Déjà

Regression Models (semiparametric)

- Model (event) ***rates*** or ***hazards*** p ***h*** reg
- Models are multiplicative in rates/hazards
  linear predictors if work in log[rate] or log[hazard] scale
- Proportionality of rates or hazards ***p h*** reg
  Constancy of rate ratio parameter over time-bands
- Avoid modelling the nuisance parts
  don't fit parameters that (a) are not our focus (b) waste "d.f."
- Use risksets & conditioning to reduce # parameters
- Choice of Time-scale and "Time-zero" is important
  (has implications for risksets)
- Models, and conditioning as a way of eliminating parameters, applicable to matched case-control studies and even to c-c and other (e.g. consumer choice*) studies with no 'time' element

---

Homework: Fruitfly longevity / Clayton Hills Exercises Ch 30

Other Resources

- Texts [http://www.epi.mcgill.ca/hanley/c681/cox]
  Kleinbaum's 'Self-Learning' textbook, Chapter 3/4
  Collett Textbook, Chapter 3/4

[http://www.epi.mcgill.ca/hanley/c681/cox]

New

Regression Models (semiparametric)

- Proportional hazards model
- Relationship b/w S[t] for z=1 and S[t] for z=0 [corner]
  one is a constant power of the other
- Two log[-log[S]] functions should be parallel
  - uses relationship S[t] = exp[-H[t]]
  - H[t] is the integrated or "cumulative" hazard
  - -log[S] = H[t], so $-\log[S_1[t]] = HR \times \{ -\log[S_0[t]] \}$
    2 -log[S] curves should be proportional (easier to judge parallel)
  - hazard functions may not be stable enough
    (so cannot assess whether 2 h[t] curves are proportional)
- Fitting proportional hazards model to data
- estimating HR by (Partial) Likelihood approach
- **"**Information": how sharp is curvature of LogL fn
- Estimating HR via SAS PROC PHREG/Stata
- Estimating $h_0(t)$ and $S_0(t)$ [the "corner"]
- Estimating $S_{\underline{X}}(t)$ [$\underline{X}$ = a specified covariate pattern]
- Split Records
  (also a way to handle time-dep. covariates)
- ML estimation for stratified survival data

Readings

[http://www.epi.mcgill.ca/hanley/c681/cox]

Clayton&Hills, Ch 30, sections 4-6

Pair of expository articles by JH

Proportional hazards model

Simplest case (1 covariate z, 2 levels/groups which we will refer to as 0 and 1)

Compared with reference individuals (group 0), who have a hazard $h_0[t]$ at time t, those in group 1 have a hazard that is a constant times $h_0[t]$, i.e.

$$\frac{h_1[t]}{h_0[t]} = \text{constant}$$

{Selvin uses 'c' and Collett uses 'ψ' for HR, the hazard ratio}.

Equivalently, one can write

$$h_1[t] = HR \bullet h_0[t]$$

The hazard ratio HR will be a number between 0 and infinity. To make it easier to fit this parameter without having to constrain it within these bounds, it helps to re express HR as

$$HR = e^{\beta} \quad \{ \text{ or } \ln[HR] = \beta \}$$

so that the model becomes

$$h_1[t] = e^{\beta} \bullet h_0[t]$$

or

$$\ln[h_{z=1}[t]] = \ln[h_{z=0}[t]] + \beta \bullet (z=1) .$$

One can think of the $\ln[h_0[t]]$ as the intercept and z as the indicator variable for group in a regression. Note that the 'intercept' here is a full hazard curve over t; Unlike the case of other regressions, here the intercept may be of interest. However we may not have enough data to estimate it well, especially if, as is often the case, it varies considerably over t, or we do not have many events.

