Modern Applied Statistical Models
Topic V: Survival Analysis

DT9209: MSc Applied Mathematics

Dr Joe Condon

School of Mathematical Sciences
Dublin Institute of technology

Introduction to survival data

DATA: Survival of women with breast cancer

stain	time	status	stain	time	status	stain	time	status
1	23	1	2	10	1	2	71	1
1	47	1	2	13	1	2	76	0
1	69	1	2	18	1	2	105	0
1	70	0	2	24	1	2	107	0
1	71	0	2	26	1	2	109	0
1	100	0	2	26	1	2	113	1
1	101	0	2	31	1	2	116	0
1	148	1	2	35	1	2	118	1
1	181	1	2	40	1	2	143	1
1	198	0	2	41	1	2	154	0
1	208	0	2	48	1	2	162	0
1	212	0	2	50	1	2	188	0
1	224	0	2	59	1	2	212	0
2	5	1	2	61	1	2	217	0
2	8	1	2	68	1	2	225	0

The stain was *Helix pomatia* agglutinin (HPA). A positive stain for a tumour indicated a tumour with the potential for metastasis (See Collett, page 6). In these data,

- stain: 1=negative staining 2=positive staining,
- time: survival time in months after mastectomy to remove tumour,
- status: 1=woman died at that time, 0=woman was still alive and last heard of at that time.

The question here is: Does a positive or negative staining predict survival after surgery? i.e. is the HPA marker a genuine tool for prognosis?

These data display some of the classic features of survival data.

- The response from the experiment is time itself. This is not like any other response (e.g. blood sugar) which can be taken over a number of individuals at an instant in time and where the metric used is independent of the response size.
  In survival data we have to wait for our response - both our metric and our response is time.
  We have to wait a long time to observe a 'large' response.
- Some of the women above yield survival times (length of time to death/failure) - but some do not.
  Women with status 0 are still alive at the end of the observation period - this is called right censoring or just censoring.

Right censoring can occur for a number of reasons:

1. When the observation period ends the subject is still alive.
2. The subject is lost to follow up - moves city or decides not to reply to questionnaires etc.
3. The subject dies from some other cause not related to the condition of interest.

Options for dealing with right censoring?

Disregard: Loss of information - you know something about these women and therefore should be able to use this info.

Include: Requires specialist techniques, i.e. survival analysis techniques.

Other types of censoring are not considered in this course.

Skew: Survival times tend to display skew - so methods based on normality assumptions cannot be used naively.

Either use non-parametric methods or parametric methods based on non-normal skewed distributions (e.g. Weibull or exponential).

Also, because the mean is sensitive to skew, the median becomes very important in survival analysis as a measure of location.

Components of Survival data:

Survival data is generally recorded as triple for each subject i, i.e. $\{t_i, d_i, x_i\}$:

t_i: this is the recorded time from time zero to death/event or censoring for subject i. The terminal event is well defined and need not be actual death - other failures can be considered, e.g.

- time to kidney failure
- time to failure of dental filling
- remission time from cancer
- time 'clean' from an addictive activity
- time to breakdown of a mechanical device

d_i: This is an indicator variable, 1=the subject experienced a failure (possibly a death) at time t_i, 0 = the subject was censored at that time for some reason.

x_i: covariate information about subject i, e.g. age, sex, length of disease onset, severity of disease, treatment etc. The covariate information is typically vector valued and can be extensive.

Modelling Survival

- Want to summarise survival times using probability models. What might be a plausible probability mechanism that gave rise to these data?
- Use three basic definitions: density function, survivor function, hazard function.

Density Function:

Survival time is a continuous *Random Variable* (RV) so it can be considered to have a density function.

Call the survival time density function - $f(t)$.

Recall from basic probability:

$$F(t) = P(T < t) = \int_0^t f(u)du \qquad (1)$$

where $f(.)$ is the density function. Its integral, $F(t)$, is called the (cumulative) distribution function.

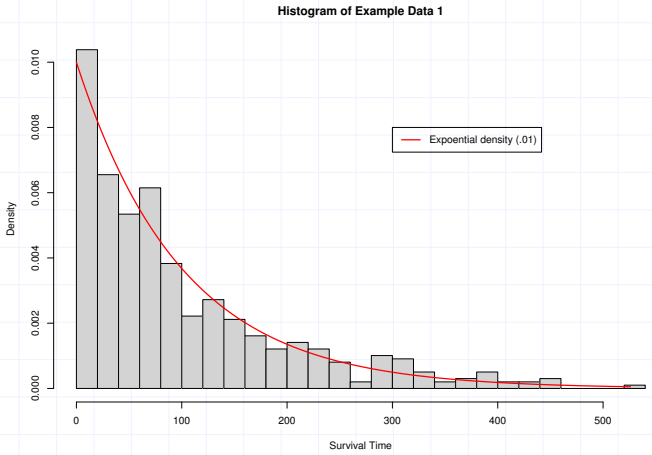
For example, the survival times could be distributed according to a exponential distribution. Again, from basic probability:

$$T \sim \text{exp}(\lambda) \qquad F(T) = 1 - e^{-\lambda t} \qquad f(t) = \lambda e^{-\lambda t}$$

In the survival of women with breast cancer data these are: t_i= time, d_i=status and x_i=stain.

The data set Example Data 1 is on the website. It consists of 496 simulated survival times (no censoring, sample mean = 101.1894, median=68.12) from an exponential distribution with $\lambda = .01$.

A histogram with the exponential density is given below.



The problem is what density to 'choose' in modelling real data.

The data might be expected to follow the adopted density and the parameters of the density might be estimated from the data using the method of maximum likelihood.

Example: Find the MLE of λ for the Example Data 1 dataset. Assume the survival times are independent.

$$\begin{aligned} \ell(\lambda) &= 496 \log \lambda - \lambda \sum_{i=1}^{496} t_i \\ \frac{\partial \ell(\lambda)}{\partial \lambda} &= \frac{496}{\lambda} - \sum_{i=1}^{496} t_i \\ &\Rightarrow \quad \hat{\lambda} = \frac{496}{50189.95} = 0.0099 \end{aligned}$$

...which is close to the 'true' value of 0.01.
There are many other skewed distributions that may be considered (Weibull, log normal, Gompertz, etc.).

Recall that $f(t) \approx P(t \leq T \leq t + \delta t)/\delta t$ - so only a measure of probability in a relative sense.

- $f(t)$ must be a non-negative,
- it can be interpreted as the rate of change of the CDF of the survival times.

Survivor function:

It is not easy to see where censored observations fit in with the density function.
We know something about censored observations. Up to the time that subject was censored they we still alive, i.e. they survived at least until time t_i.
A natural way of thinking about survival data therefore is to use the survivor function, $S(t)$:

$$S(t) = P(T > t) = 1 - F(t) = 1 - \int_0^t f(u)du \tag{2}$$

This function can be evaluated at any time t for both censored an uncensored observations.

Hazard function

It turns out that for much statistical modelling another function is very useful.

The hazard function, $h(t)$, is the instantaneous failure rate at time t, conditional on being alive just before time t. (Note the conditionality being introduced here.)

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T \leq t + \delta t | T \geq t)}{\delta t} \right\} \tag{3}$$

NB. $h(t)$ is a conditional rate of failure at time t. It is not a probability of failure. However, it can be interpreted as:

$$h(t)\delta t \approx P(t \leq T \leq t + \delta t | T \geq t)$$

i.e. the probability an individual dies in the interval $(t, t + \delta t]$ conditional on the person being alive at time t.

Relationship between density, survivor, and hazard functions

These three function are central to survival analysis techniques. They are interrelated.

$$\begin{aligned} P(t \leq T \leq t + \delta t | T \geq t) &= \frac{P(t \leq T \leq t + \delta t \cap T \geq t)}{P(T \geq t)} \\ &= \frac{P(t \leq T \leq t + \delta t)}{P(T \geq t)} \end{aligned}$$

- since one can only die in an interval if one is still alive at the start of that interval.

$$= \frac{F(t + \delta t) - F(t)}{S(t)}$$

$$\begin{aligned} \Rightarrow h(t) &= \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \frac{1}{S(t)} \right\} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{4}$$

This results immediately yields:

$$S(t) = \frac{f(t)}{h(t)} \tag{5}$$

$$f(t) = h(t)S(t) \tag{6}$$

Other useful relationships include:

$$f(t) = \frac{-dS(t)}{dt} \quad (7)$$

$$h(t) = \frac{-d\log S(t)}{dt} \quad (8)$$

$$S(t) = exp\{-\int_0^t h(u)du\} = exp\{-H(t)\} \quad (9)$$

where $\int_0^t h(u)du = H(t)$ is the integrated hazard function.

$H(t)$ is also called the **cumulative hazard function**.

Estimating of the Survivor Function

Simple case with NO censoring

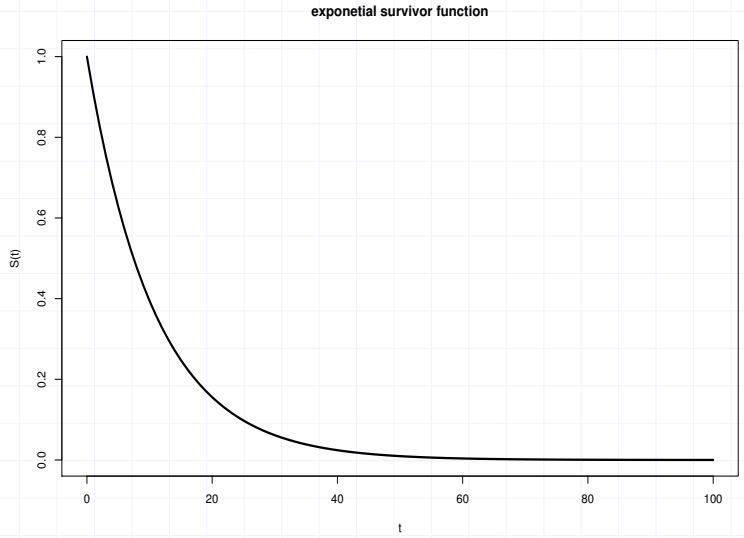
$t := \{5, 7, 7, 14, 21\}$ $\sum t_i = 54$ $\bar{t} = 10.8$

We could do as before an assume that the data are generated from an exponential density with some rate parameter λ. Using the MLE we could estimate $\hat{\lambda} = 5/54 = 0.093$.

