

A Mini Workshop on Survival Analysis Using S^{*†}

Dr. Jong Sung Kim[‡]

April 3, 2004

This work shop consists of excerpted portions of the book *Survival Analysis Using S: Analysis of Time-to-Event Data* by Mara Tableman[§] and Jong Sung Kim, published by Chapman & Hall/CRC, Boca Raton, 2004.

*Co-sponsored by Math & Stat, and Biomedical Science Ph.D. program, Wright State University, Dayton, Ohio

†Copyright(c) 2004 by Mara Tableman and Jong Sung Kim, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without written consent from the authors and the publisher.

‡Dr. Kim is an Assistant Professor of Statistics in the Department of Mathematics & Statistics, Portland State University, Portland, OR. He has an extensive experience in the various methods for analysis of censored data, and has taught Survival Analysis course several times.

§Dr. Tableman is an Associate Professor of Statistics in the Department of Mathematics & Statistics, Portland State University, Lecturer in the Seminar für Statistik at the Swiss Federal Institute of Technology, and Adjunct Associate Professor at the Oregon Health & Science University.

Software, Resources, and Topics

- Software: S-PLUS or R
- Resources: <http://www.crcpress.com/e-products/downloads/default.asp>
- Topics: Downloading R, data sets, and functions, Starting S-PLUS/R, Data entry and import/export of data files, Rationale of survival analysis, Basic identities associated with the analysis of censored data and major goals, Kaplan-Meier estimator of survival function, Weibull regression, AIC procedure for variable selection for the Cox Proportional Hazards Model, Connection to extended Cox model, competing risks model, and censored regression quantile method, provided time permits.

Downloading R, Data Sets, Functions, and Starting S-PLUS/R

- R is free to the public for download at www.r-project.org

Type `http://cran.r-project.org`

Click Windows (95 and later)

Click base

Click `rw1081.exe` and save it in your directory.

Copy `rw1081.exe` on the desktop and double click it to automatically set it up under the window.

- Data sets, functions, and updates:

Type `http://www.crcpress.com/e-products/downloads/default.asp`

Under Search This Page, select By Title

Type in Survival Analysis using S: Analysis of Time-to-Event Data.

- Double-click S-PLUS6.2 or R1.8.1

Data Entry and Import/Export of Data Files

- To save in a data frame (spreadsheet in S-PLUS): `File` → `New` → `Data Set` → `Ok`. A new (empty) `data.frame` will appear. This likens an EXCEL spreadsheet. Double click on the cell just below the column number to enter the variable name. Save the data frame as “example”. It will have the default extension “.sdd”.

	1	2	3
	weeks	group	status
1	5	0	1
2	8	0	1
3	9	1	1
4	13	1	1
5	13	1	0
6	18	1	1
7	161	1	0

- To save as an Excel file: `File` → `ExportData` → `ToFile`.
To File Name: `A:\example` Files of Type: `Excel Worksheet(xl?)`.
- To import your Excel file into S-PLUS: In S-PLUS, click on `File` → `ImportData` → `From:`
`A:\example` File Format: `Excel Worksheet(xl?)`.
- To import your text file into S or R: Now remove `example.sdd` from the S-PLUS by `> rm(example)`.
Open `example.xls` and save it as `example.txt`. Then use the `read.table` function as follows:
`> example <- read.table("A://example.txt",header=T,sep=" ")`
- If the delimiter is “~”, use `sep = "~"`. If a comma is the delimiter, use `sep = ","`.

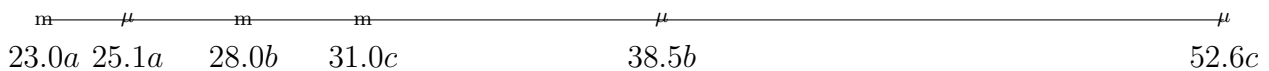
Rationale of Survival Analysis

- Time-to-Event Data: data that have as a principal endpoint the time when an event occurs. Some examples are time until an electrical component fails, time to first recurrence of a tumor (i.e., length of remission) after initial treatment, time to death, time to the learning of a skill, and promotion times for employees.
- Censoring: A “*failure*” time is not completely observed.
- Survival Analysis: The collection of statistical procedures which accommodate time-to-event censored data.
- Example: Results from a clinical trial to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukemia (AML). The study was conducted by Embury *et al.* (1977) at Stanford University. After reaching a status of remission through treatment by chemotherapy, the patients who entered the study were assigned randomly to two groups. The first group received maintenance chemotherapy; the second, or control, group did not. The objective of the trial was to see if maintenance chemotherapy prolonged the time until relapse.

Data for the AML maintenance study. A + indicates a censored value

Group	Length of complete remission (in weeks)
Maintained	9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+
Nonmaintained	5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

- Serious bias in estimated quantities, which lowers the efficacy of the study.
 - a. Throwing out the censored observations
 - b. Treating the censored observations as exact
 - c. Accounting for the censoring



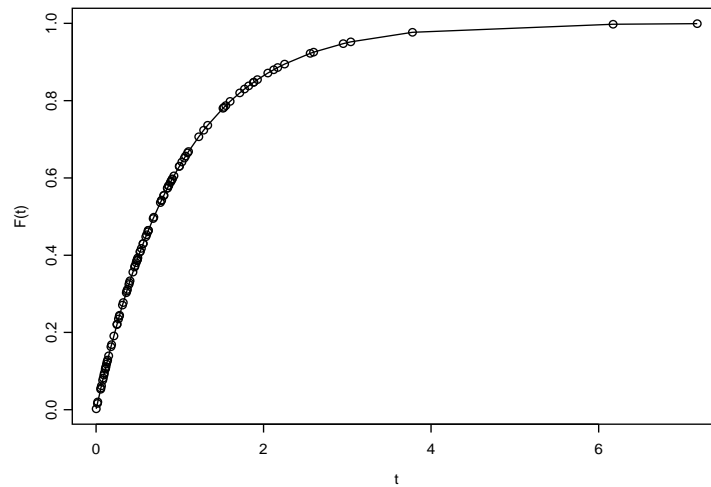
Basic Identities Associated with the Analysis of Censored Data

