Lecture 12

Semiparametric models: accelerated life, proportional hazards

12.1 Introduction to semiparametric modeling

We learned in section 6.2 how to compare observed mortality to a standard life table. In many settings, though, we are interested to compare observed mortality (or more general event times) between groups, or between individuals with different values of a quantitative covariate, and in the presence of censoring. For example,

Often we are interested to compare two (or more) different lifetime distributions. An approach that has been found to be effective is to think of there being a "standard" lifetime which may be modified in various simple ways to produce the lifetimes of the subpopulations. The standard lifetime is commonly estimated nonparametrically, while the modifications — usually the characteristic of primary interest — is reduced to one or a few parameters. The modifications may either involve a discrete collection of parameters — one parameter for each of a small number of subpopulations — or a regression-type parameter multiplied by a continuous covariate.

Examples of the former type would be clinical trials, where we compare survival time between treatment and control groups, or an observational study where we compare survival rates of smokers and non-smokers. An example of the second time would be testing time to appearance of full-blown AIDS symptoms as a function of measured T-cell counts.

There are two popular general classes of model: *Accelerated life* (AL, also called *Accelerated failure* models) and *Proportional hazards* (PH, also called *Relative risk* models).

12.2 Accelerated Life models

Suppose there are (several) groups, labelled by index i. The accelerated life model has a survival curve for each group defined by

$$S_i(t) = S_0(\rho_i t)$$

where $S_0(t)$ is some baseline survival curve and ρ_i is a constant specific to group i.

If we plot S_i against $\log t$, $i = 1, 2, \ldots, k$, then we expect to see a horizontal shift as

$$S_i(t) = S_0(\mathrm{e}^{\log \rho_i + \log t}) \, .$$

12.3 Proportional Hazards models

In this model we assume that the hazards in the various groups are proportional so that

$$h_i(t) = \rho_i h_0(t)$$

where $h_0(t)$ is the baseline hazard. Hence we see that

$$S_i(t) = S_0(t)^{\rho_i}$$

Taking logs twice we get

$$\log\big(-\log S_i(t)\big) = \log\rho_i + \log\big(-\log S_0(t)\big)$$

So if we plot the RHS of the above equation against either t or $\log t$ we expect to see a vertical shift between groups.

12.3.1 Plots

Taking both models together it is clear that we could plot

$$\log\Big(-\log\widehat{S}_i(t)\Big) \text{ against } \log t$$

as then we can check for *AL and PH in one plot.* Generally $\widehat{S}_i$ will be calculated as the Kaplan–Meier estimator for group i, and the survival function estimator for both groups will be plotted on the same graph.

- If the accelerated life model is plausible we expect to see a horizontal shift between groups.
- If the proportional hazards model is plausible we expect to see a vertical shift between groups.

Of course, if the data came from a Weibull distribution, with differences in the ρ parameter, it is simultaneously AL and PH. We see that

$$\log\big(-\log S_i(t)\big) = \log\rho_i + \alpha\log t.$$

Thus, survival curve estimates for different groups should appear approximately as parallel lines, which of course may be viewed as vertical or as horizontal shifts of one another.

In section 12.4 we illustrate this with two simulations of populations with Gompertz mortality.

12.3.2 Generalised linear survival models

Any parametric model may be turned into an AL model by replacing t by $\rho_i t$. And it may be turned into a PH model by replacing the hazard $h(t)$ by $\rho_i h(t)$.

There remains then the question, how to link ρ_i to the explanatory factors such as age, smoking status, blood pressure and so on that have been observed for the individual.

We need to incorporate these into a model using some sort of generalised regression. It is usual to do so by making ρ a function of the explanatory variables. For each observation (say individual in a clinical trial) we set the scale parameter $\rho = \rho(\beta\cdot x)$, where $\beta\cdot x$ is a linear predictor composed of a vector x of known explanatory variables (covariates) and an unknown vector β of parameters which will be estimated. The most common link function is

$$\log\rho = \beta\cdot x\ , \quad \text{equivalently} \quad \rho = \mathrm{e}^{\beta\cdot x}\ .$$

The **shape parameter** α is assumed to be the same for each observation in the study.

There are often very many covariates measured for each subject in a study. A row of data will have include

- **response**
  - event time t_i;
  - status δ_i (=1 if failure, =0 if censored);
  - possibly a left-truncation time.
- **covariates**, often a mix of quantitative and categorical variables such as
  - age;
  - sex;
  - systolic blood pressure;
  - treatment group.

