Survival Analysis

Survival Analysis

Dependent variable of interest: *time* from some initial observation until the occurrence of an event.

Examples:

- Time to failure of a light bulb
- Time to death from initial diagnosis of a disease
- Time to relapse

Properties of Survival Times

- The distribution is skewed
- Survival times of some units may not be known (i.e., *censored*).
  This may be due to loss to follow-up, drop-outs, or the event of interest not occurring when the experiment ended.

Basic assumption: Censoring mechanism is independent of failure times (i.e., cenosring is non-informative).

Characterization of the Distribution of Survival Times

Let T be survival time, with p.d.f. $f(t)$.

The *survival function* $S(t)$ is defined to be

$$S(t) = Pr[T > t]$$

The *hazard function*

The probability of failure at a specified time, given that there has been survival until that time point. Also known as *instantaneous failure rate*.

$$\lambda(t) = \lim_{u \to 0} \frac{\Pr\{t < T \leq t + u \,|\, T > t\}}{u} = \frac{f(t)}{S(t)}$$

Relationship between Survival and hazard functions:

The cumulative hazard function

$$\Lambda(t) \equiv \int_0^t \lambda(v)dv = -\log S(t)$$

$$S(t) = \exp[-\Lambda(t)]$$

Exponential distribution

$$\Lambda(t) = \lambda t$$

$$S(t) = \exp(-\Lambda(t)) = \exp(-\lambda t)$$

$$\Pr\{T > t_0 + t \mid T > t_0\} = \Pr\{T > t\}$$

Weibull distribution

$$\lambda(t) = \alpha\gamma t^{\gamma-1}$$

$$\Lambda(t) = \alpha t^{\gamma}$$

$$S(t) = \exp(-\Lambda(t)) = \exp(-\alpha t^{\gamma})$$

Objectives of Survival Analysis

- Estimation of survival curves
- Estimation of parameters of survival curves and/or hazard functions
- Comparing survival curves
- Model building and diagnostics

Estimating Survival Curves

Life-Table Methods

- Let $t_1, \cdots, t_n$ be independent survival times.
- Group the survival time into k fixed intervals.
- Let n_j, d_j and c_j be the number of units surviving at the beginning of, the number dying during, and the number censored in the j' th interval, respectively.

Estimating Survival Curves (cont.)

- Let s_j denote the probability that a unit that survived to the beginning of the j'th interval would survive through that interval. Then,

$$s_j = \frac{n_j - d_j - c_j/2}{n_j - c_j/2}$$

- The overall probability of surviving through the k'th interval

$Pr[T > t_k] = Pr[T > t_1] \cdots Pr[T > t_k \mid T > t_{k-1}]$:

$$\hat{S}(t_k) = \Pi_{j=1}^{k} s_j$$

An estimate of the variance of $\hat{S}$ may be computed based on

Greenwood's formula

$$\hat{V}ar[\hat{S}(t_k)] = \hat{S}^2(t_k) \sum_{j=1}^{k} \frac{d_j}{m_j(m_j - d_j)}$$

where $m_j = n_j - c_j/2$.

Example. The intervals correspond to year after diagnosis of a certain disease.

Year	n_j	d_j	c_j	s_j	$\hat{S}(t_j)$
0 -1	126	47	19	0.60	0.60
1- 2	60	5	17	0.90	***
2 -3	38	2	15	0.93	0.50
3 -4	21	2	9	0.88	0.44
4 -5	10	-	6	1.00	0.44
Total		56	66		

Kaplan-Meier Estimator

Approach uses the actual times, rather than grouping them into intervals.

Denote the ranked times of failure and censoring by $t_{(1)} < \cdots < t_{(n)}$.

Let n_j be the number alive just before time $t_{(j)}$, and d_j the number that died at time $t_{(j)}$. Put

$$s_j = Pr[T > t_{(j)} \mid T > t_{(j-1)}]$$

Kaplan-Meier Estimator

Then a reasonable estimate of of $1 - s_j$ is

$$1 - \hat{s}_j = \frac{d_j}{n_j}$$

Hence

$$\hat{S}(t) = \Pi_{t_{(j)} \leq t} \hat{s}_j$$

where the multiplication is over the uncensored failure times $t_{(j)}$.

