Estimation of Life Expectancies Using Continuous-Time Multi-State Models

Ardo van den Hout
Department of Statistical Science, University College London
Gower Street, London WC1E 6BT, UK
E-mail: ardo.vandenhout@ucl.ac.uk

Mei Sum Chan
University College London and University of Oxford

Fiona Matthews
Newcastle University

Abstract:
Background and Objective: There is increasing interest in multi-state modelling of health-related stochastic processes. Given a fitted multi-state model with one death state, it is possible to estimate state-specific and marginal life expectancies. This paper introduces methods and new software for computing these expectancies.
Methods: The definition of state-specific life expectancy given current age is an extension of mean survival in standard survival analysis. The computation involves the estimated parameters of a fitted multi-state model, and numerical integration. The new R package `elect` provides user-friendly functions to do the computation in the R software.
Results: The estimation of life expectancies is explained and illustrated using the `elect` package. Functions are presented to explore the data, to estimate the life expectancies, and to present results.
Conclusions: State-specific life expectancies provide a communicable representation of health-related processes. The availability and explanation of the `elect` package will help researchers to compute life expectancies and to present their findings in an assessable way.

Keywords:
Gompertz distribution, interval censoring, Markov model, panel data, sojourn time, stochastic process

1 Introduction

This paper presents methods and the R package elect for computing state-specific and marginal life expectancies using the estimated parameters of a continuous-time multi-state model with one death state. Explaining and illustrating elect is the main aim of this paper.

Multi-state models can be used to describe health-related stochastic processes over time. States can represent health stages or death, and to formulate the model, potential transitions between these states are distinguished. It is possible to have a model with multiple dead states representing different causes of death. In this package, we focus on models with one dead state only. Given such a model, it is often of interest to know how much of remaining total life expectancy at a given age subdivides into life expectancies in the living states. As an example, consider the three-state illness-death model for an older population defined by a healthy state, an ill-health state, and the dead state (Figure 1). For an individual at a specified age, we can distinguish between two types of residual life expectancies, namely expected remaining time spent in the healthy state and expected remaining time spent in the ill-health state. The sum of these expectancies constitutes the total residual life expectancy.

Important methodological work on continuous-time multi-state models for longitudinal data is presented in Kalbfleisch and Lawless (1985), Hougaard (2000), and Aalen et al. (2008). Continuous-time models are based on theory for continuous-time Markov chains as discussed in, for example, Norris (1997). Jackson (2011) and de Wreede et al. (2010) present R packages that provide a flexible framework for fitting continuous-time multi-state models to longitudinal data.

The elect package is available from the Comprehensive R Archive Network (CRAN, R Core Team, 2018) at https://CRAN.R-project.org/package=elect. The package is developed as an add-on to msm, the R package created by Jackson (2011). If msm is used to fit a Gompertz model with age as the time scale, then elect can be used to estimate state-specific life expectancies. The name "elect" is inspired by the functionality of the package: estimating life expectancies using continuous time.

Estimating life expectancies is about estimating mean sojourn time in states. The msm package has a function for estimating sojourn times: sojourn.msm. However, this function is only defined for exponential models. For estimating life expectancies, the exponential model is not suitable—models are needed that allow transition hazards to change with age. The Gompertz model is such a model and it is often used to describe morbidity and mortality; see, for example, Mueller et al. (1995), Hougaard (2000), and Blossfeld and Rohwer (2002).

The Gompertz model can be fitted in msm by defining age as a time-dependent covariate. In this case, the likelihood function involves a piecewise-constant approximation of the age dependence in the model. Nevertheless, there are no functions in msm that directly provide inference for life expectancies given a fitted Gompertz model. In addition, for marginal life expectancies, additional modelling of the state distribution is needed. Because of this, we developed elect: a user-friendly framework for estimating state-specific life expectancies using a Gompertz model that is fitted in msm. Early versions of the functions in elect show a wide range of applications; see, for example, Van den Hout et al. (2014), Robitaille et al. (2018), Van der Noordt et al. (2018), Hoogendijk et al. (2019)

Willekens and Putter (2014) discuss software for multi-state analysis in detail. Their discussion include an early version of `elect` and two other programs for estimating life expectancies: `IMaCh` (Lièvre et al., 2003) and `SPACE` (Cai et al., 2010). These three programs are also briefly reviewed in Saito et al. (2014). Both `IMaCh` and `SPACE` are based on discrete-time Markov models, whereas `elect` is based on a continuous-time model. `IMaCh` is a stand-alone program, `SPACE` is programmed in SAS.

`SPACE` works under the assumption that there is at most one transition between two successive observations. This assumption may be too restrictive for certain longitudinal data. When using `elect`, this assumption is not used, since the model underlying `elect` is a time-continuous model.

The `multistate` package in STATA by Crowther and Lambert (2017), makes it possible to estimate life expectancies based on a continuous-time multi-state model. `multistate` includes semi-Markov models, which are not available in `msm`. The software allows for a wide range of parametric distributions and has recently been extended to allow reversible transitions. `multistate` requires exactly observed transition times.

The `flexsurv` package by Jackson (2016) allows for a wide range of parametric multi-state models and can estimate total length of stay in particular states. However `flexsurv` also requires exactly observed transition times.

For estimating life expectancies, `elect` is the only software that is available in R, for Gompertz models that are estimated using interval-censored data. It allows users to compute life expectancies for any number of states and is not limited to progressive processes. Observation times can be exact or interval-censored, or a mixture of these. Because of the extensive functionality in `msm`, `elect` can work with hidden Markov models; see, for example, Robitaille et al. (2018) where misclassification of state is taken into account.

