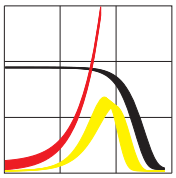


Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Konrad-Zuse-Strasse 1 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
http://www.demogr.mpg.de

MPIDR TECHNICAL REPORT 2010-003
MAY 2010

Example for a Piecewise Constant Hazard Data Simulation in R

Rainer Walke (walke@demogr.mpg.de)

This technical report has been approved for release by: Vladimir Shkolnikov (shkolnikov@demogr.mpg.de), Head of the Laboratory of Demographic Data.

Example for a Piecewise Constant Hazard Data Simulation in R

Rainer Walke
Max Planck Institute for Demographic Research, Rostock

2010-04-29

Computer simulation may help to improve our knowledge about statistics. In event-history analysis, we prefer to use the hazard function instead of the distribution function of the random variable time-to-event. In this paper, we provide a proof-of-concept that may be used to derive random times following a piecewise constant hazard function. We use the statistical software package R.

Keywords Demography, event-history analysis, hazard function, computer simulation

Background

In survival analysis, we analyze the time to an event. What is the distribution function for this random variable time-to-event? For a continuous random variable, this is equivalent to finding either the probability density function, the survival function or the hazard function. So if we know one, we know the others. The interpretation of the hazard function is clear: it describes the time-dependent risk for an event.

To compare different groups, proportional hazard models will be used often. They assume a basic common time-dependent hazard function. It will be shifted proportionally depending on the group parameters.

Here we use a proportional hazards model with a piecewise constant baseline hazard. This is a simple but powerful class of models.

Computer simulation may help to improve our knowledge about statistics. Most statistical software packages enable users to draw random samples from a number of different distribution families. I do not know whether one of them provides a piecewise constant hazard random variable directly.

In this example, we simulate event history data for a given outcome. We use the statistical software package R [R 2009].

1 Task

We have estimates from a proportional hazards model with a piecewise constant baseline hazard. We want to simulate a sample of individuals reproducing the given coefficients.

For an illustration, we take a simple data set from Ulla-Britt Lithell [Lithell 1981]. In her work she compared the infant mortality in the early 19th century for three parishes. But we restrict ourselves to the rural parish of Umeå.

R output Umeå example

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.008054   0.188950 -21.212  < 2e-16 ***
age2         0.006103   0.212686   0.029    0.977
age3        -0.964671   0.198521  -4.859 1.18e-06 ***
age4        -1.895425   0.215401  -8.800  < 2e-16 ***
period2      0.434824   0.106436   4.085 4.40e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 201.8900  on 7  degrees of freedom
Residual deviance:   3.5722  on 3  degrees of freedom
AIC: 57.425

