Analysis of Survival Data

Net Survival (2014)

1 Introduction

1.1 Context

A *net survival* analysis is a time-to-event analysis which has been adjusted for a competing, secondary event.

The most common application is in time-to-death studies, where individuals in such a study are at risk of dying of the disease of interest but are also at risk of dying from other causes. It is often impracticable to ascertain cause of death but we can work with the deaths observed in the study population and the deaths that would have been expected in a comparator population.

A net survival analysis is very like a *competing risks* analysis in that the individual is at risk of two different events, both of which are absorbing. Net survival also borrows ideas from *frailty*: in particular the at risk population changes character with time as high risk individuals are preferentially removed.

1.2 Statistical Model

Each individual is at risk of three events:

1. the event of interest A.
2. a competing event of no interest B.
3. censoring C.

If we cannot distinguish between events A and B we have a *net survival* analysis. (If we can, we have a *competing risks* analysis.)

We imagine that there are three times-to-event – T_A, T_B, T_C – which are jointly independent given a set of baseline covariates Z. The independence of T_C with the other two results in censoring being *uninformative*; the independence of T_A and T_B is an unverifiable assumption. We observe $X = T_{A\cup B\cup C} = \min(T_A, T_B, T_C)$ and we can only distinguish between $A \cup B$ and C. We will use J as a shorthand for the composite event 'A or B' so that $T_J = T_{A\cup B} = \min(T_A, T_B)$.

1.3 Medical Application

A typical net survival analyis in cancer research would have

A = death from the cancer of interest;

B = death from other causes, the Background mortality;

J = death from any cause;

C = Censoring, usually a combination of lost to follow up and the end of the study.

The baseline covariates are normally age at diagnosis, calendar year of diagnosis, gender and geographical location (the 'demographic' covariates).

In practice, it is notoriously difficult to determine cause of death. Net survival is an attempt to characterise the excess mortality due to a disease without knowing the cause of death It is very important to allow for background mortality when comparing survival for a particular disease across different populations.

2 Lecture

2.1 *Competing risks*

The hazard $\mathrm{h}_J(t)$ for the event death-from-any-cause ($A \cup B$) is given by the sum of hazards for the individual events:

$$\mathrm{h}_J(t) = \mathrm{h}_A(t) + \mathrm{h}_B(t)$$

where $\mathrm{h}_A(t)$ is defined as:

$$\mathrm{h}_A(t) = \lim_{\Delta \downarrow 0} \left\{ \frac{1}{\Delta} \mathbb{P}\left[t < T_A \leq t + \Delta | t < T_J \right] \right\},$$

with the corresponding definition for $\mathrm{h}_B(t)$, and the same relationship applies for the integrated hazards:

$$\mathrm{H}_J(t) = \mathrm{H}_A(t) + \mathrm{H}_B(t).$$

$\mathrm{h}_A(t)$, $\mathrm{h}_B(t)$, $\mathrm{h}_J(t)$ are often called the *excess, background* and *joint* hazards respectively.

2.2 Obtaining an estimate for $\mathrm{H}_A(t)$

$\mathrm{H}_J(t)$ is the integrated hazard for death that is actually observed in the *diseased* population, $\mathrm{H}_B(t)$ is the integrated hazard for deaths other than due to the disease of interest. Government life tables provide the hazard rates for all deaths broken down by the demographic covariates. The hazard for death from a particular disease in the *general* population is usually small compared to $\mathrm{H}_B(t)$. We therefore use the government published hazard rates for all deaths

as a good substitute for the hazard rates for deaths from other causes than the disease of interest.

Naïvely, an estimate for $\mathrm{H}_A(t)$ can be written in terms of known quantities:

$$\hat{\mathrm{H}}_A(t) = \hat{\mathrm{H}}_J(t) - \mathrm{H}_B(t). \tag{1}$$

(Note that there is no 'hat' above H_B as the background mortality is assumed to be known exactly from the published tables.)

