3.4 Proportional Hazards with a semi-parametric model called Cox regression

Again each subject j has a vector of covariates x_j and scale parameter $\rho_j = \rho_j(\beta.x_j)$. The basic assumption is that any two subjects have hazard functions whose ratio is a constant proportion which depends on the covariates. Hence we may write

$$h_j(t) = \rho_j h_0(t)$$

where h_0 is the baseline hazard function, β is a vector of regression coefficients to be estimated, and ρ_j again depends on the linear predictor $\beta.x_j$.

A general link could be used but in **Cox regression** $\rho_j = \mathrm{e}^{\beta.x_j}$. This model is termed semi-parametric because the functional form of the baseline hazard is not given, but is determined from the data, similarly to the idea for estimating the survival function by the Kaplan-Meier estimator.

3.4.1 Cox Regression

Suppose the event times are given by $0 < t_1 < t_2 < \cdots < t_m$. At this stage we assume no tied event times (list does not include censored times).

Let $[i]$ denote the subject with event at t_i.

Definition:Risk Set

The risk set R_i is the set of those subjects available for the event at time t_i.

Reminder: if we know that there are d subjects with hazard functions $h_1, \cdots, h_d$ then, knowing there is an event at time t_0, the probability that subject j has the event is

$$\Pr(\text{subject } \; j|t_0) = \frac{h_j(t_0)}{h_1(t_0) + \cdots + h_d(t_0)} \; .$$

Under the proportional hazards assumption we have

$$\Pr([\mathrm{i}]\,|t_i) = \frac{\rho_{[i]} h_0(t_i)}{\sum_{j \in R_i} \rho_j h_0(t_i)} = \frac{\rho_{[i]}}{\sum_{j \in R_i} \rho_j}$$

and the probability that $[i]$ has the event given it occurs at time t_i no longer depends on t_i.

Under the Cox regression model we have

$$\Pr([\mathrm{i}]\,|t_i) = \frac{\mathrm{e}^{\beta.x_{[i]}}}{\sum_{j \in R_i} \mathrm{e}^{\beta.x_j}} \; .$$

This probability only depends on the order in which subjects have the events.

The idea of the model is to specify a partial likelihood which depends only on the order in which events occur, not the times at which they occur. This means that the functional form of h_0, the baseline hazard function, is not required.

Definition: Partial Likelihood

$$L_P(\beta) = \prod_{t_i} \frac{e^{\beta . x_{[i]}}}{\sum\limits_{j \in R_i} e^{\beta . x_j}}$$

where R_i is the risk set at t_i, and subject $[i]$ is the subject with the event at t_i.

We can think of the partial likelihood as the *joint density function for subjects' ranks in terms of event order*, if there were no censoring and no tied event times.

Consequently if we use the partial likelihood for estimation of parameters we are *losing information,* because we are suppressing the actual times of events even though they are known, hence the name "partial likelihood".

Interestingly the partial likelihood acts in an exactly similar manner to the likelihood. Compute $\widehat{\beta}_P$ such that

$$L_P\left(\widehat{\beta}_P\right) = \sup_{\beta} \prod_{t_i} \frac{e^{\beta . x_{[i]}}}{\sum\limits_{j \in R_i} e^{\beta . x_j}}$$

Then $\widehat{\beta}_P$ maximises the partial likelihood and has all the usual properties.

Properties

(i) $\widehat{\beta}_P \xrightarrow{P} \beta$ as $m \longrightarrow \infty$ (and hence the number in the study tends to infinity also),

(ii) $\mathbf{var}\widehat{\beta}_P \approx I_P^{-1}$, where I_P is calculated from L_P in exactly the same way as for the usual information and likelihood,

(iii) asymptotic normality of $\widehat{\beta}_P$ also holds.

There are journal papers showing that the % information lost by ignoring actual event times is smaller than one might expect. All of the above rests on the assumption that the Cox regression model fits the data of course.

3.4.2 Relative Risk

There is a big difference between deductions from AL parametric analysis and PH semi-parametric analysis. In PH the intercept is non-identifiable and so we are estimating relative risk between subjects, not absolute risk, when we estimate the model parameters.

Definition: relative risk

The relative risk at time t between two subjects with covariates x_1, x_2 and hazard functions h_2, h_1 is defined to be

$$\frac{h_2(t)}{h_1(t)} \ .$$

For the Cox regression model this becomes time independent and is given by

$$e^{\beta . (x_2 - x_1)} \ .$$

The intercept is non-identifiable because

$$h(t;x) = \mathrm{e}^{\beta.x} h_0(t) = \mathrm{e}^{\alpha+\beta.x}\left(\mathrm{e}^{-\alpha} h_0(t)\right)$$

for any α. This means that any such intercept α included with the regression expression $\beta.x$ simply cancels out in the partial likelihood. Hence an intercept is never included in the linear regressor in this model.

However we do need to estimate the cumulative baseline hazard function and also the baseline survival function.

Definition :Breslow's estimator for the baseline cumulative hazard function

Breslow's estimator is given by

$$\widehat{H_0}(t) = \sum_{t_i \le t} \widehat{h_0}(t_i) = \sum_{t_i \le t} \frac{1}{\sum\limits_{j \in R_i} \mathrm{e}^{\beta.x_j}}$$

This is precisely what we would expect from a discrete distribution with

$$\widehat{h_0}(t_i) = \frac{1}{\sum\limits_{j \in R_i} \mathrm{e}^{\beta.x_j}}$$

where $\widehat{h_0}(t) = 0$ if t is not an event time.

Corollary: survival function estimator

$$\widehat{S_0}(t) = \mathrm{e}^{-\widehat{H_0}(t)}$$

is the estimator for the baseline survival function.

In some sense the discrete estimates for $\widehat{h_0}(t_i)$ can be thought of as the maximum likelihood estimators from the full likelihood, provided we assume that the hazard distribution is discrete (which of course it generally is not).

3.4.3 Plot for PH assumption with continuous covariate

Suppose we have a continuous covariate and we wish to check the proportional hazards assumption for that covariate. We do not have natural groups of subjects with the same value of that covariate.

Provided there is sufficient data we would group the subjects in quintiles of the covariate. Then we have 5 groups and can find the Kaplan-Meier estimator for each group. As before we plot

$$\log(-\log(\widehat{S_k}(t))) \text{ v. } \log t$$

for each $k = 1, \cdots, 5$ on the same graph. There should be a roughly constant vertical separation of groups. It generally is not a wonderful method, but is better than nothing.