Survival Analysis Using S/R*

Unterlagen für den Weiterbildungs–Lehrgang in angewandter Statistik an der ETH Zürich

---

Professor Mara Tableman†.

Fariborz Maseeh Department of Mathematics & Statistics
Portland State University
Portland, Oregon, USA

tablemanm@pdx.edu

August–September 2012

These notes are an abridged and edited version of the first six chapters 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

---

* © 2004, Mara Tableman and Jong Sung Kim, all rights reserved. This text may be freely shared among individuals, but it may not be published in any medium without written consent from the authors.

† Dr. Tableman is Professor of Statistics in the Fariborz Maseeh Department of Mathematics & Statistics, Portland State University, Lecturer in the Seminar für Statistik at ETH Zürich, and Adjunct Professor at Oregon Health & Science University

‡ Dr. Kim is Professor of Statistics in the Fariborz Maseeh Department of Mathematics & Statistics, Portland State University, Portland, Oregon.

Contents

CHAPTER 1

Introduction

The primary purpose of a survival analysis is to **model and analyze time-to-event data**; that is, data that have as a principal endpoint the time when an event occurs. Such events are generally referred to as "*failures*." 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.

In these examples we can see that it is possible that a "*failure*" time will not be observed either by deliberate design or due to **random censoring**. This occurs, for example, if a patient is still alive at the end of a clinical trial period or has moved away. The necessity of obtaining methods of analysis that accommodate censoring is the primary reason for developing specialized models and procedures for failure time data. **Survival analysis is the modern name given to the collection of statistical procedures which accommodate time-to-event censored data.** Prior to these new procedures, incomplete data were treated as missing data and omitted from the analysis. This resulted in the loss of the partial information obtained and in introducing serious systematic error (bias) in estimated quantities. This, of course, lowers the efficacy of the study. The procedures discussed here avoid bias and are more powerful as they utilize the partial information available on a subject or item.

These course notes introduce the field of survival analysis without getting too embroiled in the theoretical technicalities. Models for failure times describe either the survivor function or hazard rate and their dependence on explanatory variables. Presented here are some frequently used parametric models and methods; and the newer, very fashionable, due to their flexibility and power, nonparametric procedures. The statistical tools treated are applicable to data from medical clinical trials, public health, epidemiology, engineering, economics, psychology, and demography as well. The S/R code is woven into the text, which provides a self-learning opportunity.

Objectives of this chapter:

After studying Chapter 1, the student should be able to:

1. Recognize and describe the type of problem addressed by a survival analysis.

2. Define, recognize, and interpret a **survivor function**.
3. Define, recognize, and interpret a **hazard function**.
4. Describe the relationship between a survivor function and hazard function.
5. Interpret or compare examples of survivor or hazard curves.
6. Define what is meant by **censored data**.
7. Define or recognize three **censoring models**.
8. Know the form of the likelihood function common to these three models.
9. Give three reasons why data may be randomly censored.
10. State the **three goals of a survival analysis**.

1.1 Motivation

Example 1. AML study

The data presented in Table 1.1 are preliminary 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.

Table 1.1: *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

A naive descriptive analysis of AML study:

We consider a couple of descriptive measures to compare the two groups of data given in Example 1. The first approach is to throw out censored observations, the second is to treat the censored observations as exact ones, and the last is to use them all as they are. We at least expect to see different results among the three approaches. Let's see just how different they are.

• Analysis of AML data after throwing out censored observations

Measures	*Maintained*	*Nonmaintained*
Mean	25.1	21.7
Median	23.0	23.0

The mean for maintained group is slightly larger than that for nonmaintained group while their medians are the same. That is, the distribution of maintained group is slightly more skewed to the right than the nonmaintained group's distribution is. The difference between the two groups appears to be negligible.

• Analysis of AML data treating censored observations as exact

Measures	*Maintained*	*Nonmaintained*
Mean	38.5	21.3
Median	28.0	19.5

Both the mean and median for maintained group are larger than those for nonmaintained group. The difference between the two groups seems to be non-negligible in terms of both mean and median. The skewness of the maintained group is even more pronounced. We expect, however, that these estimates are biased in that they underestimate the true mean and median. The censored times are smaller than the true unknown failure times. The next analysis is done using a method which accommodates the censored data.

• Analysis of AML data accounting for the censoring

Measures	*Maintained*	*Nonmaintained*
Mean	52.6	22.7
Median	31.0	23.0

Both the mean and median for maintained group are larger than those for nonmaintained group. Further, the mean of the maintained group is much larger than that of the nonmaintained group. Here we notice that the distribution of

maintained group is much more skewed to the right than the nonmaintained group's distribution is. Consequently, the difference between the two groups seems to be huge. From this small example, we have learned that appropriate methods should be applied in order to deal with censored data. The method used here to estimate the mean and median is discussed in Chapter 2.1.

1.2 Basic definitions

Let T denote a nonnegative random variable representing the lifetimes of individuals in some population. ("Nonnegative" means $T \geq 0$.) We treat the case where T is continuous. For a treatment of discrete models see Lawless (1982, page 10). Let $F(\cdot)$ denote the (cumulative) **distribution function** (d.f.) of T with corresponding **probability density function** (p.d.f.) $f(\cdot)$. Note $f(t) = 0$ for $t < 0$. Then

$$F(t) = P(T \leq t) = \int_0^t f(x)dx. \tag{1.1}$$

The probability that an individual survives to time t is given by the **survivor function**

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx. \tag{1.2}$$

This function is also referred to as the **reliability function**. Note that $S(t)$ is a monotone decreasing function with $S(0) = 1$ and $S(\infty) = \lim_{t\to\infty} S(t) = 0$. Conversely, we can express the p.d.f. as

$$f(t) = \lim_{\Delta t \to 0^+} \frac{P(t < T \leq t + \Delta t)}{\Delta t} = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}. \tag{1.3}$$

The ***p*th-quantile** of the distribution of T is the value t_p such that

$$F(t_p) = P(T \leq t_p) = p. \tag{1.4}$$

That is, $t_p = F^{-1}(p)$. The pth-quantile is also referred to as the **$100 \times p$th percentile** of the distribution. The **hazard function** specifies the instantaneous rate of failure at $T = t$ given that the individual survived up to time t and is defined as

$$h(t) = \lim_{\Delta t \to 0^+} \frac{P(t < T \leq t + \Delta t \,|\, T > t)}{\Delta t} = \frac{f(t)}{S(t)}. \tag{1.5}$$

We see here that $h(t)\Delta t$ is approximately the probability of a death in $(t, t + \Delta t]$, given survival up to time t. The hazard function is also referred to as the **risk** or **mortality rate**. We can view this as a measure of intensity at time t or a measure of the potential of failure at time t. The hazard is a rate, rather than a probability. It can assume values in $[0, \infty)$.

To understand why the hazard is a rate rather than a probability, in its definition consider the expression to the right of the limit sign which gives the

ratio of two quantities. The numerator is a conditional probability and the denominator is Δt, which denotes a small time interval. By this division, we obtain a probability per unit time, which is no longer a probability but a rate. This ratio ranges between 0 and ∞. It depends on whether time is measured in days, weeks, months, or years, etc. The resulting value will give a different number depending on the units of time used. To illustrate this let $P = P(t < T \leq t + \Delta t \,|\, T > t) = 1/4$ and see the following table:

P	Δt	$\frac{P}{\Delta t}$ = rate
$\frac{1}{4}$	$\frac{1}{3}$day	$\frac{1/4}{1/3} = 0.75$/day
$\frac{1}{4}$	$\frac{1}{21}$week	$\frac{1/4}{1/21} = 5.25$/week

It is easily verified that $h(t)$ specifies the distribution of T, since

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

Integrating $h(u)$ over $(0, t)$ gives the **cumulative hazard function** $H(t)$:

$$H(t) = \int_0^t h(u)du = -\log\Big(S(t)\Big). \tag{1.6}$$

In this book, unless otherwise specified, log denotes the natural logarithm, the inverse function of the exponential function exp $= e$. Thus,

$$S(t) = \exp\Big(-H(t)\Big) = \exp\left(-\int_0^t h(u)du\right). \tag{1.7}$$

Hence, the p.d.f. of T can be expressed as

$$f(t) = h(t)\exp\left(-\int_0^t h(u)du\right).$$

Note that $H(\infty) = \int_0^\infty h(t)dt = \infty$. Figures 1.1 & 1.2 display the relationships between $h(t), H(t)$ and $S(t)$.

For a nonnegative random variable T **the mean value**, written $E(T) = \int_0^\infty t \cdot f(t)dt$, can be shown to be

$$E(T) = \int_0^\infty S(t)dt. \tag{1.8}$$

WHY! Thus, mean survival time is the total area under the survivor curve $S(t)$. It follows from expression (1.7), for a given time t, the greater the risk, the smaller $S(t)$, and hence the shorter mean survival time $E(T)$, and vice versa. The following picture should help you to remember this relationship.

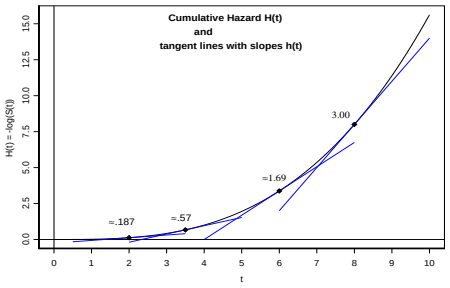


Figure 1.1 *Graph of a cumulative hazard $H(t)$ and several tangents $h(t)$.*

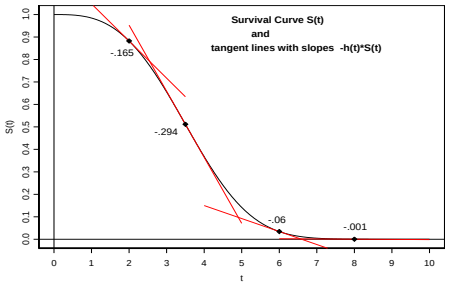
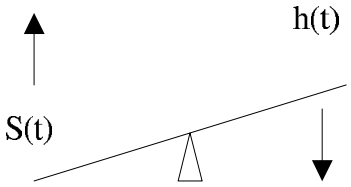


Figure 1.2 *Graph of a survivor curve $S(t)$ and several tangents $-h(t) \times S(t)$.*



Another basic parameter of interest is the **mean residual life** at time u, denoted by $\text{mrl}(u)$. For individuals of age u, this parameter measures their expected remaining lifetime. It is defined as

$$\text{mrl}(u) = E(T - u \mid T > u).$$

For a continuous random variable it can be verified that

$$\text{mrl}(u) = \frac{\int_u^\infty S(t)dt}{S(u)}. \tag{1.9}$$

WHY! The $\text{mrl}(u)$ is hence the area under the survival curve to the right of u

divided by $S(u)$. Lastly, note the mean life, $E(T) = \text{mrl}(0)$, is the total area under the survivor curve. The graph in Figure 1.3 illustrates this definition.

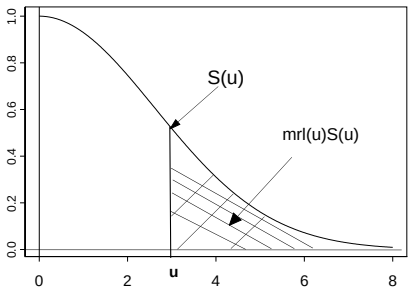


Figure 1.3 *Mean residual life at time u.*

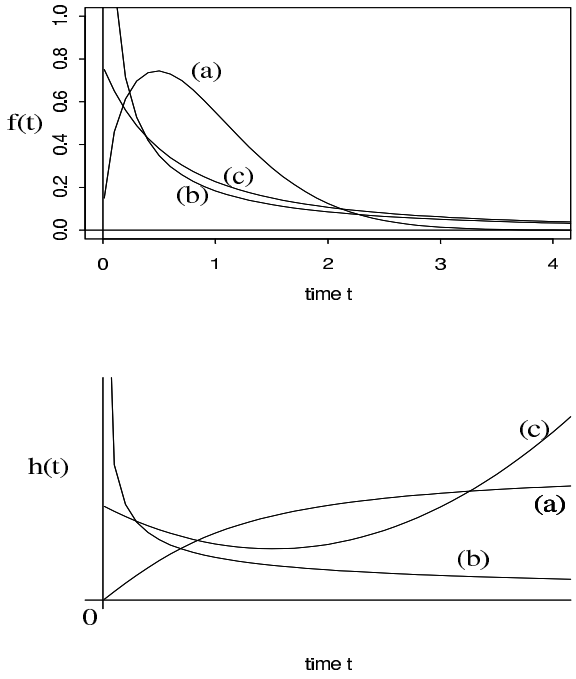


Figure 1.4 *Types of hazard rates and respective densities.*

To end this section we discuss hazard functions and p.d.f.'s for three continuous distributions displayed in Figure 1.4. Model (a) has an increasing hazard rate. This may arise when there is a natural aging or wear. Model (b) has

a decreasing hazard rate. Decreasing functions are less common but find occasional use when there is an elevated likelihood of early failure, such as in certain types of electronic devices or in patients experiencing certain types of organ transplants. Model (c) has a bathtub-shaped hazard. Most often these are appropriate for populations followed from birth. Similarly, some manufactured equipment may experience early failure due to defective parts, followed by a constant hazard rate which, in later stages of equipment life, increases. Most population mortality data follow this type of hazard function where, during an early period, deaths result, primarily from infant diseases, after which the death rate stabilizes, followed by an increasing hazard rate due to the natural aging process. Not represented in these plots is the hump-shaped hazard; i.e., the hazard is increasing early and then eventually begins declining. This type of hazard rate is often used to model survival after successful surgery where there is an initial increase in risk due to infection, hemorrhaging, or other complications just after the procedure, followed by a steady decline in risk as the patient recovers.

Remark:

Although different survivor functions can have the same basic shape, their hazard functions can differ dramatically, as is the case with the previous three models. The hazard function is usually more informative about the underlying mechanism of failure than the survivor function. For this reason, modelling the hazard function is an important method for summarizing survival data.

Hazard ratio:

For two treatment groups, say 0 and 1, their **hazard ratio** (HR) is

$$\text{HR}(t|1,0) = \frac{h(t|1)}{h(t|0)}.$$

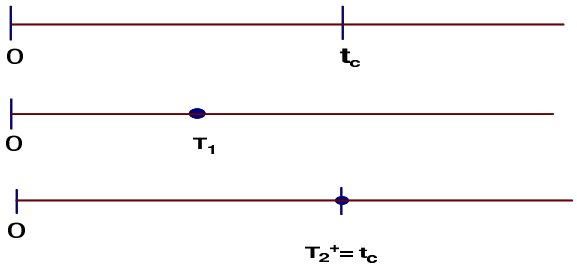
The HR is a numeric measure that describes the treatment effect over time. This descriptive measure plays a major role in a survival analysis. For example, if $\text{HR}(t^*|1,0) = .75$, this says treatment 1 cohort has three-fourths the risk of dying at time $= t^*$ than the cohort receiving treatment 0. Equivalently, the cohort receiving treatment 0 has 33% more risk of dying than the cohort receiving treatment 1.

1.3 Censoring models

We now present three types of censoring models. Let $T_1, T_2, \ldots, T_n$ be independent and identically distributed (iid) with distribution function (d.f.) F.

Type I censoring

This type arises in engineering applications. In such situations there are transistors, tubes, chips, etc.; we put them all on test at time $t = 0$ and record their times to failure. Some items may take a long time to "burn out" and we will not want to wait that long to terminate the experiment. Therefore, we terminate the experiment at a prespecified time t_c. The number of observed failure times is random. If n is the number of items put on test, then we could observe $0, 1, 2, \ldots, n$ failure times. The following illustrates a possible trial:



We call t_c the fixed censoring time. Instead of observing the T_i, we observe $Y_1, Y_2, \ldots, Y_n$ where

$$Y_i = \min(T_i, t_c) = \begin{cases} T_i & \text{if } T_i \le t_c \\ t_c & \text{if } t_c < T_i. \end{cases}$$

Notice that the d.f. of Y has positive mass $P(T > t_c) > 0$ at $y = t_c$ since the $P(Y = t_c) = P(t_c < T) = 1 - F(t_c) > 0$. That is, Y is a mixed random variable with a continuous and discrete component. The (cumulative) d.f. $M(y)$ of Y is shown in Figure 1.5. It is useful to introduce a binary random variable δ which indicates if a failure time is observed or censored,

$$\delta = \begin{cases} 1 & \text{if } T \le t_c \\ 0 & \text{if } t_c < T. \end{cases}$$

Note that $\{\delta = 0 \text{ and } T \le t_c\}$ implies that the failure time was precisely $T = t_c$, which occurs with zero probability if T is a continuous variable. (Note that for discrete distributions, we can set t_c equal to the last attainable time a failure may be observed. Hence, the probability $P(\{\delta = 0\} \cap \{T \le t_c\})$ is not equal to zero.) We then observe the iid random pairs (Y_i, δ_i).

For maximum likelihood estimation (detailed in Chapter 3.2) of any parameters of the distribution of T, we need to calculate the joint likelihood of the

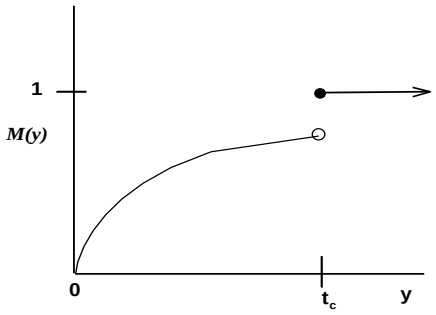


Figure 1.5 *Cumulative d.f. of the mixed random variable Y.*

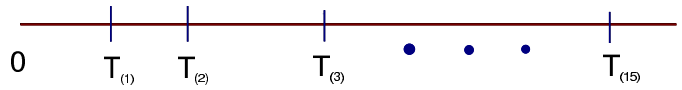
pair (Y, δ). By likelihood we mean the rubric which regards the density as a function of the parameter for a given (fixed) value (y, δ). For $y < t_c$, $P(Y \leq y) = P(T \leq y) = F(y)$ and $P(\delta = 1 \mid Y \leq y) = 1$. Therefore, the likelihood for $Y = y < t_c$ and $\delta = 1$ is the density $f(y)$. For $y = t_c$ and $\delta = 0$, the likelihood for this event is the probability $P(\delta = 0, Y = t_c) = P(T > t_c) = S(t_c)$.

We can combine these two expressions into one single expression $\big(f(y)\big)^{\delta} \times \big(S(t_c)\big)^{1-\delta}$. As usual, we define the likelihood function of a random sample to be the product of the densities of the individual observations. That is, the likelihood function for the n iid random pairs (Y_i, δ_i) is given by

$$L = \prod_{i=1}^{n} \big(f(y_i)\big)^{\delta_i} \big(S(t_c)\big)^{1-\delta_i}. \tag{1.10}$$

Type II censoring

In similar engineering applications as above, the censoring time may be left open at the beginning. Instead, the experiment is run until a prespecified fraction r/n of the n items has failed. Let $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$ denote the ordered values of the random sample $T_1, \ldots, T_n$. By plan, observations terminate after the rth failure occurs. So we only observe the r smallest observations in a random sample of n items. For example, let $n = 25$ and take $r = 15$. Hence, when we observe 15 burn out times, we terminate the experiment. Notice that we could wait an arbitrarily long time to observe the 15th failure time as $T_{(15)}$ is random. The following illustrates a possible trial:



In this trial the last 10 observations are assigned the value of $T_{(15)}$. Hence we have 10 censored observations. More formally, we observe the following full sample.

$$\begin{array}{lcl} Y_{(1)} & = & T_{(1)} \\ Y_{(2)} & = & T_{(2)} \\ \vdots & \vdots & \vdots \\ Y_{(r)} & = & T_{(r)} \\ Y_{(r+1)} & = & T_{(r)} \\ \vdots & \vdots & \vdots \\ Y_{(n)} & = & T_{(r)}. \end{array}$$

Formally, the data consist of the r smallest lifetimes $T_{(1)}, \ldots, T_{(r)}$ out of the n iid lifetimes $T_1, \ldots, T_n$ with continuous p.d.f $f(t)$ and survivor function $S(t)$. Then the likelihood function (joint p.d.f) of $T_{(1)}, \ldots, T_{(r)}$ is given

$$L = \frac{n!}{(n-r)!} f(t_{(1)}) \cdots f(t_{(r)}) \cdot \Big(S(t_{(r)})\Big)^{n-r}. \tag{1.11}$$

WHY!

Remarks:

1. In Type I censoring, the endpoint t_c is a fixed value and the number of observed failure times is a random variable which assumes a value in the set $\{0, 1, 2, \ldots, n\}$.
2. In Type II censoring, the number of failure times r is a fixed value whereas the endpoint T_r is a random observation. Hence we could wait possibly a very long time to observe the r failures or, vice versa, see all r relatively early on.
3. Although Type I and Type II censoring are very different designs, **the form of the observed likelihood function is the same in both cases**. To see this it is only necessary to note that the individual items whose lifetimes are observed contribute a term $f(y_{(i)})$ to the observed likelihood function, whereas items whose lifetimes are censored contribute a term $S(y_{(i)})$. The factor $n!/(n-r)!$ in the last equation reflects the fact that we consider the ordered observations. For maximum likelihood estimation the factor will be irrelevant since it does not depend on any parameters of the distribution function.

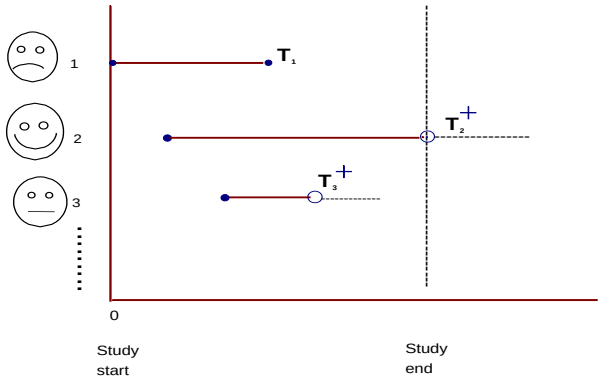
Random censoring

Right censoring is presented here. Left censoring is analogous. Random cen-

soring occurs frequently in medical studies. In clinical trials, patients typically enter a study at different times. Then each is treated with one of several possible therapies. We want to observe their "failure" time but censoring can occur in one of the following ways:

1. *Loss to Follow-up.* Patient moves away. We never see him again. We only know he has survived from entry date until he left. So his survival time is $\geq$ the observed value.
2. *Drop Out.* Bad side effects forces termination of treatment. Or patient refuses to continue treatment for whatever reasons.
3. *Termination of Study.* Patient is still "alive" at end of study.

The following illustrates a possible trial:



Here, patient 1 entered the study at $t = 0$ and died at time T_1 to give an uncensored observation; patient 2 entered the study, and by the end of the study he was still alive resulting in a censored observation T_2^+; and patient 3 entered the study and was lost to follow-up before the end of the study to give another censored observation T_3^+. The AML and CNS lymphoma studies in Examples 1 and 2 contain randomly right-censored data.

Let T denote a lifetime with d.f. F and survivor function S_f and C denote a random censor time with d.f. G, p.d.f. g, and survivor function S_g. Each individual has a lifetime T_i and a censor time C_i. On each of n individuals we observe the pair (Y_i, δ_i) where

$$Y_i = \min(T_i, C_i) \quad \text{and} \quad \delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i \\ 0 & \text{if } C_i < T_i \end{cases}.$$

Hence we observe n iid random pairs (Y_i, δ_i). The times T_i and C_i are usually assumed to be independent. This is a strong assumption. If a patient drops out

because of complications with the treatment (case 2 above), it is clearly offended. However, under the independence assumption, the likelihood function has a simple form (1.12), and even simpler in expression (1.13). Otherwise, we lose the simplicity. The likelihood function becomes very complicated and, hence, the analysis is more difficult to carry out.

Let M and S_m denote the distribution and survivor functions of $Y = \min(T, C)$ respectively. Then by the independence assumption it easily follows that the survivor function is

$$S_m(y) = P(Y > y) = P(T > y, C > y) = P(T > y)P(C > y) = S_f(y)S_g(y).$$

The d.f. of Y is $M(y) = 1 - S_f(y)S_g(y)$.

The likelihood function of the n iid pairs (Y_i, δ_i) is given by

$$\begin{aligned} L &= \prod_{i=1}^{n} \Big(f(y_i)S_g(y_i)\Big)^{\delta_i} \cdot \Big(g(y_i)S_f(y_i)\Big)^{1-\delta_i} \\ &= \left(\prod_{i=1}^{n} \Big(S_g(y_i)\Big)^{\delta_i} \Big(g(y_i)\Big)^{1-\delta_i}\right) \left(\prod_{i=1}^{n} \Big(f(y_i)\Big)^{\delta_i} \Big(S_f(y_i)\Big)^{1-\delta_i}\right) \end{aligned} \quad (1.12)$$

Note: If the distribution of C does not involve any parameters of interest, then the first factor plays no role in the maximization process. Hence, the likelihood function can be taken to be

$$L = \prod_{i=1}^{n} \Big(f(y_i)\Big)^{\delta_i} \cdot \Big(S_f(y_i)\Big)^{1-\delta_i}, \quad (1.13)$$

which has the same form as the likelihood derived for both Type I (1.10) and Type II (1.11) censoring. Thus, regardless of which of the three types of censoring is present, the maximization process yields the same estimated quantities.

Here we see how censoring is incorporated to adjust the estimates. Each observed value is (y_i, δ_i). An individual's contribution is either its p.d.f. $f(y_i)$; or $S_f(y_i)$, the probability of survival beyond its observed censored value y_i. In the complete data setting, all $\delta_i = 1$; that is, there is no censoring. The likelihood has the usual form

$$L = \prod_{i=1}^{n} f(y_i)\,.$$

The derivation of the likelihood is as follows:

$$\begin{aligned} P(Y = y, \delta = 0) &= P(C = y, C < T) = P(C = y, y < T) \\ &= P(C = y)P(y < T) \quad \text{by independence} \\ &= g(y)S_f(y). \\ P(Y = y, \delta = 1) &= P(T = y, T < C) = P(T = y, y < C) = f(y)S_g(y)\,. \end{aligned}$$

Hence, the joint p.d.f. of the pair (Y, δ) (a mixed distribution as Y is continuous and δ is discrete) is given by the single expression

$$P(y, \delta) = \Big(g(y)S_f(y)\Big)^{1-\delta} \cdot \Big(f(y)S_g(y)\Big)^{\delta}.$$

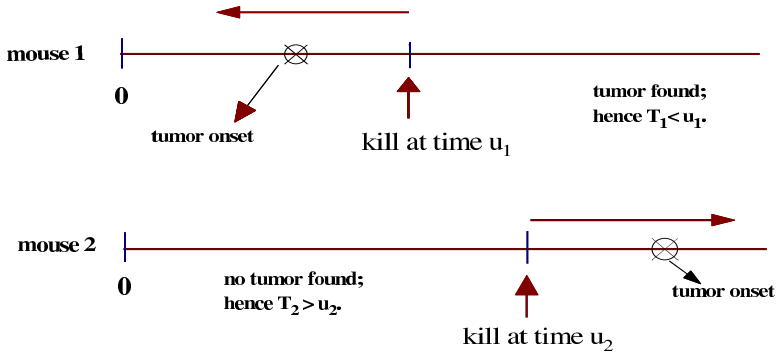
The likelihood of the n iid pairs (Y_i, δ_i) given above follows.

Case 1 Interval Censored Data: Current Status Data Consider the following two examples which illustrate how this type of censoring arises.

Example 3. Tumor free laboratory mice are injected with a tumor inducing agent. The mouse must be killed in order to see if a tumor was induced. So after a random period of time U for each mouse, it is killed and the experimenter checks to see whether or not a tumor developed. The endpoint of interest is T, "time to tumor".

Example 4. An ophthalmologist developed a new treatment for a particular eye disease. To test its effectiveness he must conduct a clinical trial on people. His endpoint of interest is "time to cure the disease". We see this trial could produce right censored data. During the course of this study he notices an adverse side-effect which impairs vision in some of the patients. So now he wants to study "time to side-effect" where he has a control group to compare to the treatment group to determine if this impairment is indeed due to the new treatment. Let's focus on the treatment group. All these patients received the new treatment. In order to determine "time to side-effect" T, he takes a snap-shot view. At a random point in time he checks all patients to see if they developed the side-effect. The records ministry keeps very precise data on when each patient received the new treatment for the disease. So the doctor can look back in time from where he takes his snap-shop to the time of first treatment. Hence for each patient we have an observed U which equals time from receiving new treatment to the time of the snap-shot. If the patient has the side-effect, then his $T \leq U$. If the patient is still free of the side-effect, then his $T > U$.

In both these examples the only available observed time is the U, the censoring time. The following illustrates a possible trial of Example 3.



More formally, we observe only the i.i.d. times $U_i, i = 1, \ldots, n$ and $\delta_i = I\{T_i \leq U_i\}$. That is, $\delta = 1$ if the event $T \leq U$ has occurred, and $\delta = 0$ if the event has not occurred. We assume the **support** (the interval over which the distribution has positive probability) of U is contained in the support of T. As before, the $T \sim F$ and the censor time $U \sim G$ and again we assume T and U are independent random times. The derivation of the joint p.d.f. of the pair of (U, δ) follows:

$$P(U = u, \delta = 0) \quad = P(\delta = 0|U = u)P(U = u) \quad = P(T > u)P(U = u) \quad = S_f(u)g(u).$$

$$P(U = u, \delta = 1) \quad = P(\delta = 1|U = u)P(U = u) \quad = P(T \leq u)P(U = u) \quad = F(u)g(u).$$

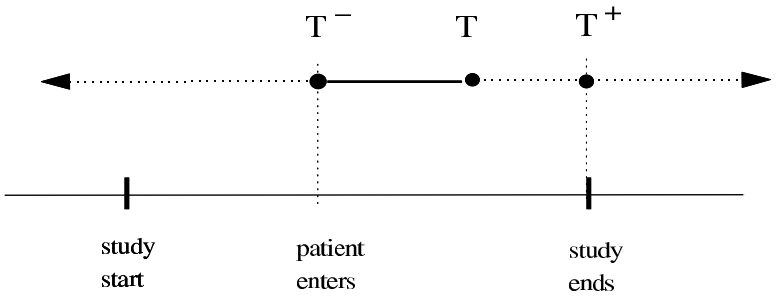
We can write this joint p.d.f. of the pair (U, δ) (again a mixed distribution) in a single expression

$$P(u, \delta) = [S_f(u)]^{1-\delta}[F(u)]^{\delta}g(u).$$

The likelihood of the n i.i.d. pairs (U_i, δ_i) easily follows.