An equivalent relationship can be written in terms of survivor functions:

$$\hat{\mathrm{F}}_A(t) = \frac{\hat{\mathrm{F}}_J(t)}{\mathrm{F}_B(t)} \tag{2}$$

where $\hat{\mathrm{F}}_A(t)$ is often referred to as the *relative* survivor function.

These relationships work at an individual level but there is a difficulty at the population level: individuals will vary with respect to their demographic covariates, the background integrated hazards $\mathrm{H}_B(t)$ at least will vary from individual to individual and great care has to be taken when combining individual hazards to give an appropriate population hazard.

Exercise Verify that (1) implies (2).

2.3 *Ederer II* estimate of $\mathrm{H}_A(t)$

The experience of the ith individual can be written in counting process notation as:

$$\mathrm{d}N_i(t) = Y_i(t)\mathrm{dH}_J^i(t) + \mathrm{d}M_i(t) \tag{3}$$

where $N_i(t)$ is the observed event indicator, $Y_i(t)$ is the at-risk indicator and $M_i(t)$ is the associated Martingale. We decompose $\mathrm{H}_J^i(t)$ into $\mathrm{H}_A^i(t) + \mathrm{H}_B^i(t)$ and use the method of moments to replace the integrated excess hazard by its estimator and $\mathrm{d}M_i(t)$ by zero:

$$\mathrm{d}N_i(t) = Y_i(t)\left(\mathrm{d}\hat{\mathrm{H}}_A^i(t) + \mathrm{dH}_B^i(t)\right)$$

The Ederer II method assumes the excess hazard is common to all indivduals. Replacing $\hat{\mathrm{H}}_A^i(t)$ by $\hat{\mathrm{H}}_A(t)$, summing over the n individuals, re-arranging and integrating gives:

$$\hat{\mathrm{H}}_A(t) = \int_0^t \frac{\sum_{i=1}^n \mathrm{d}N_i(t')}{\sum_{i=1}^n Y_i(t')} - \int_0^t \frac{\sum_{i=1}^n Y_i(t')\mathrm{dH}_B^i(t')}{\sum_{i=1}^n Y_i(t')}$$

or

$$\hat{\mathrm{H}}_A(t) = \int_0^t \frac{\mathrm{d}N_+(t')}{Y_+(t')} - \int_0^t \frac{\sum_{i=1}^n Y_i(t')\mathrm{dH}_B^i(t')}{Y_+(t')}. \tag{4}$$

Equation (4) tells us that the estimate of the integrated excess hazard is obtained by offsetting the Nelson-Aalen estimator for the integrated joint hazard by the integral of the overall background hazard, where the overall background hazard is the mean hazard of individuals still in the risk set.

2.4 *Pohar-Perme* estimate of $\mathrm{H}_A(t)$

Generally, the integrated excess hazard function $\mathrm{H}_A^i(t)$ depends on the individual i: in particular the excess hazard may depend on the demographic covariates. The Ederer II estimate $\hat{\mathrm{H}}_A(t)$ therefore is some sort of weighted average of the individual excess hazards. The question arises: is it an appropriately weighted hazard? We can find out by finding the expectation of (4) in terms of the individual hazards.

First of all we replace $\mathrm{d}N_+(t')$ by the RHS of (3), decomposing $\mathrm{H}_J^i(t)$ and writing out sums in full:

$$\hat{\mathrm{H}}_A(t) = \int_0^t \frac{\sum_{i=1}^n \left\{ Y_i(t') \left[\mathrm{dH}_A^i(t') + \mathrm{dH}_B^i(t') \right] + \mathrm{d}M_i(t') \right\}}{\sum_{i=1}^n Y_i(t')} - \int_0^t \frac{\sum_{i=1}^n Y_i(t') \mathrm{dH}_B^i(t')}{\sum_{i=1}^n Y_i(t')}.$$

