Survival Analysis

Bruce A Craig

Department of Statistics
Purdue University

Reading: Le - Section 2.2

Survival Analysis

- Statistical analysis of *time-to-event data*
  - *Survival analysis:* Patients survival time under diff trts
  - *Failure time analysis:* Lifetime of machines and parts
  - *Duration analysis:* Time to default on a credit card
  - *Event-history analysis:* Term used in sociology
- Why a special topic on survival analysis?
  - Non-Normal and skewed distributions
  - Censored and truncated data
  - Focus shift from mean to $P(\text{Time} > t + t_0 | \text{ Time} \geq t_0)$

Survival Function

- Consider continuous survival time T with
  - probability density function $f(t)$
  - cumulative probability function $F(t)$ where

$$F(t) = P(T \leq t) = \int_0^t f(s)ds$$

- The *survival function* of T is

$$S(t) \quad = \quad P(T > t) = 1 - F(t)$$

  - also called the survival rate
  - steeper CDF $\rightarrow$ shorter survivals
  - $S'(t) = -f(t)$

Mean Survival Time

- Mean survival time can be expressed in many ways

$$
\begin{aligned}
E(T) &= \int_0^\infty t f(t) \partial t = \int_0^\infty t \partial F(t) = \int_0^\infty t \partial [1 - S(t)] \\
&= \int_0^\infty \left[\int_0^t \partial x \right] \partial [1 - S(t)] = \int_0^\infty \left[\int_x^\infty \partial [1 - S(t)] \right] \partial x \\
&= \int_0^\infty S(x) \partial x \quad [= \textit{ area under } S(x) \ ]
\end{aligned}
$$

Hazard Function

- The *hazard function* of T is

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t}$$

  - instantaneous rate at which subjects experience the event given that they have survived up to time t
  - measure of 'proneness' to the event as function of time
  - $\lambda(t) \neq f(t)$: $\lambda(t)$ is conditional on survival to t
- Relates to the survival function

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \Big/ S(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}$$

$$= -\frac{\partial}{\partial t} \log S(t) \text{ (negative slope of log-survival)}$$

Cumulative Hazard Function

- The *cumulative hazard function* of T is defined:

$$\Lambda(t) = \int_0^t \lambda(s)\partial s$$

- Based on previous relationships
  - $\Lambda(t) = -\log S(t)$
  - $S(t) = \exp\{-\Lambda(t)\}$
  - $\lambda(t)$, $\Lambda(t)$ or $S(t)$ define the distribution f

Right-Censoring

- Consider a collection of n time-to-event observations
  - Survival time of ith subject is T_i
  - Censoring time of ith subject is C_i
- Observed values for the ith subject are

$$Y_i = \min(T_i, C_i) \text{ and } \delta_i = I_{\{T_i \leq C_i\}}$$

- Data reported as `(observedTime, hasEvent)`
- Assumptions:
  - C_i are predetermined and fixed, or are random and mutually independent
  - **Key: T_i and C_i are independent**

Example I: Remission Times

- 21 leukemia patients treated with *6-mercaptopurine*
- 21 matched controls
- Longer remission (i.e. disease-free time) is better
- Data set (`gehan`) available in `MASS`

```
> gehan
   pair time cens   treat
1     1    1    1 control
2     1   10    1    6-MP
3     2   22    1 control
4     2    7    1    6-MP
5     3    3    1 control
6     3   32    0    6-MP
7     4   12    1 control
8     4   23    1    6-MP
9     5    8    1 control
10    5   22    1    6-MP
11    6   17    1 control
12    6    6    1    6-MP
```

Non-Parametric Inference

Estimation of Survival Function

- Given a set of event times, the non-parametric estimate of $F(t)$ (and thus $S(t)$) is a right-continuous step function
- The Kaplan-Meier (K-M) estimator decomposes $S(t)$ into a chain of conditional probabilities
  - Accounts for censored observations at time intervals when they are not yet censored
  - Because of censoring, K-M estimator of survival doesn't always integrate to 1

    – stop plotting after the last observed event time

Kaplan-Meier Estimator

- Also called *product limit estimator*
- Best to reorganize the data by sorting on event times t

Distinct Event Times	t_0	t_1	$\cdots$	t_k
# of events at t_j	0	d_1	$\cdots$	d_k
# at Risk at t_j	n	n_1	$\cdots$	n_k

$$\begin{aligned} S(t) &= P(T > t) = P(T > t_1)P(T > t|T > t_1) \\ &= P(T > t_1)\left\{\prod_{j=2}^{k} P(T > t_j|T > t_{j-1})\right\} P(T > t|T > t_k) \end{aligned}$$

- K-M estimator $\hat{S}(t)$ produced as follows:
  - Estimate $P(T > t_j|T > t_{j-1})$ by $\hat{p}_j = (n_j - d_j)/n_j$
  - Estimate $S(t)$ by $\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j} = \prod_{j=1}^{k} \left(1 - \frac{d_j}{n_j}\right)$

Example I - R Functions

- Use function Surv to get data in censored form
  - '+' represents a censored time

```
> library(survival)
> Surv(gehan$time, gehan$cens)
 1  10  22   7   3  32+ 12  23   8  22  17   6   2  16
11  34+  8  32+ 12  25+  2  11+  5  20+  4  19+ 15   6
 8  17+ 23  35+  5   6  11  13   4   9+  1   6+  8  10+
```

- Use survfit to perform K-M estimation
  - Estimates $S(t)$ for each specified group:

```
> fit = survfit(Surv(time, cens) ~ treat, data=gehan)
> summary(fit)
```

