Statistical Methods in Cancer Epidemiology using R

Karri Seppä

Finnish Cancer Registry

Lecture 4

karri.seppa@cancer.fi

Mar 2, 2020

Time splitting

Contents:

1. Basic concepts of time to event analysis
2. Piecewise constant hazards model and
   age-specific incidence rates
3. Splitting follow-up times by age
4. Accounting for current age in rate ratio estimation

Main R functions to be covered, all in `Epi` package

- `Lexis()`
- `splitLexis()`
- `timeBand()`

Time to event analysis

Analysis of incidences = analysis of *times to event* or *failure times* or *survival times* (censored).

Mathematical concepts:

$$
\begin{aligned}
T &= \text{ time to outcome event -- random variable,} \\
S(t) = P(T > t) &= \textbf{ survival} \text{ function of } T, \\
&= \text{ probability of avoiding the event up to given time } t, \\
\lambda(t) = -S'(t)/S(t) &= \textbf{ intensity} \text{ or } \textbf{hazard} \text{ function,} \\
\Lambda(t) = \int_0^t \lambda(u)du &= -\log S(t) = \textbf{cumulative hazard}, \\
F(t) = 1 - S(t) &= 1 - \exp\{-\Lambda(t)\} = \textbf{risk} \text{ function} \\
&= \text{ probability of the outcome to occur by } t \\
&= \text{ cumulative distribution function of } T.
\end{aligned}
$$

Hazard rate or intensity function

Can be viewed as *theoretical incidence rate*. Formally

$$\lambda(t) = \lim_{\Delta \to 0} \frac{P(t < T \leq t + \Delta \mid T > t)}{\Delta}$$

$\approx$ Probability of failure occurring in a short interval $]t, t + \Delta]$, given "survival" or avoidance of event up to its start t, divided by the interval length.

This is equivalent to saying that over this short interval

$$\text{risk} \;\approx\; \text{rate} \;\times\; \text{length of interval}$$

or $$P(t < T \leq t + \Delta \mid T > t) \;\approx\; \lambda(t) \times \Delta.$$

Exponential or constant hazard model

Simplest probability model for time to event:
Exponential distribution, $\mathrm{Exp}(\lambda)$, in which

$$\lambda(t) = \lambda \text{ (constant)} \quad \Rightarrow \quad \Lambda(t) = -\log S(t) = \lambda t$$

Analysis of failure data of n individuals. For subject i let

$$y_i = \text{ time to event or time to censoring}, \quad Y = \sum y_i$$
$$d_i = \text{ indicator for observing the event}, \quad D = \sum d_i$$

$\mathrm{Exp}(\lambda)$ model $\Rightarrow$ **Likelihood function** of λ is

$$L(\lambda) = \prod_{i=1}^{n} \lambda(y_i)^{d_i} S(y_i) = \prod_{i=1}^{n} \lambda^{d_i} e^{-\lambda y_i} = \exp(D \log \lambda - \lambda Y)$$

Constant rate – Poisson model

This is actually equivalent to the *Poisson-likelihood*, *i.e.* likelihood of λ assuming that the number of cases D is distributed according to the **Poisson distribution** with expected value λY.

With randomly censored exponential times D is only approximately Poisson. This is sufficient, though, for likelihood-based (& asymptotic frequentist) inference.

Solving the *score equation*: $d \log L(\lambda)/d\lambda = 0$
$\rightarrow$ **maximum likelihood estimator** (MLE) of λ is

$$\widehat{\lambda} = \frac{D}{Y} = \frac{\text{number of cases}}{\text{total person-time}} = \text{ empirical incidence rate!}$$

Time to event – when to start the clock?

Incidence can be studied on various time scales, *e.g.*

- age (starting point = birth),
- exposure time (first exposure),
- follow-up time (entry to study),
- duration of disease (diagnosis).

Age is usully the strongest time-dependent determinant of health outcomes.

Age is also often correlated with duration of “chronic” exposure (*e.g.* years of smoking).

Therefore, adjustment for *current age* is needed rather than for *age at entry* to follow-up (like in clinical survival studies).