In what follows, we summarise the standard method of estimating life expectancies, introduce the computation in `elect`, and give two examples of using `elect`. Some familiarity with using `msm` is recommended; for details of fitting the multi-state model in `msm`, see Jackson (2011).

2 Methods

2.1 Life expectancies

State-specific life expectancy based on a multi-state model is a generalisation of mean survival in a standard survival model where there is one living state and one dead state. For an extended discussion of life expectancy and estimation for multi-state models see, for example, Izmirlian et al. (2000), Lièvre et al. (2003), and Van den Hout (2017).

Let the finite state space be given by $\{1, 2, ..., D\}$ where D is the dead state. Let Y_t denote the state at age t and let $\boldsymbol{x}$ denote the time-independent vector with covariate

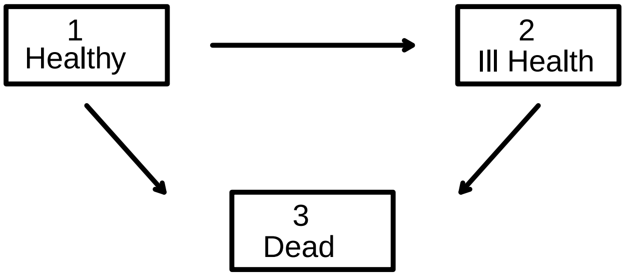

Figure 1: Three-state model for health and ill health in the older population.

values. Life expectancy in living state s given state r at age t, for $r, s \in \{1, 2, ..., D-1\}$, is defined by

$$e_{rs}(t|\boldsymbol{x}) = \int_0^\infty \mathbb{P}(Y_{t+u} = s|\ Y_t = r, \boldsymbol{x})du, \tag{1}$$

where $\mathbb{P}(Y_{t+u} = s|\ Y_t = r, \boldsymbol{x})$ is the transition probability of being in state s at age $t + u$, given starting state r at age t and covariate values $\boldsymbol{x}$. Marginal life expectancy in state s is irrespective of the initial state at age t and is defined by

$$e_{\bullet s}(t|\boldsymbol{x}) = \sum_{r \neq D} \mathbb{P}(Y_t = r|\boldsymbol{x}) e_{rs}(t|\boldsymbol{x}), \tag{2}$$

where $\mathbb{P}(Y_t = r|\boldsymbol{x})$ is the probability of being in state r at age t for $r \in \{1, 2, ..., D-1\}$. Total life expectancy at age t is defined as

$$e(t|\boldsymbol{x}) = \sum_{s \neq D} e_{\bullet s}(t|\boldsymbol{x})\,. \tag{3}$$

2.2 Models and estimation

To be able to estimate life expectancy, transition probabilities and the state distribution are estimated using longitudinal data. Using the same notation as in the previous section, we assume that data for individual i and observation j are given by (y_{ij}, t_{ij}, x_i), for $i \in \{1, ..., N\}$ and $j \in \{1, ..., n_i\}$. For ease of exposition, we use only one covariate in the example. What follows also applies to models with more than one time-independent covariate.

Transition probabilities are derived from a multi-state model where the hazards are defined by

$$h_{rs}(t_{ij}) = \exp\left(\beta_{rs} + \xi_{rs} t_{ij} + \gamma_{rs} x_i\right), \tag{4}$$

for those pairs (r, s) that define a transition in the stochastic process. Model parameters are estimated by maximum likelihood. The definition of the likelihood function can take into account whether observation times t_{ij} are exact or interval censored.

The dependency on age t in (4) defines the hazard of a Gompertz distribution, which implies that transition hazards are either increasing or decreasing exponentially. In order to use `elect`, this model has to be fitted using `msm` in R. When `msm` is used to fit a Gompertz hazard model, it will use a piecewise-constant approximation for the parametric time-dependency. Specifically, for a time interval $(t_{ij}, t_{ij+1}]$ in the data, the contribution to the log-likelihood function is defined by fixing the hazard to $h_{rs}(t_{ij})$ throughout $(t_{ij}, t_{ij+1}]$. Although piecewise-constant approximation is thus used to estimate model parameters, the fitted model is not piecewise constant. The parameter estimates are used to define a smooth Gompertz model.

When investigating an ageing process, we recommend to model the dependency of age on a shifted scale. For example, if the minimum age in the data is 65, define $t = age - 65$. In model (4), the multiplicative effect of age is $\exp\left(\xi_{rs}t\right)$ and large values of t can cause numerical problems when fitting the model. Shifting the age scale may prevent these problems.

In line with the functionality of `elect`, model (4) is restricted to time-independent covariates. It is possible to fit a model with time-dependent $x(t)$ if we use a piecewise-constant approximation. However, in that case, we would not have a model for the change of $x(t)$ over time and hence we would not be able to estimate residual life expectancies.

The distribution of the state at age t is modelled using a multinomial regression model defined by

$$\mathbb{P}(Y_t = r|x) = \frac{\exp\left(\eta_r(t)\right)}{1 + \sum_{r \neq D} \exp\left(\eta_r(t)\right)} \quad \text{with} \quad \eta_r(t) = \alpha_{r0} + \alpha_{r1}t + \alpha_{r2}x\,, \tag{5}$$

for $r \in \{1, 2, ..., D-1\}$. By restricting $\alpha_{10} = \alpha_{11} = \alpha_{12} = 0$, we make $r = 1$ the reference category. This model is estimated in `elect` using the function `multinom` in the package `nnet` (Venables and Ripley, 2002).