Number of Fisher Scoring iterations: 4

> umea1$null.deviance - umea1$deviance
[1] 198.3178
> cbind(coef(umea1),confint.default(umea1))
                             2.5 %     97.5 %
(Intercept) -4.00805406 -4.3783890 -3.6377191
age2         0.00610333 -0.4107542  0.4229609
age3        -0.96467108 -1.3537652 -0.5755770
age4        -1.89542489 -2.3176031 -1.4732467
period2      0.43482383  0.2262135  0.6434342
> cbind(exp(coef(umea1)), exp(confint.default(umea1)))
                            2.5 %     97.5 %
(Intercept) 0.01816872 0.01254555 0.02631229
age2        1.00612199 0.66314989 1.52647460
age3        0.38110853 0.25826600 0.56238031
age4        0.15025448 0.09850942 0.22918021
period2     1.54469090 1.25384334 1.90300488
>
```

Using her data, we find that the absolute hazard is about $\exp(-4.01) = 0.0181$ per week [1]. This is true for the first week (0 to 1) of life for the cohorts 1805-1807. For the rest of the first month (1 to 4.33) the relative hazard is similar (1.01), while for the remaining time to the end of the first six months (4.33 to 26) it is much lower (0.381), and for the second half-year (26 to 52) it is even lower (0.150).

interval	begin	end	relative hazard	95% confidence interval
1	0	1	1	
2	1	4.33	1.01	0.663 . . . 1.526
3	4.33	26	0.381	0.258 . . . 0.562
4	26	52	0.150	0.099 . . . 0.229

For the 1811 cohort the risk ratio is 1.54.

cohort	years	relative hazard	95% confidence interval
1	1805-1807	1	
2	1811	1.54	1.25 . . . 1.90

In the original data, the 1805-1807 cohort started with 1,082 children, and the 1811 cohort with 344 children. We simulate 1,000 samples of this size and check the confidence intervals.

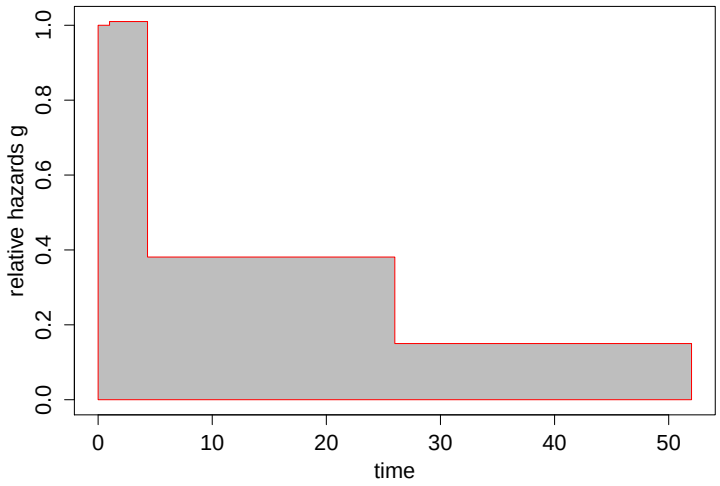
[1] 95% confidence interval: 0.0125 . . . 0.0263

2 Preparation

2.1 Piecewise constant hazard function

Given a set of time points $0 = \tau_0 < \tau_1 < \ldots < \tau_m < \tau_{m+1}$, a baseline hazard h_0 and the relative hazards $g_0 = 1, g_1 \ldots g_{m-1}, g_m$ we define a piecewise constant hazard function as

$$h(t) \;=\; h_0 \sum_{l=0}^{m} g_l I_l(t) \qquad \text{with} \quad I_l(t) = \begin{cases} 1 & \text{if} \quad \tau_l \le t < \tau_{l+1} \\ 0 & \text{if} \quad \text{elsewhere} \end{cases} .$$



The cumulated hazard reads

$$H(t) \;=\; \int_0^t h(s)ds \qquad = \quad h_0 \sum_{l=0}^{m} g_l \int_0^t I_l(s)ds.$$

Further, the survival function is

$$S(t) \;=\; \exp(-H(t)) \quad = \; \exp(-h_0 \sum_{l=0}^{m} g_l \int_0^t I_l(s)ds).$$

2.2 Piecewise exponential survival function

Determine the survival function $S_i(t)$ for a given interval $\tau_i \le t < \tau_{i+1}$. We have

$$S_i(t) \;=\; \exp\left(-h_0 \sum_{l=0}^{i-1} g_l \int_0^t I_l(s)ds - h_0 g_i \int_0^t I_i(s)ds - h_0 \sum_{l=i+1}^{m} g_l \int_0^t I_l(s)ds \right).$$

Integration simplifies to

$$S_i(t) \;=\; \exp\left(-h_0 \sum_{l=0}^{i-1} g_l(\tau_{l+1} - \tau_l) - h_0 g_i (t - \tau_i) \right),$$

in solving this equation for t, we get

$$t \;=\; \tau_i - \frac{\ln(S_i(t))}{h_0 g_i} - \frac{1}{g_i}\sum_{l=0}^{i-1} g_l(\tau_{l+1} - \tau_l). \tag{1}$$