Relationship between S[t] for z=1 versus S[t] for z=0 ["corner"]

If $h_1(t) = e^{\beta} h_0(t)$, and if H[t] is the integrated hazard,

then the integrated (or "cumulative") hazard for z=1 is

$$H_1(t) = e^{\beta} H_0(t),$$

so that the survival functions are

$$S_1(t) = e^{-H_1(t)}$$

$$= e^{-e^{\beta} H_0(t)} \quad = \left[e^{-H_0(t)} \right]^{e^{\beta}}$$

$$= \left[S_0(t) \right]^{e^{\beta}} \quad = \left[S_0(t) \right]^{HR}$$

Thus, the **log[-log]** functions should be **parallel**,

$$\log[-\log[S_1(t)]]$$

$$= \log\left[-\log\left[S_0(t)^{HR} \right] \right]$$

$$= \log\left[HR \bullet -\log\left[S_0(t) \right] \right]$$

$$= \log\left[HR \right] + \log\left[-\log\left[S_0(t) \right] \right]$$

and **separated by the quantity log[HR]**.

Thus one can **visually estimate** b = log [HR] from log[-log S] plots. If limited data, the hazard functions may be too unstable to use.

The "baseline" hazard function $h_0(t)$ can be from some parametric family [e.g. $h_0(t)$ = constant {negative exponential distribution of failure times}, Weibull, ...] or can be unspecified. In the latter case, the mixture of a parametric form for HR and a 'free' form for $h_0(t)$ is why the model is called "semi-parametric".

More general case (1 covariate z, with possibly several levels or possibly continuous; or several covariates, continuous/discrete/mixed}

For short, refer to set of covariates $\{z_1, z_2, ..., z_k\}$ as $\mathbf{z}$; without loss of generality, refer to a reference group of individuals as having$\{z_1=0, z_2=0, ..., z_k=0\}$ as $\mathbf{z}=\mathbf{0}$).

Compared with reference individuals (group with $\mathbf{z}=\mathbf{0}$), who have a hazard $h_0[t]$ at time t, those with covariate values $\{z_1, z_2, ..., z_k\}$ have a hazard that is some multiple times $h_0[t]$, where the multiple depends only on $\mathbf{z}$ i.e.

$$\frac{h_{\mathbf{z}}[t]}{h_{\mathbf{0}}[t]} = HR(\mathbf{z})$$

or

$$h_{\mathbf{z}}(t) = HR(\mathbf{z}) \bullet h_{\mathbf{0}}(t)$$

Most often, $HR(\mathbf{z})$ is taken as log-linear i.e. the log of $HR(\mathbf{z})$ is taken as linear in the k parameters $\{\beta_1, \beta_2, ..., \beta_k\}$ i.e.

$$\log[\, HR(\mathbf{z}) \,] = \{\beta_1 z_1 + \beta_2 z_2 + \ ... + \beta_k z_k\}$$

or

$$HR(\mathbf{z}) = \exp\{\beta_1 z_1 + \beta_2 z_2 + \ ... + \beta_k z_k\}.$$

Since

$$\exp\{\beta_1 z_1 + \beta_2 z_2\} = \exp\{\beta_1 z_1\} \bullet \exp\{\beta_2 z_2\},$$

we can rewrite model as

$$h_{\mathbf{z}}(t) = HR(z_1) \bullet HR(z_2) \bullet ... \bullet HR(z_k) \bullet h_{\mathbf{0}}(t)$$

or

$$S_{\mathbf{z}}(t) = [S_0(t)]^{e^{\beta \mathbf{z}}} = [S_0(t)]^{HR_1 + HR_2 + ... + HR_k}$$

where HR_1 is shorthand for $\exp\{\beta_1 z_1\}$, same for HR_2 etc.

Important to have the "corner" covariate pattern near the actual **z** values (so, might want to **'center'** the **z** values first, *before* fitting.

Not precluded from using **products** or **powers** of the z's.

Fitting proportional hazards model: Risksets

Our prime interest is in estimating the parameters of HR; we will also, as a secondary objective, estimate $h_{\mathbf{0}}(t)$. The keys to the estimation are the Risk Sets, the collections of candidates for (individuals at risk just before) each distinct failure time (event)

Simplest case (1 covariate z, 2 levels or Tx groups which we will distinguish using indicator variable z= 0 and z=1). In e.g. below, a • denotes a failure (event), a + denotes a censored observation; and time runs from left to right [*note*: to estimate HR function we do not need the failure & censoring times themselves, only their *order* with respect to z].

