The Stata Journal (2023)
23, Number 2, pp. 491–517 DOI: 10.1177/1536867X231175332

Pseudo-observations in a multistate setting

Morten Overgaard
Aarhus University
Aarhus, Denmark
moov@ph.au.dk

Per Kragh Andersen
University of Copenhagen
Copenhagen, Denmark
pka@biostat.ku.dk

Erik Thorlund Parner
Aarhus University
Aarhus, Denmark
parner@ph.au.dk

Abstract. Regression analyses of how state occupation probabilities or expected lengths of stay depend on covariates in multistate settings can be performed using the pseudo-observation method, which involves calculating jackknife pseudo-observations based on some estimator of the expected value of the outcome. In this article, we present a new command, `stpmstate`, that calculates such pseudo-observations based on the Aalen–Johansen estimator. We give examples of use of the command, and we conduct a small simulation study to offer insights into the pseudo-observation regression approach.

Keywords: st0717, stpmstate, multistate model, regression analysis, state occupation probability, length of stay, jackknife, pseudovalues, Aalen–Johansen estimator

1 Introduction

When one studies a setting where participants may transition between several states over time, multistate models provide a suitable framework. A simple example is the illness-death model, where transitions between the three states of disease free, diseased, and dead describe how disease-free study participants may get a disease and later die or how they may die before they get the disease. Inference in multistate models may be based on hazard regression models for the transition intensities—for example, the Cox proportional hazards model—though parameters such as state occupation probabilities and expected lengths of stay are also of interest and may have more direct interpretations than intensities. In the setting of the illness-death model, for example, the expected proportion of diseased individuals at a given time point is a state occupation probability and could be of interest. Even in a situation with right-censored multistate trajectories, estimates of state occupation probabilities and expected lengths of stay can be based on the Aalen–Johansen estimator of transition probabilities. This is essentially a plugin estimator based on nonparametric fits to the intensities; see (1) below. If one wants adjusted comparisons of state occupation probabilities and expected lengths of stay between groups or other types of regression analyses in terms of these outcomes, one may similarly use plugin estimates based on hazard regression models. Another possibility, however, is to use a pseudovalue approach as suggested by Andersen, Klein, and Rosthøj (2003). This has the advantage of providing estimates of parameters directly quantifying the association between each covariate and the outcome in question. Such parameters include odds ratios, risk ratios, risk differences, where odds and risk refer to state occupation, and differences in expected length of stay in certain states.

The pseudovalue approach by Andersen, Klein, and Rosthøj (2003), here called the pseudo-observation method, works by converting the potentially censored trajectories into pseudovalues of the desired outcome, also called pseudo-observations. The pseudo-observations are the jackknife pseudovalues based on leave-one-out estimates from a suitable estimator. These pseudo-observations replace the potentially unobserved individual outcomes, for instance, of state occupation or length of stay, in any regression analysis of interest. The pseudo-observation for an individual relates to the contribution of the individual to the estimate. The idea of the approach is that if the estimand is the mean of an outcome, the pseudo-observation may hold important information on the potentially unobserved outcome for that individual. This piece of intuition certainly works in the uncensored case with an average of the individual outcomes as an estimate of the outcome mean where the individual pseudo-observation turns out to be the individual outcome. The approach cannot be expected to work with any estimator but agrees well with inverse probability of censoring weighted estimators, such as the Kaplan–Meier estimator of survival, under certain assumptions according to the results of Overgaard, Parner, and Pedersen (2017, 2019). When one obtains regression parameter estimates, the suggestion by Andersen, Klein, and Rosthøj (2003) is to use a robust sandwich variance estimate for obtaining standard errors of the regression parameter estimates.

Examples of use of the pseudo-observation method in multistate models have been presented in several articles. Andersen, Klein, and Rosthøj (2003) considered an example with a logistic regression model for state occupation at several time points; the model was essentially a proportional-odds model. It was used with a state of having acute graft-versus-host disease while not having relapsed or died as the state of interest in a bone marrow transplantation example. Andersen and Klein (2007) considered the pseudo-observation method for current leukemia-free survival, which combines state occupation probabilities of two states: first and second remissions. Grand and Putter (2016) considered a regression analysis of expected length of stay based on the pseudo-observation method with an application in expected life in disability. In Spitoni, Lammens, and Putter (2018), the pseudo-observations were not used for a regression analysis but for calculation of prediction errors of predicted state occupation probabilities.

As we describe in more detail below, the pseudo-observations are conceptually not difficult to compute. To give an example, the pseudo-observations of state occupation may be based on the Aalen–Johansen-derived estimates of state occupation probabilities, which can be obtained using the `msaj` command from the `multistate` package for Stata by Crowther and Lambert (2016). However, it is desirable to have Stata commands that allow for easy specifications of states and time points of interest and more efficient calculation of the desired pseudo-observations. Parner and Andersen (2010), updated by Overgaard, Andersen, and Parner (2015), present commands for calculation of pseudo-observations for state occupation and length of stay in the simple special case of survival followed by failure of one or more types; this case is often called the competing-risks case. In this article, we present the command `stpmstate`, which calculates pseudo-observations for state occupation and length of stay based on the

Aalen–Johansen estimator in more general multistate settings. The command allows for calculation of both types of pseudo-observations at various states and several time points simultaneously. In this article, we also offer some theoretical insights into why or when the pseudo-observation method would work in a multistate setting, and this is corroborated in a simulation study. In particular, we note how the Markov assumption is not required for the method to work under the stated censoring assumptions.

Methodological and computational details are found in section 2. In section 3, we present the `stpmstate` command, and examples of its use are given in section 4. In section 5, we conduct a small simulation study to demonstrate some properties and limitations of estimates obtained using the pseudo-observation method in this setting. Finally, we have some closing remarks in section 6.

2 Method

2.1 The general pseudo-observation method

The method considered in this article deals with censoring and allows for regression analysis of censored outcomes on baseline covariates. We call this method the pseudo-observation method because it uses the jackknife pseudo-observations or pseudovalues of a relevant estimator. The general method can be described as follows. Suppose interest is in how an outcome V depends on baseline covariates $\mathbf{Z}$ but V is not always observed because of censoring. Find a reasonable estimator of the expectation $\theta = E(V)$, and calculate the jackknife pseudo-observations based on this estimator. Concretely, if $\widehat{\theta}$ is the estimate based on the entire sample and $\widehat{\theta}^{(i)}$ is the estimate obtained by using the sample where observation i has been left out, the ith jackknife pseudo-observation is $\widehat{\theta}_i = n\widehat{\theta} - (n-1)\widehat{\theta}^{(i)}$, where n is the sample size. The main idea of the method is that the pseudo-observations may carry information on the association between V and $\mathbf{Z}$, and the next step is to use the pseudo-observations as the outcomes in the relevant regression analysis, replacing the potentially censored outcomes V_i, to estimate the parameters in a model of how the expectation of V depends on covariates, $E(V \mid \mathbf{Z}) = \mu(\boldsymbol{\beta}; \mathbf{Z})$. The outcome V could be multivariate—for instance, a status at several different time points—but we will focus on the univariate case.

An example of the method described above is with V indicating $T > t$ for a survival time T and time point of interest $t > 0$. In this case, the estimate, $\widehat{\theta}$, could be the Kaplan–Meier estimate of the survival probability. Calculations in this case can be carried out using the `stpsurv` command introduced in Parner and Andersen (2010) and with an update described by Overgaard, Andersen, and Parner (2015). Other examples are handled by `stpmean` and also `stpci` and `stplost` in a competing-risks setting, as described in the referenced articles.

When the pseudo-observations have been calculated, a wide variety of models can be fit using the generalized linear models framework of the `glm` command. If g is the link function in the generalized linear model framework, the model of how the expectation of V depends on covariates is $\mu(\boldsymbol{\beta}; \mathbf{Z}) = g^{-1}(\boldsymbol{\beta}^T \mathbf{Z})$ in this case. We cannot expect the pseudo-observations to follow any standard distribution as can be specified by the `family()` option, and because we are concerned only with the aspect of the model involving the conditional expectation and not the conditional distribution of the outcome, it is appropriate to use the robust sandwich variance estimator by specifying the `vce(robust)` option to obtain robust standard errors. Using such robust standard errors was suggested by Andersen, Klein, and Rosthøj (2003) in accordance with the generalized estimating equation approach of Liang and Zeger (1986). According to Jacobsen and Martinussen (2016), Overgaard, Parner, and Pedersen (2017), and Overgaard, Parner, and Pedersen (2018), even the robust standard error is not exactly asymptotically unbiased. In the settings of those articles, the bias was seen to be upward, leading to conservative inference, and the size of the bias very much related to the size of the effect of covariates on the outcome and the amount of censoring up to a time point of interest. The bias of the robust standard error seemed to be minor in many cases. Although the articles mentioned above present coverage probabilities of corresponding 95% confidence intervals above 96% and 97% in some scenarios, these scenarios must be considered rather extreme.

