

Lecture 13

Semiparametric models: accelerated life, proportional hazards

Reading: Cox & Oakes chapter 5.1–5.7, K & M chapter 8.1–8.4, 8.8, 12.1–5

13.1 Introduction to semiparametric modeling

We learned in section 6.3 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 as in the heading above - AL and PH.

13.2 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 $\log t$, $i = 1, 2, \dots, k$, then we expect to see a horizontal shift as

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

13.2.1 Medians and Quantiles

Note too that each group has a different median lifetime, since, if $S_0(m) = 0.5$,

$$S_i\left(\frac{m}{\rho_i}\right) = S_0\left(\rho_i \frac{m}{\rho_i}\right) = 0.5,$$

giving a median for group i of $\frac{m}{\rho_i}$. Similarly if the $100\alpha\%$ quantile of the baseline survival function is t_α , then the $100\alpha\%$ quantile of group i is $\frac{t_\alpha}{\rho_i}$.

13.3 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

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

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.

13.3.1 Plots

Taking both models together it is clear that we should plot

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

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 each group will be plotted on the same graph.

(i) If the accelerated life model is plausible we expect to see a horizontal shift between groups.

(ii) If the proportional hazards model is plausible we expect to see a vertical shift between groups.

13.4 AL parametric models

There are several well-known parametric models which have the accelerated life property. These models also allow us to take account of continuous covariates such as blood pressure.

Name	Survival $S(t)$	Hazard $h(t)$	Density $f(t) = h(t)S(t)$
Weibull	$\exp(-(\rho t)^\alpha)$	$\alpha\rho^\alpha t^{\alpha-1}$	$\alpha\rho^\alpha t^{\alpha-1}e^{-(\rho t)^\alpha}$
log-logistic	$\frac{1}{1+(\rho t)^\alpha}$	$\frac{\alpha\rho^\alpha t^{\alpha-1}}{1+(\rho t)^\alpha}$	$\frac{\alpha\rho^\alpha t^{\alpha-1}}{(1+(\rho t)^\alpha)^2}$
log-normal	$1-\Phi\left(\frac{\log t+\log \rho}{\sigma}\right)$	\dots	$\frac{1}{t\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\log t + \log \rho)^2\right)$
exponential	$e^{-\rho t}$	ρ	$\rho e^{-\rho t}$

Remarks:

- (i) Exponential is a submodel of Weibull with $\alpha = 1$
- (ii) log-normal is derived from a normal distribution with mean $-\log \rho$ and variance σ^2 . In this distribution $\alpha = \frac{1}{\sigma}$ has the same role as in the Weibull and log-logistic.
- (iii) The **shape** parameter is α . The **scale** parameter is ρ .

Shape in the hazard function $h(t)$ is important.

- Weibull \dots $\begin{cases} h \text{ monotonic increasing } \alpha > 1 \\ h \text{ monotonic decreasing } \alpha < 1 \end{cases}$
- log-normal \dots $h \rightarrow 0$ as $t \rightarrow 0, \infty$, one mode only
- log-logistic \dots see problem sheet 5.

Comments:

- a) to get a "bathtub" shape we might use a mixture of Weibulls. This gives high initial probability of an event, a period of low hazard rate and then increasing hazard rate for larger values of t .
- b) to get an inverted "bathtub" shape we may have a mixture of log-logistics, or possibly a single log-normal or single log-logistic.

To check for appropriate parametric model (given AL checked)

There are some distributional ways of testing for say Weibull v. log-logistic etc., but they involve generalised F-distributions and are not in general use.

We can do a simple test for Weibull v. exponential as this simply means testing a null hypothesis $\alpha = 1$, and the exponential is a sub-model of the Weibull model. Hence we can use the likelihood ratio statistic which involves

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

13.4.1 Plots for parametric models

However most studies use plots which give a rough guide from shape. We should use a **straight-line fit** as this is the fit which the human eye spots easily.

1. **Exponential** - $S = e^{-\rho t}$, plot $\log S$ v. t
2. **Weibull** - $S = e^{-(\rho t)^\alpha}$, plot $\log(-\log S)$ v. $\log t$
3. **log-logistic** - $S = \frac{1}{1+(\rho t)^\alpha}$, plot \dots see problem sheet 6
4. **log-normal** - $S = 1 - \Phi\left(\frac{\log t+\log \rho}{\sigma}\right)$, plot $\Phi^{-1}(1 - S)$ v. $\log t$ or equivalently $\Phi^{-1}(S)$ v. $\log t$

In each of the above we would estimate S with the Kaplan-Meier estimator $\widehat{S}(t)$, and use this to construct the plots.

