Survival time and covariate data

Given covariates Z_i, denote $S_i(t)$ by $S_i(t; Z_i, \theta)$ and $\lambda_i(t)$ and $\Lambda_i(t)$ as $\lambda_i(t; Z_i, \theta)$ and $\Lambda_i(t; Z_i, \theta)$, where θ is the parameters associating the covariates to the event or failure.

- Noninformative censoring - (given Z_i) the random censoring time C_i is assumed independent of parameters of interest θ.
- Given covariates Z_i and the parameters of interest θ, the likelihood is

$$L_i(\theta) \propto \lambda_i(t_i; Z_i, \theta)^{d_i} S_i(t_i; Z_i, \theta).$$

Proportional hazards model (1)

T_0 - failure time of nonsmokers with rate $\lambda_0(t)$

T_1 - failure time of smokers with rate $\lambda_1(t)$

If we believe that the death rate among smokers is higher/different compared to the nonsmokers then $\lambda_1(t)$ can be expressed as a multiple of $\lambda_0(t)$.

Proportional hazards model (2)

Let $\theta > 0$ be the multiplicative factor

$$\begin{aligned} \lambda_1(t;\theta) &= \theta\lambda_0(t), \ \ \forall t > 0, \\ \lambda(t;Z,\theta) &= \theta^Z \lambda_0(t), \\ &= \lambda_0(t)\exp\{\log(\theta)Z\}, \\ \log(\lambda(t;Z,\theta)) &= \log(\lambda_0(t)) + \beta Z, \end{aligned}$$

where $Z = 1$ if smoker and zero otherwise and $\beta = \log(\theta)$.

Proportional hazards model (3)

The hazard ratio between two individuals at any given time

$$\frac{\lambda_1(t; Z_1, \theta)}{\lambda_2(t; Z_2, \theta)} = \exp\{\beta(Z_1 - Z_2)\}$$

is constant over time!

Proportional hazards model (4)

In general, a proportional hazards family of regression models is written as

$$\lambda_i(t_i; Z_i, \theta) \quad = \quad \lambda_0(t_i) r(Z_i; \beta).$$

Here $\theta = (\lambda_0(t), t > 0, \beta)$, $r(Z; \beta)$ is a positive valued function and $\lambda_0(t)$ is a baseline hazard rate.

Part I: Parametric regression models

Full parametric specification of $\lambda_0(t)$ and $r(Z;\beta)$

Commonly used specification for $r(Z;\beta)$ is $\exp\{\beta Z\}$.

Part II: Semiparametric models

- Semiparametric survival models
  Full parametric specification of the relative risk function $r(Z;\beta)$ but not of the baseline hazard $\lambda_0(t_i)$
  Commonly used specification for $r(Z;\beta)$ is $\exp\{\beta Z\}$.
- Cox in 1972 proposed that the conditional hzard be modelled as

$$\lambda_i(t_i; Z_i, \theta) = \lambda_0(t_i)\exp(\beta' Z_i).$$

Reference:
Cox, D. R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

Inferential problems

- Estimation of the regression parameters β
- Estimation of the baseline hazard $\lambda_0(t)$
- Estimation of functions of β and $\lambda_0(t)$ $(S(t; Z, \theta))$
- Hypothesis testing for the significant effects of the covariates

Assumptions

1. The ratio of the hazards of two individuals is the same at all times.
2. The covariates act multiplicatively on the hazard.
3. Conditionally on Z_i and Z_j, the failure times of the individuals i and j are independent.