$$f(\cdot)$$

$$F(t) = P(T \leq t) = \int_0^t f(x)dx. \quad (1)$$

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx. \quad (2)$$

$$f(t) = \lim_{\Delta t \rightarrow 0^+} \frac{P(t < T \leq t + \Delta t)}{\Delta t} = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}. \quad (3)$$



$$F(t_p) = P(T \leq t_p) = p. \quad (4)$$

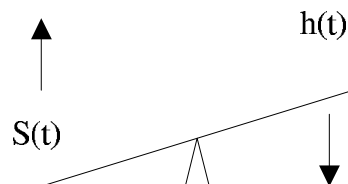
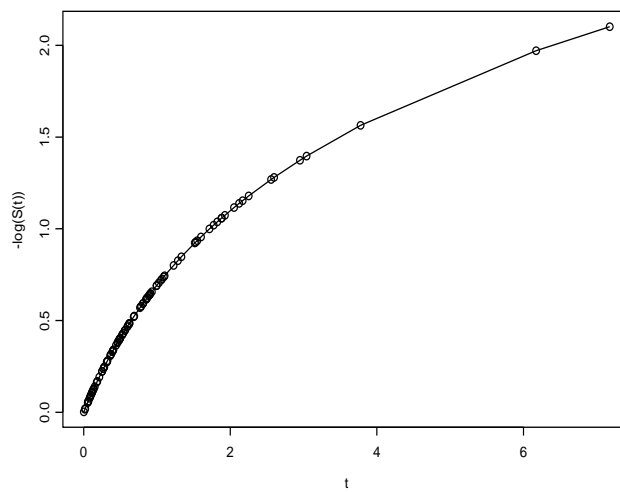
That is, $t_p = F^{-1}(p)$.

Hazard Function

$$h(t) = \lim_{\Delta t \rightarrow 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \quad (5)$$

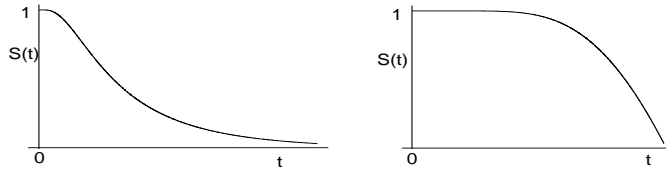
$$= -\frac{dS(t)/dt}{S(t)} = -\frac{d \log(S(t))}{dt}. \quad (6)$$

- specifies the instantaneous rate of failure at $T = t$ given that the individual survived up to time t .
- is the slope of the tangent line at $T = t$ of $-\log(S(t))$.
- specifies the distribution of T since it is a one-to-one function of $S(t)$.

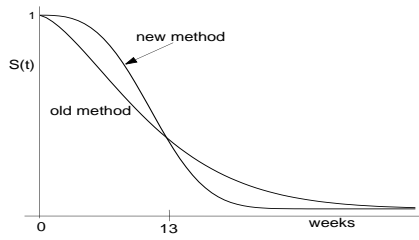


Major Goals

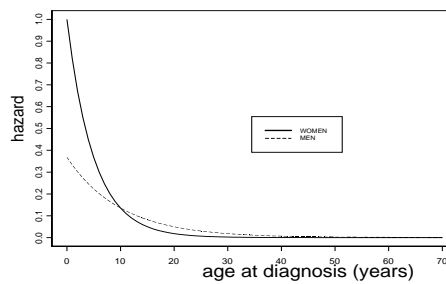
Goal 1. To estimate and interpret survivor and/or hazard functions from survival data.



Goal 2. To compare survivor and/or hazard functions.



Goal 3. To assess the relationship of explanatory variables to survival time, especially through the use of formal mathematical modelling.



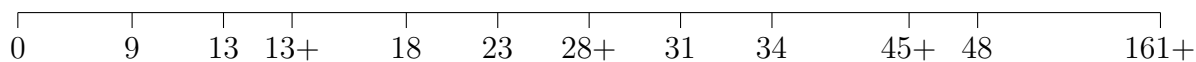
K-M Estimator of Survivor Function

Group	Data for the AML maintenance study Length of complete remission(in weeks)
Maintained	9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+
Nonmaintained	5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

The K-M estimator of the survivor function is

$$\hat{S}(t) = \prod_{y_{(i)} \leq t} \hat{p}_i = \prod_{y_{(i)} \leq t} \left(\frac{n_i - d_i}{n_i} \right) = \prod_{i=1}^k \left(\frac{n_i - d_i}{n_i} \right), \quad (7)$$

where $y_{(k)} \leq t < y_{(k+1)}$.



$$\begin{aligned}
 \hat{S}(0) &= 1 \\
 \hat{S}(9) &= \hat{S}(0) \times \frac{11-1}{11} = .91 \\
 \hat{S}(13) &= \hat{S}(9) \times \frac{10-1}{10} = .82 \\
 \hat{S}(13+) &= \hat{S}(13) \times \frac{9-0}{9} = \hat{S}(13) = .82 \\
 \hat{S}(18) &= \hat{S}(13) \times \frac{8-1}{8} = .72 \\
 \hat{S}(23) &= \hat{S}(18) \times \frac{7-1}{7} = .61 \\
 \hat{S}(28+) &= \hat{S}(23) \times \frac{6-0}{6} = \hat{S}(23) = .61 \\
 \hat{S}(31) &= \hat{S}(23) \times \frac{5-1}{5} = .49 \\
 \hat{S}(34) &= \hat{S}(31) \times \frac{4-1}{4} = .37 \\
 \hat{S}(45+) &= \hat{S}(34) \times \frac{3-0}{3} = \hat{S}(34) = .37 \\
 \hat{S}(48) &= \hat{S}(34) \times \frac{2-1}{2} = .18 \\
 \hat{S}(161+) &= \hat{S}(48) \times \frac{1-0}{1} = \hat{S}(48) = .18
 \end{aligned}$$

Estimates of Variance of $\hat{S}(t)$

Greenwood's formula (1926):

$$\widehat{\text{var}}(\hat{S}(t)) = \hat{S}^2(t) \sum_{y_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)} = \hat{S}^2(t) \sum_{i=1}^k \frac{d_i}{n_i(n_i - d_i)}, \quad (8)$$

where $y_{(k)} \leq t < y_{(k+1)}$.

