DTU

02443 Stochastic Simulation
2020-6-15
SASE,BFN/sase,bfn

Simulation and estimation in a Markov model of breast cancer

1 Introduction

Markov models are frequently used in biostatistics to model the long-term development of various diseases. The purpose of this project, is to simulate and analyze a Markov model of breast cancer, which describes possible complications following a surgery for breast cancer.

1.1 Markov Chains

Observe the random variables $X_t, t = 0, 1, \ldots$, indexed by time, which can attain the values $1, 2, \ldots, N$. We assume X_t has the Markov property, which means the future is conditionally independent of the past given the present. We may phrase this mathematically as

$$P(X_{t+1}|X_1, X_2, \ldots, X_t) = P(X_{t+1}|X_t).$$

For simulation purposes, this means we can determine the next value of X_t, i.e. X_{t+1}, by only accounting for the value of X_t. We may describe the relation between X_t and X_{t+1} as the probabilities of going from state $i \in \{1, 2, \ldots, N\}$ to state $j \in \{1, 2, \ldots, N\}$. We denote these probabilities as p_{ij}. These probabilities are collected in an $N \times N$ probability matrix:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{bmatrix}.$$

As an example, assume that X_t begins in the first state, i.e. $X_0 = 1$. The probability that $X_1 = 1$, is then p_{11}, the probabiliy that $X_1 = 2$ is p_{12}, and so on. To ensure we get an actual probability distribution, the following condition must be met:

$$\sum_{j=1}^{N} p_{ij} = 1, \quad \text{for all } i = 1, 2, \ldots, N.$$

Part 1: A discrete-time model

In this project, we will work with the following Markov model[1]:

The model follows women after they had their breast tumor removed. The cancer may reappear close to the removed tumor. This is called local recurrence, and the woman enters state 2. The cancer may also reappear distant from where it was operated. This is called distant metastatis. Both things may also occour. Death can occur from any state. In this model, once the death state has been entered, it can never be left again. This means the simulation should be terminated once this state has been reached.

Task 1

Use the following probability matrix:

$$\mathbf{P} = \begin{bmatrix} 0.9915 & 0.005 & 0.0025 & 0 & 0.001 \\ 0 & 0.986 & 0.005 & 0.004 & 0.005 \\ 0 & 0 & 0.992 & 0.003 & 0.005 \\ 0 & 0 & 0 & 0.991 & 0.009 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Assume that one time step equals one month. Simulate 1000 women, all starting in state 1, until death. Summarize the lifetime distribution of the women, after surgery, for example using a histogram. In what proportion of women does the cancer eventually reappear, locally?

Simulations can be validated by ensuring they are consistent with analytical results. One way to do this, is to ensure the distribution over the states at a certain time is consistent with what we expect.

[1] Putter, Hein, et al. ”Estimation and prediction in a multi-state model for breast cancer.” Biometrical journal 48.3 (2006): 366-380.

Let $\mathbf{p}_t$ denote the probability distribution over the states at time t. This distribution can be found from the following formula:

$$\mathbf{p}_t = \mathbf{p}_0(\mathbf{P}^t).$$

Task 2

In your simulations, what is the distribution over the states at $t = 120$? Does this correspond to what we expect? Answer the question using an appropriate statistical test.

The above approach only validates the simulation at a certain time point. A better approach is to ensure the emperical lifetime distribution matches the theoretical.
It can be shown that the lifetime, T, follows a so-called discrete phase-type distribution. This distribution has probability mass function

$$P(T = t) = \boldsymbol{\pi}(\mathbf{P_s})^t \mathbf{p_s}$$

and mean

$$E(T) = \boldsymbol{\pi}(\mathbf{I} - \mathbf{P}_s)^{-1}\mathbf{1}.$$

Where $\boldsymbol{\pi}$ is the distribution over states $1, 2, 3, 4$ at $t = 0$. $\mathbf{P}_s$ is a 4×4 sub-matrix of $\mathbf{P}$ formed by removing the last row and column. $\mathbf{p}_s$ is column vector indicating the probability of dying from states $1, 2, 3, 4$. $\mathbf{1}$ is a vector of ones of appropriate dimension.

1.2 Task 3

Does your simulated lifetimes follow this distribution?