From this it is straightforward to estimate the survivor function:

$$\begin{aligned}\hat{S}(t) &= 1 - \hat{F}(t) = 1 - \int_0^t \hat{f}(u)du \\ &= 1 - \int_0^t \hat{\lambda}e^{-\hat{\lambda}u}du \\ &= 1 - \int_0^t 0.093e^{-0.093u}du = e^{-0.093(t)}\end{aligned}$$

We can then plot this function over time.



- This has some advantages - like modelling survival beyond the end of the data.
- Disadvantages include that it may be hard to justify the use of the exponential density with small amounts of data
- The 'choice' of density can also be controversial with little to recommend one over another

An alternative therefore is to use a non-parametric method.

Kaplan-Meier Estimate (product limit)

- Consider intervals of time (non-overlapping) constructed such that there is only 1 death time in the interval (there may however, be more than 1 death at that death time).
- Easy way to construct such intervals is start each interval at each separate death time - therefore, deaths happen at start of interval.
- Intervals only defined from death times - not from censored times. The ordered death times (and hence the ordered interval start times) are denoted,
  $t_{(1)},\ t_{(2)}, t_{(3)}, \ldots, t_{(r)}$, with $t_{(0)} = 0$ and there are $j = 1, \ldots, r$ separate death times.
  So, the j^{th} interval is defined from $[t_{(j)}, t_{(j+1)})$
  - n_j = the number of individuals alive just before time $t_{(j)}$,
  - d_j the number of deaths in the interval $t_{(j)} - \delta$ to $t_{(j)}$ where δ is an infinitesimal period of time,
  - c_j is the number of individuals censored in $[t_{(j)}, t_{(j+1)})$.

- Where a censored observation is recorded at the same time as a death time, that censoring is taken to have occurred after any deaths recorded at the same time.

By this construction of the interval, the probability of surviving from $[t_{(j)},\ t_{(j+1)-\delta})$ is 1.

The probability of dying in $(t_{(j)-\delta}, t_{(j)}]$ is d_j/n_j.

Therefore the probability of survival from $[t_{(j)} - \delta,\ t_{(j+1)} - \delta)$ is $(n_j - d_j)/n_j$ and this is the probability of survival in $[t_{(j)},\ t_{(j+1)})$ as the limit of $\delta \to 0$

$$P_{\text{survival}}\left([t_{(j)},\ t_{(j+1)})\right) = \frac{n_j - d_j}{n_j}$$

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j} \tag{10}$$

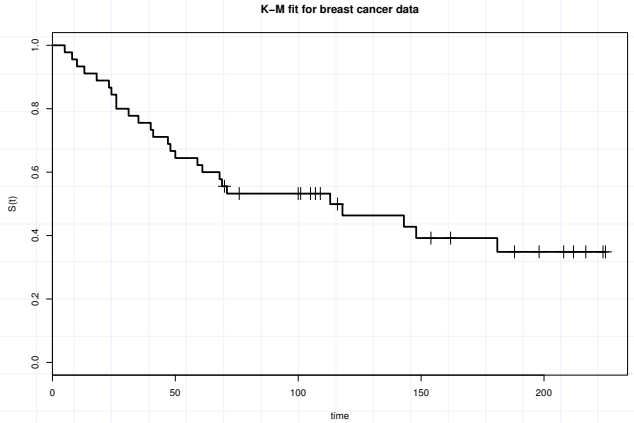
for $t_{(k)} \leq t < t_{(k+1)}$

$S(t)$ for $t < t_{(1)} = 1$
$S(t)$ for $t > t_{(r)} = 0$ iff the last observation is a death.
If the last observation (i.e. largest observed $t = t^*$) is a censored survival time, then $S(t)$ for $t \geq t^*$ is undefined.

The K-M is the limit of the life-table estimate at the intervals $\to 0$ - hence is also called the 'product limit' method.

Example: Breast cancer data:

time	status	n_j	d_j	$\hat{S}(t)$
0				1
5	1	45	1	0.978
8	1	44	1	0.956
10	1	43	1	0.933
13	1	42	1	0.911
18	1	41	1	0.889
23	1	40	1	0.867
24	1	39	1	0.844
26	1	38	2	0.800
31	1	36	1	0.778
35	1	35	1	0.756
40	1	34	1	0.733
41	1	33	1	0.711
47	1	32	1	0.689
48	1	31	1	0.667
50	1	30	1	0.644
59	1	29	1	0.622
61	1	28	1	0.600
68	1	27	1	0.578
69	1	26	1	0.556
71	1	24	1	0.532
113	1	16	1	0.499
118	1	14	1	0.463
143	1	13	1	0.428
148	1	12	1	0.392
181	1	9	1	0.349



Statistical Inference Based on Survivor Function

- The K-M estimate of survival gives only point estimates.
- Statistics also requires inference on these point estimates to aid decision making.
  e.g. If an stimated survival for a sample of transplant patients at 1 year is 0.6 - what does this tell me about the population of transplant patients at 1 year?
- Start by forming C.I.s for the survivor function - for this we need 2 things:
  - Variance of $\hat{S}(t)$,
  - Probability density of $\hat{S}(t)$.

C.I. for K-M estimate

$Var(\hat{S}(t))$:

$$\hat{S}(t) = \prod_{j=1}^{k} \hat{p}_j \Rightarrow \log \hat{S}(t) = \sum_{j=1}^{k} \log \hat{p}_j$$

$$\Rightarrow Var\{\log \hat{S}(t)\} = \sum_{j=1}^{k} Var(\log \hat{p}_j) \qquad (11)$$

[assuming independence]

The number $(n_j - d_j)$ of subjects who survive the interval $(t_{(k)},\ t_{(k+1)}]$ is a binomial RV with 'true' parameter p_j, mean $n_j p_j$ and variance $n_j p_j (1 - p_j)$.

$$\hat{p}_j = \frac{n_j - d_j}{n_j}$$

$$Var(\hat{p}_j) = Var\left\{\frac{n_j - d_j}{n_j}\right\} = \frac{Var\{n_j - d_j\}}{n_j^2}$$

$$= \frac{n_j p_j (1 - p_j)}{n_j^2} = \frac{p_j (1 - p_j)}{n_j}$$

But, we want the variance of $\log \hat{p}_j$.

In general for a RV X, $Var[f(X)] \neq f(Var[X])$.

i.e. $$Var[\log \hat{p}_j] \neq \log\left[\frac{p_j (1 - p_j)}{n_j}\right]$$

So, how do we determine the this variance?

Taylor Series and the Delta Method

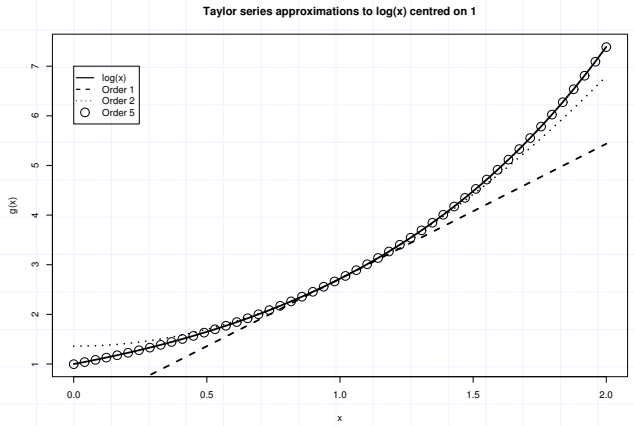
For a function $g(x)$ with derivatives of order r, the Taylor polynomial of order r about a constant a is:

$$T_r(x) = \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i$$

Taylor's theorem shows that $g(x) \equiv T_\infty(x)$ and that for $r < \infty$ the remainder always tends to zero faster than the highest-order explicit term. The upshot of all this, is that $T_r(x) \approx g(x)$ in the locality of the constant a. The approximation improves with the order of the polynomial (r).

Example:

$$\begin{aligned} g(x) &= \log(x) \\ g^{(1)}(x) &= x^{-1} \quad g^{(2)}(x) = -x^{-2} \quad g^{(3)}(x) = 2x^{-3} \\ g^{(i)}(x) &= (-1)^{i-1}x^{-i}(i-1)!, \qquad i > 0 \\ \Rightarrow T_r(x) &= \log(a) + \sum_{i=1}^{r} \frac{(-1)^{i-1}x^{-i}(i-1)!}{i!}(x-a)^i \end{aligned}$$



- all three are pretty good in the locality of a.

- We can use this to get an approximation to the variance of $g(X)$ where X is a random variable - thus avoiding the difficulty of deriving the probability density for $g(X)$.

Let $g(t)$ be some function of a parameter estimate t. Define $g(\theta)$ be the same function but of the 'true' population parameter θ for which t is an unbiased estimate.

