SSMLMNP
A Multinomial/Multiperiod Probit Program Using Maximum SRC/GHK Smoothly Simulated Likelihood

by

©Axel Börsch-Supan, University of Mannheim

(Enhancements and UNIX Port by Vassilis Hajivassiliou, London School of Economics)

Version 1.2, June 1992

1 Introduction

The SSMLMNP program, written in Fortran-77, carries out smoothly simulated maximum likelihood estimation of a Multinomial Probit Model for both Cross-Sectional and Multiperiod (panel) data sets. The estimation method is described in **A. Börsch-Supan and V. Hajivassiliou, 1993, "Smooth Unbiased Multivariate Probaiblity Simulators for maximum Likelihood Estimation of Limited Dependent Variable Models,"** *Journal of Econometrics*, **347–368**.

2 Running the Program on a UNIX System

1. Create a profile file using '`ssmlmnp.pro`' as a template, specifying the configuration of your program – see the end of this section.

2. To use the program in conjunction with the profile file '`myfile.pro`' you should issue

   ```
   ssmlmnp myfile
   ```

   The default output will be printed on the screen. In order to save the default output in file '`myfile.scn`' you should issue

   ```
   ssmlmnp myfile >& myfile.scn
   ```

   In usual unix fashion, if you would like to have the program run as a background process, you should append the ampersand '`&`' character at the end of the command line.

   For example, the command

   ```
   ssmlmnp myfile >& myfile.scn &
   ```

will run the program in the background, saving all screen messages into file myfile.scn. An even better idea is to issue

```
nice ssmlmnp myfile >& myfile.scn &
```

which will run the job at a "nice" priority depending on the usage of the machine you are employing.