Raw data...

```
z=1   •   +       •           •
z=0           •       +           •
```

It is easier to lay them out as separate time lines [in the 'early days' before computers, some investigators would represent survival data on their patients using lines of thread along a wall].

```
z=1 --•
z=1 ------+
z=1 ----------•
z=1 ----------•----------•

z=0 ----------•
z=0 ----------•----------+
z=0 ----------•----------•----•
Riskset #  1       2   3          4    5
```

Cox argued that since there are no failures (events) between the •'s, we do not know much about the hazards in these gaps [unless we want to posit parametric form for $h_{\mathbf{0}}(t)$ or $S_{\mathbf{0}}(t)$]. In any case our prime interest is in HR, and so we will concentrate just on these risk sets.

Estimating HR by (Partial) Likelihood approach

It helps to lay it out the 5 risk sets as follows (note that in the 5th riskset there is 'no contest') ...

$o=d_1$	**1**	0	**1**	**1**	-
s_1	3	2	1	0	-
n_1	4	2	2	1	
d_0	0	**1**	0	0	**1**
s_0	3	2	2	1	0
n_0	3	3	2	1	1

In the Maximum Likelihood method, we find that value of the HR which maximizes the likelihood of the **observed data pattern** (the **sequence** is indicated in **bold** above) The likelihood is a function of HR. To construct it, we need a probability model for each table (ie for the outcome in each riskset) and an assumption regarding the separate tables. In the calculation of a variance for the MH statistic (log rank test) we already assumed that the 2x2 tables were realizations of hypergeometric (urn sampling) models and that the tables could be treated as if they were independent of each other. We could do the same here to set up a likelihood.

For each risk set, we ask

> ***"Given that the event occurred, what is the chance that it ocurred to the individual it happened to, rather than to someone else in the risk set?"***

Consider a risk set where the event happened at t to a person with z=1.

If the hazard for persons with z=1 is $HR \cdot h_0(t)$ and $1 \cdot h_0(t)$ for those with z=0, and if in the risk set there are n_1 and n_0 persons respectively, then the [conditional] probability that the event happened to that particular person with z=1 out of the n_1 and n_0 'at risk' is

$$\frac{HR \cdot h_0[t]}{n_1 \cdot HR \cdot h_0[t] \quad + \quad n_0 \cdot 1 \cdot h_0[t]}$$

which simplifies to

$$\frac{HR}{n_1 \cdot HR \quad + \quad n_0 \cdot 1}$$

Conversely, in a risk set where the event happened to a person with z=0. then the [conditional] chance that the event happened to that particular person with z=0 out of the n_1 and n_0 'at risk' is

$$\frac{1}{n_1 \cdot HR \quad + \quad n_0 \cdot 1}$$

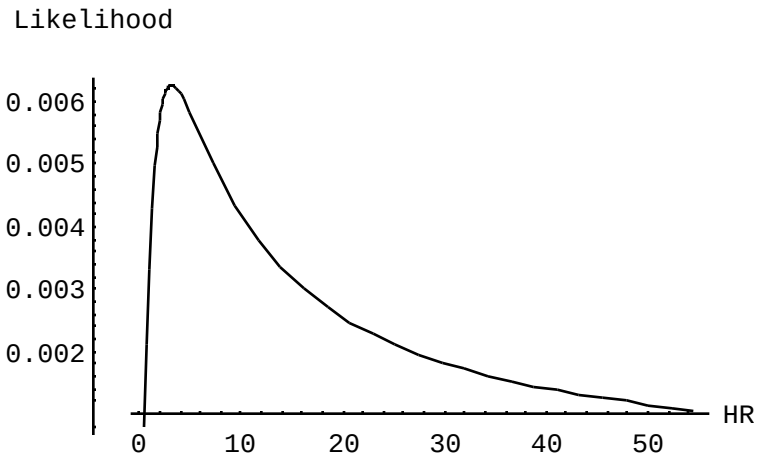
Thus, for the example above, the product of the probabilities of the observed outcome (likelihood) in each of the 4 informative risksets is

$$L = \frac{HR}{4HR+3} \cdot \frac{1}{2HR+3} \cdot \frac{HR}{2HR+2} \cdot \frac{HR}{HR+1}$$

