Project 1 part 2-3

6/14/2022

Part 2: A continous-time model

In this part, we extend our model for breast cancer such that we consider Continuous-Time Markov Chains (CTMC).

Task 7

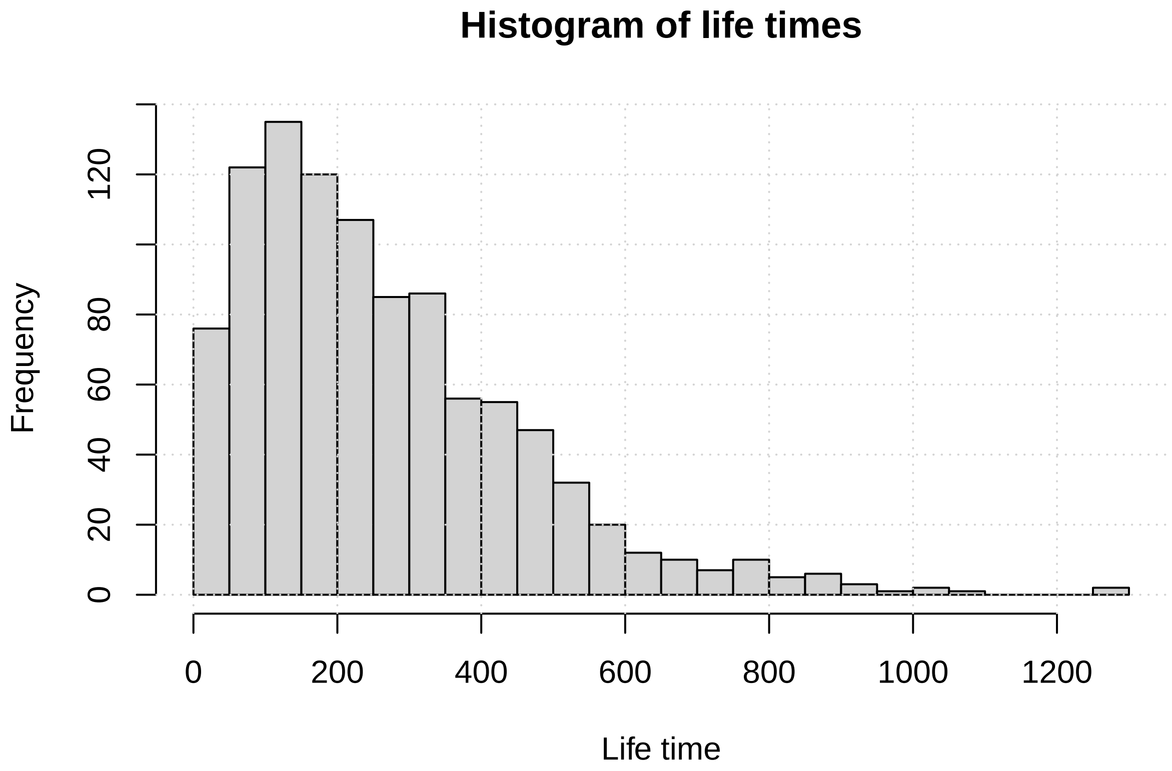
We will now simulate 1000 women all starting in state 1 until death. We will use the transition-rate matrix as given in the problem description.

```r
sim.ctmc <- function(Q) {
  ## Assume that we start in state 1
  ts = append(list(), 0)
  state = append(list(), 1)

  ii = 1
  while (Q[state[[ii]], state[[ii]]] < 0) {
    ts = append(ts,rexp(1,-Q[state[[ii]], state[[ii]]]))
    state = append(state, sample((1:5)[-state[[ii]]],size = 1,
      prob = -Q[state[[ii]],-state[[ii]] ]/Q[state[[ii]], state[[ii]]]))
    ii = ii + 1
  }
  list(ts = cumsum(unlist(ts)),
       state = unlist(state))
}

Q = t(matrix(c(-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),5,5))

set.seed(1); N.sim = 1000
sol = lapply(1:N.sim, function(i) sim.ctmc(Q))
lt = sapply(sol, function(e) e$ts[length(e$ts)])
hist(lt, breaks = 40, xlab = "Life time", main = "Histogram of life times")
grid()
```