As an example, suppose we think the Weibull distribution is generally a good fit. Then

$$S_i(t) = \mathrm{e}^{-(\rho t)^\alpha} \quad \text{and} \quad \rho = \mathrm{e}^{\beta\cdot\mathbf{x}}$$
$$\beta\cdot\mathbf{x} = \beta_0 + \beta_1\,\mathrm{age}_i + \beta_2\,\mathrm{sex}_i + \beta_3\,\mathrm{sbp}_i + \beta_4\,\mathrm{trt}_i,$$

where β_0 is the intercept and all regression coefficients β_j are to be estimated, as well as estimating α. Note this model assumes that α is the same for each subject. We have not shown interaction terms such as $x_{age} * x_{trt}$, but these could be added as well. This would allow a different effect of age according to treatment group.

Suppose subject i has covariate vector $\mathbf{x}^i$, and so scale parameter

$$\rho_i = \mathrm{e}^{\beta\cdot\mathbf{x}^i}\ .$$

This gives a likelihood

$$\begin{aligned} L(\alpha,\beta) &= \prod_i \left(\alpha\rho_i^\alpha t_j^{\alpha-1}\right)^{\delta_i} \mathrm{e}^{-(\rho_i t_i)^\alpha} \\ &= \prod_j \left(\alpha \mathrm{e}^{\alpha\beta\cdot\mathbf{x}^i} t_j^{\alpha-1}\right)^{\delta_i} \mathrm{e}^{-\left(\mathrm{e}^{\beta\cdot x_i} t_i\right)^\alpha}. \end{aligned}$$

We can now compute MLEs for α and all components of the vector β, using numerical optimisation, giving estimators $\widehat{\alpha}$, $\widehat{\beta}$ together with their standard errors $(\sqrt{\mathrm{Var}\,\widehat{\alpha}}, \sqrt{\mathrm{Var}\,\widehat{\beta}_j})$, estimated from the observed information matrix. Of course, the same could have been done for another parametric model instead of the Weibull.

In general, if we have a parametric model with cumulative hazard $H_\alpha(t)$ (where α represents the parameters of the model that do not vary between individuals), and $h_\alpha(t) = H'_\alpha(t)$ is the hazard function, we have the likelihood

$$L(\alpha,\beta) = \prod_i \left[\mathrm{e}^{\beta\cdot\mathbf{x}^i} h_\alpha\left(\mathrm{e}^{\beta\cdot\mathbf{x}^i} t_i\right)\right]^{\delta_i} \mathrm{e}^{-H(\mathrm{e}^{\beta\cdot\mathbf{x}^i} t_i)}.$$

If observations are left-truncated, say at a time s_i, the survival term includes only the cumulative hazard from s_i to t_i:

$$L(\alpha,\beta) = \prod_i \mathrm{e}^{\beta\cdot\mathbf{x}^i} h_\alpha\left(\mathrm{e}^{\beta\cdot\mathbf{x}^i} t_i\right) \mathrm{e}^{-H(\mathrm{e}^{\beta\cdot\mathbf{x}^i} t_i) + H(\mathrm{e}^{\beta\cdot\mathbf{x}^i} s_i)}.$$

We can test for $\alpha = 1$ using

$$2\log\widehat{L}_{\mathrm{weib}} - 2\log\widehat{L}_{\mathrm{exp}} \sim \chi^2(1), \text{ asymptotically.}$$

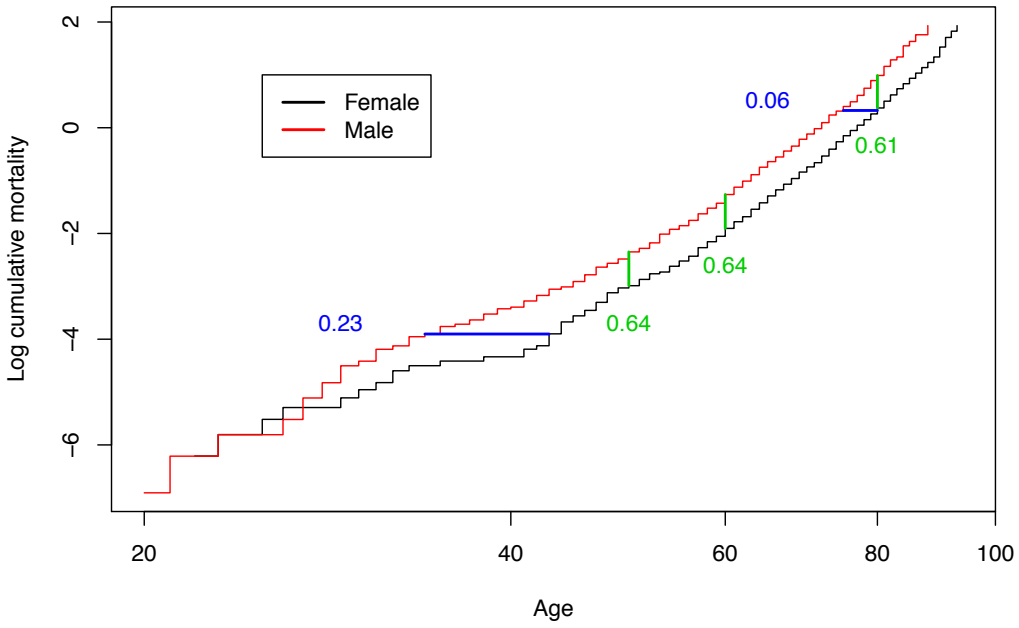
Packages allow for Weibull, log-logistic and log-normal models, sometimes others.