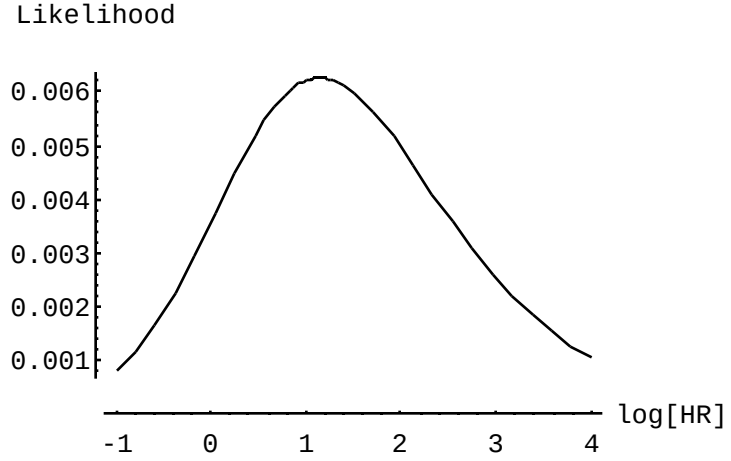
This likelihood L(HR)= prob(data | HR) can be evaluated for a range of HR values in order to find the value $\hat{HR}_{ML}$ which mazimises L. e.g.

HR	1/2	1	2	4	8	16
$Lx10^3$	1.4	3.6	5.8	6.1	4.8	3.0

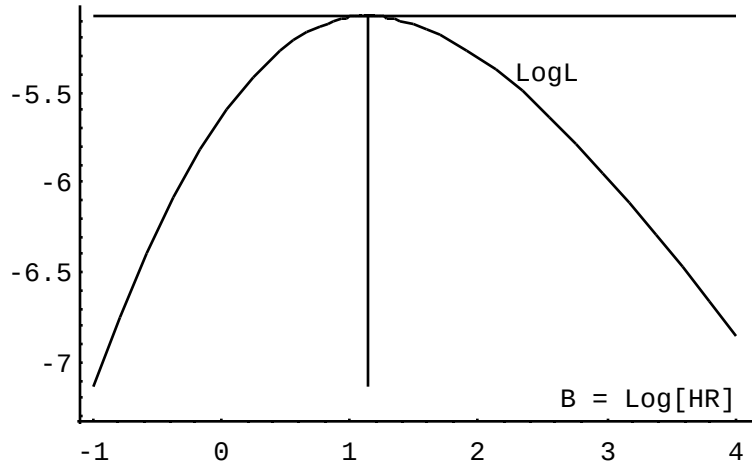
The function L & derived functions are shown graphically on next page.



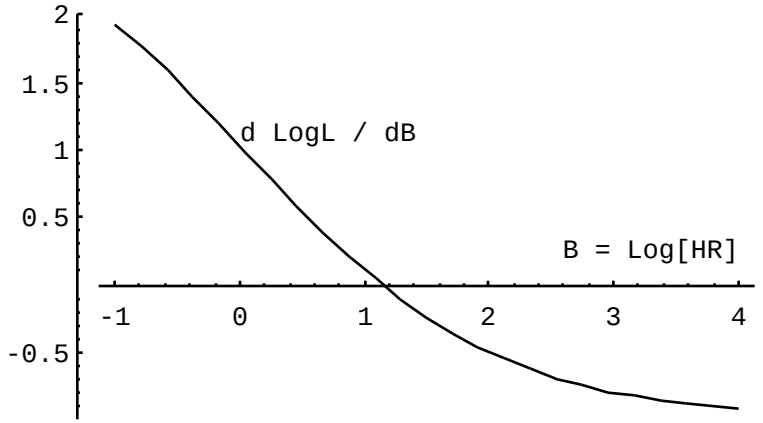
Or with the parameter B = Log[HR] ...



or in the log Likelihood scale...



