde{\boldsymbol{\gamma}}_n' \boldsymbol{x} + \tilde{\boldsymbol{d}}_n' \boldsymbol{z}^* + \tilde{\varepsilon}_n, \text{ and } \tilde{\psi}_{n,a_n-1} < \tilde{y}_n^* < \tilde{\psi}_{n,a_n}, \quad (3)$$

where $\boldsymbol{x}$ is a vector of exogenous and possibly endogenous variables as defined earlier, $\tilde{\boldsymbol{\gamma}}_n$ is a corresponding vector of coefficients to be estimated, $\tilde{\boldsymbol{d}}_n$ is an (4×1) vector of latent variable loadings on the n^{th} continuous outcome, the $\tilde{\psi}$ terms represent thresholds, and $\tilde{\varepsilon}_n$ is the standard normal random error for the n^{th} ordinal outcome. For each ordinal outcome, $\tilde{\psi}_{n,0} < \tilde{\psi}_{n,1} < \tilde{\psi}_{n,2} \ldots < \tilde{\psi}_{n,J_n-1} < \tilde{\psi}_{n,J_n}$; $\tilde{\psi}_{n,0} = -\infty$, $\tilde{\psi}_{n,1} = 0$, and $\tilde{\psi}_{n,J_n} = +\infty$.

For later use, let $\tilde{\boldsymbol{\psi}}_n = (\tilde{\psi}_{n,2}, \tilde{\psi}_{n,3} \ldots, \tilde{\psi}_{n,J_n-1})'$ and $\tilde{\boldsymbol{\psi}} = (\tilde{\boldsymbol{\psi}}_1', \tilde{\boldsymbol{\psi}}_2', \ldots, \tilde{\boldsymbol{\psi}}_N)'$. Stack the N underlying continuous variables $\tilde{y}_n^*$ into an $(N\times 1)$ vector $\tilde{\boldsymbol{y}}^*$, and the N error terms $\tilde{\varepsilon}_n$ into another $(N\times 1)$ vector $\tilde{\boldsymbol{\varepsilon}}$. Define $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_1, \tilde{\boldsymbol{\gamma}}_2, \ldots, \tilde{\boldsymbol{\gamma}}_N)'$ [$(N\times A)$ matrix] and $\tilde{\boldsymbol{d}} = \left(\tilde{\boldsymbol{d}}_1, \tilde{\boldsymbol{d}}_2, \ldots, \tilde{\boldsymbol{d}}_N\right)$ [$(N\times 4)$ matrix], and let $\mathbf{IDEN_N}$ be the identity matrix of dimension N representing the correlation matrix of $\tilde{\boldsymbol{\varepsilon}}$ (so, $\tilde{\boldsymbol{\varepsilon}} \sim MVN_N(\mathbf{0}_N, \mathbf{IDEN_N})$; again, this is for identification purposes, given the presence of the unobserved $\boldsymbol{z}^*$ vector to generate covariance. Finally, stack the lower thresholds for the decision-maker $\tilde{\psi}_{n,a_n-1}(n = 1, 2, \ldots, N)$ into an $(N\times 1)$ vector $\tilde{\boldsymbol{\psi}}_{low}$ and the upper thresholds $\tilde{\psi}_{n,a_n}(n = 1, 2, \ldots, N)$ into another vector $\tilde{\boldsymbol{\psi}}_{up}$. Then, in matrix form, the measurement equation for the ordinal outcomes (indicators) for the decision-maker may be written as:

$$\tilde{\boldsymbol{y}}^* = \tilde{\boldsymbol{\gamma}}\boldsymbol{x} + \tilde{\boldsymbol{d}}\boldsymbol{z}^* + \tilde{\boldsymbol{\varepsilon}}, \quad \tilde{\boldsymbol{\psi}}_{low} < \tilde{\boldsymbol{y}}^* < \tilde{\boldsymbol{\psi}}_{up}. \quad (4)$$

Next, let there be G nominal (unordered-response) variables for an individual, and let g be the index for the nominal variables, in our application G=2. Also, let I_g be the number of alternatives corresponding to the g^{th} nominal variable ($I_g \geq 3$) and let i_g be the corresponding index. Both nominal outcomes in our application have I_g=3. Consider the g^{th} nominal variable and assume that the individual under consideration chooses the alternative m_g. Also, assume the usual random utility structure for each alternative i_g.