An important requirement for the pseudo-observation method to work is that the pseudo-observations must have the appropriate conditional expectation. Somewhat informally, this can be stated as $E(\widehat{\theta}_i \mid \mathbf{Z}_i) \approx E(V_i \mid \mathbf{Z}_i)$. More formally, as described in Graw, Gerds, and Schumacher (2009) and Overgaard, Parner, and Pedersen (2017), the requirement is that $E\{\dot{\theta}(X_i) \mid \mathbf{Z}_i\} = E(V_i \mid \mathbf{Z}_i) - E(V_i)$ where $\dot{\theta}$ is the influence function of the estimator leading to $\widehat{\theta}$ and X_i refers to the observable information on individual i used by the estimator. There is a close connection between pseudo-observation and influence function when a reasonable, consistent estimator is considered, namely, that $\widehat{\theta}_i \approx \theta + \dot{\theta}(X_i)$ when the sample size is not too small. In the simple example where V is the binary indicator of survival to a time point t, $T > t$, and pseudo-observations are based on the Kaplan–Meier estimator and other similar examples, the requirement of $E\{\dot{\theta}(X_i) \mid \mathbf{Z}_i\} = E(V_i \mid \mathbf{Z}_i) - E(V_i)$ was seen by Graw, Gerds, and Schumacher (2009) to be fulfilled under an assumption of independence between censoring time and event time, as well as covariates. Overgaard, Parner, and Pedersen (2017) called this an assumption of completely independent censorings. Overgaard, Parner, and Pedersen (2019) demonstrate how inverse probability of censoring weighted estimators satisfies the requirement of $E\{\dot{\theta}(X_i) \mid \mathbf{Z}_i\} = E(V_i \mid \mathbf{Z}_i) - E(V_i)$ under the completely independent censorings assumption.

2.2 The pseudo-observation method in multistate settings

We will consider the case where V_i is the binary outcome of state occupation (that is, being in a certain state at a certain time) or the restricted length of stay (that is, the time spent in a certain state up to some time point) in a multistate setting. Such outcomes may be left unobserved because of censoring. The estimators considered here are the Aalen–Johansen-derived estimators of state occupation probabilities and expected (restricted) length of stay in such a multistate setting. In line with Gill and Johansen (1990), we make use of the product integral, a limit of products, with the notation $\prod$. The Aalen–Johansen-derived estimate of the state occupation probabilities is the row vector

$$\widehat{p}(t) = \widehat{p}(0) \prod_{0}^{t} \left\{ \mathbf{I} + \widehat{\Lambda}(du) \right\} \tag{1}$$

where $\widehat{p}(0)$ is the empirical estimate of the initial state occupation probabilities and $\widehat{\Lambda}$ is the matrix of Nelson–Aalen estimates of cumulative forces of transition. Because the Nelson–Aalen estimates jump only at transition times and are constant between jumps, the product integral $\prod_0^t \{\mathbf{I} + \widehat{\Lambda}(du)\}$ corresponds to the ordinary matrix product $\prod_{u \in (0,t]} \{\mathbf{I} + \Delta\widehat{\Lambda}(u)\}$, where only a transition time is a relevant u in the product. Estimates of the expected length of stay in a state j up to time t are obtained by $\int_0^t \widehat{p}_j(u)du$; or in other words, $\int_0^t \widehat{p}(u)du$ is the vector of such estimates for each state.

Using the pseudo-observation method now involves calculating the jackknife pseudo-observations based on the estimators mentioned above; for instance, $\widehat{\theta}_i = n\widehat{p}_j(t) - (n-1)\widehat{p}_j^{(i)}(t)$ would be a pseudo-observation to replace the potentially unobserved outcome of individual i being in state j at time t. Because state occupation is a binary outcome, models from binomial regression are of interest for this outcome. In the generalized linear model framework, a logit link results in the model $\mu(\boldsymbol{\beta}; \mathbf{Z}) = \text{expit}(\boldsymbol{\beta}^T\mathbf{Z})$ for estimation of state occupation odds ratios, a log link results in $\mu(\boldsymbol{\beta}; \mathbf{Z}) = \exp(\boldsymbol{\beta}^T\mathbf{Z})$ for estimation of state occupation probability ratios, and an identity link results in $\mu(\boldsymbol{\beta}; \mathbf{Z}) = \boldsymbol{\beta}^T\mathbf{Z}$ for estimation of state occupation probability differences. For the outcome of length of stay, the identity and log link are of interest for estimation of differences and ratios of expected length of stay up to some time point. Here focus has been on a single time point of interest, t, but pseudo-observations corresponding to multiple time points may be calculated and form a multivariate outcome for each individual if such an outcome is considered of interest.

In the following paragraphs, we would like to offer a few insights into why and when the pseudo-observation method is appropriate in this setting. The influence function of the estimate $\widehat{p}(t)$ can be stated as

$$\begin{aligned}\dot{p}(t; X) = \dot{p}(0; X) \prod_{0}^{t} \{\mathbf{I} + \Lambda^{\mathrm{c}}(du)\} \\ + p^{\mathrm{c}}(0) \int_0^t \prod_{0}^{s-} \{\mathbf{I} + \Lambda^{\mathrm{c}}(du)\} \, \dot{\Lambda}(ds; X) \prod_{s}^{t} \{\mathbf{I} + \Lambda^{\mathrm{c}}(du)\}\end{aligned}$$

where $\dot{p}(0; X)$ and $\dot{\Lambda}(\cdot; X)$ refer to the influence functions of the empirical estimate of the initial distribution and the Nelson–Aalen estimates and where p^{c} and Λ^{c} refer to the limits of those estimators. More details are given in the appendix. The influence function of the estimate of expected length of stay can be derived from the influence function of $\widehat{p}(t)$ and is $\int_0^t \dot{p}(u; X)du$.

When censoring is independent of the multistate process and allows for possible observation of the multistate process up to time point t, we have the consistency properties $p^{\mathrm{c}}(0) = p(0)$ and $\Lambda^{\mathrm{c}}(s) = \Lambda(s)$, which ensure that $\widehat{p}(t)$ estimates $p(t) = p(0) \prod_0^t \{\mathbf{I}+\Lambda(du)\}$ consistently. See, for instance, Overgaard (2019). Under an assumption of completely independent censoring, which here means that the censoring time is independent of the multistate process and covariates, $E\{\dot{p}(t; X_i) \mid \mathbf{Z}_i\} = p(t \mid \mathbf{Z}_i) - p(t)$ also holds, as argued in the appendix. The same property then holds for the influence function of the estimator of expected length of stay, owing to linearity of the integral. In other words, the main requirement for the pseudo-observation method to work with either of the two outcome types is fulfilled under the completely independent censoring assumption.

A Markov assumption can be used to establish the identity of the transition probability matrix $P(s,t)$ and the product integral $\prod_s^t \{\mathbf{I} + \Lambda(du)\}$; see, for example, Aalen and Johansen (1978). This helps explain the consistency of the Aalen–Johansen estimate of the transition probabilities, $\prod_s^t \{\mathbf{I} + \widehat{\Lambda}(du)\}$, and then the consistency of $\widehat{p}(t) = \widehat{p}(0) \prod_0^t \{\mathbf{I}+\widehat{\Lambda}(du)\}$ estimating $p(t) = p(0)P(0,t) = p(0) \prod_0^t \{\mathbf{I}+\Lambda(du)\}$. As noted by Datta and Satten (2001), the Aalen–Johansen-derived estimate of state occupation probabilities and thereby length of stay is consistent even without the Markov assumption. Essentially, this is because $p(t) = p(0) \prod_0^t \{\mathbf{I}+\Lambda(du)\}$ continues to hold without the Markov assumption even though $P(0,t) = \prod_0^t \{\mathbf{I} + \Lambda(du)\}$ cannot be expected to hold. Owing to continuity properties of the product integral, assumptions on the censoring mechanism to ensure consistency of $\widehat{p}(0)$ and $\widehat{\Lambda}$ are then sufficient to establish consistency of $\widehat{p}(t)$. This and some further details are discussed in Overgaard (2019), Maltzahn et al. (2020), and Niessl et al. (2020). Similarly, the pseudo-observation method in multistate models, considering state occupation probabilities and expected lengths of stay, does not rely on a Markov assumption.