We now want to determine the mean and variance of the life times based on the 1000 simulations. To do so we use bootstrapping. Specifically, we sample 1000 women with replacements 100 times and for each sample, we compute the mean and stadard deviation. Via the central limit theorem, we expect the distribution of both mean and standard deviation to follow a normal distribution.

```
N = 100; set.seed(1)
mu <- s <- rep(NA,N)
for (ii in 1:N) {
  lt.boot = sample(lt, replace = T)
  mu[ii] = mean(lt.boot)
  s[ii] = sd(lt.boot)
}
mean(mu) + qt(0.975,N)*sd(mu)/sqrt(N)*c(-1,1)
```

```
## [1] 262.1360 264.6615
```

```
mean(s) + qt(0.975,N)*sd(s)/sqrt(N)*c(-1,1)
```

```
## [1] 193.5469 196.1868
```

Via bootstrapping, we find that the mean is given, with a 95% confidence level, to:

$$\mu \in_{95\%} [262.14 \quad 264.66].$$

Additionally, we find that the standard deviation is given, with a 95% confidence level, to:

$$\sigma \in_{95\%} [193.55 \quad 196.19].$$

We are now asked to determine the proportion of women who has had cancer reapperead distantly after 30.5 months. We interpret this, as women who within the first 30.5 months has had distant metastatis or both local recurrance and distant metastatis, i.e., been in state 3 or 4 before $t = 30.5$.

To solve this, we again use bootstrapping and sample 1000 women with replacement 100 times.

```
N = 100; set.seed(1)
p <- sapply(1:N, function(zz){
        sum(sapply(sample(sol, replace = T),
             function(e) any(e$ts[e$state %in% c(3,4)] < 30.5) )) / N.sim
})
mean(p) + qt(0.975, N)*sd(p)/sqrt(N) * c(-1,1)
```

```
## [1] 0.08458494 0.08821506
```

On a 95% confidence interval, we find the proportion of women who has had cancer reapperead distantly after 30.5 months to be:

$$p \in_{95\%} [0.0846 \quad 0.0882].$$

Task 8

We are now given that the lifetime distribution now follows a continuous time phase-type distribution. We therefore want to compare the emperical lifetime distribution function, from our simulations, to the theoretical.

We start by a graphical comparison of the empirical PDF with the theoretical.

```
library('expm')
```

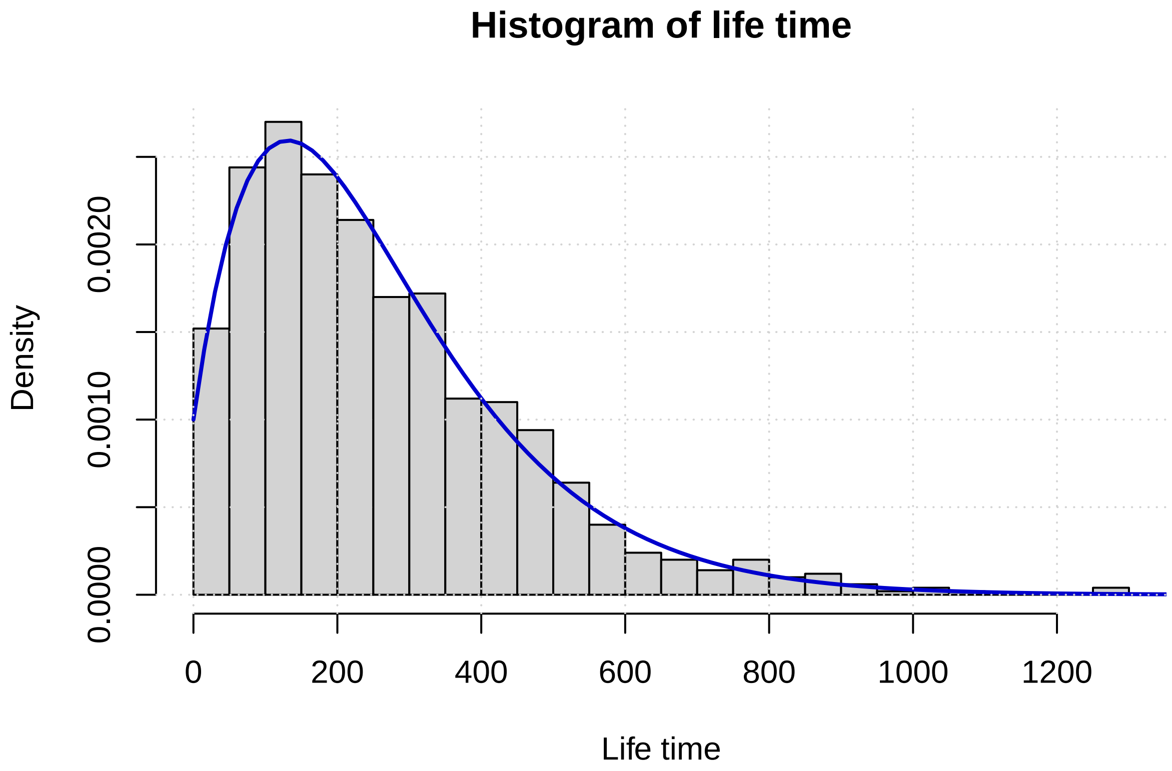
```
## Indlæser krævet pakke: Matrix
```

```
##
## Vedhæfter pakke: 'expm'
```

```
## Det følgende objekt er maskeret fra 'package:Matrix':
##
##     expm
```

```
library('numDeriv')
Qs = Q[-5,]; Qs = Qs[,-5]
Ft <- function(x) sapply(x, function(t) sum(c(1,0,0,0) %*% expm(Qs*t)))
cdf <- function(x) sapply(x, function(t) 1 - sum(c(1,0,0,0) %*% expm(Qs*t)))
pdf <- function(x) grad(cdf,x)

hist(lt, breaks = 40, xlab = "Life time", freq = F, main = "Histogram of life time")
curve(pdf, from = 0, to = 1500, add = T, col = 'blue3', lwd = 2)
grid()
```