We have to obey the condition $\tau_i \leq t < \tau_{i+1}$. This is

$$\tau_i \;\leq \tau_i - \frac{\ln(S_i(t))}{h_0 g_i} - \frac{1}{g_i}\sum_{l=0}^{i-1} g_l(\tau_{l+1} - \tau_l) \;<\; \tau_{i+1}.$$

Left condition gives

$$\ln(S_i(t)) \;\leq\; -h_0 \sum_{l=0}^{i-1} g_l(\tau_{l+1} - \tau_l) \tag{2}$$

and the right condition gives

$$-\,h_0 \sum_{l=0}^{i} g_l(\tau_{l+1} - \tau_l) \;<\; \ln(S_i(t)). \tag{3}$$

We started with a piecewise constant hazard function, and we got a piecewise exponential survival function. This piecewise exponential survival function is a strictly decreasing function. The inverse function exists, and we have provided an analytical formula for this inversion.

If we have an analytical formula for the survival function, we have the formula for the cumulative distribution function $F = 1 - S$ too.

To generate sample numbers at random for a given probability distribution, we use the inverse transform sampling. We generate uniformly in (0,1) distributed random numbers, and use the inverse function to get the random times.

2.3 Simulation recipe

- Set the time intervals, the baseline hazard, and all relative hazards.
- For every covariate combination do the following steps. (In our example we have only two sub-populations.)
  - Set the size of the sub-population.
  - Draw a uniformly (0,1) distributed random variable $S = 1 - F$.
  - Determine the right interval using the conditions 2 and 3.
  - Compute the random time t using equation 1.
- Combine the computed random times to one file.
- Censor all cases beyond the last time point.
- Compute the maximum likelihood estimates using a proportional hazard model.

3 Simulation

3.1 R code

For the calculations, we used R version 2.9.2 [R 2009] (www.r-project.org).

First we compute estimators and confidence intervals for a proportional hazard model with a piecewise constant baseline hazard function using the Umeå dataset. We may reutilize without any modification algorithms, which were developed to estimate generalized linear models with a Poisson distribution and a log-link function [Laird, Oliver 1981].

Our simulation produces 1,000 samples with the same size as our original data set. For every sample, we estimate the proportional hazard model. We check whether the resulting estimators are within the 95% confidence intervals.

The R function *survreg* does not support left-truncated data. Fortunately, we may reuse the log-linear contingency table analysis to estimate the proportional hazard model with piecewise constant baseline hazards [Laird, Oliver 1981]. Both share the same likelihood function (except a factor).

Just for comparison, we compute a proportional hazard Cox model for the simulated data, as well using the R function *coxph*.

R source code simulation program

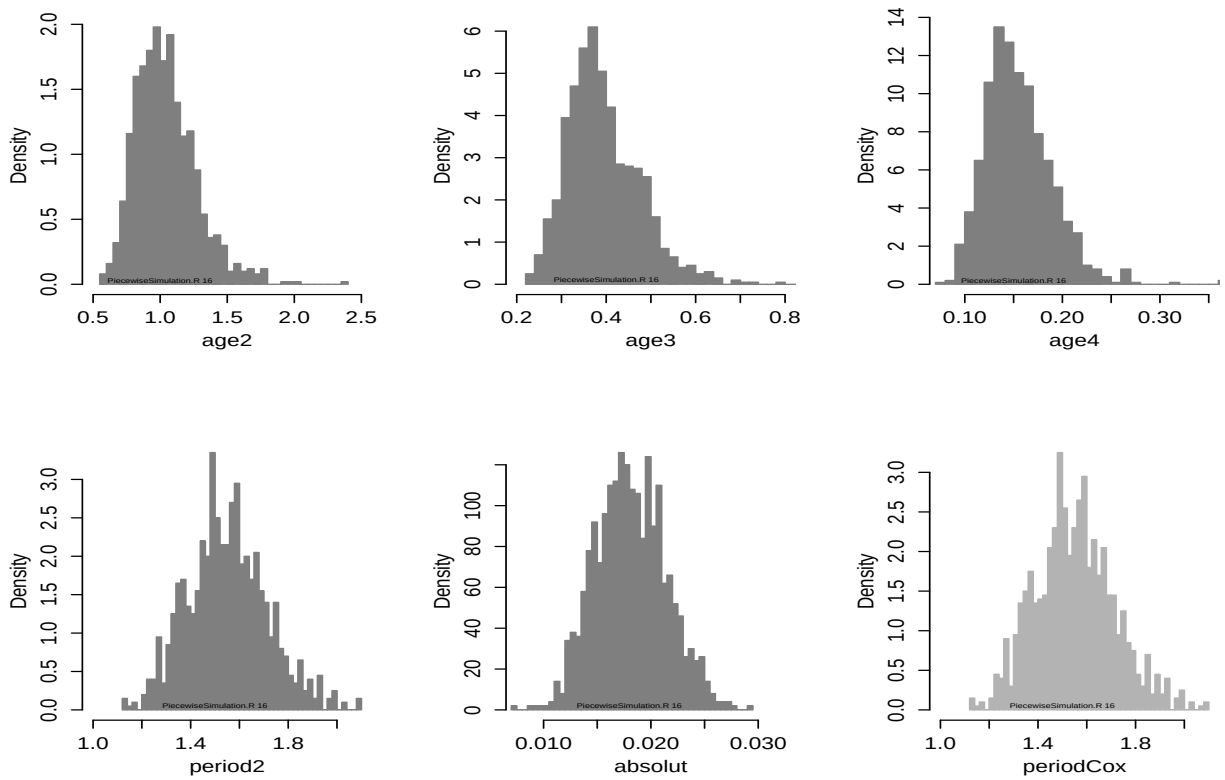
```
#
# Rainer Walke, MPI Rostock, 17-Sep-2009
# Technical Report: Piecewise constant baseline hazard data simulation
rm(list = ls())