2.3 Computational approach

The pseudo-observation method requires recalculation of an estimate n times and will be computationally demanding in larger samples. Additionally, the current Stata and Mata built-in tools do not appear to allow for a very vectorized calculation of estimates like those based on the Aalen–Johansen estimator, where a running matrix product would be useful. Specifically, $\prod_s^t \{\mathbf{I} + \widehat{\Lambda}(du)\} = \prod_{u \in (s,t]} \{\mathbf{I} + \Delta\widehat{\Lambda}(u)\}$ seemingly requires a number of matrix multiplications equal to the number of distinct transition times. We therefore consider it worthwhile to look for ways to reduce the number of required operations when calculating the pseudo-observations mentioned above.

As noted by Andersen et al. (1993) in section IV.4.1.4., the estimate $\widehat{p}(t)$ is simply the empirical distribution in the case of no censoring before time point t. The need for matrix multiplication is entirely eliminated in this case, but this is also a case where the pseudo-observations are trivial, $\widehat{\theta}_i = V_i$, and the pseudo-observation method would have no problem to solve and be of no interest. This example does, however, illustrate that the number of computations can be reduced considerably in some settings.

In our computational approach, we consider a setting where individuals may enter and exit the study several times. This is a more general setting than individuals having one potential exit due to right censoring, as considered earlier. In the following, we let $Y_j(s)$ denote the number of individuals observed to be in state j at time s and $N_{jk}(s,t)$ denote the number of transitions from j to k in the time interval $(s,t]$. With this notation, $\widehat{\Lambda}$ is given by $\widehat{\Lambda}_{jk}(t) = \int_0^t Y_j(s-)^{-1} N_{jk}(ds)$ off the diagonal and $\widehat{\Lambda}_{jj}(t) = -\sum_{k \neq j} \widehat{\Lambda}_{jk}(t)$ on the diagonal. Here $Y_j(s-)$ is then the number of individuals in state j immediately before time s and can be considered the number of individuals at risk of a transition from state j at time s. In our computational approach, we take advantage of the fact that

$$\prod_s^t \left\{ \mathbf{I} + \widehat{\Lambda}(du) \right\} = \mathbf{I} + H(s,t)$$

where $H_{jk}(s,t) = Y_j(s)^{-1} N_{jk}(s,t)$ off the diagonal and $H_{jj} = -\sum_{k \neq j} H_{jk}$ on the diagonal when there are transitions only from one state and no censorings from that state and no entries into that state in $(s,t]$. For a given time point, t, this allows for a coarse partitioning $0 = u_0 < u_1 < \cdots < u_m = t$ of the interval $(0,t]$ such that $\prod_0^t \left\{ \mathbf{I} + \widehat{\Lambda}(du) \right\} = \prod_{i=1}^m \{ \mathbf{I} + H(u_{i-1}, u_i) \}$. Concretely, the partition is found by the following procedure:

1. Start with the set of time points, $s_1 < \cdots < s_k$, of any occurrence, be it a transition, censoring, or entry.
2. Associate each s_i with an active transition state where the most recent transition is from, including potentially at s_i, if applicable.
3. Reduce to the subset of points, $s_{i_1} < \cdots < s_{i_{k'}}$ that have been marked by one of the following points:
   a. Mark s_i if a censoring occurs from or an entry occurs into the active transition state at s_i.
   b. Mark s_i if a change in the active transition state occurs at s_{i+1}.
   c. Mark s_i if transitions from different states occur at s_i.
4. Obtain the final subset, $u_1 < \cdots < u_m$, by removing s_{i_j} if no transitions occur in $(s_{i_j}, s_{i_{j+1}}]$.

With this partition in hand, $\widehat{p}(t)$ can be obtained for any t by $\widehat{p}(t) = \widehat{p}(0)\prod_{u_i<t}\{\mathbf{I} + H(u_{i-1}, u_i)\}\{\mathbf{I} + H(u_j, t)\}$, where u_j is the largest u_i smaller than t. The estimate of expected length of stay up to time t is then obtained by $\int_0^t \widehat{p}(s)ds = \sum_{s_i<t}\widehat{p}(s_{i-1})(s_i - s_{i-1}) + \widehat{p}(s_j)(t - s_j)$, where s_j is the largest s_i smaller than t. Here $s_1, \ldots, s_k$ refer to the similarly named partition mentioned above, but it could be replaced by the distinct transition times where the estimate $\widehat{p}(\cdot)$ changes.

The number of calculations of the actual pseudo-observations is reduced by taking advantage of the fact that individuals with the same type of transitions, entries, or censorings in each of the intervals $(u_{i-1}, u_i]$ contribute very similarly to the $\widehat{p}(t)$ estimates even if the transitions, entries, or censorings of the individuals in the various $(u_{i-1}, u_i]$ intervals do not occur at the same time point. If we make sure t is on the list of u_i, the contribution to the calculation of $\widehat{p}(t)$ is in fact the same and the pseudo-observations for state occupation at time t will be the same for such individuals. For pseudo-observations for expected length of stay up to time t, individuals with the same contribution to each $\widehat{p}(s_i)$ up to time t will have the same pseudo-observation according to the calculation above. This is ensured under the stricter requirement that the individuals make transitions at the same time points.

3 The stpmstate command

In the following section, we describe the new `stpmstate` command, which calculates pseudo-observations for state occupation and length of stay in a multistate setting. The command requires the data to be `stset` such that risk sets can be determined using variables `_t0` and `_t`. Transitions are specified using the `from()` and `to()` options as described below. An individual is understood to have been in the state of the variable specified by `from()` between `_t0` and `_t` and then to have made a transition to the state of the variable specified in `to()` at time `_t`. If the state of the two variables of `from()` and `to()` are the same, the individual is understood to have exited at time `_t` without making a transition; this is the way to specify a censoring. The failure information of `_d` is not used. The dataset may contain multiple transitions per individual in a long format in this manner.

3.1 Syntax

`stpmstate` *newvar* `=` {`p(`*state*`)` | `los(`*state*`)`} [*newvar* `= ...`] [*if*] [*in*] [*weight*] `,`
`at(`*numlist*`)` `from(`*varname*`)` `to(`*varname*`)` [`by(`*varlist*`)` `id(`*varname*`)`
`atnumbers` `placement(`*place*`)` `replace`]

where *newvar* is the name of the new variable to be generated containing pseudo-observations or the stub of the new variables if multiple time points are specified by `at()`. Specifying `p(`*state*`)` results in pseudo-observations for state occupation for the state with name specified by *state* at time points specified by `at()` as described below, whereas `los(`*state*`)` results in pseudo-observations for a length of stay up to time points specified by `at()`.

You must `stset` your data before using `stpmstate`; see [ST] **stset**. `fweights`, `aweights`, `iweights`, and `pweights` are allowed; see [U] **11.1.6 weight**.

3.2 Options

`at(`*numlist*`)` specifies the time points at which pseudo-observations are to be calculated. `at()` is required.

`from(`*varname*`)` specifies the variable containing the states where transitions are from. `from()` is required.

`to(`*varname*`)` specifies the variable containing the states where transitions are to. `to()` is required.

`by(`*varlist*`)` specifies that calculation of pseudo-observations be performed separately in the groups defined by *varlist*.

`id(`*varname*`)` specifies the variable identifying the individuals that the leave-one-out procedure is based on. The default is the ID variable from the preceding `stset` command if available, and observations are considered separate individuals otherwise.

`atnumbers` specifies that names of generated variables be suffixed by the corresponding time point of the `at()` list when multiple time points are specified by `at()` rather than the $1, 2, \ldots, k$ default.

`placement(`*place*`)` specifies on which row the pseudo-observations for an ID are to be placed. Possible values of *place* are `first` for placement on the earliest entry, `last` for placement on the latest entry, and `all` for placement on all entries of that ID. The default is `placement(first)`.

`replace` specifies that generated variables can replace existing variables without error.

3.3 Notes

Any weights from the preceding `stset` command will be carried over to `stpmstate` unless other weights are specified in the `stpmstate` statement itself. Weights are handled as if they were frequency weights. This means that n in the calculation is the sum of the weights and that one unit of weight of ID i is left out rather than all of i in the calculation of the pseudo-observations.

Generated variables are equipped with characteristics (see [P] **char**), with information on what type of pseudo-observation they hold, for which state, at what time point, and potentially by which variable the calculation was stratified.

The state variables of `from()` and `to()` may be numeric or string variables. The state variables and the states specified in the command are internally converted to strings.

For instance, a state denoted by a numeric `1` is considered identical to a state denoted by the string `"1"`.

The command allows depleted risk sets without error. The calculations remain in line with the description in section 2. Such depleted risk sets may cause bias in the estimates and thereby in the pseudo-observation method, and we strongly recommend applying the method only to time points in a range where ample information is available.

Any information on transitions after the last time point of interest is not used by the Aalen–Johansen estimates up to that time point and therefore has no influence on the calculation of the pseudo-observations. The command will generally censor any information after this point internally, improving computational speed.