Life expectancies (1), (2), and (3) can be derived using the parameters in the multi-state model and the multinomial regression model. The specification of $\boldsymbol{x}$ is undoubtedly important, but the specification of age t will in most cases be the most influential. To approximate the integral (1), a maximum age $t_{\max}$ has to specified such that we may safely assume that the integrand $\mathbb{P}(Y_{t+u} = s|\ Y_t = r, \boldsymbol{x})$ is zero when $t + u > t_{\max}$.

Note that with a fitted multi-state model and specified $\boldsymbol{x}$, the integrand in (1) can be computed for any u. In the computation of this integrand, a piecewise-constant approximation is used to account for changing age over time. Computationally, it is convenient to use the same grid for the piecewise-constant approximation *and* the numerical approximation of the integral. This is how it is implemented in `elect`.

The above provides a point estimate of life expectancies. To estimate the uncertainty (standard errors and/or confidence intervals) we make use of the asymptotic properties of the maximum likelihood estimator of the parameters for the multi-state model and the multinomial regression model. The two models are fitted independently from each other, and for each we define a multivariate normal distribution with expectation equal to the maximum likelihood estimate of the parameter vector and the covariance matrix equal to the estimated covariance matrix at the optimum. The sample variation in the estimation of the life expectancies is evaluated by drawing parameter values from the two multivariate distributions and computing the life expectancies for the drawn values. This simulation-based approach is a general method for deriving standard errors or confidence intervals for a complex function of a maximum likelihood estimate; see Mandel (2013).

3 Results

3.1 Example

Consider a progressive three-state illness-death process for an older population. State 1 is defined as the healthy state, state 2 is the ill-health state, and state 3 represents death; see Figure 1. For this process, simulated data are available in `elect` under the name `electData`. This data set contains simulated trajectories for 150 individuals. Transition times for moving from state 1 to state 2 are interval censored; entry times for the dead state are known. The data reflect the context of longitudinal data collection in ageing research: pre-scheduled interview times (panel data for the living states) and exact times for death during the study time.

The longitudinal data format in `electData` is such that there is one row per observation. The first 6 records in the data are:

```
R> library(elect)
R> head(electData)

  id state   age x bsline
1  1     1  7.26 1      1
2  1     1  9.36 1      0
3  1     2 11.28 1      0
4  1     2 13.38 1      0
5  1     3 14.24 1      0
6  2     1 19.61 0      1
```

The identifier for individuals is `id`. Variable `state` denote states 1, 2, and 3. Variable `age` is age in years on a shifted scale (age at the observation time minus 70 years). The binary time-independent covariate `x` is specified as 0 for women and 1 for men. Variable `bsline` is an indicator for the baseline observation and will be used to define the data for the multinomial regression model (5). The number of records per individual varies:

```
R> table(table(electData$id))

 2  3  4  5  6  7
16 24 22 13 18 57
```

An important data summary with respect to multi-state modelling is the state table which can be produced by `statetable.msm` in `msm`:

```
R> statetable.msm(state, id, data = electData)

    to
from   1   2   3
   1 358  56  53
   2   0  94  53
```

This frequency table shows the number of times each pair of states is observed at successive observation times. For example, there are 53 individuals who were observed in state 3 after being observed in state 1. Due to the interval censoring, we do not know whether these individuals moved directly from state 1 to state 3, or via state 2.

For the three-state model with the three transitions as illustrated in Figure 1, we specify model (4) as

$$h_{rs}(\texttt{age}) = \exp\big(\beta_{rs} + \xi_{rs}\texttt{age} + \gamma_{rs}\texttt{x}\big), \tag{6}$$

where $(r, s) \in \{(1, 2), (1, 3), (2, 3)\}$.

The model can be fitted using `msm`:

```
R> Q     <- rbind(c(NA, 0.01, 0.01), c(0, NA, 0.01), c(0, 0 ,NA))
R> model <- msm(state˜age, subject = id, data = electData,
                center = FALSE, qmatrix = Q, deathexact = 3,
                covariates = ˜age+x)
```

The specification of matrix `Q` defines the transitions that are possible according to the model (only the off-diagonal entries of `Q` are used). Matrix `Q` also provides the starting values for the transition intensities in the maximum likelihood estimation in `msm`; for details see Jackson (2011).

The Gompertz model is fitted using a piecewise-constant approximation; see Section 2.2. Once the model parameters are estimated, we use them to derive the smooth Gompertz hazards. A function in `elect` can be used to depict these fitted hazards of the age-dependent model. For a woman aged 70, we have `age = 0` and `x = 0` and use:

```
R> hazards(model, b.covariates = list(age = 0, x = 0),
           no.years = 20, max.haz = 0.2, age.shift = -70)
```

Figure 2 depicts the hazards on the original age scale (derived from the argument `age.shift = -70`). All three hazards increase with increasing age.

To enable using `elect`, there are prerequisite elements in the `msm` call: using names `state` and `age` in the data, and using the option `center = FALSE`. For covariates that are encoded as factors, dummy variables have to be used instead when fitting the model with `msm`.