Left Censored and Doubly Censored Data The following two examples illustrate studies where left censored, uncensored, and right censored observations could occur. When all these can occur, this is often referred to as doubly censored data.

Example 5. A child psychiatrist visits a Peruvian village to study the age at which children first learn to perform a particular task. Let T denote the age a child learns to perform a specified task. The following picture illustrates the possible outcomes:



We read the recorded values as follows: T: exact age is observed (uncensored), T^-: age is left censored as the child already knew the task when s/he was initially tested in the study, and T^+: age is right censored since the child did not learn the task during the study period.

Example 6. Extracted from Klein & Moeschberger (1997): High school boys are interviewed to determine the distribution of the age of boys when they first used marijuana. The question stated was "When did you first use marijuana?". The three possible answers and respective recorded values are given in the following table:

Possible answer:	Recorded value:
a I used it but I cannot recall just when the first time was.	a T^-: age of interview as exact age was earlier but unknown
b I first used it when I was ___.	b T: exact age since it is known (uncensored)
c I never used it.	c T^+: age of interview since exact age occurs sometime in the future

Interval Censoring The time-to-event T is known only to occur within an interval. Such censoring occurs when patients in clinical trial or longitudinal study have *periodic* follow-up. For example, women in a study are required to have yearly PAP smear exams. Each patient's event time T_i is only known to fall in an interval $(L_i, R_i]$ which represents the time interval between the visit prior to the visit when the event of interest is detected. The L_i and R_i denote respectively the left and right endpoints of the censoring interval. For example, if the ith patient shows the sign of the symptom at her first follow-up time, then L_i is zero, in other words, the origin of the study and R_i is her first follow-up time. Further, if she showed no sign of the symptom until her $i-1$th follow-up times but shows the sign of the symptom at her ith follow-up, then L_i is her $i-1$th follow-up and R_i is her ith follow-up. If she doesn't exhibit the symptom at her last follow-up, L_i is her last follow-up and R_i is ∞. Note that any combination of left, right, or interval censoring may occur in a study.

Furthermore, we see that left censoring, right censoring, and current status data are special cases of interval censoring.

Truncation Truncation is a procedure where a condition other than the main event of interest is used to screen patients; that is, only if the patient has the truncation condition prior to the event of interest will s/he be observed by the investigator. Hence, there will be subjects "rejected" from the study so that the investigator will never be aware of their existence. This truncation condition may be exposure to a certain disease, entry into a retirement home, or an occurrence of an intermediate event prior to death. In this case, the main event of interest in said to be *left-truncated*. Let U denote the time at which the truncation event occurs and let T denote the time of the main event of interest to occur. Then for left-truncated samples, only individuals with $T \geq U$ are observed. The most common type of *left truncation* occurs when subjects enter the study at a random age and are followed from this *delayed entry time* until the event of interest occurs or the subject is right-censored. In this situation, all subjects who experience the event of interest prior to the delayed entry time will not be known to the experimenter. The following example of *left-truncated* data is described in Klein & Moeschberger (1997, pages 15-17). In Chapter **??** we treat the analysis of left-truncated data.

Example 7. Death Times of Elderly Residents of a Retirement Community Age in months when members of a retirement community died or left the center (right-censored) and age when the members entered the community (the truncation event) are recorded. Individuals must survive to a sufficient age to enter the retirement community. Individuals who die at an early age are excluded from the study. Hence, the life lengths in this data set are *left-truncated*. Ignoring this truncation leads to problem of *length-biased sampling*. We want a survival analysis to account for this type of bias.

Right truncation occurs when only individuals who have experienced the main event of interest are included in the sample. All others are excluded. A mortality study based on death records is a good example of this. The following example of *right-truncated* data is described in Klein & Moeschberger (1997, page 19).

Example 8. Time to AIDS Measurement of interest is the waiting time in years from HIV infection to development of AIDS. In the sampling scheme, only individuals who have developed AIDS prior to the end of the study are included in the study. Infected individuals who have yet to develop AIDS are excluded from the sample; hence, unknown to the investigator. This is a case of *right truncation*.

1.4 Course objectives

The objectives here are to learn methods to model and analyze the data like those presented in the two examples in Section 1.1. We want these statistical procedures to accommodate censored data and to help us attain **the three basic goals of survival analysis** as so succinctly delineated by Kleinbaum (1995, page 15).

In Table 1.2, the graph for **Goal 1** illustrates the survivor functions give very different interpretations. The left one shows a quick drop in survival probabilities early in follow-up. Then the rate of decrease levels off later on. The right function, in contrast, shows a very slow decrease for quite a long while, then a sharp decrease much later on.

In Table 1.2, the plot for **Goal 2** shows that up to 13 weeks, the graph for the new method lies above that for the old. Thereafter the graph for old method is above the new. Hence, this dual graph reveals that up to 13 weeks the new method is more effective than the old; however, after 13 weeks, it becomes less effective.

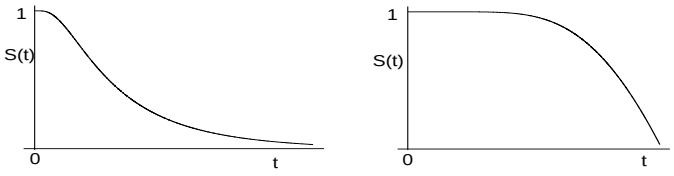
In Table 1.2, the graph for **Goal 3** displays that, for any fixed point in time, up to about 10 years of age, women are at greater risk to get the disease than men are. From 10 to about 40 years of age, men now have a slightly greater risk. For both genders the hazard function decreases as the person ages.

Remark:

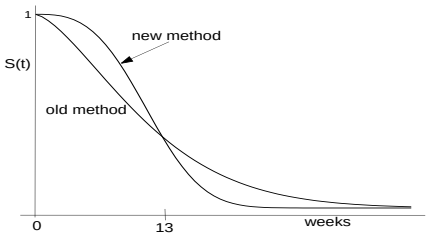
As usual, the emphasis is on modelling and inference. Modelling the hazard function or failure time in turn provides us with estimates of population features such as the mean, the mean residual life, quantiles, HR's, and survival probabilities.

Table 1.2: *Goals of survival analysis*

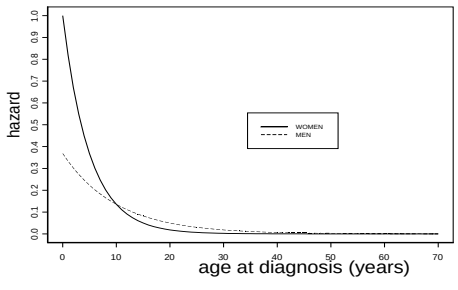
Goal 1. To estimate and interpret survivor and/or hazard functions from survival data.



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.



1.5 Data entry and import/export of data files

The layout is a typical spreadsheet format which is virtually the same for all data analytic software packages. Some examples are EXCEL, SPSS, MINITAB, SAS. The spreadsheet in S-PLUS is the data object called a `data.frame`. On the standard toolbar menu click sequentially on the white blank page at upper far left, `File` → `New` → `Data Set` → `Ok`. A new (empty) `data.frame` will appear. This likens an EXCEL spreadsheet. Double right click on the cell just below the column number to enter the variable name. Below is a table which displays our S-PLUS data set "aml.data" along with a key. This `data.frame` object contains the AML data first given in Table 1.1 under Example 1, page 2. Note that **status variable = the indicator variable δ**. This data set is saved as, e.g., "aml.sdd." You can also save this data set as an Excel file. Just click on `File` → `ExportData` → `ToFile`. Go to `Save as` and click `Type` → `MicrosoftExcelFiles` (*.xls).

	1	2	3	
	weeks	group	status	
1	9	1	1	
2	13	1	1	
3	13	1	0	**group** = 1 for maintained,
4	18	1	1	**group** = 0 for nonmaintained.
·	·	·	·	
·	·	·	·	**status** = 1 if uncensored
·	·	·	·	(relapse occurred),
11	161	1	0	**status** = 0 if censored (still in
12	5	0	1	remission; recorded with + sign).
13	5	0	1	
14	8	0	1	
·	·	·	·	
·	·	·	·	
·	·	·	·	
23	45	0	1	

It seems that EXCEL has spread itself worldwide. All the mainstream statistical packages can accept an EXCEL file. Feel free to first enter your data in an EXCEL spreadsheet. To import into S-PLUS do the following sequentially: in S-PLUS, click on File → `ImportData` → `FromFile` → `FilesofType` → MicrosoftExcelFiles (*.xl*). In `Look In`, find your way to the directory where your desired *.xls data file is. Then right-click on it and click on `Open`. It's now in an S-PLUS data sheet. You can save it in S-PLUS as an S-PLUS data file (`data.frame` object). Click on File, then on Save. It should be clear from this point. Your file will be saved as a *.sdd file.

To import your data file into S or R, first save your EXCEL file, or any other file, as a *.txt file. Be sure to open this file first to see what the delimiter is; that is, what is used to separate the data values entered on each row. Suppose your data file, called your.txt, is in the C: directory. The S and R function `read.table` imports your.txt file and creates a `data.frame` object. When a comma is the delimiter, use the following S line command:

```
> your <- read.table("C://your.txt",header = T,sep = ",")
```

If the delimiter is ~, use `sep = "~"`. If blank space separates the data values, use `sep = " "`. If the space between columns has been tabbed, omit `sep`. In R, to perform a survival analysis it is necessary to install the survival analysis library. The R command is

```
> library(survival)
```

The R function `require(survival)` accomplishes the same.

CHAPTER 2

Nonparametric Methods

We begin with nonparametric methods of inference concerning the survivor function $S(t) = P(T > t)$ and, hence, functions of it.

Objectives of this chapter:

After studying Chapter 2, the student should:

1 Know how to compute the **Kaplan-Meier** (K-M) estimate of survival and **Greenwood's** estimate of asymptotic variance of K-M at time t.

2 Know how to estimate the hazard and cumulative hazard functions.

3 Know how to estimate the pth-quantile.

4 Know how to plot the K-M curve over time t in S.

5 Know how to implement the S function `survfit` to conduct nonparamtric analyses.

6 Know how to plot two K-M curves to compare survival between two (treatment) groups.

7 Be familiar with **Fisher's exact test**.

8 Know how to compute the **log-rank test statistic**.

9 Know how to implement the S function `survdiff` to conduct the log-rank test.

10 Understand why we might **stratify** and how this affects the comparison of two survival curves.

11 Understand how the log-rank test statistic is computed when we stratify on a covariate.

2.1 Kaplan-Meier estimator of survival

We consider the AML data again introduced in Table 1.1, Chapter 1.1. The ordered data is included here in Table 2.1 for ease of discussion.

We first treat this data as if there were NO censored observations. Let t_i

Table 2.1: *Data for the AML maintenance study*

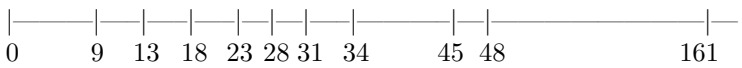
Group	Length of complete remission(in weeks)
Maintained	9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+
Nonmaintained	5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

A + indicates a censored value.

denote an ordered observed value. The **empirical survivor function (esf)**, denoted by $S_n(t)$, is defined to be

$$S_n(t) = \frac{\#\text{ of observations} > t}{n} = \frac{\#\{t_i > t\}}{n}. \tag{2.1}$$

The $S_n(t)$ is the proportion of patients still in remission after t weeks. Let's consider the AML maintained group data (AML1) on a time line:



The values of the **esf** on the maintained group are:

t	0	9	13	18	23	28	31	34	45	48	161
$S_n(t)$	$\frac{11}{11}$	$\frac{10}{11}$	$\frac{8}{11}$	$\frac{7}{11}$	$\frac{6}{11}$	$\frac{5}{11}$	$\frac{4}{11}$	$\frac{3}{11}$	$\frac{2}{11}$	$\frac{1}{11}$	0

The plot of this **esf** function in Figure 2.1 can be obtained by the following S commands. Here status is an 11×1 vector of 1's since we are ignoring that four points are censored. We store the AML data in a data frame called aml. The S function `survfit` calculates the $S_n(t)$ values.

```
> aml1 <- aml[aml$group==1, ] # maintained group only
> status <- rep(1,11)
> esf.fit <- survfit(Surv(aml1$weeks,status)~1)
> plot(esf.fit,conf.int=F,xlab="time until relapse (in weeks)",
          ylab="proportion without relapse",lab=c(10,10,7))
> mtext("The Empirical Survivor Function of the AML Data",3,-3)
> legend(75,.80,c("maintained group","assuming no censored
  data"))
> abline(h=0)
```

The estimated median is the first value t_i where the $S_n(t) \leq 0.5$. Here the

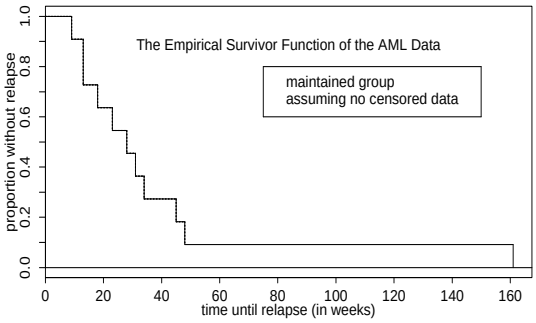


Figure 2.1 *Empirical survivor function (esf).*

$\widehat{\text{med}} = 28$ weeks. The estimated mean (expected value) is

$$\widehat{\text{mean}} = \int_0^\infty S_n(t)\, dt = \text{area under } S_n(t) = \bar{\mathbf{t}}.$$

$S_n(t)$ is a right continuous step function which steps down at each distinct t_i. The estimated mean then is just the sum of the areas of the ten rectangles on the plot. This sum is simply the sample mean. Here the $\widehat{\text{mean}} = \bar{\mathbf{t}} = 423/11 = 38.45$ weeks.

Note: The **esf** is a consistent estimator of the true survivor function $S(t)$. The exact distribution of $nS_n(t)$, for each fixed t, is binomial (n, p), where $n =$ the number of observations and $p = P(T > t)$. Further, it follows from the central limit theorem that for each fixed t,

$$S_n(t) \quad \overset{a}{\sim} \quad \text{normal}(p, p(1-p)/n),$$

where $\overset{a}{\sim}$ is read "approximately distributed as."

We now present the product-limit estimator of survival. This is commonly called the **Kaplan-Meier (K-M) estimator** as it appeared in a seminal 1958 paper.

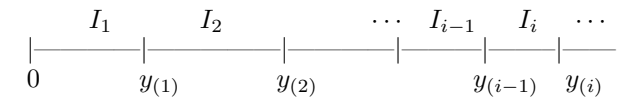
The Product-limit (PL) estimator of $S(t) = P(T > t)$:

K-M adjusts the esf to reflect the presence of right-censored observations.

Recall the random right censoring model in Chapter 1.3. On each of n individuals we observe the pair (Y_i, δ_i) where

$$Y_i = \min(T_i, C_i) \quad \text{and} \quad \delta_i = \begin{cases} 1 & \text{if } T_i \le C_i \\ 0 & \text{if } C_i < T_i. \end{cases}$$

On a time line we have



where $y_{(i)}$ denotes the ith distinct ordered censored or uncensored observation and is the right endpoint of the interval I_i, $i = 1, 2, \ldots, n' \leq n$.

- **death** is the generic word for the event of interest.
  In the AML study, a "relapse" (end of remission period) = "death"
- A **cohort** is a group of people who are followed throughout the course of the study.
- The people at risk at the beginning of the interval I_i are those people who survived (not dead, lost, or withdrawn) the previous interval I_{i-1} .
  Let $\mathcal{R}(t)$ denote the **risk set just before time t** and let

$$\begin{aligned}
n_i &= \#\text{ in } \mathcal{R}(y_{(i)}) \\
&= \#\text{ alive (and not censored) just before } y_{(i)} \\
d_i &= \#\text{ died at time } y_{(i)} \\
p_i &= P(\text{surviving through } I_i \mid \text{alive at beginning } I_i) \\
&= P(T > y_{(i)} \mid T > y_{(i-1)}) \\
q_i &= 1 - p_i \;\; = P(\text{die in } I_i \mid \text{alive at beginning } I_i).
\end{aligned}$$

Recall the general multiplication rule for joint events A_1 and A_2:

$$P(A_1 \cap A_2) = P(A_2 \mid A_1)P(A_1).$$

From repeated application of this product rule the survivor function can be expressed as

$$S(t) = P(T > t) = \prod_{y_{(i)} \leq t} p_i.$$

The estimates of p_i and q_i are

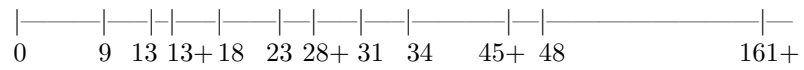
$$\widehat{q_i} = \frac{d_i}{n_i} \qquad \text{and} \qquad \widehat{p_i} = 1 - \widehat{q_i} = 1 - \frac{d_i}{n_i} = \left(\frac{n_i - d_i}{n_i}\right).$$

The **K-M estimator of the survivor function** is

$$\widehat{S}(t) = \prod_{y_{(i)} \leq t} \widehat{p_i} = \prod_{y_{(i)} \leq t} \left(\frac{n_i - d_i}{n_i}\right) = \prod_{i=1}^{k} \left(\frac{n_i - d_i}{n_i}\right), \tag{2.2}$$

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

Let's consider the AML1 data on a time line where a "+" denotes a right-censored observed value. The censored time 13+ we place to the right of the observed relapse time 13 since the censored patient at 13 weeks was still in remission. Hence, his relapse time (if it occurs) is greater than 13 weeks.



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

The K-M curve is a right continuous step function which steps down only at an uncensored observation. A plot of this together with the **esf** curve is displayed in Figure 2.2. The "+" on the K-M curve represents the survival probability at a censored time. Note the difference in the two curves. K-M is

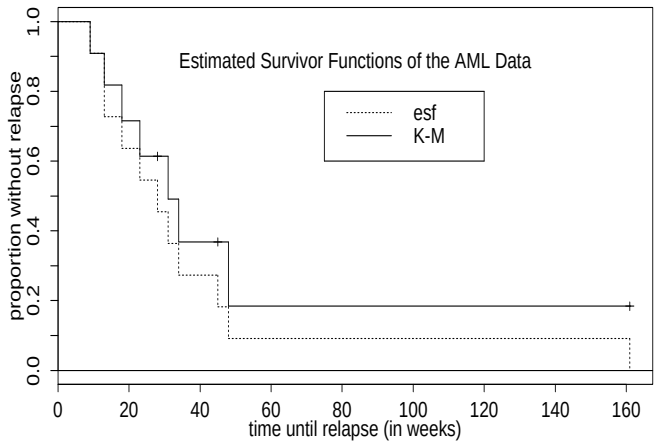


Figure 2.2 *Kaplan-Meier and esf estimates of survival.*

always greater than or equal to **esf**. 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 $\widehat{S}(t)$ as a defective survival function. Alternatively, $\widehat{F}(t) = 1 - \widehat{S}(t)$ is called a subdistribution function as the total probability is less than one.

Estimate of variance of $\widehat{S}(t)$:

Greenwood's formula (1926):

$$\widehat{\text{var}}\left(\widehat{S}(t)\right) = \widehat{S}^2(t) \sum_{y_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)} = \widehat{S}^2(t) \sum_{i=1}^{k} \frac{d_i}{n_i(n_i - d_i)}, \tag{2.3}$$

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

Example with the AML1 data:

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

The theory tells us that for each fixed value t

$$\widehat{S}(t) \overset{a}{\sim} \text{normal}\left(S(t),\ \widehat{\text{var}}\left(\widehat{S}(t)\right)\right).$$

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

$$\widehat{S}(t) \pm \text{z}_{\frac{\alpha}{2}} \times \text{s.e.}\left(\widehat{S}(t)\right), \tag{2.4}$$

where s.e. $\left(\widehat{S}(t)\right)$ is the square root of Greenwood's formula for the estimated variance.

Smith (2002), among many authors, discusses the following estimates of hazard and cumulative hazard. Let t_i denote a distinct ordered death time, $i = 1, \ldots, r \le n$.

Estimates of hazard (risk):

1 Estimate at an observed death time t_i:

$$\widetilde{h}(t_i) = \frac{d_i}{n_i}. \tag{2.5}$$

2 Estimate of hazard in the interval $t_i \leq t < t_{i+1}$:

$$\widehat{h}(t) = \frac{d_i}{n_i(t_{i+1} - t_i)}. \tag{2.6}$$

This is referred to as the K-M type estimate. It estimates the rate of death per unit time in the interval $[t_i, t_{i+1})$.

3 Examples with the AML1 data:

$$\widetilde{h}(23) = \frac{1}{7} = .143$$

$$\widehat{h}(26) = \widehat{h}(23) = \frac{1}{7 \cdot (31 - 23)} = .018$$

Estimates of $H(\cdot)$, cumulative hazard to time t:

1 Constructed with K-M:

$$\widehat{H}(t) = -\log\left(\widehat{S}(t)\right) = -\log \prod_{y_{(i)} \leq t} \left(\frac{n_i - d_i}{n_i}\right), \tag{2.7}$$

$$\widehat{\text{var}}\left(\widehat{H}(t)\right) = \sum_{y_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}\,. \tag{2.8}$$

2 Nelson-Aalen estimate (1972, 1978):

$$\widetilde{H}(t) = \sum_{y_{(i)} \leq t} \frac{d_i}{n_i}\,, \tag{2.9}$$

$$\widehat{\text{var}}\left(\widetilde{H}(t)\right) = \sum_{y_{(i)} \leq t} \frac{d_i}{n_i^2}\,. \tag{2.10}$$

The Nelson-Aalen estimate is the cumulative sum of estimated conditional probabilities of death from I_1 through I_k where $t_k \leq t < t_{k+1}$. This estimate is the first order Taylor approximation to the first estimate. To see this let $x = d_i/n_i$ and expand $\log(1 - x)$ about $x = 0$.

3 Examples with the AML1 data:

$$\widehat{H}(26) = -\log\left(\widehat{S}(26)\right) = -\log(.614) = .488$$

$$\widetilde{H}(26) = \frac{1}{11} + \frac{1}{10} + \frac{1}{8} + \frac{1}{7} = .4588$$

Kernel estimator of hazard:

The kernel estimator of $h(t)$ is given by

$$\widetilde{h}^{kernel}(t) = \frac{1}{b} \sum_{i=1}^{n'} \mathcal{K}\left(\frac{t - y_{(i)}}{b}\right) \frac{d_i}{n_i}\,. \tag{2.11}$$

The *kernel function* $\mathcal{K}$ is a bounded function which vanishes outside $[-1, 1]$ and has integral 1. The *bandwidth* or *window size* b is a positive parameter. The kernel estimator smoothes the occurrence/exposure rates - the increments d_i/n_i of the Nelson-Aalen estimator $\widetilde{H}(t)$ (2.9). In fact, it is a weighted average of the increments over $[t - b, t + b]$. This estimator was proposed and studied by Ramlau-Hansen (1983). He establishes consistency and asymptotic normality. One frequently used kernel is the Epanechnikov kernel $\mathcal{K}(t) = 0.75(1 - t^2)$, $|t| \leq 1$. Another is the biweight kernel $\mathcal{K}(t) = (15/16)(1-t^2)^2$, $|t| \leq 1$. The R function `density` in version 2.2.1 or later can be used to compute a kernel estimate. The `weights` argument is essential and is not available in S or in earlier versions of R. An example is delayed until page 42, where we compare two empirical hazard functions resulting from two treatment groups.

Estimate of quantiles:

Recall the definition:

> the ***p*th-quantile t_p** is such that $F(t_p) = p$ or $S(t_p) = 1 - p$. As usual, when S is continuous, $t_p \leq S^{-1}(1 - p)$.

As the K-M curve is a step function, the inverse is not uniquely defined. We define the estimated quantile to be

$$\widehat{t}_p = \min\{t_i : \widehat{S}(t_i) \leq 1 - p\}. \tag{2.12}$$

By applying the delta method (Chapter 3.2, page 58) to $\widehat{\text{var}}\left(\widehat{S}(\widehat{t}_p)\right)$, Collett (1994, pages 33 and 34) provides the following estimate of variance of $\widehat{t}_p$:

$$\widehat{\text{var}}(\widehat{t}_p) = \frac{\widehat{\text{var}}\left(\widehat{S}(\widehat{t}_p)\right)}{\left(\widehat{f}(\widehat{t}_p)\right)^2}, \tag{2.13}$$

where $\widehat{\text{var}}\left(\widehat{S}(\widehat{t}_p)\right)$ is Greenwood's formula for the estimate of the variance of the K-M estimator, and $\widehat{f}(\widehat{t}_p)$ is the estimated probability density at $\widehat{t}_p$. It is defined as follows:

$$\widehat{f}(\widehat{t}_p) = \frac{\widehat{S}(\widehat{u}_p) - \widehat{S}(\widehat{l}_p)}{\widehat{l}_p - \widehat{u}_p}, \tag{2.14}$$

where $\widehat{u}_p = \max\{t_i | \widehat{S}(t_i) \geq 1 - p + \epsilon\}$, and $\widehat{l}_p = \min\{t_i | \widehat{S}(t_i) \leq 1 - p - \epsilon\}$, for $i = 1, \ldots, r \leq n$ with r being the number of distinct death times, and ϵ a small value. An $\epsilon = 0.05$ would be satisfactory in general, but a larger value of ϵ will be needed if $\widehat{u}_p$ and $\widehat{l}_p$ turn out to be equal. In the following example, we take $\epsilon = 0.05$.

Example with the AML1 data:

The median $\widehat{t}_{.5} = 31$ weeks. We find $\widehat{u}_{.5} = \max\{t_i|\widehat{S}(t_i) \geq 0.55\} = 23$, $\widehat{l}_{.5} = \min\{t_i|\widehat{S}(t_i) \leq 0.45\} = 34$, and $\widehat{f}(31) = \frac{\widehat{S}(23)-\widehat{S}(34)}{34-23} = \frac{0.614-0.368}{11} = 0.0224$. Therefore, its variance and s.e. are

$$\widehat{\text{var}}(31) = \left(\frac{.1642}{.0224}\right)^2 = 53.73 \quad \text{and} \quad \text{s.e.}(31) = 7.33.$$

An approximate 95% C.I. for the median is given by

$$31 \pm 1.96 \times 7.33 \quad \Rightarrow \quad \Big(16.6 \text{ to } 45.4\Big) \text{ weeks.}$$

The truncated mean survival time:

The estimated mean is taken to be

$$\widehat{\text{mean}} \quad = \quad \int_0^{y_{(n)}} \widehat{S}(t)\, dt, \tag{2.15}$$

where $y_{(n)} = \max(y_i)$. If $y_{(n)}$ is uncensored, then this truncated integral is the same as the integral over $[0, \infty)$ since over $[y_{(n)}, \infty)$, $\widehat{S}(t) = 0$. But if the maximum data value is censored, the $\lim_{t\to\infty} \widehat{S}(t) \neq 0$. Thus, the integral over $[0, \infty)$ is undefined. That is, $\widehat{\text{mean}} = \infty$. To avoid this we truncate the integral. By taking the upper limit of integration to be the $y_{(n)}$, we redefined the K-M estimate to be zero beyond the largest observation. Another way to look at this is that we have forced the largest observed time to be uncensored. This does give, however, an estimate biased towards zero. This estimate is the total area under the K-M curve. As $\widehat{S}(t)$ is a step function, we compute this area as the following sum:

$$\widehat{\text{mean}} = \sum_{i=1}^{n'} \big(y_{(i)} - y_{(i-1)}\big)\, \widehat{S}(y_{(i-1)}), \tag{2.16}$$

where $n' = \#$ of distinct observed y_i's, $n' \leq n$, $y_{(0)} = 0$, $\widehat{S}(y_{(0)}) = 1$, and $\widehat{S}(y_{(i-1)})$ is the height of the function at $y_{(i-1)}$.

In the AML1 data, $y_{(n)} = 161$ and, from the following S output, the estimated expected survival time $\widehat{\text{mean}} = 52.6$ weeks with s.e.$(\widehat{\text{mean}}) = 19.8$ weeks. The variance formula for this estimator is given in Remark 5. An estimate of the truncated mean residual life, mrl(t), along with a variance estimate is given in Remark 6.

Note: As survival data are right skewed, the median is the preferred descriptive measure of the typical survival time.