Task 4

Estimate the expected lifetime, after surgery, of a woman who survives the first 12 months following surgery, but whose breast cancer has also reappeared within the first 12 months, either locally or distant.

Hint: Use rejection sampling. Simulate a number of women, and discard all simulations that do not meet the requirements. Do this until you have reached 1000 acceptable simulations.

Task 5

What fraction of women die within the first 350 months? Answer this by simulating 200 women, and record the fraction. Do this 100 times. Use control variates to reduce the variance.
How large a reduction in variance do you see, using control variates, as opposed to the crude Monte Carlo estimator?

Hint: Use the mean lifetime after surgery of the 200 simulations as the control variate.

Task 6

For the report, consider the following questions: What assumptions underlie the discrete time Markov chain model? Are those assumptions realistic? How may we relax some of these assumptions, possibly at the cost of increased model complexity?

Part 2: A continuous-time model

In the previous part, we assumed transitions from one state to another only happened once a month. In a more realistic model, transitions may occur at any time. Markov chains where transitions occur in continuous time are called Continuous-Time Markov Chains (CTMC). A CTMC is specified by a transition-rate matrix:

$$\mathbf{Q} = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1N} \\ q_{21} & q_{22} & \cdots & q_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N1} & q_{N2} & \cdots & q_{NN} \end{bmatrix}.$$

The non-diagonal elements of this matrix must be greater than 0. The diagonal elements, q_{ii}, are chosen as follows:

$$q_{ii} = -(q_{i1} + \ldots + q_{i(i-1)} + q_{i(i+1)} + \ldots q_{iN}), \quad \text{for } i = 1, \ldots, N. \tag{1}$$

This is done to ensure the row sums are 0 (If you're interested in why this is done, you should refer to other literature on Markov processes).

The q_{ij}'s are not probabilties. Instead, q_{ij} is the rate with which the CTMC, $X(t)$, moves from state i to state j, given that $X(t)$ is in state i.

An important property of CTMCs is that the sojourn time (that is, the time $X(t)$ remains in a state) in state i is exponentially distributed with rate $-q_{ii}$.

As an example, assume $X(0) = 1$. The CTMC will then remain in state 1, for an exponentially distributed amount of time with rate $-q_{11}$. It will then jump to state 2 with probability $-\frac{q_{12}}{q_{11}}$, to state 3 with probability $-\frac{q_{13}}{q_{11}}$, and so forth.

Task 7

As before, we choose one time-unit equal to one month. Use the following transition-rate matrix:

$$\mathbf{Q} = \begin{bmatrix} -0.0085 & 0.005 & 0.0025 & 0 & 0.001 \\ 0 & -0.014 & 0.005 & 0.004 & 0.005 \\ 0 & 0 & -0.008 & 0.003 & 0.005 \\ 0 & 0 & 0 & -0.009 & 0.009 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Simulate 1000 women, all starting in state 1, until death. Summarize the lifetime distribution after surgery, for example in a histogram. Report the mean, along with a confidence interval, and the standard deviation, also with a confidence interval. In what proportion of women has the cancer reappeared distantly after 30.5 months?

The lifetime distribution now follows a continuous time phase-type distribution. This has distribution function given by

$$F_T(t) = 1 - \mathbf{p}_0 \exp(\mathbf{Q}_s t)\mathbf{1}.$$

Where $\mathbf{Q}_s$ is a sub-matrix of $\mathbf{Q}$, where last row and column are removed. In this case,

$$\mathbf{Q}_s = \begin{bmatrix} -0.0085 & 0.005 & 0.0025 & 0 \\ 0 & -0.014 & 0.005 & 0.004 \\ 0 & 0 & -0.008 & 0.003 \\ 0 & 0 & 0 & -0.009 \end{bmatrix}.$$

$\mathbf{1}$ is a column vector of ones, of appropriate dimension. $\exp(\mathbf{Q_s}t)$ is called the matrix exponential of the matrix $\mathbf{Q_s}t$. It is defined as the infinite sum

$$\exp(\mathbf{Q}_s t) = \sum_{i=1}^{\infty} \frac{(\mathbf{Q_s}t)^i}{i!}.$$

In R, it can be calculated with the function expm() from the library expm, in MATLAB, it can be computed with the function expm(), and in Python with scipy.linalg.expm for Python.

Task 8

Compare the emperical lifetime distribution function, from your simulations, to the theoretical, using an appropriate statistical test.