# SVN identification
# This links the results with the source code.
svn <- unlist(strsplit("$Id: PiecewiseSimulation.R 16 2009-09-17 06:32:43Z walke $"," "))
note <- paste(svn[2],svn[3])

# motivation: orginal data analysis
age <-factor(c(1,1,2,2,3,3,4,4))
period <-factor(c(1,2,1,2,1,2,1,2))
occ <-c(19,10,70,23,134,69,57,27)
expo <-c(1072.5,339,3427,1075,20063,5991,21593,5941)
umea <- data.frame(age, period, occ, expo)
rm(age, period, occ, expo)

umea1 <- glm(occ ~ age + period + offset(log(expo)), family = poisson(link="log"), data = umea)
summary(umea1)
umea1$null.deviance - umea1$deviance
cbind(coef(umea1),confint.default(umea1))
cbind(exp(coef(umea1)), exp(confint.default(umea1)))


# prepare the simulation
library(survival)

# set a fixed starting point for the pseudo-random numbers
set.seed(3456998)

pcbhsim <- function(){
# set the start population size for both groups
  number1 <- 1082
  number2 <- 344
# set the time points
    TAU <- c(0, 1, 4.33, 26, 52)
    DT = TAU[2:5] - TAU[1:4]
# set the absolute risk
    h0 <- exp(-4.01)
# set the relative risks
    G <- c(1, 1.01, 0.381, 0.150)
# set the relative risks for the population groups
    P <- c(1, 1.54)
# create a helping matrix
  LD <- matrix(0,nrow=5, ncol=4)
  LD[lower.tri(LD)]<-1

```

```
# start with population group 1
  LS <- log(1-runif(number1))
  GP <- P[1]*G
# determine the ln(S) for all TAU
    LSM <- -h0 * as.vector(LD %*% (GP * DT))
# find the appropriate time interval
    t1 <- rep(NA,number1)
        for (i in 1:4) {
    t1 <- ifelse(LSM[i]>=LS & LS>LSM[i+1], TAU[i] + (LSM[i] - LS)/h0/GP[i], t1)
  }
# end of population group1

