The SPACE Program

Version 2.0

June 23, 2009

Introduction

The SPACE (Stochastic Population Analysis for Complex Events) program is a package of SAS programs that compute multi-state life table (MSLT) functions, including the widely used health expectancy (HE). The MSLT model follows either a first-order Markov process where the transition probabilities depend on current status only, or a semi-Markov process where the transition probabilities depend on both current status and duration in current status (Cai, Schenker and Lubitz 2006; included in the zip file). The SPACE program also allows the computation of standard errors for the estimated MSLT functions using the bootstrap method (Rao and Wu 1988), which has been used in two recent studies (Cai and Lubitz 2007; Cai et al. 2006) and is described in detail in an upcoming manuscript (Cai et al. 2008).

The SPACE program offers users various options to calculate MSLT functions. The estimation can be performed with or without the covariates, which are not limited to be dichotomous. It allows users to choose the appropriate method to estimate the transition probabilities and rates (multinomial logistic regression or hazard regression). It also provides different ways to calculate HE – the deterministic approach or the stochastic approach (i.e., microsimulation). Simulation offers users a high degree of flexibility to summarize various aspects of the dynamics of population health changes.

The SPACE program has six main sets of programs, giving users different combinations of its functions. The first five sets estimates a first-order Markov model:

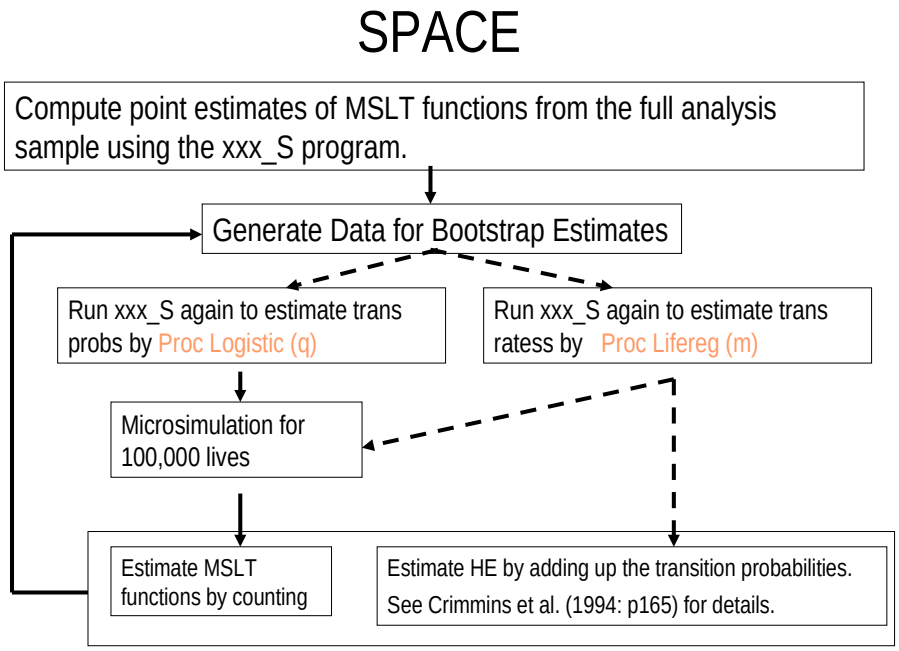
1. MSLT_RAD estimates age-specific state-dependent transition rates using the discrete-time hazard model with no covariates and calculates HE using the deterministic approach.
2. MSLT_RAD1COV estimates age-specific state-dependent transition rates using the discrete-time hazard model with one covariate and calculates HE using the deterministic approach.
3. MSLT_RAD2COV estimates age-specific state-dependent transition rates using the discrete-time hazard model with two covariates and calculates HE using the deterministic approach.
4. MSLT_SIM estimates age-specific state-dependent transition probabilities using the multinomial logistic regression with no covariate and calculates MSLT functions (including HE) using microsimulation.
5. MSLT_SIMxCOV estimates age-specific state-dependent transition probabilities using the multinomial logistic regression with one or more covariates and calculates MSLT functions (including HE) using microsimulation.

The last set estimates a semi-Markov process model:

6. SMPEM_SIMxCOV estimates age-specific state-duration-dependent transition probabilities using the multinomial logistic regression with one or more covariates, and calculates MSLT functions (including HE) using microsimulation.

Each set of programs has two components: xxx_M and xxx_S. The xxx_M program is the main control program, while the xxx_S is the main statistical module. The xxx_M program first launches the xxx_S program to calculate the point estimates from the full analysis data sample (indexed by BS=0). It then generates a large number of bootstrap samples and calculates the MSLT functions for each of the samples (indexed by BS>=1). The standard errors of the original point estimates are the standard deviations of these bootstrap

estimates. The steps of the computation can be illustrated by the following diagram.



For users with multi-CPU computers, they can use the MSLT_SIMxCOV_M_DX or the SMPEM_SIMxCOV_M_DX program. These two programs perform the estimation of MSLT function for the bootstrap samples in batches. The current version executes four sessions of SAS simultaneously for computers with four CPUs. For computers with more than four CPUs, users can modify the program to increase the number of simultaneous sessions. Note that in order to run these two programs, the users must have SAS Connect installed on their computer.