To estimate life expectancies, the user has to provide the data for the state-distribution model (5). Because `elect` uses `age` by default as a covariate for the

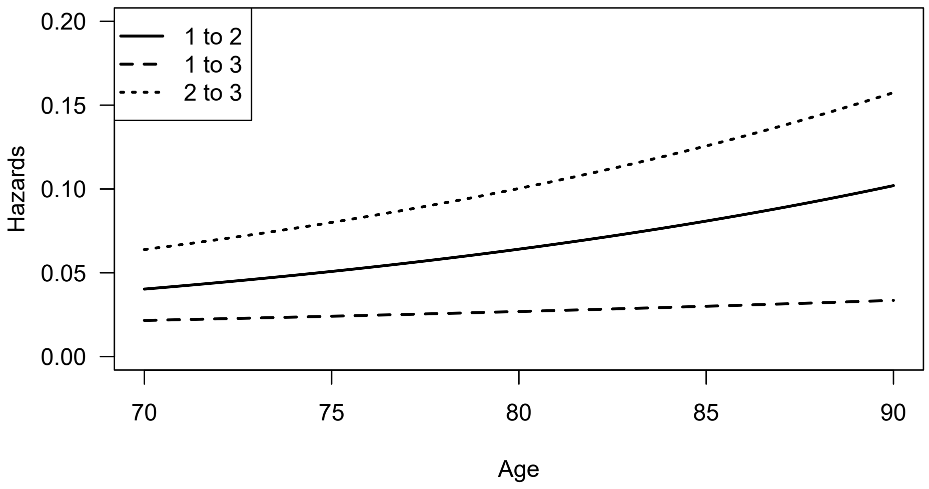

Figure 2: Fitted hazards of the age-dependent model (6) for women at age 70 and older.

state-distribution, this variable has to be included. In the current case, we assume that the baseline of `electData` is representative of the population state distribution. Given this assumption, it makes sense to use the baseline records as the data for the state-distribution model:

```
R> sddata <- electData[electData$bsline == 1,]
```

The range of the (shifted) age in this subset is from about 1 up to 23. We assume that the subset represents the population of interest for the marginal life expectancies. The definition of the data for the state-distribution model is application specific; see also the example in Section 3.2.6 and the discussion in Section 4.

Say we want the life expectancies for a man aged 70. The `elect` call and the summary commands are:

```
R> LEs    <- elect(x = model, b.covariates = list(age = 0, x = 1),
                statedistdata = sddata, h = 0.1,
                setseed = 1234, age.max = 50, S = 500)
R> summary(LEs)
```

In the above, we specify `age = 0` because of the shifted age scale, and `x = 1` for men. Value `h = 0.1` is the parameter for the grid in the integral that is used in the estimation of the life expectancies. In this case, the integral is approximated on a 0.1-year grid.

The specification of `age.max` should be such that the probability to survive beyond `age.max` is assumed to be negligible. This specification should take into account the transformation of age before the model was fitted. In the example, specifying `age.max = 50` corresponds with an assumed maximum age of $70 + 50 = 120$ years.

If no estimation of the uncertainty is required, the default `S = 0` should be used. In that case only point estimates of the life expectancies will be provided. In the call above, the estimated uncertainty is estimated using 500 replications.

The summary function provides the following output:

```
-----------------------------
elect summary
-----------------------------
Covariate values in the multi-state model:
age   x
  0   1
Covariates in the state-distribution model:
   age

Life expectancies:
Using simulation with  500 replications

Point estimates, and mean, SEs, and quantiles from simulation:
       pnt    mn    se 0.025q   0.5q 0.975q
e11  8.627 8.402 1.174  6.137  8.375 10.564
e12  2.399 2.376 0.505  1.494  2.349  3.496
e21  0.000 0.000 0.000  0.000  0.000  0.000
e22  5.360 5.387 1.379  2.823  5.351  8.191
e.1  7.496 7.151 1.201  4.766  7.125  9.432
e.2  2.787 2.828 0.616  1.795  2.799  4.146
e   10.283 9.979 1.233  7.532 10.002 12.306
-----------------------------
```

The output for the life expectancies consists of state-specific (1), marginal (2), and total life expectancies (3) derived from the maximum likelihood point estimate of the model parameters (`pnt`). In addition, means (`mn`), standard errors (`se`) and quantiles (`0.025q`, `0.5q`, and `0.975q`) of the simulated distribution are derived from the maximum likelihood estimation.

The output for the life expectancies shows that for a man aged 70, the total life expectancy is estimated at 10.283 years with a 95%-confidence interval of $(7.532, 12.306)$. If this man is healthy when aged 70, then he is expected to spend 8.627 years in good health. If we do not know whether he is healthy at age 70, he is expected to spend 7.496 years in good health. Because of the small sample size in this example, the uncertainty on these statistics is large.

Because the uncertainty is estimated by simulation, small differences are expected when rerunning `elect`. The argument `setseed` in `elect` can be used to produce the same results across reruns.

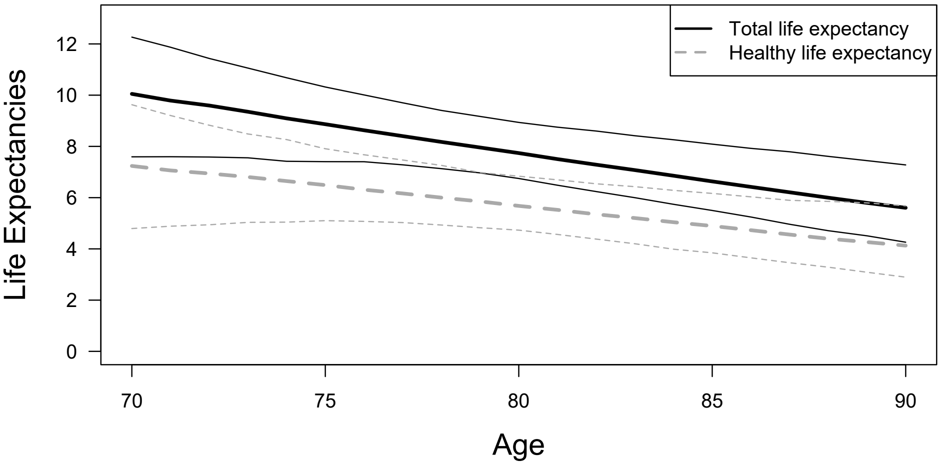

Figure 3: Life expectancies for men conditional on their age (with 95% confidence bands). Solid lines for total life expectancy, dashed lines for marginal life expectancy in the healthy state.