The intended use of the command is in the case where all individuals are available at time 0, but the command will not produce an error if this is not the case. If no individual is available at time 0, the command will look for the earliest entry to play the role as 0 in the computations mentioned earlier. Length of stay then refers to time since this earliest entry.

4 Examples

To illustrate the use of the `stpmstate` command, we consider the bone marrow transplantation dataset `ebmt4.csv` from the R package `mstate`. According to van Houwelingen and Putter (2008), who also use this dataset and are a source for the following description, the dataset is obtained from the European Group for Blood and Marrow Transplantation registry. The dataset consists of times, in days, from transplantation to events such as platelet recovery, adverse event (acute graft-versus-host disease), relapse, and death for 2,279 leukemia patients who had a bone marrow transplantation between 1985 and 1998. The variables holding these times to events are called `rec`, `ae`, `rel`, and `srv`, with `recs`, `aes`, `rels`, and `srvs` indicating whether an event occurred. The dataset also contains information on covariates for the patients such as age at transplantation (categorized as ≤ 20, 20–40, > 40), year of transplantation (categorized as 1985–1989, 1990–1994, 1995–1998), whether prophylaxis was used, and whether the donor was a gender match or mismatch. The variable names are `agecl`, `year`, `proph`, and `match`.

We will consider a multistate model with six states for these data: transplanted, which is the initial state; adverse event, where individuals have experienced an adverse event but not platelet recovery; platelet recovery, where individuals have experienced platelet recovery but not an adverse event; adverse event and platelet recovery, where individuals have experienced both an adverse event and platelet recovery; relapse, where individuals have relapsed; and death, where individuals have died. This model and possible transitions are illustrated in figure 1. In Stata, we set this up as follows:

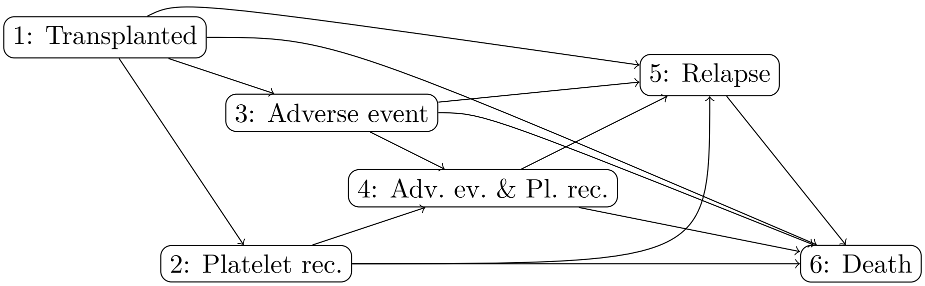

Figure 1. The multistate model of the bone marrow transplantation example

```
. import delimited
> "https://vincentarelbundock.github.io/Rdatasets/csv/mstate/ebmt4.csv"
(encoding automatically selected: ISO-8859-1)
(16 vars, 2,279 obs)
. foreach var of varlist year agecl proph match {
  2.         egen `var'_cat = group(`var'), label
  3. }
. label define match_cat 1 "mismatch" 2 "match", modify
. stset srv, failure(srvs) id(id)
Survival-time data settings
           ID variable: id
         Failure event: srvs!=0 & srvs<.
Observed time interval: (srv[_n-1], srv]
     Exit on or before: failure
--------------------------------------------------------------------------
      2,279  total observations
          0  exclusions
--------------------------------------------------------------------------
      2,279  observations remaining, representing
      2,279  subjects
        838  failures in single-failure-per-subject data
  3,826,341  total analysis time at risk and under observation
                                                At risk from t =         0
                                     Earliest observed entry t =         0
                                          Last observed exit t =     6,299
. foreach var of varlist rec ae rel srv {
  2.         stsplit post`var' if `var's == 1, at(0) after(`var')
  3.         recode post`var' (0 = 1) (else = 0)
  4. }
(1,218 observations (episodes) created)
(3,497 changes made to postrec)
(1,134 observations (episodes) created)
(4,631 changes made to postae)
(347 observations (episodes) created)
(4,978 changes made to postrel)
(no new episodes generated)
(4,978 changes made to postsrv)
. label define statelbl 1 "Transplanted" 2 "Platelet rec."
>         3 "Adverse event"
>         4 "Adv. ev. & Pl. rec."
>         5 "Relapse" 6 "Death"
```

```
. generate fromstate = 1
. replace  fromstate = 2 if postrec & !postae
(896 real changes made)
. replace  fromstate = 3 if postae  & !postrec
(961 real changes made)
. replace  fromstate = 4 if postae  &  postrec
(760 real changes made)
. replace  fromstate = 5 if postrel
(347 real changes made)
. generate tostate = fromstate
. by id (_t), sort: replace tostate = fromstate[_n + 1] if _n < _N
(2,699 real changes made)
. replace  tostate = 6 if _d == 1
(838 real changes made)
. label values fromstate tostate statelbl
```

At this stage, the 2,279 individuals are split into 4,978 rows, with 3,537 rows representing observed transitions between states, including 838 transitions into the absorbing death state, whereas the remaining 1,441 rows represent censorings. In other words, of the 2,279 individuals, 838 are eventually observed to die, and the remaining 1,441 individuals are lost to follow-up beforehand. This censoring problem is the issue we will be handling using the pseudo-observation method. These censorings essentially occur over the entire follow-up period, the maximal follow-up time being 6,299 days or 17.2 years. For example, 543 censorings occur before 5 years of follow-up.

To give an example of what the data look like when they are set up like this, we can look at the first two individuals.

```
. list id _t0 _t fromstate tostate if id == 1 | id == 2, sepby(id)
```

	id	_t0	_t	fromstate	tostate
1.	1	0	22	Transplanted	Platelet rec.
2.	1	22	995	Platelet rec.	Platelet rec.
3.	2	0	12	Transplanted	Adverse event
4.	2	12	29	Adverse event	Adv. ev. & Pl. rec.
5.	2	29	422	Adv. ev. & Pl. rec.	Relapse
6.	2	422	579	Relapse	Death

This is understood as follows. Individual 1 transitions to state 2 at day 22 and is then censored in state 2 at day 995. Individual 2 transitions to state 3 at day 12, then to state 4 at day 29, then to state 5 at day 422, and finally to state 6 at day 579. The individuals are in `fromstate` between `_t0` and `_t` and contribute to the risk set in that state in that interval. Note that this is different from the way the `multistate` package would have had the data set up. The `multistate` package requires an entry for each potential transition of the individual to work with transition-specific risk sets and other things. The way the data are set up for `stpmstate` above is sufficient to determine the size of risk sets and number of transitions necessary in the computational approach described earlier because risk sets are state specific but not transition specific.

4.1 State occupation probability example

Let us take an interest in the probability of having had platelet recovery without adverse events, relapse, or death occurring yet after 5 years, that is, being in state 2 in figure 1. Because this outcome is unknown for individuals who are lost to follow-up before 5 years, we calculate pseudo-observations for this outcome using `stpmstate` as follows:

```
. stpmstate pseudo = p(2), at(`=5*365.25') from(fromstate) to(tostate)
Computing pseudo-observations (progress dots indicate percent completed).
────┼──── 1 ────┼──── 2 ────┼──── 3 ────┼──── 4 ────┼──── 5
..................................................    50
..................................................   100
```

As described earlier, the specification `pseudo = p(2)` means we want to generate a variable called `pseudo`, which, because we have specified `p()`, should hold pseudo-observations for state occupation, and this is desired for state 2. The time of interest for the pseudo-observation is controlled by the `at()` option, where `` `=5*365.25' `` gives us approximately 5 years because the time scale was kept in days above. Note that we could have generated various variables with pseudo-observations by adding more specification, for instance, `stpmstate ps2 = p(2) ps3 = p(3)` for pseudo-observations for state occupation for both state 2 and 3. Also, if we wanted pseudo-observations at several time points of interest, we could have specified a *numlist* in `at()`—for instance, `at(100(100)400)`—if time points 100, 200, 300, and 400 are of interest. A specification with `pseudo = p(2), at(100(100)400)` would have generated 4 variables—`pseudo1`, `pseudo2`, `pseudo3`, `pseudo4`—using `pseudo` as a stub and `1`, `2`, `3`, and `4` as suffixes unless the `atnumbers` option had been used. Specifying `atnumbers` would have led to the generation of variables `pseudo100`, `pseudo200`, `pseudo300`, and `pseudo400`, which could be useful for a subsequent reshape of the data.

