Contents

What is survival analysis?

- Survival analysis[1] is a class of statistical methods for studying the occurrence and timing of **events**, e.g., death, onset of disease, return of a tumor, equipment failures, customer churn/purchase, earthquakes, stock market crashes, recessions, revolutions, job terminations, divorces, promotions, retirements, re-arrests after parole, leaving a web site
- A.K.A. history analysis (sociology), reliability analysis (engineering), failure-time analysis (engineering), duration analysis (economics), transition analysis (economics)
- Survival analysis can model data with two features that are difficult to handle with conventional methods:
  - **censoring**: we may not observe the "event" for all observations
  - **time-dependent covariates**: predictor variables may change during the study

  All survival analysis methods allow for censoring. Many allow for time-dependent covariates

[1] Here's a short list of references:

- Allison, Paul (1995), *Survival Analysis using the SAS System*, SAS Institute (newer 2nd edition too)
- Cox, D.R. and Oakes, D. (1984), *Analysis of Survival Data*, Chapman and Hall
- Crowder, M.J., Kimber, A.C., Smith, R.L., and Sweeting, T.J. (1991), *Statistical Analysis of Reliability Data*, Chapman and Hall
- Kalbfleisch, J.D. and Prentice, R.L. (1980), *The Statistical Analysis of Failure Time Data*, Wiley

Survival Analysis Terms

- T random variable giving the time of the event
- *Cumulative distribution function*: $F(t) = \mathsf{P}(T \le t)$
- *Surviver function*[2]: $S(t) = \mathsf{P}(T > t) = 1 - F(t)$
- *Probability density function* (PDF) is

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$$

- *Hazard function*:

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathsf{P}(t \le T < t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}$$

  Indicates the "proneness to failure" in the instant after time t or "instantaneous risk that the event will occur"

- The *cumulative hazard function* is

$$H(t) = \int_0^t h(u)\, du = -\log S(t)$$

Survival analysis allows us to study these functions and how covariates affect their shape/level.

[2] Books are inconsistent in where to put the equals sign. Some define $F(t) = \mathsf{P}(T < t)$ and $S(t) = \mathsf{P}(T \ge t)$.

Example: Exponential Distribution

- The PDF is $f(t) = \lambda e^{-\lambda t}$, $t \geq 0$, $\lambda > 0$
- The CDF is $F(t) = \int_0^t \lambda e^{-\lambda x} dx = 1 - e^{-\lambda t}$
- The survival function is $S(t) = 1 - F(t) = e^{-\lambda t}$
- The hazard function is

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

  which is **constant** over time

- Exponentials have the **memoryless property**: the probability of a unit that has survived until time t survives an additional u is independent of t:

$$\begin{aligned} \mathsf{P}(T > t + u | T > t) &= \frac{\mathsf{P}(T > t + u)}{\mathsf{P}(T > t)} \\ &= \frac{S(t+u)}{S(t)} = \frac{e^{\lambda(t+u)}}{e^{-\lambda t}} = e^{-\lambda u} \end{aligned}$$

- Here we assume the PDF and derive the other functions. Survival models estimate parameters (e.g., λ), or estimate functions without assuming a parametric form (like a histogram)

Example: Weibull Distribution

The *Weibull distribution* has survival function

$$S(t) = \exp\left[-(t/\alpha)^{\beta}\right], \quad \text{for } t \geq 0,$$

where $\beta > 0$ determines the **shape** and α is a **scaling** factor. We can derive the other functions: $F(t) = 1 - S(t)$,

$$f(t) = -\frac{dS(t)}{dt} = \beta\alpha^{-\beta}t^{\beta-1}S(t), \quad \text{and}$$

$$h(t) = \frac{f(t)}{S(t)} = \beta\alpha^{-\beta}t^{\beta-1}.$$

The **hazard** has the shape of a *power function*, e.g.,

- $\beta = 1 \Longrightarrow h(t) = t^0/\alpha = 1/\alpha$, which is constant. Thus the exponential distribution is a special case of the Weibull.
- $\beta = 2 \Longrightarrow h(t) = 2t^1/\alpha^2 = k_1 t$, which increases *linearly* with t. This is also called the *Rayleigh distribution*
- $\beta = 3 \Longrightarrow h(t) = 3t^2/\alpha^3 = k_2 t^2$, which shows increasing returns.
- $\beta = 1.5 \Longrightarrow h(t) = 1.5t^{0.5}/\alpha^{1.5} = k_3\sqrt{t}$, which shows decreasing returns.