NB. t is a random variable, θ is a fixed constant population parameter. Define:

$$g'(\theta) = \frac{\partial}{\partial t} g(t)\Big|_{t=\theta}$$

Take the first order Taylor series expansion around θ:

$$g(t) \approx g(\theta) + g'(\theta)(t-\theta) \tag{12}$$

which will be a 'good' approximation if the omitted higher order terms are small. Now take expectation on both sides:

$$E[g(t)] \approx g(\theta) + g'(\theta)E[(t-\theta)] = g(\theta) \tag{13}$$

Now, the variance of $g(t)$ can be approximated by:

$$\begin{aligned} Var[g(t)] &= E\left[g(t) - E\{g(t)\}\right]^2 \\ &\approx E\left[g(t) - g(\theta)\right]^2 \qquad \{\text{using eqn. } 13\} \\ &\approx E\left[g(\theta) + g'(\theta)[(t-\theta)] - g(\theta)\right]^2 \qquad \{\text{using eqn. } 12\} \\ &= E\left[g'(\theta)[(t-\theta)]\right]^2 \\ &= [g'(\theta)]^2 Var(t) \\ &\approx [g'(t)]^2 Var(t) \end{aligned} \tag{14}$$

We can now apply this to our problem:

$$\begin{aligned} Var(\log \hat{p}_j) &\approx \left[\frac{\partial \log \hat{p}_j}{\partial \hat{p}_j}\right]^2 Var(\hat{p}_j) \\ &= \frac{1-\hat{p}_j}{\hat{p}_j n_j} \end{aligned}$$

Given $p_j = (n_j - d_j)/n_j$ we get:

$$Var(\log \hat{p}_j) \approx \frac{d_j}{(n_j - d_j)n_j}$$

From equation (11):

$$Var\{\log \hat{S}(t)\} = \sum_{j=1}^{k} Var(\log \hat{p}_j) \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}$$

Now we are really interested in $Var\{\hat{S}(t)\}$ and using equation (13) once more, we can write:

$$\begin{aligned} Var\{\log \hat{S}(t)\} &\approx \frac{1}{[\hat{S}(t)]^2} Var\{\hat{S}(t)\} \\ \Rightarrow var\{\hat{S}(t)\} &\approx [\hat{S}(t)]^2 Var\{\log \hat{S}(t)\} \\ &= [\hat{S}(t)]^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} \end{aligned} \tag{15}$$

The standard error is simply the square root of the estimated variance:

$$se\{\hat{S}(t)\} \approx \hat{S}(t) \sqrt{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}} \tag{16}$$

Equation (16) is Greenwood's formula.

The Delta method takes the basic result concerning the Taylor series approximation and states the following:

$$\text{if } \sqrt{n}(\hat{\theta} - \theta) \to N(0, \sigma^2)$$

$$\text{then,} \sqrt{n}(g(\hat{\theta}) - g(\theta)) \to N(0, [g'(\theta)]^2 \sigma^2) \tag{17}$$

Less formally we can say that approximately:

$$g(\hat{\theta}) \sim N\left(g(\theta), [g'(\theta)]^2 \sigma^2/n\right)$$

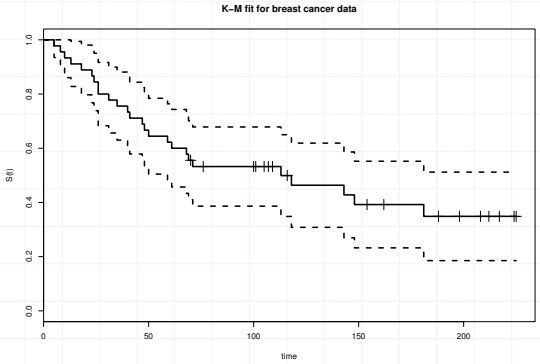
Since we don't know θ and σ^2 we estimate them from the data using $\hat{\theta}$ and $\hat{\sigma}^2$, i.e.

$$g(\hat{\theta}) \approx N\left(g(\hat{\theta}), [g'(\hat{\theta})]^2 \hat{\sigma}^2/n\right) \tag{18}$$

Putting this all together we get can calculate a 95% CI for $S(t)$ as:

$$\hat{S}(t) \pm Z_{\alpha/2}\ se\{\hat{S}(t)\} \tag{19}$$

Example: Breast Cancer Data



NB. these are pointwise CI's and not CI for the whole survivor function.

time	status	n_j	d_j	$\hat{s}(t)$	$\frac{d_j}{n_j(n_j-d_j)}$	$\sum \frac{d_j}{n_j(n_j-d_j)}$	se
0		45		1			
5	1	45	1	0.978	0.0005	0.0005	0.0220
8	1	44	1	0.956	0.0005	0.0010	0.0307
10	1	43	1	0.933	0.0006	0.0016	0.0372
13	1	42	1	0.911	0.0006	0.0022	0.0424
18	1	41	1	0.889	0.0006	0.0028	0.0468
23	1	40	1	0.867	0.0006	0.0034	0.0507
24	1	39	1	0.844	0.0007	0.0041	0.0540
26	1	38	2	0.800	0.0015	0.0056	0.0596
31	1	36	1	0.778	0.0008	0.0063	0.0620
35	1	35	1	0.756	0.0008	0.0072	0.0641
40	1	34	1	0.733	0.0009	0.0081	0.0659
41	1	33	1	0.711	0.0009	0.0090	0.0676
47	1	32	1	0.689	0.0010	0.0100	0.0690
48	1	31	1	0.667	0.0011	0.0111	0.0703
50	1	30	1	0.644	0.0011	0.0123	0.0714
59	1	29	1	0.622	0.0012	0.0135	0.0723
61	1	28	1	0.600	0.0013	0.0148	0.0730
68	1	27	1	0.578	0.0014	0.0162	0.0736
69	1	26	1	0.556	0.0015	0.0178	0.0741
71	1	24	1	0.532	0.0018	0.0196	0.0745
113	1	16	1	0.499	0.0042	0.0238	0.0769
118	1	14	1	0.463	0.0055	0.0293	0.0793
143	1	13	1	0.428	0.0064	0.0357	0.0808
148	1	12	1	0.392	0.0076	0.0432	0.0815
181	1	9	1	0.349	0.0139	0.0571	0.0833

Note: that for any upper CL calculated > 1 is set = 1 and similarly any lower CL < 0 is set = 0.

An alternative is to transform the data, calculate the CI on the transformed scale and then back tansform in such a way as to ensure a CI between (0,1).

One such transformation is the complimentary log-log, i.e. first transform: $\log(-\log \hat{S}(t))$.

$$\begin{aligned} Var\{\log(-X)\} &= \frac{1}{X^2} Var\{X\} \\ \Rightarrow Var\{\log[-\log \hat{S}(t)]\} &= \frac{1}{\{\log \hat{S}(t)\}^2} \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j))} \\ \Rightarrow CI &= \hat{S}(t)^{exp\left\{\pm z_{\alpha/2}\sqrt{Var[log(-log\hat{S}(t))]}\right\}} \end{aligned}$$

This method however is known to be anti-conservative when $\hat{S}(t)$ is close to zero or one.

Estimates of Hazard

NB. Strictly speaking the hazard function for a non-parametric step-function survivor function estimate is not defined.

K-M like estimate of hazard:

$$\hat{h}(t) = \frac{d_j}{n_j t_j}$$

where t_j is the length of time interval.

So, hazard is a rate per unit time but no estimate at final death time.

A standard error may be derived.

d_j is a binomial RV. Define p_j as the probability of dying in an interval.

where, $Var(d_j) = n_j p_j (1 - p_j)$ using $\hat{p}_j = d_j / n_j$

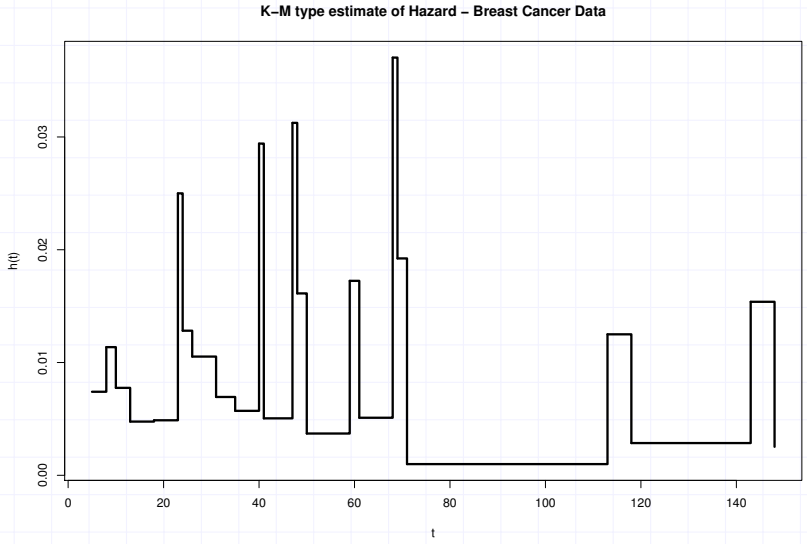
$$
\begin{aligned}
Var[\hat{h}(t)] &= \frac{Var[d_j]}{n_j^2 t_j^2} = \frac{n_j p_j (1-p_j)}{n_j^2 t_j^2} \\
&= \frac{p_j (1-p_j)}{n_j t_j^2} = \frac{1}{n_j t_j^2} \cdot \frac{d_j}{n_j} \cdot \frac{(n_j - d_j)}{n_j} \\
&= \frac{1}{n_j^2 t_j^2} \cdot \frac{d_j}{d_j} \cdot \frac{d_j (n_j - d_j)}{n_j} = \frac{d_j^2}{n_j^2 t_j^2} \cdot \frac{d_j (n_j - d_j)}{n_j d_j} \\
&= \hat{h}^2(t) \frac{d_j (n_j - d_j)}{n_j d_j}
\end{aligned}
$$

which gives:

$$se\{\hat{h}(t)\} = \hat{h}(t) \sqrt{\frac{n_j - d_j}{n_j d_j}}$$

NB. when d_j is small these CIs can be quite wide and of little practical use.

time	nj	dj	$\hat{h}(t)$	se	LCL	UCL
0	45		0			
5	45	1	0.0074	0.0073	0.000	0.0218
8	44	1	0.0114	0.0112	0.000	0.0334
10	43	1	0.0078	0.0077	0.000	0.0228
13	42	1	0.0048	0.0047	0.000	0.0140
18	41	1	0.0049	0.0048	0.000	0.0143
23	40	1	0.0250	0.0247	0.000	0.0734
24	39	1	0.0128	0.0127	0.000	0.0376
26	38	2	0.0105	0.0072	0.000	0.0247
31	36	1	0.0069	0.0068	0.000	0.0204
35	35	1	0.0057	0.0056	0.000	0.0168
40	34	1	0.0294	0.0290	0.000	0.0862
41	33	1	0.0051	0.0050	0.000	0.0148
47	32	1	0.0313	0.0308	0.000	0.0915
48	31	1	0.0161	0.0159	0.000	0.0472
50	30	1	0.0037	0.0036	0.000	0.0108
59	29	1	0.0172	0.0169	0.000	0.0504
61	28	1	0.0051	0.0050	0.000	0.0149
68	27	1	0.0370	0.0363	0.000	0.1083
69	26	1	0.0192	0.0189	0.000	0.0562
71	24	1	0.0010	0.0010	0.000	0.0029
113	16	1	0.0125	0.0121	0.000	0.0362
118	14	1	0.0029	0.0028	0.000	0.0083
143	13	1	0.0154	0.0148	0.000	0.0444
148	12	1	0.0025	0.0024	0.000	0.0073
181	9	1	.	.		