In the following, I will describe in more detail how these programs work, and the type of data sets these programs require.

Before running the program

- The SPACE program requires SAS/IML (Interactive Matrix Language). If you don't have IML installed on your computer, please install it first.
- The SPACE program is developed and tested in PC SAS 9.1.3 and 9.2. It's not been tested in earlier versions.

- In order to test the program using the sample data included in this package, please first create a library name “S” and save the sample data there.

Data

The data set should be prepared like the SAMPLE data set. Each interview observation should occupy one line of record in the data set. The xxx_S program will convert these records into the format of *annual* (not 12-month) intervals for transition probability or transition rate estimation. The interviews are not required to be equally spaced; the length of time between interviews can vary, as long as they are longer than one year.[1] If the survey has interview gaps of two or more years, then pseudo interview data will be created to “fill in” the unobserved years. If the states occupied in the two successive interviews are different, then an event is assumed to have occurred randomly between the two interviews. For example, if the interviews were conducted in 1996 and 2000, then the timing of event would be randomly assigned to 1997, 1998 or 1999 with the probability of 1/3. If the observed states are identical at two successive interviews, then no event is assumed to have occurred.

In the SAMPLE data set, these variables are included (Some variables are mandatory for any input data set. They are marked by *.):

- ID*: personal identifier
- Age*: age at interview
- Sex: 1=men, 2=women
- Race: 1=white, 2=black
- Edu: 1=0~11 years, 2=high school graduates, 3=some college
- HSQ*: categorical and mutually exclusive health measure (1=active, 2=disabled with 1+ IADL or ADL limitations, 3=dead). Please note that

[1] If the survey has more frequent interviews, the SPACE program needs to be modified. Please contact Liming Cai for such problems.

the number of health states in your study can exceed 3. But death should always be indicated by the largest integer in the defined state space.

- Strata*: indicator of strata in the sample
- PSU*: indicator of PSUs in the sample
- Weight*: cross-sectional weight for the current observation

Please note that the bootstrap samples are generated only from the first-stage sampling (i.e., at the PSU level). If the original survey is a stratified simple random sample, then the PSU variable should be identical to the ID variable so that individuals are treated at PSUs and are selected.

Program Details

MSLT_RAD2COV_M and MSLT_RAD2COV_S

These two programs estimate HE as a function of age and two other covariates. The MSLT_RAD2COV_M launches the MSLT_RAD2COV_S program to first estimates HE from the full analysis sample, then for each of the bootstrap samples. It requires these macro variables to run:

- DATA: name of the full analysis data set
- BSIZE: number of the bootstrap samples to generate
- VAR: name of the health measure variable
- NS: number of health states of the health measure, including dead as the absorbing state. In the sample data, we set ns=3. But users can define more than 3 health states, if data permits.
- COV: list of covariates
- NC: number of covariates
- STRATA: name of the strata variable
- PSU: name of the PSU variable
- WGT: name of the weight variable
- LOI: length of interview cycles (=1 yr, 2 yrs, ...)
- BEG: first age of annual transition and HE estimates

- END: last age of transition. Note that END=95 in RAD_S, RAD1COV_S and RAD2COV_S programs, while END=150 in the SIM_S and SIMxCOV_S programs.

Please note that the bootstrap samples are generated at PSU level. Once a PSU is selected, all sampled persons in that PSU are included in the bootstrap sample and their weights are recalculated by the number of time their PSUs are selected. Also, if there is only a single PSU in a particular stratum, then this single PSU is selected with certainty.

The MSLT_RAD2COV_S program requires only one macro variable DATA, which denotes the name of the input data set – either the full analysis data or the bootstrap sample. Please read the annotated MSLT_RAD2COV_S to better understand each section of the program.

The MSLT_RAD1COV programs estimates HE as a function of age and only one additional covariate. They are similar to the above programs with two covariates and thus are not further discussed here.

Note that the above programs estimate HE separately for each level of the covariates in the case of only one covariate, or each combination of the levels in the case of two covariates.

MSLT_SIMxCOV_M and MSLT_SIMxCOV_S

The MSLT_SIMxCOV programs estimates MSLT functions conditional on one or more covariates. It has three major differences from the MSLT_RAD2COV program. First, the SIMxCOV estimates transition *probabilities* using a multinomial logistic regression (Allison 1982; Allison 1984), while the RAD2COV program estimates a transition *rates* using a hazard model. Second, the SIMxCOV program estimate a variety of MSLT functions via simulation, while the RAD2COV program estimates HE only using the deterministic approach. Simulation produces a large collection of individual health trajectories, and offers researchers much greater flexibility to summarize the dynamics of population health. Third, the SIMxCOV program simulates a single large cohort for each

age. Each cohort is distributed by the health status and covariates measured at the baseline of survey. The RAD2COV program estimate HE separately for each combination of the levels of the two covariates.