12.4 Simulation example

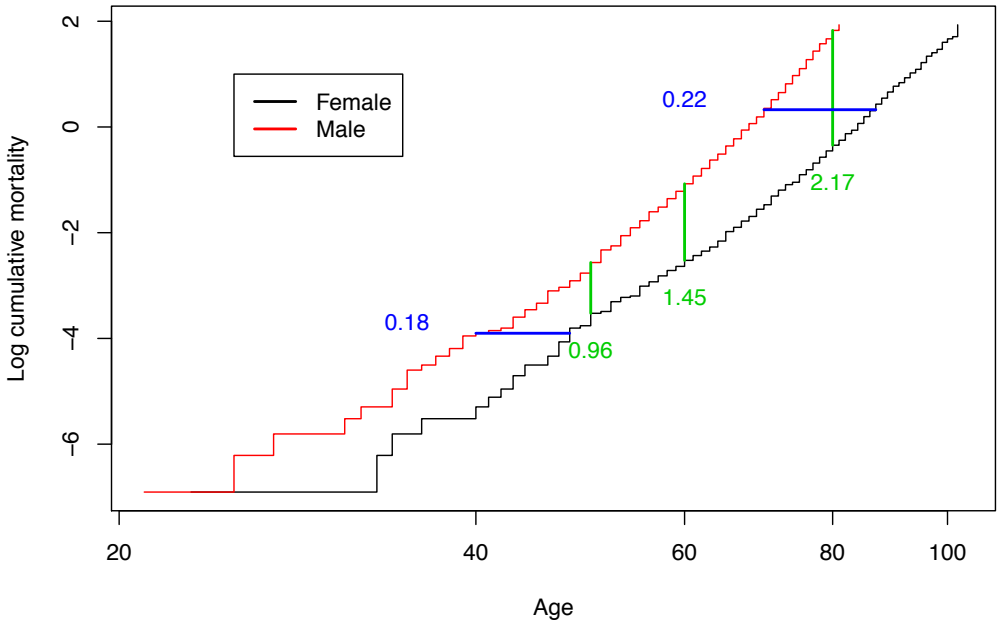
The Gompertz hazard fits naturally into either the AL or the PH framework. If individual i has hazard rate $B_i e^{\theta x}$ at age x, where θ is always the same, this is a PH model. If the hazard rate is $Be^{\theta_i x}$, this is an AL model. We can use this to illustrate the graphical approach to identifying AL or PH variation between different groups.

In Figure 12.1 we show the results of two different simulations where we have simulated 1000 survival times for each of two groups, one of which has Gompertz mortality rates approximating those of UK women, and the other somewhat higher rates, which we think of as "men". All are conditioned on survival to age 20 (since human survival ages before maturity are very different from Gompertz). In each case we plot the log estimated cumulative hazard of each group against age on a log scale, and draw the horizontal and vertical differences at various levels.

In Figure 12.1(a) the (B, θ) parameters are $(2 \times 10^{-5}, 0.11)$ and $(4 \times 10^{-5}, 0.11)$. We see, as expected, very nearly constant vertical differences, but large variation in the horizontal gap between the curves. Figure 12.1(b), on the other hand, shows the AL version, with parameters $(1 \times 10^{-5}, 0.11)$ and $(1 \times 10^{-5}, 0.14)$. The horizontal gaps are now similar, while the vertical differences change.



(a) PH Gompertz difference



(b) AL Gompertz difference

Figure 12.1: Simulations to show effect of PH and AL differences in Gompertz populations.

Lecture 13

Proportional hazards regression, Part I

It is possible to estimate AL models in a completely semi-parametric way — that is, baseline survival function S_0 is modelled non-parametrically, with no assumptions about its form, and each subject has time t_i scaled to $\rho_i t_i$. This model is beyond the scope of this course.

A simpler model — and by far the most popular survival model — is the proportional hazards semi-parametric model with log link function, generally known as Cox regression, after David Cox, who introduced the model in 1972. (Cox's paper [3] was listed in a recent survey by *Nature* as the 24th most frequently cited paper of all time, over all sciences.)

13.1 What is Cox Regression?

Each subject i has a (possibly right-censored) survival time T_i and a vector of covariates $\mathbf{x}^i$ and scale parameter $\rho_i = \rho_i\left(\beta \cdot \mathbf{x}^i\right)$. 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

$$h_i(t) = \rho_i h_0(t)$$

where h_0 is the baseline hazard function, β is a vector of regression coefficients to be estimated, and ρ_i again depends on the linear predictor $\beta \cdot \mathbf{x}^i$.