Estimating Functions With Survival Analysis

Survival analysis is a set of methods for understanding questions about the time when some "event" occurs (T) such as:

Q1 When is the event, e.g., cancelation or repurchase, likely to occur? Are there times when the event is more or less likely to happen? How long until we expect the event to occur?

Q2 How do **static characteristics** of customers affect the probability of T? e.g., acquisition source, demographics, and the length of the initial contract. Are customers acquired from one channel (e.g., telemarketing) systematically more likely to cancel than those from another (e.g., direct mail)? Are young people more likely to cancel than old?

Q3 How do things that happen during the relationship (**time-dependent covariates**) affect the probability of T?

There are many "survival analysis" methods

- Nonparametric "product-moment" estimates, e.g., Kaplan-Meier (KM) (answers Q1)
- Accelerated failure-time (AFT) model (Q1, Q2)
- Discrete-time survival model (Q1, Q2, Q3)
- Cox Proportional Hazard model (Q1, Q2, Q3)

Nonparametric "Product-Moment" Estimates

- Let n_t be the number of customers "at risk" at time t
- Let d_t be the number of customers who cancel at time t
- The product-moment estimate of the survival function is

$$\hat{S}(t) = \prod_{i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

- The *Kaplan-Meier* (KM) method counts all people who are censored at t as being at risk
- Intuition: if we had discrete time periods (e.g., months) where the probability of retaining a customer in period t is π_t then $S(t) = \mathsf{P}(T > t) = \prod_{i=1}^{t} \pi_i$. Note $1 - d_i/n_i$ estimates π_i
- Available in R with the `survfit` function
- Specify dependent variable as `Surv(T, event)`, where `event` equals one if the event happened and 0 if censored, and `T` is the time of the event/censoring.