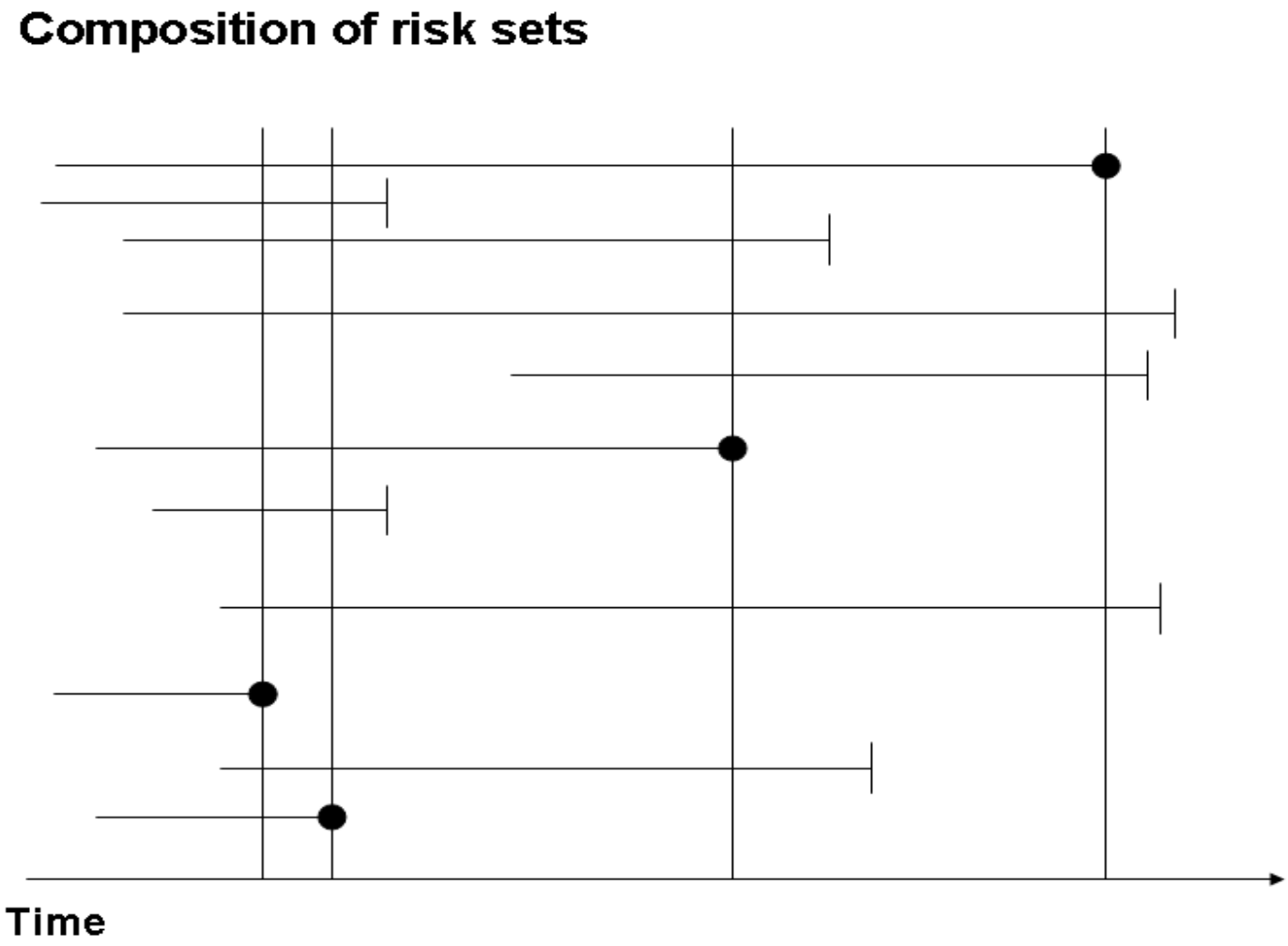
Estimation of β

Recall: $Y_i(t) = I(T_i \geq t)$ is an indicator whether an individual i is at risk at time t or not

If individuals enter the study at different times then the entry time should be before the time t and the exit time should be after time t for an individual to be at risk at time t.