Age to event split into agebands

Let T = age at which outcome event occurs.
Parametric form of $\lambda(t)$, hazard by age – usually unknown.

Piecewise exponential model or **piecewise constant hazards' model** – an approximation for $\lambda(t)$:

$$\lambda(t) = \lambda_k, \qquad t \in \,]a_{k-1}, a_k], \quad \Delta_k = a_k - a_{k-1},$$

where cutpoints $0 = a_0 < a_1 < \cdots < a_K$ divide the age range into disjoint **agebands**, each with constant rate.

In chronic disease epidemiology agebands with $\Delta_k = 5$ years (0-4, 5-9, ..., 80-84) or 10 years are commonly used.

Age-specific incidence rates

For empirical estimation of rates we calculate in each ageband

$$D_k = \text{ number of cases occurring in ageband } k,$$
$$Y_k = \sum_{i=1}^{n} y_{ik} = \text{ total person-time in ageband } k,$$

where y_{ik} is the time slot that subject i spends in ageband k out of his/her whole **follow-up time** (from **entry** to **exit**).

ML estimators of $\lambda_1, \ldots, \lambda_K$: **age-specific incidence rates**

$$\widehat{\lambda}_k = I_k = D_k / Y_k, \quad k = 1, \ldots, K$$

based on log-likelihood $\log L = \sum_k (D_k \log \lambda_k - \lambda_k Y_k)$.

Cumulative rates & risks

In this model, the cumulative hazard and risk functions are

$$\Lambda(t) = \sum_{a_j < t} \lambda_j \Delta_j + \lambda_k (t - a_{k-1}), \qquad t \in \,]a_{k-1}, a_k]$$

$$F(t) = 1 - S(t) = 1 - \exp\{-\Lambda(t)\},$$

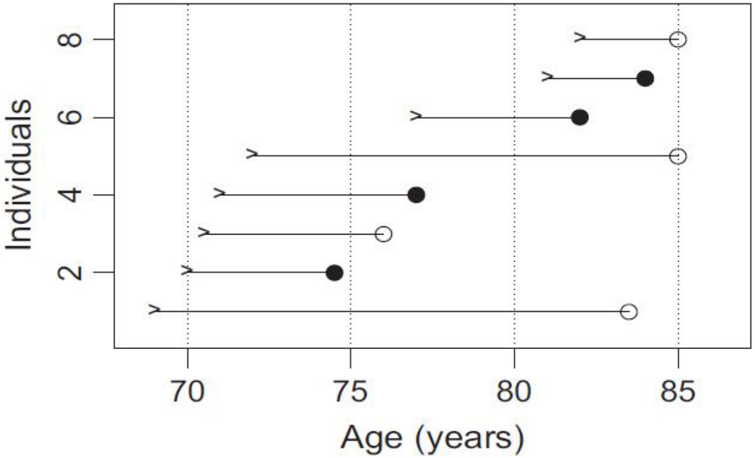
the latter assuming that no **competing risks** are present.

Estimation: Plug in empirical rates $\widehat{\lambda}_j = D_j / Y_j$ to get the cumulative rate C and incidence proportion R by t:

$$C = \widehat{\Lambda}(t) = \sum_{a_j < t} \widehat{\lambda}_j \Delta_j + \widehat{\lambda}_k (t - a_{k-1}), \qquad t \in \,]a_{k-1}, a_k]$$

$$R = \widehat{F}(t) = 1 - \widehat{S}(t) = 1 - \exp\{-\widehat{\Lambda}(t)\}$$

Example: Follow-up of a small geriatric cohort



No's of cases/p-years & rates (/100 y) in 5-y agebands:

$$1/21 = 4.8, \quad 1/16 = 6.2, \quad 2/16.5 = 12.1$$

Splitting follow-up by `Lexis()` in package `Epi`

Individual ages at entry and at exit, as well as outcomes are assigned into vectors and stored in a data frame `coh`:

```
ag.entry <- c(69, 70, 70.5, 71, 72, 76.9, 81, 81.9)
ag.exit <- c(83.5, 74.5, 76, 77, 85, 82, 84, 85)
event <- c(0,1,0,1,0,1,1,0) ; ind <- 1:8
coh <- data.frame( ind, ag.entry, ag.exit, event)
```