Example. The following data shows the length of time until remission: 9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+.

T	d_j	n_j	d_j/n_j	$\hat{S}(t)$
9	1	11	1/11	0.91
13	1	10	1/11	0.82
13+	0	9	0	0.82
18	1	8	1/8	0.72
23	1	7	1/ 7	0.61
28+	0	6	0	0.61
31	1	5	1/ 5	0.49
34	1	4	1/ 4	0.37
45+	0	3	0	0.37
48	1	2	1/2	0.18
161	0	1	0	0.18

Nelson estimator

The Nelson estimtor of the integrated hazard function is given by

$$\hat{\Lambda}_N(t) = \sum_{uncensored: t_{(j)} \leq t} \left(\frac{d_j}{n_j} \right).$$

Fleming-Harrington Estimator

$$\hat{S}_{FH}(t) = e^{-\hat{\Lambda}_N(t)}$$

Variance Estimation

Greenwood's formula

$$Var\hat{\Lambda}_N(t) = \sum_{uncensored: t_{(j)} \le t} \frac{d_j}{n_j(n_j - d_j)}$$

$$Var\hat{S}(t) = \hat{S}^2(t) Var\hat{\Lambda}_N(t)$$

Approximate confidence intervals based on the original scales:

$$\hat{S}(t) \pm Z_{\alpha/2} \times SE(\hat{S}(t))$$

may give values outside the acceptable range [0,1].

The cumulative hazard or log S scale. Guarantees positive values.

$$exp\{logS \pm Z_{\alpha/2} SE(\hat{\Lambda})\}$$

Log hazard scale. Guarantees values in [0,1].

$$exp\{exp\{log(-logS) \pm Z_{\alpha/2} SE(\hat{\Lambda}_N)\}\}$$

Commands to fit a Kaplan-Meier and plot it

```
fit <- survfit(Surv(time, status) ~ group, data=dataset)
plot(fit, lty=2:3)
```

```
> data(ovarian)
> ovarian[1:5,]
  futime   fustat     age        resid.ds rx ecog.ps
1     59      1        72.3315        2       1       1
2    115      1        74.4932        2      1       1
3    156      1        66.4658        2      1       2
4    421      0        53.3644        2      2       1
5    431      1        50.3397        2      1       1
......
```

```
fit <- survfit(formula = Surv(futime, fustat) ~ rx, data = ovarian)
```