Example with the AML1 data:

$$\begin{aligned} \widehat{\text{var}}(\hat{S}(13)) &= (.82)^2 \left(\frac{1}{11(11-1)} + \frac{1}{10(10-1)} \right) = .0136 \\ \text{s.e.}(\hat{S}(13)) &= .1166 \end{aligned}$$

The theory tells us that for each fixed value t

$$\hat{S}(t) \stackrel{a}{\sim} \text{normal}(S(t), \widehat{\text{var}}(\hat{S}(t))).$$

Thus, at time t , an approximate $(1 - \alpha) \times 100\%$ confidence interval for the probability of survival, $S(t) = P(T > t)$, is given by

$$\hat{S}(t) \pm z_{\frac{\alpha}{2}} \times \text{s.e.}(\hat{S}(t)), \quad (9)$$

where $\text{s.e.}(\hat{S}(t))$ is the square root of Greenwood's formula for the estimated variance. S function: `conf.type="plain"` in the `survfit` function.

Note:

- The **default** intervals in `survfit` are called "log" and the formula is:

$$\exp\left(\log \hat{S}(t) \pm 1.96 \text{s.e.}(\widehat{H}(t))\right), \quad (10)$$

where $\widehat{H}(t)$ is the estimated cumulative hazard function

- Sometimes, both of these intervals give limits outside the interval $[0, 1]$. This is not so appealing as $S(t)$ is a probability! Kalbfleisch & Prentice (1980) suggest using the transformation $W = \log(-\log(\hat{S}(t)))$ to estimate the log cumulative hazard parameter $\log(-\log(S(t)))$, and to then transform back. These intervals will always have limits within the interval $[0, 1]$. S function: `conf.type="log-log"` in the `survfit`.

S-PLUS/R Application

```
> library(survival) # For R users only
> km.fit <- survfit(Surv(weeks,status)~group,data=aml)
> plot(km.fit,conf.int=F,xlab="time until relapse (in weeks)",
      ylab="proportion without relapse",
      lab=c(10,10,7),cex=2,lty=1:2)
> summary(km.fit) # Displays the survival probability
                  # for each group
```

group=0

time	n.risk	n.event	survival	std.err	lower	95% CI upper	95% CI
5	12	2	0.8333	0.1076	0.6470	1.000	
8	10	2	0.6667	0.1361	0.4468	0.995	
12	8	1	0.5833	0.1423	0.3616	0.941	
23	6	1	0.4861	0.1481	0.2675	0.883	
27	5	1	0.3889	0.1470	0.1854	0.816	
30	4	1	0.2917	0.1387	0.1148	0.741	
33	3	1	0.1944	0.1219	0.0569	0.664	
43	2	1	0.0972	0.0919	0.0153	0.620	
45	1	1	0.0000	NA	NA	NA	

group=1

time	n.risk	n.event	survival	std.err	lower	95% CI upper	95% CI
9	11	1	0.909	0.0867	0.7541	1.000	
13	10	1	0.818	0.1163	0.6192	1.000	
18	8	1	0.716	0.1397	0.4884	1.000	
23	7	1	0.614	0.1526	0.3769	0.999	
31	5	1	0.491	0.1642	0.2549	0.946	
34	4	1	0.368	0.1627	0.1549	0.875	
48	2	1	0.184	0.1535	0.0359	0.944	

Properties:

- The K-M curve is a right continuous step function which steps down only at an uncensored observation.
- When there are no censored data values K-M reduces to the **esf**.
- Note the K-M curve does not jump down to zero as the largest survival time (161^+) is censored. We cannot estimate $S(t)$ beyond $t = 48$. Some refer to $\hat{S}(t)$ as a defective survival function.

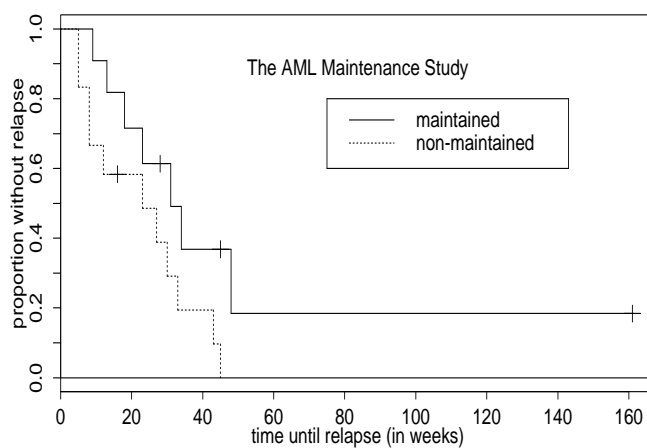


Figure 1: A comparison of two K-M curves.