The Derivative of the log Likelihood ...



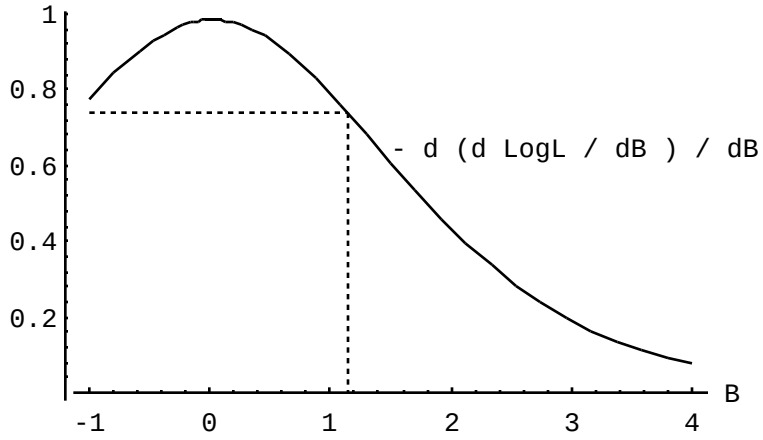
Tangent to logL curve is zero at B = 1.14 (we call this B_hat or b];

So... $\hat{HR}_{ML} = \exp[b] = 3.14$.

Uncertainty / Information concerning log [HR]

The 'sharpness' or 'flatness' of the logL(HR) curve in the vicinity of B = 1.14 gives an indication of how sensitive logL is to changes in log[HR] i.e. of how well or badly other values of log[HR] would do in producing a large likelihood. This can be measured by the 2nd derivative of logL (or if you like by the tangent to the 1st derivative curve) with respect to B. Note that the L curve increases until B = 1.14 then decreases. Thus the slope dlogL/dB goes from positive to negative over this range. ie the 2nd derivative is negative. Since we are simply interested in the curvature we use the negative of the 2nd derivative; it will be a big positive quantity when the curvature is very sharp, and a small positive quantity when the curvature is very slow.

The plot below shows that the curvature of logL is quite small (approximately 0.7412 at B = 1.14). This negative of the 2nd derivative of the log likelihood, evaluated at the ML estimate, is called the "**Information**" in the data. Its reciprocal is a good measure of the variance of the ML estimate of B.



We usually work with B =log[HR], since the sampling variability of b is more symmetric. The I[B] calculated at b = 1.14 is approximately 0.7412, yielding SE[b] = √(1/0.7412) = 1.16, yielding a 95% CI for HR=exp[B] of {0.3 to 31}.The 4 informative risk sets provide just a small amount of information about log[HR] and our confidence in values near the ML estimate is low.

Estimating HR via SAS PROC PHREG (Stata below)

```
DATA a;
INPUT event  time   tx ;  /* Note arbitrary times             */
LINES;                    /* only ORDER matters               */
       1     2     1     /* event=0 stands for censored obsn. */
       0     4     1
       1     6     0
       1     8     1
       0    10     0
       1    12     1
       1    14     0
;
title null model; proc phreg data = a   ;
                  model time*event(0) = ;
```

```
Dependent Variable: TIME       Number of Event & Censored Values
Censoring Variable: EVENT
Censoring Value(s):            Total   Event  Censored  %Censored
Ties Handling:      BRESLOW        7       5         2    28.57

NOTE: No explanatory variables in this model.  -2 LOG L = 11.27

JH: LOG L = log[{1/7} x {1/5} x {1/4} x {1/2}] = LOG[1/280] = -5.63
```

```
title model with tx; proc phreg data=a;
                     model time*event(0) = tx / RISKLIMITS;

    Testing Global Null Hypothesis: BETA=0

        Without   With Covariates        Model Chi-Square

-2 LOG L  11.27   10.15             1.12  with 1 DF (p=0.29)

ML Estimates
          Parameter Standard  Wald    Pr >     Risk*      95% CL
Variable  Estimate   Error  Chi-Sq  Chi-Sq    Ratio    Lower  Upper
TX         1.14      1.16** 0.9685   0.33     3.14      0.32    30.6
```