Example I - Output

treat=6-MP

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
6	21	3	0.857	0.0764	0.720	1.000
7	17	1	0.807	0.0869	0.653	0.996
10	15	1	0.753	0.0963	0.586	0.968
13	12	1	0.690	0.1068	0.510	0.935
16	11	1	0.627	0.1141	0.439	0.896
22	7	1	0.538	0.1282	0.337	0.858
23	6	1	0.448	0.1346	0.249	0.807

treat=control

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	21	2	0.9048	0.0641	0.78754	1.000
2	19	2	0.8095	0.0857	0.65785	0.996
3	17	1	0.7619	0.0929	0.59988	0.968
4	16	2	0.6667	0.1029	0.49268	0.902
5	14	2	0.5714	0.1080	0.39455	0.828
8	12	4	0.3810	0.1060	0.22085	0.657
11	8	2	0.2857	0.0986	0.14529	0.562
12	6	2	0.1905	0.0857	0.07887	0.460
15	4	1	0.1429	0.0764	0.05011	0.407
17	3	1	0.0952	0.0641	0.02549	0.356
22	2	1	0.0476	0.0465	0.00703	0.322
23	1	1	0.0000	NaN	NA	NA

Manual Verification

- Order observed times in '6-MP' group

```
> trt = gehan$treat=='6-MP'
> orderOfTime = order(gehan$time[trt])
> Surv(gehan$time[trt], gehan$cens[trt])[orderOfTime]
 [1]  6   6   6   6+  7   9+ 10  10+ 11+ 13  16  17+ 19+ 20+ 22  23
[17] 25+ 32+ 32+ 34+ 35+
```

- Construct the K-M table (partial results)

Event times	6	7	10	13
d_j	3	1	1	1
n_j	21	17	15	12
$\hat{S}(t)$	$1 - 3/21$	$\hat{S}(6)(1 - 1/17)$	$\hat{S}(7)(1 - 1/15)$	$\hat{S}(10)(1 - 1/12)$
	0.857	0.807	0.753	0.690

Uncertainty Quantification

- Greenwood CI for $S(t)$
  - The variance of $\hat{S}(t)$ estimated by (Greenwood, 1926)

$$\widehat{Var}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

  - CI for $S(t)$ is $\hat{S}(t) \pm 1.96 \sqrt{\widehat{Var}[\hat{S}(t)]}$
  - No guarantee interval will be in $[0, 1]$
- Preferred interval based on $\log S(t)$
  - The variance of $\log \hat{S}(t)$ estimated by

$$\widehat{Var}\left(\log \hat{S}(t)\right) = \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

  - CI for $\log S(t)$ is $\log \hat{S}(t) \pm 1.96 \sqrt{\widehat{Var}\left(\log \hat{S}(t)\right)}$
  - CI of $S(t)$ is $[\exp\{L\}, \exp\{U\}]$

Deriving the Greenwood SE

- Recall the delta method
  - Want dist of $f(T)$ given $T \sim N(\mu, \sigma)$
  - $f(T) \approx f(\mu) + f'(\mu)(T - \mu)$ (1st order Taylor series)
  - $E(f(T)) \approx f(\mu)$ and $\text{Var}(f(T)) \approx [f'(\mu)]^2 \sigma^2$
- Let's assume $d_j \overset{ind}{\sim} B(n_j, p_j)$
  - $E(d_j/n_j) = p_j$ and $\text{Var}(d_j/n_j) = p_j(1 - p_j)/n_j$
- First, focus on variance of $\log[\hat{S}(t)] = \sum_{t_j \le t} \log\left(1 - \frac{d_j}{n_j}\right)$
  - The terms of this sum are not indep as d_j affects n_{j+1}
  - However, $E\left(\frac{d_j}{n_j} - p_j | d_1, \ldots, d_{j-1}\right) = 0$
  - Therefore, can show that variance(sum)=sum(variances)

Deriving the Greenwood SE

- Using $\hat{p}_j = d_j/n_j$

$$\begin{aligned}
\widehat{\text{Var}}\left(\log[\hat{S}(t)]\right) &= \sum_{t_j \le t} \widehat{\text{Var}}\left(\log(1-\hat{p}_j)\right) \\
\text{delta method} \quad &\approx \sum_{t_j \le t} \left(\frac{1}{1-\hat{p}_j}\right)^2 \frac{\hat{p}_j(1-\hat{p}_j)}{n_j} \\
&= \sum_{t_j \le t} \frac{\hat{p}_j}{(1-\hat{p}_j)n_j} \\
&= \sum_{t_j \le t} \frac{d_j}{(n_j-d_j)n_j}
\end{aligned}$$

- Apply delta method again to get $\widehat{\text{Var}}[\hat{S}(t)] = \exp\{\log[\hat{S}(t)]\}$

CI Based on Log-Log

- Consider $LCH(t) = \log\left(-\log[S(t)]\right)$
- Range now unrestricted $-\infty < LCH(t) < \infty$
- Construct CI for $LCH(t)$: $\widehat{LCH}(t) \pm 1.96 SE[\widehat{LCH}(t)]$
- Get CI for $S(t)$ by back-transform

$$\left([\hat{S}(t)]^{\exp\{1.96SE\}}, [\hat{S}(t)]^{\exp\{-1.96SE\}}\right)$$

- Applying delta method again

$$\widehat{\text{Var}}[\widehat{LCH}(t)] = \frac{1}{(\log[\hat{S}(t)])^2} \sum_{t_j \le t} \frac{d_j}{(n_j - d_j)n_j}$$

Survival CIs in R

```
> fit = survfit(Surv(time, cens) ~ treat, data=gehan, conf.type="log")
> summary(fit)
                treat=6-MP
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
    7     17       1    0.807  0.0869        0.653        0.996
   10     15       1    0.753  0.0963        0.586        0.968
   13     12       1    0.690  0.1068        0.510        0.935
   16     11       1    0.627  0.1141        0.439        0.896
   22      7       1    0.538  0.1282        0.337        0.858
   23      6       1    0.448  0.1346        0.249        0.807
> fit = survfit(Surv(time, cens) ~ treat, data=gehan, conf.type="log-log")
> summary(fit)
                treat=6-MP
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.620        0.952
    7     17       1    0.807  0.0869        0.563        0.923
   10     15       1    0.753  0.0963        0.503        0.889
   13     12       1    0.690  0.1068        0.432        0.849
   16     11       1    0.627  0.1141        0.368        0.805
   22      7       1    0.538  0.1282        0.268        0.747
   23      6       1    0.448  0.1346        0.188        0.680
```