Figure 3 shows estimated life expectancies for 70 up to 90 years old. This graph is constructed by using `elect` repeatedly on single years of age between 70 and 90. Confidence bands are based on 500 simulations.

For the state-distribution model, the default in `elect` is to use a multinomial logistic regression model with age as the only covariate. This can be extended to include additional covariates. This option is relevant for the estimation of the marginal life expectancies only. Further to the example above, when adding `statedist.covariates = c("age","x")` to the `elect` call, we obtain point estimates `e.1` = 7.677, and `e.2` = 2.725, which are quite close the previous estimates. More information about the fitted logistic regression can be obtained by augmenting the above command: `summary(LEs, sd.model = TRUE)`.

3.2 Additional functionality in `elect`

3.2.1 Data exploration

Predicting life expectancies typically involves extrapolation of the model beyond the age range in the data. For this reason, it is important to know the distribution of age in the data. To explore the data before fitting the multi-state model, the function `explore` produces basic statistics with respect to sample size, number of observations, and age. The function also provides information on the length of the time interval between observations. The latter is important given the piecewise-constant approximation that is underlying the estimating of the age-dependent model.

For the example data set, the command is `explore(electData)`. In this data set, the median length of time intervals is 1.98 years. This has to be interpreted relative to the expected rate of change in the health process that is of interest. For example, if ill health is defined as cognitive impairment in older age, following individuals up about every 2 years seems suitable; see, for example, the English Longitudinal Study of Ageing (ELSA, Taylor et al., 2007).

3.2.2 Numerical settings

The user can specify the method for the numerical approximation of the integral in (1). In the `elect` call, specification `method = "step"` is the default. Additional options are `"MiddleRiemann"` and `"Simpson"`. Typically, this makes little difference if a small `h` is specified in the `elect` call; see the example in Section 3.2.6.

3.2.3 More complex multi-state models

The function `msm` allows for a wide range of multi-state models. For example, it is possible to fit hidden Markov models that allow for misclassification of state. In this case, `elect` can still be used and it will compute life expectancies for the latent process (as opposed to the manifest process which is assumed to be confounded by misclassification).

It is also possible to fit multi-state models in `msm` with restrictions on parameters. These can be equal-to-zero restrictions (e.g. $\gamma_{23} = 0$), constraints across transitions (e.g. $\xi_{13} = \xi_{23}$), or a combination thereof. Equal-to-zero restrictions are taken into account in `elect` automatically. Constraints, however, have to be formulated explicitly in the `elect` call by using the argument `RestrAndConst`. The function `check.RestrAndConst` can help the user to check the specification.

Typically, `msm` is used to fit a model for interval-censored data, however, the software can also be used for data with exact times of transitions for all states. The use of `elect` does not have to be adapted when exact-time data are used.

3.2.4 Functions of life expectancies

It is possible to compute functions of life expectancies. Further to the example in Section 3.1, one might want to know about the difference of life expectancies given either state 1 or state 2 at a given age. In the notation in Section 2.1 this is $\big(e_{11}(t|\boldsymbol{x}) + e_{12}(t|\boldsymbol{x})\big) - \big(e_{21}(t|\boldsymbol{x}) + e_{22}(t|\boldsymbol{x})\big)$ given age t. To obtain the uncertainty, the function `plusmin` can be used:

```
R>  plusmin(LEs, index = c(1, 2, 3, 4),
               func  = c("plus", "minus", "minus"), digits = 2)

            pnt   mn   se 0.025q 0.5q 0.975q
func(LEs) 5.67 5.39 1.66   1.96 5.51   8.34
```

3.2.5 Plotting the estimated life expectancies

The computation of the uncertainty of the estimated life expectancies is based on the asymptotic properties of the maximum likelihood estimator. However, assuming that the uncertainty of the parameters can be described by a normal distribution does not imply that the uncertainty of the life expectancies can be described by a normal distribution. The life expectancies are a non-linear function of the model parameters, and their estimated distribution can be skewed. For a plot of this distribution in the example in Section 3.1: `plot(LEs)`.

3.2.6 Additional example

To illustrate some of the additional functionality in `elect`, we use data from heart transplant patients on cardiac allograft vasculopathy (CAV); see Sharples et al. (2003). This dataset contains information for 622 patients and is available in the `msm` package as `cav`. Four states are defined: state 1 for no CAV, state 2 for mild/moderate CAV, state 3 for severe CAV, and state 4 for death. Each patient is assigned to state 1 at the baseline examination. Subsequent interval-censored data are available from examinations that are approximately a year apart. Death times during the follow-up are known exactly. Sharples et al. (2003) assumed that CAV is a progressive process, and fitted a four-state model that allows misclassification of state. The misclassification makes it possible to explain backward transitions in the data by attributing these transitions to measurement error in CAV diagnosis.

With the CAV data, we illustrate using `elect` with a model with four states, misclassification of state, and parameter restrictions. The following model is for illustrative purpose only—extended models can be investigated in a similar way.

We start by defining the data using variables names that link up with the `elect` requirements:

```
R>  library(elect)
R>  dta <- as.data.frame(cbind(id       = cav$PTNUM, age = cav$years,
                               state    = cav$state, x   = cav$dage,
                               firstobs = cav$firstobs))
R>  Q <- rbind(c(NA, 0.01, 0,  0.01), c(0, NA, 0.01, 0.01),
               c(0 , 0    , NA, 0.01),  c(0, 0, 0,    NA))
```

Note that `age` represents time since the heart transplant in years and that matrix `Q` defines the five allowable transitions for a progressive four-state model with state 4 as an absorbing state.

