**Code Documentation for the Estimation of a Probit Kernel-based Spatial Discrete Choice Model Accommodating Residential Self-selection Effect**

This addendum serves to document the estimation code for a probit kernel-based spatial interaction discrete choice model, which accommodates the residential self-selection effect. An explanation of the parameters in the model and the different settings available within the code are documented below. Please refer to Bhat, and Dubey (2013) for the notations and the model structure.

Bhat, C.R., and S.K. Dubey (2013) "**A New Spatial (Social) Interaction Discrete Choice Model Accommodating Self-Selection in Group Formation**," Technical paper, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, 2013.

1. **Data Generation Settings**

* The current version of the code allows a user to generate the data within the code as well as specify the external dataset.
* If the user wants to simulate the data inside the code, choose the option “Simulation\_data” (line 24). On the other hand, if the user wants to specify his or her own dataset, choose the option “Own\_data” (line 23) and specify the file name in the section “How to specify your own dataset” in the code. The user can also change the seed for data generation process under the simulation option by changing the value of variable “Run\_No” (line 58).

1. **Dataset Specifications**

The dataset should be a csv file with no header. The dataset should include the columns in the following order.

1. Columns of exogenous variables for each of the alternatives.
2. A column filled with the value 1.
3. A column filled with the value 0.
4. A column indicating the chosen alternative number. For example: if the choice model has three alternatives- car, rail, and transit, the corresponding row in the column should have a value 1 if the person chooses car, 2 for rail and 3 for transit.
5. Since the dataset used in this program is a panel dataset, the dataset should contain a panel record of each individual arranged in ascending choice occasion order.
6. **Simulation Data**

The data in the code is generated using the simulation configuration discussed in the paper titled “**A New Spatial (Social) Interaction Discrete Choice Model Accommodating Self-Selection in Group Formation”.**  Please refer to the paper for details of simulation configuration.

1. **Code Settings**

The user must specify the value of following variables (lines 30 – 44).

**Table 1: Description of Variables**

|  |  |
| --- | --- |
| **Variables** | **Description** |
| nind | Number of individuals |
| nchocc | Number of choice occasion per individual |
| nvar | Number of variables in the utility equation per alternative |
| nc | Number of alternatives in the model |
| nran | Number of random parameters in the model |
| cutoff\_distance | Distance band for pairing of observations in the likelihood expression |
| Psi\_Cholesky | Used to determine, whether the user wants to obtain the Cholesky elements of the error covariance matrix |

1. **Sample Data**

The data consist of 300 individual record with five choice occasions each (thus, the number of rows in the dataset equals 1500 [300\*5]). The number of alternatives in the dataset is equal to four with three explanatory variables each. The following table presents the content and the structure of the sample dataset.

**Table 2: Sample Data Description**

|  |  |
| --- | --- |
| **Column No** | **Explanation** |
| 1 to 12 | Explanatory variables for each of the four alternatives |
| 13 | A column of 1’s |
| 14 | A column of 0’s |
| 15 | Column indicating the chosen alternative number |

We do not provide labels for the explanatory variables because the labels are generated internally by the program. Please refer to line number 126 for how to define labels for variables. The sample data follows the same structure as discussed in the simulation exercise of the paper.

1. **Estimation Results**

The code has two likelihood expressions as some sort of parameterization is necessary to ensure that all the spatial parameters have a value between zero and one at the end of the estimation. Thus, the first likelihood expression estimates the parameterized coefficients and then passes the un-parameterized values to the second likelihood expression.

Since the standard error calculations for spatial models are slightly different from non-spatial models, please ignore all the standard error values printed by Gauss. The correct standard error and the corresponding t-stat value for the un-parameterized coefficients are printed at the end under the section titled “Final Result,” as shown below.

Here we assume generic coefficients for three exogenous variables. We also consider the first variable to be fixed, but allow spatial drift effect on remaining two variables. Thus, we estimate three generic coefficients (mean effect represented by X01, X02 and X03), three Cholesky elements (lower triangular matrix) of the random parameter covariance matrix (represented by Tild01, Tild02 and Tild03), one spatial lag parameter (Del01) and two spatial drift parameters (Lamda02 and Lamda03). Please note that for all the Cholesky decomposition elements, the corresponding position of the elements in the matrix is added at the end of their name. For example: Tild01 indicates that this element is the 1st element of the Cholesky decomposition (lower triangular matrix) of the random parameter covariance matrix.

-----------------Final Result---------------------------------------------------------------------------------

Log-likelihood value : -30562.8

Parameter Estimate ST.Error T-Stat

X01 0.523 0.0034 153.31

X02 0.862 0.0052 165.24

X03 1.043 0.0065 161.22

Tild01 0.992 0.0085 116.16

Tild02 0.459 0.0083 55.58

Tild03 0.920 0.0097 94.56

Del01 0.199 0.0027 74.44

Lamda02 0.249 0.1885 1.32

Lamda03 0.250 0.1172 2.13