Estimating the Cumulative Hazard

1. Can use the K-M estimator of $\hat{S}(t)$
   - $\hat{\Lambda}(t) = -\log \hat{S}(t)$
   - $\widehat{\text{Var}}(\hat{\Lambda}(t)) = \widehat{\text{Var}}(\log \hat{S}(t)) = \sum_{t_j \leq t} \frac{d_j}{(1-d_j)n_j}$
2. Alternatively, can use the Nelson-Aalen Estimator:

$$\tilde{\Lambda}(t) = \begin{cases} 0 & \text{if } t \leq t_1 \\ \sum_{j:t_j \leq t} \frac{d_j}{n_j} & \text{if } t \geq t_j \end{cases}$$

$$\widehat{\text{Var}}(\tilde{\Lambda}(t)) = \sum_{j:t_j \leq t} \frac{d_j}{n_j^2}$$

   - Use to obtain Flemington-Harrington estimator of $S(t)$

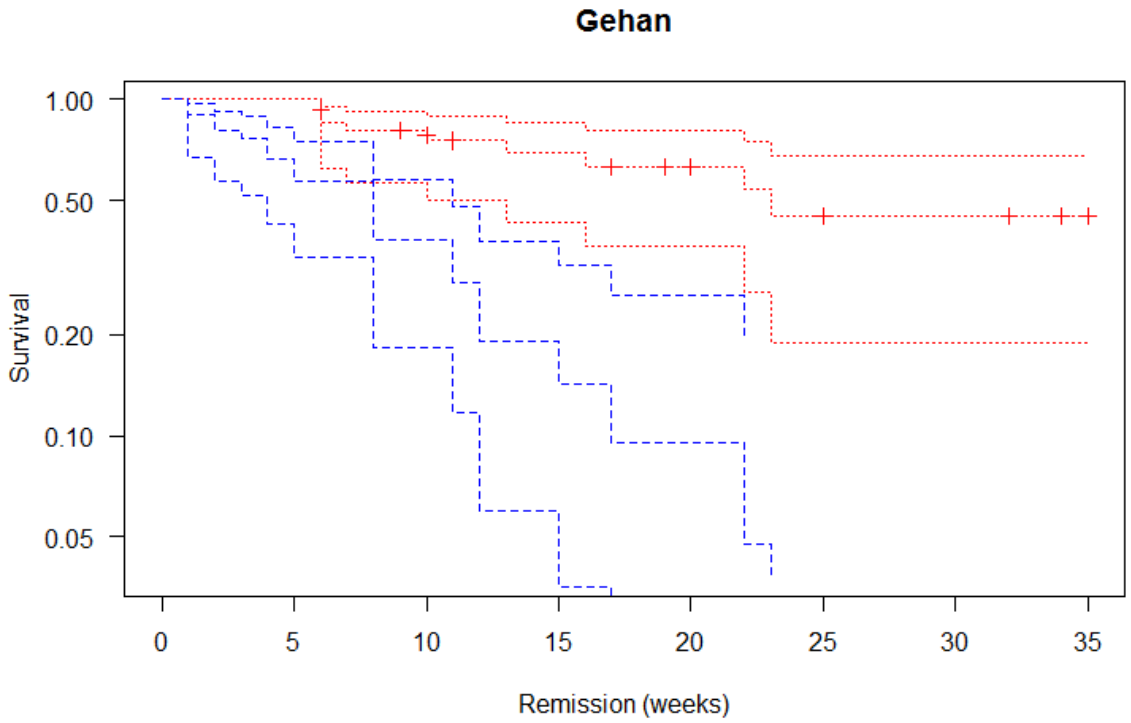
$$\tilde{S}(t) = \exp\left(-\tilde{\Lambda}(t)\right)$$

   - Has better small-sample-size performance
   - Used to check parameteric model selections

Plotting $\hat{S}(t)$

- Plot K-M curves and CI (default CI based on log)
- 'log=TRUE' plots y-axis on log scale

```
> plot(fit, conf.int=TRUE, lty=3:2, col=c("red","blue"), log=TRUE,
+     xlab="Remission (weeks)", ylab="Log-Survival", main="Gehan",
+     mark.time=T)
```



Log-Rank Test of Homogeneity

- Test the hypotheses

$$H_0 : S_1(t) = S_2(t) \quad \forall t \quad 0 < t < \infty$$
$$H_a : S_1(t) \neq S_2(t) \quad \text{for some } t$$

- Targets on the hazard (not survival) functions
- Adapted from test for stratified 2×2 tables
- Procedure:
  - Construct **pooled** and group tables with the k distinct event times

Event times	t_1	$\cdots$	t_k
Pooled	d_1	$\cdots$	d_k
Sample	n_1	$\cdots$	n_k
Sample	d_{11}	$\cdots$	d_{1k}
1	n_{11}	$\cdots$	n_{1k}
Sample	d_{21}	$\cdots$	d_{2k}
2	n_{21}	$\cdots$	n_{2k}

  - Note $d_j = d_{1j} + d_{2j}$, $n_j = n_{1j} + n_{2j}$, $j = 1, \cdots, k$

Log-Rank Test of Homogeneity

- At time t_j, can form 2×2 table

	Group		
Outcome	1	2	Total
Failure	d_{1j}	d_{2j}	d_j
Success	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
At risk	n_{1j}	n_{2j}	n_j

- Under H_o, $d_{1j}|n_j, d_j, n_{1j} \sim Hypergeometric(n_j, d_j, n_{1j})$

$$P(d_{1j} = d|n_j, d_j, n_{1j}) = \frac{\begin{pmatrix} d_j \\ d \end{pmatrix}\begin{pmatrix} n_j - d_j \\ n_{1j} - d \end{pmatrix}}{\begin{pmatrix} n_j \\ n_{1j} \end{pmatrix}}$$

$$E(d_{1j}|\text{marginals}) = n_{1j}\frac{d_j}{n_j}$$

$$\text{Var}(d_{1j}|\text{marginals}) = n_{1j}\frac{d_j}{n_j}\frac{n_j - d_j}{n_j}\frac{n_j - n_{1j}}{n_j - 1}$$