13.4.2 Regression in parametric AL models (assuming right censoring only)

In general studies each observation will have measured explanatory factors such as age, smoking status, blood pressure and so on. 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

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

Censoring is assumed to be independent mechanism and is sometimes referred to as non-informative.

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 perhaps:-

response - event time t_i , status δ_i (=1 if failure, =0 if censored)

covariates - age, sex, systolic blood pressure, treatment, and so a mixture of categorical variables and continuous variables amongst the covariates.

Suppose that Weibull is a good fit. Then

$$\begin{aligned} S(t) &= e^{-(\rho t)^\alpha} \quad \text{and} \quad \rho = e^{\beta \cdot x} \\ \beta \cdot x &= b_0 + b_1 x_{age} + b_2 x_{sex} + b_3 x_{sbp} + b_4 x_{trt} \end{aligned}$$

where b_0 is the intercept and all regression coefficients b_i 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, but could have, interaction terms such as $x_{age} * x_{trt}$. This interaction would allow a different effect of age according to treatment group.

Suppose subject j has covariate vector x_j and so scale parameter

$$\rho_j = e^{\beta \cdot x_j}.$$

This gives a likelihood

$$\begin{aligned} L(\alpha, \beta) &= \prod_j \left(\alpha \rho_j^\alpha t_j^{\alpha-1} \right)^{\delta_j} e^{-(\rho_j t_j)^\alpha} \\ &= \prod_j \left(\alpha e^{\alpha \beta \cdot x_j} t_j^{\alpha-1} \right)^{\delta_j} e^{-(e^{\beta \cdot x_j} t_j)^\alpha}. \end{aligned}$$

We can now compute MLEs for α and all components of the vector β , using numerical optimisation, giving estimators $\hat{\alpha}$, $\hat{\beta}$ together with their standard errors ($=\sqrt{\text{var}\hat{\alpha}}$, $\sqrt{\text{var}\hat{\beta}_j}$) calculated from the observed information matrix (see problem sheet 5). Of course, the same could have been done for another parametric model instead of the Weibull.

As already noted we can test for $\alpha = 1$ using

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

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

13.4.3 Linear regression in parametric AL models

The idea is to mirror ordinary linear regression and find a baseline distribution which does not depend on ρ , similar to looking at the error term in least squares regression. We give the derivation just for the Weibull distribution, but similar arguments work for all AL parametric models. We have

$$\begin{aligned} S(t) &= e^{-(\rho t)^\alpha} = \mathbb{P}\{T > t\} \\ &= \mathbb{P}\{\log T > \log t\} \\ &= \mathbb{P}\{\alpha(\log T + \log \rho) > \alpha(\log t + \log \rho)\} \end{aligned}$$

Now let $Y = \alpha(\log T + \log \rho)$ and $y = \alpha(\log t + \log \rho)$.

$$\begin{aligned} \mathbb{P}\{Y > y\} &= S_Y(y) \\ &= S(t) \\ &= e^{-(\rho t)^\alpha} \\ &= \exp(-e^y) \end{aligned}$$

Hence we have

$$\log T = -\log \rho + \frac{1}{\alpha}Y, \quad \text{where } S_Y(y) = \exp(-e^y)$$

The distribution of Y is independent of the parameters ρ and α . And in the case of the Weibull distribution its distribution is called the **extreme value distribution** and is as above.

In general we will write $\log T = -\log \rho + \frac{1}{\alpha}Y$ for all AL parametric models, and Y has a distribution in each case which is independent of the model parameters.

Name	$S(t)$	Y	$S_Y(y)$ distribution
Weibull	$\exp(-(\rho t)^\alpha)$	$\log T = -\log \rho + \frac{1}{\alpha}Y$	$\exp(-e^y)$: extreme value distrib.
log-logistic	$\frac{1}{1+(\rho t)^\alpha}$	$\log T = -\log \rho + \frac{1}{\alpha}Y$	$(1 + e^y)^{-1}$: logistic distribution
log-normal	$1 - \Phi\left(\frac{\log t + \log \rho}{\sigma}\right)$	$\log T = -\log \rho + \sigma Y$	$1 - \Phi(y)$: $\mathcal{N}(0, 1)$

as before $\alpha = \frac{1}{\sigma}$, for the log-normal.

In recent years, a semi-parametric model has been developed in which the baseline survival function S_0 is modelled non-parametrically, and each subject has time t scaled to $\rho_j t$. This model is beyond the scope of this course.