S/R application:

`survfit:`

This is the main S nonparametric survival analysis function. Its main argument takes a `Surv(time,status)` object. We have modified some of the output. Data for both groups in the AML study are in a data frame called **aml**. The "group" variable = 1 for maintained group, = 0 for nonmaintained.

```
> aml1 <- aml[aml$group == 1,  ] # Creates a data frame with
                                 # maintained group data only.
> Surv(aml1$weeks,aml1$status) # Surv object
 [1]   9   13   13+  18   23   28+  31   34   45+  48  161+
> km.fit <- survfit(Surv(weeks,status),type="kaplan-meier",
            data = aml1)
> plot(km.fit,conf.int=F,xlab="time until relapse (in weeks)",
       ylab="proportion in remission",lab=c(10, 10, 7))
> mtext("K-M survival curve for the AML data",3,line=-1,cex=2)
> mtext("maintained group",3,line = -3)
> abline(h=0) # Figure 2.3 is now complete.
> km.fit
    n events mean se(mean) median 0.95LCL 0.95UCL
   11    7   52.6    19.8     31      18      NA
> summary(km.fit) # survival is the estimated S(t).


 time n.risk n.event survival std.err    95% LCL    95% UCL
   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
> attributes(km.fit) # Displays the names of objects we can
                     # access.
$names:
 [1] "time" "n.risk" "n.event" "surv" "std.err" "upper"
 [7] "lower" "conf.type" "conf.int" "call"
$class: [1] "survfit"
 # Example: to access "time" and "surv"
> t.u <- summary(km.fit)$time # t.u is a vector with the
                              # seven uncensored times.
> surv.u <- summary(km.fit)$surv # Contains the estimated
                                 # S(t.u).
```

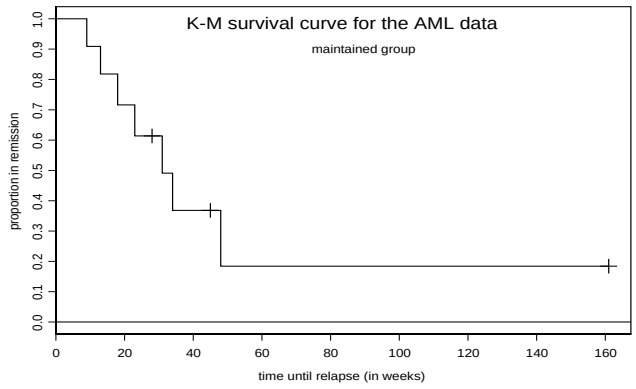


Figure 2.3 *Kaplan-Meier survival curve. A + indicates a censored value.*

Remarks:

1 Notice the effect of accommodating the censored data points. The median time in complete remission is increased from 28 weeks to 31 weeks. The expected time is increased from 38.45 weeks to 52.6 weeks. This explains the third method alluded to in the **A naive descriptive analysis of AML study** presented in Chapter 1.1, page 2.

2 `survfit` uses a simple graphical method of finding a confidence interval for the median. Upper and lower confidence limits for the median are defined in terms of the confidence intervals for $S(t)$: the upper confidence limit is the smallest time at which the upper confidence limit for $S(t)$ is ≤ 0.5. Likewise, the lower confidence limit is the smallest time at which the lower confidence limit for $S(t)$ is ≤ 0.5. That is, draw a horizontal line at 0.5 on the graph of the survival curve, and use intersections of this line with the curve and its upper and lower confidence bands. If, for example, the `UCL` for $S(t)$ never reaches 0.5, then the corresponding confidence limit for the median is unknown and it is represented as an NA. See pages 242 and 243, S-PLUS 2000, Guide to Statistics, Vol.II.

3 Confidence intervals for pth-quantile without using an estimate of the density (2.14) at $\widehat{t}_p$ are also available. See Chapter 4.5, Klein & Moeschberger (1997).

4 The default confidence intervals for $S(t)$ produced by `survfit` are **not** constructed solely with the Greenwood's standard errors (`std.err`) provided in the output. To obtain confidence intervals which use the Greenwood's s.e. directly, you must specify `conf.type="plain"` in the `survfit` function. These correspond to the formula (2.4).

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

$$\exp\left(\log\left(\widehat{S}(t)\right) \pm 1.96\,\text{s.e.}\left(\widehat{H}(t)\right)\right), \tag{2.17}$$

where $\widehat{H}(t)$ is the estimated cumulative hazard function (2.7) and $\text{s.e.}\left(\widehat{H}(t)\right)$ is the square root of the variance (2.8). These `"log"` intervals are derived using the delta method defined in Chapter 3.2, page 58. The log-transform on $\widehat{S}(t)$ produces more efficient intervals as we remove the source of variation due to using $\widehat{S}(t)$ in the variance estimate. Hence, this approach is preferred.

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 $W = \log(-\log(\widehat{S}(t)))$ to estimate the log cumulative hazard parameter $\log(-\log(S(t)))$, and to then transform back. Using the delta method, an estimate of the asymptotic variance of this estimator is given by

$$\widehat{\text{var}}(W) \approx \frac{1}{\left(\log\left(\widehat{S}(t)\right)\right)^2}\widehat{\text{var}}\left(-\log\left(\widehat{S}(t)\right)\right) = \frac{1}{\left(\log\left(\widehat{S}(t)\right)\right)^2}\sum_{y_{(i)}\le t}\frac{d_i}{n_i(n_i-d_i)}. \tag{2.18}$$

An approximate $(1-\alpha)\times 100\%$ C.I. for the quantity $S(t)$ is given by

$$\left(\widehat{S}(t)\right)^{\exp\{z_{\frac{\alpha}{2}}\,\text{s.e.}(W)\}} \le S(t) \le \left(\widehat{S}(t)\right)^{\exp\{-z_{\frac{\alpha}{2}}\,\text{s.e.}(W)\}}. \tag{2.19}$$

To get these intervals specify `conf.type="log-log"` in the `survfit` function. These intervals will always have limits within the interval $[0, 1]$.

5 The variance of the estimated truncated mean survival time (2.15) is

$$\widehat{\text{var}}(\widehat{\text{mean}}) = \sum_{i=1}^{n'}\left(\int_{y_{(i)}}^{y_{(n)}}\widehat{S}(u)du\right)^2\frac{d_i}{n_i(n_i-d_i)}. \tag{2.20}$$

The quantity `se(mean)` reported in the `survfit` output is the square root of this estimated variance.

6 An estimate of the truncated mean residual life at time t (1.9), denoted by $\widehat{\text{mrl}}(t)$, is taken to be

$$\widehat{\text{mrl}}(t) = \frac{\int_t^{y_{(n)}}\hat{S}(u)du}{\widehat{S}(t)} \tag{2.21}$$

with estimated variance

$$\widehat{\text{var}}\left(\widehat{\text{mrl}}(t)\right) = \frac{1}{\widehat{S}^2(t)}\left(\sum_{t\le y_{(i)}\le y_{(n)}}\left(\int_{y_{(i)}}^{y_{(n)}}\widehat{S}(u)du\right)^2\frac{d_i}{n_i(n_i-d_i)} - \left(\widehat{\text{mrl}}(t)\right)^2\sum_{y_{(i)}\le t}\frac{d_i}{n_i(n_i-d_i)}\right). \tag{2.22}$$

The `hazard.km` and `quantile.km` functions:

The function `hazard.km` takes a `survfit` object for its argument. It outputs $\hat{h}(t)$, $\tilde{h}(t_i)$, $\widehat{H}(t)$, se$(\widehat{H}(t))$, $\widetilde{H}(t)$, and se$(\widetilde{H}(t))$. The function `quantile.km` computes an estimated pth-quantile along with its standard error and an approximate $(1-\alpha)\times 100\%$ confidence interval. It has four arguments: `(data,p,eps,z)`, where `data` is a `survfit` object, `p` is a scalar between 0 and 1, `eps` (ϵ) is .05 or a little larger, and `z` is the standard normal z-score needed for the desired confidence level.

```
> hazard.km(km.fit)
    time ni di  hihat hitilde   Hhat se.Hhat Htilde se.Htilde
  1    9 11  1 0.0227  0.0909 0.0953  0.0953 0.0909    0.0909
  2   13 10  1 0.0200  0.1000 0.2007  0.1421 0.1909    0.1351
  3   18  8  1 0.0250  0.1250 0.3342  0.1951 0.3159    0.1841
  4   23  7  1 0.0179  0.1429 0.4884  0.2487 0.4588    0.2330
  5   31  5  1 0.0667  0.2000 0.7115  0.3345 0.6588    0.3071
  6   34  4  1 0.0179  0.2500 0.9992  0.4418 0.9088    0.3960
  7   48  2  1     NA  0.5000 1.6923  0.8338 1.4088    0.6378
> quantile.km(km.fit,.25,.05,1.96) # the .25th-quantile
 [1] "summary"
   qp se.S.qp   f.qp  se.qp   LCL    UCL
 1 18  0.1397 0.0205 6.8281 4.617 31.383  # in weeks
```

Remarks:

1 In the case of no censoring, `quantile.km` differs from the S function `quantile`. Try `quantile(1:10,c(.25,.5,.75))` and compare `quantile.km` after using `survfit(Surv(1:10,rep(1,10)))`.

2 If we extend the `survfit` graphical method to find the confidence limits for a median to the .25th quantile, we get `13` and `NA` as the lower and upper limits, respectively. WHY! See Remark 2, page 33.

2.2 Comparison of survivor curves: two-sample problem

For the AML data the variable "weeks" contains all 23 observations from both groups.

There is now the variable group:

$$\text{group} = \begin{cases} 1 & \text{for maintained} \\ 0 & \text{for nonmaintained.} \end{cases}$$

A plot of the K-M curves for both groups is displayed in Figure 2.4. A summary of the survival estimation using the `survfit` function follows:

```
> 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) # This displays the survival probability
            # table for each group. The output is omitted.
> km.fit
          n  events  mean  se(mean) median 0.95LCL 0.95UCL
 group=0 12     11   22.7     4.18     23       8      NA
 group=1 11      7   52.6    19.83     31      18      NA
```

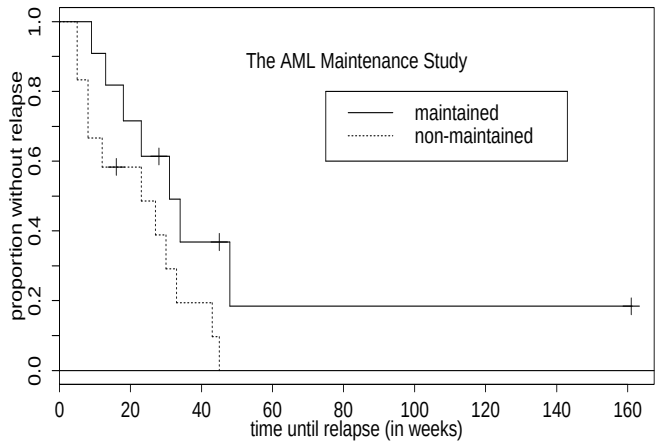


Figure 2.4 *A comparison of two K-M curves.*

- Notice the estimated mean, median, and survivor curve of "maintained" group are higher than those of the other group.
- Is there a significant difference between the two survivor curves? Does maintenance chemotherapy statistically prolong time until relapse?

To test $H_0 : F_1 = F_2$, we present the Mantel-Haenszel (1959) test, also called the **log-rank test**. Another well known test is the Gehan (1965) test, which is an extension of the Wilcoxon test to accommodate right-censored data. See Miller (1981, Chapter 4.1) for a presentation of this test. To motivate

the construction of the Mantel-Haenszel test statistic, we first briefly study Fisher's exact test.

Comparing two binomial populations:

Suppose we have two populations, and an individual in either population can have one of two characteristics. For example, Population 1 might be cancer patients under a certain treatment and Population 2 cancer patients under a different treatment. The patients in either group may either die within a year or survive beyond a year. The data are summarized in a 2×2 contingency table. Our interest here is to compare the two binomial populations, which is common in medical studies.

	Dead	Alive	
Population 1	a	b	n_1
Population 2	c	d	n_2
	m_1	m_2	n

Denote

$$\begin{aligned} p_1 &= P\{\text{Dead}|\text{Population 1}\}, \\ p_2 &= P\{\text{Dead}|\text{Population 2}\}. \end{aligned}$$

Want to test

$$H_0 : p_1 = p_2.$$

Fisher's exact test:

The random variable A, which is the entry in the $(1,1)$ cell of the 2×2 table, has the following **exact discrete conditional distribution under $\mathbf{H_0}$**: Given n_1, n_2, m_1, m_2 fixed quantities, it has a **hypergeometric distribution** where

$$P\{A = a\} = \frac{\binom{n_1}{a}\binom{n_2}{m_1 - a}}{\binom{n}{m_1}}.$$

The test based on this exact distribution is called the **Fisher's exact test**. The S function `fisher.test` computes an exact p-value. The mean and variance of the hypergeometric distribution are

$$\begin{aligned} E_0(A) &= \frac{n_1 m_1}{n}, \\ Var_0(A) &= \frac{n_1 n_2 m_1 m_2}{n^2(n-1)}. \end{aligned}$$

We can also conduct an approximate chi-square test when samples are large as

$$\chi^2 = \left(\frac{a - E_0(A)}{\sqrt{Var_0(A)}} \right)^2 \overset{a}{\sim} \chi^2_{(1)},$$

where $\chi^2_{(1)}$ denotes a chi-square random variable with 1 degree of freedom.

Mantel-Haenszel/log-rank test:

Now suppose we have a sequence of 2×2 tables. For example, we might have k hospitals; at each hospital, patients receive either Treatment 1 or Treatment 2 and their responses are observed. Because there may be differences among hospitals, we do not want to combine all k tables into a single 2×2 table. We want to test

$$H_0 : p_{11} = p_{12}, \text{ and} \ldots, \text{ and } p_{k1} = p_{k2},$$

where

$$\begin{aligned} p_{i1} &= P\{\text{Dead}|\text{Treatment 1}, \text{Hospital } i\}, \\ p_{i2} &= P\{\text{Dead}|\text{Treatment 2}, \text{Hospital } i\}. \end{aligned}$$

	Dead	Alive	
Treatment 1	a_1		n_{11}
Treatment 2			n_{12}
	m_{11}	m_{12}	n_1

Hospital 1

$\vdots$

	Dead	Alive	
Treatment 1	a_k		n_{k1}
Treatment 2			n_{k2}
	m_{k1}	m_{k2}	n_k

Hospital k

Use the **Mantel-Haenszel (1959) statistic**

$$\text{MH} = \frac{\sum_{i=1}^{k}(a_i - E_0(A_i))}{\sqrt{\sum_{i=1}^{k} Var_0(A_i)}}. \tag{2.23}$$

If the tables are independent, then MH $\overset{a}{\sim}$ N(0, 1) either when k is fixed and $n_i \to \infty$ or when $k \to \infty$ and the tables are also identically distributed.

In survival analysis the MH statistic is applied as follows: Combine the two samples, order them, and call them $z_{(i)}$. Construct a 2×2 table for each uncensored time point $z_{(i)}$. Compute the MH statistic for this sequence of tables to test $H_0 : F_1 = F_2$. The theory tells us that asymptotic normality still holds even though these tables are clearly not independent.

We illustrate how to compute the MH with the following fictitious data:

Treatment Old	3, 5, 7, 9+, 18
Treatment New	12, 19, 20, 20+, 33+

Computations for the MH are given in the following table. Denote the combined ordered values by z. Note that n is the total number of patients at risk in both groups; m_1 the number of patients who died at the point z; n_1 the number at risk in treatment Old at time z; a equals 1 if death in Old or 0 if death in New. Remember that

$$E_0(A) = \frac{m_1 n_1}{n} \quad \text{and} \quad Var_0(A) = \frac{m_1(n-m_1)}{n-1} \times \frac{n_1}{n}\left(1 - \frac{n_1}{n}\right).$$

trt	z	n	m_1	n_1	a	$E_0(A)$	r	$\frac{m_1(n-m_1)}{n-1}$	$\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)$
Old	3	10	1	5	1	.50	.50	1	.2500
Old	5	9	1	4	1	.44	.56	1	.2469
Old	7	8	1	3	1	.38	.62	1	.2344
Old	9+		0		0				
New	12	6	1	1	0	.17	−.17	1	.1389
Old	18	5	1	1	1	.20	.80	1	.1600
New	19	4	1	0	0	0	0	1	0
New	20	3	1	0	0	0	0	1	0
New	20+								
New	33+								
Total					4	1.69	2.31		1.0302

where $r = (a - E_0(A))$. Then

$$\begin{aligned} \text{MH} &= \frac{\text{sum of } (a - E_0(A))}{\sqrt{\text{sum of } \left(\frac{m_1(n-m_1)}{n-1} \times \frac{n_1}{n}(1 - \frac{n_1}{n}) \right)}} \\ &= \frac{2.31}{1.02} = 2.26 \\ p\text{-value} &= 0.012 \quad (\text{one-tailed} \quad Z \text{ test}). \end{aligned}$$

The S function `survdiff` provides the log-rank (= MH) test by default. Its first argument takes a `Surv` object. It gives the square of the MH statistic which is then an approximate chi-square statistic with 1 degree of freedom. This is a two-tailed test. Hence, the p-value is twice that of the MH above. Except for round-off error, everything matches.

```
> grouph <- c(1,1,1,1,1,2,2,2,2,2) # groups: 1=old; 2=new
> hypdata <- c(3,5,7,9,18,12,19,20,20,33) # the data
> cen <- c(1,1,1,0,1,1,1,1,0,0) # censor status:
                                # 1=uncensored; 0=censored
> survdiff(Surv(hypdata,cen)~grouph)

           N  Observed Expected (O-E)^2/E  (O-E)^2/V
 grouph=1  5       4     1.69      3.18       5.2
 grouph=2  5       3     5.31      1.01       5.2

 Chisq = 5.2  on 1 degrees of freedom, p = 0.0226
 # This p-value corresponds to a two-tailed Z-test
 # conducted with MH.
> sqrt(5.2) # square root of log-rank test statistic.
[1] 2.280351 # MH.
 # .0226 = (1 - pnorm(2.280351))*2: p-value for two-sided test
> .0226/2
[1] 0.0113 # p-value for one-sided test.
```

The log-rank test on the AML data is:

```
> 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 $.0653/2 = .033$.

Remark:

The `survdiff` function contains a "rho" parameter. The default value, rho = 0, gives the log-rank test. When 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, $\widetilde{h}(t_i)$ and $\widehat{h}(t)$, defined on page 28, to form ratios and then interpret them in the context of the AML study. The function `emphazplot` contains an abridged form of the `hazard.km` function (page 35) and produces two plots, one for each of the two types of hazard estimates. Modified output and plots follow.

```
> 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
 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()
```

Consider the following two hazard ratios of nonmaintained to maintained:

$$\frac{\widehat{h}_{\mathrm{nm}}(15)}{\widehat{h}_{\mathrm{m}}(15)} = \frac{.011}{.020} = .55 \qquad \text{and} \qquad \frac{\widehat{h}_{\mathrm{nm}}(25)}{\widehat{h}_{\mathrm{m}}(25)} = \frac{.042}{.018} = 2.33\,.$$

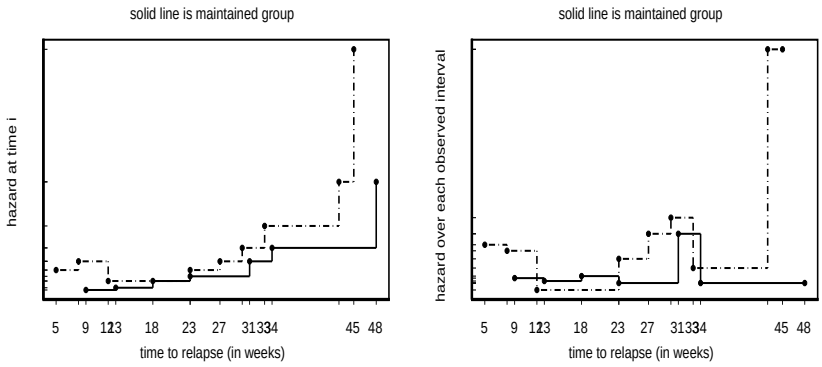


Figure 2.5 *A comparison of empirical hazards. Left plot displays* $\widetilde{h}(t_i)$. *Right plot displays* $\widehat{h}(t)$.

The nonmaintained group has 55% of the risk of those maintained of relapsing at 15 weeks. However, on the average, those nonmaintained have 2.33 times the risk of those maintained of relapsing at 25 weeks.

Neither of the two plots in Figure 2.5 displays roughly parallel curves over time. In the second plot, the hazard curves cross over time. One group's risk is not always lower than the other's with respect to time. This causes the above HR's to change values. **Both plots indicate the hazard ratio is not constant with respect to follow-up time**, which says the hazard functions of the two groups are not proportional. The notion of proportional hazards is a central theme threaded throughout survival analyses. It is discussed in detail in Chapters 4, 5, and 6.

With larger datasets the plots in Figure 2.5 will be chaotic. The smoothed d_i/n_i obtained via the kernel estimator (2.11) provide a far clearer picture of hazard and are very useful when comparing curves. The essential pieces of R code follow: Let $g = 0, 1$.

```
> fit.g <- summary(survfit(Surv(weeks,status),subset=group==g,
                conf.type="n",data=aml),censor=T)
> u.g <- fit.g$time
> weight.g <- fit.g$n.event/fit.g$n.risk
> smooth.g <- density(u.g,kernel="epanechnikov",
                weights=weight.g,n=50,from=0,to=50)
> plot(smooth.g$x,smooth.g$y,type="l",...)
```

Figure 2.6 shows the maintained group always has lower risk. Both hazards increase linearly until about 26 weeks. At about 40 weeks the nonmaintained group's risk increases quadratically with a maximum at 40 weeks, whereas the hazard for the maintained group is essentially constant after 26 weeks.

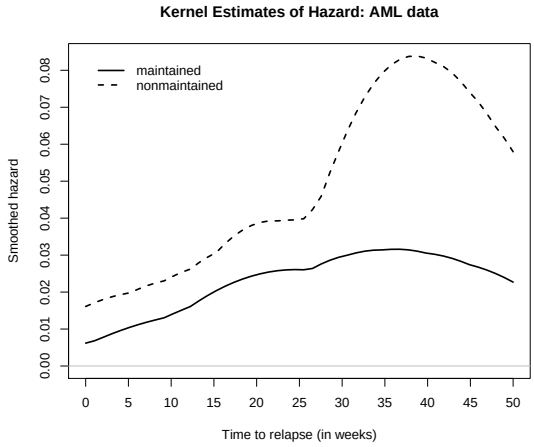


Figure 2.6 *Smoothed estimates,* $\widetilde{h}_g^{kernel}(t)$, $g = 0, 1$, *of hazards. The Epanechnikov kernel* $\mathcal{K}(t) = 0.75(1 - t^2), |t| \leq 1$ *was used.*

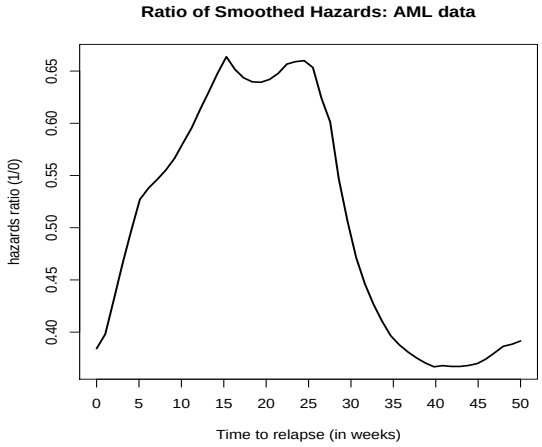


Figure 2.7 *Ratio of smoothed hazards for AML data.*

Figure 2.7 clearly shows that the hazard functions are not proportional as their ratio is not constant over time. At 15 weeks we estimate the maintained group has about 66% of the risk of those nonmaintained of relapsing; or, those nonmaintained have 1.52 times the risk of those maintained of relapsing at 15 weeks. At 25 weeks the risk is slightly higher.

The plot in Figure 2.7 is only an illustration of how to visualize and interpret HR's. Of course, statistical accuracy (confidence bands) should be incorpo-

rated as these comments may not be statistically significant. Pointwise 95% bootstrap confidence limits for the log-HR are commonly reported.

Stratifying on a covariate:

- Stratifying on a particular covariate is one method that can account for (adjust for) its possible confounding and/or interaction effects with the treatment of interest on the response.
- Confounding and/or interaction effects of other known factors with the treatment variable **can mask the "true" effects of the treatment of interest.** Thus, stratification can provide us with stronger (or weaker) evidence, or more importantly, reverse the sign of the effect. That is, it is possible for the aggregated data to suggest treatment is favorable when in fact, in every subgroup, it is highly unfavorable; and vice versa. This is known as **Simpson's paradox** (Simpson, 1951).

Let's consider the fictitious data again and see

1 What happens when we stratify by sex?

2 How is the log-rank statistic computed?

Recall:

```
grouph <- c(1,1,1,1,1,2,2,2,2,2) # groups: 1 = old 2 = new
hypdata <- c(3,5,7,9,18,12,19,20,20,33) # the data
cen <- c(1,1,1,0,1,1,1,1,0,0) # censor status:
                              1 = uncensored; 0 = censored
```

How to:

Separate the data by sex. Then, within each sex stratum, construct a sequence of tables as we did above. Then combine over the two sexes to form $(\text{MH})^2$. According to the sex vector

$$\texttt{sex <- c(}\overbrace{1,1,1,2,2}^{\text{old}},\overbrace{2,2,2,1,1}^{\text{new}}\texttt{)}, \quad \text{where } 1 = \text{male} \quad 2 = \text{female}.$$

Within each stratum, n is the total number at risk, m_1 the number who die at point z, n_1 the number at risk in treatment Old at time z, and a equals 1 if death in Old or 0 if death in New.

MALE : Old 3, 5, 7
New 20+, 33+

trt	z	n	m_1	n_1	a	$E_0(A)$	$\frac{m_1(n-m_1)}{n-1}$	$\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)$
Old	3	5	1	3	1	.60	1	.24
Old	5	4	1	2	1	.50	1	.25
Old	7	3	1	1	1	.333333	1	.222222
New	20+	2						
New	33+	1						
Total					3	1.433333		.712222

Note: $E_0(A) = \frac{n_1 m_1}{n}$ and $Var_0(A) = \frac{m_1(n-m_1)}{n-1} \times \frac{n_1}{n}\left(1-\frac{n_1}{n}\right)$.

FEMALE : Old 9+, 18
New 12, 19, 20

trt	z	n	m_1	n_1	a	$E_0(A)$	$\frac{m_1(n-m_1)}{n-1}$	$\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)$
Old	9+	5						
New	12	4	1	1	0	.25	1	.1875
Old	18	3	1	1	1	.333333	1	.222222
New	19	2	1	0	0	0		0
New	20	1	1	0	0	0		0
Total					1	.583333		.409722

Then pooling by summing over the two tables, we have $a = 4$, $E_0(A) = 1.433333 + .583333 = 2.016666$, and $Var_0(A) = .712222 + .409722 = 1.121944$. The log-rank statistic is

$$(\text{MH})^2 = \frac{(4 - 2.016666)^2}{1.121944} = 3.506,$$

which matches the following S output from `survdiff`. Note the `strata(sex)` term that has been included in the model statement within the `survdiff` function.