Log-Rank Test of Homogeneity

- Define test statistic

$$U = \sum_{j=1}^{k} \left(d_{1j} - \frac{n_{1j} d_j}{n_j} \right) = \sum_{j=1}^{k} (O_j - E_j)$$

- Assuming H_0 and independence among k 2×2 tables

$$\begin{aligned} E(U) &= 0 \\ \text{Var}(U) &= \sum_{j=1}^{k} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \end{aligned}$$

- Compute P-value using

$$\frac{U - 0}{\sqrt{\text{Var}(U)}} \sim N(0, 1)$$

Log-Rank Test in R

```
> survdiff(Surv(time, cens) ~ treat, data=gehan)

               N Observed Expected (O-E)^2/E (O-E)^2/V
treat=6-MP    21        9     19.3      5.46      16.8
treat=control 21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

- Reject H_0. Very strong evidence just looking at $\hat{S}(t)$'s
- Only rank of event times matter in this analysis
- Test does not adjust for differences in covariates and therefore can be inappropriate or misleading
  - OK here because individuals were matched

Log-Rank Test

- Power of the test depends on the number of observed failures
- Most powerful when detecting alternatives

$$H_a : S_2(t) = S_1(t)^{\exp\{\beta\}} \leftrightarrow h_2(t) = h_1(t)\exp\{\beta\}$$

- Can approximate power using

$$\Phi\left(\beta\sqrt{Dn_1n_2/n^2} - 1.96\right)$$

D is the expected number of failures

Targeting the Hazard Function

- Consider the test statistic U

$$\begin{aligned}
\sum_{j=1}^{k}\left(d_{1j}-\frac{n_{1j}d_j}{n_j}\right) &= \sum_{j=1}^{k}\frac{d_{1j}n_j-n_{1j}d_j}{n_j}\\
&= \sum_{j=1}^{k}\frac{d_{1j}n_{1j}+d_{1j}n_{2j}-n_{1j}d_{1j}-n_{1j}d_{2j}}{n_j}\\
&= \sum_{j=1}^{k}\frac{d_{1j}n_{2j}-n_{1j}d_{2j}}{n_j}\\
&= \sum_{j=1}^{k}\frac{n_{1j}n_{2j}}{n_j}\left(\frac{d_{1j}}{n_{1j}}-\frac{d_{2j}}{n_{2j}}\right)\\
&= \int_0^{\infty}\frac{n_1(s)n_2(s)}{n_1(s)+n_2(s)}(\hat{\lambda}_1(s)-\hat{\lambda}_2(s))\partial s
\end{aligned}$$

What if $p > 2$ Groups?

- Follow same basic procedure
  - Determine all unique event times
  - Construct pooled and individual group tables
  - At time t_j, now have a $2 \times p$ contingency table
- With $p > 2$, must consider **cond** hypergeometric dists

$$P(\mathbf{d}|\mathbf{n}, d_j) = P(d_{1j}|\mathbf{n}, d_j)P(d_{2j}|\mathbf{n}, d_j, d_{1j}) \cdots P(d_{(p-1)j}|\mathbf{n}, d_j, d_{1j}, ..., d_{(p-2)j})$$

- Can show

$$\begin{aligned} E(d_{ij}|\text{marginals}) &= n_{ij}\frac{d_j}{n_j} \\ \text{Var}(d_{ij}|\text{marginals}) &= n_{ij}\frac{d_j}{n_j}\frac{n_j - d_j}{n_j}\frac{n_j - n_{ij}}{n_j - 1} \\ \text{Cov}(d_{ij}, d_{kj}) &= -\frac{n_{ij}n_{kj}d_j(n_j - d_j)}{n_j^2(n_j - 1)} \end{aligned}$$

What if $p > 2$ Groups?

- For each event time t_j:
  - Observed vector $\mathbf{O}_j = \left(d_{1j}, d_{2j}, ..., d_{(p-1)j}\right)$
  - Mean vector $\mathbf{E}_j$ and covariance matrix $\mathbf{V}_j$ are compted using the marginals
- Compute the test statistic

$$(\mathbf{O} - \mathbf{E})' \mathbf{V}^{-1} (\mathbf{O} - \mathbf{E}) \sim \chi^2_{p-1}$$

  where each vector or matrix is the sum over the k distinct event times
- Similar extension can be done if desire is to stratify across subpopulations (e.g., stages of cancer)

Example: 3 Groups

```
> grp = factor(c(rep(1,6),rep(2,6),rep(3,6)))
> time = c( 9.0, 9.5, 9.0, 8.5,10.0,10.5,
+           10.0,12.0,12.0,11.0,12.0,10.5,
+           12.0,12.0,12.0,12.0,12.0,12.0)
> cens = c(1,1,1,1,1,1,1,1,0,1,1,1,1,0,0,0,0,0)
>
> ex2 = data.frame(grp,time,cens)
> survdiff(Surv(time,cens)~grp, data=ex2)
Call:
survdiff(formula = Surv(time, cens) ~ grp, data = ex2)

      N Observed Expected (O-E)^2/E (O-E)^2/V
grp=1 6        6     1.57   12.4463   17.2379
grp=2 6        5     4.53    0.0488    0.0876
grp=3 6        1     5.90    4.0660    9.4495