Next we define the potential misclassification by defining a 4 × 4 matrix:

```
R>  E <- rbind(c(0, 0.1, 0, 0), c(0.1, 0, 0.1, 0),
               c(0, 0.1, 0, 0), c(0,   0, 0,   0))
```

The off-diagonal entries in matrix `E` that are not zero define the misclassification model. A latent state 1 can be observed as state 2, a latent state 2 can be observed as state 1 or 3, and a latent state 3 can be observed as state 2. Value `0.1` is used here as starting value for the maximum likelihood estimation—other choices within

(0,1) are possible. To fit the multi-state model with misclassification, we call the `msm` function:

```
R> model <- msm(state ~ age, subject = id, data = dta,
                covariates = ~  age + x, center = FALSE, ematrix = E,
                qmatrix = Q, fixedpars = c(7:10), obstrue = firstobs,
                deathexact = 4, constraint = list(x = c(1,2,1,2,2)))
```

The argument `obstrue = firstobs` implies that we assume that the first state 1 of each patient is not misclassified. With `age` and `x` as covariates, there are three sets of five parameters in this model. The first set includes the intercepts for the five transitions; the order is 1 → 2, 1 → 4, 2 → 3, 2 → 4, and 3 → 4. The next set includes the effects of `age` in the same order. The argument `fixedpars` is used to fix covariate effects to their initial values, which are zero by default. The specification `fixedpars = c(7:10)` uses the ordering above to restrict the effect of `age` to be zero for all the transitions but the first one. The last five parameters are for the effect of `x`. The argument `constraint = list(x=c(1,2,1,2,2))` uses again the same ordering and implies that transitions 1 → 2 and 2 → 3 have the same effect of `x`, *and* that transitions 1 → 4, 2 → 4, and 3 → 4 have the same effect of `x`. In other words, there are two effects for `x`, one for moving forward through the living states, and one for dying.

The estimated misclassification matrix is

```
> round(model$Ematrices$baseline,2)
        State 1 State 2 State 3 State 4
State 1    0.97    0.03    0.00       0
State 2    0.17    0.77    0.07       0
State 3    0.00    0.10    0.90       0
State 4    0.00    0.00    0.00       1
```

This implies, for example, that the probability to classify a latent state 2 as an observed state 1 is 0.17. Estimated effects of `age` and `x` on the hazards can be assessed in a standard way and we will not discuss them here.

The estimation of life expectancies in `elect` is based on the fitted latent progressive four-state process; the estimated misclassification is not taken into account. This is because we are interested in duration in state after we have take into account the measurement error. The code for `elect` starts with defining the data for the state-distribution model and specifying the parameter restrictions:

```
R>  sddata <- dta[dta$firstobs == 1,]
R>  RestrAndConst <- c(1,2,3,4,5, 6,0,0,0,0, 7,8,7,8,8)
R>  CHECK <- check.RestrAndConst(model, RestrAndConst, PRINT = FALSE)

<RestrAndConst> correctly defined.
```

Note that the vector `RestrAndConst` is in line with the parameter restrictions in the `msm` call. The function `check.RestrAndConst` is used to check this vector with the fitted model produced by `msm`. The function has an argument `PRINT = TRUE` to provide more details. Estimated life expectancies are next obtained by

```
R>  LEs    <- elect(x = model, b.covariates = list(age=0, x=1),
                statedistdata = sddata, h = 0.01, setseed = 1234,
                RestrAndConst = RestrAndConst, age.max = 50, S = 500)
R> summary(LEs, sd.model = TRUE)

-----------------------------
elect summary
-----------------------------
Covariate values in the multi-state model:
age   x
  0   1

No state-distribution model was fitted.
Life expectancies:
Using simulation with  1000 replications

Point estimates, and mean, SEs, and quantiles from simulation:
       pnt     mn    se 0.025q   0.5q 0.975q
e11  7.983  8.024 0.738  6.670  7.987  9.578
e12  3.711  3.615 0.662  2.491  3.524  5.077
e13  1.969  1.974 0.642  0.950  1.895  3.550

...
-----------------------------
```

Because all patients start in state 1 at `age = 0`, `elect` does not need to fit a model for the distribution of state at `age = 0`. In this case, $e_{\bullet s}(t|\boldsymbol{x}) = e_{1s}(t|\boldsymbol{x})$, for $s = 1, 2, 3$; that is, the marginal life expectancies are equal to the life expectancies for state s given initial state 1.

Expected duration in state 3 for severe CAV is shorter than expected duration in state 2 for mild/moderate CAV:

```
R> plusmin(LEs, index = c(2, 3), func  = "minus", digits=2)
           pnt   mn   se 0.025q 0.5q 0.975q
func(LEs) 1.74 1.64 0.77   0.08 1.66   3.33
```

Since the `elect` estimation is based on numerical integration, arguments in the `elect` call can be used to specify the approximation that is involved. In general, smaller `h` and larger `age.max` will give better results, but are computationally intensive.

For the CAV example, Table 1 shows the impact of varying the specification of `h`, `age.max`, and `method`. The specification of `h` should take the time scale of the model into account. In the model for CAV, years since heart transplant is the time scale; for example, if `h = 0.5`, then the integral is approximated on a half-year grid. The specification of `age.max` should be such that the probability to survive beyond `age.max` is assumed to be negligible.