```
 # sex = 1 for male, sex = 2 for female
 # group = 1 for old, group = 2 for new treatment

> survdiff(Surv(hypdata,cen)~grouph+strata(sex))

           N Observed Expected (O-E)^2/E (O-E)^2/V
 grouph=1  5    4       2.02     1.951     3.51
 grouph=2  5    3       4.98     0.789     3.51

 Chisq= 3.5  on 1 degrees of freedom, p= 0.0611
```

Note that the p-value of a one-sided alternative is $0.0611/2 = .031$. Although there is still significant evidence at the .05 level that the new treatment is better, it is not as strong as before we stratified. That is, after taking into account the variation due to sex, the difference between treatments is not as strong.

At `www.mth.pdx.edu/~mara/ndk_August_2006.htm`, the interested reader may download Example of Simpson's paradox.

CHAPTER 3

Parametric Methods

Objectives of this chapter:

After studying Chapter 3, the student should:

1 Be familiar with six distributional models.

2 Be able to describe the behavior of their hazard functions.

3 Know that the log-transform of three of these lifetime distributions transforms into a familiar **location and scale family**; and know the relationships between the parameters of the transformed model and those in the original model.

4 Know how to construct a **Q-Q plot** for each of these log(time) distributions.

5 Know the definition of a **likelihood function**.

6 Understand the method of **maximum likelihood estimation** (MLE).

7 Know how to apply the **delta method**.

8 Understand the concept of **likelihood ratio test (LRT)**.

9 Know the general form of the likelihood function for randomly censored data.

10 Understand how to apply the above estimation and testing methods under the exponential model to one sample of data containing censored values. Hence, be familiar with the example of fitting the AML data to an exponential model.

11 Be familiar with the S function `survReg` used to provide a parametric description and analysis of censored data; in particular, how to fit data to the Weibull, log-logistic, and log-normal models.

12 Know how to apply `survReg` to the one-sample and two-sample problems. Be familiar with the additional S functions `anova`, `predict`, and the functions `qq.weibull`, `qq.loglogistic`, `qq.weibreg`, `qq.loglogisreg`, and `qq.lognormreg`, which produce Q-Q plots for one or several samples.

3.1 Frequently used (continuous) models

The exponential distribution

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

The outstanding simplicity of this model is its constant hazard rate. We display some p.d.f.'s and survivor functions for three different values of λ in Figure 3.1. The relationship between the cumulative hazard and the survivor

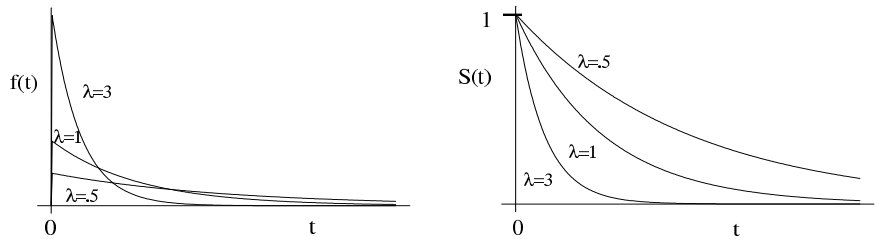


Figure 3.1 *Exponential density and survivor curves.*

function (1.6) is

$$\log\big(H(t)\big) = \log\big(-\log(S(t))\big) = \log(\lambda) + \log(t)$$

or, equivalently expressed with $\log(t)$ on the vertical axis,

$$\boxed{\log(t) = -\log(\lambda) + \log\big(-\log(S(t))\big).} \tag{3.1}$$

Hence, the plot of $\log(t)$ versus $\log\big(-\log(S(t))\big)$ is a straight line with slope 1 and y-intercept $-\log(\lambda)$. At the end of this section we exploit this linear relationship to construct a Q-Q plot for a graphical check of the goodness of fit of the exponential model to the data. Since the hazard function, $h(t) = \lambda$, is constant, plots of both empirical hazards, $\widetilde{h}(t_i)$ and $\widehat{h}(t)$ (page 28), against time provide a quick graphical check. For a good fit, the plot patterns should resemble horizontal lines. Otherwise, look for another survival model. The parametric approach to estimating quantities of interest is presented in Section 3.4. There we first illustrate this with an uncensored sample. Then the same approach is applied to a censored sample. The exponential is a special case of both the Weibull and gamma models, each with their shape parameter equal to 1.

The 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 $Var(T)$	pth-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 3.2 displays p.d.f.'s and hazard functions, respectively.

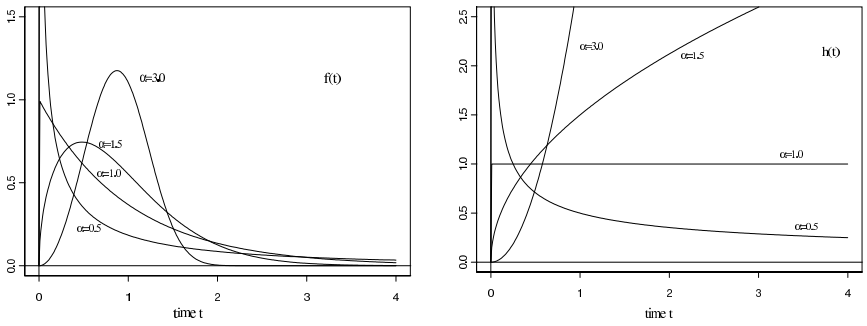


Figure 3.2 *Weibull density and hazard functions with* $\lambda = 1$.

Note that the Weibull hazard function is monotone increasing 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. This is clearly seen in Figures 3.2. 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.

This model is very flexible and has been found to provide a good description of many types of time-to-event data. We might expect 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. We might expect some 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.

The relationship between the cumulative hazard $H(t)$ and the survivor $S(t)$ (1.6) is seen to be

$$\log\big(H(t)\big) = \log\big(-\log(S(t))\big) = \alpha(\log(\lambda) + \log(t)) \tag{3.2}$$

or equivalently expressed as

$$\boxed{\log(t) = -\log(\lambda) + \sigma \log\big(-\log(S(t))\big)}, \tag{3.3}$$

where $\sigma = 1/\alpha$. The plot of $\log(t)$ versus $\log\big(-\log(S(t))\big)$ is a straight line with slope $\sigma = 1/\alpha$ and y-intercept $-\log(\lambda)$. Again, we can exploit this linear relationship to construct a Q-Q plot.

An example of fitting data to the Weibull model using S, along with its Q-Q plot, is presented in Section 3.4. This distribution is intrinsically related to the extreme value distribution which is the next distribution to be discussed. The natural log transform of a Weibull random variable produces an extreme value random variable. This relationship is exploited quite frequently, particularly in the statistical computing packages and in diagnostic plots.

The 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 $S(y)$	
$\sigma^{-1}\exp\big(\frac{y-\mu}{\sigma} - \exp\big(\frac{y-\mu}{\sigma}\big)\big)$	$\exp\big(-\exp\big(\frac{y-\mu}{\sigma}\big)\big)$	
mean $E(Y)$	variance $Var(Y)$	pth - quantile y_p
$\mu - \gamma\sigma$	$\frac{\pi^2}{6}\sigma^2$	$y_p = \mu$ $+\sigma\log\big(-\log(1-p)\big)$

Here γ denotes Euler's constant, $\gamma = 0.5772...$, the location parameter μ is the 0.632th quantile, and y can also be negative so that $-\infty < y < \infty$. Further, the following relationship can be easily shown:

Fact: If T is a Weibull random variable with parameters α and λ, then $Y = \log(T)$ follows an extreme value distribution with $\mu = -\log(\lambda)$ and $\sigma = \alpha^{-1}$. The r.v. Y can be represented as $Y = \mu + \sigma Z$, where Z is a standard extreme value r.v., as the extreme value distribution is a location and scale family of distributions.

As values of μ and σ different from 0 and 1 do not effect the shape of the p.d.f., but only location and scale, displaying only plots of the standard extreme value p.d.f. and survivor function in Figure 3.3 suffices.

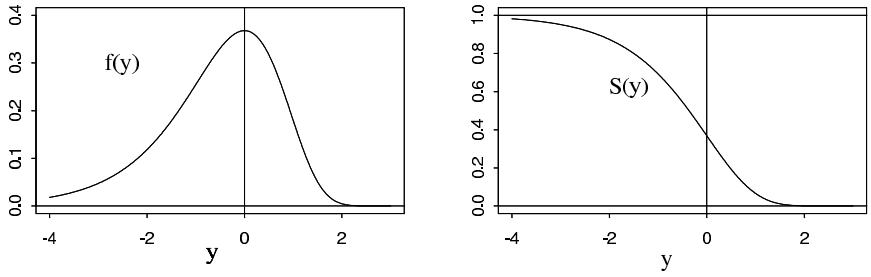


Figure 3.3 *Standard extreme value density and survivor functions.*

The log-normal distribution

This distribution is most easily characterized by saying the lifetime T is log-normally distributed if $Y = \log(T)$ is normally distributed with mean and variance specified by μ and σ^2, respectively. Hence, Y is of the form $Y = \mu + \sigma Z$ where Z is a standard normal r.v. We have the following table for T with $\alpha > 0$ and $\lambda > 0$ and where $\Phi(\cdot)$ denotes the standard normal d.f.:

p.d.f. $f(t)$	survivor $S(t)$	hazard $h(t)$
$(2\pi)^{-\frac{1}{2}}\alpha t^{-1}\exp\left(\frac{-\alpha^2(\log(\lambda t))^2}{2}\right)$	$1-\Phi\Big(\alpha\log(\lambda t)\Big)$	$\frac{f(t)}{S(t)}$
mean $E(T)$	variance $Var(T)$	**Note**:
$\exp(\mu+\frac{\sigma^2}{2})$	$(\exp(\sigma^2)-1)\times$ $\exp(2\mu+\sigma^2)$	$\mu = -\log(\lambda)$ and $\sigma = \alpha^{-1}$

The hazard function has value 0 at $t = 0$, increases to a maximum, and then decreases, approaching zero as t becomes large. Since the hazard decreases for large values of t, it seems implausible as a lifetime model in most situations. But, it can still be suitable for representing lifetimes, particularly when large values of t are not of interest. We might also expect this hazard to describe tuberculosis patients well. Their potential for dying increases early in the disease and decreases later. Lastly, the log-logistic distribution, to be presented next, is known to be a good approximation to the log-normal and is often a preferred survival time model. Some p.d.f's and hazard functions are displayed in Figure 3.4.

The log-logistic distribution

The lifetime T is log-logistically distributed if $Y = \log(T)$ is logistically distributed with location parameter μ and scale parameter σ. Hence, Y is also

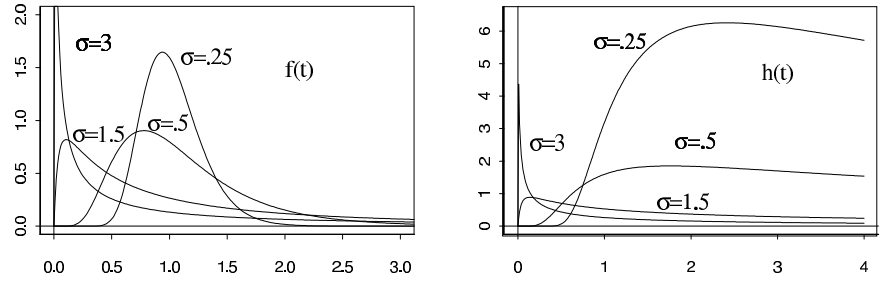


Figure 3.4 *Log-normal densities and hazards with* $\mu = 0$ *and* $\sigma = .25, .5, 1.5,$ *and* 3.

of the form $Y = \mu + \sigma Z$ where Z is a standard logistic r.v. with density

$$\frac{\exp(z)}{\Big(1 + \exp(z)\Big)^2}, \quad -\infty < z < \infty.$$

This is a symmetric density with mean 0 and variance $\pi^2/3$, and with slightly heavier tails than the standard normal, the excess in kurtosis being 1.2. We have the following table for T with $\alpha > 0$ and $\lambda > 0$:

p.d.f. $f(t)$	survivor $S(t)$	hazard $h(t)$
$\lambda\alpha(\lambda t)^{\alpha-1}\Big(1 + (\lambda t)^\alpha\Big)^{-2}$	$\frac{1}{1+(\lambda t)^\alpha}$	$\frac{\lambda\alpha(\lambda t)^{\alpha-1}}{1+(\lambda t)^\alpha}$
Note:		pth-quantile t_p
$\mu = -\log(\lambda)$ and $\sigma = \alpha^{-1}$		$\lambda^{-1}\left(\frac{p}{1-p}\right)^{\frac{1}{\alpha}}$

This model has become popular, for like the Weibull, it has simple algebraic expressions for the survivor and hazard functions. Hence, handling censored data is easier than with the log-normal while providing a good approximation to it except in the extreme tails. The hazard function is identical to the Weibull hazard aside from the denominator factor $1 + (\lambda t)^\alpha$. For $\alpha < 1$ ($\sigma > 1$) it is monotone decreasing from ∞ and is monotone decreasing from λ if $\alpha = 1$ ($\sigma = 1$). If $\alpha > 1$ ($\sigma < 1$), the hazard resembles the log-normal hazard as it increases from zero to a maximum at $t = (\alpha - 1)^{1/\alpha}/\lambda$ and decreases toward zero thereafter. In Section 3.4 an example of fitting data to this distribution using S along with its Q-Q plot is presented. Some p.d.f.'s and hazards are displayed in Figure 3.5.

We exploit the simple expression for the survivor function to obtain a relationship which is used for checking the goodness of fit of the log-logistic model to the data. The odds of survival beyond time t are

$$\frac{S(t)}{1 - S(t)} = (\lambda t)^{-\alpha}. \tag{3.4}$$

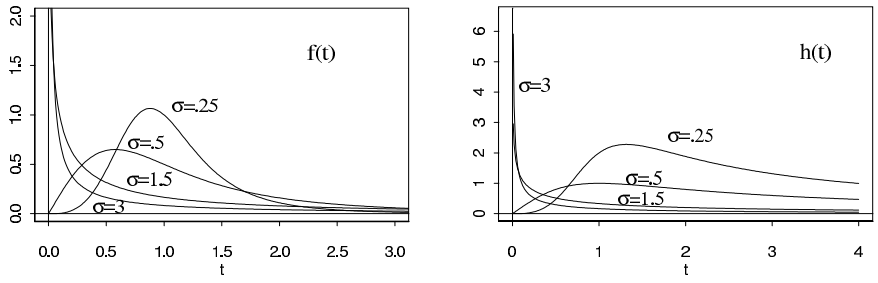


Figure 3.5 *Log-logistic densities and hazards with* $\mu = 0$ *and* $\sigma = .25, .5, 1.5,$ *and* 3.

It easily follows that $\log(t)$ is a linear function of the log-odds of survival beyond t. The precise linear relationship is

$$\boxed{\log(t) = \mu + \sigma\Big(- \log\Big(\tfrac{S(t)}{1-S(t)}\Big)\Big)}, \tag{3.5}$$

where $\mu = -\log(\lambda)$ and $\sigma = 1/\alpha$. The plot of the $\log(t)$ against $-\log\{S(t)/(1-S(t))\}$ is a straight line with slope σ and y-intercept μ. At the end of this section, the Q-Q plot is constructed using this linear relationship.

The gamma distribution

Like the Weibull, this distribution has a scale parameter $\lambda > 0$ and shape parameter $k > 0$ and contains the exponential distribution as a special case; i.e., when shape $k = 1$. As a result, this model is also more flexible than the exponential. We have the following table for this distribution:

p.d.f. $f(t)$	survivor $S(t)$	hazard $h(t)$
$\frac{\lambda^k t^{k-1}}{\Gamma(k)} \exp(-\lambda t)$	no simple form	no simple form
mean $E(T)$	variance $Var(T)$	
$\frac{k}{\lambda}$	$\frac{k}{\lambda^2}$	

The hazard function is monotone increasing from 0 when $k > 1$, monotone decreasing from ∞ if $k < 1$, and in either case approaches λ as t increases.

The model for $Y = \log(T)$ can be written $Y = \mu + Z$, where Z has density

$$\frac{\exp\left(kz - \exp(z)\right)}{\Gamma(k)}\,. \tag{3.6}$$

The r.v. Y is called a log-gamma r.v. with parameters k and $\mu = -\log(\lambda)$. The quantity Z has a negatively skewed distribution with skewness decreasing with k increasing. When $k = 1$, this is the exponential model and, hence, Z has the standard extreme value distribution. With the exception of $k = 1$, the

log-gamma is not a member of the location and scale family of distributions. It is, however, a member of the location family. Figure 3.6 shows some gamma p.d.f.'s and hazards. We display some log-gamma p.d.f.'s in Figure 3.7. See Klein & Moeschberger (1997, page 44) and Kalbfleisch & Prentice (1980, page 27) for a discussion of the generalized gamma and corresponding generalized log-gamma distributions.

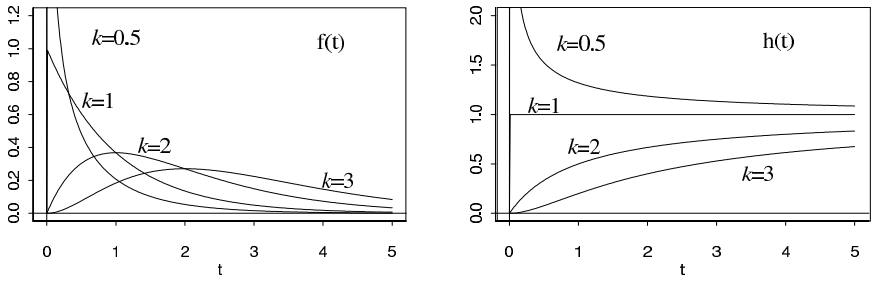


Figure 3.6 *Gamma densities and hazards with* $\lambda = 1$ *and* $k = 0.5, 1, 2,$ *and* 3.

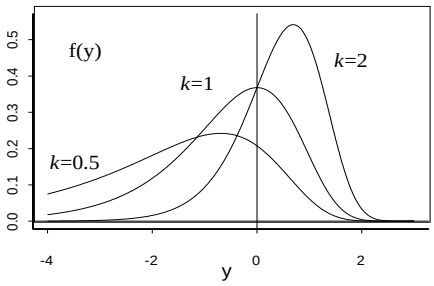


Figure 3.7 *Log-gamma density with* $k = 0.5, 1, 2,$ *and* $\lambda = 1$.

Summary

Except for the gamma distribution, all distributions of lifetime T we work with have the property that the distribution of the log-transform $\log(T)$ is a member of the location and scale family of distributions. The common features are:

- The time T distributions have two parameters −

$$\text{scale} = \lambda \quad \text{and} \quad \text{shape} = \alpha\,.$$

- In log-time, $Y = \log(T)$, the distributions have two parameters −

$$\text{location} = \mu = -\log(\lambda) \quad \text{and} \quad \text{scale} = \sigma = \frac{1}{\alpha}\,.$$

- Each can be expressed in the form

$$Y = \log(T) = \mu + \sigma Z\,, \tag{3.7}$$

  where Z is the standard member; that is,

$$\mu = 0 \ (\lambda = 1) \quad \text{and} \quad \sigma = 1 \ (\alpha = 1)\,.$$

- They are log-linear models.

The three distributions considered in our examples are summarized as follows:

T	$\Longleftrightarrow$	$Y = \log(T)$
Weibull	$\Longleftrightarrow$	extreme minimum value
log-normal	$\Longleftrightarrow$	normal
log-logistic	$\Longleftrightarrow$	logistic

If the true distribution of $Y = \log(T)$ is one of the above, then the pth-quantile y_p is a linear function of z_p, the pth-quantile of the standard member of the specified distribution. The straight line has slope σ and y-intercept μ. Let t_p denote an arbitrary pth-quantile. In light of the foregoing discussion, the linear relationships for $y_p = \log(t_p)$ reported in expressions (3.3), (3.5), (3.7) take on new meaning. This is summarized in Table 3.1.

Table 3.1: *Relationships to exploit to construct a graphical check for model adequacy*

t_p quantile	$y_p = \log(t_p)$ quantile	form of standard quantile z_p
Weibull	extreme value	$\log(-\log(S(t_p))) = \log(H(t_p))$ $= \log(-\log(1-p))$
log-normal	normal	$\Phi^{-1}(p)$, where Φ denotes the standard normal d.f.
log-logistic	logistic	$-\log\left(\dfrac{S(t_p)}{1-S(t_p)}\right) = -\log(\text{odds})$ $= -\log\left(\frac{1-p}{p}\right)$

Construction of the quantile-quantile (Q-Q) plot

Let $\widehat{S}(t)$ denote the K-M estimator of survival probability beyond time t. Let $t_i, i = 1, \ldots, r \leq n$, denote the ordered uncensored observed failure times. For each uncensored sample quantile $y_i = \log(t_i)$, the estimated failure probability is $\hat{p}_i = 1 - \widehat{S}(t_i)$. The parametric standard quantile z_i is obtained by using the $\hat{p}_i$ to evaluate the expression for the standard quantile given in Table 3.1.

Thus, $F_{0,1}(z_i) = P(Z \leq z_i) = \hat{p}_i$, where $F_{0,1}$ is the d.f. of the standard parametric model ($\mu = 0, \sigma = 1$) under consideration. As the K-M estimator is distribution free and consistently estimates the "true" survival function, for large sample sizes n, the z_i should reflect the "true" standard quantiles, if F is indeed the "true" lifetime d.f.. Hence, if the proposed model fits the data adequately, the points (z_i, y_i) should lie close to a straight line with slope σ and y-intercept μ. **The plot of the points $(\mathbf{z_i}, \mathbf{y_i})$ is called a quantile-quantile (Q-Q) plot**. An appropriate line to compare the plot pattern to is $\widehat{y}_p = \hat{\mu} + \widehat{\sigma} z_p$ (3.7), where $\widehat{\mu}$ and $\widehat{\sigma}$ denote the maximum likelihood estimates to be discussed in the next section. Plot patterns grossly different from this straight line indicate the proposed model is inadequate. The more closely the plot pattern follows this line, the more evidence there is in support of the proposed model. The Q-Q plot is a major diagnostic tool for checking model adequacy.

A cautionary note: Fitting the uncensored points (z_i, y_i) to a least squares line alone can be very misleading in deeming model adequacy. Our first example of this is discussed in Section 3.4, where we first construct Q-Q plots to check and compare the adequacy of fitting the AML data to the exponential, Weibull, and log-logistic distributions.

Equivalently, we can plot the points (z_i, e_i) where the e_i is the ith ordered residual

$$e_i = \frac{y_i - \widehat{\mu}}{\widehat{\sigma}}$$

and z_i is the corresponding log-parametric standard quantile of either the Weibull, log-normal, or log-logistic distribution. If the model under study is appropriate, the points (z_i, e_i) should lie very close to the 45^o-line through the origin.

3.2 Maximum likelihood estimation (MLE)

Our assumptions here are that the $T_1, \ldots, T_n$ are iid from a continuous distribution with p.d.f. $f(t|\theta)$, where θ belongs to some parameter space Ω. Here, θ could be either a real-valued or vector-valued parameter. The **likelihood function** is the joint p.d.f. of the sample when regarded as a function of θ for a given value $(t_1, \ldots, t_n)$. To emphasize this we denote it by $L(\theta)$. For a random sample, this is the product of the p.d.f.'s. That is, the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f(t_i|\theta).$$

The **maximum likelihood estimator** (MLE), denoted by $\widehat{\theta}$, is the value of θ in Ω that maximizes $L(\theta)$ or, equivalently, maximizes the log-likelihood

$$\log L(\theta) = \sum_{i=1}^{n} \log f(t_i|\theta).$$

MLE's possess the *invariance property*; that is, the MLE of a function of θ, say $\tau(\theta)$, is $\tau(\widehat{\theta})$. For a gentle introduction to these foregoing notions, see DeGroot (1986). Under the random censoring model, we see from expression (1.13) that if we assume that the censoring time has no connection to the survival time, then the log-likelihood for the maximization process can be taken to be

$$\log L(\theta) = \log \prod_{i=1}^{n} \Big(f(y_i|\theta)\Big)^{\delta_i} \cdot \Big(S_f(y_i|\theta)\Big)^{1-\delta_i} = \sum_{u} \log f(y_i|\theta) + \sum_{c} \log S_f(y_i|\theta), \tag{3.8}$$

where u and c mean sums over the uncensored and censored observations, respectively. Let $I(\theta)$ denote the **Fisher information matrix**. Then its elements are

$$I(\theta) = \left(\left(-E(\frac{\partial^2}{\partial\theta_j \partial\theta_k} \log L(\theta))\right)\right),$$

where E denotes expectation. As we are working with random samples (iid) we point out that $I(\theta)$ can be expressed as

$$I(\theta) = nI_1(\theta),$$

where $I_1(\theta) = \left(\left(-E(\frac{\partial^2}{\partial\theta_j \partial\theta_k} \log f(y_1|\theta))\right)\right)$ is the Fisher information matrix of any one of the observations.

The MLE $\widehat{\theta}$ has the following **large sample distribution**:

$$\widehat{\theta} \overset{a}{\sim} \mathbf{MVN}(\theta, I^{-1}(\theta)), \tag{3.9}$$

where MVN denotes multivariate normal and $\overset{a}{\sim}$ is read "is asymptotically distributed." The asymptotic covariance matrix $I^{-1}(\theta)$ is a $d \times d$ matrix, where d is the dimension of θ. The ith diagonal element of $I^{-1}(\theta)$ is the asymptotic variance of the ith component of $\hat{\theta}$. The off-diagonal elements are the asymptotic covariances of the corresponding components of $\hat{\theta}$. If θ is a scalar (real valued), then the asymptotic variance, denoted var_a, of $\hat{\theta}$ is

$$\text{var}_a(\hat{\theta}) = \frac{1}{I(\theta)},$$

where $I(\theta) = -E\Big(\partial^2 \log L(\theta)/\partial\theta^2\Big)$. For censored data, this expectation is a function of the censoring distribution G as well as the survival time distribution F. Hence, it is necessary to approximate $I(\theta)$ by the **observed information matrix** $i(\theta)$ evaluated at the MLE $\hat{\theta}$, where

$$i(\theta) = \left(\left(-\frac{\partial^2}{\partial\theta_j \partial\theta_k} \log L(\theta)\right)\right). \tag{3.10}$$

For the univariate case,

$$i(\theta) = -\frac{\partial^2 \log L(\theta)}{\partial \theta^2}. \tag{3.11}$$

Hence, $\text{var}_a(\hat{\theta})$ is approximated by $\left(i(\hat{\theta})\right)^{-1}$.

The **delta method** is useful for obtaining limiting distributions of smooth functions of an MLE. When variance of an estimator includes the parameter of interest, the delta method can be used to remove the parameter in the variance. This is called the variance-stabilization. We describe it for the univariate case.

Delta method:

Suppose a random variable Z has a mean μ and variance σ^2 and suppose we want to approximate the distribution of some function $g(Z)$. Take a first order Taylor expansion of $g(Z)$ about μ and ignore the higher order terms to get

$$g(Z) \approx g(\mu) + (Z - \mu)g'(\mu).$$

Then the $\text{mean}(g(Z)) \approx g(\mu)$ and the $\text{var}(g(Z)) \approx (g'(\mu))^2 \sigma^2$. Furthermore, if

$$Z \overset{a}{\sim} \text{normal}(\mu, \sigma^2),$$

then

$$g(Z) \overset{a}{\sim} \text{normal}(g(\mu), (g'(\mu))^2 \sigma^2). \tag{3.12}$$

Example: Let $X_1, \ldots, X_n$ be iid from a Poisson distribution with mean λ. Then the MLE of λ is $\widehat{\lambda} = \overline{X}$. We know that the mean and variance of $Z = \overline{X}$ are λ and λ/n, respectively. Take $g(Z) = \overline{X}^{\frac{1}{2}}$. Then $g(\lambda) = \lambda^{\frac{1}{2}}$ and

$$\overline{X}^{\frac{1}{2}} \overset{a}{\sim} \text{normal with mean} \approx \lambda^{\frac{1}{2}} \text{ and variance} \approx \frac{1}{4n}.$$

There are multivariate versions of the delta method. One is stated in Section 3.6.

3.3 Confidence intervals and tests

For some estimators we can compute their small sample exact distributions. However, for most, in particular when censoring is involved, we must rely on the large sample properties of the MLE's. For confidence intervals or for testing $\text{H}_0 : \theta = \theta_0$, where θ is a scalar or a scalar component of a vector, we can construct the asymptotic z-intervals with the standard errors (s.e.) taken from the diagonal of the asymptotic covariance matrix which is the inverse of the information matrix $I(\theta)$ evaluated at the MLE $\widehat{\theta}$ if necessary. The s.e.'s are, of course, the square roots of these diagonal values. In summary:

An approximate $(1-\alpha)\times 100\%$ confidence interval for the parameter θ is given by

$$\hat{\theta} \pm z_{\frac{\alpha}{2}} \text{s.e.}(\hat{\theta}), \tag{3.13}$$

where $z_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution and, by (3.11), s.e. is the square root of $\text{var}_a(\hat{\theta}) \approx -\left(\partial^2 \text{log}L(\theta)/\partial\theta^2\right)^{-1} = \left(i(\widehat{\theta})\right)^{-1}$.

However, if we are performing joint estimation or testing a vector-valued θ, we have three well known procedures: Assume θ_0 has d-components, $d \geq 1$. Unless otherwise declared, $\widehat{\theta}$ denotes the MLE.

- The **Wald** statistic:

$$(\widehat{\theta} - \theta_0)' I(\theta_0)(\widehat{\theta} - \theta_0) \overset{a}{\sim} \chi^2_{(d)} \text{ under } \text{H}_0.$$

- The **Rao** statistic:

$$\frac{\partial}{\partial\theta}\text{log}L(\theta_0)' I^{-1}(\theta_0) \frac{\partial}{\partial\theta}\text{log}L(\theta_0) \overset{a}{\sim} \chi^2_{(d)} \text{ under } \text{H}_0.$$

  Note that Rao's method does not use the MLE. Hence, no iterative calculation is necessary.

- The Neyman-Pearson/Wilks **likelihood ratio test (LRT)**:
  Let the vector $\underline{t}$ represent the n observed values; that is, $\underline{t}' = (t_1, \ldots, t_n)$. The LRT statistic is given by

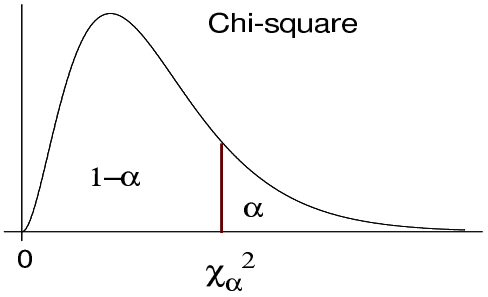
$$r^*(\underline{t}) = -2 \log\left(\frac{L(\theta_0)}{L(\widehat{\theta})}\right) \overset{a}{\sim} \chi^2_{(d)} \text{ under } \text{H}_0. \tag{3.14}$$

  To test $\text{H}_0 : \theta = \theta_0$ against $\text{H}_A : \theta \neq \theta_0$, we reject for small values of $L(\theta_0)/L(\widehat{\theta})$ (as this ratio is less than or equal to 1). Equivalently, we reject for large values of $r^*(\underline{t})$.

For **joint confidence regions** we simply take the region of values that satisfy the elliptical region formed with either the Wald or Rao statistic with $I(\theta)$ or $i(\theta)$ evaluated at the MLE $\widehat{\theta}$. For example, an approximate $(1-\alpha)\times 100\%$ joint confidence region for θ is given by

$$\{\theta; \text{Wald} \leq \chi^2_\alpha\},$$

where χ^2_α is the chi-square upper αth-quantile with d degrees of freedom. The following picture explains:



3.4 One-sample problem

3.4.1 Fitting data to the exponential model

Let u, c, and n_u denote uncensored, censored, and number of uncensored observations, respectively. The n observed values are now represented by the vectors $\underline{y}$ and $\underline{\delta}$, where $\underline{y}' = (y_1, \ldots, y_n)$ and $\underline{\delta}' = (\delta_1, \ldots, \delta_n)$. Then

- **Likelihood:** See expressions (1.13), (3.8).

$$\begin{aligned} L(\lambda) &= \prod_u f(y_i|\lambda) \cdot \prod_c S_f(y_i|\lambda) \\ &= \prod_u \lambda \exp(-\lambda y_i) \prod_c \exp(-\lambda y_i) \\ &= \lambda^{n_u} \exp\Big(-\lambda \sum_u y_i\Big) \exp\Big(-\lambda \sum_c y_i\Big) \\ &= \lambda^{n_u} \exp\Big(-\lambda \sum_{i=1}^{n} y_i\Big) \end{aligned}$$

- **Log-likelihood:**

$$\log L(\lambda) = n_u \log(\lambda) - \lambda \sum_{i=1}^{n} y_i$$

The MLE $\widehat{\lambda}$ solves $\dfrac{\partial \log L(\lambda)}{\partial \lambda} = \dfrac{n_u}{\lambda} - \sum_{i=1}^{n} y_i = 0.$

$$\frac{\partial^2 \log L(\lambda)}{\partial \lambda^2} = -\frac{n_u}{\lambda^2} = -i(\lambda), \text{the negative of the observed information.}$$

- **MLE:**

$$\widehat{\lambda} = \frac{n_u}{\sum_{i=1}^{n} y_i} \qquad \text{and} \qquad \text{var}_a(\widehat{\lambda}) = \Big(-E\Big(-\frac{n_u}{\lambda^2}\Big)\Big)^{-1} = \frac{\lambda^2}{E(n_u)},$$

where $E(n_u) = n \cdot P(T \leq C)$. From expression (3.9),

$$\frac{\widehat{\lambda} - \lambda}{\sqrt{\lambda^2 / E(n_u)}} \overset{a}{\sim} N(0,1).$$

We replace $E(n_u)$ by n_u since we don't usually know the censoring distribution $G(\cdot)$. Notice the dependence of the asymptotic variance on the unknown parameter λ. We substitute in $\hat{\lambda}$ and obtain

$$\text{var}_a(\hat{\lambda}) \approx \frac{\hat{\lambda}^2}{n_u} = \frac{1}{i(\hat{\lambda})},$$

where $i(\lambda)$ is just above. The MLE for the mean $\theta = 1/\lambda$ is simply $\hat{\theta} = 1/\hat{\lambda} = \sum_{i=1}^{n} y_i / n_u$.

On the AML data, $n_u = 7$,

$$\widehat{\lambda} = \frac{7}{423} = 0.0165, \qquad \text{and} \qquad \text{var}_a(\widehat{\lambda}) \approx \frac{\hat{\lambda}^2}{7} = \frac{0.0165^2}{7}.$$

- **A 95% C.I. for $\boldsymbol{\lambda}$** (3.13) is given by

$$\widehat{\lambda} \pm z_{0.025} \cdot \text{se}(\widehat{\lambda}) =: \ 0.0165 \pm 1.96 \cdot \frac{0.0165}{\sqrt{7}} =: \ [0.004277 \ , \ 0.0287].$$

- **A 95% C.I. for $\boldsymbol{\theta}$**, the mean survival, can be obtained by inverting the previous interval for λ. This interval is: [34.8, 233.808] weeks. Both intervals are very skewed. However, as $\widehat{\theta} = 1/\widehat{\lambda} = 60.42856$ weeks, we have $\theta = g(\lambda) = 1/\lambda$ and we can use the delta method to obtain the asymptotic variance of $\widehat{\theta}$. As $g'(\lambda) = -\lambda^{-2}$, the asymptotic variance is

$$\text{var}_a(\widehat{\theta}) = \frac{1}{\lambda^2 E(n_u)} \approx \frac{1}{\widehat{\lambda}^2 \cdot n_u} = \frac{\hat{\theta}^2}{n_u}. \tag{3.15}$$

Hence, a second 95% C.I. for θ, the mean survival, is given by

$$\widehat{\theta} \pm z_{0.025} \text{se}(\widehat{\theta}) =: \ 60.42856 \pm 1.96 \cdot \frac{1}{0.0165 \cdot \sqrt{7}} =: \ [15.66246, 105.1947] \text{ weeks.}$$

Notice this is still skewed, but much less so; and it is much narrower. Here we use the asymptotic variance of $\widehat{\theta}$ directly, and hence, eliminate one source of variation. However, the asymptotic variance still depends on λ.

- **The MLE of the pth-quantile:**

$$\widehat{t_p} = -\frac{1}{\widehat{\lambda}} \log(1-p) = -\frac{\sum_{i=1}^{n} y_i}{n_u} \log(1-p).$$

Thus, the MLE of the median is

$$\widehat{\text{med}} = -\frac{423}{7} \log(0.5) = 41.88 \ \text{ weeks.}$$

Notice how much smaller the median is compared to the estimate $\widehat{\theta} =$

60.43. The median reflects a more typical survival time. The mean is greatly influenced by the one large value 161+. Note that

$$\mathrm{var}_a(\widehat{t_p}) = \Big(\log(1-p)\Big)^2 \cdot \mathrm{var}_a\Big(\widehat{\lambda}^{-1}\Big) \approx \Big(\log(1-p)\Big)^2 \cdot \frac{1}{\hat{\lambda}^2 \cdot n_u}\,.$$

The $\mathrm{var}_a\Big(\widehat{\lambda}^{-1}\Big)$ is given in expression (3.15). Thus, **a 95% C.I. for the median** is given by

$$\widehat{t}_{0.5} \pm 1.96 \cdot \frac{-\log(0.5)}{\widehat{\lambda} \cdot \sqrt{n_u}} =: \quad 41.88 \pm 1.96 \cdot \frac{-\log(0.5)}{0.0165 \cdot \sqrt{7}} =: \quad [10.76, 73] \ \text{ weeks.}$$

- With the delta method (3.12) we can construct intervals that are less skewed and possibly narrower by finding transformations which eliminate the dependence of the asymptotic variance on the unknown parameter of interest. For example, the natural log-transform of $\widehat{\lambda}$ accomplishes this. This is because for $g(\lambda) = \log(\lambda)$, $g'(\lambda) = 1/\lambda$ and, thus, $\mathrm{var}_a(\log(\widehat{\lambda})) = \lambda^{-2}\{\lambda^2/E(n_u)\} = 1/E(n_u)$. Again we replace $E(n_u)$ by n_u. Therefore, we have

$$\log(\widehat{\lambda}) \quad \overset{a}{\sim} \quad N\left(\log(\lambda), \frac{1}{n_u}\right). \tag{3.16}$$

A 95% C.I. for $\log(\lambda)$ is given by

$$\log(\widehat{\lambda}) \pm 1.96 \cdot \frac{1}{\sqrt{n_u}}$$

$$\log\left(\frac{7}{423}\right) \pm 1.96 \cdot \frac{1}{\sqrt{7}}$$

$$[-4.84, -3.36].$$

Transform back by taking exp(endpoints). This second 95% C.I. for λ is

$$[.0079, .0347],$$

which is slightly wider than the previous interval for λ. Invert and reverse endpoints to obtain a third 95% C.I. for the mean θ. This yields $[28.81, 126.76]$ weeks, which is also slightly wider than the second interval for θ.

Analogously, since $\mathrm{var}_a(\widehat{\theta}) \approx \hat{\theta}^2/n_u$ (3.15), the delta method provides large sample distributions for $\log(\widehat{\theta})$ and $\log(\widehat{t_p})$ with the same variance, which is free of the parameter θ. They are

$$\log(\widehat{\theta}) \quad \overset{a}{\sim} \quad N\left(\log(\theta), \frac{1}{n_u}\right) \tag{3.17}$$

$$\log(\widehat{t_p}) \quad \overset{a}{\sim} \quad N\left(\log(t_p), \frac{1}{n_u}\right). \tag{3.18}$$

Analogously, we first construct C.I.'s for the log(parameter), then take exp(endpoints) to obtain C.I.'s for the parameter. Most statisticians prefer

this approach. Using the AML data, we summarize 95% C.I.'s in Table 3.2.

Table 3.2: *Preferred 95% confidence intervals for mean and median (or any quantile) of an exponential survival model based on the log-transform*

parameter	point estimate	log(parameter)	parameter	
mean	60.43 weeks	[3.361, 4.84]	[28.81, 126.76]	weeks
median	41.88 weeks	[2.994, 4.4756]	[19.965, 87.85]	weeks

- **The MLE of the survivor function $S(t) = \exp(-\lambda t)$**:

$$\widehat{S}(t) = \exp(-\widehat{\lambda}t) = \exp(-0.0165\,t).$$

For any fixed t, $\widehat{S}(t)$ is a function of $\widehat{\lambda}$. We can get its approximate distribution by using the delta method. Alternatively, we can take a log-log transformation that usually improves the convergence to normality. This is because the var_a is free of the unknown parameter λ. This follows from (3.16) and the relationship

$$\log\Big(-\log(\widehat{S}(t))\Big) = \log(\widehat{\lambda}) + \log(t)\ .$$

Hence,

$$\text{var}_a\Big\{\log\Big(-\log(\widehat{S}(t))\Big)\Big\} = \text{var}_a\Big(\log(\widehat{\lambda})\Big) \approx \frac{1}{n_u}\ .$$

It follows from the delta method that for each fixed t,

$$\log\Big(-\log(\widehat{S}(t))\Big) \overset{a}{\sim} N\left(\log(-\log(S(t))) = \log(\lambda t),\ \frac{1}{n_u}\right).$$

It then follows, with some algebraic manipulation, a **$(1-\alpha)\times 100\%$ C.I. for the true probability of survival beyond time t, $S(t)$**, is given by

$$\exp\left\{\log\Big(\widehat{S}(t)\Big)\exp\left(\frac{z_{\alpha/2}}{\sqrt{n_u}}\right)\right\} \le S(t) \le \exp\left\{\log\Big(\widehat{S}(t)\Big)\exp\left(\frac{-z_{\alpha/2}}{\sqrt{n_u}}\right)\right\}. \tag{3.19}$$

WHY!

- **The likelihood ratio test** (3.14):

$$\begin{aligned} r^*(\underline{y}) &= -2\log L(\lambda_0) + 2\log L(\widehat{\lambda}) \\ &= -2n_u\log(\lambda_0) + 2\lambda_0\sum_{i=1}^{n} y_i + 2n_u\log\Big(\frac{n_u}{\sum_{i=1}^{n} y_i}\Big) - 2n_u \\ &= -2\cdot 7\cdot\log(\frac{1}{30}) + \frac{2}{30}\cdot 423 + 2\cdot 7\cdot\log\Big(\frac{7}{423}\Big) - 2\cdot 7 \\ &= 4.396. \end{aligned}$$

The p-value $= P(r^*(\underline{y}) \geq 4.396) \approx 0.036$. Therefore, here we reject $H_0: \theta = 1/\lambda = 30$ and conclude that mean survival is > 30 weeks.

A computer application:

We use the S function `survReg` to fit parametric models (with the MLE approach) for censored data. The following S program is intended to duplicate some of the previous hand calculations. It fits an exponential model to the AML data, yields point and 95% C.I. estimates for both the mean and the median, and provides a Q-Q plot for diagnostic purposes. Recall that the exponential model is just a Weibull with shape $\alpha = 1$ or, in log(time), is an extreme value model with scale $\sigma = 1$. The function `survReg` fits log(time) and outputs the coefficient $\widehat{\mu} = -\log(\widehat{\lambda})$, the MLE of μ, the location parameter of the extreme value distribution. Hence, the MLE(λ)=$\widehat{\lambda} = \exp(-\widehat{\mu})$ and the MLE(θ)= $\widehat{\theta} = \exp(\widehat{\mu})$. Unnecessary output has been deleted. The S function `predict` is a companion function to `survReg`. It provides estimates of quantiles along with their s.e.'s. One of the arguments of the `predict` function is `type`. Set `type="uquantile"` to produce estimates based on the log-transform as in Table 3.2. The default produces intervals based on the variance for quantiles derived on page 62. The function `qq.weibull` produces a Q-Q plot. The pound sign # denotes our inserted annotation. We store the data for the maintained group in a `data.frame` object called aml1. The two variables are `weeks` and `status`.

 # Exponential fit

