**Online supplement to**

**“On Accommodating Spatial Interactions in a Generalized Heterogeneous Data Model (GHDM) of Mixed Types of Dependent Variables”**

Chandra R. Bhat (corresp. author), Abdul R. Pinjari, Subodh K. Dubey, and Amin S. Hamdi

**Section 1**

**1.1 Description of Matrix M used to Derive the Distribution of Vector *yu***

Create a matrix **M** of size  whose elemens are all initialized to zero. Then, for every individual *q*, consider a matrix  of size filled with zeros. Insert an identity matrix of size *E* into the first *E* rows and *E* columns of the matrix . Next, consider the rows from , and columns from  These rows and columns correspond to the first nominal variable. Insert an identity matrix of size  after supplementing with a column of ‘-1’ values in the column corresponding to the chosen alternative. Next, consider the rows  through and columns  through correspond to the second nominal variable. Again position an identity matrix of size  after supplementing with a column of ‘-1’ values in the column corresponding to the chosen alternative for the second nominal variable. Continue this procedure for all *G* nominal variables. Insert this matrix  in the matrix **M** occupying the rows to and columns to.

**1.2 Rearrangement Matrices  and  used to Partition  and **

Consider a rearrangement matrix **R** of size filled with zeros. Then, for every individual *q*, consider a identity matrix  of size . Insert first *H* rows of matrix  into matrix **R** occupying rows to and columns to. Next insert the remaining  rows of matrix  into matrix **R** occupying rows to and columns to. Divide the matrix **R** into two submatrices  and . For example: Consider the case with two individuals and one continous, two ordinal, one count and two nominal variables with three alternatives each. Then the matrix **R** may be written as:

 (S.1)

Now, consider the following re-arranged vectors and matrices:

, , , , and .

**Section 2**

**The CML Estimation Approach**

To develop the CML function corresponding to Equation (B.2) in Appendix B of the main paper, first we extract from the mean vector  and covariance matrix , relevant components  and  (as well as  from  and  from ) for each (and every) pair of individuals  and . To do so, define a selection matrix  of size and  of size  (see Section A.1 Step-1 and Step-2 of Supplement Appendix A for construction details). Then the vectors , ,  and matrix  can be extracted from , ,  and  as follows: , , , and . Next, for each pair of individuals, we partition vector  and matrix  into components relevant to continuous variable components  and ordinal, count and nominal variable components  To do so, define a selection matrix  of size (see Supplement Appendix A.1 Step-3 for construction details). Then vector  and matrix  can be partitioned into the components corresponding to continuous and non-continuous variable components as follows: ,, ,, and . That is,  and. (S.2)

Thus the conditional distribution of  given , may be written as: mean  and variance . In the last step, we arrange the elements inside the vector , ,  and matrix  so that the elements corresponding to all ordinal variables of both individuals  and  (i.e., 2*N* ordinal variables) are stacked together, followed by elements corresponding to all count variables (i.e., 2*C* count variables) and in the end the elements corresponding to all nominal variables (i.e., 2*G* nominal variables). This arrangement makes it easy to enumerate pairs of observed outcomes for forming the CML function. To achieve such ordering, define a matrix  of size  (see Supplement Appendix A.1 Step-4 for construction details). Then the elements can be rearranged as follows: , , , and . Finally, replace the last elements of the lower threshold vector  from  to zero. That is, .

Now, to explicitly write the CML function in terms of standard (multivariate) normal density and cumulative distribution functions, define  as the diagonal matrix of standard deviations of the matrix **Δ**, using  for the multivariate standard normal density function of dimension *R* and correlation matrix  (), and  for the multivariate standard normal cumulative distribution function of dimension *E* and correlation matrix . Also, define a set of two selection matrices:  of size  and  of size , constructed as described in Supplement Appendix A.2. Using these selection matrices, let   , , and , where  represents the *v*th element of  (and similarly for other vectors), and represents the  element of the matrix . Then, one may write the CML function as follows:

**** (S.3)

In Equation (S.3), for each pair of individuals  and , the first component corresponds to the marginal likelihood of the continuous outcomes for the two individuals, the second component corresponds to the likelihood of pairs of outcomes across all ordinal and count outcomes, the third component corresponds to the pairwise likelihood of ordinal/count outcomes and nominal outcomes, and the last component corresponds to the pairwise likelihood for the nominal outcomes.

