BIST P8110: Applied Regression II

6. Comparing $S(t)$ of Two or More Groups

Qixuan Chen

Department of Biostatistics
Columbia University

This lecture's big ideas

- Log-rank test
- Wilcoxon test
- Generalized log-rank test
- Generalized Wilcoxon test

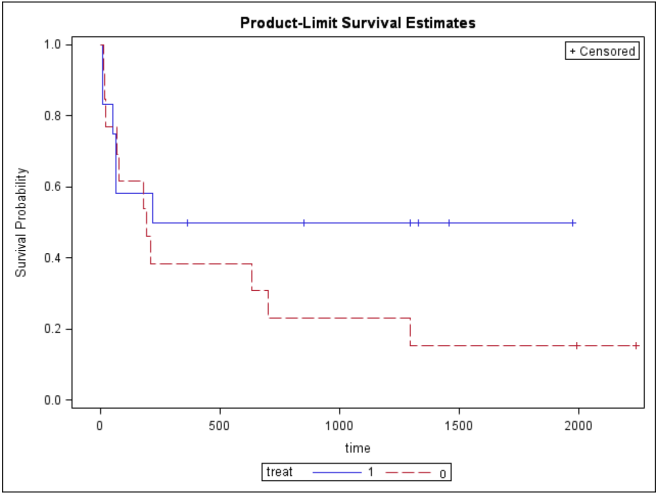
Comparison of Survival Functions

- It is often of interest to compare the survival functions of 2 or more groups.
- For example, we are interested in knowing if patients in one treatment group survive longer than those in the other treatment group.
- When comparing survival time of patients in two groups, we should begin with a graphical display of the data in each group, using Kaplan-Meier estimator of survival function.

Data: Survival Time for Patients with Myelomatosis (Peto et al. 1977)

OBS	DUR	STATUS	TREAT	RENAL
1	8	1	1	1
2	180	1	0	0
3	632	1	0	0
4	852	0	1	0
5	52	1	1	1
6	2240	0	0	0
7	220	1	1	0
8	63	1	1	1
9	195	1	0	0
10	76	1	0	0
11	70	1	0	0
12	8	1	1	0
13	13	1	0	1
14	1990	0	0	0
15	1976	0	1	0
16	18	1	0	1
17	700	1	0	0
18	1296	0	1	0
19	1460	0	1	0
20	210	1	0	0
21	63	1	1	1
22	1328	0	1	0
23	1296	1	0	0
24	365	0	1	0
25	23	1	0	1

Graphical Display of Myelomatosis Data



Comparison of the Survival Curves

- The survival probability decreases quickly at the beginning and slows down after about 220 days.
  - Before $t = 220$ days, the two survival curves are virtually indistinguishable, with little visual evidence of a treatment effect.
  - The gap that develops after 220 days reflects the fact that no additional deaths occur in treatment group 1 after that time.
- Question: Whether the observed difference between the two estimated survival functions are significant?

Comparing Two Survival Functions

- Our goal is to compare $S_1(t)$ to $S_0(t)$ and to formally test

$$H_0 : S_1(t) = S_0(t), \text{ for all } t \leq \tau$$
$$H_\alpha : S_1(t) \neq S_0(t), \text{ for some } t \leq \tau$$

  - τ is the largest observed time
  - H_0 reflects no difference between groups in survival functions and H_α states that this is not true.
  - We need to formulate a test statistic that measures the extent to which the observed data depart from the null hypothesis.

Two Possible Approaches

- We will consider two methods:
  - The log-rank test
  - The Wilcoxon test
- Both approaches
  - are non-parametric
  - can accommodate censored data
  - are a weighted comparison of the observed and expected failures at each failure time across treatment groups.

The Log-Rank Test

- The log-rank test statistic is the most commonly used statistic to test for difference between 2 groups.
- To construct the log-rank test, we will order the unique observed survival times as done previously:

$$t_{(1)}, t_{(2)}, \ldots, t_{(m)}$$

- For each observed survival time, $t_{(j)}$, create a contingency table of group by status.

Group	1	0	Total
Death	d_{1j}	d_{0j}	d_j
Alive	$n_{1j} - d_{1j}$	$n_{0j} - d_{0j}$	$n_j - d_j$
At Risk	n_{1j}	n_{0j}	n_j

The Log-Rank Test

- The contribution to the test statistic at each time point is obtained by calculating the expected number of events in group 1 or 0, assuming that the survival function is the same in each of the two groups.
  - The expectation of d_{1j} (the number of events in group 1) is estimated as
$$\hat{e}_{1j} = \frac{n_{1j} d_j}{n_j}$$
  - The variance of d_{1j} is estimated as
$$\hat{\nu}_{1j} = \frac{n_{1j} n_{0j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

The Log-Rank Test

- One way to assess the validity of the null hypothesis is to sum the difference between the observed number of events in each group and the expected number of events in that group if the null hypothesis is true at each observed survival time point.
- The log-rank test statistic is defined as

$$Q = \frac{\left[\sum_{j=1}^{m}(d_{1j} - \hat{e}_{1j})\right]^2}{\sum_{j=1}^{m} \hat{\nu}_{1j}}$$

- Assuming that the censoring is independent of group, and the total number of observed events as well as the sum of the expected number of events are large, Q follows a **chi-square** distribution with **one** degree of freedom.