```
> attach(aml1)
> exp.fit <- survReg(Surv(weeks,status)~1,dist="weib",scale=1)
> exp.fit
Coefficients:
 (Intercept)
    4.101457
Scale fixed at 1 Loglik(model)= -35.7 n= 11
```

The `Intercept` $= 4.1014$, which equals $\widehat{\mu} = -\log(\widehat{\lambda}) = \log(\widehat{\theta})$. The next five line commands produce a 95% C.I. for the mean θ.

```
> coeff <- exp.fit$coeff # muhat
> var <- exp.fit$var
> thetahat <- exp(coeff) # exp(muhat)
> thetahat
     60.42828
> C.I.mean1 <- c(thetahat,exp(coeff-1.96*sqrt(var)),
                 exp(coeff+1.96*sqrt(var)))
> names(C.I.mean1) <- c("mean1","LCL","UCL")
> C.I.mean1
```

```
     mean1       LCL       UCL
  60.42828  28.80787  126.7562
```

Estimated median along with a 95% C.I. (in weeks) using the `predict` function.

```
> medhat <- predict(exp.fit,type="uquantile",p=0.5,se.fit=T)
> medhat1 <- medhat$fit[1]
> medhat1.se <- medhat$se.fit[1]
> exp(medhat1)
[1] 41.88569

> 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
  41.88569  19.96809  87.86072
```

Point and 95% C.I. estimates for $S(t)$, the probability of survival beyond time t, at the uncensored maintained group's survival times.

```
> muhat <- exp.fit$coeff
> weeks.u <- weeks[status == 1]
> nu <- length(weeks.u)
> scalehat <- rep(exp(muhat),nu)
> Shat <- 1 - pweibull(weeks.u,1,scalehat)
  # In S, Weibull's scale argument is exp(muhat) = 1/lambdahat,
    # which we call scalehat.
> LCL <- exp(log(Shat)*exp(1.96/sqrt(nu)))#See expression (3.19)
> UCL <- exp(log(Shat)*exp(-1.96/sqrt(nu)))
> C.I.Shat <- data.frame(weeks.u,Shat,LCL,UCL)
> round(C.I.Shat,5)
   weeks.u     Shat      LCL      UCL  # 95% C.I.'s
 1       9  0.86162  0.73168  0.93146
 2      13  0.80644  0.63682  0.90253
 4      18  0.74240  0.53535  0.86762
 5      23  0.68344  0.45005  0.83406
 7      31  0.59869  0.34092  0.78305
 8      34  0.56970  0.30721  0.76473
10      48  0.45188  0.18896  0.68477
```

The next line command produces the Q-Q plot in Figure 3.8 using the `qq.weibull` function. The scale=1 argument forces an exponential to be fit.

```
> qq.weibull(Surv(weeks,status),scale=1)
[1] "qq.weibull:done"
```

The following table summarizes the estimates of the mean and the median.

Exponential fit with MLE to AML1 data		
	Point Estimate	95% C.I.
median1	41.88569	[19.968, 87.86] weeks
mean1	60.42828	[28.81, 126.76] weeks

This table's results match those in Table 3.2. In Figure 3.8 a Q-Q plot is displayed. The following S program performs a likelihood ratio test (LRT) of

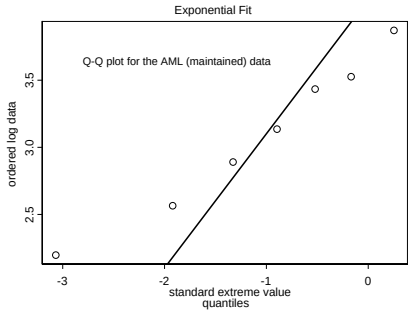


Figure 3.8 *Exponential Q-Q plot. The line has MLE intercept* $\widehat{\mu}$ *and slope 1.*

the null hypothesis $H_0 : \theta = 1/\lambda = 30$ weeks. To compute the value of the log likelihood function $L(\theta)$ at $\theta = 30$, we use the function `weib.loglik.theta`. It has four arguments: `time`, `status`, `shape`, `theta`. A shape value (α) of 1 forces it to fit an exponential and theta is set to $1/\lambda = 30$. The results match those hand-calculated back on page 63.

```
> weib.loglik.theta(weeks,status,1,30)
[1] -37.90838
> rstar <- - 2*(weib.loglik.theta(weeks,status,1,30) -
                exp.fit$loglik[1])
> rstar
[1] 4.396295
> pvalue <- 1 - pchisq(rstar,1)
> pvalue
[1] 0.0360171
```

3.4.2 Fitting data to the Weibull and log-logistic models

The following S program fits the AML data to the Weibull and log-logistic models both using the MLE approach via the `survReg` function. The `survReg`

function uses by default a log link function which transforms the problem into estimating location $\mu = -\log(\lambda)$ and scale $\sigma = 1/\alpha$. In the output from `> summary(weib.fit)`,

$\widehat{\mu}$ (= `Intercept`) `<- weib.fit$coeff`, and $\widehat{\sigma}$ (= `Scale`) `<-weib.fit$scale`.

This holds for any `summary(fit)` resulting from `survReg` evaluated at the `"Weibull"`, `"loglogistic"`, and `"lognormal"` distributions. The S output has been modified in that the extraneous output has been deleted.

Once the parameters are estimated via `survReg`, we can use S functions to compute estimated survival probabilities and quantiles. These functions are given in Table 3.3 for the reader's convenience.

Table 3.3: *S distribution functions*

	Weibull	logistic ($Y = \log(T)$)	normal ($Y = \log(T)$)
$F(t)$	`pweibull`(q, α, λ^{-1})	`plogis`(q, μ, σ)	`pnorm`(q, μ, σ)
t_p	`qweibull`(p, α, λ^{-1})	`qlogis`(p, μ, σ)	`qnorm`(p, μ, σ)

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

 # Log-logistic fit

```
> loglogis.fit<-survReg(Surv(weeks,status)~1,dist="loglogistic")
> summary(loglogis.fit)
             Value Std. Error      z         p
(Intercept)  3.515      0.306  11.48 1.65e-030
 Log(scale) -0.612      0.318  -1.93 5.39e-002
Scale= 0.542
```

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

```
> medhat <- predict(loglogis.fit,type="uquantile",p=0.5,se.fit=T)
> medhat1 <- medhat$fit[1]
> medhat1.se <- medhat$se.fit[1]
> exp(medhat1)
[1] 33.60127
> 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
  33.60127  18.44077  61.22549
> qq.loglogistic(Surv(weeks,status)) # Produces a Q-Q plot.
[1] "qq.loglogistic:done"
> detach()
```

Discussion

In order to compare some of the output readily, we provide a summary in the following table:

MLE's fit to AML1 data at the models:

model	$\widehat{\mu}$	median1	95% C.I.	$\widehat{\sigma}$
exponential	4.1	41.88	[19.97, 87.86] weeks	1
Weibull	4.1	42.29	[20.22, 88.44] weeks	.969
log-logistic	3.52	33.60	[18.44, 61.23] weeks	.542

The log-logistic gives the narrowest C.I. among the three. Further, its estimated median of 33.60 weeks is the smallest and very close to the K-M estimated median of 31 weeks on page 32. The Q-Q plots in Figure 3.10 are useful for distributional assessment. It "appears" that a log-logistic model fits adequately and is the best among the three distributions discussed. The estimated log-logistic survival curve is overlayed on the K-M curve for the AML1 data in Figure 3.9. We could also consider a log-normal model here. The cautionary note, page 56, warns that we must compare the plot pattern to the MLE line with slope $\widehat{\sigma}$ and y-intercept $\widehat{\mu}$. For without this comparison,

the least squares line alone fitted only to uncensored times would lead us to judge the Weibull survival model adequate. But, as we see in Figure 3.10, this is wrong. We do see that the least squares line in the Q-Q plot for the log-logistic fit is much closer to the MLE line with slope $\widehat{\sigma}$ and y-intercept $\widehat{\mu}$.

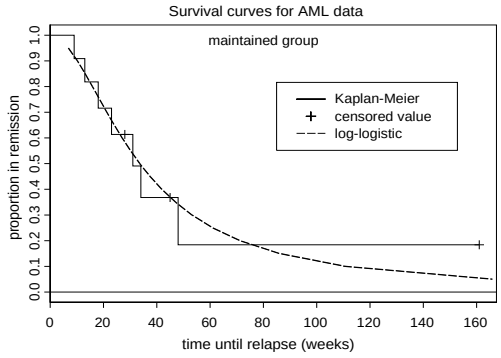


Figure 3.9 *K-M and log-logistic survival curves for AML data.*

Figure 3.11 displays Q-Q plots of (z_i, e_i). We delay the description of the function `qq.reg.resid.r`, which draws the plot, until page 125, where we discuss checking the adequacy of a regression model. Some R code for the Q-Q plot follows:

```
> fit.lognorm <- survreg(Surv(weeks,status)~1,dist="lognormal",
                         data=aml1)
> qq.reg.resid.r(aml1,aml1$weeks,aml1$status,fit.lognorm,"qnorm",
              "standard normal quantile")
```

3.5 Two-sample problem

In this section we compare two survival curves from the same parametric family. We focus on comparing the two scale (λ) parameters. In the log-transformed problem, this compares the two location, $\mu = -\log(\lambda)$, parameters. We picture this in Figure 3.12. We continue to work with the AML data. The nonparametric log-rank test (page 40) detected a significant difference (p-value= 0.03265) between the two K-M survival curves for the two groups, maintained and nonmaintained. We concluded maintenance chemotherapy prolongs remission period. We now explore if any of the log-transform distributions, which belong to the location and scale family (3.7), fit this data

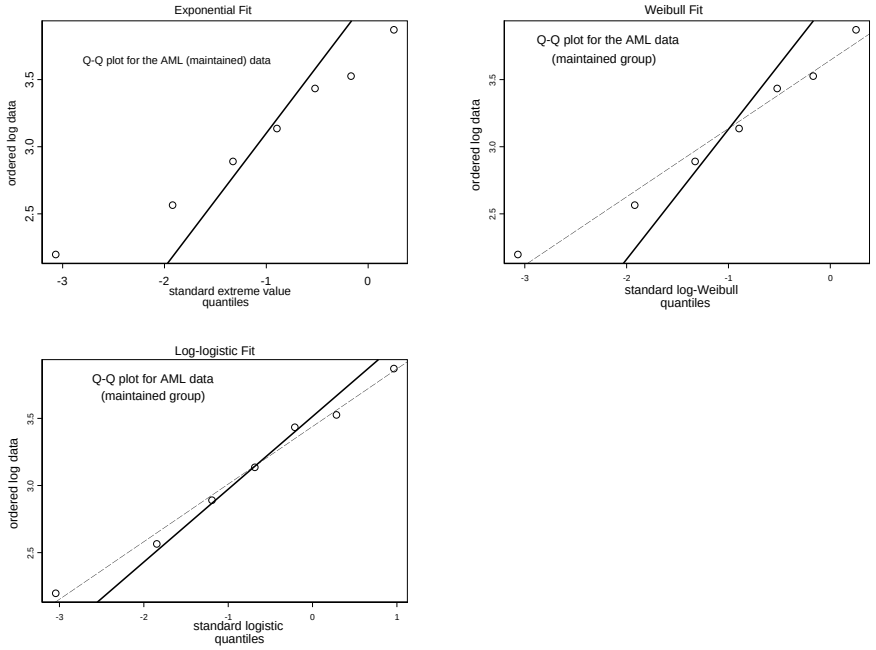


Figure 3.10 *Q-Q plots for the exponential, the Weibull, and the log-logistic. Each solid line is constructed with MLE's* $\widehat{\mu}$ *and* $\widehat{\sigma}$*. The dashed lines are least squares lines.*

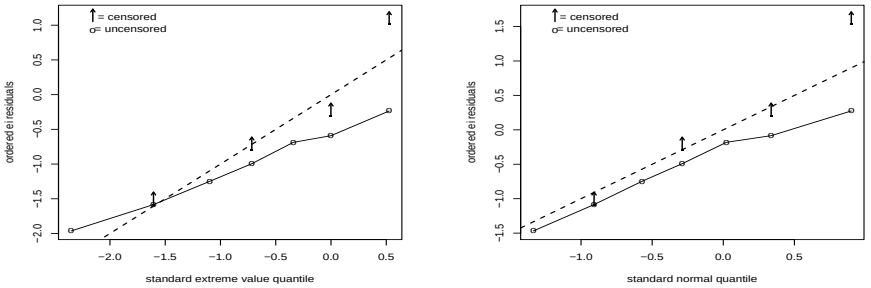


Figure 3.11 *Q-Q plots of the ordered residuals* $e_i = (y_i - \hat{\mu})/\hat{\sigma}$ *where* y_i *denotes the log-data. Dashed line is the* 45^o*-line through the origin.*

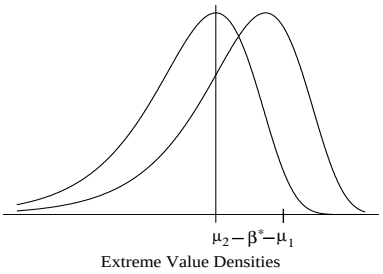


Figure 3.12 *Comparison of two locations.*

adequately. The **full model** can be expressed as a log-linear model as follows:

$$\begin{aligned} Y &= \log(T) \\ &= \widetilde{\mu} + \text{error} \\ &= \theta + \beta^* \text{group} + \text{error} \\ &= \begin{cases} \theta + \beta^* + \text{error} & \text{if group} = 1 \text{ (maintained)} \\ \theta + \text{error} & \text{if group} = 0 \text{ (nonmaintained).} \end{cases} \end{aligned}$$

The $\widetilde{\mu}$ is called the *linear predictor*. In this two groups model, it has two values $\mu_1 = \theta + \beta^*$ and $\mu_2 = \theta$. Further, we know $\widetilde{\mu} = -\log(\widetilde{\lambda})$, where $\widetilde{\lambda}$ denotes the scale parameter values of the distribution of the target variable T. Then $\widetilde{\lambda} = \exp(-\theta - \beta^* \text{group})$. The two values are $\lambda_1 = \exp(-\theta - \beta^*)$ and $\lambda_2 = \exp(-\theta)$. The **null hypothesis** is:

$$H_0 : \lambda_1 = \lambda_2 \quad \text{if and only if} \quad \mu_1 = \mu_2 \quad \text{if and only if} \quad \beta^* = 0\,.$$

Recall that the scale parameter in the log-transform model is the reciprocal of the shape parameter in the original model; that is, $\sigma = 1/\alpha$. We test H_0 under each of the following cases:

Case 1: Assume equal shapes (α); that is, we assume equal scales $\sigma_1 = \sigma_2 = \sigma$. Hence, error $= \sigma Z$, where the random variable Z has either the standard extreme value, standard logistic, or the standard normal distribution. Recall by standard, we mean $\mu = 0$ and $\sigma = 1$.

Case 2: Assume different shapes; that is, $\sigma_1 \neq \sigma_2$.

Fitting data to the Weibull, log-logistic, and log-normal models

In the following S program we first fit the AML data to the Weibull model and conduct formal tests. Then we fit the AML data to the log-logistic and log-normal models. Quantiles in the log-linear model setting are discussed. Lastly, we compare Q-Q plots. The S function `anova` conducts LRT's for hierarchical

models; that is, each reduced model under consideration is a subset of the full model under consideration. Extraneous output has been deleted. The AML data is stored in the data frame aml.

Model 1: Data come from same distribution. **The Null Model** is $Y = \log(T) = \theta + \sigma Z$, where Z is a standard extreme value random variable.

```
> attach(aml)
> weib.fit0 <- survReg(Surv(weeks,status) ~ 1,dist="weib")
> summary(weib.fit0)
              Value Std. Error       z         p
(Intercept)  3.6425      0.217  16.780 3.43e-063 Scale= 0.912
Loglik(model)= -83.2   Loglik(intercept only)= -83.2
```

Model 2: Case 1: With different locations and equal scales σ, we express this model by

$$Y = \log(T) = \theta + \beta^* \text{group} + \sigma Z. \tag{3.20}$$

```
> weib.fit1 <- survReg(Surv(weeks,status) ~ group,dist="weib")
> summary(weib.fit1)
             Value Std. Error      z         p
(Intercept)  3.180      0.241  13.22 6.89e-040
      group  0.929      0.383   2.43 1.51e-002
Scale= 0.791 Loglik(model)= -80.5 Loglik(intercept only)= -83.2
    Chisq= 5.31 on 1 degrees of freedom, p= 0.021
> weib.fit1$linear.predictors # Extracts the estimated mutildes.
 4.1091 4.1091 4.1091 4.1091 4.1091 4.1091 4.1091 4.1091
 4.1091 4.1091 4.1091 3.1797 3.1797 3.1797 3.1797 3.1797
 3.1797 3.1797 3.1797 3.1797 3.1797 3.1797 3.1797
 # muhat1=4.109 and muhat2=3.18 for maintained and
 # nonmaintained groups respectively.
```

Model 3: Case 2: $Y = \log(T) = \theta + \beta^* \text{group} + \text{error}$, different locations, different scales.

Fit each group separately. On each group run a `survReg` to fit the data. This gives the MLE's of the two locations μ_1 and μ_2, and the two scales σ_1 and σ_2.

```
> weib.fit20 <- survReg(Surv(weeks,status) ~ 1,
                  data=aml[aml$group==0,],dist="weib")
> weib.fit21 <- survReg(Surv(weeks,status) ~ 1,
                  data=aml[aml$group==1,],dist="weib")
```

```
> summary(weib.fit20)
             Value  Std.Error      z         p
(Intercept)  3.222      0.198  16.25 2.31e-059  Scale=0.635
> summary(weib.fit21)
             Value  Std.Error      z         p
(Intercept)  4.1        0.366  11.19 4.74e-029  Scale=0.969
```

To test the reduced model against the full model we use the **LRT**. The `anova` function is appropriate for hierarchical models.