When working with these types of models, one is often interested in the effect of a certain treatment. To compare two treatments visually, it is common to plot their survival functions, $S(t)$, in the same plot. The survival function is defined as the proportion of women alive at time t, i.e.

$$S(t) = P(T > t).$$

An unbiased estimator of the survival function, $\widehat{S}(t)$, is the Kaplan-Meier estimator,

$$\widehat{S}(t) = \frac{N - d(t)}{N}.$$

Where N is the total number of women, and $d(t)$ is the number of women who have died at time t.

Task 9

A certain preventitive treatment results in the following transition-rate matrix instead:

$$\mathbf{Q} = \begin{bmatrix} * & 0.0025 & 0.00125 & 0 & 0.001 \\ 0 & * & 0 & 0.002 & 0.005 \\ 0 & 0 & * & 0.003 & 0.005 \\ 0 & 0 & 0 & * & 0.009 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Simulate 1000 women who have received this treatment. Plot the Kaplan-Meier estimate of the survival function. In the same figure, plot the Kaplan-Meier estimate of the survival function for women, who have not received this treatment. Does the treatment appear to have an effect?

The log-rank test is a statistical test to compare the survival functions of two samples. Read about it here: `https://en.wikipedia.org/wiki/Log-rank_test`

Task 10 (Optional)

Does the preventitive treatment have a significant effect on the survival function? Answer the question using a log-rank test.

Task 11

For the report, consider the following questions: What assumptions have been eliminated, by going from the discrete to the continuous time model? What have been added? How could the model be extended, such that the sojourn times are Erlang distributed?

Part 3: Estimation

In practice, **Q** is unknown. The aim of this part is to estimate it from the kind of observations we may encounter in practice.

In practice, the state of the women after surgery is monitored at screenings in the doctor's office every few years. The observations, for each women, is a short time series of states. For this project we assume the state is observed every 4'th year (48 months).

Task 12

Simulate 1000 women, starting in state 1, until death, using the same $\mathbf{Q}$ as in the previous part. For each of the women, create a vector (or time series) of her observed states $\mathbf{Y}^{(i)}$. A time series will consist of the values $\mathbf{Y}^{(i)} = (X(0), X(48), X(96), \ldots,)$. The time series should continue until death, thus the last value in each of the time series should be 5.

For the remainder of this part, we will assume those 1000 time series is all we have observed.
It can be shown that an unbiased estimator of the transition rates, q_{ij}, is

$$q_{ij} = \frac{N_{ij}}{S_i}, \quad \text{for } i \neq j. \tag{2}$$

Where N_{ij} is the total number of jumps (for all of the women) from state i to j, and S_i is the total sojourn time in state i (for all of the women). The diagonal elements are found from equation (1). The problem is that N_{ij} and S_i are unknown, all we know are the timeseries of observations every 4'th year.
One approach to estimation is to recreate N_{ij} and S_i from the partial information that we have observed.

Task 13

Implement the following algorithm to estimate the q_{ij}'s:

Select $\mathbf{Q}^{(0)}$ as some initial guess. In the k'th iteration, do the following

1. For all of the time series, simulate a possible complete trajectory, taking the observations into account, using $\mathbf{Q} = \mathbf{Q}^{(k)}$.
2. Summarize the trajectories in the variables $N_{ij}^{(k)}$ and $S_i^{(k)}$.
3. Find $\mathbf{Q^{(k+1)}}$ using equation (2).

The above should be repeated until some convergence criterion is reached, for example until $||\mathbf{Q}^{(k)} - \mathbf{Q}^{(k+1)}||_\infty < 10^{-3}$. Using this method, it is possible to approximately recreate the parameters used in the original simulation. The above is an example of a Monte Carlo Expectation Maximization algorithm.

Hint: The toughest part of the above algorithm is the first step. It can be done as follows. Simulate the Markov process between each observation. For example, initialize the process at the first observation $y_1^{(i)}$. Then simulate the Markov process until 48 months have passed. The value $X(48)$ should equal the observed value, $y_2^{(i)}$. If it does not, reject the simulation and try again, until it does. Then, initialize the Markov process at the second observed value, $y_2^{(i)}$. Simulate until $X(96) = y_3^{(i)}$. This should be done for the interval between all the observations, and for all women.