By simple visual inspection, the empirical densities seems to follow the theory relatively well. Since we are dealing with a continous distribution, a relevant statistical test would be the Kolmogorov-Smirnov test, where we compare the empiricial distribution with the theoretical CDF (which can be found from the survival function).

```
ks.test(lt, cdf)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  lt
## D = 0.01735, p-value = 0.9242
## alternative hypothesis: two-sided
```

We get a p-value of 0.92. Therefore we cannot reject that the empirical data corresponds to the theoretical CDF.

Since we are already working with the theoretical PDF we might as well find the theoretical mean and standard deviation of the life times.

```
m1 = integrate(function(x) x*grad(cdf,x), 0, 10000)
m2 = integrate(function(x) x^2*grad(cdf,x), 0, 10000)
m1$value
```

```
## [1] 262.3716
```

```
sqrt(m2$value - m1$value^2)
```

```
## [1] 190.7735
```

From the theoretical PDF we find the mean life time to be:

$$\mu = 262.37.$$

And the standard deviation:

$$\sigma = 190.77.$$

Notice they are quite close to the estimate parameters from the previous exercise, albeit not equal.

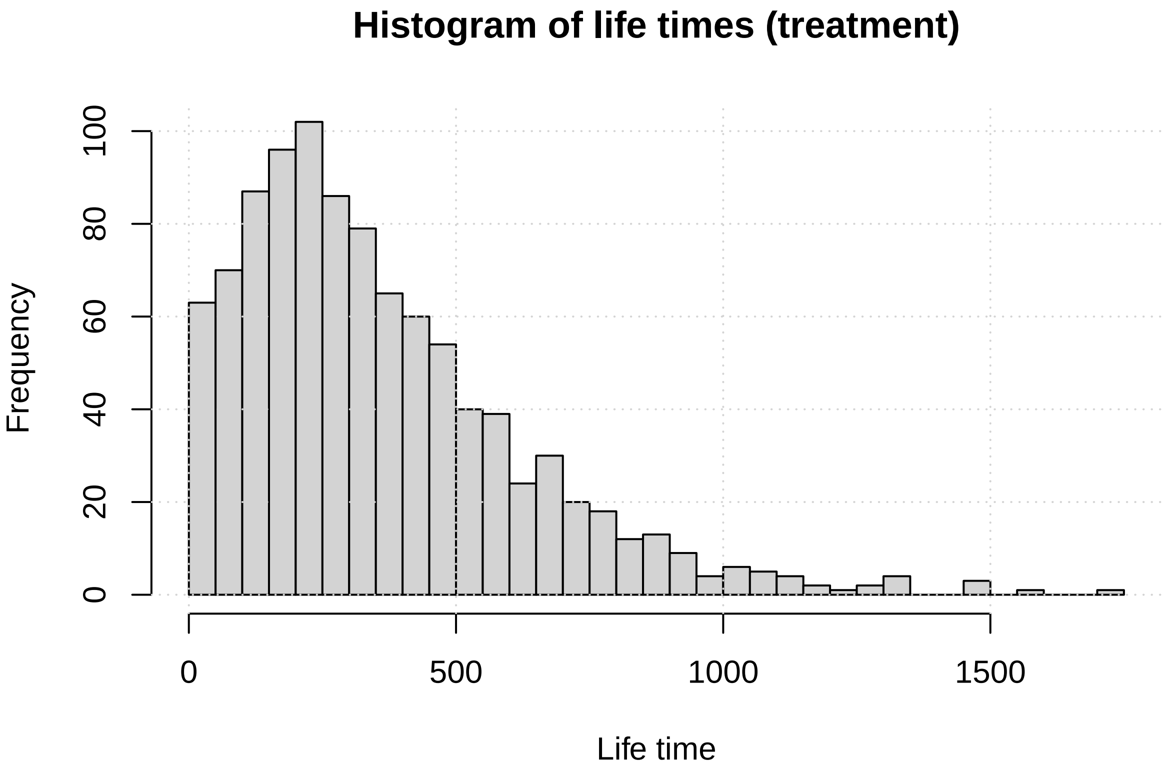
Task 9

A certain preventitive treatment results in a different transition-rate matrix. We now want to simulate 1000 new women, who have received this treatment and compare with the original women.

We start by simulating the 1000 women who has undergone treatment.

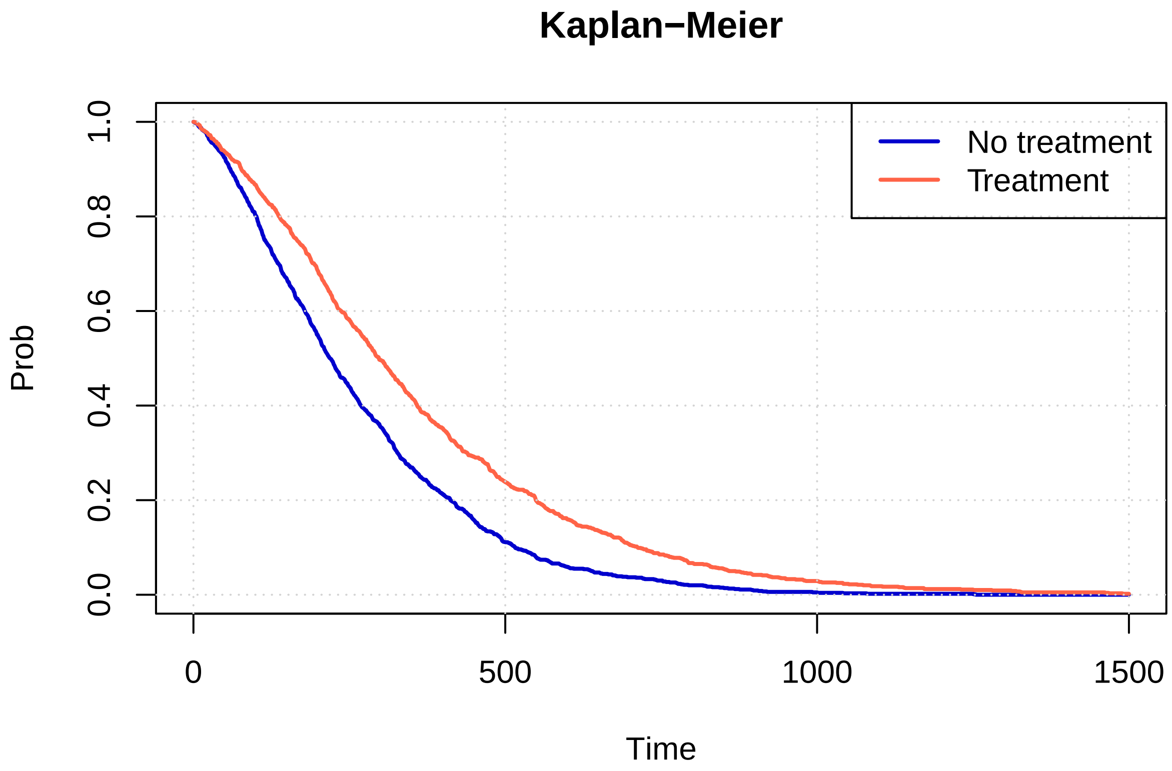
```
Q.treat = t(matrix(c(NA, 0.0025, 0.00125,     0, 0.001,
               0,NA,  0, 0.002, 0.005,
               0,     0, NA, 0.003, 0.005,
               0,     0,      0,NA, 0.009,
               0,     0,      0,     0,    0),5,5))
diag(Q.treat) = -apply(Q.treat,1,sum, na.rm = T)

N = 1000; set.seed(1)
sol.treat = lapply(1:N, function(i) sim.ctmc(Q.treat))
lt.treat = sapply(sol.treat, function(e) e$ts[length(e$ts)])
hist(lt.treat, breaks = 40, xlab = "Life time",
     main = "Histogram of life times (treatment)")
grid()
```