* Technically speaking, should be called Hazard Ratio; Obtained as exp[1.14]

** See 2nd Derivative graph on left: SE[b] = Sqrt[var] = sqrt[1/Information]

Stata (after input, using same variable names as above)

```
stset time , failure(event)
* null model
stcox, estimate
* model with tx .. gives beta_hats, not HR_hats
stcox tx, nohr
* model with tx .. gives HR_hats, not beta_hats
stcox tx
```

Estimating $h_0(t)$ and $S_0(t)$ [see Collett §3.8]

Once one has estimated HR, using $\hat{HR}_{ML} = \exp[\hat{\beta}_{ML}]$, one can estimate the baseline hazard (and S) via a procedure similar to the Kaplan-Meier product method. One might have expected this type of non-parametric approach, since no form is specified for $h_0(t)$.

One uses all the events (in both groups) even though the estimate is supposed to represent individuals with z=0. The reason for this is that in a dataset with continuous covariates, there may be nobody with the specific configuration of z's that one considers the 'reference' population.

As with all modelling and regression, we are being 'synthetic' and borrowing strength from all the data. As Collett explains, the derivation is complex, but one can get some sense of the logic from the 2-sample case where there is one event at a time [see Collett's 'particular case' following his equation 3.16]. The way JH thinks of it is to imagine a two sample situation where we were given 2 samples of death times, 1 for males and 1 for females (reference group), and told that the ratio (HR) of the death rates in the population of males and females was say 2. Would you just estimate a K-M curve for females using the data for females and call it your best estimate of the 'reference' of female S or would you try to use all the data, including the deaths from males, to estimate a better K-M curve for females?

Cox [and later Kalbfleisch and Prentice] take the 'synthetic' approach. One estimates a quantity Collett calls ξ for each riskset. This is the the 'conditional probability of survival'; estimates of these various success probabilities are multiplied together to give the unconditional probability ξ_1 of surviving past the time of the 1st riskset, $\xi_1\xi_2$ for surviving past the 2nd, etc as in the K-M approach.

If there are multiple events per riskset, one must iteratively solve equation 3.16 for ξ_1. If there is only one, ξ_1 can be calculated directly as

$$\hat{\xi} = \left\{ 1 - \frac{\mathbf{HR}}{\sum \mathrm{HR}} \right\}^{1/\mathbf{HR}}$$

where **HR** is *the calculated HR for the individual who suffered the event,* and the summation of the HR's for all the persons in the risk set.

To go back to our example of males and females and a HR of 2 for males relative to a "1" for females: suppose the risk set had 100 men and 50 women. From a hazard point of view, one can think of this as

$$100 \times 2 \; + \; 50 \times 1 = 250 \text{ "}\textit{women equivalents}\text{"}$$

at risk. Now if the one event occurs to a woman, that is like saying that we had a failure of 1/250 and thus

$$\hat{\xi} = \left\{ 1 - \frac{\mathbf{1}}{\sum \mathrm{HR}} \right\}^{1/\mathbf{1}} = \left\{ 1 - \frac{\mathbf{1}}{250} \right\}^{1/\mathbf{1}} = \frac{249}{250}$$

If however the one event occurs to a man, that is like saying that we had a failure of 2/250 (or a success of 248/250) in two trials, so that in 1 trial of 250, we should have a success of

$$\hat{\xi} = \left\{ 1 - \frac{\mathbf{2}}{\sum \mathrm{HR}} \right\}^{1/\mathbf{2}} = \left\{ 1 - \frac{\mathbf{2}}{250} \right\}^{1/\mathbf{2}} \qquad \frac{248.998}{250}$$