Among all the multivariate normal cumulative distribution (MVNCD) integrals in Equation (S.3), the maximum dimension of integration is the sum of alternatives of nominal variables with the two highest numbers of alternatives minus 2. To solve such MVNCD integrals within the CML estimation routine, Bhat (2011) proposed the use of an analytic approximation method. This approach, labelled the MACML approach, as demonstrated by Bhat and Sidharthan (2011), is at least as accurate as simulation based approaches in retrieving model parameters, albeit computationally much faster and robust in that the approximate CML surface is smoother and easier to maximize than the traditional simulation-based likelihood surfaces.

We write the resulting equivalent of Equation (S.3) computed using the analytic approximation for MVNCD as . The MACML estimator is then obtained by maximizing the logarithm of , which involves computation of pairwise log-likelihoods (i.e., ) for  pairs. The asymptotic covariance matrix of the parameters  may be estimated by the inverse of Godambe’s (1960) sandwich information matrix.

 (S.4)

The reader is referred to Zhao and Joe (2005), Bhat (2014), and Sidharthan and Bhat (2012) for more details on the calculations of the Hessian matrix () and the Jacobian matrix () in the above expression (Sidharthan and Bhat, 2012 provide these details for spatial models).

As spatial dependency decreases with an increase in the distance between any two individuals, one can reduce the number of pairings (between individual observations) in the MACML function by neglecting all the pairings beyond a certain threshold distance. To determine this threshold distance, analyst can estimate the model with different threshold distances and choose the one that minimizes the total variance across all parameters as given by the trace of the asymptotic covariance matrix.

One final consideration relevant to model estimation is that the matrix  for each observation has to be positive definite. The simplest way to guarantee this is to ensure that the (*L*×*L*) correlation matrix **Γ** is positive definite, and each matrix (*g*=1,2,…,*G*) is also positive definite. To do so, we parameterize the CML function in terms of the Cholesky parameters for these matrices. Further, because the matrix **Γ** is a correlation matrix, we write each diagonal element (say the *aath* element) of the lower triangular Cholesky matrix of **Γ** as , where the  elements are the Cholesky factors to be estimated. In addition, note that the top diagonal element of each  matrix has to be normalized to one (as discussed in Appendix A of the main paper), which implies that the first element of the Cholesky matrix of each  is fixed to the value of one. Also, the spatial autoregressive parameter  should be constrained between 0 and 1 , for which we parameterize the spatial autoregressive parameter as . However, in our case, we expect the parameter to be positive, because the spatial dependence is being introduced in the latent pschological constructs, and this impose the ‘+’ sign.

**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. (2014). The composite marginal likelihood (CML) inference approach with applications to discrete and mixed dependent variable models. *Foundations and Trends in Econometrics*, 7(1), 1-117.

Bhat, C.R., Sidharthan, R. (2011). A simulation evaluation of the maximum approximate composite marginal likelihood (MACML) estimator for mixed multinomial probit models. *Transportation Research Part B*,45(7), 940-953.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics* 31(4), 1208-1211.

Sidharthan, R., Bhat, C.R. (2012). Incorporating spatial dynamics and temporal dependency in land use change models. *Geographical Analysis*, 44(4), 321-349.

Zhao, Y., Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *The Canadian Journal of Statistics* 33(3), 335-356.

**Supplement Appendix A: Construction of Selection Matrices**

**A.1. To Extract Relevant Components from**  **and**  **for a pair of individuals** *q* **and , and rearrange elements in the order of ordinal, count and nominal variables**

To build the CML function, we use a set of selection matrices such that one can extract the relevant components from vectors  and matrix  for a pair of individuals *q* and . The selection matrices are described below:

**Step-1:** A matrix  of size  filled with zeros. Insert an identity matrix of size  in first  rows and columns to. Next, insert another identity matrix of size  in rows  to  and columns to.

**Step-2:** A matrix  of size  filled with zeros. Insert an identity matrix of size  in first  rows and columns to. Next, insert another identity matrix of size  in rows  to  and columns to.

**Step-3:** A matrix  of size  filled with zeros. Then, for individual , consider an identity matrix  of size . Insert first  rows of the matrix  into matrix  occupying first *H* rows and columns to. Next insert the remaining  rows of matrix  into matrix  occupying rows to and columns to. For individual , Insert first *H* rows of matrix  into matrix occupying rows  to  and remaining  rows of matrix  into matrix  occupying rows to and columns to. Divide the matrix  into two submatrices  and .

**Step-4:** Now, define a matrix  of size  filled with zeros. Next, perform the steps described below to arrange the elements inside the vector  and matrix  in the order of ordinal, count and nominal variables.