Cumulative Hazard

The cumulative hazrad function $H(t)$ is defined as:

$$H(t) = \int_0^t h(u)du = -\log S(t)$$

the last is more convenient generally.

E.g. using the K-M estimate, $\hat{S}(t)$ we get:

$$\hat{H}(t) = -\sum_{j=1}^{k} \log \hat{S}(t) = -\sum_{j=1}^{k} \log\left(\frac{n_j - d_j}{n_j}\right)$$

time	nj	dj	$\hat{S}(t)$	$H(t)$ -KM	$H(t)$ - NA
0	45		1.000	0.000	0.000
5	45	1	0.978	0.022	0.022
8	44	1	0.956	0.045	0.045
10	43	1	0.933	0.069	0.068
13	42	1	0.911	0.093	0.092
18	41	1	0.889	0.118	0.116
23	40	1	0.867	0.143	0.141
24	39	1	0.844	0.169	0.167
26	38	2	0.800	0.223	0.220
31	36	1	0.778	0.251	0.247
35	35	1	0.756	0.280	0.276
40	34	1	0.733	0.310	0.305
41	33	1	0.711	0.341	0.336
47	32	1	0.689	0.373	0.367
48	31	1	0.667	0.405	0.399
50	30	1	0.644	0.439	0.433
59	29	1	0.622	0.474	0.467
61	28	1	0.600	0.511	0.503
68	27	1	0.578	0.549	0.540
69	26	1	0.556	0.588	0.578
71	24	1	0.532	0.630	0.620
113	16	1	0.499	0.695	0.682
118	14	1	0.463	0.769	0.754
143	13	1	0.428	0.849	0.831
148	12	1	0.392	0.936	0.914
181	9	1	0.349	1.054	1.025

Estimating the median survival times and other percentiles

- The median is preferred as a measure of location over the mean because of the skew in the survival times.
- How this is done - depends on whether parametric or nonparametric methods are used.

Nonparametric methods

Define: $t(50)$ as the time beyond which 50% of the subjects survive

$t(50) = s\{(t(50)\} = .5$

Since the $S(t)$ is a step function and only changes at death times then there may not be a death time $t*$ such that:

$$S(t*) = 0.5$$

So refine definition:

$$\begin{aligned} \hat{t}(50) &= min\{t_i; \hat{S}(t_i) < 0.5\} \\ &= min\{t_{(j)}; \hat{S}(t_{(j)}) < 0.5\} \end{aligned}$$

for some death time $t_{(j)}$.

There may be cases where there is a death time for which $\hat{S}(t) = 0.5$. This means that an interval of time is defined such that:

$$\hat{S}\left([t_{(k)}, t_{(k+1)})\right) = 0.5$$

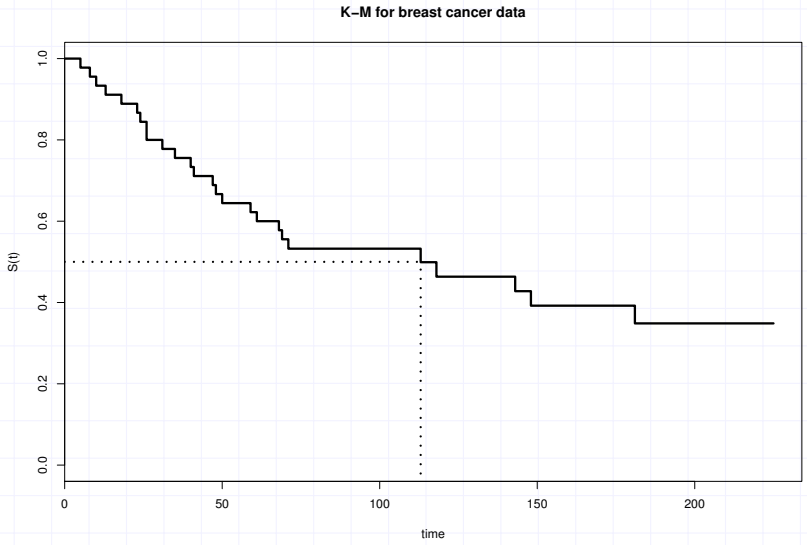
i.e. the survivor function is 0.5 throughout the interval. In such a case the median is taken as the midpoint of this interval.

Example: Breast cancer data

time	nj	dj	S(t)
⋮			
68	27	1	0.578
69	26	1	0.556
71	24	1	0.532
113	16	1	0.499
118	14	1	0.463
143	13	1	0.428
⋮			

So the median is 113 months.

This may be visualised from the plot of the survivor function as follows:



There is nothing mathematically special about $t(50)$.

Might be interested in $t(30)$ or $t(90)$ or other percentiles, i.e. $t(30)$ is the survival time by which 30% of people have died etc.

Define in general:

$$\begin{aligned}\hat{t}(p) &= min\{t_i; \hat{S}(t_i) < 1 - p/100\} \\ &= min\{t_{(j)}; \hat{S}(t_{(j)}) < 1 - p/100\}\end{aligned}$$

Comparing Two Groups

Log Rank Test

- Given two groups, take a particular death time $t_{(j)}$:

Group	# of deaths at $t_{(j)}$	# surviving beyond $t_{(j)}$	# at risk at $t_{(j)}$
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
total	d_j	$n_j - d_j$	n_j

If the margins are fixed, then the table is determined by d_{1j}.

$d_{1j} \sim$ Hypergeometric

Mean and variance of Hypergeometric

$$P(d_{ij}) = \binom{n_{ij}}{d_{1j}}\binom{n_{2j}}{d_{2j}} \Big/ \binom{n_j}{d_j}$$

$$e_{1j} = E(d_{1j}) = \frac{n_{1j}d_j}{n_j}$$

$$U_l = \sum_{j=1}^{r}(d_{1j} - e_{1j})$$

where the sum is over the $j = 1, \ldots, r$ individual death times.

NB. this test statistic takes the form observed-expected under H_0 :

So, a small U_l does not supply evidence against the H_0 :, a large U_l may.

U_l is the log rank test statistic.

$$Var[U_l] = \sum_{j=1}^{r} v_{1j} = \sum_{j=1}^{r} \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$$

under the assumption that the number of deaths at each time are independent.

Where the number of deaths is not too small, it can be shown that:

$$U_l \sim N(0, var(U_l))$$

$$\Rightarrow W_l = \frac{U_l}{\sqrt{Var(U_l)}} \sim N(0,1)$$

$$\text{or, } \frac{U_l^2}{Var(U_l)} \sim \chi_1^2$$

under the H_0 : that the survivor functions in the two groups are equal.

Example: Acute Myelogenous Leukemia

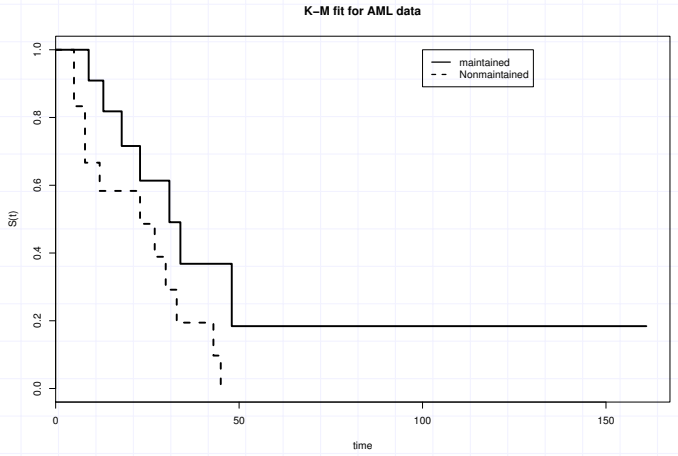
time	status	treatment	time	status	treatment
9	1	Maintained	5	1	Nonmaintained
13	0	Maintained	5	1	Nonmaintained
13	1	Maintained	8	1	Nonmaintained
18	1	Maintained	8	1	Nonmaintained
23	1	Maintained	12	1	Nonmaintained
28	0	Maintained	16	0	Nonmaintained
31	1	Maintained	23	1	Nonmaintained
34	1	Maintained	27	1	Nonmaintained
45	0	Maintained	30	1	Nonmaintained
48	1	Maintained	33	1	Nonmaintained
161	0	Maintained	43	1	Nonmaintained
			45	1	Nonmaintained

Scientific question: is the survival experience for maintained group (chemotherapy) the same as the nonmaintained group?