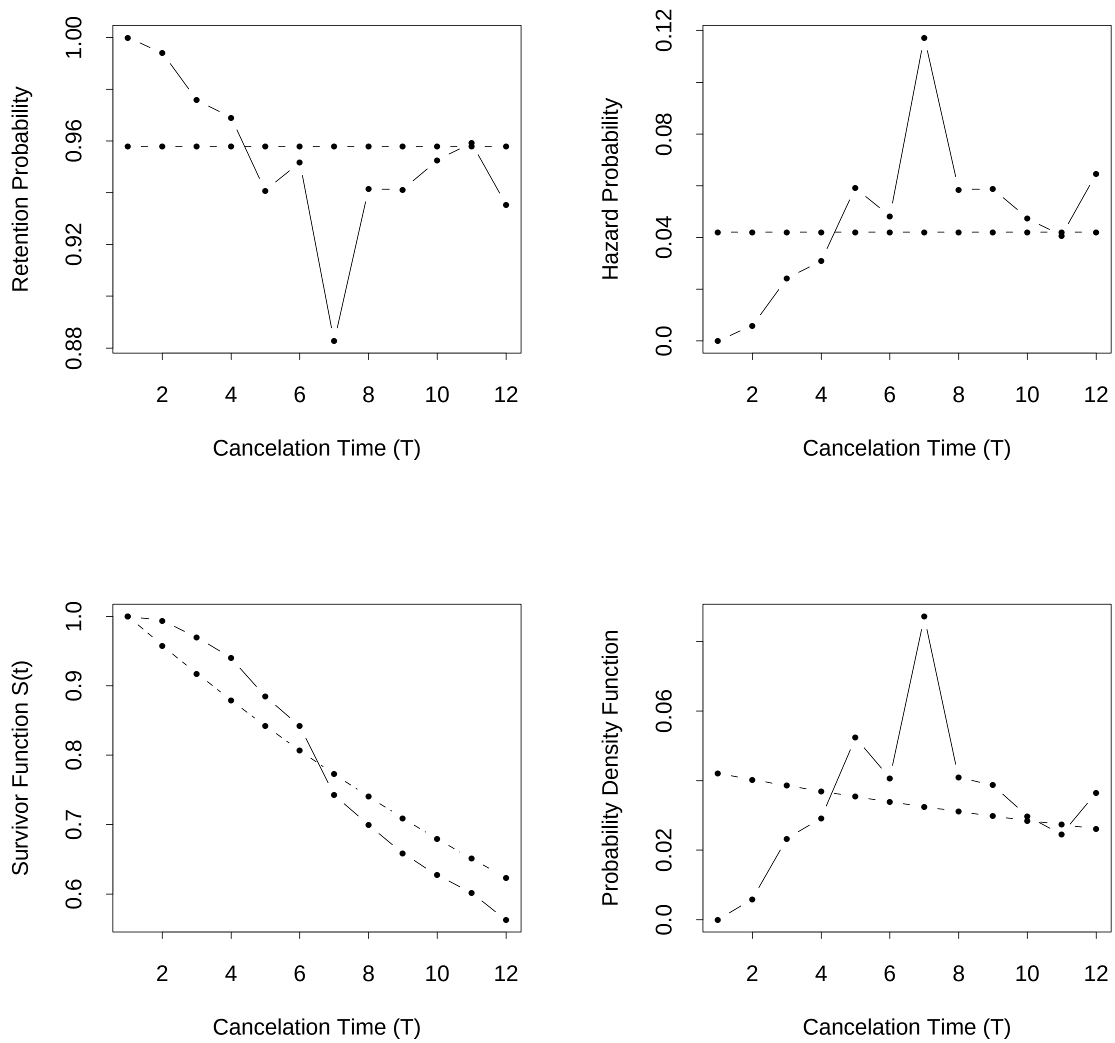
Educational Service Example

A Japanese cram school acquires junior high school children as subscribers to its service. For a sample of kids acquired within the past year, the following are the times of cancelation or censoring:

	Time of Cancelation/Censoring												
Censor	1	2	3	4	5	6	7	8	9	10	11	12	Total
No	0	4	16	20	37	28	61	24	19	13	10	13	245
Yes	3	0	2	1	7	33	49	63	30	16	34	188	426
Total	3	4	18	21	44	61	110	87	49	29	44	201	671

			Kaplan-Meier			Life Table	
t	Number Cancel d_t	Number Censor c_t	Number at Risk n_t	Retention Rate $1 - d_t/n_t$	Survivor Function $S(t)$	Number at Risk n_t	Survivor Function $S(t)$
1	0	3	671	1.0000	1.0000	669.5	1.0000
2	4	0	668	0.9940	0.9940	668	0.9940
3	16	2	664	0.9759	0.9701	663	0.9700
4	20	1	646	0.9690	0.9400	645.5	0.9400
5	37	7	625	0.9408	0.8844	621.5	0.8840
6	28	33	581	0.9518	0.8418	564.5	0.8402
7	61	49	520	0.8827	0.7430	495.5	0.7367
8	24	63	410	0.9415	0.6995	378.5	0.6900
9	19	30	323	0.9412	0.6584	308	0.6474
10	13	16	274	0.9526	0.6271	266	0.6158
11	10	34	245	0.9592	0.6015	228	0.5888
12	13	188	201	0.9353	0.5626	107	0.5173