Function `Lexis()` specifies the time scale(s) to be considered. It creates an enriched data frame belonging to class `Lexis`.

```
library(Epi)
coh.L <- Lexis(entry = list(age = ag.entry),
               exit = list(age = ag.exit),
               exit.status = event, data = coh, id = ind)
```

Data frame of class Lexis

```
coh.L
```

```
   age lex.dur lex.Cst lex.Xst lex.id ind ag.entry ag.exit event
1 69.0    14.5       0       0      1   1     69.0    83.5     0
2 70.0     4.5       0       1      2   2     70.0    74.5     1
3 70.5     5.5       0       0      3   3     70.5    76.0     0
4 71.0     6.0       0       1      4   4     71.0    77.0     1
5 72.0    13.0       0       0      5   5     72.0    85.0     0
6 76.9     5.1       0       1      6   6     76.9    82.0     1
7 81.0     3.0       0       1      7   7     81.0    84.0     1
8 81.9     3.1       0       0      8   8     81.9    85.0     0
```

Interpretation of new columns

$$\begin{aligned} \texttt{age} &= \text{age at entry to follow-up},\\ \texttt{lex.dur} &= \text{duration of follow-up},\\ \texttt{lex.Cst} &= \text{current status at entry},\\ \texttt{lex.Xst} &= \text{status at exit from follow-up}. \end{aligned}$$

Splitting follow-up times by agebands

Function `splitLexis()` splits individual follow-up times into given agebands and expands the data frame.

```
coh.A <- splitLexis(coh.L,
       br = c(70,75,80,85), time.scale="age")
```

```
coh.A[-c(8:12),]
```

```
   lex.id  age lex.dur lex.Cst lex.Xst ind ag.entry ag.exit event
1       1 69.0     1.0       0       0   1     69.0    83.5     0
2       1 70.0     5.0       0       0   1     69.0    83.5     0
3       1 75.0     5.0       0       0   1     69.0    83.5     0
4       1 80.0     3.5       0       0   1     69.0    83.5     0
5       2 70.0     4.5       0       1   2     70.0    74.5     1
6       3 70.5     4.5       0       0   3     70.5    76.0     0
7       3 75.0     1.0       0       0   3     70.5    76.0     0
13      6 76.9     3.1       0       0   6     76.9    82.0     1
14      6 80.0     2.0       0       1   6     76.9    82.0     1
15      7 81.0     3.0       0       1   7     81.0    84.0     1
16      8 81.9     3.1       0       0   8     81.9    85.0     0
```

Splitted `Lexis` object

- Function `splitLexis()` expanded the original data frame such that for all cohort members one or more rows were created, one for each ageband into which a subject contributes person time.
- Ex: Subject 1 has been under follow-up in all agebands considered, but subjects 7 and 8 only in $80- < 85$ y.
- Function `timeBand()` converts variable age into factor `ageband`. Also, shorthand names for person-time slots and occurrence of outcome event are given.

```
coh.A$ageband <- timeBand(coh.A, "age", "factor")
coh.A$y_ik <- coh.A$lex.dur # person-time slot
coh.A$d_ik <- coh.A$lex.Xst # occurrence of outcome
```

Splitted Lexis object (cont'd)

```r
coh.A[, c(1,10:12)]
```

```
   lex.id   ageband y_ik d_ik
1       1 (-Inf,70]  1.0    0
2       1   (70,75]  5.0    0
3       1   (75,80]  5.0    0
4       1   (80,85]  3.5    0
5       2   (70,75]  4.5    1
6       3   (70,75]  4.5    0
7       3   (75,80]  1.0    0
8       4   (70,75]  4.0    0
9       4   (75,80]  2.0    1
10      5   (70,75]  3.0    0
11      5   (75,80]  5.0    0
12      5   (80,85]  5.0    0
13      6   (75,80]  3.1    0
14      6   (80,85]  2.0    1
15      7   (80,85]  3.0    1
16      8   (80,85]  3.1    0
```

`lex.id` = subject index in original data frame,

`ageband` = ageband and its limits,

`y_ik` = person-time slot spent in ageband

`d_ik` = indicator for event occurring in ageband.