# start with population group 2
  LS <- log(1-runif(number2))
  GP <- P[2]*G
# determine the ln(S) for all TAU
    LSM <- -h0 * as.vector(LD %*% (GP * DT))
# find the appropriate time interval
    t2 <- rep(NA,number1)
    for (i in 1:4) {
    t2 <- ifelse(LSM[i]>=LS & LS>LSM[i+1], TAU[i] + (LSM[i] - LS)/h0/GP[i], t2)
  }
# end of population group 2

# combine both populations
  sim.data <- data.frame( rbind(cbind(t=t1,period=1),cbind(t2,2)))
  sim.data$occ <- ifelse(is.na(sim.data$t), 0, 1)
  sim.data$t <- ifelse(is.na(sim.data$t), TAU[5], sim.data$t)
  sim.data$id <- row(sim.data)[,1]
# the data set is ready


# compute a proportional hazards Cox model
  sim.cox <- coxph(Surv(t,occ) ~ period, data=sim.data)


# compute a proportional hazards model with piecewise constant baseline hazard
# split the process time
  sim.split <- survSplit(sim.data, cut=c(1, 4.33, 26), end="t", start="t0", event="occ", episode="age")

# +1 just for convenience
  sim.split$age <- factor(sim.split$age+1)
  sim.split$period <- factor(sim.split$period)

# compute the hazards using the poisson regression
# it requires to collapse the exposure times and occurences
  sim.split$expo <- sim.split$t - sim.split$t0
  sim.tab <- aggregate(sim.split[c("occ","expo")], by=list(age = sim.split$age, period = sim.split$period), sum)

  sim.poisson <- glm(occ ~ age + period +offset(log(expo)), family = poisson(link="log"), data = sim.tab)

# prepare both results for return
  return(c(exp(coef(sim.poisson)), exp(coef(sim.cox))))

} # end of function pcbhsim

# start the simulation
results <- NULL
reps <- 1000

for (i in 1:reps){
  results <- rbind(results, pcbhsim())
}
# end of simulation

summary(results)

# count the number of trials outside the confidence band
m1 <- exp(confint.default(umea1))
Cabsolut <- ifelse(m1[1,2] < results[,1] | results[,1] < m1[1,1], 1, 0)
Cage2 <- ifelse(m1[2,2] < results[,2] | results[,2] < m1[2,1], 1, 0)
Cage3 <- ifelse(m1[3,2] < results[,3] | results[,3] < m1[3,1], 1, 0)
Cage4 <- ifelse(m1[4,2] < results[,4] | results[,4] < m1[4,1], 1, 0)
Cperiod2 <- ifelse(m1[5,2] < results[,5] | results[,5] < m1[5,1], 1, 0)
CperiodCox <- ifelse(m1[5,2] < results[,6] | results[,6] < m1[5,1], 1, 0)

cbind(table(Cabsolut), table(Cage2), table(Cage3), table(Cage4), table(Cperiod2), table(CperiodCox))


# graph the simulation results
# save default
```

```
def.par <- par(no.readonly = TRUE)
layout(matrix(c(1,2,3,4,5,6),2,3,byrow=TRUE), c(1,1,1), c(1,1), FALSE)

# windows(8, 6,  record = TRUE)
hist(results[,2], freq=FALSE, breaks=40, col="gray50", border = "gray50",
  xlim=c(0.5,2.5), xlab="age2", main="", cex.axis=1.3, cex.lab=1.3, mgp=c(2.4,1,0))
  text(1,0.03,note,cex=0.5)

hist(results[,3], freq=FALSE, breaks=40, col="gray50", border = "gray50",
  xlim=c(0.2,0.8), xlab="age3", main="", cex.axis=1.3, cex.lab=1.3, mgp=c(2.4,1,0))
  text(0.4,0.1,note,cex=0.5)

hist(results[,4], freq=FALSE, breaks=40, col="gray50", border = "gray50",
  xlim=c(0.075,0.35), xlab="age4", main="", cex.axis=1.3, cex.lab=1.3, mgp=c(2.4,1,0))
  text(0.15,0.2,note,cex=0.5)