Define $R(t) = \{j : Y_j(t) = 1\} = \{j : T_j \geq t\}$ a risk set of all the individuals who are at risk at time t

Risk set



Conditional probability

- Let $T_1, \ldots T_k$ be the failure times of k failures observed among n individuals.
- Note that $(n-k)$ observations are censored.
- The conditional probability that individual i fails at time T_i given that exactly one individual failed at T_i and the covariates and the risk set at T_i can be written as

$$\begin{aligned} L_i(\beta) &= \frac{\lambda(T_i; Z_i, \theta)}{\sum_{j \in R(T_i)} \lambda(T_i; Z_j, \theta)} \\ &= \frac{\exp(\beta' Z_i)}{\sum_{j \in R(T_i)} \exp(\beta' Z_j)} \end{aligned}$$

Partial likelihood (1)

β is now estimated using the likelihood which is the product of $L_i(\beta)$ over all the failures

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} (L_i(\beta))^{d_i} \\ &= \prod_{i=1}^{k} \frac{\exp(\beta' Z_i)}{\sum_{j \in R(T_i)} \exp(\beta' Z_j)} \end{aligned}$$

Partial likelihood (2)

- Does not depend on $\lambda_0(t)$!
- The denominator is the contribution by those who are at risk at time T_i

Why partial? (1)

Let $T_{(1)}, \ldots, T_{(k)}$ denote the observed and ordered failure times (distinct) where $k \leq n$.

Let U_j represents information on censoring in the interval $[T_{(j-1)}, T_{(j)})$ plus the fact that one individual fails at $T_{(j)}$, and let (j) specify the individual failing at $T_{(j)}$.

The information contained in the observed data $\{(T_i, d_i), i = 1, \ldots, n\}$ is the same as the information contained in $\{U_1, (1), \ldots, U_k, (k)\}$.

Why partial? (2)

The part of the likelihood based on $\{(1), \ldots, (k)\}$ in the sequence $\{U_1, (1), \ldots, U_k, (k)\}$ is

$$\prod_{j=1}^{k} P((j) \mid U_j, (U_l, (l)), l = 1, \ldots, j-1)$$

$$= \prod_{j=1}^{k} \frac{\exp(\beta' Z_{(j)})}{\sum_{l \in R(T_{(j)})} \exp(\beta' Z_{(l)})}$$

Estimation of baseline hazard (1)

Recall: the contribution of individual i to the log-likelihood is the sum of contributions for each time band and it has the Poisson form:

$$L(\theta) = \sum_{i,t} d_i^t \log(\lambda_i(t; Z_i, \theta)) - y_i^t \lambda_i(t; Z_i, \theta)$$

where y_i^t is the observation time in time-band t and d_i^t indicates whether failure occured $(= 1)$ or not $(= 0)$.

$$\begin{aligned} L(\theta) &= \sum_{i,t} [d_i^t \log(\lambda_0(t) \exp(\beta' Z_i)) - y_i^t \lambda_0(t) \exp(\beta' Z_i)] \\ &= \sum_{i,t} [d_i^t \log(\lambda_0(t)) + d_i^t \beta' Z_i - y_i^t \lambda_0(t) \exp(\beta' Z_i)] \end{aligned}$$

Estimation of baseline hazard (2)

Taking derivative w.r.t. $\lambda_0(t)$ and equating it to zero, we obtain the most likely value of $\lambda_0(t)$

$$\lambda_0(t) = \frac{\sum_i d_i^t}{\sum_i y_i^t \exp(\beta' Z_i)}$$

By substituting the most likely value of the baseline hazard in the log-likelihood results into *profile* likelihood for β.

Estimation of baseline hazard (3)

By substituting the estimates of β, the baseline hazard is estimated.

- Generalise the above idea and dividing the time scale into clicks which contain no more than one event.
- y_i^t can be understood as either 1 or 0 depending on whether the individual i was observed at click t - at risk indicator
- If the duration of the time band is h then the observation time becomes hy_i^t

Aalen-Breslow estimator

- The cumulative baseline hazard rate is either zero or

$$\frac{1}{\sum_i y_i^t \exp(\beta' Z_i)}$$

  when failure occurs.
- It is estimated by a step function which jumps at the observed failure times.
- The height of the jump at each failure is

$$\frac{1}{\sum_{\text{Risk set}} \exp(\beta' Z_i)}$$

  rather than 1/(Number of individuals at risk).