Clearly, specifications of `h` and `age.max` have an impact. However, Table 1 also shows that beyond some point, a smaller `h` or a larger `age.max` does not lead to a

Table 1: Exploring the effect of argument specifications on the `elect` estimation in the CAV example. Point estimates only.

Argument specification		`e11`	`e12`	`e13`	Fixed arguments
`h`	0.005	7.984	3.711	1.969	`age.max` = 50
	0.010	7.983	3.711	1.969	`method` = `"step"`
	0.050	7.970	3.710	1.969	
	0.100	7.954	3.708	1.968	
	0.500	7.829	3.693	1.961	
	1.000	7.676	3.669	1.952	
`age.max`	100	7.983	3.714	1.974	`h` = 0.010
	75	7.983	3.714	1.974	`method` = `"step"`
	50	7.983	3.711	1.969	
	25	7.976	3.387	1.601	
	10	6.418	1.289	0.383	
`method`	`"step"`	7.983	3.711	1.969	`h` = 0.010
	`"MiddleRiemann"`	7.981	3.711	1.969	`age.max` = 50
	`"Simpson"`	7.978	3.711	1.969	

relevant change in the estimated life expectancies. For example, given `age.max` = 50 and `method` = `"step"`, switching from `h` = 0.010 to `h` = 0.005 is not worth the extra computational effort (with `h` = 0.005 the estimation takes about twice as long as with `h` = 0.010). Options `"step"`, `"MiddleRiemann"`, and `"Simpson"` represent standard methods for numerical integration. Table 1 shows that once `h` and `age.max` are fine-tuned, the results are consistent across the three methods. We see this as an indication of successfull numerical integration.

In Table 1, we investigated the numerical integration using point estimates only; that is, with specification `S` = 0. When `S` > 0, simulation is used in `elect` to obtain standard errors and confidence intervals, the same numerical integration is used repeatedly with different parameter vectors. The extra computational effort increases linear with increasing `S`: if the number of simulations is doubled, the computational time will also double.

4 Discussion

We have demonstrated that `elect` can be used to compute state-specific and marginal life expectancies on the basis of continuous-time multi-state models fitted with `msm`. There is extended functionality to estimate confidence intervals and functions of these life expectancies.

The package `elect` only works with multi-state Gompertz models that are fitted with `msm`. The underlying methodology, however, extends to any other parametric hazard model. We are planning an extension of `elect` that allows for parametric models that cannot be fitted in `msm`, such as the Weibull model and the log-normal model.

It is important to understand that the estimation of life expectancies is completely reliant on the fit of the estimated multi-state model. As stated in Van den Hout (2017), if there is bias in the estimated model, then this bias is propagated in the estimation of the life expectancies. Even if there is no bias in the estimated model, one has to take into account the age range in the data to which the model is fitted. In most cases, the definition of the life expectancies implies an extrapolation of the model beyond the study time. Say a longitudinal study is set up where all individuals are 75 years old at baseline. If the follow-up time is 15 years, then estimated life expectancies are based on extrapolation of the model beyond the age range in the data. In addition, if the study started in 2000, and the results are used to compute life expectancies for individuals who are 75 years of age in 2015, then there is also an extrapolation across birth cohorts.

If the only data available are the longitudinal data for the transition model, we can use a subset of the data to fit the multinomial regression model (5) for the state distribution. If the baseline of the longitudinal data is representative of the population state distribution, then it makes sense to use the baseline records as the data for the state-distribution model. This was illustrated in the examples. It is also possible to use the longitudinal data instead (bar the observations in the dead state). This second option might be of interest when at the study baseline, all individuals are in one state at the same age but inference is aimed at marginal residual life expectancies at a later age. Yet another possibility is to use a second dataset. Any data that contains information on the age of being in the living states can be used. This may be of interest if one is interested in a population with the same transition process but with another state distribution.

In addition to age, `elect` allows the user to include other covariates in the multinomial regression model for the state distribution. In general, we advise to include the covariates that are also used in the transition model. Note however, that a significant effect on transition hazards does not imply a significant effect on state prevalence. The latter can be investigated in `elect` by using the argument `sd.model = TRUE` in the function `summary`.

The specification of the multinomial regression model and the option to specify the data for that model, can be used to investigate the extrapolation that underlies the estimation of the marginal life expectancies. If small changes in the estimated state distribution imply big changes in estimated life expectancies, then extra care is needed with respect to the model specification and the inference.

It is important to check whether the piecewise-constant approximation for the Gompertz model is realistic. The function `explore` can be used to check the time between observations. Some preliminary knowledge of the process of interest is needed to assess whether the piecewise-constant approximation is realistic. Note that this concerns the fitting of the model only. The estimation of the life expectancies also uses the piecewise-constant approximation, but the grid for this approximation is specified in the `elect` call by argument `h`, which can be chosen as small as required.

There are no restrictions on the number of states in `elect` or on the pattern of the transitions between the living states; that is, the transition process can be progressive or reversible, or a combination thereof; see for example, the four-state model in Robitaille et al. (2018) and the five-state model in Van den Hout (2017, Chapter 7). The methods for the estimation of life expectancies can be applied to processes with multiple death states, but the current version of `elect` does not support this.

Typically, the research questions and the available multi-state data will define the number of states. Although it is relatively straightforward to define models for a stochastic process with many states, estimation of model parameters may be hampered by lack of information in the data. It is important to explore the data with regard to the multi-state process before fitting models. For example, if the sample size is large with respect to the number of individuals but there are just a few transitions between states, then estimation of model parameters will be problematic. The same holds for covariate information; for example, if there are men and women in the study but only women transition to other states, then adding gender as a covariate will lead to estimation problems. The package `elect` is developed as an add-on to `msm` and is only of use when `msm` is able to fit a model to the data. For instance, if `msm` is not able to estimate the uncertainty of the parameter estimates because if cannot evaluate an intermediate quantity (e.g., the second-order partial derivatives of the log-likelihood function), then `elect` will not compute life expectancies, and will output a warning message instead.

