Nonparametric regression models

2019.04.15

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.

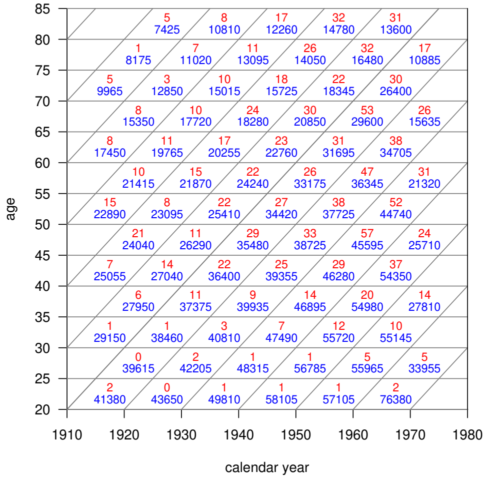
Piecewise constant model (4)

- 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.

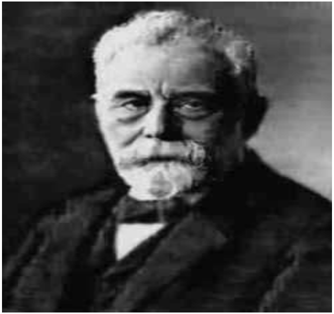
Data in a Lexis diagram



Wilhelm Lexis

Born: 17 July 1837 in Eschweiler (near Aachen), Germany

Died: 25 Oct 1914 in Gttingen, Germany



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 $\sum_{i=1}^{n} d_{ijk} \sim \mathrm{Poisson}\left(\lambda_{jk} \sum_{i=1}^{n} y_{ijk}\right)$.
- 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.