2.1 Notes on the profile file

```
   ** PAUSE [SEC]: number of seconds to wait between screens; if
   non-interactive, set to -1

   ** SPOOL FILE : name of output file

   ** DATA FILE  : name of data file.  See Subsection~\ref{Data}
   for a detailed explanation as to how the data should be
   organized.  In summary, the first line should state in parentheses
   a FORTRAN statement to read one complete observation on all
   variables.  The NX data for the variables that vary across
   alternatives (X variables -- see below) should appear first,
   followed by the value for the dependent variable, plus the NY
   values of the Y variables that do not vary across alternatives.
   Example 1:
   ---------
   NALT=3, NX=0, NY=4, first 9 observations
(5g13.6)
 3.00000     0.000000     0.000000      1.00000      47.0000
 1.00000      1.00000     0.000000      1.00000      48.0000
 3.00000     0.000000     0.000000      1.00000      49.0000
 1.00000      1.00000     0.000000      1.00000      50.0000
 3.00000     0.000000     0.000000      1.00000      47.0000
 1.00000      1.00000     0.000000      1.00000      48.0000
 3.00000     0.000000     0.000000      1.00000      49.0000
 1.00000      1.00000     0.000000      1.00000      50.0000
 3.00000     0.000000     0.000000      1.00000      49.0000
 1.00000      1.00000     0.000000      1.00000      50.0000
   Example 2:
   ---------
   NALT=3, NX=2, NY=3, first 9 observations
```

```
(8g8.4)
  0.2342  5.0000  3.0000  0.0000  0.0000  1.0000  47.0000
  1.0000  5.2222
  2.1111  5.4444
  3.2222  7.6666  1.0000  1.0000  0.0000  1.0000  48.0000
  2.3333  7.3333
  1.4444  7.0000
  4.5555  2.2222  3.0000  0.0000  0.0000  1.0000  49.0000
  5.1111  2.0000
  6.2222  2.3333

  ** OUTPARMS FILE : output file with results saved as possible
  inputs in future runs

  ** INITIAL ESTIMATION parameters:
     ** DERIVATIVE CHECK OF OMEGA MATRIX (0=NO, 1=YES) **
     ** METHOD:(0=NONE,1=LOGIT,2=CLARK)
        Checks of Derivatives: (0=NONE,1=LIK,2=YES)
        ITER : Maximum number of iterations
        ACCU : Convergence accuracy
        IALG : Algorithm to use: (2=DFP,4=BHHH)

  ** SML ESTIMATION parameters:
     ** NSIMUL : number of simulations
     ** ISEED  : starting seed for random number generator
     ** Derivative check: (0=NONE,1=LIK,2=YES)
     ** ITER   : maximum number of iterations
     ** ACCU   : convergence accuracy
     ** IALG   : Algorithm to use: (2=DFP,4=BHHH)

  ** COVARIANCE CORRECTION:
     ** RECOMPUTATION OF ROBUST COVARIANCE: 0=NO, 1=YES/OLD H, 2=YES/NEW H
     Let "OLD_H" = Hessian already calculated by optimization
                   algorithm.  This may be far from the Hessian
                   evaluated at the latest parameter values, because
                   algorithms like DFP recalculate the Hessian only
                   every several iterations, using rough
                   approximations in-between.
```

```
   Let "NEW_H" = Hessian reevaluated at the latest parameter values.
   Let "OFP_H" = BHHH approximation of Hessian based on Outer
                 Product of First Partial derivatives.
   Then the final variance-covariance matrix will be based on:
   0 = -inv\{OLD_H\}
   1 = inv\{ OFP_H * inv(-OLD_H) * OFP_H \}
   2 = inv\{ OFP_H * inv(-NEW_H) * OFP_H \}

** DIMENSIONS:
   ** NINDIV : number of individuals
   ** NALT   : number of discrete alternatives
   ** NPER   : number of time periods per individual
               NB:  Note that this implies that your panel data set
               --   must be a BALANCED one, i.e., it must have the
                    same number of time periods per individual.
   ** RAN/NO_RAN : 1=random effects present, 0=no random effects
   ** AR1/NO_AR1 : 1=AR1 present, 0=no AR1
   ** MNP/IIA    : 1=MNP, 0="IIA"

** INITIALVALUES:
    1) NX                X-VARIABLES
    2) NY*(NALT-1)       Y-VARIABLES
    3) NALT-1            CONSTANTS
    4) NALT              STANDARD DEVIATIONS OF UNOBSERVED UTIL'S
    5) NALT*(NALT-1)/2   CORRELATIONS OF UNOBSERVED UTIL'S