Covariate effects and their significance

- Follow-up for certain type of cancer
- Age group 25-74 at the baseline survey
- Time scale - time in the study
- Covariates measure at the baseline (say we are interested in 4 of them)
- To examine whether the cancer rates are different in different age groups with regard to the effects of the covariates

Comparison between age-groups (1)

- Using the age at the time of the baseline examination, define four groups $[-, 45)$, $[45, 55)$, $[55, 65)$, and $[65, -]$
- The hazard model for the age-group s can be specified as

$$\lambda_{is}(t_i; Z_i, \theta) = \lambda_s(t_i) \exp(\beta_s' Z_i),$$

  where $\beta_s' = (\beta_{s1}, \ldots, \beta_{s4})$, and $s = 1, \ldots, 4$.
- The above model is fitted for each age-group to evaluate covariate effects.

Comparison between age-groups (2)

- A global hypothesis of equality of covariate effects can be stated as

$$\beta_{1l} = \beta_{2l} = \beta_{3l} = \beta_{4l} = \beta_l, \forall l = 1, 2, 3, 4.$$

- Under this hypothesis, the hazard model reduces to

$$\lambda_{is}(t_i; Z_i, \theta) = \lambda_s(t_i) \exp(\beta' Z_i),$$

  where $\beta' = (\beta_1, \beta_2, \beta_3, \beta_4)$.
- The hypothesis is tested using the likelihood ratio test using chi-square distribution with $16 - 4 = 12$ degrees of freedom.

Comparison between age-groups (3)

- If this hypothesis is rejected then the hypothesis of equal covariate effects for each covariate is tested separately.
- The model under the equal effect of Z_2 across the age-groups can be specified as

$$\lambda_{is}(t_i; Z_i, \theta) = \lambda_s(t_i) \exp(\beta'_{s2} Z_i),$$

  where $\beta'_{s2} = (\beta_{s1}, \beta_2, \beta_{s3}, \beta_{s4})$.
- The likelihood ratio statistic has a chi-square distribution with $16 - 13 = 3$ degrees of freedom.

Comarison between age-groups (4)

- Let Z_1 be the age at baseline
- Note that the hypothesis of equality of age effects across the age-groups corresponds to the linearity assumption for age in the proportional hazards model.
- The model under the equal effects of (Z_1, Z_2) across the age-groups can be specified as

$$\lambda_{is}(t_i; Z_i, \theta) = \lambda_s(t_i) \exp(\beta'_{s12} Z_i),$$

  where $\beta'_{s12} = (\beta_1, \beta_2, \beta_{s3}, \beta_{s4})$.
- The likelihood ratio statistic has a chi-square distribution with $16 - 10 = 6$ degrees of freedom.

Checking proportionality

Again, the loglog plot can be used, because

$$\begin{aligned} \log(-\log(S_i(t; Z_i, \theta))) &= log(\Lambda(t; Z_i, \theta)) \\ &= \log(\Lambda_0(t) \exp(\beta Z_i)) \\ &= \log(\Lambda_0(t)) + \beta Z_i \end{aligned}$$

If the model is apppropriate, the curves for values Z_1 and Z_2 are parallel and their (vertical) distance is $\beta(Z_2 - Z_1)$.

Also these can be plotted from the Kaplan-Meier estimates. Note that the curves need not be straight lines.

R functions

- libraray(survival)
- *coxph* - Cox proportional hazards model

```
coxph(formula = Surv(time, status) ~ Z1 + Z2 , data=trial)
```

- *extractAIC*(object) - Model comparisons using information criterion
- *anova*(object1, object2) - Likelihood ratio test for nested model where object1 corresponds to the results from the model with less parameters and object2 corresponds to the results from the model which includes model 1.

Part III: Nonparametric models

A toy example: Entry end exit dates for the cohort of four subjects

Subject	Born	Entry	Exit	Outcome
1	1904	1943	1952	Lost
2	1924	1948	1955	Failure
3	1914	1945	1961	Study ends
4	1920	1948	1956	Unrelated death