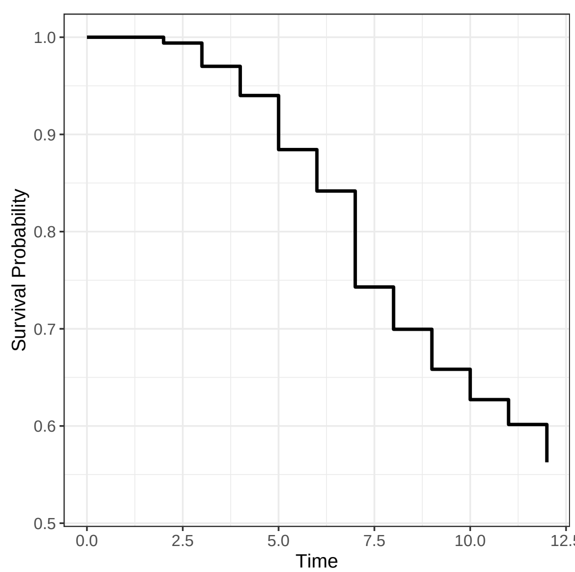
Educational Service Example Continued



Product-Moment Estimates in R

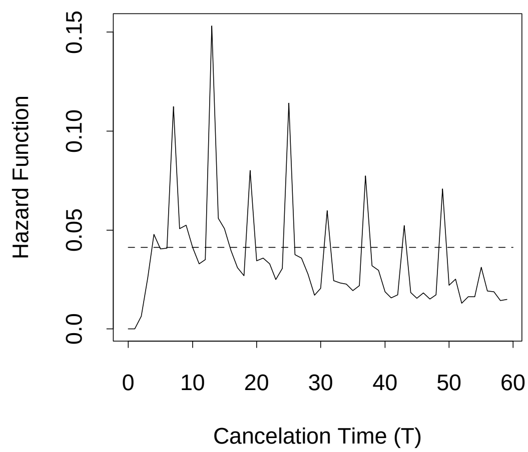
```
library(survival)
dat= read.table("service1yr.txt", header=T)
dat = data.frame(
  bigT = c(2:12, 1, 3:12),
  cancel = c(rep(1,11), rep(0,11)),
  count = c(4,16,20,37,28,61,24,19,13,10,13,3,2,1,7,33,49,63,30,16,34,188)
)
fit = survfit(Surv(bigT, cancel) ~ 1, data=dat, weight=count)
summary(fit)
Call: survfit(formula = Surv(bigT, cancel) ~ 1, data = dat, weights = count)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2    668       4    0.994 0.00299        0.988        1.000
    3    664      16    0.970 0.00659        0.957        0.983
    4    646      20    0.940 0.00919        0.922        0.958
    5    625      37    0.884 0.01239        0.860        0.909
    6    581      28    0.842 0.01417        0.814        0.870
    7    520      61    0.743 0.01725        0.710        0.778
    8    410      24    0.700 0.01838        0.664        0.736
    9    323      19    0.658 0.01958        0.621        0.698
   10    274      13    0.627 0.02048        0.588        0.669
   11    245      10    0.602 0.02118        0.561        0.645
   12    201      13    0.563 0.02239        0.520        0.608
```

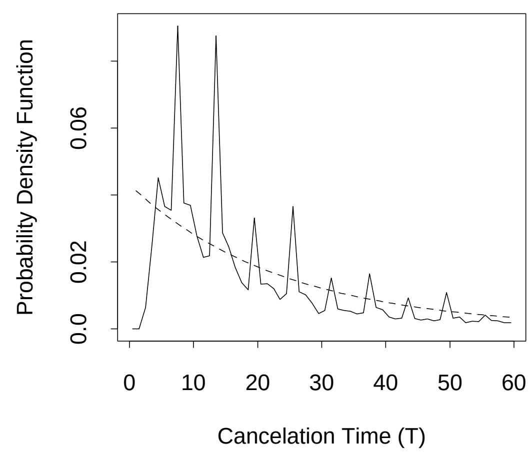
```
plot(fit) # basic version

library(ggsurvfit) # ggplot
fit %>%
  ggsurvfit(size = 1) +
  add_confidence_interval()
```

