Survival Analysis - part 2

Testing the Equality of Two Survival Curves

- The **log rank test**, a special case of the **Cochran-Mantel-Haenszel test**, is used to test $H_0 : S_T(t) = S_C(t)$.
- The null hypothesis states that the survival functions are the same for each time t.
- Calculation of the test statistic is shown below.
- At the i'th observed failure time t_i, let
  - M_i be the number at risk in the treatment group
  - T_i be the total number at risk (for both groups)
  - a_i be the number of deaths in the treatment group
  - N_i be the total number of deaths (for both groups)

- At each failure time t_i, we construct a 2 by 2 table comparing the number of failures in the two groups.

	Dead	Surviving	At Risk
Treat	a_i	$M_i - a_i$	M_i
Control	$N_i - a_i$	$T_i - N_i - M_i + a_i$	$T_i - M_i$
	N_i	$T_i - N_i$	T_i

- If the failure rate is the same in both groups, the expected number of deaths in the Treatment group is $M_i N_i / T_i$, which is the number at risk M_i times the combined proportion of deaths.

- The test statistic compares the observed to expected number of deaths in the treatment group, standardized by an estimate of its variance

$$Z = \sum_i (a_i - E_i) / \sqrt{\sum_i V_i}$$

where

$$E_i = \frac{M_i N_i}{T_i}$$

and

$$V_i = \frac{M_i N_i (T_i - M_i)(T_i - N_i)}{T_i^2 (T_i - 1)}$$

- The p-value against the two-sided alternative is

$$2P(Z > |Z_{obs}|)$$

- For the leukemia study, the necessary information to construct these tables is as follows:

	Num at Risk		Num of Deaths			
	Treat	Total	Treat	Total		
t_i	M_i	T_i	a_i	N_i	E_i	V_i
1	21	42	0	2	1.00	0.49
2	21	40	0	2	1.05	0.49
3	21	38	0	1	0.55	0.25
4	21	37	0	2	1.14	0.48
5	21	35	0	2	1.20	0.47
6	21	33	3	3	1.91	0.65
7	17	29	1	1	0.59	0.24
8	16	28	0	4	2.29	0.87
10	15	23	1	1	0.65	0.23
11	13	21	0	2	1.24	0.45
12	12	18	0	2	1.33	0.42
13	12	16	1	1	0.75	0.19
15	11	15	0	1	0.73	0.20
16	11	14	1	1	0.79	0.17
17	10	13	0	1	0.77	0.18
22	7	9	1	2	1.56	0.30
23	6	7	1	2	1.71	0.20
Total			9		19.25	6.26

- The test statistic is

$$Z = (9 - 19.25)/\sqrt{6.26} = -4.098$$

- The P value is $2P(Z > |-4.098|) = 4.17 \times 10^{-5}$, so we conclude that there is very strong evidence against the null hypothesis that the survival curves are the same.

- Note that $Z^2 = 16.79$ which equals the χ^2 value obtained from the computer in the last set of notes.

Proportional hazards model

- The **hazard function** is the rate of failure in a small interval Δ after time t, given that the subject has survived until t

$$h(t)\Delta = P(t \leq T < t + \Delta | T \geq t)$$

- If the failure time T has cumulative distribution function $F(t)$, density $f(t) = F'(t)$ and survival function $S(t) = 1 - F(t)$, then the hazard function is

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

- The simplest probability model for survival is the exponential, with density

$$f(t) = \lambda e^{-\lambda t}$$

  The cumumlative distribution function is

$$F(t) = 1 - e^{-\lambda t}$$

  and survival function

$$S(t) = e^{-\lambda t}$$

- The hazard function in this case is constant over time

$$h(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

- More realistic hazard functions are increasing, decreasing or 'bathtub' shaped - first decreasing, then constant, then increasing.

- To compare two groups, like Treatment and Control, we can compare their hazard functions.