The Wilcoxon Test

- The Wilcoxon test statistic is a weighted log-rank test statistic, which has the following general form:

$$Q = \frac{[\sum_{j=1}^{m} w_j (d_{1j} - \hat{e}_{1j})]^2}{\sum_{j=1}^{m} w_j^2 \hat{\nu}_{1j}}$$

  - The Log-rank test: $w_j = 1$
  - The Wilcoxon test: $w_j = n_j$ = #. of subjects at risk at time t_j
  - It follows that $Q \sim \chi_1^2$ under the null hypothesis.

Computing the Test Statistics: Myelomatosis Data Set

Grp 1: $8, 8, 52, 63, 63, 220, 365^+, 852^+, 1296^+, 1328^+, 1460^+, 1976^+$
Grp 0: $13, 18, 23, 70, 76, 180, 195, 210, 632, 700, 1296, 1990^+, 2240^+$

- Step 1: Sort the unique observed survival time points from both groups

t_j	d_{1j}	d_{0j}	$d_j = d_{1j} + d_{0j}$	n_{1j}	n_{0j}	$n_j = n_{1j} + n_{0j}$
8	2	0	2	12	13	25
13	0	1	1	10	13	23
18	0	1	1	10	12	22
23	0	1	1	10	11	21
52	1	0	1	10	10	20
63	2	0	2	9	10	19
70	0	1	1	7	10	17
76	0	1	1	7	9	16
180	0	1	1	7	8	15
195	0	1	1	7	7	14
210	0	1	1	7	6	13
220	1	0	1	7	5	12
632	0	1	1	5	5	10
700	0	1	1	5	4	9
1296	0	1	1	4	3	7

- Step 2: Construct 15 (2×2) tables corresponding to 15 distinct failure times, and calculate $\hat{e}_{1j}$ and $\hat{\nu}_{1j}$.
  - For example, $j = 1 (t = 8$ days$)$

$j = 1$ (8 days)	treat=1	treat=0	Total
Dead	$d_{1j} = 2$	$d_{0j} = 0$	$d_j = 2$
Alive	$n_{1j} - d_{1j} = 10$	$n_{0j} - d_{0j} = 13$	$n_j - d_j = 23$
At Risk	$n_{1j} = 12$	$n_{0j} = 13$	$n_j = 25$

$$\hat{e}_{1j} = \frac{n_{1j} d_j}{n_j} = \frac{12 \times 2}{25} = 0.96$$

$$\hat{\nu}_{1j} = \frac{n_{1j} n_{0j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} = \frac{12 \times 13 \times 2 \times 23}{25^2 \times (25 - 1)} = 0.4784$$

Computing the Test Statistics: Myelomatosis Data Set

- Continue Step 2 and get the following table

t_j	d_{1j}	n_j	$\hat{e}_{1j}$	$\hat{\nu}_{1j}$	$n_j\hat{e}_{1j}$	$n_j^2\hat{\nu}_{1j}$
8	2	25	0.96000	0.47840	24	299
13	0	23	0.43478	0.24575	10	130
18	0	22	0.45455	0.24793	10	120
23	0	21	0.47619	0.24943	10	110
52	1	20	0.50000	0.25000	10	100
63	2	19	0.94737	0.47091	18	170
70	0	17	0.41176	0.24221	7	70
76	0	16	0.43750	0.24609	7	63
180	0	15	0.46667	0.24889	7	56
195	0	14	0.50000	0.25000	7	49
210	0	13	0.53846	0.24852	7	42
220	1	12	0.58333	0.24306	7	35
632	0	10	0.50000	0.25000	5	25
700	0	9	0.55556	0.24691	5	20
1296	0	7	0.57143	0.24490	4	12

- Step 3: Calculate the test statistics
  - The log-rank test
  $$\chi^2_{\text{log-rank}} = \frac{[\sum_{j=1}^{15}(d_{1j} - \hat{e}_{1j})]^2}{\sum_{j=1}^{15}\hat{\nu}_{1j}} = \frac{[2.3375973]^2}{4.1630128} = 1.3126$$
  $$\text{P-value} = Pr(\chi^2_1 \geq 1.3126) = 0.2519$$
  - The Wilcoxon test
  $$\chi^2_{\text{Wilcoxon}} = \frac{[\sum_{j=1}^{15} n_j(d_{1j} - \hat{e}_{1j})]^2}{\sum_{j=1}^{15} n_j^2\hat{\nu}_{1j}} = \frac{[18]^2}{1301} = 0.2490$$
  $$\text{P-value} = Pr(\chi^2_1 \geq 0.2490) = 0.6178$$
  - Conclusion: Fail to reject the null hypothesis at a significance level of 0.05.