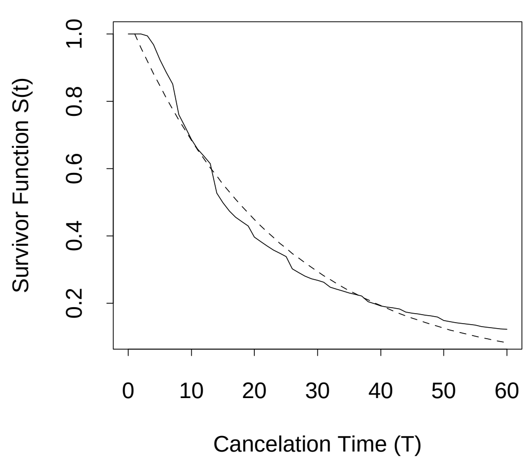


Service Provider, Five Years

```
service5yr = read.table("service5yr.txt", header=T)
fit2 = survfit(Surv(bigT, cancel) ~ 1, data=service5yr, weight=count)
```



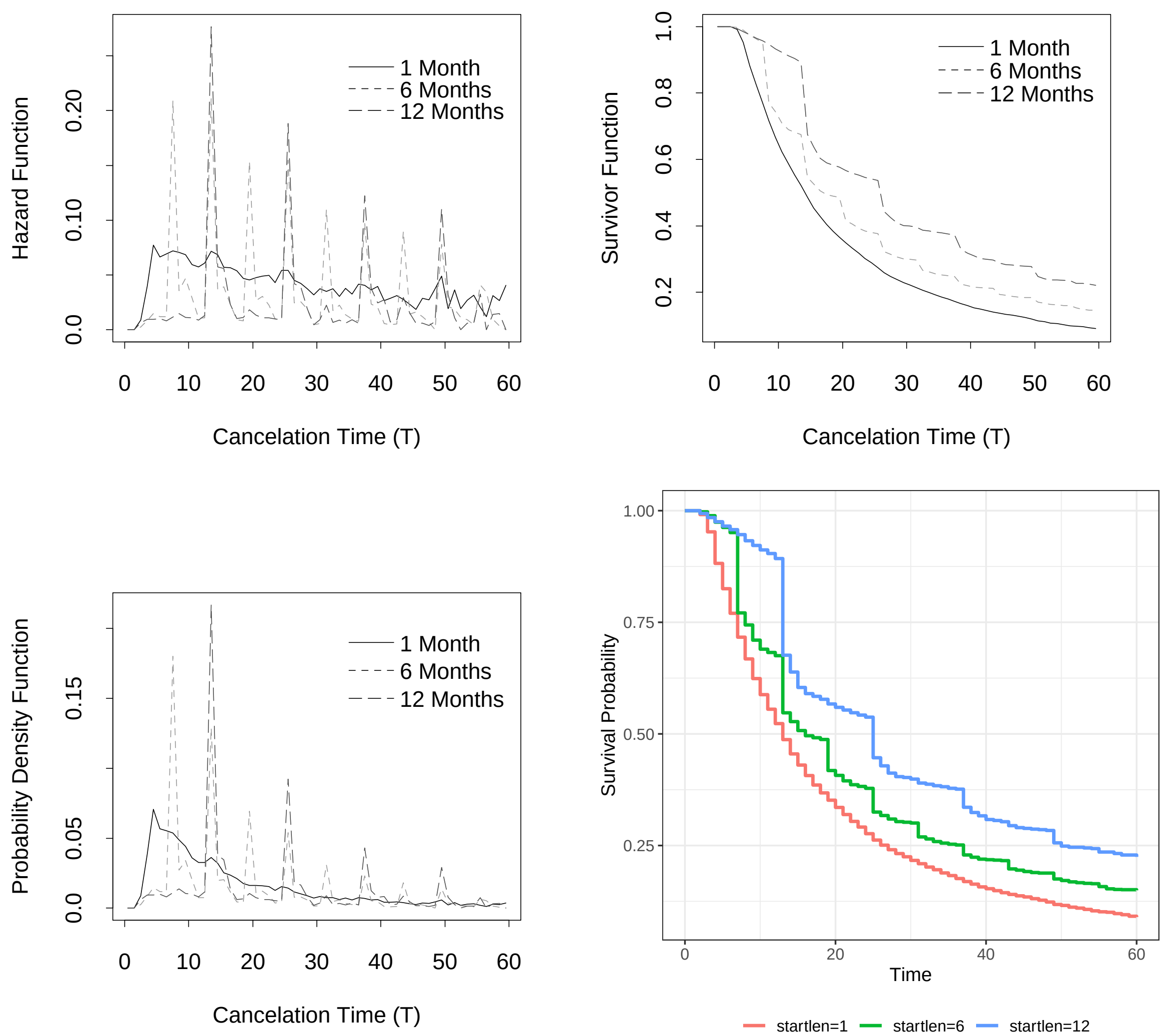




- Dashed line shows simple retention model (SRM).
- There is a distinct seasonal pattern, with large spikes every 12 months and smaller ones 6 months after the big ones.
- People sign 1-, 6- or 12-month contracts.
- The SRM fit of the hazard and PDF is poor, but the survival function looks OK.

Service Provider, Five Years, Stratifying on Starting Contract Length

```
fit3 = survfit(Surv(bigT, cancel) ~ startlen, data=service5yr, weight=count)
fit3 %>% ggsurvfit(size = 1) + add_confidence_interval()
```



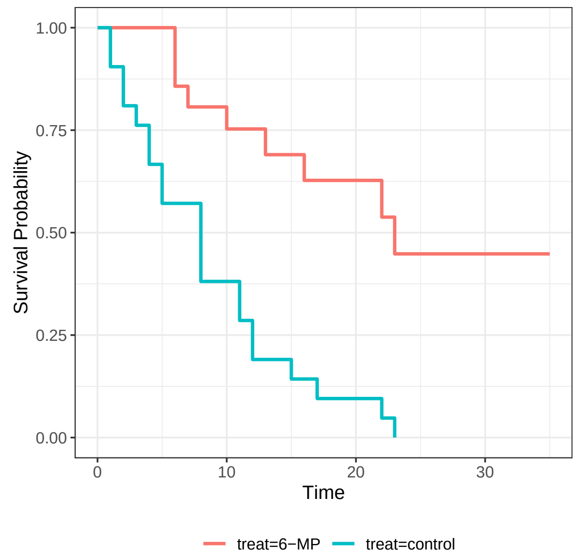
Gehan Leukemia Example

21 leukemia patients treated with new drug (6-mercaptopurine) and 21 matched controls (See Venables and Ripley, ch 13)

```
> library(survival)
> library(MASS)
> head(gehan) # cens poorly labeled, 1 if event happened and 0 for censored
  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
> fit = survfit(formula = Surv(time, cens) ~ treat, data = gehan)
> with(gehan, Surv(time, cens))
 [1]  1  10  22   7   3  32+ 12  23   8  22  17   6   2  16  11  34+  8  32+ 12
[20] 25+  2  11+  5  20+  4  19+ 15   6   8  17+ 23  35+  5   6  11  13   4   9+
[39]  1   6+  8  10+

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

Plotting and Inference on KM

```
> fit %>%
    ggsurvfit(size = 1) +
    add_confidence_interval()
```



- $\mathbb{V}(\hat{S}(t))$ can be estimated by **Greenwood's formula**:

$$\mathbb{V}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n - d_i)}$$

- A 95% CI for $S(t)$ is $\hat{S}(t) \pm 1.96\sqrt{\mathbb{V}[S(t)]}$
- **Log-rank test** evaluates $H_0 : S_1(t) = S_2(t), \forall t$.