Obviously, in smaller n's the differences would be more dramatic. For example, with the data above, we had $\hat{HR}_{ML} = 3.14$ *which for simplicity we will round to* $\hat{HR}_{ML} = 3$. Thus in the 5 risksets, we had the following structure

$o=d_1$	**1**	0	**1**	**1**	-
s_1	3	2	1	0	-
n_1	4	2	2	1	
d_0	0	**1**	0	0	**1**
s_0	3	2	2	1	0
n_0	3	3	2	1	1

or in "z=0 equivalents", replacing each person in z=1 group with HR=3

$o=d_1$	**3**	0	**3**	**3**	-
s_1	9	6	3	0	-
n_1	12	6	6	3	
d_0	0	**1**	0	0	**1**
s_0	3	2	2	1	0
n_0	3	3	2	1	1

or, summing and putting all individuals into "**z=0 equivalents**"...

d_0	**3**	**1**	**3**	**3**	**1**
s_0	12	8	5	1	0
n_0	15	9	8	4	1

Then the $\hat{\xi}$'s are estimated as

$$\left\{\frac{12}{15}\right\}^{1/3} \quad \frac{8}{9} \quad \left\{\frac{5}{8}\right\}^{1/3} \quad \left\{\frac{1}{4}\right\}^{1/3} \quad 0$$

or 0.93 0.89 0.86 0.63 0

yielding a Product Limit estimate of the S[ ] function for the **z=0** group:

1.00 0.93 0.83 0.71 0.44 0

For the **z= 1** group, the corresponding estimate is

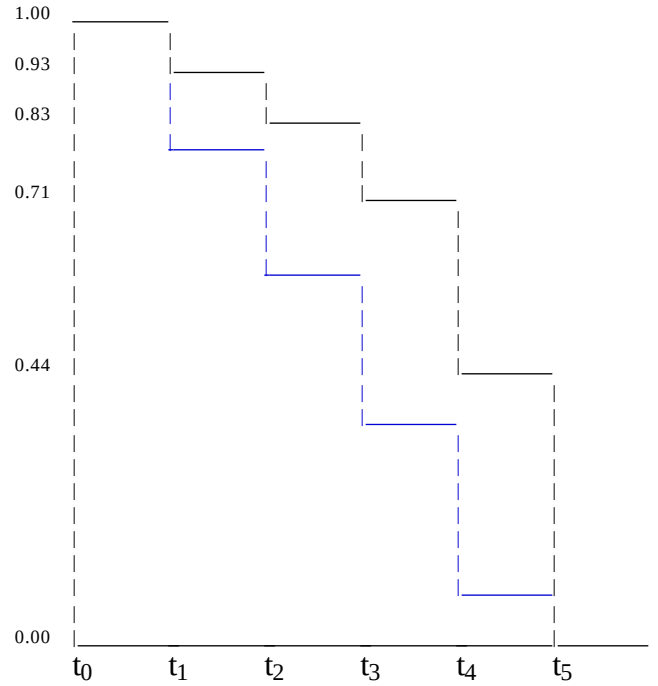
1.00 0.93^3 0.83^3 0.71^3 0.44^3 0

or

1.00 0.80 0.57 0.36 0.09 0

or (roughly)

$\hat{S}_0(t)$ (estimated from 5 risk sets) and $\hat{S}_1(t) = \{ \hat{S}_0(t) \}^3$



Following is the **estimate of $S_0(t)$ and $S_1(t)$** produced by **PHREG**.

First one must set up a file with the **covariate patterns** for which one wants survival (and other) curves.. Here there is just one covariate z, with 2 values, so there are only 2 possible covariate patterns.