Statistical version: Does the data display any significant evidence that the survivor function are different between the two groups?

H_0 : survivor functions are the same
H_a : survivor functions are not the same

First; take a look at K-M estimates of the survivor functions.



So, it looks at though the estimates of the survivor functions are different.
But - is there enough evidence in the data to determine if the difference is significant?

Calculate the log-rank statistic...

time	d_{1j}	d_{2j}	n_{1j}	n_{2j}	n_j	d_j	e_{1j}	u_{1j}	v_{ij}
5	0	2	11	12	23	2	0.957	-0.957	0.476
8	0	2	11	10	21	2	1.048	-1.048	0.474
9	1	0	11	8	19	1	0.579	0.421	0.244
12	0	1	10	8	18	1	0.556	-0.556	0.247
13	1	0	10	7	17	1	0.588	0.412	0.242
18	1	0	8	6	14	1	0.571	0.429	0.245
23	1	1	7	6	13	2	1.077	-0.077	0.456
27	0	1	6	5	11	1	0.545	-0.545	0.248
30	0	1	5	4	9	1	0.556	-0.556	0.247
31	1	0	5	3	8	1	0.625	0.375	0.234
33	0	1	4	3	7	1	0.571	-0.571	0.245
34	1	0	4	2	6	1	0.667	0.333	0.222
43	0	1	3	2	5	1	0.600	-0.600	0.240
45	0	1	3	1	4	1	0.750	-0.750	0.188
48	1	0	2	0	2	1	1.000	0.000	0.000
								-3.689	4.008

Log-rank test statistic: $U_l = \frac{(-3.689)^2}{4.008} = 3.396$.

Compare this with a χ^2_1: p-value = 0.065

Comparing G-groups

There is an intuitive extension from the log-rank to K groups of survival data.

Lets say that there are g groups.

Define for $k = 1, \ldots, g-1$:

$$U_{Lk} = \sum_{i=1}^{r} \left(d_{kj} - \frac{n_{kj} d_j}{n_j} \right) \tag{20}$$

NB. Group g is determined by the other groups if the margins are fixed.

Each of the $k = 1, \ldots, g-1$ components are the stacked into a column vector denoted U_L.

The variance of the vector U_L are given by their respective variance-covariance matrices.

The entries in this matrix are calculated by:

$$V_{Lkk'} = \sum_{i=1}^{r} \frac{n_{kj} d_j (n_j - d_j)}{n_j (n_j - 1)} \left(\delta_{kk'} - \frac{n_{k'j}}{n_j} \right) \tag{21}$$

$$\tag{22}$$

where $\delta_{kk'} = 1$ if $k = k'$ and zero otherwise.

Denoting this matrix as V_L and assuming that $g = 4$ then we get:

$$\begin{pmatrix} V_{L11} & V_{L12} & V_{L13} \\ V_{L21} & V_{L22} & V_{L23} \\ V_{L31} & V_{L32} & V_{L33} \end{pmatrix}$$

Note: this matrix will be symmetric as $V_{Lij} = V_{Lji}$.

The test statistics for the G groups is:

$$U_L' V_L^{-1} U_L \tag{23}$$

which is distributed as χ^2 with $g-1$ degrees of freedom.

Parametric and Semi-parametric Survival Models

- Need to consider the likelihood for both censored and uncensored observations.
- A regression type formulation would also help in modelling covariates.

Recall: the likelihood is (proportional to) the probability of the data considered as a function of unknown parameters.

For each observation there are two random processes involved: The survival process and the censoring process.

Denote:
$f(t)$ and S(t) as the density and survivor functions of the survival process
and
$f_c(t)$ and $S_c(t)$ as the density and survivor functions of the censoring process

Observe	survival process	censoring process
$(t_i, d_i = 1)$	$f(t_i)$	$S_c(t_i)$
$(t_i, d_i = 0)$	$S(t_i)$	$f_c(t_i)$

Therefore, the likelihood for one observation can be written:

$$L_i(\theta, \theta_c | t_i, d_i) = \left\{ f(t_i)^{d_i} S(t_i)^{(1-d_i)} \right\} \left\{ f_c(t_i)^{(1-d_i)} S_c(t_i)^{d_i} \right\}$$

where θ is the parameter vector for the survivor process and θ_c is the parameter vector for the censoring process.

The likelihood over $i = 1, \ldots, n$ observations (assuming independence) is:

$$L(\theta, \theta_c | t, d) = \prod_{i=1}^{n} \left\{ f(t_i)^{d_i} S(t_i)^{(1-d_i)} \right\} \left\{ f_c(t_i)^{(1-d_i)} S_c(t_i)^{d_i} \right\} \tag{24}$$

where t and d are the vectors of survival times and censoring indicators.

If the survival process is independent of the censoring process then the likelihood for θ can be written:

$$L(\theta | t, d) = \prod_{i=1}^{n} \left\{ f(t_i)^{d_i} S(t_i)^{(1-d_i)} \right\} C_i \tag{25}$$

where the C_i is some constant in the likelihood for θ, i.e. carries the information on θ_c and no information on θ.

Take the log likelihood for θ,

$$\log L(\theta | t, d) = \sum_{i=1}^{n} \log \left\{ f(t_i)^{d_i} S(t_i)^{(1-d_i)} \right\} + \sum_{i=1}^{n} \log C_i \tag{26}$$

The maximisation of equation (26) WRT θ will not involve $\sum_{i=1}^{n} \log C_i$ and so this term may be omitted to give:

$$\log L(\theta | t, d) = \sum_{i=1}^{n} \log \left\{ f(t_i)^{d_i} S(t_i)^{(1-d_i)} \right\} \tag{27}$$

Since that $f(t_i) = h(t_i) S(t_i)$ we can rewrite equation (27) as:

$$\log L(\theta | t, d) = \sum_{i=1}^{n} \log \left\{ h(t_i)^{d_i} S(t_i) \right\} \tag{28}$$

$$= \sum_{i=1}^{n} d_i \log h(t_i) + \log S(t_i) \tag{29}$$

Example:

t_i	8	99	8	11	68	33	9	100	141	4	$\sum = 481$
d_i	0	0	1	0	0	0	0	1	1	1	$\sum = 4$

Assume that $t_i \sim$ Exponential(λ), then the likelihood is:

$$\begin{aligned} L(\lambda|t,d) &= \prod_{i=1}^{10} \{\lambda e^{-\lambda t_i}\}^{d_i} \{e^{-\lambda t_i}\}^{(1-d_i)} \\ &= \prod_{i=1}^{10} \lambda^{d_i} e^{-\lambda t_i} \\ \log L(\lambda|t,d) &= \sum_{i=1}^{10} [d_i \log \lambda - \lambda t_i] \end{aligned}$$

and hence we find that $\hat{\lambda} = 4/481$

Proportional Hazard (PH) Regression

- The PH assumption for two hazards is:

$$h_1(t) = \psi h_2(t)$$

- This can be used as the staring point for a regression type model.
- Assume that there is a baseline hazard - call this $h_0(t)$.
- The model relates all other observations to this baseline in a simple way.

Data: Ovarian cancer patients

patient	time	status	treat	age	rdisease	perf	age2
1	156	1	0	66	1	1	16
2	1040	0	0	38	1	1	-12
3	59	1	0	72	1	0	22
4	421	0	1	53	1	0	3
5	329	1	0	43	1	0	-7
6	769	0	1	59	1	1	9
7	365	1	1	64	1	0	14
8	770	0	1	57	1	0	7
9	1227	0	1	59	0	1	9
10	268	1	0	74	1	1	24
11	475	1	1	59	1	1	9
12	1129	0	1	53	0	0	3
13	464	1	1	56	1	1	6
14	1206	0	1	44	1	0	-6
15	638	1	0	56	0	1	6
16	563	1	1	55	0	1	5
17	1106	0	0	44	0	0	-6
18	431	1	0	50	1	0	0
19	855	0	0	43	0	1	-7
20	803	0	0	39	0	0	-11
21	115	1	0	74	1	0	24
22	744	0	1	50	0	0	0
23	477	0	0	64	1	0	14
24	448	0	0	56	0	1	6
25	353	1	1	63	0	1	13
26	377	0	1	58	0	0	8

Age:	Age of patients at treatment
Treat:	1=cyclophosphamide, 2=cyclophosphamide + adriamycin
rdisease:	residual disease (0=partial, 1=fully excised)
Perf:	performance (0=good, 1=poor)
Age2:	(Age of patients at treatment -50)

We want to know how each or all of these variables if any, are related to survival times.

Start with age: Is age related to the survival experience of women?

Could specify the following model for the hazard for an individual i.

$$h_i(t) = h_0(t)e^{age_i \times \beta}$$

Where $h_0(t)$ is the hazard for a woman with age zero.

The problem then is to specify a full likelihood for this model.

Again, assume that the density of the survival times are exponential. Then, the hazard at the baseline is $h_0(t) = \lambda$ and for a person not at the baseline it is:

$$h_i(t) = \lambda e^{age_i \times \beta}$$

If this is the hazard, the the survivor function is given by:

$$S_i(t) = exp\{-\int_0^t \lambda e^{age_i \times \beta} dt\} = exp\{-\lambda e^{age_i \times \beta} t_i\}$$

and using equation (28) we can specify a log likelihood for the data as:

$$\begin{aligned} \log L(\theta; t, d, x) &= \sum_{i=1}^{n} [d_i \log h_i(t_i) + \log S_i(t_i)] \\ &= \sum_{i=1}^{n} \left[d_i \log \lambda + d_i\{x_i\beta\} - \lambda e^{x_i\beta} t_i\right] \end{aligned} \tag{30}$$

where θ is a vector of the unknown parameters, i.e. $\theta' = (\lambda\ \beta)$

This log likelihood is solved for the unknown parameters using any one of a number of techniques - the Newton-Raphson algorithm is often used.