Log-Rank Test vs. Wilcoxon Test

- The log-rank test weights the difference between the observed and the expected number of events equally across all event times, while the Wilcoxon test weights this difference by the number at risk.
- The log-rank test is more sensitive than the Wilcoxon test to difference between groups in later points in time.
- The Wilcoxon test is more sensitive than the log-rank test to difference between groups in early points in time.

Log-Rank Test vs. Wilcoxon Test

- The log-rank test is more powerful than the Wilcoxon test when the proportional hazard assumption holds.
- If the survival curves cross, neither the log-rank test nor the Wilcoxon test is helpful.
- **Conclusion**: Use the log-rank test unless there is evidence that the proportional hazard assumption does not hold.

Comparison of $S(t)$ for Three or More Groups

- It may be desirable to compare survival functions among $K \geq 3$ groups
  - Hypothesis
    $H_0 : S_1(t) = S_2(t) = ... = S_K(t)$ for all $t \leq \tau$;
    H_α : at least one of the $S_k(t)$ is different for some $t \leq \tau$;
    where, τ is the largest observed time.
  - Both the log-rank test and the Wilcoxon test can be extended for multiple-group comparison.

The Generalized Tests

- To construct the test, we will order the unique observed survival times as done previously:

$$t_{(1)}, t_{(2)}, \dots, t_{(m)}$$

- For each observed survival time, $t_{(j)}$, create a two by K contingency table.

Group	1	2	...	K	Total
Death	d_{1j}	d_{2j}	...	d_{Kj}	d_j
Alive	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	...	$n_{Kj} - d_{Kj}$	$n_j - d_j$
At Risk	n_{1j}	n_{2j}	...	n_{Kj}	n_j

The Generalized Tests

- We compare the observed and expected number of events for the first $K-1$ of the K groups.
  - The observed and expected number of events are

$$\underline{d}_j^T = (d_{1j}, d_{2j}, \ldots, d_{(K-1)j})$$
$$\underline{\hat{e}}_j^T = (\hat{e}_{1j}, \hat{e}_{2j}, \ldots, \hat{e}_{(K-1)j})$$

  - Under H_0, the expectation of d_{kj} is estimated as

$$\hat{e}_{kj} = \frac{n_{kj} d_j}{n_j}, k = 1, 2, \ldots, (K-1)$$

- The difference between the observed and expected number of events are

$$(\underline{d}_j - \underline{\hat{e}}_j)^T = (d_{1j} - \hat{e}_{1j}, d_{2j} - \hat{e}_{2j}, \ldots, d_{(K-1)j} - \hat{e}_{(K-1)j})$$

The Generalized Tests

- To obtain a test statistic, we need to estimate the covariance matrix of $\underline{d}_j$.
  - $\underline{\hat{V}}_j$ denotes the $(K-1) \times (K-1)$ covariance matrix of $\underline{d}_j$.
  - The diagonal elements of $\underline{\hat{V}}_j$ are

$$\hat{\nu}_{kkj} = \frac{n_{kj}(n_j - n_{kj})d_j(n_j - d_j)}{n_j^2(n_j - 1)}, k = 1, 2, \ldots, K - 1.$$

  - The off-diagonal elements of $\underline{\hat{V}}_j$ are

$$\hat{\nu}_{klj} = -\frac{n_{kj}n_{lj}d_j(n_j - d_j)}{n_j^2(n_j - 1)}, k, l = 1, 2, \ldots, K - 1, k \neq l.$$

Test Statistic

- The test statistic to compare the survival experience of the K groups is

$$Q = \Big[\sum_{j=1}^{m} \underline{W}_j (\underline{d}_j - \underline{\hat{e}}_j) \Big]^T \Big[\sum_{j=1}^{m} \underline{W}_j \underline{\hat{V}}_j \underline{W}_j \Big]^{-1} \Big[\sum_{j=1}^{m} \underline{W}_j (\underline{d}_j - \underline{\hat{e}}_j) \Big]$$

  - $\underline{W}_j$ is a $(K-1) \times (K-1)$ diagonal matrix, that is, $\underline{W}_j = \text{diag}(w_j)$. For the generalized log-rank test, $w_j = 1$. For the generalized Wilcoxon test, $w_j = n_j$.
  - Under H_0, if the summed expected number of events is large, then Q will be approximately distributed as chi-square with $K-1$ degrees of freedom, and p-value is equal to $\Pr(\chi^2_{K-1} \geq Q)$.

Suggested Readings

- Chapter 2.4 (Hosmer et al.)
- Chapter 6.8 (Dupont)