```
> survdiff(Surv(time, cens)~treat, gehan)  # log-rank test of differences
Call: survdiff(formula = 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= 4.17e-05
```

The Proportional Hazard Model

$$h_i(t) = \lambda_0(t)\exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

- $h_i(t)$ is the hazard for individual i at time t
- $\lambda_0(t)$ is the unspecified baseline hazard function
- x_{ij} is the value of covariate j for individual i
- β_j is the effect of covariate j on the hazard function
- If we take the log of both sides, we get something that looks more like a linear regression model. Note that the (log) baseline hazard function determines the intercept.

$$\log[h_i(t)] = \log[\lambda_0(t)] + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

- The ratios of the hazards of two individuals i and i' is some constant (independent of time). This is called the *proportional hazards property.*

$$\frac{h_i(t)}{h_{i'}(t)} = \exp[\beta_1(x_{i1} - x_{i'1}) + \cdots + \beta_p(x_{ip} - x_{i'p})]$$

- With one predictor having two levels, the **hazard ratio** is

$$\frac{h(t|x=1)}{h(t|x=0)} = \frac{\lambda_0(t)\exp(\beta_1)}{\lambda_0(t)} = \exp(\beta_1)$$

Gehan Data with Cox PH

```
> fit4 = coxph(Surv(time, cens) ~ treat, gehan, method="exact") # cox
> plot(survfit(fit4),xlab="Remission (weeks)", ylab="Survival", cex=1.5)
> summary(fit4)
Call:
coxph(formula = Surv(time, cens) ~ treat, data = gehan, method = "exact")

  n= 42, number of events= 30

               coef exp(coef) se(coef)     z Pr(>|z|)
treatcontrol 1.6282    5.0949   0.4331 3.759  0.00017 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

Rsquare= 0.321   (max possible= 0.98 )
Likelihood ratio test= 16.25  on 1 df,   p=5.544e-05
Wald test            = 14.13  on 1 df,   p=0.0001704
Score (logrank) test = 16.79  on 1 df,   p=4.169e-05
```

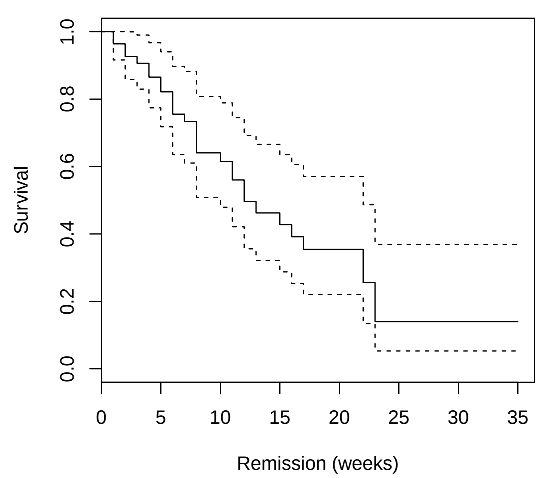
Let $x = 1$ for control, 0 for treatment

$$\log[h(t)] = \log[\lambda_0(t)] + 1.63x$$

$$h(t) = \lambda_0(t)\exp(1.63x)$$

Or, the hazard $h(t)$ is 5.095 times greater for those in the control group compared with the treatment group

KM estimate for average patient



Gehan Data with pairs as blocking variable

```
> fit5 = coxph(Surv(time, cens) ~ treat + factor(pair), gehan, method="exact") # cox
> summary(fit5)
Call:
coxph(formula = Surv(time, cens) ~ treat + factor(pair), data = gehan,
    method = "exact")

  n= 42, number of events= 30

                    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)3  -3.598195  0.027373  1.547371 -2.325 0.020053 *
....

               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)3   0.027373   36.53222 1.319e-03   0.56813
...
> anova(fit4, fit5)
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 .

> drop1(fit5, test="Chisq")
Single term deletions

Surv(time, cens) ~ treat + factor(pair)
             Df    AIC    LRT  Pr(>Chi)
<none>          161.83
treat         1 190.16 30.328 3.648e-08 ***
factor(pair) 20 151.09 29.256   0.08283 .
```