**Online supplement for the paper**

**“****Sharing the Road with Autonomous Vehicles: Perceived Safety and Regulatory Preferences”**

Gopindra S. Nair, Chandra R. Bhat (corresponding author)

## Mathematical formulation of GHDM model with only ordinal outcomes

Consider the case of an individual . Let  be the index of the latent constructs and let  be the value of the latent variable *l* for the individual *q*.  is expressed as a function of its explanatory variables as,

,

where  is a column vector of the explanatory variables of latent variable *l* and is a vector of its coefficients.  is the unexplained error term and is assumed to follow a standard normal distribution. Equation (1) can be expressed in the matrix form as,

,

where is a column vector of all the latent variables, is a matrix formed by vertically stacking the vectors  and  is formed by vertically stacking .  follows a multivariate normal distribution centered at the origin and having a correlation matrix of , i.e., , where  is a vector of zeros. The variance of all the elements in  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 denote the index of the outcome variables (including the indicator variables). Let  be the underlying continuous measure associated with the outcome variable . Then,

,

where  denotes the ordinal category assumed by  and  denotes the lower boundary of the *k*th discrete interval of the continous measure associated with the *j*th outcome.  for all *j* and all *k*. Since  may take any value in , we fix the value of and  for all *j*. Since the location of the thresholds on the real-line is not uniquely identifiable, we also set .  is expressed as a function of its explanatory variables as,

,

where is a vector of size of explanatory variables for the continuous measure .  is a column vector of the coefficients associated with  and  is the vector of coefficeints of the latent variables for outcome *j*.  is a stochastic error term that captures the effect of unobserved variables on the value of .  is assumed to follow a standard normal distribution. Jointly, the continuous measures of the *J* outcome variables may be expressed as,

,

where  and  are the vectors formed by vertically stacking and  respectively of the *J* dependent variables.  is a matrix formed by vertically stacking the vectors  and  is a matrix formed by vertically stacking .  follows a multivariate normal distribution centered at the origin with an identity matrix as the covariance matrix (independent error terms). . We assume the terms in  to be independent because it is not possible to uniquely identify all the correlations between the elements in and all the correlations between the elements in . Further, because of the ordinal nature of the outcome variables, the scale of  cannot be uniquely identified. Therefore, the variances of all elements in  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),  can be expressed in the reduced form as,

,

.

In the R.H.S. of Equation (7),  and  are random vectors that follow the multivariate normal distribution and the other variables are constants. Therefore,  also follows the multivariate normal distribution with a mean of  (all the elements of  and  have a mean of zero) and a covariance matrix of.

.

The parameters that are to be estimated are the elements of , strictly upper triangular elements of **Γ**, elements of ***β***, elements of ***d*** and  for all *j* and . Let ***θ*** 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,

,

where,  denotes the probability density of a *J* dimensional multivariate normal distribution centered at the origin with a covariance matrix **Σ** at the point . 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.

## Output predictions

To predict the outcome for an individual, first the random error terms for the latent variables are drawn from a multivariate normal distribution that has a mean of  and a covariance matrix of **Γ**. Following this the latent variables are predicted using Equation (2). Then the random error terms of the outcome variables are drawn from independent standard normal distributions. The outcomes are then predicted in a sequential manner using Equation (3) and Equation (4). Since some of the outcome variables are used as explanatory variables for other outcome variables, the sequence used for estimating the outcomes is such that the outcome that is predicted will only have explanatory variables that are exogenous or that have been predicted in one of the previous steps.

## Average treatment effects

The expression for computing the ATE is as follows,

,

where  denotes the ATE on the *i*th level of the *j*th outcome variable by applying a treatment that changes condition of individuals from *A* to *B*, is a function that computes the probability for the individual *q*. Since the computation of this probability is cumbersome when the outcome variable under consideration is explained by other outcome variables that are also affected by the treatment, we use Monte Carlo methods to simulate the required probabilities. For approximating the probability, we make 1000 predictions for the individual *q* where each prediction is made using a different random draw. The process for making the prediction is the same as that described in the previous section except that the variables that are affected by the condition *B* is set as per the condition irrespective of the value generated by the prediction process. The estimate for  will then be the proportion of the number of cases where  assumes the value *i* to the total number of predictions.

## References

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.

https://doi.org/10.1016/j.trb.2015.05.017

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.

https://doi.org/10.1016/j.trb.2018.01.011

## Latent variable loadings and thresholds of indicators

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable (Minimum level – Maximum level)** | **Loading** | **Constant** | **Threshold 2|3** | **Threshold 3|4** |
| **Estimate** | **t-stat** | **Estimate** | **t-stat** | **Estimate** | **t-stat** | **Estimate** | **t-stat** |
| Tech-savviness indicators |   |   |   |   |   |   |   |   |
| Computer use frequency (Never/Almost never – Everyday) | 0.803 | 17.59 | 2.620 | 27.86 | 0.952 | 17.43 | 1.727 | 25.49 |
| Frequency of internet use (Less than once a day – Many times a day) | 0.853 | 18.83 | 1.794 | 24.99 | 1.272 | 27.61 |   |   |
| Uses laptop (No – Yes) | 0.494 | 15.99 | 0.769 | 20.29 |     |     |   |   |
| Uses tablet or e-book reader (No – Yes) | 0.287 | 12.23 | 0.049 | 1.85 |   |   |
| Used voice activated digital assistant in smart devices (No – Yes) | 0.176 | 7.26 | -0.669 | -26.41 |   |   |
| Enthusiasm about riding in / sharing the road with AVs |  |  |  |  |  |  |  |  |
| Enthusiasm about the development of AVs (Not at all enthusiastic – Very enthusiastic) | 1.850 | 18.30 | 2.574 | 19.82 | 2.420 | 21.78 | 4.599 | 23.23 |
| Anxiety about riding in / sharing road with AVs |   |   |   |   |   |   |   |   |
| Worried about the development of AVs (Not at all worried – Very worried) | 1.229 | 7.73 | 1.487 | 12.18 | 1.848 | 12.49 | 3.886 | 12.72 |