 Chisq= 20.4  on 2 degrees of freedom, p= 4e-05
```

Parametric Inference

Common Survival Distributions

- Exponential distribution:
  - $\lambda(t) = \lambda > 0$ (constant)
  - $S(t) = \exp\{-\lambda t\} \Rightarrow f(t) = -S'(t) = \lambda \exp\{-\lambda t\}$
- Weibull distribution:
  - $\lambda(t) = p\lambda^p t^{p-1}$ (power of t)
  - $S(t) = \exp\{-(\lambda t)^p\} \Rightarrow f(t) = \exp\{-(\lambda t)^p\} p\lambda^p t^{p-1}$
  - Is the Exponential distribution when $p = 1$
- Gompertz distribution:
  - $\lambda(t) = \alpha \exp\{\beta t\}$
  - Is the Exponential distribution when $\beta = 0$
- Gompertz-Makeham distribution:
  - $\lambda(t) = \lambda + \alpha \exp\{\beta t\}$;
  - Hazard increasing from $\alpha + \lambda$ at $t = 0$

The Likelihood Function

- Recall that $Y_i = \min(T_i, C_i)$ and $\delta_i = 1$ if $Y_i = T_i$
  - For T: define pdf $f(t)$, and $S(t) = P(T > t)$
  - For C, define pdf $g(t)$, and $R(t) = P(C > t)$
  - **Key assumption: T_i and C_i are independent**
- For observed event times (i.e., $\delta_i = 1$):

$$P(t \leq T < t + dt, C > t) = R(t)f(t)dt$$

- For censored times (i.e., $\delta_i = 0$):

$$P(t \leq C < t + dt, T > t) = S(t)g(t)dt$$

- The likelihood over all the observations

$$L = \prod_{i=1}^{n} [R(t_i)f(t_i)]^{\delta_i} [S(t_i)g(t_i)]^{1-\delta_i}$$

The Likelihood Function

- Can rewrite likelihood as

$$\begin{aligned} L &= \prod_{i=1}^{n} R(t_i)^{\delta_i} g(t_i)^{1-\delta_i} \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \\ &\propto \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \end{aligned}$$

- Because (T_i, C_i) are independent, we can ignore the first term when maximizing the likelihood wrt to the parameters of $S(t)$

MLE for Exponential Dist

- The likelihood is

$$\begin{aligned} L(\lambda) &\propto \prod_{i=1}^{n} f(t_i|\lambda)^{\delta_i} S(t_i|\lambda)^{1-\delta_i} \\ (using \lambda(t) = \frac{f(t)}{S(t)}) &= \prod_{i=1}^{n} \lambda(t_i|\lambda)^{\delta_i} S(t_i|\lambda) \\ &= \prod_{i=1}^{n} \lambda^{\delta_i} \exp\{-\lambda t_i\} \end{aligned}$$

- Maximize (log-)likelihood: first derivative

$$\begin{aligned} 0 &= \frac{\partial}{\partial \lambda} \log L = \frac{\partial}{\partial \lambda} \log \lambda \sum_i \delta_i - \lambda \sum_i t_i = \frac{\sum_i \delta_i}{\lambda} - \sum_i t_i \\ \hat{\lambda} &= \sum_i \delta_i / \sum_i t_i \text{ (deaths per unit time)} \end{aligned}$$

MLE for Exponential Dist

- Var($\hat{\lambda}$): second derivative

$$Var(\hat{\lambda}) = \left\{ E\left(-\frac{\partial^2}{\partial \lambda^2} \log L \right) \right\}^{-1} = \left\{ E\frac{\sum_i \delta_i}{\lambda^2} \right\}^{-1} = \frac{\lambda^2}{E\{\sum_i \delta_i\}}$$

- Plugging our estimates

$$\begin{aligned} \widehat{Var}(\hat{\lambda}) &= \frac{(\sum_i \delta_i)^2}{(\sum_i t_i)^2} \cdot \frac{1}{\sum_i \delta_i} \\ &= \frac{\sum_i \delta_i}{(\sum_i t_i)^2} \end{aligned}$$

Accounting for Covariates I

- Typically focus is on hazard function
- General form: $\lambda(t) = \lambda_0(t)\exp\{\mathbf{x}\boldsymbol{\beta}\}$
  - $\lambda_0(t)$: the baseline hazard
  - $\exp\{\mathbf{x}\boldsymbol{\beta}\}$: multiplicative effect, independent of time
- Assumption regarding censoring:
  - (T_i, C_i) are conditionally independent, given $\mathbf{X}_i$
- Assumption of proportional hazard implies:
  - Multiplicative effects of time and of covariates
  - The hazard ratio for subjects i and j with different covariates is constant over time

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)}{\lambda_0(t)} \frac{\exp\{\beta_0 + \beta_1 X_i\}}{\exp\{\beta_0 + \beta_1 X_j\}} = \exp\{(X_i - X_j)\,\beta_1\}$$

Accounting for Covariates I

- Exponential distribution, no predictors:
  - $\lambda(t) = \lambda$
- Exponential distribution, one predictor X:
  - $\lambda(t) = \exp\{\beta_0 + \beta_1 X\} = \exp\{\beta_0\}\exp\{\beta_1 X\}$
  - $\exp\{\beta_0\}$ is the baseline hazard $\lambda_0(t)$
- Weibull distribution, no predictors:
  - $\lambda(t) = p\lambda^p t^{p-1}$
- Weibull distribution, one predictor X:
  - $\lambda(t) = pt^{p-1}\exp\{(\beta_0 + \beta_1 X)p\}$
  - λ is replaced with $\exp\{\beta_0 + \beta_1 X\}$ in overall hazard
  - $pt^{p-1}\exp\{p\beta_0\}$ is the baseline hazard $\lambda_0(t)$

Proportional Hazard Assumption

- Impact of covariates on the survival function:

$$\begin{aligned} \Lambda(t) &= \int_0^t \lambda_0(s)\exp\{\mathbf{x}\boldsymbol{\beta}\}ds = \Lambda_0(t)\exp\{\mathbf{x}\boldsymbol{\beta}\} \\ \log\Lambda(t) &= \log[-\log S(t)] = \log\Lambda_0(t) + \mathbf{x}\boldsymbol{\beta} \end{aligned}$$

- When predictor is categorical (i.e., treatment indicator), can compare assumed $\lambda(t)$ with KM estimates of $S(t)$
  - Plot $\log[-\log\hat{S}(t)]$ for different groups
  - Assumption reasonable if roughly parallel lines
- Exponential model and a binary predictor:

$$\log\Lambda(t) = \log\left(t\exp\{\beta_0+\beta_1 X\}\right) = \log t + \beta_0 + \beta_1 X$$

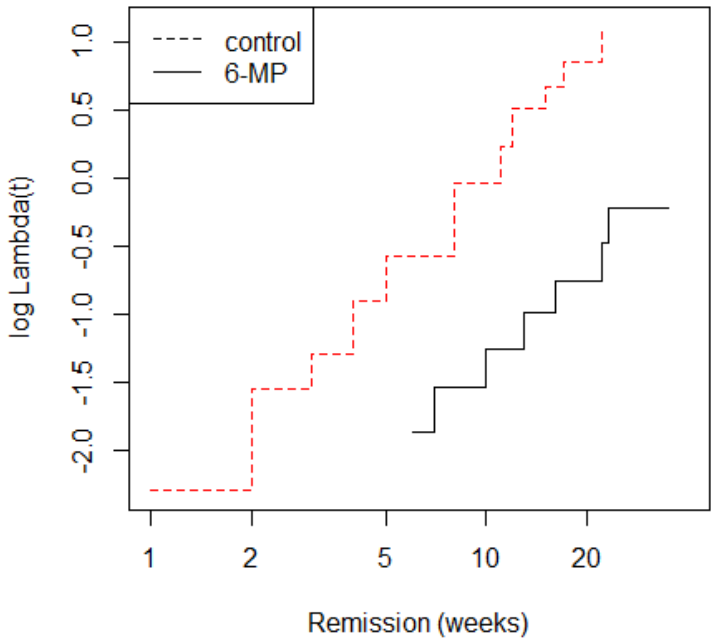
- Weibull model and a binary predictor:

$$\log\Lambda(t) = \log\left(t^p\exp\{p\beta_0+p\beta_1 X\}\right) = p\log t + p\beta_0 + p\beta_1 X$$

Checking Proportional Hazard Assumption

- Graphically check the Gehan data set