Log-rank Test

```
> survdiff(Surv(week,status)~group,data=aml)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
group=1	11	7	10.69	1.27	3.4
group=2	12	11	7.31	1.86	3.4

Chisq= 3.4 on 1 degrees of freedom, p= 0.0653

There is mild evidence to suggest that maintenance chemotherapy prolongs the remission period since the one-sided test is appropriate and its p -value is $.0653/2 = .033$.

Remarks:

1. One might ask why the MH test is also called the log-rank test. Suppose there is only one covariate x so that the model is $Y = \beta_0^* + \beta^*x + \sigma Z$. Miller (1981), Chapter 6.2.1, derives the locally most powerful rank statistic for testing $H_0 : \beta^* = 0$ against $H_A : \beta^* \neq 0$ when evaluated at the extreme value distribution for error. He says that Peto and Peto (1972) first derived this test and named it the **log-rank test** as it involves a log-transformation. When x is binary, that is, $x = 1$ if in group 1 and $= 0$ if in group 2, this statistic simplifies to a quantity that is precisely a rescaled version of the MH statistic when there are no ties.
2. The `survdif` function contains a “rho” parameter. The default value, $\text{rho} = 0$, gives the log-rank test. When $\text{rho} = 1$, this gives the Peto test. This test was suggested as an alternative to the log-rank test by Prentice and Marek (1979). The Peto test emphasizes the beginning of the survival curve in that earlier failures receive larger weights. The log-rank test emphasizes the tail of the survival curve in that it gives equal weight to each failure time. Thus, choose between the two according to the interests of the study. The choice of emphasizing earlier failure times may rest on clinical features of one’s study.

Hazard Ratio as a Measure of Effect

The hazard ratio is a descriptive measure of the treatment (group) effect on survival. Here we use the two types of empirical hazard functions, $\tilde{h}(t_i) = \frac{d_i}{n_i}$ and $\hat{h}(t) = \frac{d_i}{n_i(t_{i+1}-t_i)}$, to form ratios and then interpret them in the context of the AML study.

```

> attach(aml)
> Surv0 <- Surv(weeks[group==0],status[group==0])
> Surv1 <- Surv(weeks[group==1],status[group==1])
> data <- list(Surv0,Surv1)
> emphazplot(data,text="solid line is maintained group")

```

nonmaintained				maintained							
time	hitilde	hihat	time	hitilde	hihat	time	hitilde	hihat			
1	5	0.167	0.056	1	9	0.091	0.023				
2	8	0.200	0.050	2	13	0.100	0.020				
3	12	0.125	0.011	3	18	0.125	0.025				
4	23	0.167	0.042	4	23	0.143	0.018				
5	27	0.200	0.067	5	31	0.200	0.067				
6	30	0.250	0.083	6	34	0.250	0.018				
7	33	0.333	0.033	7	48	0.500	0.018				
8	43	0.500	0.250								
9	45	1.000	0.250								

```

> detach()

```

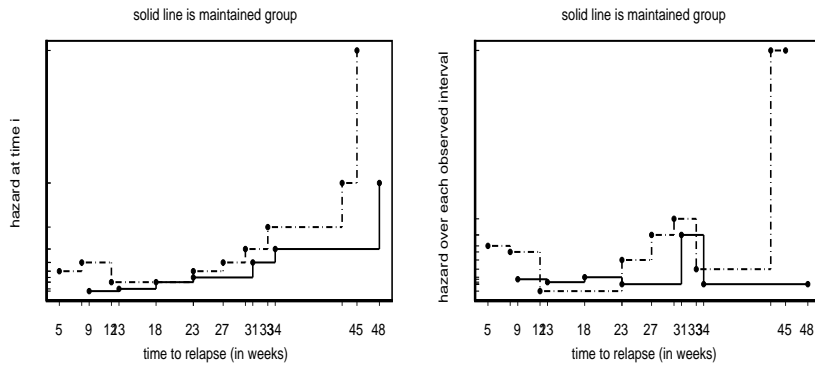


Figure 2: A comparison of empirical hazards. Left plot displays $\tilde{h}(t_i)$. Right plot displays $\hat{h}(t)$.

Weibull Distribution

p.d.f. $f(t)$	survivor $S(t)$	hazard $h(t)$
$\lambda\alpha(\lambda t)^{\alpha-1} \times \exp(-(\lambda t)^\alpha)$	$\exp(-(\lambda t)^\alpha)$	$\lambda\alpha(\lambda t)^{\alpha-1}$
mean $E(T)$	variance $Var(T)$	p th-quantile t_p
$\lambda^{-1}\Gamma(1 + \frac{1}{\alpha})$	$\lambda^{-2}\Gamma(1 + \frac{2}{\alpha}) - \lambda^{-2}(\Gamma(1 + \frac{1}{\alpha}))^2$	$\lambda^{-1}(-\log(1-p))^\frac{1}{\alpha}$ $\lambda > 0$ and $\alpha > 0$

The $\Gamma(k)$ denotes the gamma function and is defined as $\int_0^\infty u^{k-1}e^{-u}du, k > 0$.



Figure 3: Weibull density and hazard functions with $\lambda = 1$.

$$\log(t) = -\log(\lambda) + \sigma \log(-\log(S(t))), \quad (11)$$

where $\sigma = 1/\alpha$.

- Monotone increasing hazard when $\alpha > 1$, decreasing when $\alpha < 1$, and constant for $\alpha = 1$. The parameter α is called the shape parameter as the shape of the p.d.f., and hence the other functions, depends on the value of α .
- The λ is a scale parameter in that the effect of different values of λ is just to change the scale on the horizontal (t) axis, not the basic shape of the graph.

- An increasing Weibull hazard to be useful for modelling survival times of leukemia patients not responding to treatment, where the event of interest is death. As survival time increases for such a patient, and as the prognosis accordingly worsens, the patient's potential for dying of the disease also increases.
- A decreasing Weibull hazard to well model the death times of patients recovering from surgery. The potential for dying after surgery usually decreases as the time after surgery increases, at least for a while.
- Plot of $\log(t)$ versus $\log(-\log(S(t)))$ is a straight line with slope $\sigma = 1/\alpha$ and y -intercept $-\log(\lambda)$.

Extreme (Minimum) Value Distribution

The interest in this distribution is not for its direct use as a lifetime distribution, but rather because of its relationship to the Weibull distribution. Let μ , where $-\infty < \mu < \infty$, and $\sigma > 0$ denote location and scale parameters, respectively. The standard extreme value distribution has $\mu = 0$ and $\sigma = 1$.

p.d.f. $f(y)$	survivor $S(y)$	
$\sigma^{-1} \exp\left(\frac{y-\mu}{\sigma} - \exp\left(\frac{y-\mu}{\sigma}\right)\right)$	$\exp\left(-\exp\left(\frac{y-\mu}{\sigma}\right)\right)$	
mean $E(Y)$	variance $Var(Y)$	p th - quantile y_p
$\mu - \gamma\sigma$	$\frac{\pi^2}{6}\sigma^2$	$y_p = \mu + \sigma \log(-\log(1-p))$

Construction of the Quantile-Quantile (Q-Q) Plot

t_p quantile	$y_p = \log(t_p)$ quantile	form of standard quantile z_p
Weibull	extreme value	$\log(-\log(S(t_p))) = \log(H(t_p)) = \log(-\log(1-p))$

Fitting Data to the Weibull