After generating the `pseudo` variable above, suppose we are interested in how the state occupation probability for state 2 at 5 years depends on how the covariates age at transplantation and gender match. In that case, we can now do a `glm` with the pseudo-observations as the outcome variable and with the covariates specified as usual. As discussed above, using robust standard errors is expected to be fairly appropriate. We specify a log link below to consider a model where covariates influence the probability by ratios, $p_j(t \mid \mathbf{Z}) = \exp(\boldsymbol{\beta}^T\mathbf{Z})$ for state $j = 2$ at $t = 5$ years, where $\mathbf{Z}$ denotes the vector of age category and gender match indicators in addition to a constant term. In its current format, the dataset has multiple rows per individual, but because the pseudo-observation variable is nonmissing only on one of these by default, the earliest, we can proceed with the `glm` procedure without reformatting or making restrictions.

```
. glm pseudo i.agecl_cat i.match_cat, eform link(log) vce(robust) baselevels

Iteration 0:  Log pseudolikelihood = -2658.4499
Iteration 1:  Log pseudolikelihood =  -1811.734
Iteration 2:  Log pseudolikelihood = -1083.2436
Iteration 3:  Log pseudolikelihood = -1082.8065
Iteration 4:  Log pseudolikelihood = -1082.8061
Iteration 5:  Log pseudolikelihood = -1082.8061

Generalized linear models                         Number of obs   =      2,279
Optimization     : ML                             Residual df     =      2,275
                                                  Scale parameter =   .1516965
Deviance         =  345.1094552                   (1/df) Deviance =   .1516965
Pearson          =  345.1094552                   (1/df) Pearson  =   .1516965

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =    .953757
Log pseudolikelihood = -1082.806075               BIC             =  -17244.03

------------------------------------------------------------------------------
             |               Robust
      pseudo |     exp(b)   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
   agecl_cat |
       20-40 |          1  (base)
        <=20 |   1.104623   .1204473     0.91   0.361     .8920714    1.367818
         >40 |   1.013094   .1178623     0.11   0.911     .8065327    1.272558
             |
   match_cat |
    mismatch |          1  (base)
       match |   1.266397   .1476224     2.03   0.043     1.007735    1.591451
             |
       _cons |   .1451825   .0160501   -17.46   0.000     .1168995    .1803083
------------------------------------------------------------------------------
```

According to this analysis, to take an example, the probability of being in the platelet recovery state at 5 years since transplantation in the $\leq$ 20 years category is a factor 1.10 (confidence interval $[0.89, 1.37]$) higher than in the reference category 20–40 years when comparing on a fixed level of gender match category. The intercept parameter, estimated at 0.145 (confidence interval $[0.117, 0.180]$), refers to the probability of being in the platelet recovery state at 5 years since transplantation in the reference category of 20–40 years at transplantation and gender mismatch.

We could have specified another link function in the `glm` command if we wanted to study a different model. For this outcome, an interesting option is the logit link function if state occupation odds and odds ratios are desired.

4.2 Length-of-stay example

To illustrate the use of pseudo-observations of length of stay, suppose we are now interested in the amount of time spent in remission having had an adverse event within the first five years after transplantation and how this amount of time depends on the use of prophylaxis. In the multistate model of figure 1, we consider two states in remission having had an adverse event, namely, one with platelet recovery and one without. In the following, we calculate pseudo-observations for length of stay restricted to five years after transplantation for each of these states. The two pseudo-observation variables are then combined to form the relevant pseudo-observation. Because the amount of time spent in one of the two states is the sum of the amounts of time spent in the two states, or more precisely because a natural estimate of the expected length of stay is also the sum of the two separate estimates, the pseudo-observation can also be obtained by the corresponding sum.

```
. stpmstate ps_los_ae = los(3) ps_los_recae = los(4), at(`=5*365.25')
>         from(fromstate) to(tostate)
Computing pseudo-observations (progress dots indicate percent completed).
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
..................................................   100
. generate pseudo_los = ps_los_ae + ps_los_recae
(2,699 missing values generated)
```

With the pseudo-observations in hand, we can now fit a model of our choosing. If we wanted to estimate the difference in expected length of stay associated with prophylaxis use while adjusting for factors such as age at transplantation and whether the donor was a gender match or mismatch, we could consider a linear model, $E(V_i \mid \mathbf{Z}_i) = \boldsymbol{\beta}^T\mathbf{Z}_i$, where covariates $\mathbf{Z}_i$ consist of observations on the mentioned variables in addition to a constant term and where V_i refers to the potentially unobserved length of stay up to five years. Such a model can be fit using the default identity link of the `glm` command as follows.

```
. glm pseudo_los i.proph_cat i.agecl_cat i.match_cat, vce(robust) baselevels
Iteration 0:  Log pseudolikelihood = -18474.294
Generalized linear models                         Number of obs   =      2,279
Optimization     : ML                             Residual df     =      2,274
                                                  Scale parameter =   644965.1
Deviance         =   1466650545                   (1/df) Deviance =   644965.1
Pearson          =   1466650545                   (1/df) Pearson  =   644965.1
Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]
                                                  AIC             =   16.21702
Log pseudolikelihood = -18474.29432               BIC             =   1.47e+09

------------------------------------------------------------------------------
             |               Robust
  pseudo_los | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
   proph_cat |
         no  |          0  (base)
        yes  |  -192.6659     36.836    -5.23   0.000    -264.8631   -120.4687
             |
   agecl_cat |
      20-40  |          0  (base)
       <=20  |    13.8786    41.9529     0.33   0.741    -68.34757    96.10477
        >40  |  -28.78528    41.5118    -0.69   0.488    -110.1469    52.57635
             |
   match_cat |
   mismatch  |          0  (base)
      match  |   9.842872   39.10605     0.25   0.801    -66.80358    86.48932
             |
       _cons |   616.5214   39.25449    15.71   0.000     539.5841    693.4588
------------------------------------------------------------------------------
```

We see that, in this adjusted analysis, prophylaxis use is associated with a decrease of 193 (confidence interval $[120, 265]$) days spent having had an adverse event in remission within the first 5 years after transplantation.

On second thought, we might suspect that year of transplantation is an important confounder. We do have information on year of transplantation in categories, and suppose we now want to adjust for this categorical variable. We can simply use the same pseudo-observations in the new model.

```
. glm pseudo_los i.proph_cat i.agecl_cat i.match_cat i.year_cat, vce(robust)
> baselevels
Iteration 0:  Log pseudolikelihood = -18467.843
Generalized linear models                         Number of obs   =      2,279
Optimization     : ML                             Residual df     =      2,272
                                                  Scale parameter =   641888.6
Deviance         =   1458370888                   (1/df) Deviance =   641888.6
Pearson          =   1458370888                   (1/df) Pearson  =   641888.6
Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]
                                                  AIC             =   16.21311
Log pseudolikelihood = -18467.84329               BIC             =   1.46e+09

------------------------------------------------------------------------------
             |               Robust
  pseudo_los | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
   proph_cat |
         no  |          0  (base)
        yes  |   -163.394   38.70929    -4.22   0.000    -239.2628   -87.52515
             |
   agecl_cat |
      20-40  |          0  (base)
       <=20  |   17.47392     41.812     0.42   0.676    -64.47609    99.42393
        >40  |  -45.61846   42.51294    -1.07   0.283    -128.9423    37.70538
             |
   match_cat |
   mismatch  |          0  (base)
      match  |   9.053128   39.15273     0.23   0.817     -67.6848    85.79106
             |
    year_cat |
  1985-1989  |          0  (base)
  1990-1994  |    153.408   42.54835     3.61   0.000     70.01479    236.8013
  1995-1998  |   84.95325   44.82027     1.90   0.058    -2.89286    172.7994
             |
       _cons |   524.7723   48.66241    10.78   0.000     429.3957    620.1488
------------------------------------------------------------------------------
```

This changes the conclusion. Now prophylaxis use is associated with a decrease of 163 (confidence interval $[88, 239]$) days spent having had an adverse event in remission within the first 5 years after transplantation in this new adjusted analysis.