$$U_{gi_g} = \boldsymbol{b}_{gi_g}' \boldsymbol{x} + \boldsymbol{\vartheta}_{gi_g}' \boldsymbol{z}^* + \varsigma_{gi_g}, \quad (5)$$

where $\boldsymbol{x}$ is as defined earlier, $\boldsymbol{b}_{gi_g}$ is an $(A\times 1)$ column vector of corresponding coefficients, and ς_{gi_g} is a normal error term, and $\boldsymbol{\vartheta}_{gi_g}$ is an $(N_{gi_g}\times 1)$-column vector of coefficients capturing the effects of latent variables. Let $\boldsymbol{\varsigma_g}=(\varsigma_{g1},\varsigma_{g2},...\varsigma_{gI_g})'$ ($I_g\times 1$ vector), and $\boldsymbol{\varsigma_g}\sim MVN_{I_g}(\mathbf{0},\boldsymbol{\Lambda_g})$. Taking the difference with respect to the first alternative, the only estimable elements are found in the covariance matrix $\breve{\boldsymbol{\Lambda}}_g$ of the error differences, $\breve{\boldsymbol{\varsigma}}_{\boldsymbol{g}}=(\breve{\varsigma}_{g2},\breve{\varsigma}_{g3},...,\breve{\varsigma}_{gI_g})$ (where $\breve{\varsigma}_{gi}=\varsigma_{gi}-\varsigma_{g1},\ i\neq 1$). Further, the variance term at the top left diagonal of $\breve{\boldsymbol{\Lambda}}_g$ ($g=1,\ 2$) is set to 1 to account for scale invariance. $\boldsymbol{\Lambda_g}$ is constructed from $\breve{\boldsymbol{\Lambda}}_{\mathbf{g}}$ by adding a row on top and a column to the left. All elements of this additional row and column are filled with values of zero. In addition, the usual identification restriction is imposed such that one of the alternatives serves as the base when introducing alternative-specific constants and variables that do not vary across alternatives (that is, whenever an element of $\boldsymbol{x}$ is individual-specific and not alternative-specific, the corresponding element in $\boldsymbol{b}_{gi_g}$ is set to zero for at least one alternative i_g).

To proceed, define $\boldsymbol{U}_g=(U_{g1},U_{g2},U_{g3})'$ and $\boldsymbol{b}_g=(\boldsymbol{b}_{g1},\boldsymbol{b}_{g2},\boldsymbol{b}_{g3})'$. Also, define the $\left(I_g\times\sum_{i_g=1}^{I_g}N_{gi_g}\right)$ matrix $\boldsymbol{\vartheta}_{\boldsymbol{g}}$, which is initially filled with all zero values. Then, position the $(1\times N_{g1})$ row vector $\boldsymbol{\vartheta}'_{g1}$ in the first row to occupy columns 1 to N_{g1}, position the $(1\times N_{g2})$ row vector $\boldsymbol{\vartheta}'_{g2}$ in the second row to occupy columns N_{g1}+1 to $N_{g1}+N_{g2}$, and $\boldsymbol{\vartheta}'_{g3}$ to occupy columns N_{g2}+1 to $N_{g1}+N_{g2}+N_{g3}$. Further, define $\vec{G}=\sum_{g=1}^{G}I_g$, $\boldsymbol{U}=\left(\boldsymbol{U}'_1,\boldsymbol{U}'_2\right)'$ ($\vec{G}\times 1$ vector), $\boldsymbol{\varsigma}=(\boldsymbol{\varsigma}_1,\boldsymbol{\varsigma}_2,...\boldsymbol{\varsigma}_G)'$ ($\vec{G}\times 1$ vector), $\boldsymbol{b}=(\boldsymbol{b}'_1,\boldsymbol{b}'_2)'$ ($\vec{G}\times A$ matrix), and $\boldsymbol{\vartheta}=\text{Vech}(\boldsymbol{\vartheta}_1,\boldsymbol{\vartheta}_2)$ (that is, $\boldsymbol{\vartheta}$ is a column vector that includes all elements of the matrices $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$). Then, in matrix form, we may write Equation (3) as:

$$\boldsymbol{U}=\boldsymbol{bx}+\boldsymbol{\vartheta z}^{*}+\boldsymbol{\varsigma} \tag{6}$$

where $\boldsymbol{\varsigma}\sim MVN_{\vec{G}}(\mathbf{0}_{\vec{G}},\boldsymbol{\Lambda})$. As earlier, to ensure identification, we specify $\boldsymbol{\Lambda}$ as follows:

$$\boldsymbol{\Lambda}=\begin{bmatrix}\boldsymbol{\Lambda}_1 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}_2\end{bmatrix}\ (\vec{G}\times\vec{G}\ \text{matrix}). \tag{7}$$

In the general case, this allows the estimation of $\sum_{g=1}^{G}\left(\frac{I_g * (I_g - 1)}{2} - 1\right)$ terms across all the G nominal variables, as originating from $\left(\frac{I_g * (I_g - 1)}{2} - 1\right)$ terms embedded in each $\breve{\boldsymbol{\Lambda}}_g$ matrix.