```
# Weibull fit

> weib.fit <- survReg(Surv(weeks,status)~1,dist="weib")
> summary(weib.fit)
              Value Std. Error      z      p
(Intercept)  4.0997      0.366  11.187 4.74e-029
Log(scale)  -0.0314      0.277  -0.113 9.10e-001
Scale= 0.969

# Estimated median along with a 95% C.I. (in weeks).

> medhat <- predict(weib.fit,type="uquantile",p=0.5,se.fit=T)
> medhat1 <- medhat$fit[1]
> medhat1.se <- medhat$se.fit[1]
> exp(medhat1)
[1] 42.28842
> C.I.median1 <- c(exp(medhat1),exp(medhat1-1.96*medhat1.se),
                  exp(medhat1+1.96*medhat1.se))
> names(C.I.median1) <- c("median1","LCL","UCL")
> C.I.median1
  median1      LCL      UCL
42.28842  20.22064  88.43986
```

Weibull fit to AML1 data

model	$\hat{\mu}$	median1	95% C.I.	$\hat{\sigma}$
Weibull	4.1	42.29	[20.22, 88.44] weeks	.969

Q: What would happen if we fitted only the uncensored observations (z_i, y_i) to a least squares line alone?

```
> qq.weibull(Surv(weeks,status)) # Produces a Q-Q plot  
[1] "qq.weibull:done"
```

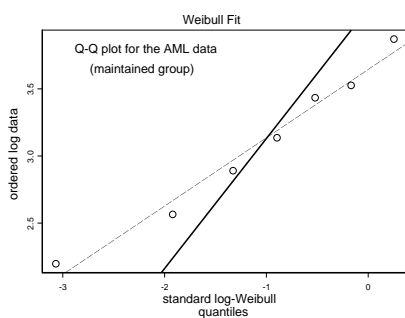


Figure 4: Q-Q plot for the Weibull. Solid line is constructed with MLE's $\hat{\mu}$ and $\hat{\sigma}$. The dashed line is the least squares line.

Weibull Regression Model

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

To include the covariate vector \underline{x} we now write the hazard for a given \underline{x} as

$$\begin{aligned} h(t|\underline{x}) &= h_0(t) \cdot \exp(\underline{x}'\underline{\beta}) \\ &= \alpha \lambda^\alpha t^{\alpha-1} \exp(\underline{x}'\underline{\beta}) = \alpha \left(\lambda \cdot (\exp(\underline{x}'\underline{\beta}))^{\frac{1}{\alpha}} \right)^\alpha t^{\alpha-1} \\ &= \alpha (\tilde{\lambda})^\alpha t^{\alpha-1}, \end{aligned} \tag{12}$$

where $\tilde{\lambda} = \lambda \cdot (\exp(\underline{x}'\underline{\beta}))^{\frac{1}{\alpha}}$.

From (11) and (12), given \underline{x} , $Y = \log(T) = \tilde{\mu} + \sigma Z$, where

$$\tilde{\mu} = -\log(\tilde{\lambda}) = -\log(\lambda \cdot (\exp(\underline{x}'\underline{\beta}))^{\frac{1}{\alpha}}) = -\log(\lambda) - \frac{1}{\alpha} \underline{x}'\underline{\beta}, \tag{13}$$

$\sigma = \frac{1}{\alpha}$, and $Z \sim$ standard extreme value distribution. Therefore,

$$Y = \underbrace{\beta_0^* + \underline{x}'\underline{\beta}^*}_{\tilde{\mu}} + \sigma Z, \tag{14}$$

where $\beta_0^* = -\log(\lambda)$ and $\underline{\beta}^* = -\sigma \underline{\beta}$.

$$\begin{aligned} S(t|\underline{x}) &= \exp\left(-(\tilde{\lambda}t)^\alpha\right) \\ H(t|\underline{x}) &= -\log(S(t|\underline{x})) = (\tilde{\lambda}t)^\alpha \\ \log(H(t|\underline{x})) &= \alpha \log(\tilde{\lambda}) + \alpha \log(t) \\ &= \alpha \log(\lambda) + \alpha \log(t) + \underline{x}'\underline{\beta} \\ &= \log(H_0(t)) + \underline{x}'\underline{\beta}, \end{aligned} \tag{15}$$

where $H_0(t) = -\log(S_0(t)) = (\lambda t)^\alpha$. The log of the cumulative hazard function is linear in $\log(t)$ and in the β coefficients. Thus, for a fixed \underline{x} value, the plot of $H(t|\underline{x})$ against t on a log-log scale is a straight line with slope α and intercept $\underline{x}'\underline{\beta} + \alpha \log(\lambda)$.

AIC Procedure for Variable Selection for the Cox PH Model

• **Proportional hazards model:** At a given covariate value \underline{x} , $h(t|\underline{x}) = h_0(t) \exp(\underline{x}'\beta)$, where $h_0(t)$ is the baseline hazard.

$$\text{AIC} = -2 \times \log(\text{maximum partial likelihood}) + 2 \times b, \quad (16)$$

where b is the number of β coefficients in each model under consideration. The smaller the AIC value the better the model is.

The `stepAIC` function requires an object representing a model of an appropriate class. This is used as the initial model in the stepwise search. Useful optional arguments include `scope` and `direction`. The scope defines the range of models examined in the stepwise search. The direction can be one of “both,” “backward,” or “forward,” with a default of “both.” If the direction argument is missing, the default for direction is “backward.” We consider an example fitting CNS data to Cox PH model.

Example:

1. PT.NUMBER: patient number
2. GROUP: 0=no prior radiation with respect to 1st blood brain barrier disruption (BBBD) procedure to deliver chemotherapy ; 1=prior radiation
3. SEX: 0=male ; 1=female
4. AGE: at time of 1st BBBD, recorded in years; instead, AGE60 = 1 if AGE \leq 60 and = 0 otherwise
5. STATUS: 0=alive ; 1=dead
6. B3TODeath: time from 1st BBBD to death in years
7. KPS.PRE.: Karnofsky performance score before 1st BBBD, numerical value 0 – 100
8. LESSING: Lesions: single=0 ; multiple=1
9. LESDEEP: Lesions: superficial=0 ; deep=1
10. LESSUP: Lesions: supra=0 ; infra=1 ; both=2
11. PROC: Procedure: subtotal resection=1 ; biopsy=2 ; other=3
12. CHEMOPRIOR: no=0 ; yes=1
13. RESPONSE: Tumor response to chemotherapy - complete=1; partial=2; blanks represent missing data

Step I: stepAIC to select the best model according to AIC statistic