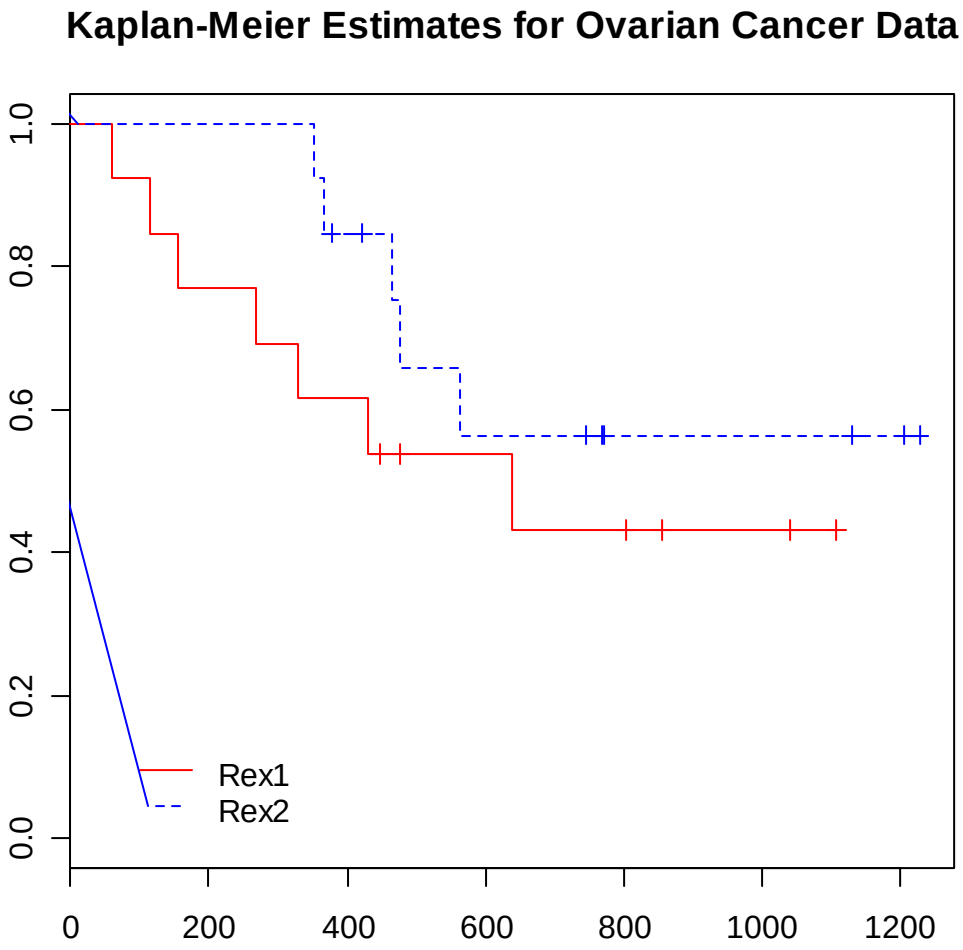
> **summary(fit)**

rx=1

time	n.risk	n.event	survival	std.err	95% CI	
59	13	1	0.923	0.0739	0.789	1.000
115	12	1	0.846	0.1001	0.671	1.000
156	11	1	0.769	0.1169	0.571	1.000
268	10	1	0.692	0.1280	0.482	0.995
329	9	1	0.615	0.1349	0.400	0.946
431	8	1	0.538	0.1383	0.326	0.891
638	5	1	0.431	0.1467	0.221	0.840

rx=2

time	n.risk	n.event	survival	std.err	95% CI	
353	13	1	0.923	0.0739	0.789	1.000
365	12	1	0.846	0.1001	0.671	1.000
464	9	1	0.752	0.1256	0.542	1.000
475	8	1	0.658	0.1407	0.433	1.000
563	7	1	0.564	0.1488	0.336	0.946


Kaplan-Meier Estimates for Ovarian Cancer Data
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
1200
Rex1
Rex2

Comparing Survival Curves

Two Samples

Suppose we are interested in comparing the median survival times corresponding to two distributions.

Let $X_1, \cdots, X_n$, and $Y_1, \cdots Y_m$ be independent samples of survival times.

1. *Gehan test*

An extension of the Mann-Whitney (Wilcoxon rank-sum) test. Recall that the Mann-Whitney test is based on

$$U_{jk} = \begin{cases} +1 & \text{if } X_j > Y_k \\ 0 & \text{if } X_j = Y_k \\ -1 & \text{if } X_j < Y_k \end{cases}$$

- For censored data:

$$U_{jk} = \begin{cases} +1 & \textit{if we know } X_j > Y_k \\ & \textit{i.e., } X_j > Y_k \textit{ or } X_j^+ \geq Y_k \\ 0 & \textit{otherwise} \\ -1 & \textit{if we know } X_j < Y_k \end{cases}$$

Put

$$U = \sum_j \sum_k U_{jk}$$

The Gehan test rejects for large values of U.

2. *Mantel-Haenszel or log-rank test*

Consider a sequence of r 2×2 tables.

