The log-rank test

Patrick Breheny

September 11

Introduction

- Last week, we discussed point and interval estimation for survival curves
- By visually comparing the curves and bands, one can get a rough sense of whether there is a statistically significant difference the survival of two groups
- Clearly, however, it would also be nice to test for a difference between two groups; this is our subject for today

Setup

We will use the same notation as last week's Kaplan-Meier estimator lectures, with straightforward extensions to accommodate multiple groups:

- Let $0 = t_0 < t_1 < t_2 < \cdots < t_J < t_{J+1} = \infty$ denote the pooled failure times (times at which any subject in either group was observed to fail)
- Let d_{1j} denote the number of failures at time t_j in group 1, d_{2j} the number of failures at time t_j in group 2, and d_j denote the total number of failures at time t_j across all groups
- And so on for $n_{1j}, c_{1j}, \ldots$

Table for time t_j

Consider the following contingency table for all subjects in the risk set at time t_j:

	Group 1	Group 2	Total
Deaths	d_{1j}	d_{2j}	d_j
Survivors	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
Total	n_{1j}	n_{2j}	n_j

Hypergeometric distribution

- Conditional on the margins of the table on the previous slide, under the null hypothesis that $S_1 = S_2$ the random variable D_{1j} follows a hypergeometric distribution with mean

$$e_{1j} = n_{1j} \frac{d_j}{n_j}$$

and variance

$$v_{1j} = n_{1j} \frac{d_j}{n_j} \frac{n_j - d_j}{n_j} \frac{n_j - n_{1j}}{n_j - 1}$$

- However, we still need some way of combining information across all the failure times

Combining over failure times

- The sum $\sum D_{1j}$ does not follow any known distribution; however, since each $w_j = d_{1j} - e_{1j}$ follows an approximate normal distribution with zero mean and variance $v_j = v_{1j}$ under the null,

$$W \dot{\sim} \mathrm{N}(0, V),$$

  where $W = \sum_j w_j$ and $V = \sum_j v_j$, provided that failures are conditionally independent

- Or equivalently,

$$\frac{W^2}{V} = \frac{\left(\sum_j w_j\right)^2}{\sum_j v_j} \dot{\sim} \chi^2_1$$

Remarks

- This test is known as the *log-rank test*
- The idea behind the test is essentially the same as that of the Cochran-Mantel-Haenszel test in categorical data analysis, with time as the stratification variable
- The log-rank test is the most widely used test for comparing two survival time distributions, in part because the test statistic has a simple "observed - expected" form
- The log-rank test is particularly powerful when the ratio between the two hazard functions being compared is constant across time

Extension to multiple groups

- It is fairly straightforward to extend the log-rank test to compare an arbitrary number of groups
- Suppose we have $K+1$ groups, with one group arbitrarily chosen as the reference group
- Let $\mathbf{w}_j$ denote the vector $(d_{1j} - e_{1j}, \ldots, d_{Kj} - e_{Kj})$
- The conditional covariance matrix of $\mathbf{w}_j$, $\mathbf{V}_j$, has diagonal elements as given previously and off-diagonal elements

$$(V_j)_{ik} = -\frac{n_{ij} n_{kj} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

Extension to multiple groups (cont'd)