Subject 1's follow-up time (14.5 y $= 1 + 5 + 5 + 3.5$ y) is split into 4 agebands, . . . , subject 8 contributes only to 1 ageband.

Tabulation of cases, rates etc. by ageband

Event indicators & person-time slots are summed over the rows of the split-expanded data frame in categories of `ageband`:

```
D <- with(coh.A, tapply(d_ik, ageband, sum))
Y <- with(coh.A, tapply(y_ik, ageband, sum))
```

Incidence rates (I), cumulative rates (C) and incidence proportions (R), the latter two by the end of each `ageband`:

```
I <- 100*D/Y; C <- cumsum((D/Y)*5); R <- 1-exp(-C)
tab <- round(cbind(D,Y,I,C,R),3)[2:4, ]
```

Age standardised incidence rate

Direct age standardisation

- e.g. to the World Standard population

```r
library(popEpi)
stdpop18
```

```
    agegroup world europe nordic
 1:      0-4 12000   8000   5900
 2:      5-9 10000   7000   6600
 3:    10-14  9000   7000   6200
 4:    15-19  9000   7000   5800
 5:    20-24  8000   7000   6100
 6:    25-29  8000   7000   6800
 7:    30-34  6000   7000   7300
 8:    35-39  6000   7000   7300
 9:    40-44  6000   7000   7000
10:    45-49  6000   7000   6900
11:    50-54  5000   7000   7400
12:    55-59  4000   6000   6100
13:    60-64  4000   5000   4800
14:    65-69  3000   4000   4100
15:    70-74  2000   3000   3900
16:    75-79  1000   2000   3500
17:    80-84   500   1000   2400
18:       85   500   1000   1900
```

Age standardised incidence rate (cont'd)

Weighted average of age-specific incidence rates I_k

$$\sum_k w_k I_k, \quad \text{where} \quad \sum_k w_k = 1$$

```r
tab <- cbind(tab,stdpop18[15:17,1:2])
tab$w <- tab$world/sum(tab$world)
tab
```

```
   D    Y      I     C     R agegroup world         w
1: 1 21.0  4.762 0.238 0.212    70-74  2000 0.5714286
2: 1 16.1  6.211 0.549 0.422    75-79  1000 0.2857143
3: 2 16.6 12.048 1.151 0.684    80-84   500 0.1428571
```

```r
#age-standardised incidence per 100 person years
sum(tab$w*tab$I)
```

```
[1] 6.216857
```

```r
#non-standardised incidence per 100 person years
100*sum(tab$D)/sum(tab$Y)
```

```
[1] 7.44879
```

Age standardised incidence rate (cont'd)

$$\mathrm{SE}\left[\log\left(\sum_k w_k I_k\right)\right] \approx \frac{1}{\sum_k w_k I_k}\sqrt{\sum_k \frac{w_k^2 d_k}{y_k^2}}$$

```
std.rate <- sum(tab$w*tab$D/tab$Y)
SE.std.rate <- sqrt( sum(tab$w^2*tab$D/tab$Y^2) )
exp( log(std.rate)+1.96*(1/std.rate)*SE.std.rate*c(-1,1) )
```

```
[1] 0.02082472 0.18559491
```

```
rate(tab,obs="D",pyrs="Y",adjust="agegroup",weights=tab$w)[,c(3,5:6)]
```

```
Adjusted rates (agegroup; 3) and 95% confidence intervals:

     rate.adj rate.adj.lo rate.adj.hi
1: 0.06216882  0.02082472   0.1855949
```

- Age-standardised incidence in age group 70-84 is:
  - 6.2 per 100 person years, 95% CI (2.1, 18.6)

Example: The Diet Study (see C&H)

A cohort of 337 men in three occupational groups in England, aged 30 to 67 y at entry, recruited in '50s and '60s, followed-up until mid '70s for incidence of CHD events.