```
> plot(fit, lty=1:2, col=1:2, fun="cloglog",xlab="Remission (weeks)",
+ ylab="log Lambda(t)",xlim=c(1,40))
> legend("topleft", c("control", "6-MP"), lty=2:1)
```



Accounting for Covariates II

- Instead of multiplying the hazard by some constant, could have covariate accelerate or decelerate the life course
- This is called Accelerated Failure Time model
- In terms of the survival function:

$$S(t) = S_0\left(t \exp\{\mathbf{x}\boldsymbol{\beta}\}\right)$$

  - S_0 is the baseline survival function
  - Covariates accelerate or contract the time to event
  - Survival time $T_0 = T \exp\{\mathbf{x}\boldsymbol{\beta}\}$ has a fixed distribution
  - Additive difference in times on the log scale

$$\log T \quad = \quad \log T_0 - \mathbf{X}\boldsymbol{\beta}$$

  - Weibull / Exponential are the only distributions that can be simultaneously (and equivalently) specified as proportional hazard and accelerated failure models.

Accelerated Failure Model

- Exponential distribution:
  - $\lambda(t) = \lambda$, $S(t) = \exp\{-\lambda t\}$
  - Incorporating covariates:
    - $S(t) = S_0(t \exp\{\beta_0 + \beta_1 X\})$
    - If $T_0 \sim exp(1)$, and T are the observed times

$$\log T = -\beta_0 - \beta_1 X + \log T_0$$

- In R, use regression to estimate model:

$$\log T = \beta_0 + \beta_1 X + \sigma \log \varepsilon$$

  - σ is a scale parameter fixed at $\sigma = 1$
  - $\varepsilon \sim Exp(1)$
  - Use opposite sign of $\hat{\beta}$'s to estimate survival:

$$\hat{S}(t) = S_0(t \exp\{-\hat{\beta}_0 - \hat{\beta}_1 X\})$$

Fitting Exponential AFT Model

```
> fit.exp = survreg(Surv(time,cens)~treat,data=gehan,dist="exponential")
> summary(fit.exp)

              Value Std. Error     z       p
(Intercept)   3.686      0.333 11.06 < 2e-16
treatcontrol -1.527      0.398 -3.83 0.00013

Scale fixed at 1

Exponential distribution
Loglik(model)= -108.5   Loglik(intercept only)= -116.8
Chisq= 16.49 on 1 degrees of freedom, p= 4.9e-05
```

- Parameter interpretation:
  - $-\beta_{\texttt{treatcontrol}} = 1.53 > 0$
  - time until remission is shorter for controls
- Predicted survival times $T > 10$ for trt and control:
  - `exp(-10*exp(-3.69))=0.779`
  - `exp(-10*exp(-3.69+1.53))=0.316`
  - comparable to KM estimates of 0.753 and $(0.286, 0.381)$

Accelerated Failure Model

- Weibull distribution:
  - $\lambda(t) = \lambda^p p t^{p-1}$, $S(t) = \exp\{-(\lambda t)^p\}$
  - Can be viewed as the survival function of the exponential random variable $T' = T^p \sim exp(\lambda^p)$
  - Incorporating covariates:
    - Identical to the exponential: $S(t') = S_0(t' \exp\{\beta_0 + \beta_1 X\})$
    - If $T_0 \sim exp(1)$, and T' are the observed times

$$\log T_0 = \log T' + \beta_0 + \beta_1 X$$

    - Returning to the original notation $T' = T^p$:

$$\log T = -\frac{1}{p}\beta_0 - \frac{1}{p}\beta_1 X + \frac{1}{p}\log T_0$$

Fitting Weibull AFT Model

- In R, Weibull is the default distribution:

$$\log T = \beta_0 + \beta_1 X + \sigma \cdot \log \varepsilon$$

  - σ is a scale parameter, $\sigma = \frac{1}{p}$, $\varepsilon \sim exp(1)$
  - Use $-\hat{\beta}/\hat{\sigma}$ to estimate survival:

$$\hat{S}(t) = S_0(t^{1/\sigma} \exp\{-\hat{\beta}_0/\sigma - \hat{\beta}_1 X/\sigma\})$$

```
> fit.weibull <- survreg(Surv(time,cens)~treat, data=gehan)
> summary(fit.weibull)
               Value Std. Error     z        p
(Intercept)    3.516      0.252 13.96 2.61e-44
treatcontrol  -1.267      0.311 -4.08 4.51e-05
Log(scale)    -0.312      0.147 -2.12 3.43e-02

Scale= 0.732
Weibull distribution
Loglik(model)= -106.6   Loglik(intercept only)= -116.4
Chisq= 19.65 on 1 degrees of freedom, p= 9.3e-06
```

Fitting Weibull AFT Model

- $-\beta_{\texttt{treatcontrol}} = 1.267 > 0$
- time until remission is shorter for controls
- Predicted survival time $T > 10$:
  - `exp(-10^(1/0.732)*exp(-3.516/0.732))=0.731`
  - `exp(-10^(1/0.732)*exp((-3.516+1.267)/0.732))=0.341`
  - These again are quite comparable to the KM estimates

Variable Selection and Prediction

- Should `pair` be included as a blocking factor?