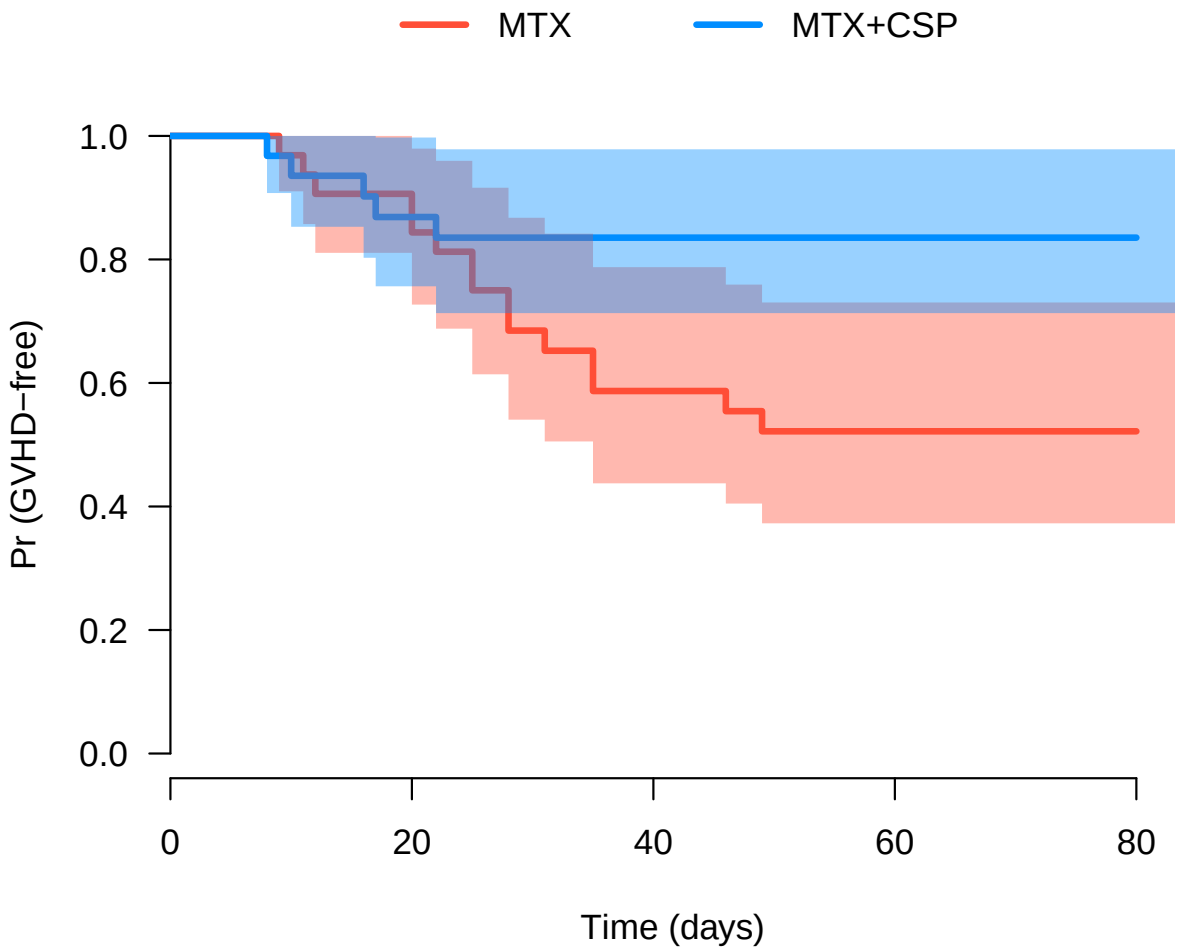
- Then, letting $\mathbf{w} = \sum_j \mathbf{w}_j$ and $\mathbf{V} = \sum_j \mathbf{V}_j$, we have

$$\mathbf{w}^T \mathbf{V}^{-1} \mathbf{w} \dot{\sim} \chi^2_K$$

- Note that in this setup, we have $K + 1$ groups but only K independent counts in the contingency table because we are conditioning on the margins, and therefore only K degrees of freedom
- An alternative (and more elegant, at least in my opinion) way of dealing with this would be to include all groups in the test statistic, but use the generalized inverse of $\mathbf{V}$ to construct the test statistic

GVHD data

Recall our GVHD data from last week:



Observed vs. Expected

- The survival curves suggest a difference between the two groups, at least for day 20 onwards, but the confidence intervals are fairly wide at that point and it isn't obvious whether the difference could be explained by chance alone
- For the MTX + CSP group, under the null we would expect

$$\sum_j e_{1j} = 10.2$$

  subjects to experience GVHD
- In the actual experiment, however, only 5 subjects developed GVHD in the MTX+CSP group
- Conversely, we would expect 9.8 subjects to develop GVHD in the MTX alone group, but 15 subjects did