To develop the reduced form equations, replace the right side of Equation (2) for $\boldsymbol{z}^*$ in Equations (4) and (6) to obtain the following system:

$$\tilde{\boldsymbol{y}}^* = \tilde{\boldsymbol{\gamma}}\boldsymbol{x} + \tilde{\boldsymbol{d}}\boldsymbol{z}^* + \tilde{\boldsymbol{\varepsilon}} = \tilde{\boldsymbol{\gamma}}\boldsymbol{x} + \tilde{\boldsymbol{d}}(\boldsymbol{\alpha}\boldsymbol{w} + \boldsymbol{\eta}) + \tilde{\boldsymbol{\varepsilon}} = \tilde{\boldsymbol{\gamma}}\boldsymbol{x} + \tilde{\boldsymbol{d}}\boldsymbol{\alpha}\boldsymbol{w} + \tilde{\boldsymbol{d}}\boldsymbol{\eta} + \tilde{\boldsymbol{\varepsilon}}, \tag{8}$$

$$\boldsymbol{U} = \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\vartheta}\boldsymbol{z}^* + \boldsymbol{\varsigma} = \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\vartheta}(\boldsymbol{\alpha}\boldsymbol{w} + \boldsymbol{\eta}) + \boldsymbol{\varsigma} = \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\vartheta}\boldsymbol{\alpha}\boldsymbol{w} + \boldsymbol{\vartheta}\boldsymbol{\eta} + \boldsymbol{\varsigma}. \tag{9}$$

Now, consider the $[(N + \vec{G}) \times 1)]$ vector $\boldsymbol{yU} = \left[\tilde{\boldsymbol{y}}^{*\prime}, \boldsymbol{U}'\right]'$. Define

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{y}}^* \boldsymbol{x} + \tilde{\boldsymbol{d}}\boldsymbol{\alpha}\boldsymbol{w} \\ \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\vartheta}\boldsymbol{\alpha}\boldsymbol{w} \end{bmatrix} \text{ and } \boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_1 & \boldsymbol{\Omega}_{12}' \\ \boldsymbol{\Omega}_{12} & \boldsymbol{\Omega}_2 \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{d}}\boldsymbol{\Gamma}\tilde{\boldsymbol{d}}' + \tilde{\boldsymbol{\Sigma}} & \tilde{\boldsymbol{d}}\boldsymbol{\Gamma}\boldsymbol{\vartheta}' \\ \boldsymbol{\vartheta}\boldsymbol{\Gamma}\tilde{\boldsymbol{d}}' & \boldsymbol{\vartheta}\boldsymbol{\Gamma}\boldsymbol{\vartheta}' + \boldsymbol{\Lambda} \end{bmatrix}. \tag{10}$$

Then $\boldsymbol{yU} \sim \boldsymbol{MVN}_{N+\tilde{\boldsymbol{G}}}(\boldsymbol{B}, \boldsymbol{\Omega})$.

The model estimation is performed using Bhat's (2011) MACML. We refer the reader to Bhat (2015) for the detailed explanation as well as information on model identification criteria.

REFERENCES

Bhat, C.R., 2011. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part* B, 45(7), 923-939.

Bhat, C.R., 2015. A new generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. *Transportation Research Part B,* 79, 50-77.

Table 1. Thresholds and constants of indicators and loadings of latent variables on indicators

Attitudinal and lifestyle indicators	Threshold 2		Threshold 3		Threshold 4		Constant		Latent variable loading	
	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat
Privacy-sensitivity										
I don't mind sharing a ride with strangers if it reduces my costs (inverse scale)	2.523	19.85	3.598	21.06	5.123	19.08	2.504	12.84	1.792	14.09
Having privacy is important to me when I make a trip	0.922	12.13	1.799	22.17	3.076	33.69	2.101	23.01	0.575	16.21
I feel uncomfortable sitting close to strangers	0.954	17.55	1.737	25.04	2.777	25.44	1.409	22.24	0.427	6.19
Tech-savviness										
I frequently use online banking services	1.133	8.67	2.606	18.136	4.099	28.56	2.559	12.83	1.601	55.44
I frequently purchase products online	0.506	6.475	1.017	11.17	1.849	19.27	1.861	14.69	0.681	26.15
Learning how to use new smartphone apps is easy for me	1.138	9.685	1.993	16.22	2.859	23.18	2.255	15.08	0.787	30.61
Variety-seeking lifestyle propensity (VSLP)										
I think it is important to have all sorts of new experiences and I am always trying new things	1.159	13.78	2.374	26.41	3.676	35.80	2.631	19.33	0.930	22.40
Looking for adventures and taking risks is important to me	1.195	2.45	2.468	2.33	3.834	2.17	1.739	2.67	1.033	23.83
I love to try new products before anyone else	0.910	6.67	1.859	7.37	2.934	7.38	1.908	6.88	0.704	2.69
Green lifestyle propensity (GLP)										
When choosing my commute mode, there are many things that are more important than being environmentally friendly (inverse scale)	1.045	15.37	1.860	16.49	2.746	15.00	0.988	12.66	0.158	1.84
I don't give much thought to saving energy at home (inverse scale)	0.708	10.87	1.182	16.44	2.203	25.18	1.910	21.34	0.132	1.80