***Step-4.1*** Insert an identity matrix of size *N* in first *N* rows and *N* columns. Insert another identity matrix of size *N* in rows (*N+1*) to (2*N*) and columns  to.

***Step-4.2*** Again, insert an identity matrix of size *C* in rows (2*N+1*) to  and columns  to . Insert another identity matrix of size *C* in rows  to  and columns  to.

***Step-4.3*** Finally, insert an identity matrix of size  in rows  to  and columns  to . Insert another identity matrix of size  in rows  to  and columns  to .

**A.2. To write the CML function in terms of standard normal density and cumulative distribution functions**

Define a set of two selection matrices as follows: (1)  is a  selection matrix with an entry of ‘1’ in the first row and the column and an identity matrix of size  occupying the last  rows and the through columns (with the convention that ), and entries of ‘0’ everywhere else, (2)  is a  selection matrix with an identity matrix of size () occupying the first () rows and the through columns (with the convention that ), and another identity matrix of size  occupying the last  rows and the through columns; all other elements of  take a value of zero.

**Section 3**

**Table 1. Descriptive statistics of independent variables**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Categories** | **Percentage** |
| Age of school going children | 5-10 years old | 47.37 |
| 11-15 years old | 40.83 |
| 16-18 years old | 11.80 |
| Gender of school going children | Boys | 51.80 |
| Girls | 48.20 |
| Household monthly income | Less than 25K | 7.70 |
| 25K – 49,999 | 16.18 |
| 50K – 74,999 | 16.68 |
| 75K – 99,999 | 16.77 |
| 100K or more | 42.67 |
| Race | Caucasian | 69.78 |
| African-American | 7.39 |
| Asian | 9.72 |
| Hispanic | 12.07 |
| Others | 1.04 |
| Households with fraction of adults (25 or more) with | High school degree or less | 25.45 |
| Some college degree | 25.73 |
| Bachelor’s degree | 27.69 |
| Graduate degree | 21.13 |
| Households with fraction of adults in age group | 19-30 years | 7.87 |
| 31-45 years | 51.70 |
| 46-60 years | 36.43 |
| 61 or more | 4.00 |
| Households with number of full-time workers | 0 | 5.21 |
| 1 | 58.66 |
| 2 | 33.36 |
| 3 or more | 2.77 |
| Households with number of part-time workers | 0 | 67.93 |
| 1 | 28.80 |
| 2 or more | 3.27 |
| Households with number of workers with the option to work from home | 0 | 75.62 |
| 1 | 21.43 |
| 2 or more | 2.95 |
| Households with number of workers with flexible work time | 0 | 42.26 |
| 1 | 44.52 |
| 2 or more | 13.22 |
| Family type | Nuclear | 93.41 |
| Single-parent | 6.59 |
| Housing type | Detached | 81.75 |
| Duplex | 6.54 |
| Apartment or townhouse | 11.71 |
| Tenure | Own | 80.05 |
| Rent | 19.95 |
| Distance to school | Less than ¼ mile | 11.29 |
| Between ¼ to ½ mile | 10.37 |
| Greater than ½ mile and less than or equal to 1 mile | 14.47 |
| Greater than 1 mile and less than or equal to 2 miles | 21.43 |
| More than 2 miles | 42.44 |

**Section 4**

**Table 2. Parameter estimates of exogenous variable effects on non-nominal variables**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dependent variables** | Constants | | Thresholds for ordinal variables between… | | | | | | Dispersion parameter\* | | Variance | |
| A little bit of an issue & somewhat of an issue | | Somewhat of an issue & very much an issue | | Very much an issue & a serious issue | |
| **Coeff** | **T-stat** | **Coeff** | **T-stat** | **Coeff** | **T-stat** | **Coeff** | **T-stat** | **Coeff** | **T-stat** | **Coeff** | **T-stat** |
| **Household average commute distance (miles)** | 2.428 | 3.80 | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | 0.770 | 4.06 |
| **Walk/bike issue: violence/crime along the route** | 0.072 | 1.67 | 0.419 | 10.22 | 0.779 | 13.91 | 1.035 | 17.54 | ---- | ---- | ---- | ---- |
| **Walk/bike issue: speed of traffic along the route** | 3.543 | 2.37 | 1.154 | 6.71 | 2.757 | 11.16 | 4.117 | 15.65 | ---- | ---- | ---- | ---- |
| **Walk/bike issue: amount of traffic along the route** | 3.667 | 2.52 | 1.353 | 10.33 | 2.848 | 15.56 | 4.268 | 18.16 | ---- | ---- | ---- | ---- |
| **Number of biking episodes in past week** | -2.355 | -3.01 | ---- | ---- | ---- | ---- | ---- | ---- | 0.062 | 86.63 | ---- | ---- |
| **Number of walking episodes in past week** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Number of times public transit used in past week** | 0.424 | 2.65 | ---- | ---- | ---- | ---- | ---- | ---- | 0.098 | 86.00 | ---- | ---- |
| **Vehicle ownership** | 1.384 | 4.79 | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **At least one working adult uses public transit/walk/bike as mode to work** | -0.097 | -2.31 | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |

* The t-stat for dispersion parameter is calculated with respect to numerical value of 5 as oppose to zero. This is due to the fact that negative binomial count model collapses to Poisson count model for a dispersion parameter value of 5 or more.

**Table 2 (cont.) Parameter estimates of exogenous variable effects on non-nominal variables**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dependent variables** | Number of children (18 years or less) in the household | | Household income (base: 75K or more) | | | | | | Number of full-time workers | | Number of part-time workers | |
| Less than 25K | | 25K-49,999 | | 50K-74,999 | |
| **Coeff** | **T-stat** | **Coeff** | **T-stat** | **Coeff** | **T-stat** | **Coeff** | **T-stat** | **Coeff** | **T-stat** | **Coeff** | **T-stat** |
| **Household average commute distance (miles)** | 0.052 | 1.73 | -0.720 | -6.37 | -0.372 | -4.96 | -0.118 | -2.66 | ---- | ---- | -0.162 | -3.38 |
| **Walk/bike issue: violence/crime along the route** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Walk/bike issue: speed of traffic along the route** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Walk/bike issue: amount of traffic along the route** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Number of biking episodes in past week** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Number of walking episodes in past week** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Number of times public transit used in past week** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Vehicle ownership** | 0.017 | 2.27 | -0.308 | -1.91 | -0.110 | -2.18 | ---- | ---- | 0.229 | 2.63 | 0.234 | 4.78 |
| **At least one working adult uses public transit/walk/bike as mode to work** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |

**Table 2 (Cont.) Parameter estimates of exogenous variable effects on non-nominal variables**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dependent variables** | Number of workers with the option to work from home | | Housing type  (base: detached) | | | | Tenure  (base: owned) | |
| Duplex | | Apartment or Townhouse | | Rented | |
| **Coeff** | **T-stat** | **Coeff** | **T-stat** | **Coeff** | **T-stat** | **Coeff** | **T-stat** |
| **Household average commute distance (miles)** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Walk/bike issue: violence/crime along the route** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Walk/bike issue: speed of traffic along the route** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Walk/bike issue: amount of traffic along the route** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Number of biking episodes in past week** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Number of walking episodes in past week** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Number of times public transit used in past week** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| **Vehicle ownership** | -0.039 | -3.08 | -0.257 | -2.42 | -0.610 | -5.65 | -0.176 | -2.34 |
| **At least one working adult uses public transit/walk/bike as mode to work** | ---- | ---- | ---- | ---- | ---- | ---- | ---- | ---- |

**Table 3. Parameter estimates of exogenous variable effects on residential location choice**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variables** | Residential location  (base: less than 1000 housing units per sq mile) | | | | | |
| 1000-1999 | | 2000-3999 | | 4000 or more | |
| **Coeff** | **T-stat** | **Coeff** | **T-stat** | **Coeff** | **T-stat** |
| Constants | 0.584 | 6.42 | 0.563 | 11.04 | 0.238 | 1.75 |
| Family type (base: nuclear family) |  |  |  |  |  |  |
| Single-parent | ---- | ---- | 0.018 | 1.44 | 0.018 | 1.44 |
| Household income (base: less than 100K) |  |  |  |  |  |  |
| 100K or more | ---- | ---- | -0.037 | -2.47 | ---- | ---- |
| Housing type (base: detached) |  |  |  |  |  |  |
| Duplex | ---- | ---- | ---- | ---- | 0.099 | 6.60 |
| Apartment or Townhouse | ---- | ---- | 0.072 | 2.44 | 0.302 | 4.65 |
| Number of workers with the option to work from home | ---- | ---- | -0.028 | -2.15 | -0.028 | -2.15 |

In addition to the above observed effects of exogenous variables, we estimated the following covariance matrix (t-statistics in parenthesis) between the differences of error terms with respect to the error term of the lowest density residential location alternative:

****

Note that the terms in the parentheses represent the t-statistic values. Since this is a differenced error-covariance matrix (**Λ**), a clear interpretation of the elements of this matrix is not straightforward.