We note, first of all, that the integrals in $\mathrm{dH}_B^i(t)$ vanish (it would be embarassing if they did not), giving:

$$\hat{\mathrm{H}}_A(t) = \int_0^t \frac{\sum_{i=1}^n \left\{ Y_i(t') \mathrm{dH}_A^i(t') + \mathrm{d}M_i(t') \right\}}{\sum_{i=1}^n Y_i(t')}. \tag{5}$$

We would now like to take expectations. The expectation of $Y_i(t)$ is the probability that both the censoring time T_C and the joint event time T_J are both greater than t. Censoring being uninformative means that (i) we can multiply the survivor functions for T_C and T_J and (ii) the survivor function for T_C does not depend on the individual. The expectation of $Y_i(t)$ is given therefore by:

$$\mathbb{E}Y_i(t) = \mathrm{F}_A^i(t)\mathrm{F}_B^i(t)\mathrm{F}_C(t) \tag{6}$$

where we have also used the independence of events A and B. We can now take the expectations of both sides of (5):

$$\mathbb{E}\hat{\mathrm{H}}_A(t) = \int_0^t \frac{\sum_{i=1}^n \mathrm{F}_A^i(t')\mathrm{F}_B^i(t')\mathrm{dH}_A^i(t')}{\sum_{i=1}^n \mathrm{F}_A^i(t')\mathrm{F}_B^i(t')} \tag{7}$$

where we have used the Martingale property $\mathbb{E}\mathrm{d}M_i(t) = 0$ and then cancelled the $\mathrm{F}_C(t)$.

Expectation (7) is not satisfactory. The expectation of the estimate of the overall integrated excess hazard depends not only on the individual excess hazards (desirable and necessary) but also on the background hazards (undesirable as the underlying motivation is to remove the effect of background hazard).

We can remove the dependence by noting that the expected hazard in (7) is a weighted sum of the individual hazards. The $\mathrm{F}_B^i(t)$ are known quantities so we can adjust the weights to eliminate them. We define $Y_i^*(t)$ and $N_i^*(t)$by:

$$Y_i^*(t) = Y_i(t)/\mathrm{F}_B^i(t)$$

and

$$N_i^*(t) = N_i(t)/\mathrm{F}_B^i(t)$$

respectively. The Pohar-Perme $\tilde{\mathrm{H}}_A(t)$ estimate of the net survival is, by analogy with (4):

$$\tilde{\mathrm{H}}_A(t) = \int_0^t \frac{\mathrm{d}N_+^*(t')}{Y_+^*(t')} - \int_0^t \frac{\sum_{i=1}^n Y_i^*(t')\mathrm{dH}_B^i(t')}{Y_+^*(t')}.$$

The expectation of $\tilde{\mathrm{H}}_A(t)$ can be obtained by first noting that, by (6):

$$\mathbb{E}Y_i^*(t) = \mathbb{E}Y_i(t)/\mathrm{F}_B^i(t) = \mathrm{F}_A^i(t)\mathrm{F}_C(t)$$

and then following through the derivation of (7) to obtain:

$$\mathbb{E}\tilde{\mathrm{H}}_A(t) = \int_0^t \frac{\sum_{i=1}^n \mathrm{F}_A^i(t')\mathrm{dH}_A^i(t)}{\sum_{i=1}^n \mathrm{F}_A^i(t')} \tag{8}$$

which has a much more appropriate form as it is a weighted sum of the excess hazards.

Exercise The hazard experienced by the ith of n individuals is $\mathrm{h}^i(t)$. Show that the overall hazard $\bar{\mathrm{h}}(t)$ experienced by the population of n individuals is given by:

$$\bar{\mathrm{h}}(t) = \frac{\sum_{i=1}^n \mathrm{F}^i(t)\mathrm{h}^i(t)}{\sum_{i=1}^n \mathrm{F}^i(t)} \tag{9}$$

where $\mathrm{F}^i(t) = \exp\left[-\int_0^t \mathrm{h}^i(t')\mathrm{d}t'\right]$. (Note that this is essentially a *frailty* problem with a discrete frailty distribution $g_i = 1/n$.) Use (9) to interpret (8).