```
> anova(survreg(Surv(time,cens) ~ treat, data=gehan),
+       survreg(Surv(time,cens)~factor(pair)+treat, data=gehan))
                  Terms Resid. Df -2*LL Test Df Deviance Pr(>Chi)
1                 treat        39 213.2      NA       NA       NA
2 factor(pair) + treat        19 181.3    = 20    31.82   0.0453
```

- Prediction for the median survival
  - On the linear predictor scale

```
> fit.weibull.nopairs = survreg(Surv(time,cens)~treat, data=gehan)
> pred.ctrl = predict(fit.weibull.nopairs,data.frame(treat=
+             'control'),type="uquantile",p=0.5, se=TRUE)
```

  - On the survival function scale

```
> exp(c(L=pred.ctrl$fit-2*pred.ctrl$se.fit,
+       U=pred.ctrl$fit+2*pred.ctrl$se.fit) )
 L.1   U.1
 5.046 10.396
```

Semi-Parametric Inference

Cox Proportional Hazards Model

- Alternative approach to include covariates
- Still assume C_i and T_i conditionally indep given $\mathbf{X}_i$
- Still consider there are k distinct event times (i.e., $\delta_i = 1$) such that $t_1 < t_2 < \cdots < t_k$
  - But now assume no ties (i.e., $t_i \neq t_{i'}$ if $\delta_i = \delta_{i'} = 1$)
- Also return to proportional hazard model....

$$\lambda(t) = \lambda_0(t) \exp\{\mathbf{x}\boldsymbol{\beta}\}$$

  - But treat $\lambda_0(t)$ as a nuisance function
- Approach proposed by Cox (1975):
  - Given proportional hazard model and conditional on an event occurring at time t, the probability it is subject i that fails out of all those still as risk is

$$\frac{\lambda_0(t) \exp\{\mathbf{x_i}\boldsymbol{\beta}\}}{\sum_l I(T_l \geq t)\lambda_0(t) \exp\{\mathbf{x_l}\boldsymbol{\beta}\}} = \frac{\exp\{\mathbf{x_i}\boldsymbol{\beta}\}}{\sum_l I(T_l \geq t) \exp\{\mathbf{x_l}\boldsymbol{\beta}\}}$$

Cox Proportional Hazards Model

- Conditional probability is the proportion of the total sum of hazards rates for those at risk that is attributed to subject i
- Because of the proportional hazards assumption
  - This conditional probability is free of $\lambda_0(t)$
  - It's also proportional to the multinomial likelihood
- Called semi-parametric because no need to specify a form or shape for $\lambda_0(t)$
- Partial likelihood is the product of these observed conditional probabilities given the risk sets at the distinct event times

Estimation of β

- Maximize the **partial likelihood**

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp\{\mathbf{x_i}\beta\}}{\sum_l I(T_l \geq t_j)\exp\{\mathbf{x_l}\beta\}} = \prod_{i=1}^{n} \left\{ \frac{\exp\{\mathbf{x_i}\beta\}}{\sum_l I(T_l \geq t_i)\exp\{\mathbf{x_l}\beta\}} \right\}^{\delta_i}$$

  - Cox argued that the partial likelihood has all properties of the likelihood (i.e., contains all the info about β)
  - Standard inference (Wald test, LRT) applies
  - $\exp\{\mathbf{x}_i\beta\}$ is the parametric part of the model
- Interpretation of β when $x = 1$ if exposure
  - Define hazard ratio as

$$HR = \frac{\lambda(t \mid \text{exposure})}{\lambda(t \mid \text{no exposure})} = \frac{\lambda_0(t)\exp\{\beta_1\}}{\lambda_0(t)} = \exp\{\beta_1\}$$

  - Wald CI for β : $[L, U] \Rightarrow$ CI for HR : $[\exp\{L\}, \exp\{U\}]$

Prediction

- For prediction, we need $\Lambda_0(t) = -\log[S(t)]$
- Recall under proportional hazards model that

$$\begin{aligned} \Lambda(t) &= \Lambda_0(t)\exp\{\mathrm{x}\boldsymbol{\beta}\} \\ &\downarrow \\ S(t) &= [\exp\{-\Lambda_0(t)\}]^{\exp\{\mathrm{x}\boldsymbol{\beta}\}} \end{aligned}$$

- Estimation of $\Lambda_0(t)$:
  - Related to nonparametric part of Cox model
  - Breslow's estimator:

$$\begin{aligned} \hat{\Lambda}_0(t) &= \sum_{j:t_j \le t} \hat{\lambda}_0(t_j) \\ &= \sum_{j:t_j \le t} \frac{1}{\sum_{i=1}^{n} I(T_i \ge t_j)\exp\{\mathbf{x}_i\hat{\boldsymbol{\beta}}\}} \end{aligned}$$

Prediction

- Breslow's estimator derived by maximizing full likelihood with respect to $\lambda_o(t)$ when $\hat{\boldsymbol{\beta}}$ replaces $\boldsymbol{\beta}$

$$L \propto \prod_{i=1}^{n} \left[\lambda_0(t_i)\exp\{\mathbf{x}_i\hat{\boldsymbol{\beta}}\}\right]^{\delta_i} \exp\{-\int_0^{t_i} \lambda_0(s)\exp\{\mathbf{x}_i\hat{\boldsymbol{\beta}}\}ds\}$$

  - Second component tells us to make $\lambda_0(t)$ small
  - First component tells us to make $\lambda_0(t)$ large
  - Leads to $\hat{\lambda}_0(t) = 0$ except at event times ($\delta_i = 1$)
- Take log, differentiate with respect to $\lambda_0(t_j)$ to get $\hat{\lambda}_0(t_j)$

Extensions

- Tied event times are conceptually difficult
  - For continuous distributions, $P(t_i = t_{i'}) = 0$
- Various methods to adjust

Method	R Options	Comment
Exact	`method="exact"`	accurate, long, considers all possible orders
Efron's	`method="efron"`	approximate, better, alters denominator
Breslow's	`method="breslow"`	approximate, alters denominator

- Different baseline hazards for different groups
  - R option `strata(predictor)`

Using R

- If feasible, use exact method to account for ties