We have mentioned how covariate-independent censoring is a requirement as part of the completely independent censorings assumption. We ought to consider whether the completely independent censorings assumption is reasonable and, in particular, check whether censoring is in fact independent of covariates. In this regard, the added variable concerning the year of transplantation is particularly suspect because the censoring time may well simply be the time from transplantation to an end-of-study calendar time and is thus completely determined by the time of transplantation. As we discuss further in section 6, approaches to alleviate the problem exist, and a suggestion by Andersen and Pohar Perme (2010) is to base pseudo-observations on a mixture estimator, combining estimates from strata of a variable that censoring depends on. This approach is equivalent to calculating pseudo-observations in each stratum, and it is an approach we can also take in the multistate setting considered here. The calculation is easily carried out using the `by()` option of `stpmstate`, as demonstrated below.

```
. stpmstate ps_los_ae_by = los(3) ps_los_recae_by = los(4) , at(`=5*365.25')
>         from(fromstate) to(tostate) by(year_cat)
Computing pseudo-observations (progress dots indicate percent completed).
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
..................................................   100

. generate pseudo_los_by = ps_los_ae_by + ps_los_recae_by
(2,699 missing values generated)
```

With the new pseudo-observations in hand, we can fit the same model as above.

```
. glm pseudo_los_by i.proph_cat i.agecl_cat i.match_cat i.year_cat, vce(robust)
> baselevels

Iteration 0:  Log pseudolikelihood = -18469.275

Generalized linear models                         Number of obs   =      2,279
Optimization     : ML                             Residual df     =      2,272
                                                  Scale parameter =   642695.4
Deviance         =   1460203922                   (1/df) Deviance =   642695.4
Pearson          =   1460203922                   (1/df) Pearson  =   642695.4

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   16.21437
Log pseudolikelihood = -18469.27464               BIC             =   1.46e+09

------------------------------------------------------------------------------
             |               Robust
pseudo_los~y | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
   proph_cat |
         no  |          0  (base)
        yes  |  -164.5535   38.64405    -4.26   0.000    -240.2944   -88.81251
             |
   agecl_cat |
      20-40  |          0  (base)
       <=20  |   16.56764   41.77769     0.40   0.692    -65.31513     98.4504
        >40  |  -45.70776   42.61102    -1.07   0.283    -129.2238     37.8083
             |
   match_cat |
   mismatch  |          0  (base)
      match  |   9.399726   39.16746     0.24   0.810    -67.36709    86.16654
             |
    year_cat |
  1985-1989  |          0  (base)
  1990-1994  |   152.6126   42.30561     3.61   0.000     69.69513    235.5301
  1995-1998  |   84.72187   44.88182     1.89   0.059    -3.24489     172.6886
             |
       _cons |    525.942   48.51066    10.84   0.000     430.8628    621.0211
------------------------------------------------------------------------------
```

Based on this adjusted analysis, prophylaxis use is associated with a decrease of 165 (confidence interval [89–240]) days spent having had an adverse event in remission within the first 5 years after transplantation. This is only slightly different from the conclusion above, and the potential problem of covariate-dependent censoring due to year of transplantation seems to be minor here, at least up to five years after transplantation.

5 Simulations

To further illustrate the properties of the pseudo-observation method in multistate models, we conduct a simulation study. We want to illustrate that the method does produce reasonable parameter estimates under the requirements discussed in section 2, that some bias can be expected when the completely independent censoring assumption is not met, but that reasonable parameter estimates can be expected even in the non-Markov case when assumptions are met. We also want to illustrate how the robust sandwich variance estimator fares in these scenarios.

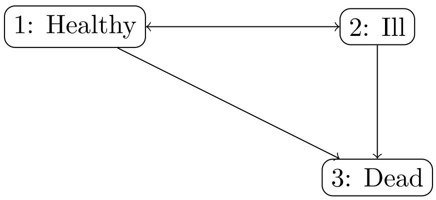

Figure 2. The multistate model of the simulation study

The setting considered in this simulation study is as follows. A multistate model with three states is considered. The three states can be thought of as healthy, ill, and dead, and the model is an illness-death model with recovery as illustrated in figure 2. We consider three scenarios of such a model. Common to all scenarios are the following features. We have two covariates, $\mathbf{Z} = (Z_1, Z_2)^T$, where Z_1 is a group variable, simulated as $b(1, 0.5)$, that we aim at estimating the effect of and Z_2, independently simulated as log-normal(0, 0.3), is a factor that we want to adjust for. Initially, all individuals are healthy. Conditional on covariates, we have the transition rates $\lambda_{12}(s) = (0.2Z_1 + 0.2Z_2)^{-1}$, $\lambda_{13}(s) = \lambda_{23}(s) = \log(2)/5$, whereas $\lambda_{21}(s)$ depends on the scenario and $\lambda_{31}(s) = \lambda_{32}(s) = 0$. Trajectories are right-censored, but the censoring rates depend on the scenario. A sample size of 400 independent individuals is considered in all scenarios. The three specific scenarios are as follows.

Sc.1 Constant $\lambda_{21}(s) = 1$ and completely independent censoring with censoring rate $\lambda_C(s) = 1/5$.

Sc.2 Constant $\lambda_{21}(s) = 1$ and covariate-dependent censoring with rate $\lambda_C(s) = (2.5Z_1 + 3.75Z_2)^{-1}$ independent of the multistate process.

Sc.3 A non-Markovian process where a latent log-normal(0, 0.3) waiting time determines when a transition back to healthy from ill occurs if death has not occurred in the mean time and completely independent censoring with censoring rate $\lambda_C(s) = 1/5$.

The marginal mean time to censoring is 5 or roughly 5 in either scenario. We consider two outcomes of interest—namely, state occupation for state 2, the ill state, at time 5 and length of stay in state 2 up to time 5. The scenarios are simple to simulate, but for either outcome the conditional expectation of the outcome does not depend in a simple way on covariates. We fit simple models that are wrong nonetheless to illustrate how the best fit to the pseudo-observations match the best fit in uncensored data. For the state occupation outcome, we consider the model $p_2(5 \mid Z_1, Z_2) = \text{expit}(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2)$, corresponding to a logit link in the generalized linear model framework. For the length of stay outcome, we consider the model $\text{los}_2(5 \mid Z_1, Z_2) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2$, corresponding to an identity link in the generalized linear model framework. In either case, the parameter of interest is considered to be β_1, which has the interpretation as either an adjusted log odds-ratio concerning odds of state occupation at time 5 or an adjusted difference in expected length of stay up to time 5 associated with the group variable Z_1, adjusting for Z_2. The parameter β_1 is estimated using the pseudo-observation method based on the Aalen–Johansen-derived estimate of state occupation probabilities and expected lengths of stay as described above. The robust sandwich variance estimate is used as a variance estimate. For comparison, a similar parameter estimate is obtained based on the uncensored sample where state occupation and length of stay are directly observed. For each scenario and for each outcome, we make 10,000 iterations of this procedure.

The results of this simulation study are presented in table 1, which shows averages of β_1 estimates in uncensored samples to get an idea of what kind of best fit we are trying to estimate in that particular scenario. Of perhaps primary interest are the presented averages of parameter estimates in censored samples obtained using the pseudo-observation method. The variance observed in parameter estimates across iterations is also presented, as well as the average of variance estimates across iterations.

Table 1. Results of simulations in six scenarios. "Est. uncens." refers to average of parameter estimates in the uncensored samples, "Est. cens." refers to parameter estimates in the censored samples, "Obs. var." refers to the variance in parameter estimates, and "Var. est." refers to averages of variance estimates.

	Est. uncens.	Est. cens.	Obs. var.	Var. est.
State occupation, sc. 1	−0.2533	−0.2597	0.0989	0.0965
State occupation, sc. 2	−0.2508	−0.2141	0.0949	0.1014
State occupation, sc. 3	−0.2421	−0.2403	0.0952	0.0958
Length of stay, sc. 1	−0.4879	−0.4865	0.0275	0.0276
Length of stay, sc. 2	−0.4856	−0.4600	0.0278	0.0276
Length of stay, sc. 3	−0.4499	−0.4498	0.0262	0.0266

In table 1, we see that, for scenarios 1 and 3 for either outcome, very similar averages are obtained in censored and uncensored samples, indicating no or very limited bias of the method in these scenarios. In contrast, a considerable bias in this sense can be seen for either outcome in scenario 2, where completely independent censoring is not fulfilled. Across scenarios, the observed variance and the average variance estimate seem to be in a reasonable correspondence. Figure 3 illustrates the approximate normality of the β_1 parameter estimates of the model for state occupation under scenario 1.

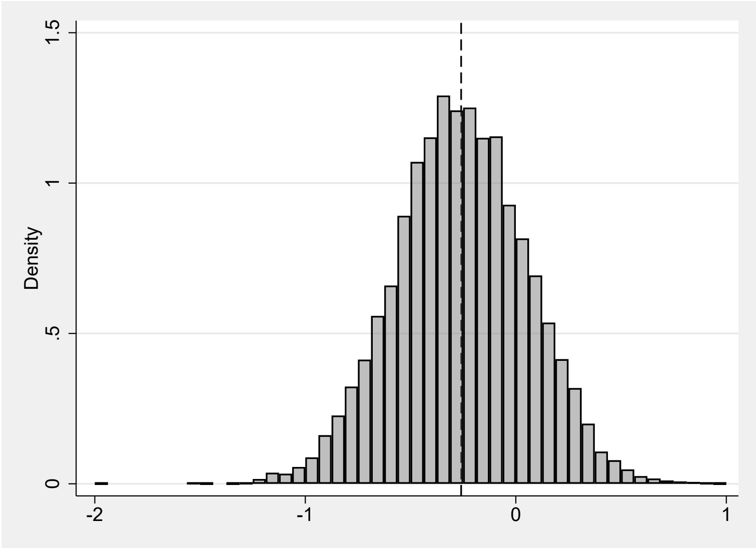

Figure 3. Estimates of β_1 in simulations of scenario 1 for state occupation. The average of estimates is at the dashed line.