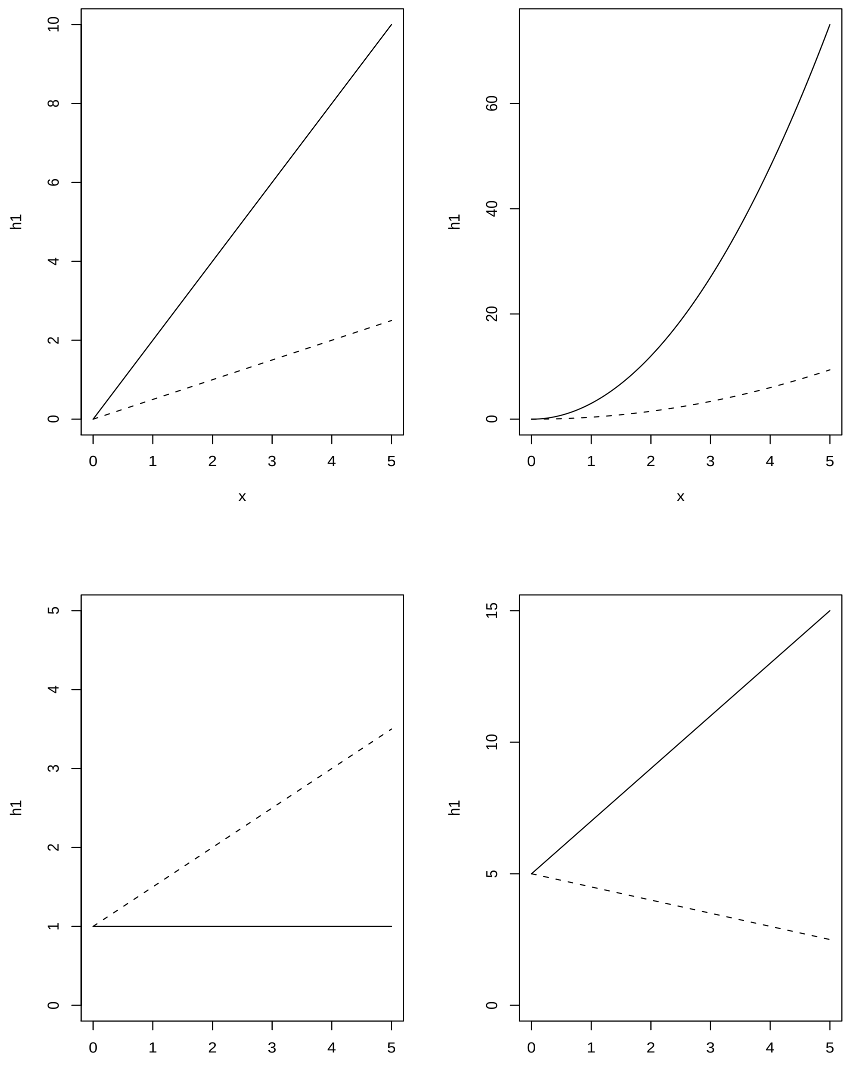
  - A smaller hazard indicates a slower rate of failures.

- Often it is assumed that hazard functions for two groups are *proportional*, so that

$$h_T(t) = k h_C(t)$$

  for some k.

- The following shows two cases with proportional hazards (top) and two where the hazards are not proportional (bottom).



- Cox's proportional hazard regression model is used to model survival as a function of predictors or covariates $X_1, \ldots, X_p$.
- Cox's model says that, if an individual has predictors $X_1, \ldots, X_p$, then their hazard is

$$h(t) = h_0(t) exp(b_1 X_1 + \ldots + b_p X_p)$$

- $h_0(t)$ is the baseline hazard, estimated nonparametrically.
- The term $exp(b_1 X_1 + \ldots + b_p X_p)$ is 1 if all X's are zero, and positive otherwise.
- The probability of survival at time t is estimated by

$$S(t) = exp(-H(t))$$

  where $H(t)$ is the cumulative hazard, obtained by integrating $h(s)$ up to time t
- The *hazard ratio* for two values of a covariate X_i (with all other covariates held the same) is

$$\frac{h_1(t)}{h_2(t)} = exp(b_i x_{i1} - b_i x_{i2}) = exp[b_i(x_{i1} - x_{i2})]$$

- Equivalently

$$log\left(\frac{h_1(t)}{h_2(t)}\right) = b_i(x_{i1} - x_{i2})$$

- and we see that b_i is the logarithm of the hazard ratio associated with a unit increase in X_i, with all other variables held constant.
- If X_i is binary, such as an indicator equal to 1 for the treatment group and 0 for the control group, then

$$\frac{h_1(t)}{h_2(t)} = exp(b_i)$$

- A hazard ratio greater than 1 implies subjects with X_{i1} fare less well than those with X_{i2}.

- Computer output for the leukemia data is shown below.

```
> leuktr.Surv=Surv(leuk.t,1-leuk.cen)
> leuk.ph=coxph(leuktr.Surv~leuktr)
> leuk.ph=coxph(leuktr.Surv~leuk.tr)
> print(leuk.ph)
Call:
coxph(formula = leuktr.Surv ~ leuk.tr)


          coef exp(coef) se(coef)     z       p
leuk.tr -1.57     0.208    0.412 -3.81 0.00014

Likelihood ratio test=16.4  on 1 df, p=5.26e-05
n= 42, number of events= 30
```

- In this case the only covariate is an indicator for Treatment vs Control.
- A test for difference between Treatment and Control is given by a test that the β coefficient is zero.
- The output gives us the Z statistic (coef/se) and P-value.
- Note that this test statistic is close to the log rank statistic obtained above.
- One reason they are slightly different is that this approach assumes that the hazards are proportional whereas the log rank test does not.