```
> fit.cox1 <-coxph(Surv(time,cens)~treat, method="exact", data=gehan)
> summary(fit.cox1)

               coef exp(coef) se(coef)     z Pr(>|z|)
treatcontrol 1.6282    5.0949   0.4331 3.759  0.00017 ***

             exp(coef) exp(-coef) lower .95 upper .95
treatcontrol     5.095     0.1963      2.18     11.91

Concordance= 0.69  (se = 0.041 )
Likelihood ratio test= 16.25  on 1 df,   p=6e-05
Wald test            = 14.13  on 1 df,   p=2e-04
Score (logrank) test = 16.79  on 1 df,   p=4e-05
```

- Given $\hat{\beta} > 0$, survival curve drops more steeply to 0 for treatment group

Using R

- Consider adding pair as a blocking factor

```
> fit.cox2 <- coxph(Surv(time,cens)~treat+factor(pair),method="exact",
+              data=gehan)
> summary(fit.cox2)
                     coef exp(coef)  se(coef)      z Pr(>|z|)
treatcontrol     3.314679 27.513571  0.742620  4.463 8.06e-06 ***
factor(pair)2   -5.015219  0.006636  1.550131 -3.235 0.001215 **
:
factor(pair)21 -3.651486  0.025953  1.516753 -2.407 0.016065 *
-
                exp(coef) exp(-coef) lower .95 upper .95
treatcontrol   27.513571    0.03635 6.418e+00 117.94128
factor(pair)2   0.006636  150.68919 3.180e-04   0.13848
:
factor(pair)21  0.025953   38.53190 1.328e-03   0.50727

Concordance= 0.826  (se = 0.055 )
Likelihood ratio test= 45.51  on 21 df,   p=0.001
Wald test            = 27.42  on 21 df,   p=0.2
Score (logrank) test = 39.73  on 21 df,   p=0.008
```

Using R

- LR test compares `log(partial likelihoods)`, but the test has similar properties

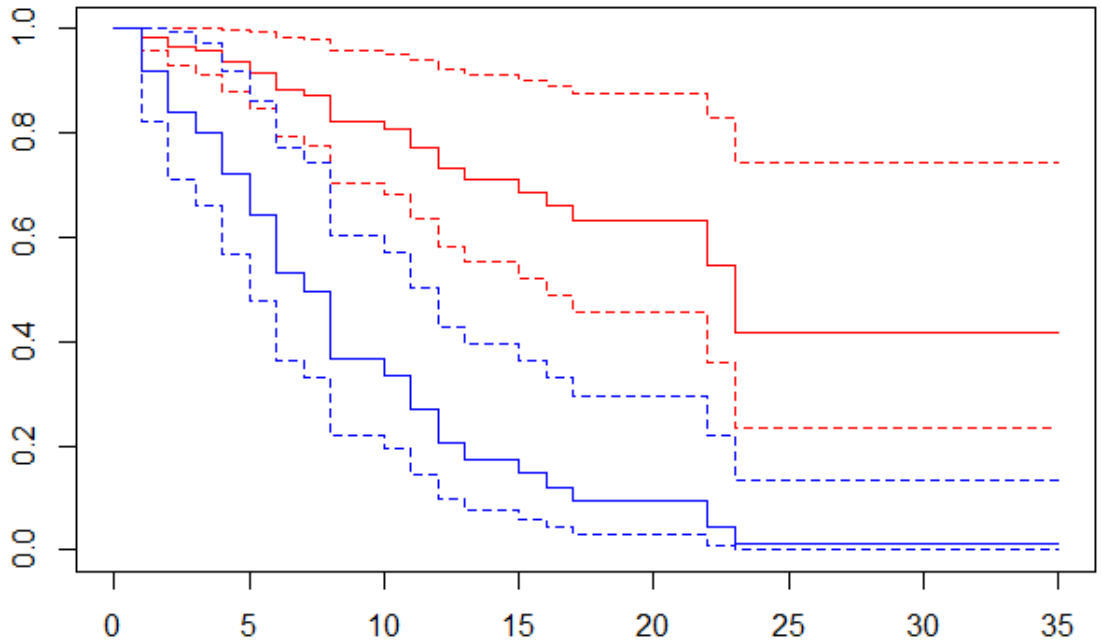
```
> anova(fit.cox1, fit.cox2, test="Chisq")

Analysis of Deviance Table
 Cox model: response is  Surv(time, cens)
 Model 1: ~ treat
 Model 2: ~ treat + factor(pair)
   loglik  Chisq Df P(>|Chi|)
1 -74.543
2 -59.915 29.256 20   0.08283 .
```

- Result does not suggest inclusion of blocking factor is needed

Visualize Model Fit

- Survival curve and CI

```
> plot(survfit(fit.cox1,newdata=data.frame(treat=c('6-MP'))), col='red')
> lines(survfit(fit.cox1,newdata=data.frame(treat=c('control'))), col='blue')
```



Model Diagnostics: Residuals

- Martingale residuals
  - $r_i = \delta_i - \hat{\Lambda}(t_i)$ (Flemming & Harrington, 1991)
  - Fit the Cox model with all predictors except the one of interest, and plot the martingale residuals against the predictor of interest
  - A loess curve fitted to the residuals suggests the functional form of the predictor – Works best for continuous predictors
  - max(r_i)=1, skewed towards negative values
  - In R: `resid(fit.cox1)`
- Deviance residuals
  - $sign(r_i)\sqrt{2[-r_i - \delta_i log(\delta_i - r_i)]}$
  - A transformation of martingale residuals
  - Reduces skewness
  - Help find data points with poor fit