In order to compare women who has undergone treatment with women who has not, we plot the Kaplan-Meier estimator for both.

```r
kaplan.meier <- function(t,lt) sapply(t, function(t) (length(lt)-sum(lt<t))/length(lt))
plot(0:1500, kaplan.meier(0:1500,lt), type = 'l', col = 'blue3', lwd = 2,
     xlab = "Time", ylab = "Prob", main = "Kaplan-Meier")
lines(0:1500, kaplan.meier(0:1500,lt.treat), col = "tomato", lwd = 2)
grid()
legend('topright',c('No treatment','Treatment'), col = c('blue3','tomato'),
       lty = 1, lwd = 2)
```



From the Kaplan-Meier it clearly seems that the treatment does prolong the life time of some women.

Task 10

To actually determine if the treatmenet has any effect, we perform a log-rank test.

```
m = 100
N <- O <- matrix(NA, m, 2) -> E

N[,1] = sapply(seq(0,1500, length.out = m), function(t) sum(lt>t))
N[,2] = sapply(seq(0,1500, length.out = m), function(t) sum(lt.treat>t))

O[,1] = sapply(seq(0,1500, length.out = m), function(t) sum(lt<=t))
O[,2] = sapply(seq(0,1500, length.out = m), function(t) sum(lt.treat<=t))

Nj = apply(N,1,sum)
Oj = apply(O,1,sum)

E = N*Oj/Nj
V = E*(Nj-Oj)/Nj*((Nj-N)/(Nj-1))

(Z = sum(O[,1] - E[,1]) / sqrt(sum(abs(V[,1]))))
```

```
## [1] 64.84072
```

```r
1 - pnorm(Z)
```

```
## [1] 0
```

We find a test statistic, Z, to be

$$Z = 64.84$$

which we expect to follow $Z \sim N(0,1)$. This yields a p-value of 0. We therefore reject the null hypothosis that the survival functions are equal. In other words, the treatment does have a significant effect.

Task 11

When going from a discrete model to a continous model, we remove the assumption that women only change state, i.e., get cancer, at specific time points. In the continous model women can get cancer at any time, and the do so with a constant rate. We have added the assumption that the time spend in each state is exponentially distributed. If we were to extend the model such that you only go from one state to another when multiple exponentially distributed events has occured, then the sojourn times would be Erlang.

Part 3

In the final part, we consider the case where women are only monitored every 48 months.

Task 12

We now want to create 1000 time series containing the state of a women, representing the doctor visits made by 1000 women. The time series for woman i, $Y^{(i)}$, will therefore look like

$$Y^{(i)} = (X(t_1), X(t_2), ...), \quad t_1 = 0,\ t_2 = 48, ...$$

where $X(t)$ denotes the state at time t.

```r
get.ts <- function(e) sapply(0:(ceiling(e$ts[length(e$ts)] / 48)),
            function(ii) ifelse(any(e$ts<ii*48),
                                e$state[tail(which(e$ts<ii*48),1)], 1))
set.seed(1); N = 1000
tss = lapply(lapply(1:N, function(i) sim.ctmc(Q)), function(e) get.ts(e))
```

Task 13