It turns out that using the likelihood given in equation (30) in the N-R step is prone to failure because of the requirement that $\lambda > 0$.

There is a simple fix however - replace λ with e^{α} where $\log \lambda = \alpha$.

$$\log L(\theta; t, d, x) = \sum_{i=1}^{n} \left[d_i\alpha + d_i\{x_i\beta\} - e^{\alpha + x_i\beta} t_i\right]$$

The scalar version of N-R: given a function $f(z) = 0$ to find the roots, start with a guess and iterate updating each time using the formula;

$$z^{update} = z^{old} - \frac{f(z^{old})}{f'(z^{old})}$$

For example; What is the value of z such that $exp(-z^2) = z$? Then we have, $f(z) = z - e^{-z^2} = 0$ and $f'(z) = 1 + 2ze^{-z^2}$, so starting at a guess $z = 2$ we iterate to a solution,

z^{old}	$f(z^{old})$	$f'(z^{old})$	z^{new}
2	1.981684361	1.073262556	0.153588466
0.153588466	-0.823098172	1.300015606	0.786733306
0.786733306	0.24822335	1.847327436	0.652364409
0.652364409	-0.001026743	1.852498265	0.652918656
0.652918656	2.97708E-08	1.852605505	0.652918640

Convergence is reached where there is little change in the estimate. So, decide to stop the algorithm where e.g. there is no change correct to 5 decimal places in two successive iterations.

In the case of PH regression $f(z) = 0$ is derived from the log likelihood. The slope of the log likelihood evaluated at the maximum likelihood value of the parameters estimates should be zero - since it is a maximum. So the function we need to use is the first derivative of the log likelihood with respect of the unknown parameters.

These derivatives may be expressed in terms of the unknown parameters as a gradient vector (i.e. a vector of partial derivatives),

$$U = \begin{pmatrix} \frac{\partial \log L(\theta;y)}{\partial \alpha} \\ \\ \frac{\partial \log L(\theta;y)}{\partial \beta} \end{pmatrix} = 0$$

where y here is collection of data, i.e. (d, t, x). U plays a large role in PH regression models and is called the score vector.

$f'(z)$ is the derivative of this score vector - i.e. is the Hessian matrix of second partial derivatives,

$$U' = H = \begin{pmatrix} \frac{\partial^2 \log L(\theta;y)}{\partial \alpha^2} & \frac{\partial^2 \log L(\theta;y)}{\partial \alpha \partial \beta} \\ \\ \frac{\partial^2 \log L(\theta;y)}{\partial \beta \partial \alpha} & \frac{\partial^2 \ell(\theta;y)}{\partial \beta^2} \end{pmatrix}$$

This leads to the multivariate version of the scalar N-R scheme being applied to PH regression as follows;

Given some prior estimate of the parameters $\hat{\theta}^{(m-1)}$ calculate an update as,

$$\hat{\theta}^{(m)} = \hat{\theta}^{(m-1)} - {H^{(m-1)}}^{-1} U^{(m-1)}$$

Where the score and hessian matrices are evaluated at the $(m-1)^{th}$ estimates. Iterate using this scheme until there is little change in the parameter estimates.

This is the multivariable version of N-R.

In this case we get:

$$\begin{aligned}
\log L(\theta; y) &= \sum_{i=1}^{n} \left[d_i \alpha + d_i \{x_i \beta\} - e^{\alpha + x_i \beta} t_i \right] \\
\frac{\partial \log L(\theta; y)}{\partial \alpha} &= \sum_i d_i - \sum_i e^{\alpha + x_i \beta} t_i \\
\frac{\partial^2 \log L(\theta; y)}{\partial \alpha^2} &= -\sum_i e^{\alpha + x_i \beta} t_i \\
\frac{\partial \log L(\theta; y)}{\partial \beta} &= \sum_i d_i x_i - \sum_i x_i e^{\alpha + x_i \beta} t_i \\
\frac{\partial^2 \log L(\theta; y)}{\partial \beta^2} &= -\sum_i x_i^2 e^{\alpha + x_i \beta} t_i \\
\frac{\partial^2 \log L(\theta; y)}{\partial \alpha \partial \beta} &= -\sum_i x_i e^{\alpha + x_i \beta} t_i
\end{aligned}$$

So, get starting values e.g. $\alpha^0 = 1, \beta^0 = 0$

$$\begin{pmatrix}-0.001\\0\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix} - \begin{pmatrix}-42372.577 & -2209147.642\\-2209147.642 & -118375177.1\end{pmatrix}^{-1}\begin{pmatrix}-42360.577\\-2208415.642\end{pmatrix}$$

$$\begin{pmatrix}-1.005\\0\end{pmatrix} = \begin{pmatrix}-0.001\\0\end{pmatrix} - \begin{pmatrix}-15592.416 & -812969.366\\-812969.366 & -43564241.364\end{pmatrix}^{-1}\begin{pmatrix}-15580.416\\-812237.366\end{pmatrix}$$

$$\begin{pmatrix}-2.016\\0\end{pmatrix} = \begin{pmatrix}-1.005\\0\end{pmatrix} - \begin{pmatrix}-5740.547 & -299344.216\\-299344.216 & -16042844.601\end{pmatrix}^{-1}\begin{pmatrix}-5728.547\\-298612.216\end{pmatrix}$$

$$\vdots$$

$$\begin{pmatrix}-13.874\\0.118\end{pmatrix} = \begin{pmatrix}-13.874\\0.118\end{pmatrix} - \begin{pmatrix}-12 & -732.006\\-732.006 & -45525.409\end{pmatrix}^{-1}\begin{pmatrix}0\\-0.006\end{pmatrix}$$

$$\begin{pmatrix}-13.874\\0.118\end{pmatrix} = \begin{pmatrix}-13.874\\0.118\end{pmatrix} - \begin{pmatrix}-12 & -732\\-732 & -45525.093\end{pmatrix}^{-1}\begin{pmatrix}0\\0\end{pmatrix}$$

So the final answer is $\hat{\alpha} = -13.8741899$ and $\hat{\beta} = 0.1184563$.

This also gives $\hat{\lambda} = e^{-13.874} = 9.432 \times 10^{-7}$

How do we interpret these values?

- The lifespan of women at age 0 (birth) with ovarian cancer is exponentially distributed with $\hat{\lambda} = e^{-13.874}$. This gives an average lifespan of $1/\hat{\lambda} = 1,060,233$ days = 2,905 years! Clearly ridiculous but not surprising - no data for any women under 38 years - and no newborn has ovarian cancer...
- What is the projected mean and median lifespan for women with ovarian cancer aged 50 years?

$$\begin{aligned}\hat{\lambda}_i &= e^{\hat{\alpha}+(50)\hat{\beta}} = e^{-13.8741899+(50)0.1184563} = exp(-7.9514)\\ &\Rightarrow \text{ mean } = 1/\hat{\lambda}_i = 2,839 = 7.8 \text{ years}\\ &\Rightarrow \text{ median } = 1/\hat{\lambda}_i \log 2 = 5.4 \text{ years}\end{aligned}$$

So, the model is 'sensible' when we consider reasonable scenarios.

- What is the β parameter estimate? Compare the hazard for woman aged 50 years with a woman aged 51 years. Recall the hazard for a woman is λ_i, so we get the estimates:

50 year old:$\hat{\lambda}_1 = e^{-13.8741899+(50)0.1184563}$
51 year old:$\hat{\lambda}_2 = e^{-13.8741899+(51)0.1184563}$

and we can see that the estimated hazard ratio for the hazard for the 51 year old over the 50 year old is:

$$\frac{e^{-13.8741899+(51)0.1184563}}{e^{-13.8741899+(50)0.1184563}} = e^{0.1184563} = e^{\hat{\beta}}$$

Therefore, the parameter estimate for age is the hazard ratio for an increase in age of one year - NB. you will get the same result for ages 38 and 39, 73 and 74 etc.

- So, do not try to interpret $\hat{\lambda}$ directly, it is like the intercept in linear models - needed for the fit but not in general to be interpreted on its own.
- The β parameter (regression parameter or slope) is the increase in the hazard ratio for a unit increase in the covariate.

There are three issues/problems associated with the above:

(1). What if you want to see if the other covariates are important in terms of survival of women with ovarian cancer?

We, you can just extend the PH model with more regression parameter, e.g.
$h_i(t) = \lambda\, exp\{age_i \times \beta_1 + treatment_i \times \beta_2\}$

and in general for a set of covariates,
$h_i(t) = \lambda\, exp\{x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p\}$

where x_{i1} is the value of the 1st covariate for the i^{th} individual and x_{i2} s the value of the 2nd covariate and so on.

There is nothing new here, except the dimensions of the vectors and matrices in the NR step increases.

We can more compactly write the above as:

$$h_i(t) = \lambda \, exp\{x_i'\beta\} \tag{31}$$

where x_i is a $1 \times p$ vector of covariates for the individual and β is here taken to be a $p \times 1$ vector of unknown regression parameters.

(2) If fitting a statistical model we want to reach conclusions - by conducting hypothesis tests and constructing confidence intervals - how do we do this here?

We can use the result that at convergence the parameter estimates have a variance-covariance matrix which is:

$$\hat{Var}(\theta) = -H^{-1} \tag{32}$$

where H is evaluated at $\hat{\theta}$. Further it can be shown with standard asymptotic results that the parameter estimates are approximately normally distributed.

So, lets say you want to test $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ in the ovarian data above. First calculate an appropriate test statistic:

$$\text{Test statistic } : \frac{\hat{\beta} - 0}{\sqrt{Var[\hat{\beta}]}} = \frac{0.118}{\sqrt{0.00115}} = 3.48$$

and compare with a standard normal distribution: p-value = 0.0003

Nearly all interesting hypotheses can be done using extensions to this basic method.