In conclusion, this simulation study illustrates how reasonable parameter estimates can be obtained using this method when assumptions are met but that some bias can occur when the completely independent censoring assumption is violated. The simulation study also illustrates how the robust sandwich variance estimator provides reasonable estimates of the variance of parameter estimates in these scenarios.

6 Conclusions

We have presented the `stpmstate` command and demonstrated how it can be used for obtaining jackknife pseudo-observations in multistate settings and how these pseudo-observations can be of use in regression analyses of interest. The approach will be appropriate even without a Markov property of the underlying multistate process.

However, there are limitations on the usability of this pseudo-observation approach. We have mentioned how covariate-independent censoring is a requirement that can be quite strict in practice, where attrition may be associated with certain characteristics of study participants that also influence state occupation. This limitation of the pseudo-observation method stems from the fact that the Aalen–Johansen estimator, which the pseudo-observations are based on, does not account for covariates, and the limitation can likely be removed by basing pseudo-observations instead on more involved estimators that account for covariates, such as the estimator suggested by Datta and Satten (2002). Another option is to generalize the approaches of Binder, Gerds, and Andersen (2014) and Overgaard, Parner, and Pedersen (2019), where the censoring distribution is also modeled. As mentioned in section 4, another suggestion can be found in Andersen and Pohar Perme (2010), where pseudo-observations are based on a mixture estimator, which combines estimates from strata of a variable that censoring is considered to depend on. The pseudo-observations obtained from the mixture estimator are in fact identical to what is obtained by calculating pseudo-observations based on the original estimator in each stratum. In section 4, we saw how this approach can be taken in the multistate setting considered in this article simply by using the `by()` option of `stpmstate`. The completely independent censoring assumption is equivalent to an assumption of conditional independence of censoring time and an underlying multistate process given covariates and independence of censoring time and covariates. The conditional independence is usually impossible to check with the available data, but the independence of censoring time and covariates can be checked, at least if the conditional independence of censoring time and underlying multistate process holds. Another potential limitation is that of delayed entry. Preliminary investigations into the theory indicate that the selection that happens if some potential study participants do not actually enter the study because they reach an absorbing state before coming under observation can cause bias. But this bias may be minor in many cases and was not detected when studied by Grand et al. (2019).

The examples and the method presented here focused on one outcome of interest at a given time point. The pseudo-observation approach can be taken when considering more than one time point and also more than one outcome at the same time. The `stpmstate` command allows for simultaneous calculation of the required pseudo-observations. To fit the desired model, one may need to reshape the data into a long format and then obtain parameter estimates by using `glm` while accounting for the fact that more observations on the same individuals are used when calculating standard errors by specifying `vce(cluster id)` for the relevant `id` variable. If separate models and model parameters are considered for different outcomes, the seemingly unrelated estimation approach, as carried out by the `suest` command, may be useful for simultaneous inference about the various parameters.

An alternative to using jackknife pseudo-observations and the pseudo-observation method as demonstrated here would be to use weighting according to the inverse probability of complete observation, that is, to use the approach of inverse probability of censoring weighting. Weights can be applied to how each individual enters the procedure of obtaining regression parameter estimates—for example, in the `glm` step—or weights can be applied to the outcome variable specifically, as suggested by Scheike and Zhang (2007). This latter approach is known in some settings as direct binomial regression; see also Scheike, Zhang, and Gerds (2008). How these approaches compare with the pseudo-observation approach does not seem to be known.

7 Acknowledgment

Morten Overgaard is supported by the Novo Nordisk Foundation, grant NNF17OC0028276.