```
> anova(weib.fit0,weib.fit1,test="Chisq")
Analysis of Deviance Table Response: Surv(weeks, status)

  Terms Resid. Df    -2*LL Test Df  Deviance    Pr(Chi)
1     1        21     166.3573
2 group        20     161.0433   1  5.314048  0.02115415
# Model 2 is a significant improvement over the null
# model (Model 1).
```

To construct the appropriate likelihood function for Model 3 to be used in the LRT:

```
> loglik3 <- weib.fit20$loglik[2]+weib.fit21$loglik[2]
> loglik3
[1] -79.84817
> lrt23 <- -2*(weib.fit1$loglik[2]-loglik3)
> lrt23
[1] 1.346954
> 1 - pchisq(lrt23,1)
[1] 0.2458114  # Retain Model 2.
```

The following table summarizes the **three models weib.fit0, 1, and 2**:

Model	Calculated Parameters	The Picture
1 (0)	θ, σ	same location, same scale
2 (1)	θ, β^*, $\sigma \equiv \mu_1$, μ_2, σ	different locations, same scale
3 (2)	μ_1, μ_2, σ_1, σ_2	different locations, different scales

We now use the log-logistic and log-normal distribution to estimate Model 2. The form of the log-linear model is the same. The distribution of error terms is what changes.

$$Y = \log(T) = \theta + \beta^* \text{group} + \sigma Z,$$

where $Z \sim$ standard logistic or standard normal.

```
> loglogis.fit1 <- survReg(Surv(weeks,status) ~ group,
                           dist="loglogistic")
```

```
> summary(loglogis.fit1)
             Value Std. Error      z         p
(Intercept)  2.899      0.267  10.84 2.11e-027
      group  0.604      0.393   1.54 1.24e-001
Scale= 0.513 Loglik(model)= -79.4 Loglik(intercept only)= -80.6
Chisq= 2.41 on 1 degrees of freedom, p= 0.12 # p-value of LRT.
   # The LRT is test for overall model adequacy. It is not
   # significant.

> lognorm.fit1 <- survReg(Surv(weeks,status) ~ group,
                          dist="lognormal")
> summary(lognorm.fit1)
             Value Std. Error       z         p
(Intercept)  2.854      0.254  11.242 2.55e-029
      group  0.724      0.380   1.905 5.68e-002
Scale= 0.865 Loglik(model)= -78.9 Loglik(intercept only)= -80.7
Chisq= 3.49 on 1 degrees of freedom, p= 0.062 # p-value of LRT.
    # Here there is mild evidence of the model adequacy.
```

Quantiles

Let $\widehat{y}_p = \log(\widehat{t}_p)$ denote the estimated pth-quantile. For Model 2 (3.20) the quantile lines are given by

$$\widehat{y}_p = \widehat{\theta} + \widehat{\beta}^* \text{group} + \widehat{\sigma} z_p \,, \tag{3.21}$$

where z_p is the pth-quantile from either the standard normal, standard logistic, or standard extreme value tables. As p changes from 0 to 1, the standard quantiles z_p increase and $\widehat{y}_p$ is linearly related to z_p. The slope of the line is $\widehat{\sigma}$. There are two intercepts, $\widehat{\theta} + \widehat{\beta}^*$ and $\widehat{\theta}$, one for each group. Hence, we obtain two parallel quantile lines. Let us take z_p to be a standard normal quantile. Then if $p = .5$, $z_{.5} = 0$. Hence, $\widehat{y}_{.5} = \widehat{\theta} + \widehat{\beta}^* \text{group}$ represents the estimated median, and the mean as well, for each group. We see that if T is log-normal, then the estimated linear model $\widehat{y}_{.5} = \log(\widehat{t}_{.5}) = \widehat{\theta} + \widehat{\beta}^* \text{group}$ resembles the least squares line where we regress y to the group; that is, $\widehat{y} = \widehat{\theta} + \widehat{\beta}^* \text{group}$ is the estimated mean response for a given group. In Table 3.4 we provide the estimated .10, .25, .50, .75, .90 quantiles for the three error distributions under consideration. Plot any two points $(z_p, \widehat{y}_p)$ for a given group and distribution. Then draw a line through them. This is the MLE line drawn on the Q-Q plots in Figure 3.13.

The following S code computes point and C.I. estimates for the medians and draws Q-Q plots for the three different estimates of Model 2 (3.21). This recipe works for any desired estimated quantile. Just set `p=desired quantile` in the `predict` function.

Table 3.4: *Five quantiles for the AML data under Model 2* (3.21)

		extreme value			logistic			normal		
g	p	z_p	$\widehat{y}_p$	$\hat{t}_p$	z_p	$\widehat{y}_p$	$\hat{t}_p$	z_p	$\widehat{y}_p$	$\hat{t}_p$
	.10	-2.25	1.40	4.05	-2.20	1.77	5.88	-1.28	1.75	5.73
	.25	-1.25	2.19	8.98	-1.10	2.34	10.33	-.67	2.27	9.68
0	.50	-.37	2.89	18	0	2.9	18.16	0	2.85	17.36
	.75	.33	3.44	31.14	1.10	3.46	31.9	.67	3.44	31.12
	.90	.83	3.84	46.51	2.20	4.03	56.05	1.28	3.96	52.6
	.10	-2.25	2.33	10.27	-2.20	2.38	10.76	-1.28	2.47	11.82
	.25	-1.25	3.12	22.73	-1.10	2.94	18.91	-.67	2.99	20
1	.50	-.37	3.82	45.56	0	3.50	33.22	0	3.58	35.8
	.75	.33	4.37	78.84	1.10	4.07	58.36	.67	4.16	64.16
	.90	.83	4.77	117.77	2.2	4.63	102.53	1.28	4.69	108.5

g denotes group.

```
> medhat <- predict(weib.fit1,newdata=list(group=0:1),
                                type="uquantile",se.fit=T,p=0.5)
> medhat
$fit:
        1       2
    2.889819   3.81916
$se.fit:
    0.2525755 0.3083033
> medhat0 <- medhat$fit[1]
> medhat0.se <- medhat$se.fit[1]
> medhat1 <- medhat$fit[2]
> medhat1.se <- medhat$se.fit[2]

> C.I.median0 <- c(exp(medhat0),exp(medhat0-1.96*medhat0.se),
                     exp(medhat0+1.96*medhat0.se))
> names(C.I.median0) <- c("median0","LCL","UCL")
> 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")
# Weibull 95% C.I.'s follow.
> C.I.median0
  median0       LCL       UCL
 17.99005  10.96568  29.51406
> C.I.median1
  median1       LCL       UCL
 45.56593  24.90045  83.38218
# Similarly, log-logistic 95% C.I.'s follow.
> C.I.median0
```

```
  median0       LCL       UCL
 18.14708  10.74736  30.64165
> C.I.median1
  median1       LCL       UCL
 33.21488  18.90175  58.36648
# Log-normal 95% C.I.'s follow.
> C.I.median0
  median0       LCL       UCL
 17.36382  10.55622  28.56158
> C.I.median1
  median1       LCL       UCL
 35.83274  20.50927  62.60512
# The Q-Q plots are next.
> t.s0 <- Surv(weeks[group==0],status[group==0])
> t.s1 <- Surv(weeks[group==1],status[group==1])
> qq.weibull(Surv(weeks,status))
> qq.weibreg(list(t.s0,t.s1),weib.fit1)
> qq.loglogisreg(list(t.s0,t.s1),loglogis.fit1)
> qq.lognormreg(list(t.s0,t.s1),lognorm.fit1)
> detach()
```

Results:

- The LRT per the `anova` function provides evidence that Model 2 (3.20), weib.fit1, which assumes equal scales, is adequate.
- We summarize the distributional fits to Model 2 (3.20) in the following table:

distribution	max. log-likeli $\log(L(\widehat{\theta}, \widehat{\beta}^*))$	p-value for model adequacy	$\widehat{\theta}$	$\widehat{\beta}^*$	p-value for group effect
Weibull	-80.5	0.021	3.180	0.929	0.0151
log-logistic	-79.4	0.12	2.899	0.604	0.124
log-normal	-78.9	0.062	2.854	0.724	0.0568

- For the Weibull fit we conclude that there is a significant "group" effect (p-value= 0.0151). The maintained group tends to stay in remission longer, with estimated extreme value location parameters $\widehat{\mu}_1 = 4.109$ and $\widehat{\mu}_2 = 3.18$.
- The median of the maintained group is 45.6 weeks whereas the median of the nonmaintained group is only about 18 weeks. Corresponding 95% confidence intervals are (24.9, 83.38) weeks, and (10.96, 29.51) weeks, respectively.

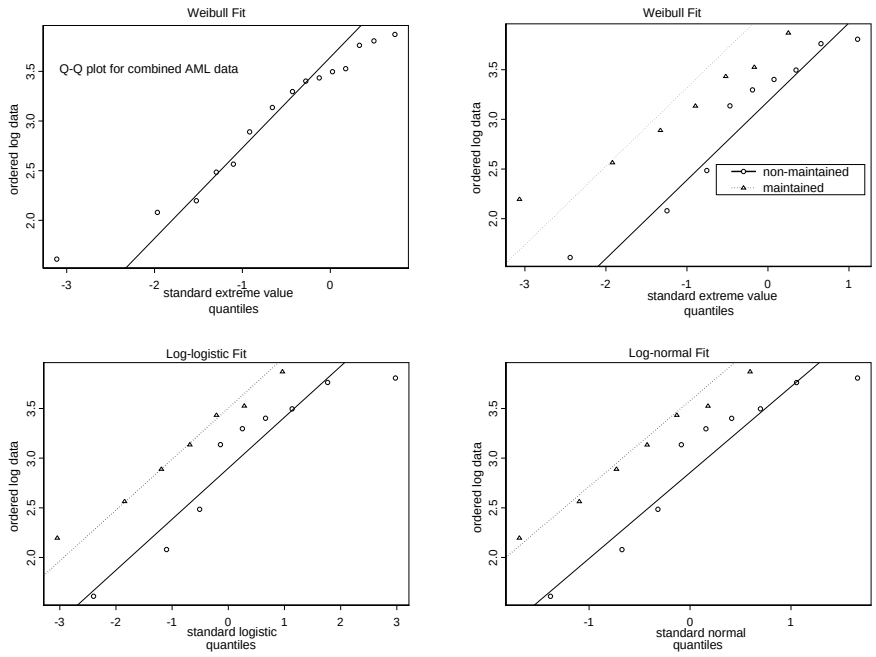


Figure 3.13 *Q-Q plots for the Weibull, the log-logistic, and the log-normal fit to Model 2:* $y = \theta + \beta^*\text{group} + \sigma\text{Z}$. *Each line constructed with the MLE's* $\widehat{\theta}, \widehat{\beta}^*$, *and* $\widehat{\sigma}$. *In each plot, the lines have same slope* $\widehat{\sigma}$ *and different intercepts, either* $\widehat{\theta}$ *or* $\widehat{\theta} + \widehat{\beta}^*$.

- The log-normal has largest maximized likelihood, whereas the Weibull has the smallest. But the LRT for overall model fit is significant only for the Weibull; i.e., its p-value is the only one less than 0.05.
- The estimated linear predictor $\widehat{\widetilde{\mu}} = \widehat{\theta} + \widehat{\beta}^*\text{group}$. As $\widehat{\widetilde{\mu}} = -\log(\widehat{\widetilde{\lambda}}), \widehat{\widetilde{\lambda}} = \exp(-\widehat{\widetilde{\mu}}) = \exp(-\widehat{\theta} - \widehat{\beta}^*\text{group})$. $\widehat{\alpha} = 1/\widehat{\sigma}$. We summarize the estimated parameters for each group and distributional model in the following table:

	Weibull		log-logistic		log-normal	
group	$\widehat{\widetilde{\lambda}}$	$\widehat{\alpha}$	$\widehat{\widetilde{\lambda}}$	$\widehat{\alpha}$	$\widehat{\widetilde{\lambda}}$	$\widehat{\alpha}$
0	0.042	1.264	0.055	1.95	0.058	1.16
1	0.0164	1.264	0.030	1.95	0.028	1.16

- The Q-Q plots in Figure 3.13 suggest that the log-logistic or log-normal models fit the maintained group data better than the Weibull model. However, they do not improve the fit for the nonmaintained.
- The nonparametric approach based on K-M, presented in Chapter 2, may give the better description of this data set.

Prelude to parametric regression models

As a prelude to parametric regression models presented in the next chapter, we continue to explore Model 2 (3.20) under the assumption that $T \sim$ Weibull. That is, we explore

$$\begin{aligned} Y &= \log(T) \\ &= \theta + \beta^* \text{group} + \sigma \text{Z} \\ &= \widetilde{\mu} + \sigma Z \,, \end{aligned}$$

where Z is a standard extreme minimum value random variable. Let the linear predictor $\widetilde{\mu} = -\log(\widetilde{\lambda})$ and $\sigma = 1/\alpha$. It follows from page 49 that the hazard function for the Weibull in this context is expressed as

$$\begin{aligned} h(t|\text{group}) &= \alpha \widetilde{\lambda}^{\alpha} t^{\alpha-1} \\ &= \alpha \lambda^{\alpha} t^{\alpha-1} \exp(\beta \text{group}) \\ &= h_0(t) \exp(\beta \text{group}) \,, \end{aligned} \tag{3.22}$$

when we set $\lambda = \exp(-\theta)$ and $\beta = -\beta^*/\sigma$. WHY! The $h_0(t)$ denotes the baseline hazard; that is, when group $= 0$ or $\beta = 0$. Thus, $h_0(t)$ is the hazard function for the Weibull with scale parameter λ, which is free of any covariate.

The hazard ratio (HR) of group 1 to group 0 is

$$\text{HR} = \frac{h(t|1)}{h(t|0)} = \frac{\exp(\beta)}{\exp(0)} = \exp(\beta) \,.$$

If we believe the Weibull model is appropriate, the HR is constant over follow-up time t. That is, the graph of HR is a horizontal line with height $\exp(\beta)$. We say the Weibull enjoys the *proportional hazards property* to be formally introduced in Chapter 4.3. On the AML data,

$$\widehat{\beta} = \frac{-\widehat{\beta}^*}{\widehat{\sigma}} = \frac{-0.929}{0.791} = -1.1745 \,.$$

Therefore, the estimated HR is

$$\widehat{\text{HR}} = \frac{\hat{h}(t|1)}{\hat{h}(t|0)} = \exp(-1.1745) \approx 0.31 \,.$$

The maintained group has 31% of the risk of the control group's risk of relapse. Or, the control group has (1/0.31)=3.23 times the risk of the maintained group of relapse at any given time t. The HR is a measure of effect that describes the relationship between time to relapse and group.

If we consider the ratio of the estimated survival probabilities, say at $t = 31$ weeks, since $\widehat{\widetilde{\lambda}} = \exp(-\widehat{\widetilde{\mu}})$, we obtain

$$\frac{\widehat{S}(31|1)}{\widehat{S}(31|0)} = \frac{0.652}{0.252} \approx 2.59 \,.$$

The maintained group is 2.59 times more likely to stay in remission at least 31 weeks. The Weibull survivor function $S(t)$ is given in a table on page 49.

3.6 A bivariate version of the delta method

$$\begin{pmatrix} x \\ y \end{pmatrix} \quad \overset{a}{\sim} \quad MVN\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}; \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}\right)$$

and suppose we want the asymptotic distribution of $g(x, y)$. Then the 1^{st} order Taylor approximation for scalar fields is

$$g(x,y) \approx g(\mu_x, \mu_y) + (x - \mu_x)\frac{\partial}{\partial x} g(\mu_x, \mu_y) + (y - \mu_y)\frac{\partial}{\partial y} g(\mu_x, \mu_y).$$

Note that we expand about $(x, y) = (\mu_x, \mu_y)$. The $g(\cdot, \cdot)$ is a bivariate function that yields a scalar, i.e., a univariate. Then

$$\begin{aligned} g(x,y) \quad \overset{a}{\sim} \quad & \text{normal with} \\ & \text{mean} \approx g(\mu_x, \mu_y) \\ & \text{asymptotic variance} \approx \\ & \qquad \sigma_x^2(\tfrac{\partial}{\partial x} g)^2 + \sigma_y^2(\tfrac{\partial}{\partial y} g)^2 + 2\sigma_{xy}(\tfrac{\partial}{\partial x} g)(\tfrac{\partial}{\partial y} g). \end{aligned} \tag{3.23}$$

WHY!

3.7 General version of the likelihood ratio test

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a population with p.d.f. $f(x|\theta)$, (θ may be a vector), where $\theta \in \Omega$, its parameter space. The **likelihood function** is given by

$$L(\theta) = L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta), \text{ where } \mathbf{x} = (x_1, x_2, \ldots, x_n).$$

Let Ω_0 denote the null space. Then $\Omega = \Omega_0 \cup \Omega_0^c$.

Definition 3.7.1 ***The likelihood ratio test statistic***

for testing $H_0 : \theta \in \Omega_0$ (reduced model) against $H_A : \theta \in \Omega_0^c$ (full model) is given by

$$r(\mathbf{x}) = \frac{\sup_{\Omega_0} L(\theta)}{\sup_{\Omega} L(\theta)}.$$

Note that $r(\mathbf{x}) \leq 1$. Furthermore, this handles hypotheses with nuisance parameters. Suppose $\theta = (\theta_1, \theta_2, \theta_3)$. We can test for example $H_0 : (\theta_1 = 0, \theta_2, \theta_3)$ against $H_A : (\theta_1 \neq 0, \theta_2, \theta_3)$. Here θ_2 and θ_3 are nuisance parameters. Most

often, finding the sup amounts to finding the MLE's and then evaluating $L(\theta)$ at the MLE. Thus, for the denominator, obtain the MLE over the whole parameter space Ω. We refer to this as the full model. For the numerator, we maximize $L(\theta)$ over the reduced (restricted) space Ω_0. Find the MLE in Ω_0 and put into $L(\cdot)$. As $r(\mathbf{x}) \leq 1$, we reject H_0 for small values. Or, equivalently, we reject H_0 for large values of

$$r^*(\mathbf{x}) = -2 \log r(\mathbf{x}).$$

Theorem 3.7.1 ***Asymptotic distribution of the r*(x) test statistic.***

Under $H_0 : \theta \in \Omega_0$, the distribution of $r^*(\mathbf{x})$ converges to a $\chi^2_{(df)}$ as $n \to \infty$. The degrees of freedom (df) = (# of free parameters in Ω) − (# of free parameters $\in \Omega_0$).

That is,

$$r^*(\mathbf{x}) \quad \overset{a}{\sim} \quad \chi^2_{(df)}.$$

Proof: See Bickel & Doksum (2001, Chapter 6.3, Theorem 6.3.2).

Thus, an approximate size$-\alpha$ test is: reject H_0 iff $r^*(\mathbf{x}) = -2 \log r(\mathbf{x}) \geq \chi^2_\alpha$.

To compute approximate p-value: if $r^*(\mathbf{x}) = r^*$, then

$$p\text{-value} \approx P(r^*(\mathbf{x}) \geq r^*),$$

the area under the Chi-square curve to the right of r^*; that is, the upper tail area.

CHAPTER 4

Regression Models

Let T denote failure time and $\underline{x} = (x^{(1)}, \ldots, x^{(m)})'$ represent a vector of available **covariates**. We are interested in modelling and determining the relationship between T and $\underline{x}$. Often this is referred to as prognostic factor analysis. These $\underline{x}$ are also called regression variables, regressors, factors, or explanatory variables. The primary question is: Do any subsets of the m covariates help to explain survival time? For example, does age at first treatment and/or gender increase or decrease (relative) risk of survival? If so, how and by what estimated quantity?

Example 1. Let

- $x^{(1)}$ denote the sex ($x_i^{(1)} = 1$ for males and $x_i^{(1)} = 0$ for females),
- $x^{(2)} =$ Age at diagnosis,
- $x^{(3)} = x^{(1)} \cdot x^{(2)}$ (interaction),
- $T =$ survival time.

We introduce four models: the exponential, the Weibull, the Cox proportional hazards, and the accelerated failure time model, and a variable selection procedure.

Objectives of this chapter:

After studying Chapter 4, the student should:

1 Understand that the hazard function is modelled as a function of available covariates $\underline{x} = (x^{(1)}, \ldots, x^{(m)})'$.

2 Know that the **preferred link function** for $\eta = \underline{x}'\underline{\beta}$ is $\boldsymbol{k(\eta) = \exp(\eta)}$ and why.

3 Recognize the **exponential** and **Weibull regression models**.

4 Know the definition of the **Cox proportional hazards model**.

5 Know the definition of an **accelerated failure time model**.

6 Know how to compute the **AIC** statistic.

7 Know how to implement the S functions `survReg` and `predict` to estimate and analyze a parametric regression model and obtain estimated quantities of interest.

8 Know how to interpret the effects of a covariate on the risk and survivor functions.

4.1 Exponential regression model

We first generalize the exponential distribution. Recall that for the exponential distribution the hazard function, $h(t) = \lambda$, is constant with respect to time and that $E(T) = \frac{1}{\lambda}$. We model the hazard rate λ as a function of the covariate vector $\underline{x}$.

We assume the hazard function at time t for an individual has the form

$$h(t|\underline{x}) = h_0(t) \cdot k(\underline{x}'\underline{\beta}) = \lambda \cdot k(\underline{x}'\underline{\beta}) = \lambda \cdot k(\beta_1 x^{(1)} + \cdots + \beta_m x^{(m)}),$$

where $\underline{\beta} = [\beta_1, \beta_2, \ldots, \beta_m]'$ is a vector of regression parameters (coefficients), $\lambda > 0$ is a constant, and k is a specified *link function*. The function $h_0(t)$ is called the baseline hazard. It's the value of the hazard function when the covariate vector $\underline{x} = \underline{0}$ or $\underline{\beta} = \underline{0}$. Note that this hazard function is constant with respect to time t, but depends on $\underline{x}$.

The most natural choice for k is $k(x) = \exp(x)$, which implies

$$\begin{aligned} h(t|\underline{x}) &= \lambda \cdot \exp(\underline{x}'\underline{\beta}) \\ &= \lambda \cdot \exp\left(\beta_1 x^{(1)} + \cdots + \beta_m x^{(m)}\right) \\ &= \lambda \cdot \exp\left(\beta_1 x^{(1)}\right) \times \exp\left(\beta_2 x^{(2)}\right) \times \cdots \times \exp\left(\beta_m x^{(m)}\right). \end{aligned}$$

This says that the covariates act multiplicatively on the hazard rate. Equivalently, this specifies

$$\log(h(t|\underline{x})) = \log(\lambda) + \eta = \log(\lambda) + (\underline{x}'\underline{\beta}) = \log(\lambda) + \beta_1 x^{(1)} + \cdots + \beta_m x^{(m)}.$$

That is, the covariates act additively on the log failure rate – a log-linear model for the failure rate. The quantity $\eta = \underline{x}'\underline{\beta}$ is called the *linear predictor of the log-hazard*. We may consider a couple of other k functions that may appear natural, $k(\eta) = 1+\eta$ and $k(\eta) = 1/(1+\eta)$. The first one has a hazard function $h(t|\underline{x}) = \lambda \times (1 + \underline{x}'\underline{\beta})$ which is a linear function of $\underline{x}$ and the second has the mean $E(T|\underline{x}) = 1/h(t|\underline{x}) = (1 + \underline{x}'\underline{\beta})/\lambda$ which is a linear function of $\underline{x}$. Note that both proposals could produce **negative** values for the hazard (which is a violation) unless the set of $\underline{\beta}$ values is restricted to guarantee $k(\underline{x}'\underline{\beta}) > 0$ for all possible $\underline{x}$. Therefore, **$k(\eta) = \exp(\eta)$ is the most natural since it will always be positive no matter what the $\underline{\beta}$ and $\underline{x}$ are**.

The survivor function of T given $\underline{x}$ is

$$S(t|x) = \exp\Big(-h(t|\underline{x})t\Big) = \exp\Big(-\lambda \exp(\underline{x}'\underline{\beta})t\Big).$$

Thus, the p.d.f. of T given $\underline{x}$ is

$$f(t|\underline{x}) = h(t|\underline{x})S(t|\underline{x}) = \lambda \exp(\underline{x}'\underline{\beta}) \exp\left(-\lambda \exp(\underline{x}'\underline{\beta})t\right).$$

Recall from **Fact**, Chapter 3.1, page 50, that if T is distributed exponentially, $Y = \log(T)$ is distributed as the extreme (minimum) value distribution with scale parameter $\sigma = 1$. Here, given $\underline{x}$, we have

$$\tilde{\mu} = -\log(h(t|\underline{x})) = -\log\left(\lambda \exp(\underline{x}'\underline{\beta})\right) = -\log(\lambda) - \underline{x}'\underline{\beta} \quad \text{and} \quad \sigma = 1.$$

Therefore, given $\underline{x}$,

$$Y = \log(T) = \tilde{\mu} + \sigma Z = \beta_0^* + \underline{x}'\underline{\beta}^* + Z \ ,$$

where $\beta_0^* = -\log(\lambda)$, $\underline{\beta}^* = -\underline{\beta}$, and $Z \sim f(z) = \exp(z - e^z)$, $-\infty < z < \infty$, the standard extreme (minimum) value distribution. The quantity $\widetilde{\mu} = \beta_0^* + \underline{x}'\underline{\beta}^*$ is called the *linear predictor of the log-time.*

In summary, $h(t|\underline{x}) = \lambda \exp(\underline{x}'\underline{\beta})$ is a log-linear model for the failure rate and transforms into a **linear** model for $Y = \log(T)$ in that the covariates act additively on Y.

Example 1 continued: The exponential distribution is usually a poor model for human survival times. We use it anyway for illustration. We obtain

hazard function:	$h(t\|\underline{x}) = \lambda \exp(\underline{x}'\underline{\beta})$
log(hazard):	$\log(h(t\|\underline{x})) = \log(\lambda) + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \beta_3 x^{(3)}$
survivor function:	$S(t\|\underline{x}) = \exp(-\lambda \exp(\underline{x}'\underline{\beta})t)$

	Male	Female
hazard	$\lambda \exp\left(\beta_1 + (\beta_2 + \beta_3)\text{age}\right)$	$\lambda \exp(\beta_2 \text{ age})$
log(hazard)	$(\log(\lambda) + \beta_1) + (\beta_2 + \beta_3)\text{age}$	$\log(\lambda) + \beta_2 \text{ age}$
survivor	$\exp\left(-\lambda \exp(\beta_1 + (\beta_2 + \beta_3)\text{age})t\right)$	$\exp\left(-\lambda \exp(\beta_2 \text{age})t\right)$

Take $\lambda = 1, \beta_1 = -1, \beta_2 = -0.2, \beta_3 = 0.1$. Then

	Male	Female
hazard	$\exp(-1 - .1 \cdot \text{ age})$	$\exp(-0.2 \text{ age})$
log(hazard)	$-1 - 0.1 \cdot \text{ age}$	$-0.2 \cdot \text{ age}$
survivor	$\exp\left(-\exp(-1 - 0.1 \cdot \text{age})t\right)$	$\exp\left(-\exp(-0.2 \cdot \text{age})t\right)$

Plots for this example are displayed in Figure 4.1.

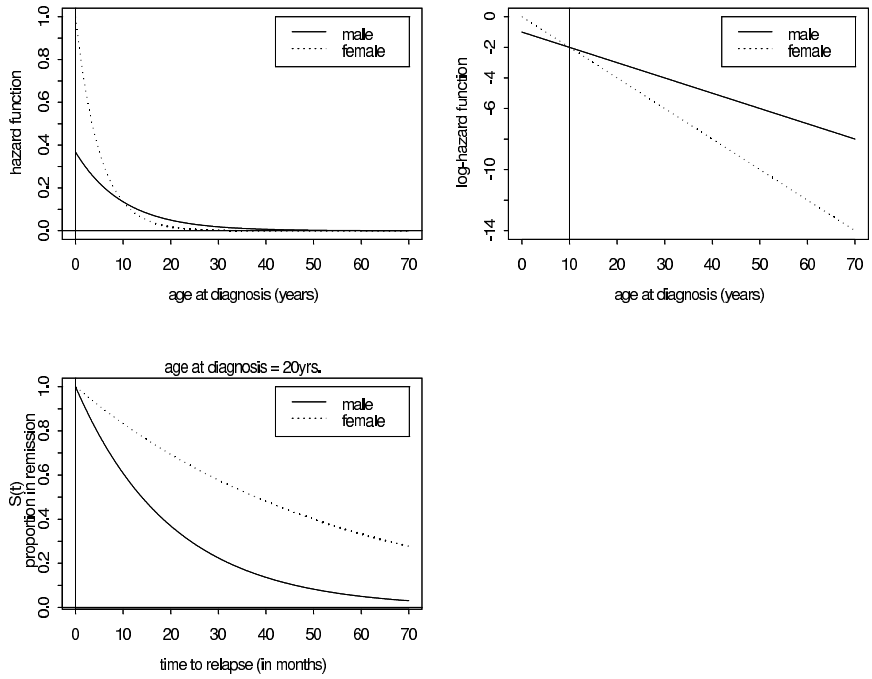


Figure 4.1 *Plots for Example 1.*

4.2 Weibull regression model

We generalize the Weibull distribution to regression in a similar fashion. Recall that its hazard function is $h(t) = \alpha\lambda^\alpha t^{\alpha-1}$.

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

$$\begin{aligned} h(t|\underline{x}) &= h_0(t) \cdot \exp(\underline{x}'\underline{\beta}) \\ &= \alpha\lambda^\alpha t^{\alpha-1} \exp(\underline{x}'\underline{\beta}) = \alpha \left(\lambda \cdot \left(\exp(\underline{x}'\underline{\beta})\right)^{\frac{1}{\alpha}}\right)^\alpha t^{\alpha-1} \\ &= \alpha(\tilde{\lambda})^\alpha t^{\alpha-1}, \end{aligned} \tag{4.1}$$

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

Again notice that

$$\begin{aligned} \log(h(t|\underline{x})) &= \log(\alpha) + \alpha\log(\tilde{\lambda}) + (\alpha-1)\log(t) \\ &= \log(\alpha) + \alpha\log(\lambda) + \underline{x}'\underline{\beta} + (\alpha-1)\log(t)\ . \end{aligned}$$

From **Fact**, Chapter 3.1, page 50, if $T \sim$ Weibull, then given $\underline{x}$, $Y = \log(T)$ $= \tilde{\mu} + \sigma Z$, where

$$\tilde{\mu} = -\log(\tilde{\lambda}) = -\log(\lambda \cdot (\exp(\underline{x}'\underline{\beta}))^{\frac{1}{\alpha}}) = -\log(\lambda) - \frac{1}{\alpha}\underline{x}'\underline{\beta}\,, \tag{4.2}$$

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

$$Y = \underbrace{\beta_0^* + \underline{x}'\underline{\beta}^*}_{\tilde{\mu}} + \sigma Z\,, \tag{4.3}$$

where $\beta_0^* = -\log(\lambda)$ and $\underline{\beta}^* = -\sigma\underline{\beta}$. It then follows from the table on page 49 that the survivor function of T given $\underline{x}$ is

$$S(t|\underline{x}) = \exp\left(-(\tilde{\lambda}t)^{\alpha}\right). \tag{4.4}$$

It follows from the relationship between the cumulative hazard and survivor functions given in expression (1.6) that, for a given $\underline{x}$, $H(t|\underline{x}) = -\log(S(t|\underline{x}))$. An expression for the log-cumulative hazard function follows from expression (4.2) for $\log(\widetilde{\lambda})$.

$$\begin{aligned} \log\Big(H(t|\underline{x})\Big) &= \alpha\log(\widetilde{\lambda}) + \alpha\log(t) \\ &= \alpha\log(\lambda) + \alpha\log(t) + \underline{x}'\underline{\beta} \\ &= \log\Big(H_0(t)\Big) + \underline{x}'\underline{\beta}\,, \end{aligned} \tag{4.5}$$

where $H_0(t) = -\log\Big(S_0(t)\Big) = (\lambda t)^{\alpha}$. 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 $\underline{x}'\underline{\beta} + \alpha\log(\lambda)$. Expression (4.5) can also be derived by noting expression (4.1) and definition (1.6) give

$$H(t|\underline{x}) = H_0(t)\exp(\underline{x}'\underline{\beta}) = (\lambda t)^{\alpha}\exp(\underline{x}'\underline{\beta})\,. \tag{4.6}$$

In summary, for both the exponential and Weibull regression model, the effects of the covariates $\underline{x}$ act multiplicatively on the hazard function $h(t|\underline{x})$ which is clear from the form

$$\begin{aligned} h(t|\underline{x}) &= h_0(t)\cdot\exp(\underline{x}'\underline{\beta}) = h_0(t)\cdot\exp\left(\beta_1 x^{(1)} + \cdots + \beta_m x^{(m)}\right) \\ &= h_0(t)\cdot\exp\left(\beta_1 x^{(1)}\right)\times\exp\left(\beta_2 x^{(2)}\right)\times\cdots\times\exp\left(\beta_m x^{(m)}\right). \end{aligned}$$

This suggests the more general **Cox proportional hazards model**, presented in the next section. Further, both are log-linear models for T in that these models transform into a linear model for $Y = \log(T)$. That is, the covariates $\underline{x}$ act additively on $\log(T)$ (multiplicatively on T), which is clear from the form

$$Y = \log(T) = \tilde{\mu} + \sigma Z = \beta_0^* + \underline{x}'\underline{\beta}^* + \sigma Z\,.$$

This suggests a more general class of log-linear models called **accelerated failure time models** discussed in Section 4.4 of this chapter.

The difference from an ordinary linear regression model for the log-transformed target variable T, $Y = \log(T)$, is the distribution of the errors Z, which here is an extreme value distribution rather than a normal one. Therefore, least-squares methods are not adequate. Furthermore, there will be methods to deal with censored values, which is rarely discussed for ordinary linear regression.

4.3 Cox proportional hazards (PH) model

For the Cox (1972) PH model, the hazard function is

$$h(t|\underline{x}) = h_0(t) \cdot \exp(\underline{x}'\underline{\beta}), \tag{4.7}$$

where $h_0(t)$ is an unspecified baseline hazard function free of the covariates $\underline{x}$. The covariates act multiplicatively on the hazard. Clearly, the exponential and Weibull are special cases. At two different points $\underline{x}_1, \underline{x}_2$, the proportion

$$\frac{h(t|\underline{x}_1)}{h(t|\underline{x}_2)} = \frac{\exp(\underline{x}_1'\underline{\beta})}{\exp(\underline{x}_2'\underline{\beta})} = \exp\Big((\underline{x}_1' - \underline{x}_2')\underline{\beta}\Big), \tag{4.8}$$

called the **hazards ratio (HR)**, is constant with respect to time t. This defines the *proportional hazards property.*

Remark:

As with linear and logistic regression modelling, **a statistical goal of a survival analysis is to obtain some measure of effect that will describe the relationship between a predictor variable of interest and time to failure, after adjusting for the other variables we have identified in the study and included in the model.** In linear regression modelling, the measure of effect is usually the regression coefficient β. In logistic regression, the measure of effect is an odds ratio, the log of which is β for a change of 1 unit in x. **In survival analysis, the measure of effect is the hazards ratio (HR).** As is seen above, this ratio is also expressed in terms of an exponential of the regression coefficient in the model.

For example, let β_1 denote the coefficient of the group covariate with group = 1 if received treatment and group = 0 if received placebo. Put treatment group in the numerator of **HR**. A HR of 1 means that there is no effect. A hazards ratio of 10, on the other hand, means that the treatment group has ten times the hazard of the placebo group. Similarly, a HR of 1/10 implies that the treatment group has one-tenth the hazard or risk of the placebo group.

Recall the relationship between hazard and survival is $S(t) = \exp\big(-H(t)\big)$. If the HR is less than one, then the ratio of corresponding survival probabilities is larger than one. Hence, the treatment group has larger probability

of survival at **any** given time t, after adjusting for the other covariates. WHY!

For any PH model, which includes the Weibull model as well as the Cox model, the **survivor function** of T given $\underline{x}$ is

$$\begin{aligned}\boxed{S(t|\underline{x})} &= \exp\left(-\int_0^t h(u|\underline{x})du\right) = \exp\left(-\exp(\underline{x}'\underline{\beta})\int_0^t h_0(u)du\right) \\ &= \left(\exp\left(-\int_0^t h_0(u)du\right)\right)^{\exp(\underline{x}'\underline{\beta})} = \boxed{(S_0(t))^{\exp(\underline{x}'\underline{\beta})}},\end{aligned}$$

where $S_0(t)$ denotes the baseline survivor function.

The p.d.f. of T given $\underline{x}$ is

$$f(t|\underline{x}) = h_0(t)\exp(\underline{x}'\underline{\beta})\,(S_0(t))^{\exp(\underline{x}'\underline{\beta})}\ .$$

There are two important generalizations:

(1) The baseline hazard $h_0(t)$ can be allowed to vary in specified subsets of the data.

(2) The regression variables $\underline{x}$ can be allowed to depend on time; that is, $\underline{x} = \underline{x}(t)$.

Chapter 5 is devoted to an example of a Cox PH prognostic factor analysis. A data set referred to as the **CNS lymphoma data** is extensively analyzed using various S/R functions.

4.4 Accelerated failure time model

This model is a log-linear regression model for T in that we model $Y = \log(T)$ as a linear function of the covariate $\underline{x}$. Suppose

$$Y = \underline{x}'\underline{\beta}^* + Z^*,$$

where Z^* has a certain distribution. Then

$$T = \exp(Y) = \exp(\underline{x}'\underline{\beta}^*)\cdot\exp(Z^*) = \exp(\underline{x}'\underline{\beta}^*)\cdot T^*,$$

where $T^* = \exp(Z^*)$. Here the covariate $\underline{x}$ acts multiplicatively on the survival time T. Suppose further that T^* has hazard function $h_0^*(t^*)$ which is independent of $\underline{\beta}^*$; that is, free of the covariate vector $\underline{x}$. The hazard function of T for a given $\underline{x}$ can be written in terms of the baseline function $h_0^*(\cdot)$ according to

$$h(t|\underline{x}) = h_0^*(\exp(-\underline{x}'\underline{\beta}^*)t)\cdot\exp(-\underline{x}'\underline{\beta}^*). \tag{4.9}$$

We see here that the covariates $\underline{x}$ act multiplicatively on both t and the hazard function. The log-logistic and log-normal regression models are examples of accelerated failure time models as well as the exponential and Weibull regression models.

It follows from expressions (1.6) and (4.9) that the **survivor function** of T given $\underline{x}$ is

$$S(t|\underline{x}) = \exp\left(-\exp(-\underline{x}'\underline{\beta}^*)\int_0^t h_0^*\left(\exp(-\underline{x}'\underline{\beta}^*)u\right)du\right). \tag{4.10}$$

Change the integration variable to $v = \exp(-\underline{x}'\underline{\beta}^*)u$. Then $dv = \exp(-\underline{x}'\underline{\beta}^*)du$ and $0 < v < \exp(-\underline{x}'\underline{\beta}^*)t$. Then for the accelerated failure time model,

$$\boxed{S(t|\underline{x})} = \exp\left(-\int_0^{\exp(-\underline{x}'\underline{\beta}^*)t} h_0^*(v)dv\right) = \boxed{S_0^*\Big(\exp(-\underline{x}'\underline{\beta}^*)t\Big)} = S_0^*(t^*), \tag{4.11}$$

where $S_0^*(t)$ denotes the baseline survivor function. Here we notice that the role of the covariate $\underline{x}$ changes the scale of the horizontal (t) axis. For example, if $\underline{x}'\underline{\beta}^*$ increases, then the last term in expression (4.11) increases. In this case it has decelerated the time to failure. This is why the log-linear model defined here is called the accelerated (decelerated) failure time model.

Remarks:

1 We have seen that the Weibull regression model, which includes the exponential, is a special case of both the Cox PH model and the accelerated failure time model. It is shown on pages 34 and 35 of Kalbfleisch and Prentice (1980) that the only log-linear models that are also PH models are the Weibull regression models.

2 Through the **partial likelihood** (Cox, 1975) we obtain estimates of the coefficients $\underline{\beta}$ that require no restriction on the baseline hazard $h_0(t)$. The S function `coxph` implements this. This partial likelihood is heuristically derived in Chapter 6.

3 For the accelerated failure time models we specify the baseline hazard function $h_0(t)$ by specifying the distribution function of Z^*.

4 Hosmer and Lameshow (1999) well present the *proportional odds and proportional times properties* of the log-logistic regression model. From expression (4.11) and page 52 we can express the log-logistic's survivor function as

$$S(t|\underline{x}, \beta_0^*, \underline{\beta}^*, \alpha) = \frac{1}{1 + \exp(\alpha(y - \beta_0^* - \underline{x}'\underline{\beta}^*))}, \tag{4.12}$$

where $y = \log(t)$, $\beta_0^* = -\log(\lambda)$, and $\alpha = 1/\sigma$. WHY! The odds of survival beyond time t is given by

$$\frac{S(t|\underline{x}, \beta_0^*, \underline{\beta}^*, \alpha)}{1 - S(t|\underline{x}, \beta_0^*, \underline{\beta}^*, \alpha)} = \exp(-\alpha(y - \beta_0^* - \underline{x}'\underline{\beta}^*)). \tag{4.13}$$

Note that $-\log(\text{odds})$ is both a linear function of $\log(t)$ and the covariates $x^{(j)}$'s, $j = 1, \ldots, m$. The odds-ratio of survival beyond time t evaluated at

$\underline{x}_1$ and $\underline{x}_2$ is given by

$$\text{OR}(t|\underline{x} = \underline{x}_2, \underline{x} = \underline{x}_1) = \exp(\alpha(\underline{x}_2 - \underline{x}_1)'\underline{\beta}^*). \tag{4.14}$$

The odds-ratio is commonly used as a measure of the effects of covariates. Note that the ratio is independent of time, which is referred to as the *proportional odds property.* For example, if OR = 2, then the odds of survival beyond time t among subjects with $\underline{x}_2$ is twice that of subjects with $\underline{x}_1$, and this holds for all t. Alternatively, some researchers prefer to describe the effects of covariates in terms of the survival time. The $(p \times 100)$th percentile of the survival distribution is given by

$$t_p(\underline{x}, \beta_0^*, \underline{\beta}^*, \alpha) = \Big(p/(1-p)\Big)^{\sigma} \exp(\beta_0^* + \underline{x}'\underline{\beta}^*). \tag{4.15}$$

WHY! Then, for example, the times-ratio at the median is

$$\text{TR}(t_{.5}|\underline{x} = \underline{x}_2, \underline{x} = \underline{x}_1) = \exp((\underline{x}_2 - \underline{x}_1)'\underline{\beta}^*). \tag{4.16}$$

This holds for any p. The TR is constant with respect to time, which is referred to as the *proportional times property.* Similarly, if TR = 2, then the survival time among subjects with $\underline{x}_2$ is twice that of subjects with $\underline{x}_1$, and this holds for all t. The upshot is that $\text{OR} = \text{TR}^{\alpha}$. That is, the odds-ratio is the power of the time ratio. Hence, the rate of change of OR is controlled by α, the shape parameter of the log-logistic distribution. For $\alpha = 1$, OR = TR. If $\alpha = 2$ and TR = 2, then $\text{OR} = 2^2 = 4$. For one unit increase in a single component, fixing the other components in $\underline{x}$, $\text{OR} \to +\infty$ or 0 as $\alpha \to \infty$ depending on the sign of the corresponding component of $\underline{\beta}^*$, and $\to 1$ as $\alpha \to 0$. Finally, Cox and Oakes (1984, page 79) claim that the log-logistic model is the only accelerated failure time model with the *proportional odds property*; equivalently, the only model with the *proportional times property.*

4.5 Summary

Let Z denote either a standard extreme value, standard logistic, or standard normal random variable. That is, each has location $\mu = 0$ and scale $\sigma = 1$.

- $$Y = \log(T) = \widetilde{\mu} + \sigma Z = \beta_0^* + \underline{x}'\underline{\beta}^* + \sigma Z$$

  accelerated failure time model
  log-linear model

$\nearrow\swarrow$	$\uparrow\downarrow$	$\searrow\nwarrow$
T	T	T
Weibull	log-logistic	log-normal
$\downarrow$	$\downarrow$	
PH property	proportional odds property	
	$\updownarrow$	
	proportional times property	

  The $\tilde{\mu}$ is called the *linear predictor* and σ is the *scale parameter*. In the target variable T distribution, $\widetilde{\lambda} = \exp(-\widetilde{\mu})$ and the shape $\alpha = 1/\sigma$. The S function `survReg` estimates β_0^*, $\underline{\beta}^*$, and σ. The `predict` function provides estimates of $\tilde{\mu}$ at specified values of the covariates. For example, returning to the AML data, where we have one covariate "group" with two values 0 or 1, to estimate the linear predictor (`lp`) for the maintained group, use `> predict(fit,type="lp",newdata=list(group=1),se.fit=T)`.

- The Weibull regression model is the only log-linear model that has the proportional hazards property. For both the Cox PH model and the Weibull regression model, we model the hazard function

  $$h(t|\underline{x}) = h_0(t) \cdot \exp(\underline{x}'\underline{\beta}),$$

  where $h_0(t)$ is the baseline hazard function. For the Weibull model, the baseline hazard $h_0(t) = \alpha\lambda^\alpha t^{\alpha-1}$, the baseline cumulative hazard $H_0(t) = (\lambda t)^\alpha$, and the log-cumulative hazard

  $$\log\Big(H(t|\underline{x})\Big) = \alpha\log(\lambda) + \alpha\log(t) + \underline{x}'\underline{\beta}\,.$$

  For the Weibull model, the relationship between the coefficients in the log-linear model and coefficients in modelling the hazard function is

  $$\underline{\beta} = -\sigma^{-1}\underline{\beta}^* \quad \text{and} \quad \lambda = \exp(-\beta_0^*)\,.$$

  The S function `survReg` estimates β_0^*, $\underline{\beta}^*$, and σ. The hazard ratio is

  $$\mathrm{HR}(t|\underline{x} = \underline{x}_2, x = \underline{x}_1) = \frac{h(t|\underline{x}_2)}{h(t|\underline{x}_1)} = \Big(\exp\Big((\underline{x}_1' - \underline{x}_2')\underline{\beta}^*\Big)\Big)^{\frac{1}{\sigma}}\,.$$

Fitting data to a Cox PH model is presented in detail in Chapter 5. The Cox procedure estimates the $\underline{\beta}$ coefficients directly.

- The log-logistic regression model is the only log-linear model that has the proportional odds property. The survivor function is

$$S(t|\underline{x}) = S_0^*\big(\exp(-\underline{x}'\beta^*)t\big) = \frac{1}{1+\big(\exp(y-\beta_0^*-\underline{x}'\beta^*)\big)^{\frac{1}{\sigma}}},$$

where $S_0^*(t)$ is the baseline survivor function, $y = \log(t)$, $\beta_0^* = -\log(\lambda)$, and $\alpha = 1/\sigma$.

The odds of survival beyond time t is given by

$$\frac{S(t|\underline{x})}{1-S(t|\underline{x})} = \Big(\exp(y-\beta_0^*-\underline{x}'\underline{\beta}^*)\Big)^{-\frac{1}{\sigma}}.$$

The $(p\times 100)$th percentile of the survival distribution is given by

$$t_p(\underline{x}) = \Big(p/(1-p)\Big)^{\sigma}\exp(\beta_0^*+\underline{x}'\underline{\beta}^*).$$

The odds-ratio of survival beyond time t evaluated at $\underline{x}_1$ and $\underline{x}_2$ is given by

$$\mathrm{OR}(t|\underline{x}=\underline{x}_2,\underline{x}=\underline{x}_1) = \Big(\exp\Big((\underline{x}_2-\underline{x}_1)'\underline{\beta}^*\Big)\Big)^{\frac{1}{\sigma}} = \Big(\mathrm{TR}\Big)^{\frac{1}{\sigma}},$$

where TR is the times-ratio. The reciprocal of the OR has the same functional form as the HR in the Weibull model with respect to $\underline{\beta}^*$ and σ.

- The upshot is: to obtain the estimated measures of effect, $\widehat{\mathrm{HR}}$ and $\widehat{\mathrm{OR}}$, we need only the estimates given by `survReg`.

4.6 AIC procedure for variable selection

Akaike's information criterion (AIC):

Comparisons between a number of possible models, which need not necessarily be nested nor have the same error distribution, can be made on the basis of the statistic

$$\mathrm{AIC} = -2\times\log(\text{maximum likelihood}) + k\times p,$$

where p is the number of parameters in each model under consideration and k a predetermined constant. This statistic is called **Akaike's (1974) information criterion (AIC)**; the smaller the value of this statistic, the better the model. This statistic trades off goodness of fit (measured by the maximized log-likelihood) against model complexity (measured by p). Here we shall take k as 2. For other choice of values for k, see the remarks at the end of this section.

We can rewrite the AIC to address parametric regression models considered in the text. For the parametric models discussed, the AIC is given by

$$\text{AIC} = -2 \times \log(\text{maximum likelihood}) + 2 \times (a + b), \tag{4.17}$$

where a is the number of parameters in the specific model and b the number of one-dimensional covariates. For example, $a = 1$ for the exponential model, $a = 2$ for the Weibull, log-logistic, and log-normal models.

Here we manually step through a sequence of models as there is only one one-dimensional covariate. But in Chapter 5 we apply an automated model selection procedure via an S function `stepAIC` as there are many one-dimensional covariates.

Motorette data example:

The data set given in Table 4.1 below was obtained by Nelson and Hahn (1972) and discussed again in Kalbfleisch and Prentice (1980), on pages 4, 5, 58, and 59. Hours to failure of motorettes are given as a function of operating temperatures 150^0C, 170^0C, 190^0C, or 220^0C. There is severe (Type I) censoring, with only 17 out of 40 motorettes failing. Note that the stress (temperature) is constant for any particular motorette over time. The primary purpose of the experiment was to estimate certain percentiles of the failure time distribution at a design temperature of 130^0C. We see that this is an accelerated process. The experiment is conducted at higher temperatures to speed up failure time. Then they make predictions at a lower temperature that would have taken them much longer to observe. The authors use the single regressor variable $x = 1000/(273.2+\text{Temperature})$. They also omit all ten data points at temperature level of 150^0C. We also do this in order to compare our results with Nelson and Hahn and Kalbfleisch and Prentice. We entered the data into a data frame called **motorette**. It contains

time	status	temp	x
hours	1 if uncensored 0 if censored	^{0}C	$1000/(273.2+^0\text{C})$

We now fit the exponential, Weibull, log-logistic, and log-normal models. The log likelihood and the AIC for each model are reported in Table 4.2. The S code for computing the AIC follows next. For each of these models the form is the same:

$$\begin{aligned} \text{intercept only: } Y &= \log(T) = \beta_0^* + \sigma Z \\ \text{both: } Y &= \log(T) = \beta_0^* + \beta_1^* + \sigma Z, \end{aligned}$$

where the distributions of Z are standard extreme (minimum) value, standard logistic, and standard normal, respectively.

Table 4.1: *Hours to failure of Motorettes*

Temperature	Times
150^0C	All 10 motorettes without failure at 8064 hours
170^0C	1764, 2772, 3444, 3542, 3780, 4860, 5196 3 motorettes without failure at 5448 hours
190^0C	408, 408, 1344, 1344, 1440 5 motorettes without failure at 1680 hours
220^0C	408, 408, 504, 504, 504 5 motorettes without failure at 528 hours
$n = 40$,	n_u = no. of uncensored times = 17

Table 4.2: *Results of fitting parametric models to the Motorette data*

Model	log-likelihood		AIC
exponential	intercept only	-155.875	311.750 + 2(1) = 313.750
	both	-151.803	303.606 + 2(1 + 1) = 307.606
Weibull	intercept only	-155.681	311.363 + 2(2) = 315.363
	both	-144.345	288.690 + 2(2 + 1) = 294.690
log-logistic	intercept only	-155.732	311.464 + 2(2) = 315.464
	both	-144.838	289.676 + 2(2 + 1) = 295.676
log-normal	intercept only	-155.018	310.036 + 2(2) = 314.036
	both	-145.867	291.735 + 2(2 + 1) = 297.735

The S code for computing the AIC for a number of specified distributions