    6) NALT              STANDARD DEVIATIONS OF RANDOM EFFECTS
    7) NALT*(NALT-1)/2   CORRELATIONS OF RANDOM EFFECTS
    8) NALT              AUTOCORRELATIONS

    See Boersch-Supan, Hajivassiliou, Kotlikoff, and Morris 1992,
    Health, Children, and Elderly Living Arrangements:  A
    Multiperiod Model with Unobserved Heterogeneity and
    Autocorrelated Errors " in D. Wise, ed., "Topics in the
    Economics of Aging," Chicago, University of Chicago Press,
    for a discussion of necessary identification assumptions.

    NB:  The program WILL NOT impose automatically any restrictions
         necessary to achieve identification.  It is your
```

```
            responsibility to impose such restrictions by enough
            "DO=0" values -- see below.

Example:
-------

DO=1 means "estimate parameter with PARVAL as starting value"
DO=0 means "hold parameter fixed during estimation at PARVAL value"

PARNAM..DO..PARVAL......
12345678..123456789012345
--------==---------------
PREMV1   1  23.4164
PREMV2   1  23.2921
LFS1     1   0.3107
LFS2     1   0.2940
ONE1     1  -0.5455
ONE2     1  -0.1827
sd-nu1   1   1.8197
sd-nu2   1   0.4088
sd-nu3   0   1.0000
co-nu1   1  -0.8801
co-nu2   1   0.2698
co-nu3   1   0.7702
sd-alfa1 1   0.7026
sd-alfa2 1   0.5484
sd-alfa3 0   0.0000
co-alfa1 1   0.0000
co-alfa2 1   0.0000
co-alfa3 1   0.0000
rho1     1   0.0000
rho2     1   0.0000
rho3     1   0.0000


***********************************************************************
sample ssmlmnp.pro file
***********************************************************************
```

```
**************** PROGRAM MODE: ************************************************
** PAUSE [SEC]: if non-interactive, set to -1 **
   -1
** SPOOL FILE **
   ssmlmnp.spo
** DATA FILE **
   ssmlmnp.dat
** OUTPARMS FILE **
   ssmlmnp.out
**************** INITIAL ESTIMATION: *******************************************
** DERIVATIVE CHECK OF OMEGA MATRIX (0=NO, 1=YES) **
   0
** METHOD                   D-CHECK                ITER   ACCU   IALG
   (0=NONE,1=LOGIT,2=CLARK)  (0=NONE,1=LIK,2=YES)                (2=DFP,4=BHHH)
   0                        0                      100    .00001   2
**************** SML ESTIMATION: ***********************************************
** NSIMUL      ISEED        D-CHECK                ITER   ACCU   IALG
                            (0=NONE,1=LIK,2=YES)                 (2=DFP,4=BHHH)

   100          1234567       0                     100    .00001   2
**************** COVARIANCE CORRECTION: ****************************************
** RECOMPUTATION OF ROBUST COVARIANCE (0=NO, 1=YES/OLD HESS, 2=YES/NEW HESS) **
   2
**************** DIMENSIONS: ***************************************************
** NINDIV **
   81
** NALT  NX  NY  CONST  WEIGHT **
   3     0   4   1      0
** NPER  RAN/NO_RAN  AR1/NO_AR1  MNP/IIA  (1/0) **
   10    1           1           1
**************** INITIALVALUES: ************************************************
1) NX                X-VARIABLES
2) NY*(NALT-1)       Y-VARIABLES
3) NALT-1            CONSTANTS
4) NALT              STANDARD DEVIATIONS OF UNOBSERVED UTIL'S
5) NALT*(NALT-1)/2  CORRELATIONS OF UNOBSERVED UTIL'S
6) NALT              STANDARD DEVIATIONS OF RANDOM EFFECTS
```

```
7) NALT*(NALT-1)/2  CORRELATIONS OF RANDOM EFFECTS
8) NALT             AUTOCORRELATIONS
--------------------------------------------------------------------------------
PARNAM..DO..PARVAL......
12345678..123456789012345
--------==---------------
PREMV1   1  23.4164
PREMV2   1  23.2921
PREOWN1  1   1.7004
PREOWN2  1   0.9207
LFS1     1   0.3107
LFS2     1   0.2940
HAGE1    1   0.0400
HAGE2    1  -0.0293
ONE1     1  -0.5455
ONE2     1  -0.1827
sd-nu1   1   1.8197
sd-nu2   0   1.0000
sd-nu3   0   1.0000
co-nu1   1  -0.8801
co-nu2   0   0.0000
co-nu3   0   0.0000
sd-alfa1 1   0.7026
sd-alfa2 1   0.5484
sd-alfa3 0   0.0000
co-alfa1 1   0.0000
co-alfa2 1   0.0000
co-alfa3 1   0.0000
rho1     1   0.0000
rho2     1   0.0000
rho3     1   0.0000
```