The MSLT_SIMxCOV_S program requires these macro variables:

- DATA: name of the input data set, either the full analysis sample or the bootstrap data
- S: if users use MSLT_SIMxCOV_M then S is always 0; if users use MSLT_SIMxCOV_M_DX then S varies: S=0 indicates it is for the full analysis sample, and S=1,2,3,4 indicates it is for one of the four bootstrap samples.
- VAR: name of the health measure variable
- NS: number of health states of the health measure, including death as the absorbing state. In the sample data, we set ns=3. But users can define more than 3 health states, if data permits.
- COV: list of covariates
- NC: number of covariates
- STRATA: name of the strata variable
- PSU: name of the PSU variable
- WGT: name of the weight variable
- LOI: length of interview cycles (=1 yr, 2 yrs, ...)
- BEG: first age of annual transition and HE estimates
- END: last age of transition.
- SIMSIZE: size of simulated cohort. The size is usually set to 100000. For small population subgroups and/or measures of rare events, users may need a higher SIMSIZE. Please note that simulation is computationally intensive and will likely produce a large output file that requires a lot of disk space.

For more details, users can read the annotated PDF version of the program.

The SIMxCOV_M requires these macro variables to generate the bootstrap samples:

- BSIZE: number of the bootstrap samples to generate
- VAR: name of the health measure variable
- NS: number of health states of the health measure, including dead as the absorbing state. In the sample data, we set ns=3. But users can define more than 3 health states, if data permits.
- STRATA: name of the strata variable
- PSU: name of the PSU variable
- WGT: name of the weight variable

The MSLT_SIM programs estimates MSLT functions by age only. They are easier to understand and are not further discussed here.

MSLT_RAD_M and MSLT_RAD_S

These two programs estimate transition rates as a function of age only and calculate MSLT functions via the deterministic approach.

MSLT_SIM_M and MSLT_SIM_S

These two programs estimate transition probabilities as a function of age only and calculate MSLT functions via simulation. They are not discussed in detail here.

SMPEM_SIMxCOV_M_DX and SMPEM_SIMxCOV_S

The SMPEM_SIMxCOV programs estimate a semi-Markov process multi-state model that the transition probabilities are dependent on both current status and the duration of current status. The methodology is described in detail in Cai, Schenker and Lubitz (2006). Like the MSLT programs, the SMPEM_SIMxCOV_M_DX program first launches the SMPEM_SIMxCOV_S program to estimate MSLT functions from the full analysis sample. It then performs estimation in batches of four bootstrap samples. The SMPEM_SIMxCOV_S requires these macro variables:

- DATA: name of the input data set, either the full analysis sample or one of the four bootstrap data sets.
- S: S=0 indicates it is for the full analysis sample, and S=1,2,3,4 indicates it is for one of the four bootstrap samples.
- VAR: name of the health measure variable
- NS: number of health states of the health measure, including death as the absorbing state. In the sample data, we set ns=3. But users can define more than 3 health states, if data permits.
- COV: list of covariates
- NC: number of covariates
- STRATA: name of the strata variable
- PSU: name of the PSU variable
- WGT: name of the weight variable
- LOI: length of interview cycles (=1 yr, 2 yrs, ...)
- BEG: first age of annual transition and HE estimates
- END: last age of transition.
- LOOP: number of iterations for the EM algorithm and the estimation of MSLT functions. See Cai, Schenker and Lubitz (2006) for details.

For more details, users can read the annotated PDF version of the program.

The SMPEM_SIMxCOV_M_DX program requires these macro variables:

- BSIZE: number of the bootstrap samples to generate
- VAR: name of the health measure variable
- NS: number of health states of the health measure, including dead as the absorbing state. In the sample data, we set ns=3. But users can define more than 3 health states, if data permits.
- STRATA: name of the strata variable
- PSU: name of the PSU variable
- WGT: name of the weight variable

SUMMARY

This manual only provides a brief overview of the SPACE program. It is strongly recommended that users test the programs on their computer first using the sample data to become familiar with the program. I will provide as much as trouble shooting as my time allows, but I cannot guarantee to respond to your questions within a certain time window. Please email your questions to me at lcai@cdc.gov.

Enjoy!

Reference:

Allison, P.D. 1982. "Discrete-Time Methods for the Analysis of Event Histories." Pp. 61-98 in *Sociological Methodology*.

—. 1984. *Event history analysis : regression for longitudinal event data*. Beverly Hills, Calif.: Sage Publications.

Cai, L.and J. Lubitz. 2007. "Was There Compression of Disability for Older Americans From 1992 to 2003?" *Demography* 44(3):479-495.

Cai, L., J. Lubitz, M. Hayward, Y. Saito, A. Hagedorn, and E. Crimmins. 2008. "Estimation of Multi-State Life Table Functions and Their Variances Using the SPACE Program." in *Population Association of America*. New Orleans, LA.

Cai, L., N. Schenker, and J. Lubitz. 2006. "Analysis of functional status transitions by using a semi-Markov process model in the presence of left-censored spells." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 55(4):477-491.

Rao, J.N.K.and C.F.J. Wu. 1988. "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83(401):231-241.