Log-rank test: Results

- Furthermore,

$$V = \sum_j v_{1j} = 4.92$$

- Thus, what we observed was 2.34 standard deviations away from what we would expect, which yields a two-sided p-value of $2\Phi(-2.34) = 0.02$
- Or equivalently,

$$\frac{(5 - 10.2)^2}{4.92} = 5.49$$
$$\mathbb{P}\{\chi_1^2 > 5.49\} = 0.02$$

survdiff

The survival package provides the function survdiff to carry out log-rank tests for differences between survival curves; its syntax is more or less identical to survfit:

```
> survdiff(Surv(Time, Status) ~ Group, Data)

                N Observed Expected (O-E)^2/E (O-E)^2/V
Group=MTX      32       15      9.8      2.75      5.49
Group=MTX+CSP  32        5     10.2      2.65      5.49

 Chisq= 5.5  on 1 degrees of freedom, p= 0.0192
```

PBC data

- Let's also try out a multi-sample comparison
- In your homework, I asked you to look at the `pbc` data in `R`, a study of progression-free survival in patients with primary biliary cholangitis
- Specifically, you calculated survival curves for these patients, broken down by stage (1-4); let's now carry out a formal test of whether the observed differences could be due to chance

PBC data: Results

```
> survdiff(Surv(time, status!=0) ~ stage, pbc)

n=412, 6 observations deleted due to missingness.

          N Observed Expected (O-E)^2/E (O-E)^2/V
stage=1  21        2     13.3      9.58     10.41
stage=2  92       28     51.4     10.66     15.05
stage=3 155       58     71.2      2.44      4.02
stage=4 144       94     46.1     49.69     67.58

 Chisq= 73.9  on 3 degrees of freedom, p= 6.66e-16
```

The results agree with what we would have expected: many more failures than would be expected among stage 4 patients, fewer failures in stages 1, 2, and 3, and a highly significant result

Remarks

- It is worth noting that, unlike in the two-sample case, one cannot reconstruct the test statistic from the table provided, as it depends on covariances
- To calculate the test statistic "by hand", we would need:

```
> lrt <- survdiff(Surv(time, status!=0) ~ stage, pbc)
> w <- lrt$obs[1:3] - lrt$exp[1:3]
> V <- lrt$var[1:3, 1:3]
> t(w) %*% solve(V) %*% w
          [,1]
[1,] 73.92355
```

- As remarked earlier, you get the same result regardless of which group you leave out, or if you include all groups but take the generalized inverse of `V`

Weighted log-rank tests

- The log-rank test statistic is of the form

$$\frac{\left(\sum_j w_j\right)^2}{\sum_j v_j}$$

- It is not obvious, however, that each time point should receive the same weight when we construct this linear combination
- One natural extension, then, is the family of *weighted log-rank tests*, which have the form

$$\frac{\left(\sum_j \alpha_j w_j\right)^2}{\sum_j \alpha_j^2 v_j},$$

where $\{\alpha_j\}$ are weights chosen to emphasize or deemphasize various time points

The Gehan and Peto-Prentice tests

- For example, one reasonable weighting scheme would be to weigh time points by the number at risk at time t_j:

$$\alpha_j = n_j;$$

  This is known as the *Gehan*, *Gehan-Breslow*, or *Gehan-Wilcoxon* test
- A somewhat similar idea is to weigh time points according to the (pooled) survival estimate:

$$\alpha_j = \hat{S}(t_j);$$

  This is known as the *Peto-Prentice*, or *Peto-Peto*, test
- Note that both tests place more emphasis on earlier failure times compared to the log-rank test

Example: GVHD data

- The `survdiff` function has an option, `rho`, that can be set to 1 in order to perform the Peto-Prentice test
- For example, returning to the GVHD data,

```
> survdiff(Surv(Time, Status) ~ Group, Data, rho=1)

N Observed Expected (O-E)^2/E (O-E)^2/V
Group=MTX      32    12.38     8.40      1.88      4.37
Group=MTX+CSP  32     4.65     8.63      1.83      4.37

 Chisq= 4.4  on 1 degrees of freedom, p= 0.0366
```

- The fact that we get a larger p-value compared to the log-rank test makes sense: for the GVHD study, the differences between the two survival curves occurred mainly at later failure times