Risk factors of interest, measured by dietary survey at entry.

$$
\begin{aligned}
\texttt{energy} &= \text{total energy intake (kcal/d)},\\
\texttt{energy.grp} &= \texttt{ energy} \text{ dichotomized:}\\
&\quad \text{1: “<=2750 KCals”, 2: “>2750 KCals”},\\
\texttt{fat} &= \text{fat intake (g/d)},\\
\texttt{fibre} &= \text{dietary fibre intake (g/d)},\\
&\quad \texttt{height, weight, bmi, } \textit{etc.}
\end{aligned}
$$

Important dates and outcome event

The data set `diet` in `Epi` contains three dates:

`dob` = date of **birth**,
`doe` = date of **entry** into follow-up,
`dox` = date of **exit**, end of follow-up.

These are given in format `yyyy-mm-dd` but implicitly stored as *number of days since 1.1.1970*.

In addition, the outcome event is represented by

`chd` = indicator for **status** at exit:
1 = CHD event occurred, 0 = censored.

Data diet: creating a Lexis object

First convert all dates into fractional calendar years using function `cal.yr()` in `Epi`

```
data(diet)
diet <- transform(diet, doe = cal.yr(doe),
                  dox = cal.yr(dox), dob = cal.yr(dob) )
```

Convert the data frame into a `Lexis` object.

```
dietL <- Lexis( entry = list(age = doe-dob),
                exit = list(age = dox-dob),
                exit.status = chd, data = diet )
```

In the nexty step the `Lexis` object is splitted according to 3 agebands (y): $30- < 50$, $50- < 60$, $60- < 70$

Splitting the Lexis object into agebands

```
dietA <- splitLexis(dietL, br = c(30,50,60,70),
                    time.scale = "age")
dietA$ageband <- timeBand(dietA, "age", "factor")
dietA$y_ik <- dietA$lex.dur ; dietA$d_ik <- dietA$lex.Xst
```

```
dietA[1:8,c("id","dob","doe","dox","y","chd","energy.grp","ageband",
            "age","y_ik","d_ik")]
```

```
   id    dob    doe    dox    y chd   energy.grp ageband  age y_ik d_ik
1 102 1939.2 1976.0 1986.9 10.9   0 <=2750 KCals (30,50] 36.9 10.9    0
2  59 1912.5 1973.5 1982.5  9.0   0 <=2750 KCals (60,70] 61.0  9.0    0
3 126 1920.0 1970.2 1984.2 14.0   1 <=2750 KCals (50,60] 50.2  9.8    0
4 126 1920.0 1970.2 1984.2 14.0   1 <=2750 KCals (60,70] 60.0  4.2    1
5  16 1906.7 1969.4 1970.0  0.6   1 <=2750 KCals (60,70] 62.7  0.6    1
6 247 1918.5 1968.2 1979.5 11.3   1 <=2750 KCals (30,50] 49.7  0.3    0
7 247 1918.5 1968.2 1979.5 11.3   1 <=2750 KCals (50,60] 50.0 10.0    0
8 247 1918.5 1968.2 1979.5 11.3   1 <=2750 KCals (60,70] 60.0  1.0    1
```

Properties of the original data frame and the expanded object:

```
> str(diet)
'data.frame':   337 obs. of  17 variables:
> str(dietA)
Classes Lexis and data.frame  729 obs. of 25 variables
```

Relevelling of energy.grp and some tabulations

The `energy.grp` variable is relevelled such that “high energy” is taken as the reference or “unexposed” category and “low energy” as the “exposed” one.

```
dietA$eg2 <- Relevel( dietA$energy.grp,
           ref = ">2750 KCals" )
```

Tabulation of cases, person-years and rates:

```
tab.ae <- stat.table( list( ageband, eg2),
      list( D = sum(d_ik), Y = sum(y_ik),
            I = ratio(d_ik, y_ik, 1000) ),
       margin = T, data = dietA )
print(tab.ae, digits= c(sum=0, ratio=1))
```