```
> library(MASS) # Call in a collection of library functions
# provided by Venables and Ripley
> attach(cns2)
> cns2.coxint<-coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP+SEX+
AGE60+LESSING+LESDEEP+factor(LESSUP)+factor(PROC)+CHEMOPRIOR)
# Initial model
> cns2.coxint1 <- stepAIC(cns2.coxint,~.^2) # Up to two-way interaction

> cns2.coxint1$anova # Shows stepwise model path with the
# initial and final models
```

	Step	Df	AIC
			246.0864
+	SEX:AGE60	1	239.3337
-	factor(PROC)	2	236.7472
-	LESDEEP	1	234.7764
-	factor(LESSUP)	2	233.1464
+	AGE60:LESSING	1	232.8460
+	GROUP:AGE60	1	232.6511

Step II: LRT to further reduce

The following output shows p -values corresponding to variables selected by stepAIC. AGE60 has a large p -value, .560, while its interaction terms with SEX and LESSING have small p -values, .0019 and .0590, respectively.

```
> cns2.coxint1 # Check which variable has a moderately large p-value
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.0471	0.9540	0.014	-3.362	0.00077
GROUP	2.0139	7.4924	0.707	2.850	0.00440
SEX	-3.3088	0.0366	0.886	-3.735	0.00019
AGE60	-0.4037	0.6679	0.686	-0.588	0.56000
LESSING	1.6470	5.1916	0.670	2.456	0.01400
CHEMOPRIOR	1.0101	2.7460	0.539	1.876	0.06100
SEX:AGE60	2.8667	17.5789	0.921	3.113	0.00190
AGE60:LESSING	-1.5860	0.2048	0.838	-1.891	0.05900
GROUP:AGE60	-1.2575	0.2844	0.838	-1.500	0.13000

In statistical modelling, an important principle is that an interaction term should only be included in a model when the corresponding main effects are also present. (Hierarchical model!) We now see if we can eliminate the variable AGE60 and its interaction terms with other variables. We use the LRT. Here the LRT is constructed on the partial likelihood function rather than the full likelihood function. Nonetheless the large sample distribution theory holds. The LRT test shows strong evidence against the reduced model and so we retain the model selected by `stepAIC`.

```
> cns2.coxint2 <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP
+SEX+LESSING+CHEMOPRIOR) # Without AGE60 and its interaction terms
> -2*cns2.coxint2$loglik[2] + 2*cns2.coxint1$loglik[2] = 13.42442
> 1 - pchisq(13.42442,4) = 0.00938 # Retain the model selected by stepAIC
```

Step III : one variable at a time reduction

Since the variable GROUP:AGE60 has a moderately large p -value, .130, we delete it.

```
> cns2.coxint3 # Check which variable has a moderately large p-value
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.0436	0.9573	0.0134	-3.25	0.0011
GROUP	1.1276	3.0884	0.4351	2.59	0.0096
SEX	-2.7520	0.0638	0.7613	-3.61	0.0003
AGE60	-0.9209	0.3982	0.5991	-1.54	0.1200
LESSING	1.3609	3.8998	0.6333	2.15	0.0320
CHEMOPRIOR	0.8670	2.3797	0.5260	1.65	0.0990
SEX:AGE60	2.4562	11.6607	0.8788	2.79	0.0052
AGE60:LESSING	-1.2310	0.2920	0.8059	-1.53	0.1300

As AGE60:LESSING has a moderately large p -value, .130, we remove it.

```
> cns2.coxint4 # Check which variable has a moderately large p-value
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.0371	0.9636	0.0124	-3.00	0.00270
GROUP	1.1524	3.1658	0.4331	2.66	0.00780
SEX	-2.5965	0.0745	0.7648	-3.40	0.00069
AGE60	-1.3799	0.2516	0.5129	-2.69	0.00710
LESSING	0.5709	1.7699	0.4037	1.41	0.16000
CHEMOPRIOR	0.8555	2.3526	0.5179	1.65	0.09900
SEX:AGE60	2.3480	10.4643	0.8765	2.68	0.00740

We eliminate the term LESSING as it has a moderately large p -value, .160.

```
> cns2.coxint5 # Check which variable has a moderately large p-value
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.0402	0.9606	0.0121	-3.31	0.00093
GROUP	0.9695	2.6366	0.4091	2.37	0.01800
SEX	-2.4742	0.0842	0.7676	-3.22	0.00130
AGE60	-1.1109	0.3293	0.4729	-2.35	0.01900
CHEMOPRIOR	0.7953	2.2152	0.5105	1.56	0.12000
SEX:AGE60	2.1844	8.8856	0.8713	2.51	0.01200

We eliminate the variable CHEMOPRIOR as it has a moderately large p -value, .120. Since all the p -values in the reduced model fit below are small enough at the .05 level, we finally stop here and retain these five variables: KPS.PRE., GROUP, SEX, AGE60, and SEX:AGE60.