The combined usage of `msm` and `elect` has potential in medical studies, epidemiology, demography, health economics, ecology, and actuarial science. Increasingly, research interest is not limited to residual total life expectancy but also includes questions about life expectancy in specific health states. For example, what proportion of total life expectancy will be spent in ill health? Or, given states that describe disease progression or disease combinations, how many years will be spent in each of the disease states? If longitudinal data are available, `elect` can contribute to this research.

Acknowledgments

Mei Sum Chan was supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) North Thames at Bart's Health NHS Trust. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. Fiona Matthews and the initial `elect` development were funded by the Medical Research (grant WBS/U.1052.00.013).

References

Aalen, O. O., Borgan, O., and Gjessing, H. (2008). *Survival and Event History Analysis.* New York: Springer.

Blossfeld, H.-P. and Rohwer, G. (2002). *Techniques of Event History Modeling. New Approaches To Causal Analysis. Second Edition.* Mahwah: Lawrence Erlbaum Associates.

Cai, L., Lubitz, J., Hayward, M. D., Hagedorn, A., Saito, Y., and Crimmins, E. (2010). Estimation of multi-state life table functions and their variability from complex survey data using the SPACE Program. *Demographic Research*, 22:129–158.

Crowther, M. J. and Lambert, P. C. (2017). Parametric multistate survival models: Flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. *Statistics in Medicine*, 36(29):4719–4742.

de Wreede, L. C., Fiocco, M., and Putter, H. (2010). The mstate Package for Estimation and Prediction in Non- and Semi-Parametric Multi-State and Competing Risks Models. *Computer Methods and Programs in Biomedicine*, 99:261–274.

Hoogendijk, E. O., van der Noordt, M., Onwuteaka-Philipsen, B. D., Deeg, D. J. H., Huisman, M., Enroth, L., and Jylha, M. (2019). Sex differences in healthy life expectancy among nonagenarians: A multistate survival model using data from the vitality 90+ study. *Experimental Gerontology*, 116:80 – 85.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer.

Izmirlian, G., Brock, D., Ferrucci, L., and Phillips, C. (2000). Active life expectancy from annual follow-up data with missing responses. *Biometrics*, 56:244–248.

Jackson, C. H. (2011). Multi-state models for panel data: the msm package for R. *Journal of Statistical Software*, 38.

Jackson, C. H. (2016). flexsurv: A platform for parametric survival modeling in r. *Journal of Statistical Software, Articles*, 70(8):1–33.

Kalbfleisch, J. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80:863–871.

Lièvre, A., Brouard, N., and Heathcote, C. (2003). The estimation of health expectancies from cross-longitudinal surveys. *Mathematical Population Studies*, 10:211–248.

Mandel, M. (2013). Simulation-based confidence intervals for functions with complicated derivatives. *The American Statistician*, 67:76–81.

Mueller, L. D., Nusbaum, T. J., and Rose, M. R. (1995). The gompertz equation as a predictive tool in demography. *Experimental Gerontology*, 30:553–569.

Norris, J. R. (1997). *Markov Chains*. Cambridge: Cambridge University Press.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robitaille, A., Van den Hout, A., Machado, R. J. M., Bennett, D. A., Cukic, I., Deary, I. J., Hofer, S. M., Hoogendijk, E. O., Huisman, M., Johansson, B., Koval, A. V., Van der Noordt, M., Piccinin, A. M., Rijnhart, J. J. M., Singh-Manoux, A., Skoog, J., Skoog, I., Starr, J., Vermunt, L., Clouston, S., and Muniz-Terrera, G.

(2018). Transitions across cognitive states and death among older adults in relation to education: A multistate survival model using data from six longitudinal studies. *Alzheimer's & Dementia*, 14(4):462 – 472.

Saito, Y., Robine, J.-M., and Crimmins, E. M. (2014). The methods and materials of health expectancy. *Statistical Journal of the IAOS*, 30:209–223.

Sharples, L. D., Jackson, C. H., Parameshwar, J., Wallwork, J., and Large, S. R. (2003). Diagnostic accuracy of coronary angiography and risk factors for post-heart-transplant cardiac allograft vasculopathy. *Transplantation*, 76:679–682.

Taylor, R., Conway, L., Calderwood, L., Lessof, C., Cheshire, H., Cox, K., and Scholes, S. (2007). *Technical Report (wave 1): Health, Wealth and Lifestyles of the Older Population in England: The 2002 English Longitudinal Study of Ageing.* National Institute for Social Research.

Van den Hout, A. (2017). *Multi-State Survival Models for Interval-Censored Data.* Boca Raton, FL: Chapman & Hall/CRC.

Van den Hout, A., Ogurtsova, E., Gampe, J., and Matthews, F. E. (2014). Investigating healthy life expectancy using a multi-state model in the presence of missing data and misclassification. *Demographic Research*, 30:1219–1244.

Van der Noordt, M., Van der Pas, S., Van Tilburg, T. G., Van den Hout, A., and Deeg, D. J. H. (2018). Changes in working life expectancy with disability in the Netherlands, 1992-2016. *Scandinavian Journal of Work, Environment & Health, (Online First, 31 August 2018).*

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S (4th edition).* New York: Spinger.

Willekens, F. and Putter, H. (2014). Software for multistate analysis. *Demographic Research*, 31:381–420.