Rates by ageband and energy intake

```
-----------------------------------
           -----------eg2----------
ageband          >2750  <=2750  Total
                 KCals   KCals
-----------------------------------
(-Inf,30]           NA      NA     NA
 ...
(30,50]              4       2      6
                   622     381   1003
                   6.4     5.2    6.0
(50,60]              6      12     18
                  1128     979   2107
                   5.3    12.3    8.5
(60,70]              8      14     22
                   794     699   1493
                  10.1    20.0   14.7
(70,Inf]            NA      NA     NA
...
Total               18      28     46
                  2544    2059   4604
                   7.1    13.6   10.0
-----------------------------------
```

Crude rate ratio

```
tab.ae[3, 6, 2]/
  tab.ae[ 3, 6, 1]
```

```
[1] 1.922497
```

Rate ratios by ageband:

```
IRs <- tab.ae[3, 2:4, 2]/
  tab.ae[3, 2:4, 1]
round(IRs,2)
```

```
(30,50] (50,60] (60,70]
   0.82    2.30    1.99
```

- Low intake risky?
- No effect in young?

Poisson model on age and exposure

Let D_{kj}, Y_{kj}, and I_{kj} be cases, p-years & rate in ageband k & exposure category j (1="unexposed", 2="exposed").
Piecewise Exp-model in both exposure categories assumed:

$$\lambda_{kj} = \text{ theoretical rate in cell } kj.$$

Theoretical rate ratio $\rho_k = \lambda_{k2}/\lambda_{k1}$,
comparing exposed *vs.* unexposed.

(a) What are the "true" values of ρ_k?

(b) Can we assume $\rho_k = \rho$, same rate ratio in all agebands?

(c) What is the value of the common rate ratio ρ?

Poisson model (cont'd)

Assuming common rate ratio the true rates are modelled

$$\log \lambda_{kj} = \alpha_k + \beta_j = \sum_{k=1}^{K} \alpha_k A_k + \sum_{j=1}^{2} \beta_j X_j,$$

where A_k and X_j are indicator $(1/0)$ variables for level k of ageband and level j of exposure. In exponential form

$$\lambda_{kj} = \exp(\alpha_k + \beta_j) = e^{\alpha_k} e^{\beta_j}.$$

Set $\beta_1 = 0$ ("unexposed" as reference) $\Rightarrow$ Interpretation:

$$\alpha_k = \log(\lambda_{k1}) = \text{ log-rate of unexposed in ageband } k$$
$$\beta_2 = \log(\lambda_{k2}/\lambda_{k1}) = \log(\rho) = \text{ log-common rate ratio}$$

Fitting the Poisson model

Use function `glm()` on the expanded data frame:

```
m.ea <- glm( d_ik/y_ik ~ ageband + eg2,
             fam = poisson, w = y_ik, data = dietA )
```

```
round(ci.lin(m.ea, Exp=T)[ , -(3:4)], 4 )
```

```
                 Estimate StdErr exp(Est.)   2.5%  97.5%
(Intercept)       -5.4033 0.4390    0.0045 0.0019 0.0106
ageband(50,60]     0.3027 0.4721    1.3535 0.5366 3.4145
ageband(60,70]     0.8456 0.4613    2.3294 0.9431 5.7535
eg2<=2750 KCals    0.6233 0.3027    1.8651 1.0306 3.3753
```

The estimated rate ratio for “low” vs. “high” energy consumption, adjusted for age, is thus 1.87 [1.03 to 3.38], only slightly lower than the unadjusted one 1.92 [1.06 to 3.47].

Concluding remarks

- Modelling could continue from this to include other confounders, continuous covariates, interactions, *etc.*
- Agebands may well be much narrower than in our example. With infinitely narrow bands Poisson regression equals the famous Cox model.
- Splitting by many time scales (*e.g.* age, calendar time, time since first exposure, *etc.*) simultaneously and the corresponding data frame expansion is straightforward using these tools. More about this in the next lecture.