3 Compiling 'ssmlmnp' on a UNIX System

You must have the following files on the same directory:

```
ssmlmnp.f
ssmlmnp.fad
```

```
ssmlmnp.cad
ssmlmnp.inc
ssmlmnp.doc
```

Then issue the command

'`sh create`'

to create the executable file '`ssmlmnp`' (this requires about 5 Meg of RAM to run successfully).

The file '`ssmlmnp.doc`' gives basic instructions as to how to use the program.

This program was originally written by Axel Börsch-Supan of the University of Mannheim. Vassilis Hajivassiliou of the London School of Economics carried out the main port to a unix system and added several enhancements.

4 The Data Input Files `INDATA.DAT`

`SSMLMNP` is intended to be addressed by an external data handling program that generates and manipulates the data that will be fed into `SSMLMNP` for estimation and analysis.

Data can be read in from a file with ASCII numbers.

We will first describe some generalities, then the format of an ASCII raw data file.

4.1 Size of Data Set

`SSMLMNP` does not limit the number of observations. However, if the data exceeds the size of the internal worksheet ("buffer"), data has to be moved between the `SSMLMNP` binary scratch file and memory ("paging") which substantially slows the program down.

The size of the internal worksheet to store data depends on the specific installation of SSMLMNP. In the standard version, `MAXBUF`=16000 numbers, which corresponds to 1 Meg of RAM. The number of observations that fit in this internal worksheet depends on the number of alternatives (`NALT`), the number of explanatory variables (`NX` and `NXD`), and whether a weight and a choice set variable are included or not (`NWT` and `NCS`). It can be computed by the formula:

`MAXOBS = MAXBUF/NDA - 2`

where the number of data items per observation is

`NDA = (NX+NCS)*NALT+1+NWT+NXD`

All data is read at the beginning of the program.

4.2 Variable Types

Six types of variables should be distinguished:

1. The dependent variable. For each observation, this is the index of the chosen alterative. (See below for indexation.)

2. If `NX` > 0:

   `NX` alternative-specific explanatory variables ("attributes") which vary by `NALT` alternatives and (possibly) by observation. For each observation, this is a two-dimensional array of `NALT` rows and `NX` columns.

3. If `NXD` > 0:

   `NXD` agent-specific variables ("characteristics") which vary only by observation but not by alternative. For each observation, this is a row-vector of length `NXD`.

4. If `NXA` $= 1$:

   A constant term. It will be generated internally, so one must not include it in the data.

5. If `NWT` $= 1$:

   The weights. For each observation, this is a number carrying a weighting factor. (See below for examples.)

6. If `NCS` $= 1$:

   The choice set variable. For each observation, this is a column vector of zeroes and ones of length `NALT`, in which a one (zero) indicates that the corresponding alternative is included (excluded) in the observation's choice set.

NOTE: In oder to conserve space in `SSMLMNP` data sets, the above order of variables does NOT correspond to the actual order in `SSMLMNP` data sets.

The dependent variable indicates the alternative chosen and carries a value between 1 and `NALT`, corresponding to the order in which the alternatives are stored.

The weights are applied to each observation's likelihood as well as to summary statistics, predictions and elasticities. They can also be used as replication factors for grouped data analysis since the sum of the weights across observations need not sum up to one. In all summary statistics, `SSMLMNP` divides by the sum of weights, not the number of physical observations.

4.3 Examples

1. In simple random sampling with balanced choice sets, the weights in each observation and alternative are 1.0. Therefore, it is unnecessary to include weights in the data, set `NWT`=0.

2. In simple choice based sampling, the weights are the population share of the alternative chosen by the observation divided by the sample share of this alternative.

3. In stratified random sampling, the weights represent the weight of the corresponding stratum.

4. The user can combine (2) and (3).

4.4 ASCII Raw Data Files

The first line of the ASCII data file must be a valid REAL FORTRAN format description. The format has to be enclosed in parentheses. Including the parentheses, the format must not exceed 72 characters.

The program then expects `NALT` lines for each observation. Each line carries
`NCS+NX+NWT+1+NXD`
entries in the following order using the above user-specified format.

1. 1, if the corresponding alternative is included in the choice set of the observation, 0 else. (Omit this entry, if `NCS`=0).

2. `NX` alternative-specific characteristics for the corresponding alternative and observation. (Omit these entries, if `NX`=0).

3. The dependent variable (=index of the chosen alternative. This entry must always be present)

4. The weight (Omit this entry, if `NWT`=0)

5. `NXD` agent-specific attributes. (Omit these entries, if `NXD`=0).

NOTE: Items (3)–(5) are the same for all alternatives of an observation.

NOTE: For each observation, the first line is read completely with the complete format description. In the remaining `NALT-1` lines of each observation, only the first `NCS+NX` items are read with the first `NCS+NX` items of the above format. The index of the chosen alternative and the `NXD` agent-specific attributes are ignored, since they are identical to those in the first alternative. (Accordingly, they can also be omitted in the data file, creating a non-rectangular file structure).

NOTE: If NCS+NX=0, each observation must have only one line of data.