(3) What if the assumption of an exponential distribution for the density of survival times is wrong?

One of the consequences of the choice of an exponential distribution is that the hazard function for a person over time is constant, i.e. $h_i(t) = \lambda_i$.

What about say, the increase in hazard after surgery? Or a decreasing hazard after a transplant? Or something more complicated?

We could choose another baseline distribution for the survival times, e.g. a very widely used distribution is the Weibull:

$$\begin{aligned} \text{Weibull: } f(t) &= \lambda\gamma t^{\gamma-1} e^{-\lambda t^{\gamma}} \\ S(t) &= e^{-\lambda t^{\gamma}} \\ h(t) &= \lambda\gamma t^{\gamma-1} \end{aligned} \tag{33}$$

The Weibull has increasing hazard over time for $\gamma > 1$, decreasing hazard for $\gamma < 1$ and constant hazard at $\gamma = 1$. A Weibull with $\gamma = 1$ is the exponential distribution. The parameter γ is called the shape parameter and λ the scale parameter.

We can even construct a PH regression model for the Weibull starting with:

$$h_i(t) = \lambda\gamma t^{\gamma-1} exp\{x_i'\beta\}$$

But the Weibull has limitations as well - the hazard in monotone increasing or decreasing.

Solution? One possible answer is to not use a parametric model, but to allow the hazard function to be arbitrary in shape - i.e. as simple or as complicated as it likes. This leads to the Cox PH model

Cox PH regression Model

Start with a PH regression Model and the PH assumption, then the HR for two sets of covariates is:

$$\frac{h_0(t)exp\{x_1'\beta\}}{h_0(t)exp\{x_2'\beta\}} = exp\{(x_1' - x_2')\beta\}$$

In particular,

$$\frac{h_0(t)exp\{x_1'\beta\}}{h_0(t)} = exp\{(x_1')\beta\}$$

So, if the hazard ratio for a particular set of covariates is modelled relative to the baseline hazard (all covariates zero) then the baseline hazard cancels.

Note that an intercept cannot be included in this model.

The advantage? - We can still get parameter estimates for the regression parameters, while the distribution of baseline hazard can be arbitrary - as it cancels from our HR model.

The problem? Before (exponential model) the likelihood involved $h_0(t)$ and we need the likelihood to fit the model.

The solution? Cox (1972, 1975) in two seminal papers gave the following likelihood.

- The likelihood is proportional to the probability of the data under the assumed model.
- Consider survival data and one death time. What is the probability that individual i dies at this time conditional on there being one death at this time?

Well, $P(A|B) = \frac{P(A \cap B)}{P(B)}$ so,

$$\frac{P(\text{person } i \text{ dies at time } t_j \cap \text{one death at time } t_j)}{P(\text{there is one death at time } t_j)}$$

$$= \frac{P(\text{person } i \text{ dies at time } t_j)}{P(\text{there is one death at time } t_j)} \quad (34)$$

Consider a small period of time $(t_j, t_j + \delta t)$. Replace t_j in equation (34) with this interval and divide both numerator and denominator by δt:

$$\frac{P(\text{person } i \text{ dies in interval } (t_j, t_j + \delta t))/\delta t}{P(\text{there is one death in interval } (t_j, t_j + \delta t))\delta t} \tag{35}$$

Now, take the limit of equation (35) as $\delta t \to 0$

$$\frac{\text{hazard for person } i \text{ at time } t_j}{\sum_{l \epsilon R(t_j)} \text{hazard for person } l \text{ at time } t_j} \tag{36}$$

where $R(t_j)$ is the risk set of individuals at time t_j, i.e. the set of individuals alive in the instant just before the death at time t_j.

Using equation (36) and the PH assumption we can write:

$$\frac{h_0(t)e^{x_i'\beta}}{\sum_{l \epsilon R(t_j)} h_0(t)e^{x_l'\beta}} = \frac{e^{x_i'\beta}}{\sum_{l \epsilon R(t_j)} e^{x_l'\beta}} \tag{37}$$

Which is proportional to the probability that person i dies at time t_j conditional on there being a death at time t_j.

The likelihood for the survival data is therefore (assuming independence between death times):

$$L(\beta; y) = \prod_{j=1}^{r} \frac{e^{x_{(j)}'\beta}}{\sum_{l \epsilon R(t_j)} e^{x_l'\beta}} \tag{38}$$

where the product is taken over the $j = 1, \ldots, r$ distinct death times and $x_{(j)}$ is the vector of covariates for the individual who dies at time t_j.

Using the death indicator (d_i) it is possible to rewrite equation (38) is the product over individuals rather than death times:

$$L(\beta; y) = \prod_{i=1}^{n} \left[\frac{e^{x_i'\beta}}{\sum_{l \epsilon R(t_i)} e^{x_l'\beta}} \right]^{d_i} \tag{39}$$

where $R(t_i)$ is the risk set at the survival time for the i^{th} individual.

The corresponding log likelihood is:

$$\log L(\beta; y) = \sum_{i=1}^{n} d_i \left[x_i'\beta - \log \left\{ \sum_{l \epsilon R(t_i)} e^{x_i'\beta} \right\} \right] \tag{40}$$

Notice that this will only have non-zero values for the $j = 1, \ldots, r$ death times.

This is called the Cox PH model log likelihood, and each term is called a partial likelihood. The MLE's for β is found using NR as before.

Notice also, that equation (40) is not a function of time.

It only involves the order of the death times and the risk sets at those death times.

The gap between death times is not relevant here because of the arbitrary nature of the hazrad function for the Cox model - between death times conceivably it could be zero.

The Cox model therefore, only updates at each death time and other information is not relevant.

A note of Cox PH model

This development is like that given by Cox in his original papers.

A more formal justification for the model is to view the Cox model as a profile likelihood. This puts the Cox PH model in the framework of a parametric model as discussed above.

For a discussion of profile likelihood and its application to the Cox PH model see:

Pawitan, Y. In all likelihood: statistical modelling and inference using likelihood. Oxford University Press, Oxford, 2001.

Fitting the Cox PH model

The procedure is very similar to fitting a parametric model.

Use the multivariate form of the NR algorithm by calculating the score function and hessian of the partial likelihoods from the Cox model and iterate to solution.

EG. the ovarian cancer data: fit a model with both age and treatment as covariates.

The model is:

$$h_i(t) = h_0(t)e^{age_i\beta_1 + treatment_i\beta_2}$$

So, we are modelling the hazard.

More particularly we actually model the hazard ratio between the baseline hazard and the hazard for a woman with particular values of the covariates, i.e.

$$\frac{h_i(t)}{h_0(t)} = e^{x_{i1}\beta_1 + x_{i2}\beta_2}$$

where x_{i1} is the age of woman i and x_{i2} is the treatment to which woman i is assigned, which is coded,
1=cyclophosphamide
2=cyclophosphamide + adriamycin

The partial likelihood is:

$$\log L(\beta; y) = \sum_{i=1}^{n} d_i \left[x_{i1}\beta_1 + x_{i2}\beta_2 - \log \left\{ \sum_{l \epsilon R(t_i)} e^{x_{l1}\beta_1 + x_{l2}\beta_2} \right\} \right]$$

The score vector (S) is:

$$\begin{pmatrix} \frac{\partial \log L(\beta;y)}{\partial \beta_1} \\ \frac{\partial \log L(\beta;y)}{\partial \beta_2} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} d_i \left\{ x_{i1} - \frac{\sum_{l \epsilon R(t_i)} x_{l1} e^{x_{l1}\beta_1 + x_{l2}\beta_2}}{\sum_{l \epsilon R(t_i)} e^{x_{l1}\beta_1 + x_{l2}\beta_2}} \right\} \\ \sum_{i=1}^{n} d_i \left\{ x_{i2} - \frac{\sum_{l \epsilon R(t_i)} x_{l2} e^{x_{l1}\beta_1 + x_{l2}\beta_2}}{\sum_{l \epsilon R(t_i)} e^{x_{l1}\beta_1 + x_{l2}\beta_2}} \right\} \end{pmatrix}$$

The hessian matrix takes the form:

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$$

where:

$$H_{11} = \sum_{i=1}^{n} d_i \left\{ \frac{\left[\sum_{l \epsilon R(t_i)} x_{l1} e^{x_{l1}\beta_1 + x_{l2}\beta_2}\right]^2}{\left[\sum_{l \epsilon R(t_i)} e^{x_{l1}\beta_1 + x_{l2}\beta_2}\right]^2} - \frac{\sum_{l \epsilon R(t_i)} x_{l1}^2 e^{x_{l1}\beta_1 + x_{l2}\beta_2}}{\sum_{l \epsilon R(t_i)} e^{x_{l1}\beta_1 + x_{l2}\beta_2}} \right\}$$

$$H_{12} = H_{21} = \sum_{i=1}^{n} d_i \left\{ \frac{\sum_{l \epsilon R(t_i)} x_{l1} e^{x_{l1}\beta_1 + x_{l2}\beta_2} \sum_{l \epsilon R(t_i)} x_{l2} e^{x_{l1}\beta_1 + x_{l2}\beta_2}}{\left[\sum_{l \epsilon R(t_i)} e^{x_{l1}\beta_1 + x_{l2}\beta_2}\right]^2} - \frac{\sum_{l \epsilon R(t_i)} x_{l1} x_{l2} e^{x_{l1}\beta_1 + x_{l2}\beta_2}}{\sum_{l \epsilon R(t_i)} e^{x_{l1}\beta_1 + x_{l2}\beta_2}} \right\}$$

$$H_{22} = \sum_{i=1}^{n} d_i \left\{ \frac{\left[\sum_{l \epsilon R(t_i)} x_{l2} e^{x_{l1}\beta_1 + x_{l2}\beta_2}\right]^2}{\left[\sum_{l \epsilon R(t_i)} e^{x_{l1}\beta_1 + x_{l2}\beta_2}\right]^2} - \frac{\sum_{l \epsilon R(t_i)} x_{l2}^2 e^{x_{l1}\beta_1 + x_{l2}\beta_2}}{\sum_{l \epsilon R(t_i)} e^{x_{l1}\beta_1 + x_{l2}\beta_2}} \right\}$$