	Number Dead	Number Alive	
Group 1	a_k	b_k	n_{k1}
Group 2	c_k	d_k	n_{k2}
	m_{k1}	m_{k2}	n_k

The hypothesis of interest may be formulated as

$$H_0 : p_{k1} = p_{k2}, k = 1, \cdots, r,$$

where p_{k1} and p_{k2} are stratum specific death probabilities for Group 1 and 2, respectively.

When the strata are independent, the Mantel-Haenszel test is given by

$$T_{MH} = \frac{\Sigma_{k=1}^{r}(a_k - E_k)}{\sqrt{\Sigma_{k=1}^{r} V_k}}$$

where $E_k = E[A_k \mid H_0]$ and $V_k = Var[A_k \mid H_0]$.

T_{MH} is approximately standard normal under H_o, when the strata are independent.

Now assume that the strata correspond to uncensored (failure) time points. The test thus obtained is called the log-rank test.

It is noted that:

- *Although the strata are not independent, under regularity conditions, asymptotic normality still holds.*
- *Unlike the Gehan test, the long-rank test is not affected by unequal censoring patterns.*
- *This test is known to be most powerful under proportional hazards alternatives.*

The test may be generalized to Tarone-Ware and Fleming-Harrington family of tests. Define

$$T = \frac{\Sigma_{k=1}^{r} W_k(a_k - E_k)}{\sqrt{\Sigma_{k=1}^{r} W_k^2 V_k}}$$

- $W_k = 1$: *The log-rank test*
- $W_k = n_k$: *Gehan test*
- $W_k = \sqrt{n_k}$: *Tarone-Ware class of tests. Intermediate between Gehan and log-rank.*
- $W_k = \hat{S}^{\rho}(t_k)$. *When* $\rho = 1$, *we have Peto-Peto, while for* $\rho \epsilon (0, 1)$, *we get Fleming-Harrington tests.*

Implementation:

survdiff(Surv(time, status) ~ group, data=dataset, rho=0)

> survdiff(Surv(futime, fustat) ~ rx, rho=0, data=ovarian)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
rx=1	13	7	5.23	0.596	1.06
rx=2	13	5	6.77	0.461	1.06

Chisq= 1.1 on 1 degrees of freedom, p= 0.303

Regression Approaches in Survival Analysis

In the presence of covariates, the standard linear model formulation is not appropriate for survival times, due to censoring and the skewed nature of the distributions.

```
> library(splines)
> library(survival)
> data(ovarian)
......
```
Some useful R functions:
```
> help(package="survival")
Surv             Package a survival variable
coxph            Proportional Hazards Regression
lines.survfit    Add lines to a survival plot
plot.survfit     Plot method for survfit.
print.survfit    Short summary of a survival curve
survfit(formula, data, weights, subset, na.action,
        newdata, individual=F, conf.int=.95, se.fit=T,
            type=c("kaplan-meier","fleming-harrington", "fh2"),
            error=c("greenwood","tsiatis"),
            conf.type=c("log","log-log","plain","none"),
            conf.lower=c("usual", "peto", "modified"))
    'print.survfit', 'plot.survfit', 'lines.survfit',
    'summary.survfit', 'survfit.object' 'coxph', 'Surv', 'strata'.
fit <- survfit(formula = Surv(futime, fustat) ~ rx, data = ovarian)
```

Problem Set 10

Reading Assignment
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2394262/

Consider the *colon data in the R* package "survival". It gives adjuvant chemotherapy data for colon cancer. Levamisole is a low-toxicity compound previously used to treat worm infestations in animals; 5-FU is a moderately toxic (as these things go) chemotherapy agent. There are two records per person, one for recurrence (etype=1) and one for death (etype=2). Other important variables include:

rx:	Treatment - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU
sex:	1=male
age:	in years
time:	days until event or censoring
status:	censoring status

For the following, consider survival to be “Days until Death”, i.e., etype=2.

1. Using the Kaplan-Meier method, estimate the survival curve for each treatment group using the following methods:
   - Kaplan-Meier
   - Fleming-Harrington
2. For each case in 1, estimate the median survival time, using the estimated survival curves.