```
> cns2.coxint6 # Check which variable has a moderately large p-value
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.0307	0.970	0.0102	-2.99	0.0028
GROUP	1.1592	3.187	0.3794	3.06	0.0022
SEX	-2.1113	0.121	0.7011	-3.01	0.0026
AGE60	-1.0538	0.349	0.4572	-2.30	0.0210
SEX:AGE60	2.1400	8.500	0.8540	2.51	0.0120

However, it is important to compare this model to the model chosen by `stepAIC` in Step I as we have not compared them. The p -value based on LRT is between .05 and .1 and so we select the reduced model with caution.

```
> -2*cns2.coxint6$loglik[2] + 2*cns2.coxint1$loglik[2] = 8.843838  
> 1 - pchisq(8.843838,4) = 0.06512354 # Selects the reduced model
```

The following output is based on the model with KPS.PRE., GROUP, SEX, AGE60, and SEX:AGE60. It shows that the three tests – LRT, Wald, and efficient score test – indicate there is an overall significant relationship between this set of covariates and survival time. That is, they are explaining a significant portion of the variation.

```
> summary(cns2.coxint6)
```

Likelihood ratio test=	27.6	on 5 df,	p=0.0000431
Wald test	= 24.6	on 5 df,	p=0.000164
Score (logrank) test =	28.5	on 5 df,	p=0.0000296

Remarks:

1. The model selection procedure may well depend on the purpose of the study. In some studies there may be a few variables of special interest. In this case, we can still use Step I and Step II. In Step I we select the best set of variables according to the smallest AIC statistic. If this set includes all the variables of special interest, then in Step II we have only to see if we can further reduce the model. Otherwise, add to the selected model the unselected variables of special interest and go through Step II.
2. It is important to include interaction terms in model selection procedures unless researchers have compelling reasons why they do not need them. As the following illustrates, we could end up with a quite different model when only main effects models are considered. This is what Dahlborg *et al.* (1996) had.

Step I: stepAIC to select the best model according to AIC statistic

```
> cns2.cox <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP+SEX
+AGE60+LESSING+LESDEEP+factor(LESSUP)+factor(PROC)
+CHEMOPRIOR) # Initial model with all variables
> cns2.cox1 <- stepAIC(cns2.cox,~.) # Backward elimination
# procedure from full model to intercept only
> cns2.cox1$anova # Shows stepwise model paths with the
# initial and final models
```

	Step	Df	AIC
			246.0864
-	factor(PROC)	2	242.2766
-	LESDEEP	1	240.2805
-	AGE60	1	238.7327
-	factor(LESSUP)	2	238.0755
-	LESSING	1	236.5548

Step III: one variable at a time reduction

The p -values corresponding to GROUP and CHEMOPRIOR are very close. This implies that their effects adjusted for the other variables are about the same.

```
> cns2.cox1 # Check which variable has a large p-value
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.0432	0.958	0.0117	-3.71	0.00021
GROUP	0.5564	1.744	0.3882	1.43	0.15000
SEX	-1.0721	0.342	0.4551	-2.36	0.01800
CHEMOPRIOR	0.7259	2.067	0.4772	1.52	0.13000

We first eliminate GROUP. Since all the p -values in the reduced model are small enough at .05 level, we finally stop here and retain these three variables: KPS.PRE., SEX, and CHEMOPRIOR.

```
> cns2.cox2 # Check which variable has a moderately large p-value
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.0491	0.952	0.011	-4.46	8.2e-006
SEX	-1.2002	0.301	0.446	-2.69	7.1e-003
CHEMOPRIOR	1.0092	2.743	0.440	2.30	2.2e-002

Now let us see what happens if we eliminate CHEMOPRIOR first instead of GROUP. Since all the p -values in the reduced model are either smaller or about the same as .05 level, we stop here and retain these three variables: KPS.PRE., GROUP, and SEX.

```
> cns2.cox3 # Check which variable has large p-value
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.0347	0.966	0.010	-3.45	0.00056
GROUP	0.7785	2.178	0.354	2.20	0.02800
SEX	-0.7968	0.451	0.410	-1.94	0.05200