```
> attach(motorette) # attach the data frame motorette to avoid
                    # continually referring to it.
 # Weibull fit
> weib.fit <- survReg(Surv(time,status)~x,dist="weibull")
> weib.fit$loglik # the first component for intercept only and
                  # the second for both
[1] -155.6817 -144.3449
> -2*weib.fit$loglik # -2 times maximum log-likelihood
[1] 311.3634 288.6898
 # exponential fit
> exp.fit <- survReg(Surv(time,status)~x,dist="exp")
> -2*exp.fit$loglik
[1] 311.7501 303.6064
```

```
 # log-normal fit
> lognormal.fit <- survReg(Surv(time,status)~x,
                           dist="lognormal")
> -2*lognormal.fit$loglik
[1] 310.0359 291.7345
 # log-logistic fit
> loglogistic.fit <- survReg(Surv(time,status)~x,
                             dist="loglogistic")
> -2*loglogistic.fit$loglik
[1] 311.4636 289.6762
> detach() # Use this to detach the data frame when no
           # longer in use.
```

Nelson and Hahn applied a log-normal model, and Kalbfleisch and Prentice applied a Weibull model. Kalbfleisch and Prentice state that the Weibull model is to some extent preferable to the log-normal on account of the larger maximized log likelihood. From Table 4.2, we find that the Weibull distribution provides the best fit to this data, the log-logistic distribution is a close second, and the log-normal distribution is the third.

When there are no subject matter grounds for model choice, the model chosen for initial consideration from a set of alternatives might be the one for which the value of AIC is a minimum. It will then be important to confirm that the model does fit the data using the methods for model checking described in Chapter 6. We revisit AIC in the context of the PH regression model in Chapter 5.

Remarks:

1 In his paper (1974), Akaike motivates the need to develop a new model identification procedure by showing the standard hypothesis testing procedure is not adequately defined as a procedure for statistical model identification. He then introduces AIC as an appropriate procedure of statistical model identification.

2 Choice of k in the AIC seems to be flexible. Collett (1994) states that the choice $k = 3$ in the AIC is roughly equivalent to using a 5% significance level in judging the difference between the values of $-2\times$log(maximum likelihood) for two nested models which differ by one to three parameters. He recommends $k = 3$ for general use.

3 There are a variety of model selection indices similar in spirit to AIC. These are, going by name, BIC, Mallow's C_p, adjusted R^2, $R_a^2 = 1 - (1 - R^2)(n - 1)/(n - p)$, where p is the number of parameters in the least squares regression, and some others. These all adjust the goodness of fit of the model by penalizing for complexity of the model in terms of the number of parameters.

4 Efron (1998) cautions that the validity of the selected model through currently available methods may be doubtful in certain situations. He illustrates an example where a bootstrap simulation study certainly discourages confidence in the selected model. He and his student find that from 500 bootstrap sets of data there is only one match to the originally selected model. Further, only one variable in the originally selected model appears in more than half (295) of the bootstrap set based models.

5 Bottom line in model selection: Does it make sense!

Estimation and testing: fitting the Weibull model

The S function `survReg` fits the times T as log-failure times $Y = \log(T)$ to model (4.3)

$$Y = \beta_0^* + \underline{x}'\underline{\beta}^* + \sigma Z,$$

where Z has the standard extreme value distribution. Further, when we re-express Y as

$$Y = \underline{x}'\underline{\beta}^* + Z^* \,,$$

where $Z^* = \beta_0^* + \sigma Z$, we see this model is an accelerated failure time model. Here $Z^* \sim$ extreme value with location β_0^* and scale σ. The linear predictor given on page 85 is

$$\tilde{\mu} = -\log(\tilde{\lambda}) = \beta_0^* + \underline{x}'\underline{\beta}^* \tag{4.18}$$

with $\beta_0^* = -\log(\lambda)$ and $\underline{\beta}^* = -\sigma\underline{\beta}$, where the vector $\underline{\beta}$ denotes the coefficients in the Weibull hazard on page 84 and, $\sigma = 1/\alpha$, where α denotes the Weibull shape parameter. Let $\widehat{\beta}_0^*$, $\underline{\widehat{\beta}}^{*\prime}$, and $\widehat{\sigma}$ denote the MLE's of the parameters. Recall that the theory tells us MLE's are approximately normally distributed when the sample size n is large. To test $H_0 : \beta_j^* = \beta_j^{*0}$, $j = 1, \ldots, m$, use

$$\frac{\widehat{\beta}_j^* - \beta_j^{*0}}{\text{s.e.}(\widehat{\beta}_j^*)} \quad \overset{a}{\sim} \quad N(0,1) \quad \text{under } H_0.$$

An approximate $(1-\alpha)\times 100\%$ confidence interval for β_j^* is given by

$$\widehat{\beta}_j^* \pm \mathrm{z}_{\frac{\alpha}{2}}\, \text{s.e.}(\widehat{\beta}_j^*),$$

where $\mathrm{z}_{\frac{\alpha}{2}}$ is taken from the $N(0,1)$ table. Inferences concerning the intercept β_0^* follow analogously.

Notes:

1 It is common practice to construct $(1-\alpha)\times 100\%$ confidence intervals for the coefficients in the Weibull model by multiplying both endpoints by $-\widehat{\sigma}^{-1}$ and reversing their order. However, we suggest constructing confidence intervals using the bivariate delta method stated in Chapter 3.6 to obtain a more appropriate standard error for $\widehat{\beta}_j$. The reason is that the bivariate delta method takes into account the variability due to $\widehat{\sigma}$ as well

as $\widehat{\beta}_j^*$. The common approach does not, and hence, could seriously underestimate the standard error. The explicit expression for the variance of $\widehat{\beta}_1$ is as follows:

$$\widehat{\text{var}}(\widehat{\beta}_1) = \frac{1}{\widehat{\sigma}^2}\left(\text{var}(\widehat{\beta}_1^*) + \widehat{\beta}_1^{*2}\text{var}(\log(\widehat{\sigma})) - 2\widehat{\beta}_1^*\text{cov}(\widehat{\beta}_1^*, \log(\widehat{\sigma}))\right). \quad (4.19)$$

WHY! We use this expression to compute a 95% confidence interval for β_1 at the end of this chapter.

2 It is common practice to compute a $(1-\alpha)\times 100\%$ confidence interval for the true parameter value of λ by multiplying LCL and UCL for the intercept β_0^* by -1, then taking the $\exp(\cdot)$ of both endpoints, and then, reversing their order. This may end up with too wide a confidence interval as we show at the end of this chapter. Again we recommend the delta method to obtain the variance estimate of $\widehat{\lambda}$. By applying the delta method to $\widehat{\lambda} = \exp(-\widehat{\beta}_0^*)$, we obtain $\widehat{\text{var}}(\widehat{\lambda}) = \exp(-2\widehat{\beta}_0^*)\text{var}(\widehat{\beta}_0^*)$. WHY!

At the point $\underline{x} = \underline{x}_0$, the MLE of the $(p \times 100)$th percentile of the distribution of $Y = \log(T)$ is

$$\widehat{Y}_p = \widehat{\beta}_0^* + \underline{x}_0'\underline{\widehat{\beta}}^* + \widehat{\sigma}z_p = (1, \underline{x}_0', z_p)\begin{pmatrix} \widehat{\beta}_0^* \\ \underline{\widehat{\beta}}^* \\ \widehat{\sigma} \end{pmatrix},$$

where z_p is the $(p \times 100)$th percentile of the error distribution, which, in this case, is standard extreme value. The estimated variance of $\widehat{Y}_p$ is

$$\text{var}(\widehat{Y}_p) = (1, \underline{x}_0', z_p)\widehat{\Sigma}\begin{pmatrix} 1 \\ \underline{x}_0 \\ z_p \end{pmatrix}, \quad (4.20)$$

where $\widehat{\Sigma}$ is the estimated variance-covariance matrix of $\widehat{\beta}_0^*$, $\widehat{\beta}_1^*$, and $\widehat{\sigma}$. WHY! Then an approximate $(1-\alpha)\times 100\%$ confidence interval for the $(p \times 100)$th percentile of the log-failure time distribution is given by

$$\widehat{Y}_p \pm \text{z}_{\frac{\alpha}{2}}\,\text{s.e.}(\widehat{Y}_p),$$

where $\text{z}_{\frac{\alpha}{2}}$ is taken from the $N(0,1)$ table. These are referred to as the `uquantile` type in the S function `predict`. The MLE of t_p is $\exp(\widehat{Y}_p)$. To obtain confidence limits for t_p, take the exponential of the endpoints of the above confidence interval.

The function `predict`, a companion function to `survReg`, conveniently reports both the quantiles in time and the uquantiles in log(time) along with their respective s.e.'s. We often find the confidence intervals based on uquantiles are shorter than those based on quantiles. See, for example, the results at the end of this section.

Doing the analysis using S:

In S, we fit the model

$$Y = \log(T) = \beta_0^* + \beta_1^* x + \sigma Z,$$

where $Z \sim$ standard extreme value distribution. The $(p \times 100)$th percentile of the standard extreme (minimum) value distribution, Table 3.1, is

$$z_p = \log\big(-\log(1-p)\big).$$

The function `survReg` outputs the estimated variance-covariance matrix $\widehat{V}$ for the MLE's $\widehat{\beta}_0^*$, $\widehat{\beta}_1^*$, and $\widehat{\tau} = \log\widehat{\sigma}$. However, internally it computes $\widehat{\Sigma}$ to estimate the var$(\widehat{Y}_p)$.

The following is an S program along with modified output. The function `survReg` is used to fit a Weibull regression model. Then the 15th and 85th percentiles as well as the median failure time are estimated with corresponding standard errors. We also predict the failure time in hours at $x_0 = 2.480159$, which corresponds to the design temperature of 130^0C. Four plots of the estimated hazard and survivor functions are displayed in Figure 4.2. Three Q-Q plots are displayed in Figure 4.3, where intercept is $\widehat{\beta}_0^* + \widehat{\beta}_1^* x$ and slope is $\widehat{\sigma}$. Since there are three distinct values of x, we have three parallel lines. Lastly, the results are summarized.