Starting at an initial estimate ($\beta_1 = 0, \beta_2 = 0$) we iterate as follows:

$$\beta^{new} = \beta^{old} - H^{old^{-1}} S^{old}$$

$$\begin{pmatrix} 0.158 \\ -1.583 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} -814.991 & -18.231 \\ -18.231 & -2.936 \end{pmatrix}^{-1} \begin{pmatrix} 100.001 \\ -1.766 \end{pmatrix}$$

$$\begin{pmatrix} 0.145 \\ -0.684 \end{pmatrix} = \begin{pmatrix} 0.158 \\ -1.583 \end{pmatrix} - \begin{pmatrix} -470.467 & 2.993 \\ 2.993 & -2.073 \end{pmatrix}^{-1} \begin{pmatrix} -9.063 \\ 1.905 \end{pmatrix}$$

$$\begin{pmatrix} 0.147 \\ -0.795 \end{pmatrix} = \begin{pmatrix} 0.145 \\ -0.684 \end{pmatrix} - \begin{pmatrix} -491.242 & 6.429 \\ 6.429 & -2.605 \end{pmatrix}^{-1} \begin{pmatrix} 1.672 \\ -0.302 \end{pmatrix}$$

$$\begin{pmatrix} 0.147 \\ -0.796 \end{pmatrix} = \begin{pmatrix} 0.147 \\ -0.795 \end{pmatrix} - \begin{pmatrix} -489.729 & 5.999 \\ 5.999 & -2.57 \end{pmatrix}^{-1} \begin{pmatrix} 0.025 \\ -0.002 \end{pmatrix}$$

So, we get convergence in 4 iteration here.

The final estimates are: $\hat{\beta}_1 = 0.147$, $\hat{\beta}_2 = -0.796$.

We can also get the variance-covariance matrix for the parameter estimates as $-H^{-1}$ where H is calculated at the values of the parameters at convergence, i.e.

$$-\begin{pmatrix} -489.729 & 5.999 \\ 5.999 & -2.57 \end{pmatrix}^{-1} = \begin{pmatrix} 0.0021 & 0.0049 \\ 0.0049 & 0.4006 \end{pmatrix}$$

$$= \begin{pmatrix} Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) \\ Cov(\hat{\beta}_1, \hat{\beta}_2) & Var(\hat{\beta}_2) \end{pmatrix}$$

We can the conduct hypotheses as follows:

$$z = \frac{\hat{\beta} - \beta^0}{s.e.\hat{\beta}}$$

to test the $H_0 : \beta = \beta^0$, where approximately

$$z \sim N(0, 1)$$

under the assumption that the null hypothesis is true.

An alternative version of the same test is:

$$\chi = \frac{(\hat{\beta} - \beta^0)^2}{Var(\hat{\beta})}$$

where approximately $\chi \sim \chi^2_1$.

For the two parameters in the example we get:

$$z_1 = \frac{0.147 - 0}{\sqrt{0.0021}} = 3.21$$

$$z_2 = \frac{-0.796 - 0}{\sqrt{0.4006}} = -1.26$$

Using $\alpha = 0.05$ we reject the H_0 : in the case of β_1 but fail to reject in the case of β_2.

Treatment of ties in Cox PH model

The likelihood equations we have just used to fit the model are based on continuous time models - i.e. ties in death times are not possible.

But in real data we can have tied recorded death and censored times (to the nearest day, week etc.).

We need to amend the above partial likelihoods to handle times.

(1) death time tied with one or more censored times?
- Easy, just assume that the censoring occurs after the death time in all cases.
(2) two or more death times tied?
- need to question what the probability mechanism might be.

Imagine that two individuals have the same death time (t_j) recorded. Call them a and b with covariate vectors x_a and x_b respectively

The partial likelihoods for these deaths either:

$$\frac{e^{x'_a\beta}}{\sum_{l \epsilon R(t_j)} e^{x'_l\beta}} \times \frac{e^{x'_b\beta}}{\sum_{l \epsilon R(t_j)} e^{x'_l\beta} - e^{x'_a\beta}}$$

OR

$$\frac{e^{x'_b\beta}}{\sum_{l \epsilon R(t_j)} e^{x'_l\beta}} \times \frac{e^{x'_a\beta}}{\sum_{l \epsilon R(t_j)} e^{x'_l\beta} - e^{x'_b\beta}}$$

And we don't know which.....

The question can be put in terms of the denominator patterns.

Lets say that at the tied death time there were 5 individuals at risk and denote $e^{x_i'\beta} = r_i$ for $i = 1, \ldots, 5$.

NB. r_i is called the risk score for person i.

Further assume that person 1 and person 2 have the tied death times.

Then we either have:

$$\frac{r_1}{r_1+r_2+r_3+r_4+r_5} \times \frac{r_2}{r_2+r_3+r_4+r_5}$$

OR

$$\frac{r_2}{r_1+r_2+r_3+r_4+r_5} \times \frac{r_1}{r_1+r_3+r_4+r_5}$$

NB. the numerators remain the same.

There are a number of proposed solutions - we will look at three.

Discrete option

The probability that two individuals die from the risk set should be divided by sum of every possible combination of two individuals. In our case this is:

$$\frac{r_1 r2}{\begin{array}{c} r_1r_2 + r_1r_3 + r_1r_4 + r_1r_5 + r_2r_3 + r_2r_4 + r_2r_5 \\ +r_3r_4 + r_3r_5 + r_4r_5 \end{array}}$$

The denominator calculated here therefore is the sum over all possible combinations of d_j individuals in the risk set.

Problem: this can become very computationally expensive: e.g. given 5 tied death times among 300 at risk this is 1.958×10^{10} combinations and that's just for one of the partial likelihoods!

Breslow Method (also called after Peto):

Keep the denominators the same -

$$\frac{r_1}{r_1+r_2+r_3+r_4+r_5} \times \frac{r_2}{r_1+r_2+r_3+r_4+r_5}$$

$$= \frac{r_1 r_2}{(r_1+r_2+r_3+r_4+r_5)^2}$$

This is the simplest method to program and therefore also the fastest. Traditional default on many computer packages, e.g. SAS.

Can lead to biased results when the number of ties is large.

Efron Mehtod:

Use a sort of 'averaged' denominator for second or subsequent term in the partial likelihood.

In our example: each of $(r_1,\ r_2)$ have a 0.5 probability of being in the second deminator - so weight them accordingly:

$$\frac{r_1}{r_1+r_2+r_3+r_4+r_5} \times \frac{r_2}{\frac{1}{2}r_1+\frac{1}{2}r_2+r_3+r_4+r_5}$$

If there are more ties - imagine $(r_1,\ r_2,\ r_3)$ all tied then:

$$\frac{r_1}{r_1+r_2+r_3+r_4+r_5} \times \frac{r_2}{\frac{2}{3}r_1+\frac{2}{3}r_2+\frac{2}{3}r_3+r_4+r_5} \times \frac{r_3}{\frac{1}{3}r_1+\frac{1}{3}r_2+\frac{1}{3}r_3+r_4+r_5}$$

and so on...

There is also an Exact method which we will not discuss...

Interpretation of Model Parameters

For the ovarian data we got the following parameter estimates; $\hat{\beta}_1 = 0.1478$ and $\hat{\beta}_2 = -0.796$.

How do we interpret these parameters?

(1) Age: The parameter gives the effect of a person's age on the hazard of death:
i.e. $h_i(t) = h_0(t)e^{age_i\beta_1}$ - where the baseline hazard is not estimated (yet) and has a arbitrary form. Under this formulation the baseline hazard is the hazard for a person with a zero value of the covariate (i.e. age=0).

The parameter is also the increase in the log hazard ratio for a unit increase in age:

$$\log \frac{h_0(t)e^{(x+1)\beta}}{h_0(t)e^{(x)\beta}} = \log(e^{\beta}) = \beta$$

So, for a 1 year increase in age, a persons hazard increases by $(e^{0.1478} - 1) \times 100\% = 16\%$ and for an increase in age of 10 years it is $(e^{1.478} - 1) \times 100\% = 338\%$.

So, the hazard of death for a 70 year old is 3.38 times that of a 60 year old.

This interpretation however only works where the values of any other covariates are assumed to be the same.

You can if you like recode continuous covariates like age to make the baseline hazard more meaningful - e.g. in the ovarian data use age2=age-50.

In this case baseline person is a 50 year old instead of a newborn.

However, this is not necessary as the parameter estimates remain unchanged.

- important just to know what the definition of the baseline is for any fitted model.

(2) Treat: The covariate is coded 1 or 2. With this coding the same basic interpretation occurs as with continuous covariates.

The parameter is the log hazard ratio for a woman in the treatment=2 group over the a woman in the treatment=1 group (of the same age!).

e.g. take two women both aged 50, one in treatment 1 and one in treatment 2. The hazard ratio is:

$$\frac{h_0(t)exp\{0.147 \times 50 - 0.796 \times 2\}}{h_0(t)exp\{0.147 \times 50 - 0.796 \times 1\}} = e^{-0.796} = 0.45$$

...and this will work with any ages once they are both the same.

So the hazard for the treatment 2 group is less than half that of the treatment 1 group for women of the same age.

Not that this coding of a categorical (grouping) variable only works where there are 2 groups.

If there are more than two groups then we must use dummy (or indicator) variables. It is also common practice to use these for the two group situation as well.