```
> detach()
```

In summary, depending on the order of elimination, we retain either SEX, KPS.PRE., and CHEMOPRIOR, or KPS.PRE., GROUP, and SEX. These two models are rather different in that one includes CHEMOPRIOR where the other includes GROUP instead. More importantly, note that none of these sets include the variable AGE60, which is a very important prognostic factor in this study evidenced by its significant interaction effect with SEX on the response (cns2.coxint6). In addition, the significance of the GROUP effect based on the interaction model is more pronounced (p -value 0.0022 versus 0.028), which was the primary interest of the study.

Discussion

- KPS.PRE., GROUP, SEX, AGE60, and SEX:AGE60 appear to have a significant effect on survival duration. Here it is confirmed again that there is a significant difference between the two groups' (0=no prior radiation,1=prior radiation) survival curves.
- The estimated coefficient for KPS.PRE. is -0.0307 with p -value 0.0028. Hence, fixing other covariates, patients with high KPS.PRE. scores have a decreased hazard, and, hence, have longer expected survival time than those with low KPS.PRE. scores.
- The estimated coefficient for GROUP is 1.1592 with p -value 0.0022. Hence, with other covariates fixed, patients with radiation prior to first BBBD have an increased hazard, and, hence, have shorter expected survival time than those in Group 0.
- Fixing other covariates, the hazard ratio between Group 1 and Group 0 is

$$\frac{\exp(1.1592)}{\exp(0)} = 3.187.$$

This means that, with other covariates fixed, patients with radiation prior to first BBBD are 3.187 times more likely than those without to have shorter survival.

- Fixing other covariates, if a patient in Group 1 has 10 units larger KPS.PRE. score than a patient in Group 0, the ratio of hazard functions is

$$\begin{aligned} \frac{\exp(1.1592) \exp(-0.0307 \times (k + 10))}{\exp(0) \exp(-0.0307 \times k)} &= \frac{\exp(1.1592) \exp(-0.0307 \times 10)}{\exp(0)} \\ &= 3.187 \times 0.7357 = 2.345, \end{aligned}$$

where k is an arbitrary number. This means that fixing other covariates, a patient in Group 1 with 10 units larger KPS.PRE. score than a patient in Group 0 is 2.34 times more likely to have shorter survival. In summary, fixing other covariates, whether a patient gets radiation therapy prior to first BBBD is more important than how large his/her KPS.PRE. score is.

- There is significant interaction between AGE60 and SEX. The estimated coefficient for SEX:AGE60 is 2.1400 with p -value 0.0120. Fixing other covariates, a male patient who is younger than 60 years old has 34.86% of the risk a male older than 60 years old has of succumbing to the disease, where

$$\frac{\exp(-2.113 \times 0 - 1.0538 \times 1 + 2.14 \times 0)}{\exp(-2.113 \times 0 - 1.0538 \times 0 + 2.14 \times 0)} = \exp(-1.0538) = .3486.$$

Whereas, fixing other covariates, a female patient who is younger than 60 years old has 2.963 times the risk a female older than 60 years old has of succumbing to the disease, where

$$\frac{\exp(-2.113 \times 1 - 1.0538 \times 1 + 2.14 \times 1)}{\exp(-2.113 \times 1 - 1.0538 \times 0 + 2.14 \times 0)} = \exp(1.0862) = 2.963.$$

In Figure 5, we plot the interaction between SEX and AGE60 based on the means computed using the S function `survfit` for the response and AGE60, fixing female and male separately. It shows a clear pattern of interaction, which supports the prior numeric results using Cox model `cns2.coxint6`.

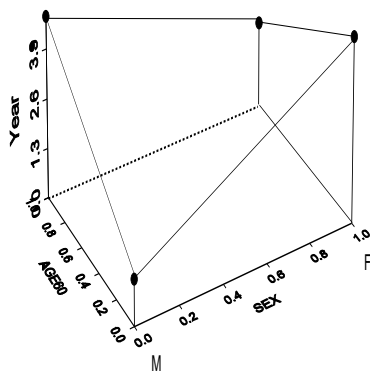


Figure 5: Interaction between SEX and AGE60.

In Figure 6 we first fit the data to the model

```
> cox.fit <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP+
                    strata(factor(SEX),factor(AGE60)))
```

which adjusts for the GROUP and KPS.PRE. effects. We then set GROUP = 1 and KPS.PRE. = 80 and obtain the summary of the adjusted quantiles and means in the same four strata (cells) as in Figure 5 using `survfit` as follows:

```
> survfit(cox.fit,data.frame(GROUP=1,KPS.PRE.=80))
> summary(survfit(cox.fit,data.frame(GROUP=1,KPS.PRE.=80)))
```

Figure 6 displays both ordinal and disordinal interactions. The survival curve for the females who are older than 60 years never steps below 0.50 (see `> summary` above). In order to produce the median plot, we set the median survival time since 1st BBBD for this stratum at 1.375 years, which is the .368-quantile.

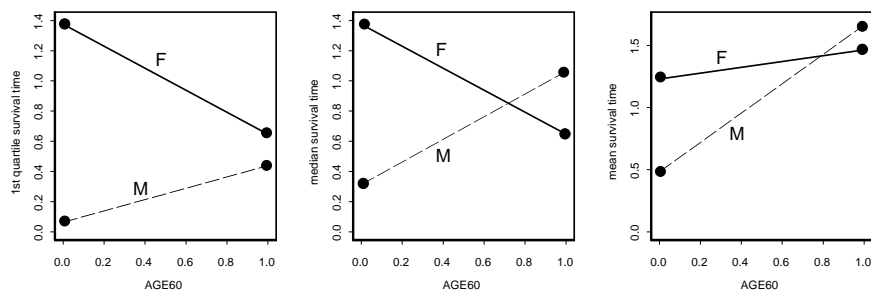


Figure 6: *Interaction between SEX and AGE60 adjusted for KPS.PRE. and GROUP via coxph and then evaluated at GROUP=1 and KPS.PRE.=80.*

Note: The “Censored Regression Quantiles” approach introduced by Portnoy (2002) enables one to study each of the estimated quantiles as a function of the targeted covariates. This non-parametric methodology is presented in Chapter 8 of my co-authored book, *Survival Analysis Using S: Analysis of Time-to- Event Data*, Chapman & Hall/CRC, 2003.

Connection to extended Cox model, competing risks model, and censored regression quantile method, provided time permits.

Example: Extracted from Kleinbaum (1996, pages 109 – 111).

A study in which cancer patients are randomized to either surgery or radiation therapy without surgery is considered. We have a $(0, 1)$ exposure variable E denoting surgery status, with 0 if a patient receives surgery and 1 if not (i.e., receives radiation). Suppose further that this exposure variable is the only variable of interest.

Is the Cox PH model appropriate?

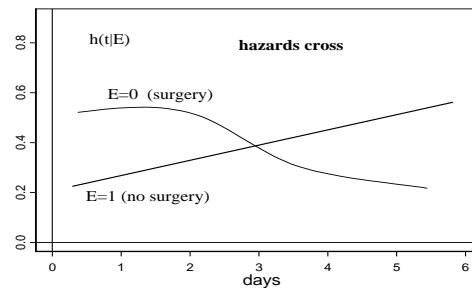


Figure 7: Hazards crossing over time.

If the Cox PH model is inappropriate,

- analyze by **stratifying** on the exposure variable; that is, do not fit any regression model, and, instead obtain the Kaplan-Meier curve for each group separately;
- start the analysis at three days, and use a Cox PH model on three-day survivors;
- fit a Cox PH model for less than three days and a different Cox PH model for greater than three days to get two different hazard ratio estimates, one for each of these two time periods;
- fit a Cox PH model that includes a time-dependent variable which measures the interaction of exposure with time. This model is called an **extended Cox model** and is presented in Chapter 7.1.
- use the **censored regression quantile** approach presented in Chapter 8 allowing crossover effects. This approach is nonparametric and is free of the PH assumption for its validity.