Exponential regression

- Assume that the event time is exponential in each band with rate λ_{jk} in the band (j, k).
- Likelihood for the band (j, k) is

$$\begin{aligned} & \prod_{i=1}^{n} \{\lambda_{jk}\}^{\delta_i^{jk}} \exp\{-\lambda_{jk} Y_i^{jk}\} \\ = & \{\lambda_{jk}\}^{\sum_i \delta_i^{jk}} \exp\{-\lambda_{jk} \sum_i Y_i^{jk}\} \\ = & \{\lambda_{jk}\}^{D_{jk}} \exp\{-\lambda_{jk} Y_{jk}\} \end{aligned}$$

where $D_{jk} = \sum_i \delta_i^{jk}$ and $Y_{jk} = \sum_i Y_i^{jk}$

Poisson regression

Does the likelihood for the band (j, k) look familiar?

$$D_{jk} = \sum_i \delta_i^{jk} \sim Poisson(\lambda_{jk} Y_{jk})$$

where
D_{jk} is the number of events in the band (j, k) and
$Y_{jk} = \sum_i Y_i^{jk}$ is the total person-year observed in the band (j, k).

Nonparametric?

- Conceptual paradox: often nonparametric means more parameters.
- In traditional Cox regression analysis, baseline hazard function on one time scale is modelled as a nonparametric function while the multiplicative regression part is modelled parametrically.

Piecewise constant model (1)

- Rather than trying to replicate the Cox regression approach, let's look at survival analysis from the viewpoint of the piecewise constant model / Poisson regression.
- In this approach, piecewise constant hazard rate is considered as a reasonable approximation for the underlying "true" hazard function.
- Here time scales do not have any special status compared to other covariates; this allows the use of several time scales and makes it easy to relax parametric assumptions where it is most needed.

Piecewise constant model (2)

- Cox model can be fitted using standard Poisson regression technique by splitting the data finely into time bands or bins and specifying the model by rate parameter for each band/bin.
- Fixed covariates of an individual are carried over to all bins while time-dependent covariates are computed for each bin.

Piecewise constant model (3)

- For each bin, the person-year (amount of time spend in the bin) and failure status are coded for each individual. In the absence of individual-level covariate data, the total person-year and the total number of failures per bin are coded.
- The number of deaths in each bin is then independent observation from the Poisson distribution with mean equal to the product of the person-year and the rate for that bin.

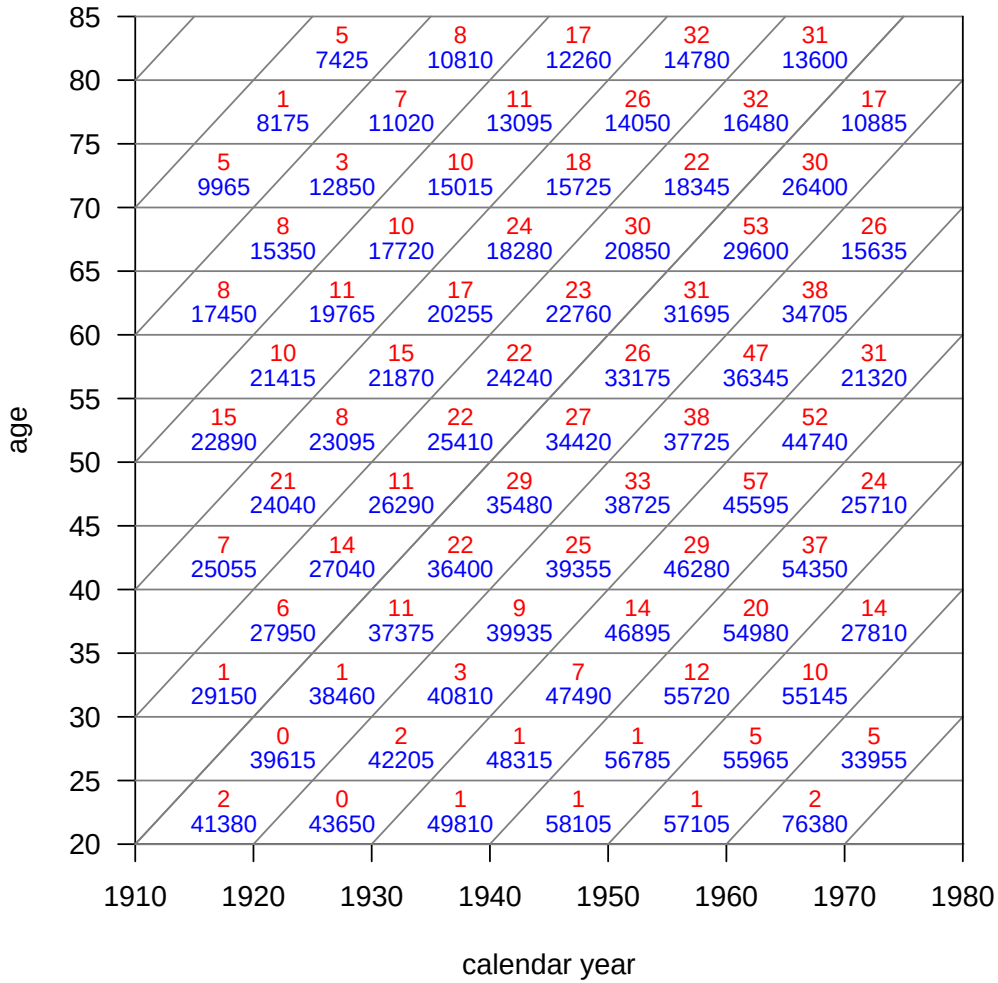
Piecewise constant model (4)