A general link could be used but in **Cox regression** we use the log link, so that $\rho_j = \mathrm{e}^{\beta \cdot \mathbf{x}^i}$. 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.

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 $[j]$ denote the subject with event at t_j.

Definition: Risk Set

The risk set R_j is the set of those subjects available for the event at time t_j.

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 i has the event is

$$\mathrm{P}\big\{\text{subject } \ j \,\big|\, t_0\big\} = \frac{h_i(t_0)}{h_1(t_0) + \cdots + h_d(t_0)} \ .$$

Under the proportional hazards assumption we have

$$\mathrm{P}\big\{[\mathrm{i}] \,\big|\, t_j\big\} = \frac{\rho_{[i]} h_0(t_j)}{\sum\limits_{i \in R_j} \rho_i h_0(t_j)} = \frac{\rho_{[j]}}{\sum\limits_{i \in R_j} \rho_i}$$

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

Under the Cox regression model we have

$$\mathrm{P}\{[\mathrm{i}] \mid t_j\} = \frac{\mathrm{e}^{\beta \cdot \mathbf{x}^{[j]}}}{\sum_{i \in R_j} \mathrm{e}^{\beta \cdot \mathbf{x}^i}} \ .$$

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

$$L_P(\beta) = \prod_{t_j} \frac{\mathrm{e}^{\beta \cdot x_{[i]}}}{\sum\limits_{i \in R_j} \mathrm{e}^{\beta \cdot x_j}}$$

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

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

$$L_P\left(\widehat{\beta}_P\right) = \sup_{\beta} \prod_{t_j} \frac{\mathrm{e}^{\beta \cdot x_{[i]}}}{\sum\limits_{i \in R_j} \mathrm{e}^{\beta \cdot x_j}}$$

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. We will discuss the question of how to test the fit of survival models in Lecture 16.

13.2 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

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

The intercept is non-identifiable because

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

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

13.3 Baseline hazard

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
Suppose the baseline survival is given by

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

where the discrete hazard estimation $\widehat{h_0}$ is given by

$$\widehat{h_0}(t_j) = \frac{1}{\sum\limits_{i\in R_j} \mathrm{e}^{\beta\cdot x_j}}$$

Breslow's estimator is given by

$$\widehat{h_0} = \frac{1}{\sum\limits_{i\in R_j} \mathrm{e}^{\beta\cdot x_j}} \tag{13.1}$$

In some sense the discrete estimates for $\widehat{h_0}(t_j)$ 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). When $\hat{\beta} = 0$ or when the covariates are all 0, this reduces simply to the Nelson-Aalen estimator. Otherwise, we see that this is equivalent to a modified Nelson-Aalen estimator, where the size of the risk set is weighted by the relative risks of the individuals. In other words, the estimate of $\widehat{h_0}$ is equivalent to the standard estimate # events/time at risk, but now time at risk is weighted by the relative risk.

The estimator may be loosely derived as follows:

$$\begin{aligned}\ell(h) &= \sum_{t_j} \log(1 - e^{-h_{[i]}(t_j)}) - \sum_{\substack{i\in R_j \\ j\neq[i]}} h_j \\ &= \sum_{t_j} \log(1 - e^{-\hat{\rho}_{[j]} h_0(t_j)}) - \sum_{\substack{i\in R_j \\ j\neq[j]}} \hat{\rho}_i h_0(t_j).\end{aligned}$$

We estimate $h_0(t_j)$ by

$$\begin{aligned} 0 &= \frac{\hat{\rho}_{[j]} e^{-\hat{\rho}_{[j]}\hat{h}_0(t_j)}}{1 - e^{-\hat{\rho}_{[j]}\hat{h}_0(t_j)}} - \sum_{\substack{i \in R_j \\ j \neq [j]}} \hat{\rho}_i \\ &\approx \frac{\hat{\rho}_{[j]}(1 - \hat{\rho}_{[j]}\hat{h}_0(t_j))}{\hat{\rho}_{[j]}\hat{h}_0(t_j)} - \sum_{\substack{i \in R_j \\ j \neq [j]}} \hat{\rho}_i \\ &= \frac{(1 - \hat{\rho}_{[j]}\hat{h}_0(t_j))}{\hat{h}_0(t_j)} - \sum_{\substack{i \in R_j \\ i \neq [j]}} \hat{\rho}_i. \end{aligned}$$

(In the second line we have assumed $h_0(t_j)$ to be small.) Thus

$$1 \approx \hat{h}_0(t_j) \left(\sum_{i \in R_j} \hat{\rho}_i \right),$$

which is the same as (13.1).