1.

```
DATA cov_vals;
INPUT        tx ;
LINES;
             0
             1
;
RUN;
```

2.

```
title model with tx, and obtain curves;
proc phreg data=a;
 model time*event(0) = tx / RISKLIMITS;
 BASELINE OUT = curves
          COVARIATES  =  cov_vals
              SURVIVAL =  SURVIVAL
              LOGSURV  =  LOGSURV
              LOGLOGS  =  LOGLOGS;
```

PROC PRINT DATA=curves ROUND; RUN;

TX	TIME	SURVIVAL	LOGSURV	LOGLOGS	
0	0	*1.00*	0.00	.	
0	2	*0.93*	-0.07	**-2.63**	*
0	6	*0.83*	-0.19	**-1.68**	**
0	8	*0.71*	-0.34	**-1.08**	***
0	12	*0.45*	-0.79	**-0.23**	****
0	14	*0.00*	.	.	
1	0	*1.00*	0.00	.	(difference)#
1	2	*0.80*	-0.23	**-1.49**	* (1.14)
1	6	*0.56*	-0.58	**-0.54**	** (1.14)
1	8	*0.35*	-1.06	**0.06**	*** (1.14)
1	12	*0.08*	-2.48	**0.91**	**** (1.14)
1	14	*0.00*	.	.	
0.57*	0	*1.00*	0.00	.	
0.57	2	*0.87*	-0.14	-1.98	
0.57	6	*0.70*	-0.36	-1.03	
0.57	8	*0.52*	-0.65	-0.43	
0.57	12	*0.22*	-1.52	0.42	
0.57	14	*0.00*	.	.	

* 0.57 is the average, in the dataset, of the z values

It is not a coincidence that there is a constant difference of 1.14 between the two FITTED log[- log[S]] curves: this is a **consequence** of the proportional hazards assumption..

Plotting the EMPIRICAL log[- log[S]] curves to see if they are reasonably parallel allows a visual check on the proportional hazards assumption.

Stata

* store fitted survival for baseline group into new variable called s

stcox tx, basesurv(s)

* generate corresponding curve for tx=1 .. ^ = 'to power of')

gen s_1 = s^(exp(1.14))

* graph -log[S] i.e., cumulative hazard curves (na = Nelson-Aalen)

sts graph, na by(tx)

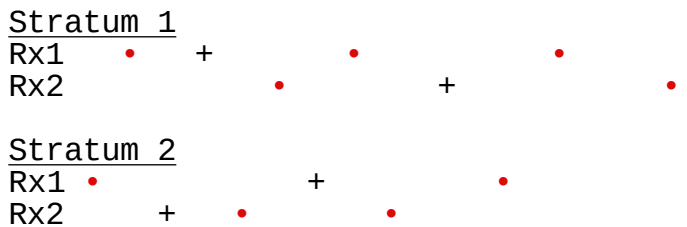
* graph -log[-log[S]] versus time, so check if parallel

stsphplot, by(tx)

* **stcoxkm** plots Kaplan-Meier observed survival curves and compares them to the Cox predicted curves for the same variable.

ML estimates for stratified survival data
(exercise.. follow ***Fig 3 in part II of JH's expository article***)

Consider the following pattern of observations for two treatments where • denotes a failure (event) and + denotes a censored observation and time runs from left to right. The observations are in 2 strata.



The above calculations used the data for stratum 1.

For the second stratum

a set up the risk sets.

b set up the likelihood contribution from each set and the overall likelihood for the stratum (follow e.g. of stratum 1)

c calculate the likelihood for several values of β

d draw a smooth sketch of the likelihood function (the numbers may be so tiny that you prefer to plotting the log of the likelihood function)

e at what value (approx) of β is the function a maximum?

f calculate numerically the 1st and 2nd derivatives of the log likelihood function in the neighbourhood of β_hat

Multiply the likelihood (or add the log Likelihoods) from the 1st stratum and the likelihood from b to produce the overall likelihood (or log Likelihood) from the 2 strata combined. Then maximize the combined likelihood (or log likelihood).

Individuals from different strata cannot be in same riskset
(but, if strata too fine, may have uninformative risksets)