- Poisson modelling of split data can be used for everything that can be done using Cox model.
- Easy to use more than one time-scale.
- A natural way to handle time-varying covariates.

Example data

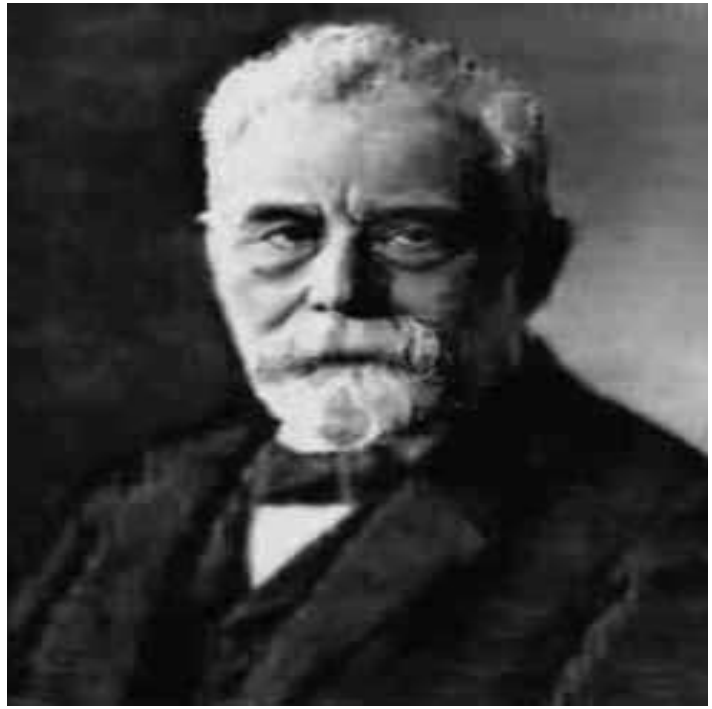
- Let's illustrate the use of piecewise constant model with Icelandic breast cancer data studied by Breslow and Clayton (1993).
- Here population level data on the number of incident cases and person years are grouped into 10-year birth cohorts from 1840-1849 to 1940-1949 and 5-year age groups from 20-24 to 80-84.
- Reference: N. E. Breslow and D. G. Clayton. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **88**, No. 421 (Mar., 1993), pp. 9-25.

Data in a Lexis diagram



Wilhelm Lexis

Born: 17 July 1837 in Eschweiler (near Aachen), Germany
Died: 25 Oct 1914 in Gttingen, Germany



http://www-history.mcs.st-andrews.ac.uk/Biographies/Lexis.html

Likelihood

- Likelihood for individual level data is of the familiar Poisson form, with each bin (j, k) with its own likelihood contribution:

$$\prod_{i=1}^{n} \prod_{j=1}^{13} \prod_{k=1}^{11} \lambda_{ijk}^{d_{ijk}} \exp\{-\lambda_{ijk} y_{ijk}\}.$$

- Here λ_{ijk} is the constant hazard rate, d_{ijk} is an event indicator and y_{ijk} is the follow-up time in years in the bin (j, k) for individual i.