hist(results[,5], freq=FALSE, breaks=40, col="gray50", border = "gray50",
  xlim=c(1,2.1), xlab="period2", main="", cex.axis=1.3, cex.lab=1.3, mgp=c(2.4,1,0))
  text(1.5,0.05,note,cex=0.5)

hist(results[,1], freq=FALSE, breaks=40, col="gray50", border = "gray50",
  xlim=c(0.0075,0.0325), xlab="absolut", main="", cex.axis=1.3, cex.lab=1.3, mgp=c(2.4,1,0))
  text(0.018,2,note,cex=0.5)

hist(results[,6], freq=FALSE, breaks=40, col="gray70", border = "gray70",
  xlim=c(1,2.1), xlab="periodCox", main="", cex.axis=1.3, cex.lab=1.3, mgp=c(2.4,1,0))
  text(1.5,0.05,note,cex=0.5)

# reset the graphics device
par(def.par)
#
```

3.2 Results

Every simulation estimates another set of five estimators of the piecewise constant model, and one estimator for the Cox model. The following histograms show the distribution of these estimates.



The following table displays the mean values for the estimates and the number of estimates outside the 95% confidence intervals.

R output simulation results

```
> summary(results)
  (Intercept)           age2             age3             age4            period2          period
 Min.   :0.007151   Min.   :0.5590   Min.   :0.2204   Min.   :0.07433   Min.   :1.128   Min.   :1.129
 1st Qu.:0.015790   1st Qu.:0.8824   1st Qu.:0.3373   1st Qu.:0.13092   1st Qu.:1.443   1st Qu.:1.443
 Median :0.017967   Median :1.0190   Median :0.3813   Median :0.15067   Median :1.545   Median :1.545
 Mean   :0.018109   Mean   :1.0455   Mean   :0.3946   Mean   :0.15528   Mean   :1.551   Mean   :1.551
 3rd Qu.:0.020340   3rd Qu.:1.1732   3rd Qu.:0.4425   3rd Qu.:0.17462   3rd Qu.:1.656   3rd Qu.:1.655
 Max.   :0.029462   Max.   :2.3837   Max.   :0.9137   Max.   :0.36729   Max.   :2.099   Max.   :2.096
>
...
> cbind(table(Cabsolut), table(Cage2), table(Cage3), table(Cage4), table(Cperiod2), table(CperiodCox))
  [,1] [,2] [,3] [,4] [,5] [,6]
0  957  948  946  957  954  954
1   43   52   54   43   46   46
>
```

We see that 43, 52, 54, 43, 46, and 46 cases out of 1,000 simulations produce an estimator outside the computed 95% confidence intervals.

Comparing the mean simulation result with the original data estimates, we have

interval	begin	end	relative hazard	95% confidence interval	mean simulation result
1	0	1	1		1
2	1	4.33	1.01	0.663 . . . 1.526	1.046
3	4.33	26	0.381	0.258 . . . 0.562	0.395
4	26	52	0.150	0.099 . . . 0.229	0.155

.

The simulation gives 0.01811 per week [2] for the mean absolute risk.

cohort	years	relative hazard	95% confidence interval	mean simulation result
1	1805-1807	1		1
2	1811	1.54	1.25 ... 1.90	1.540

The Cox model gives a very similar result for the birth cohort comparison. This may change if we analyze another class of random variables.

Summary

We simulated data for a given set of estimates. For a large number of simulations, the resulting mean estimates approximate the estimates from the original dataset. About 95% of all simulated estimates are within the 95% confidence intervals of the given data estimates. We may use such simulations to gain insights into the sampling variability.

References

[R 2009] R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[Lithell 1981] Ulla-Britt Lithell: 'Breast-Feeding habits and their relation to infant mortality and marital fertility', Journal of Family History, pages 182-194, Summer 1981.

[Laird, Oliver 1981] Nan Laird; Donald Olivier, Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques. Journal of the American Statistical Association, Vol. 76, No 374 pp. 231-240, June 1981.

[2] original data: 0.0181 per week with 95% confidence interval: 0.0125 ... 0.0263