8 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 23-2
. net install st0717      (to install program files, if available)
. net get st0717          (to install ancillary files, if available)
```

9 References

Aalen, O. O., and S. Johansen. 1978. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* 5: 141–150.

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding. 1993. *Statistical Models Based on Counting Processes*. New York: Springer. https://doi.org/10.1007/978-1-4612-4348-9.

Andersen, P. K., and J. P. Klein. 2007. Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies. *Scandinavian Journal of Statistics* 34: 3–16. https://doi.org/10.1111/j.1467-9469.2006.00526.x.

Andersen, P. K., J. P. Klein, and S. Rosthøj. 2003. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* 90: 15–27. https://doi.org/10.1093/biomet/90.1.15.

Andersen, P. K., and M. Pohar Perme. 2010. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research* 19: 71–99. https://doi.org/10.1177/0962280209105020.

Binder, N., T. A. Gerds, and P. K. Andersen. 2014. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Analysis* 20: 303–315. https://doi.org/10.1007/s10985-013-9247-7.

Crowther, M. J., and P. Lambert. 2016. multistate: Stata module to perform multi-state survival analysis. Statistical Software Components S458207, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458207.html.

Datta, S., and G. A. Satten. 2001. Validity of the Aalen–Johansen estimators of stage occupation probabilities and Nelson–Aalen estimators of integrated transition hazards for non-Markov models. *Statistics and Probability Letters* 55: 403–411. https://doi.org/10.1016/S0167-7152(01)00155-9.

———. 2002. Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. *Biometrics* 58: 792–802. https://doi.org/10.1111/j.0006-341x.2002.00792.x.

Gill, R. D., and S. Johansen. 1990. A survey of product-integration with a view toward application in survival analysis. *Annals of Statistics* 18: 1501–1555. https://doi.org/10.1214/aos/1176347865.

Grand, M. K., and H. Putter. 2016. Regression models for expected length of stay. *Statistics in Medicine* 35: 1178–1192. https://doi.org/10.1002/sim.6771.

Grand, M. K., H. Putter, A. Allignol, and P. K. Andersen. 2019. A note on pseudo-observations and left-truncation. *Biometrical Journal* 61: 290–298. https://doi.org/10.1002/bimj.201700274.

Graw, F., T. A. Gerds, and M. Schumacher. 2009. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis* 15: 241–255. https://doi.org/10.1007/s10985-008-9107-z.

Jacobsen, M., and T. Martinussen. 2016. A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics* 43: 845–862. https://doi.org/10.1111/sjos.12212.

Liang, K.-Y., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13–22. https://doi.org/10.1093/biomet/73.1.13.

Maltzahn, N., R. Hoff, O. O. Aalen, I. S. Mehlum, H. Putter, and J. M. Gran. 2020. A hybrid landmark Aalen–Johansen estimator for transition probabilities in partially non-Markov multi-state models. ArXiv Working Paper No. arXiv:2007.00974. https://arxiv.org/abs/2007.00974.

Niessl, A., A. Allignol, C. Müller, and J. Beyersmann. 2020. Estimating state occupation and transition probabilities in non-Markov multi-state models subject to both random left-truncation and right-censoring. ArXiv Working Paper No. arXiv:2004.06514. https://arxiv.org/abs/2004.06514.

Overgaard, M. 2019. State occupation probabilities in non-Markov models. *Mathematical Methods of Statistics* 28: 279–290. https://doi.org/10.3103/S1066530719040033.

Overgaard, M., P. K. Andersen, and E. T. Parner. 2015. Regression analysis of censored data using pseudo-observations: An update. *Stata Journal* 15: 809–821. https://doi.org/10.1177/1536867X1501500313.

Overgaard, M., E. T. Parner, and J. Pedersen. 2017. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *Annals of Statistics* 45: 1988–2015. https://doi.org/10.1214/16-AOS1516.

———. 2018. Estimating the variance in a pseudo-observation scheme with competing risks. *Scandinavian Journal of Statistics* 45: 923–940. https://doi.org/10.1111/sjos.12328.

———. 2019. Pseudo-observations under covariate-dependent censoring. *Journal of Statistical Planning and Inference* 202: 112–122. https://doi.org/10.1016/j.jspi.2019.02.003.

Parner, E. T., and P. K. Andersen. 2010. Regression analysis of censored data using pseudo-observations. *Stata Journal* 10: 408–422. https://doi.org/10.1177/1536867X1001000308.

Scheike, T. H., and M.-J. Zhang. 2007. Direct modelling of regression effects for transition probabilities in multistate models. *Scandinavian Journal of Statistics* 34: 17–32. https://doi.org/10.1111/j.1467-9469.2006.00544.x.

Scheike, T. H., M.-J. Zhang, and T. A. Gerds. 2008. Predicting cumulative incidence probability by direct binomial regression. *Biometrika* 95: 205–220. https://doi.org/10.1093/biomet/asm096.

Spitoni, C., V. Lammens, and H. Putter. 2018. Prediction errors for state occupation and transition probabilities in multi-state models. *Biometrical Journal* 60: 34–48. https://doi.org/10.1002/bimj.201600191.

van Houwelingen, H. C., and H. Putter. 2008. Dynamic predicting by landmarking as an alternative for multi-state modeling: An application to acute lymphoid leukemia data. *Lifetime Data Analysis* 14: 447–463. https://doi.org/10.1007/s10985-008-9099-8.

About the authors

Morten Overgaard is an assistant professor at the Research Unit for Biostatistics, Department of Public Health, Aarhus University. His current primary research focus is on various aspects of the pseudo-observation method.

Per Kragh Andersen is a professor at the Section of Biostatistics, Department of Public Health, University of Copenhagen. His research interests are in analysis of survival data and follow-up studies and other applications of statistics in epidemiology.

Erik Thorlund Parner is a professor at the Research Unit for Biostatistics, Department of Public Health, Aarhus University. His research fields are time-to-event analysis and statistical methods in epidemiology.

Appendix

In this appendix, we want to clarify a few theoretical details of the pseudo-observation method, particularly those regarding the assumptions to ensure that the pseudo-observations have the correct conditional expectation.

We let X denote the observed multistate process where we let $X(s) = 0$ indicate that the underlying process is not under observation at time s, which could be due to censoring before time s. The vector of initial distributions $p(0)$ is estimated by the empirical version of $p^{\mathrm{c}}(0)$, where $p_j^{\mathrm{c}}(0) = P\{X(0) = j \mid X(0) \neq 0\}$. The cumulative forces of transition matrix Λ is given by $\Lambda_{jk}(s) = \int_0^s p_j(u-)^{-1} F_{jk}(du)$ for off-diagonal elements and $\Lambda_{jj}(s) = -\sum_{k \neq j} \Lambda_{jk}(s)$ on the diagonal, where $F_{jk}(s)$ is the expected number of transitions from j to k up to time s for an individual. Each $\Lambda_{jk}(s)$ is estimated by Nelson–Aalen estimates, where the total number of individuals observed in state j just before time u replaces $p_j(u-)$ and the total number of observed transitions from j

to k up to time s replaces $F_{jk}(s)$. Without further assumptions, we denote the limit of the matrix of estimates $\widehat{\Lambda}$ by Λ^{c}, an observable cumulative forces of transition matrix. It should be clear that under an assumption of independent censoring, we have both of the consistency properties $p^{\mathrm{c}}(0) = p(0)$ and $\Lambda^{\mathrm{c}} = \Lambda$.

Now, because the Aalen–Johansen-derived estimates of the state occupation probabilities are $\widehat{p}(t) = \widehat{p}(0) \prod_0^t \{\mathbf{I} + \widehat{\Lambda}(du)\}$, the influence function of the estimate $\widehat{p}(t)$ will depend on the influence functions of both $\widehat{p}(0)$ and $\widehat{\Lambda}$. More precisely, the influence function of $\widehat{p}(t)$ can be stated as

$$\begin{aligned}\dot{p}(t;X) = \dot{p}(0;X) \prod_0^t \{\mathbf{I} + \Lambda^{\mathrm{c}}(du)\} \\ + p^{\mathrm{c}}(0) \int_0^t \prod_0^{s-} \{\mathbf{I} + \Lambda^{\mathrm{c}}(du)\} \, \dot{\Lambda}(ds;X) \prod_s^t \{\mathbf{I} + \Lambda^{\mathrm{c}}(du)\}\end{aligned}$$

and involves two terms: one term is related to the influence function $\dot{p}(0;X)$ of $\widehat{p}(0)$, and a second term involves the influence function $\dot{\Lambda}(\cdot;X)$ of $\widehat{\Lambda}$. The form of the latter term comes from the Duhamel equation; see, for example, Gill and Johansen (1990). To be precise, we have $\dot{p}(0;X) = 1(X(0) = j)/\sum_{k\neq 0} \widetilde{p}_k(0) - p_j^{\mathrm{c}}(0) \sum_{k\neq 0} 1(X(0) = k)/\sum_{k\neq 0} \widetilde{p}_k(0)$, where $\widetilde{p}_j(0) = P(X(0) = j)$ is the initial probability of being observed to be in state j and, for $j \neq k$,

$$\dot{\Lambda}_{jk}(s;X) = \int_0^s \frac{1}{\widetilde{p}_j(u-)} N_{X,jk}(du) - \int_0^s \frac{1(X(u-) = j)}{\widetilde{p}_j(u-)} \Lambda_{jk}^{\mathrm{c}}(du)$$

where $N_{X,jk}(s)$ counts the number of transitions from j to k of the observable multistate process X before time s.

Under an assumption of completely independent censoring—that is, the censoring time is independent of multistate process and covariates—we have, as mentioned, that $p^{\mathrm{c}}(0) = p(0)$ and $\Lambda^{\mathrm{c}} = \Lambda$ but also that $p^{\mathrm{c}}(0 \mid \mathbf{Z}) = p(0 \mid \mathbf{Z})$, $\Lambda^{\mathrm{c}}(\cdot \mid \mathbf{Z}) = \Lambda(\cdot \mid \mathbf{Z})$, and $\widetilde{p}_j(s \mid \mathbf{Z})/\widetilde{p}_j(s) = p_j(s \mid \mathbf{Z})/p_j(s)$ for the conditional distributions given covariates $\mathbf{Z}$. In addition to $E\{\dot{p}(0;X) \mid \mathbf{Z}\} = p(0 \mid \mathbf{Z}) - p(0)$, this implies that $E\{\dot{\Lambda}_{jk}(s;X) \mid \mathbf{Z}\} = \int_0^s p_j(u- \mid \mathbf{Z})/p_j(u-)\{\Lambda_{jk}(du \mid \mathbf{Z}) - \Lambda_{jk}(du)\}$ or, in matrix form, that

$$E\{\dot{\Lambda}(s;X) \mid \mathbf{Z}\} = \int_0^s \mathrm{diag}\{p(s)\}^{-1} \, \mathrm{diag}\,\{p(s \mid \mathbf{Z})\} \{\Lambda(du \mid \mathbf{Z}) - \Lambda(du)\}$$

where $\mathrm{diag}(a)$ is the diagonal matrix with the vector a on the diagonal. Because

$$\begin{aligned}& p(0) \prod_0^s \{\mathbf{I} + \Lambda(du)\} \ \mathrm{diag}\,\{p(s)\}^{-1} \ \mathrm{diag}\,\{p(s \mid \mathbf{Z})\} \\ &= p(s)\mathrm{diag}\,\{p(s)\}^{-1} \ \mathrm{diag}\,\{p(s \mid \mathbf{Z})\} = p(s \mid \mathbf{Z}) \\ &= p(0 \mid \mathbf{Z}) \prod_0^s \{\mathbf{I} + \Lambda(du \mid \mathbf{Z})\}\end{aligned}$$

we see that

$$
\begin{aligned}
E\{\dot{p}(t;X)\mid \mathbf{Z}\} &= \{p(0\mid \mathbf{Z}) - p(0)\} \prod_0^t \{\mathbf{I} + \Lambda(du)\} \\
&\quad + p(0) \int_0^t \prod_0^{s-} \{\mathbf{I} + \Lambda(du)\} \operatorname{diag}\{p(s)\}^{-1} \\
&\quad \operatorname{diag}\{p(s\mid \mathbf{Z})\} \{\Lambda(du\mid \mathbf{Z}) - \Lambda(du)\} \prod_s^t \{\mathbf{I} + \Lambda(du)\} \\
&= \{p(0\mid \mathbf{Z}) - p(0)\} \prod_0^t \{\mathbf{I} + \Lambda(du)\} \\
&\quad + p(0\mid \mathbf{Z}) \int_0^t \prod_0^{s-} \{\mathbf{I} + \Lambda(du\mid \mathbf{Z})\} \\
&\quad \{\Lambda(du\mid \mathbf{Z}) - \Lambda(du)\} \prod_s^t \{\mathbf{I} + \Lambda(du)\} \\
&= \{p(0\mid \mathbf{Z}) - p(0)\} \prod_0^t \{\mathbf{I} + \Lambda(du)\} + p(0\mid \mathbf{Z}) \\
&\quad \left[\prod_0^t \{\mathbf{I} + \Lambda(du\mid \mathbf{Z})\} - \prod_0^t \{\mathbf{I} + \Lambda(du)\}\right] \\
&= p(0\mid \mathbf{Z}) \prod_0^t \{\mathbf{I} + \Lambda(du\mid \mathbf{Z})\} - p(0) \prod_0^t \{\mathbf{I} + \Lambda(du)\} = p(t\mid \mathbf{Z}) - p(t)
\end{aligned}
$$

where the Duhamel equation is also used. This shows how the main requirement for the pseudo-observation method to work is fulfilled under the completely independent censoring assumption. A similar argument is given in the supplement of Spitoni, Lammens, and Putter (2018).