Likelihood (2)

- With the absence of other individual level covariate data, we set $\lambda_{ijk} = \lambda_{jk}$ and the likelihood can be written as

$$\prod_{j=1}^{13} \prod_{k=1}^{11} \lambda_{jk}^{\sum_{i=1}^{n} d_{ijk}} \exp\left\{ -\lambda_{jk} \sum_{i=1}^{n} y_{ijk} \right\},$$

  which is the representation for population level data.
- Now $\sum_{i=1}^{n} d_{ijk} \sim \text{Poisson}\left(\lambda_{jk} \sum_{i=1}^{n} y_{ijk}\right)$.
- Note that not all possible (j, k) combinations are present in the data (only 77 of 143).

Model parameterisation

- Suppose we are interested in age and birth cohort trends in the hazard rate. The simplest way to parameterise the model for this purpose is to assume that the two trends are independent. This corresponds to

$$\log(\lambda_{jk}) = \alpha_j + \beta_k,$$

  where $\alpha_j, j = 1, \ldots, 13$ are the age parameters and $\beta_k, k = 1, \ldots, 11$ the birth cohort parameters.
- Some of the birth cohorts included very little data, so it may not be sensible to estimate independent parameters for these.

Autoregressive smoothing

- In some cases it may be reasonable to assume that hazard rate is varying smoothly over time.
- Second order normal random walk model for the birth cohort parameters:

$$\beta_k \mid \beta_{k-1}, \beta_{k-2}, \ldots \sim N(2\beta_{k-1} - \beta_{k-2}, \phi).$$

- Precision parameter ϕ describes the variation of the β_ks. Usually a gamma distribution is assumed for these kind of parameters.

How to carry out the analysis in R? (1)

- R Epi package required (http://staff.pubhealth.ku.dk/ bxc/Epi/).
- Install the package using zip file given in the bin directory. Go to Packages menu in R and then click Install package(s) from the local zip files.
- The data set can be found from the R Epi package:

```
library(Epi)
?diet
data(diet)
str(diet)
Lexis.diagram()
Lexis.lines(entry.date = diet$doe,
    exit.date = diet$dox,
    birth.date = diet$dob,
    fail = diet$chd)
```

How to carry out the analysis in R? (2)

- To adjust the axes of the Lexis diagram, use (for example):

```
Lexis.diagram(age = c(30, 75),
    date = c(1950, 1995))
```

- To convert dates into number of days

```
diet$bdat<-as.numeric(diet$dob)
diet$exitd<-as.numeric(diet$dox)
diet$entryd<-as.numeric(diet$doe)
```

  The number of days since 1/1/1970, which is a negative number for dates prior to 1/1/1970.

- Compute the age at the time of entry and exit in years

```
diet$eage<-(diet$entryd-diet$bdat)/365.25
diet$xage<-(diet$exitd-diet$bdat)/365.25
```

How to carry out the analysis in R? (3)

```
diet$bt.yr <- cal.yr( diet$dob, format="%d/%m/%Y")
diet$bdays <- Days <- (diet$bt.yr-1800)*365.25

diet$en.yr <- cal.yr( diet$doe, format="%d/%m/%Y")
diet$endays <- Days <- (diet$en.yr-1800)*365.25

diet$ex.yr <- cal.yr( diet$dox, format="%d/%m/%Y")
diet$exdays <- Days <- (diet$ex.yr-1800)*365.25

diet$eage<-(diet$endays-diet$bdays)/365.25
diet$xage<-(diet$exdays-diet$bdays)/365.25
```

Splitting the follow-up

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

# spliting follow-up into two time scales
splitdiet1 = splitLexis(dietL, breaks = seq(1966,1990,5), time.scale="per")
splitdiet2 = splitLexis(splitdiet1,breaks = seq(30,70,5), time.scale="age")
```

More details at

http://bendixcarstensen.com/AdvCoh/papers/Plummer.2011.pdf