NOTE: The alternatives are assumed to be in ascending order independently of the structure of the tree that will be estimated. The dependent variable must be an integer between 1 and `NALT` corresponding to the alternatives in ascending order. In other words, the index of each alternative is defined by the order in the data file.

NOTE: The number and the order of alternatives must always be the same for each observation. This should be emphasized in the case of unbalanced choice sets. In this case, also the alternative-specific variables in the alternatives excluded from the choice set should be given some value to avoid a read error although this value will never be used.

EXAMPLE: Imagine the following data for a four-alternative choice problem (`NALT`=4):

1. A choice set variable (assume, observation one has no alternative 3) (`NCS`=1).

2. Two alternative-specific attributes `P` ("price of alternative") and `Q` ("quality of alternative"). `P` should be attached with two different coefficients, one for alternatives 1 and 2, the other for alternatives 3 and 4. Q should be attached with a common coefficient in all four alternatives:

```
(NX=2,  NXM=2, LMAP=1 1 0 0
                    0 0 1 1
        NXM=1, LMAP=1 1 1 1)
```

3. The dependent variable (Assume, observation 1 has chosen alternative 2 and observation 2 has chosen alternative 3)

4. Weights are included (assume, observation 1 has a third of the weight of observation 2) (`NWT`=1)

5. Three agent-specific characteristics `Y`, `S`, `A` ("income, household size, and age") which are interacted by two dummies

```
(NXD=3, NXM=2, LMAP=1 0 0 0
                    0 1 1 0)
```

6. A constant which is interacted with three dummies

```
        (NXA=1, NXM=3, LMAP=1 0 0 0
                             0 1 0 0
                             0 0 1 0)

The ASCII-data file should look like:     or, more economically, like:

(F3.1,2F10.3,2F5.1,3F10.3)                (F1.0,2F9.3,F2.0,F4.1,3F9.3)
1.0  P11  Q11  2.0  0.5  Y1  S1  A1       1 P11 Q11 2 0.5 Y1 S1 A1
1.0  P12  Q12  2.0  0.5  Y1  S1  A1       1 P12 Q12
0.0  P13  Q13  2.0  0.5  Y1  S1  A1       0 P13 Q13
1.0  P14  Q14  2.0  0.5  Y1  S1  A1       1 P14 Q14
1.0  P21  Q21  3.0  1.5  Y2  S2  A2       1 P21 Q21 3 1.5 Y2 S2 A2
1.0  P22  Q22  3.0  1.5  Y2  S2  A2       1 P22 Q22
1.0  P23  Q23  3.0  1.5  Y2  S2  A2       1 P23 Q23
1.0  P24  Q24  3.0  1.5  Y2  S2  A2       1 P24 Q24
etc.                                      etc.

{\bf NOTE}: The interaction feature will expand the explanatory variables
      internally to the following array:

      P11  0    Q11  Y1  0   S1  0   A1  0   1  0  0
      P12  0    Q12  0   Y1  0   S1  0   A1  0  1  0
      0    P13  Q13  0   Y1  0   S1  0   A1  0  0  1
      0    P14  Q14  0   0   0   0   0   0   0  0  0
      P21  0    Q21  Y2  0   S2  0   A2  0   1  0  0
      P22  0    Q22  0   Y2  0   S2  0   A2  0  1  0
      0    P23  Q23  0   Y2  0   S2  0   A2  0  0  1
      0    P24  Q24  0   0   0   0   0   0   0  0  0
      etc.
```

5 Comments About Optimization

It is recommended to employ DFP or BHHH until convergence, then recalculate the covariance matrix by using covariance option 2.

The output file includes (in parentheses) the following return codes:

-9	=	Hessian singular
-8	=	Eigenvalues did not converge
-7	=	Numerical saddlepoint
-6	=	Cannot find improving step
-5	=	Too many function errors
-4	=	*** Out of memory ***
-3	=	Function error in gradient evaluation
-2	=	Initial values not admissible
-1	=	Iteration limit exceeded
0	=	*** Unknown error ***
1	=	Stepsize convergence achieved
2	=	Gradient convergence achieved
3	=	Function value convergence achieved
4	=	Gradient*direction convergence achieved