```
> attach(motorette)
> weib.fit <- survReg(Surv(time,status)~x,dist="weibull")
> summary(weib.fit)
             Value Std. Error     z         p
(Intercept) -11.89      1.966 -6.05 1.45e-009
          x   9.04      0.906  9.98 1.94e-023
 Log(scale)  -1.02      0.220 -4.63 3.72e-006

> weib.fit$var # The estimated covariance matrix of the
               # coefficients and log(sigmahat).
             (Intercept)           x  Log(scale)
(Intercept)  3.86321759 -1.77877653  0.09543695
          x -1.77877653  0.82082391 -0.04119436
 Log(scale)  0.09543695 -0.04119436  0.04842333

> predict(weib.fit,newdata=list(x),se.fit=T,type="uquantile",
     p=c(0.15,0.5,0.85)) # newdata is required whenever
     # uquantile is used as a type whereas quantile uses the
     # regression variables as default. This returns the
     # estimated quantiles in log(t) along with standard
     # error as an option.
```

```
 # Estimated quantiles in log(hours) and standard errors in
 # parentheses. The output is edited because of redundancy.

x=2.256318    7.845713    8.369733     8.733489
            (0.1806513) (0.12339772) (0.1370423)
x=2.158895    6.965171    7.489190     7.852947
            (0.1445048) (0.08763456) (0.1189669)
x=2.027575    5.778259    6.302279     6.666035
            (0.1723232) (0.14887233) (0.1804767)

> predict(weib.fit,newdata=data.frame(x=2.480159),se.fit=T,
     type="uquantile",p=c(0.15,0.5,0.85)) # Estimated
     # quantiles in log(hours) at the new x value =
     # 2.480159; i.e., the design temperature of 130
     # degrees Celsius.

x=2.480159   9.868867    10.392887    10.756643
            (0.3444804) (0.3026464)  (0.2973887)

> sigmahat <- weib.fit$scale
> alphahat <- 1/sigmahat # estimate of shape
> coef <- weib.fit$coef
> lambdatildehat <- exp(- coef[1] - coef[2]*2.480159)
     # estimate of scale
> pweibull(25000,alphahat,1/lambdatildehat) # Computes the
     # estimated probability that a motorette failure time
     # is less than or equal to 25,000 hours. pweibull is
     # the Weibull distribution function in S.

 [1] 0.2783054 # estimated probability

> Shatq <- 1 - 0.2783054 # survival probability at 25,000
     # hours. About 72% of motorettes are still working
     # after 25,000 hours at x=2.480159; i.e., the design
     # temperature of 130 degrees Celsius.

> xl <- levels(factor(x)) # Creates levels out of the
                          # distinct x-values.
> ts.1 <- Surv(time[as.factor(x)==xl[1]],
                 status[as.factor(x)==xl[1]]) # The first
                                              # group of data
> ts.2 <- Surv(time[as.factor(x)==xl[2]],
                 status[as.factor(x)==xl[2]]) # The second
> ts.3 <- Surv(time[as.factor(x)==xl[3]],
```

```
                status[as.factor(x)==xl[3]]) # The third
> par(mfrow=c(2,2)) # divides a screen into 2 by 2 pieces.
> Svobj <- list(ts.1,ts.2,ts.3) # Surv object
> qq.weibreg(Svobj,weib.fit) # The first argument takes
      # a Surv object and the second a survReg object.
      # Produces a Weibull Q-Q plot.
> qq.loglogisreg(Svobj,loglogistic.fit) # log-logistic
                                        # Q-Q plot
> qq.lognormreg(Svobj,lognormal.fit) # log-normal Q-Q plot
> detach()
```

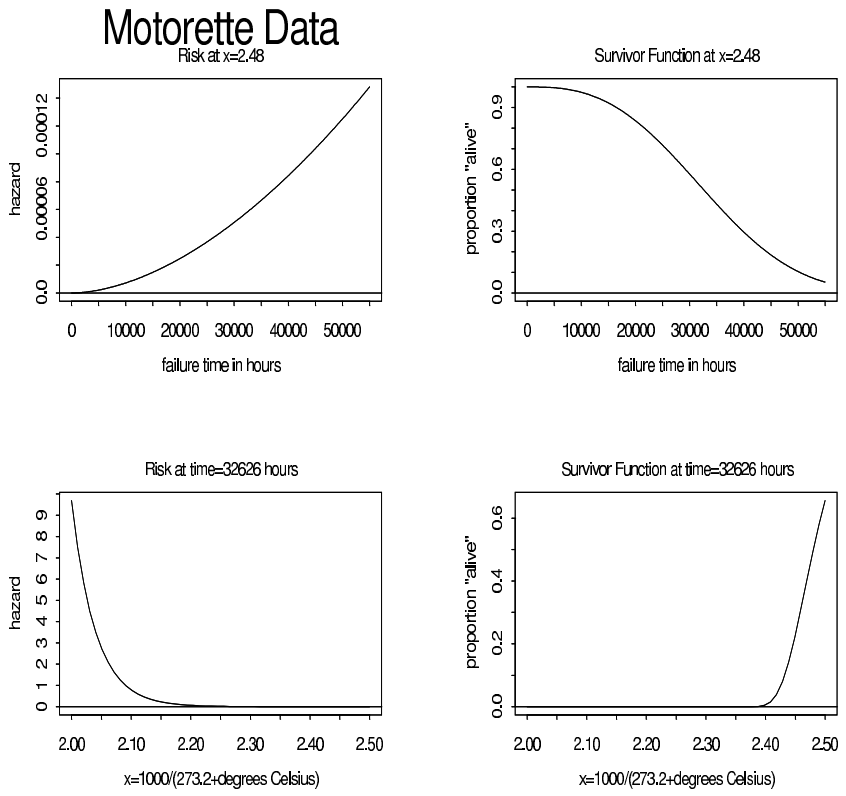


Figure 4.2 *Weibull hazard and survival functions fit to motorette data.*

Results:

- From `summary(weib.fit)`, we learn that $\widehat{\sigma} = \exp(-1.02) = .3605949$, and $\widehat{\widetilde{\mu}} = -\log(\widehat{\widetilde{\lambda}}) = \widehat{\beta}_0^* + \widehat{\beta}_1^* x = -11.89 + 9.04x$.

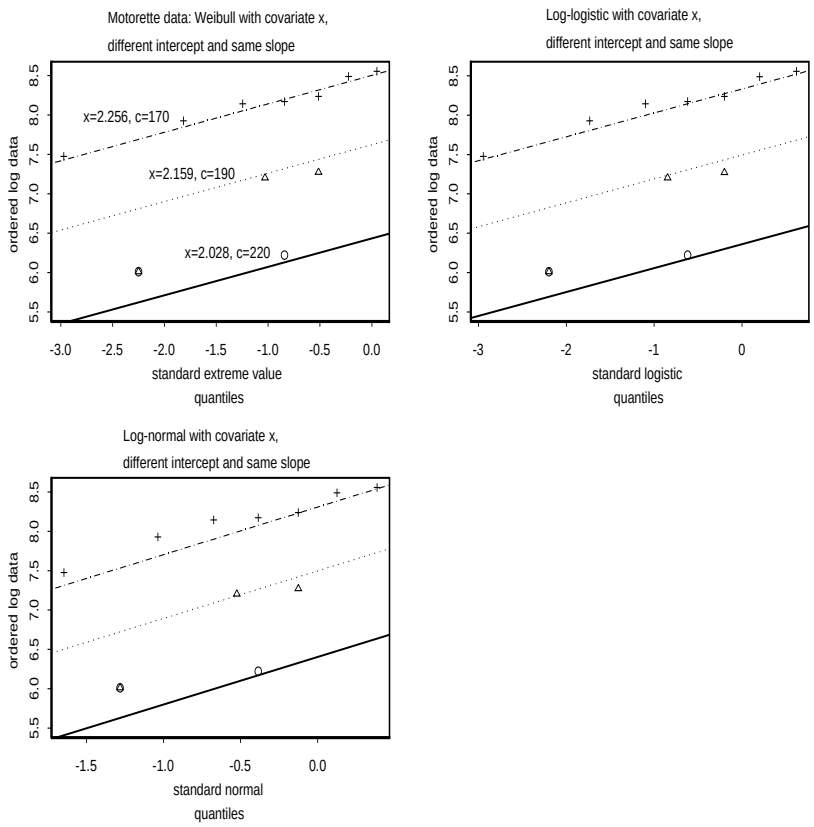


Figure 4.3 *Weibull, log-logistic, and log-normal Q-Q plots of the motorette data. Lines constructed with MLE's.*

Thus, we obtain $\widehat{\alpha} = \frac{1}{.3605949} = 2.773195$ and $\widehat{\widehat{\lambda}} = \exp(11.89 - 9.04 \times 2.480159) = 0.0000267056$ at $x = 2.480159$. Note also that both the intercept and covariate x are highly significant with p-values 1.45×10^{-9} and 1.94×10^{-23}, respectively.

- It follows from Chapter 4.2 that the estimated hazard function is

$$\widehat{h}(t|x) = \frac{1}{\hat{\sigma}} \cdot t^{\frac{1}{\hat{\sigma}}-1} \cdot (\exp(-\widehat{\widehat{\mu}}))^{\frac{1}{\hat{\sigma}}}$$

and the estimated survivor function is

$$\widehat{S}(t|x) = \exp\Big\{ - \Big(\exp(-\widehat{\widehat{\mu}})t \Big)^{\frac{1}{\hat{\sigma}}} \Big\}.$$

- The point estimate of β_1, $\widehat{\beta}_1$, is $-\widehat{\sigma}^{-1}\widehat{\beta}_1^*$. A 95% C.I. for β_1 based on the delta method is given by $[-37.84342, -12.29594]$. Whereas the one based

on the common approach is given by

$$[-\hat{\sigma}^{-1}(10.82), -\hat{\sigma}^{-1}(7.26)] = [-29.92, -20.09],$$

where $\hat{\sigma} = .3605949$ and the 95% C.I. for β_1^* is $[7.26, 10.81] = [9.04 - 1.96 \times .906,\ 9.04 + 1.96 \times .906]$. It is clear that the latter interval is much shorter than the former as it ignores the variability of $\widehat{\sigma}$.

- A 95% C.I. for λ based on the delta method is given by $[-416023.7, 707626.3]$. We see this includes negative values, which is not appropriate because λ is restricted to be positive. Therefore, we report the truncated interval $[0, 707626.3]$. The one based on the common approach is given by

$$[\exp(8.04), \exp(15.74)] = [3102.61,\ \ 6851649.6],$$

where the 95% C.I. for β_0^* is [-11.89 − 1.96 × 1.966, -11.89 + 1.96 × 1.966] = [-15.74, -8.04]. Although the common approach ends up with an unreasonably wide confidence interval compared to the one based on the delta method, this approach always yields limits within the legal range of λ.

- At $x = 2.480159$, the design temperature of 130^0C, the estimated 15th, 50th, and 85th percentiles in log(hours) and hours, respectively based on `uquantile` and `quantile`, along with their corresponding 90% C.I.'s in hours are reported in the following table.

type	percentile	Estimate	Std.Err	90% LCL	90% UCL
`uquantile`	15	9.868867	0.3444804	10962.07	34048.36
	50	10.392887	0.3026464	19831.64	53677.02
	85	10.756643	0.2973887	28780.08	76561.33
`quantile`	15	19319.44	6655.168	9937.174	37560.17
	50	32626.72	9874.361	19668.762	54121.65
	85	46940.83	13959.673	28636.931	76944.21

The 90% C.I.'s based on `uquantile`, exp(estimate ± 1.645 × std.err), are shorter than those based on quantile at each x value. However, we also suspect there is a minor bug in `predict` in that there appears to be a discrepancy between the standard error estimate for the 15th percentile resulting from `uquantile` and ours based on the delta method which follows. The other two standard error estimates are arbitrarily close to ours. Our standard error estimates are .3174246, .2982668, and .3011561 for the 15th, 50th, and 85th percentiles, respectively. Applying the trivariate delta method, we obtain the following expression:

$$\begin{aligned} \widehat{\text{var}}(\widehat{y_p}) &= \text{var}(\widehat{\beta_0^*}) + \text{var}(\widehat{\beta_1^*})x_0^2 + z_p^2\widehat{\sigma}^2\text{var}(\log\widehat{\sigma}) \\ &+ 2x_0\text{cov}(\widehat{\beta_0^*}, \widehat{\beta_1^*}) + 2z_p\widehat{\sigma}\text{cov}(\widehat{\beta_0^*}, \log\widehat{\sigma}) + 2x_0 z_p\widehat{\sigma}\text{cov}(\widehat{\beta_1^*}, \log\widehat{\sigma}). \end{aligned} \tag{4.21}$$

WHY!

- At the design temperature 130^0C, by 25,000 hours about 28% of the motorettes have failed. That is, after 25,000 hours, about 72% are still working.
- As $\hat{\alpha} = \frac{1}{\hat{\sigma}} = \frac{1}{.3605949} = 2.773195$, then for fixed x the hazard function increases as time increases. The upper two graphs in Figure 4.2 display estimated hazard and survivor functions. The covariate x is fixed at 2.480159 which corresponds to the design temperature 130^0C.
- The estimated coefficient $\hat{\beta}_1 = -\frac{1}{\hat{\sigma}}\hat{\beta}_1^* = -\frac{1}{.3605949}(9.04) = -25.06968 < 0$. Thus, for time fixed, as x increases, the hazard decreases and survival increases. The lower two graphs in Figure 4.2 display these estimated functions when time is fixed at 32,626 hours.
- For $x_1 < x_2$,
$$\frac{h(t|x_2)}{h(t|x_1)} = \exp((x_2 - x_1)(-25.06968)).$$
For example, for $x = 2.1$ and 2.2,
$$\frac{h(t|2.2)}{h(t|2.1)} = \exp(.1(-25.06968)) = .08151502.$$
Thus, for .1 unit increase in x, the hazard becomes about 8.2% of the hazard before the increase. In terms of Celsius temperature, for 21.645 degree decrease from 202.9905^0C to 181.3455^0C, the hazard becomes about 8.2% of the hazard before the decrease.
- The Q-Q plots in Figure 4.3 show that the Weibull fit looks slightly better than the log-logistic fit at the temperature 170^0C, but overall they are the same. On the other hand, the Weibull fit looks noticeably better than the log-normal fit at the temperature 170^0C and is about the same at the other two temperatures. This result coincides with our finding from AIC in Table 4.2; that is, among these three accelerated failure time models, the Weibull best describes the motorette data.

CHAPTER 5

The Cox Proportional Hazards Model

In this chapter we discuss some features of a prognostic factor analysis based on the Cox proportional hazards (PH) model. We present a detailed analysis of the CNS lymphoma data.

Example: CNS lymphoma data

The data result from an observational clinical study conducted at Oregon Health Sciences University (OHSU). The findings from this study are summarized in Dahlborg *et al.* (1996). Fifty-eight non-AIDS patients with central nervous system (CNS) lymphoma were treated at OHSU from January 1982 through March of 1992. Group 1 patients (n=19) received cranial radiation prior to referral for blood-brain barrier disruption (BBBD) chemotherapy treatment; Group 0 (n=39) received, as their initial treatment, the BBBD chemotherapy treatment. Radiographic tumor response and survival were evaluated. Table 5.1 describes the variables obtained for each patient.

The primary endpoint of interest here is survival time (in years) from first blood brain barrier disruption (BBBD) to death (B3TODEATH). Some questions of interest are:

1 Is there a difference in survival between the two groups (prior radiation, no radiation prior to first BBBD)?

2 Do any subsets of available covariates help explain this survival time? For example, does age at time of first treatment and/or gender increase or decrease the hazard of death; hence, decrease or increase the probability of survival; and hence, decrease or increase mean or median survival time?

3 Is there a dependence of the difference in survival between the groups on any subset of the available covariates?

Objectives of this chapter:

After studying Chapter 5, the student should:

1 Know and understand the definition of a Cox PH model including the assumptions.

2 Know how to use the S function `coxph` to fit data to a Cox PH model.

3 Know how to use the S function `stepAIC` along with `coxph` to identify an appropriate model.

4 Know how to use the **stratified Cox PH model**.

5 Know how to interpret the estimated β coefficients with respect to hazard and other features of the distribution.

6 Understand how to interpret the estimated hazards ratio HR. That is, understand its usefulness as a measure of effect that describes the relationship between the predictor variable(s) and time to failure. Further, the HR can be used to examine the relative likelihood of survival.

We first plot the two Kaplan-Meier (K-M) survivor curves using S. Figure 5.1 displays a difference in survival between the two groups. The higher K-M curve for the no prior radiation group suggests that this group has a higher chance of long term survival. The following S output confirms this. The S function `survdiff` yields a **log-rank** test statistic value of 9.5 which confirms this difference with an approximate p-value of .002. Further note the estimated mean and median given in the output from the S function `survfit`. Much of the output has been deleted where not needed for discussion. The CNS data is stored in a data frame named cns2.

```
> cns2.fit0 <- survfit(Surv(B3TODEATH,STATUS)~GROUP,data=cns2,
      type="kaplan-meier")
> plot(cns2.fit0,lwd=3,col=1,type="l",lty=c(1,3),cex=2,
    lab=c(10,10,7),xlab="Survival Time in Years from
     First BBBD",ylab="Percent Surviving",yscale=100)
> text(6,1,"Primary CNS Lymphoma Patients",lwd=3)
> legend(3,0.8,type="l",lty=c(1,3,0),c("no radiation prior
    to BBBD (n=39)","radiation prior to BBBD (n=19)",
       "+ = patient is censored"),col=1)

> survdiff(Surv(B3TODEATH,STATUS)~GROUP,data=cns2)
          N Observed Expected (O-E)^2/E (O-E)^2/V
 GROUP=0 39       19    26.91      2.32      9.52
 GROUP=1 19       17     9.09      6.87      9.52
 Chisq= 9.5  on 1 degrees of freedom, p= 0.00203

> cns2.fit0

          n events mean se(mean) median 0.95LCL 0.95UCL
 GROUP=0 39     19 5.33    0.973  3.917   1.917      NA
 GROUP=1 19     17 1.57    0.513  0.729   0.604    2.48
```

Since the two survival curves are significantly different, we assess the factors that may play a role in survival and in this difference in survival duration. Recall from expression (1.5) the hazard (risk) function $h(t)\Delta t$ is approximately

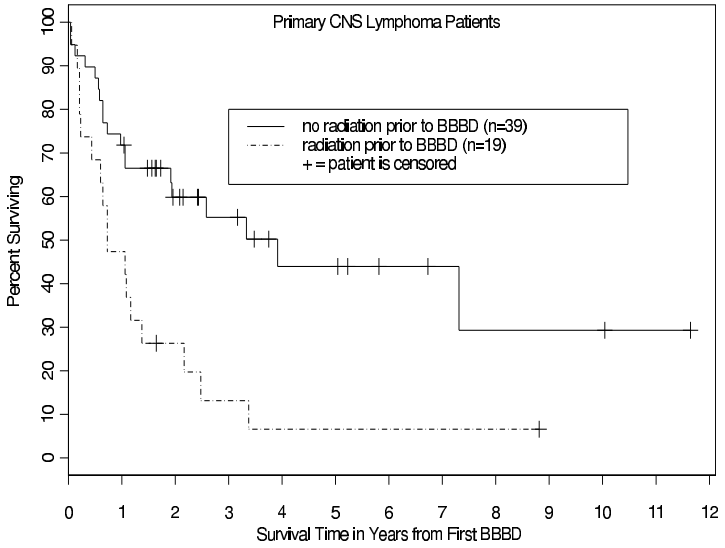


Figure 5.1 *Kaplan-Meier survivor curves.*

the conditional probability of failure in the (small) interval from t to $t + \Delta t$ given survival until time t. Here t is the length of time a patient lives from the point of his/her first BBBD. **Assuming that the baseline hazard function is the same for all patients in the study, a Cox PH model** seems appropriate. That is, we model the hazard rate as a function of the covariates $\underline{x}$. Recall from Chapter 4.3 that the **hazard function** has the form

$$\begin{aligned} h(t|\underline{x}) = h_0(t) \cdot \exp(\underline{x}'\underline{\beta}) &= h_0(t) \cdot \exp\left(\beta_1 x^{(1)} + \beta_2 x^{(2)} + \cdots + \beta_m x^{(m)}\right) \\ &= h_0(t) \cdot \exp\left(\beta_1 x^{(1)}\right) \times \exp\left(\beta_2 x^{(2)}\right) \cdots \times \exp\left(\beta_m x^{(m)}\right), \end{aligned}$$

where $h_0(t)$ is an unspecified baseline hazard function free of the covariates $\underline{x}$. The covariates act multiplicatively on the hazard. At two different points $\underline{x}_1$ and $\underline{x}_2$, the proportion

$$\begin{aligned} \frac{h(t|\underline{x}_1)}{h(t|\underline{x}_2)} &= \frac{\exp(\underline{x}_1'\underline{\beta})}{\exp(\underline{x}_2'\underline{\beta})} \\ &= \frac{\exp\left(\beta_1 x_1^{(1)}\right) \times \exp\left(\beta_2 x_1^{(2)}\right) \times \cdots \times \exp\left(\beta_m x_1^{(m)}\right)}{\exp\left(\beta_1 x_2^{(1)}\right) \times \exp\left(\beta_2 x_2^{(2)}\right) \times \cdots \times \exp\left(\beta_m x_2^{(m)}\right)} \end{aligned}$$

is constant with respect to time t. As we are interested in estimating the coefficients $\underline{\beta}$, the baseline hazard is really a nuisance parameter. Through the **partial likelihood** (Cox, 1975) we obtain estimates of the coefficients $\underline{\beta}$ without regard to the baseline hazard $h_0(t)$. Note that in the parametric regression setting of Chapter 4, we specify the form of this function since we

must specify a distribution for the target variable T. Remember the hazard function completely specifies the distribution of T; but the power of the PH model is that it provides a fairly wide family of distributions by allowing the baseline hazard $h_0(t)$ to be arbitrary. The S function `coxph` implements Cox's partial likelihood function. In Chapter 6.3 we offer a heuristic derivation of Cox's partial likelihood.

5.1 AIC procedure for variable selection

Akaike's information criterion (AIC) for the Cox PH model:

We revisit AIC in the context of the Cox PH regression model. Comparisons between a number of possible models can be made on the basis of the statistic

$$\text{AIC} = -2 \times \log(\text{maximum likelihood}) + 2 \times b, \tag{5.1}$$

where b is the number of β coefficients in each model under consideration. The maximum likelihood is replaced by the maximum partial likelihood. The smaller the AIC value the better the model is.

We apply an automated model selection procedure via an S function `stepAIC` included in MASS, a collection of functions and data sets from *Modern Applied Statistics with S* by Venables and Ripley (2002). Otherwise, it would be too tedious because of many steps involved.

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 illustrate how to use `stepAIC` together with LRT to select a best model. We fit the CNS data to a Cox PH model. In Chapter 1.2 we established the relationship that the smaller the risk, the larger the probability of survival, and hence the greater mean survival.

The estimates from fitting a Cox PH model are interpreted as follows:

- A positive coefficient increases the risk and thus decreases the expected (average) survival time.
- A negative coefficient decreases the risk and thus increases the expected survival time.
- The ratio of the estimated risk functions for the two groups can be used to examine the likelihood of Group 0's (no prior radiation) survival time being longer than Group 1's (with prior radiation).

Table 5.1: *The variables in the CNS lymphoma 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
5. STATUS: 0=alive ; 1=dead
6. DXTOB3: time from diagnosis to 1st BBBD in years
7. DXTODeath: time from diagnosis to death in years
8. B3TODeath: time from 1st BBBD to death in years
9. KPS.PRE.: Karnofsky performance score before 1st BBBD, numerical value $0 - 100$
10. LESSING: Lesions: single=0 ; multiple=1
11. LESDEEP: Lesions: superficial=0 ; deep=1
12. LESSUP: Lesions: supra=0 ; infra=1 ; both=2
13. PROC: Procedure: subtotal resection=1 ; biopsy=2 ; other=3
14. RAD4000: Radiation > 4000: no=0 ; yes=1
15. CHEMOPRIOR: no=0 ; yes=1
16. RESPONSE: Tumor response to chemotherapy - complete=1; partial=2; blanks represent missing data

The two covariates LESSUP and PROC are categorical. Each has three levels. The S function `factor` creates indicator variables. Also, the variable AGE60 is defined as AGE60 = 1 if AGE ≤ 60 and = 0 otherwise. The S code `> cns2$AGE60 <- as.integer(cns2$AGE<=60)` creates this variable and stores it in the cns2 data frame. We implement the S functions `stepAIC` and `coxph` to select appropriate variables according to the AIC criterion based on the proportional hazards model.

Let us consider the two-way interaction model, which can be easily incorporated in the `stepAIC`. Three-way or four-way interaction models can be considered but the interpretation of an interaction effect, if any, is not easy. The initial model contains all 11 variables without interactions. The scope is up to two-way interaction models. These are listed in the S code under Step I that follows. The direction is "both." The AIC for each step is reported in Table 5.2. The first AIC value is based on the initial model of 11 variables without interactions. "+" means that term was added at that step and "-" means that term was removed at that step. The final model retains the following variables: KPS.PRE., GROUP, SEX, AGE60, LESSING, CHEMOPRIOR, SEX:AGE60, AGE60:LESSING, and GROUP:AGE60.

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
```

Table 5.2: *Stepwise model path for two-way interaction model on the CNS lymphoma data*

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. 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]
[1] 13.42442
> 1 - pchisq(13.42442,4)
[1] 0.009377846 # Retain the model selected by stepAIC
```

Now we begin the process of one variable at a time reduction. This can be based on either the p-value method or the LRT. Asymptotically they are equivalent. Since the variable GROUP:AGE60 has a moderately large p-value, .130, we delete it. The following LRT test shows no evidence against the reduced model (p-value = .138) and so we adopt the reduced model.

```
> cns2.coxint3 <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP
     +SEX+AGE60+LESSING+CHEMOPRIOR+SEX:AGE60+AGE60:LESSING)
        # Without GROUP:AGE60
> -2*cns2.coxint3$loglik[2] + 2*cns2.coxint1$loglik[2]
[1] 2.194949
> 1 - pchisq(2.194949,1)
[1] 0.1384638 # Selects the reduced model

> 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
```

From this point on we use the p-value method to eliminate one term at a time. 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]
[1] 8.843838
> 1 - pchisq(8.843838,4)
[1] 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
```

This model is substantially different from that reported in Dahlborg *et al.* (1996). We go through model diagnostics in Chapter 6 to confirm that the model does fit the data.

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.

We reexamine the CNS Lymphoma data. The AIC for each model without interaction terms is reported in Table 5.3. The first AIC is based on the initial model including all the variables. The final model is selected by applying backward elimination procedure with the range from the full model with all the variables to the smallest reduced model with intercept only. It retains these four variables: KPS.PRE., GROUP, SEX, and CHEMOPRIOR.

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
```

Table 5.3: *Stepwise model path for the main effects model*

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 II: LRT to further reduce

The following output shows p-values corresponding to variables selected by `stepAIC`. 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. Therefore, we choose the interaction model cns2.coxint6 on page 110 to discuss.

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 $-.0307$ 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
$$\begin{aligned}\frac{\exp(1.1592)\exp(-0.0307\times(k+10))}{\exp(0)\exp(-.0307\times k)} &= \frac{\exp(1.1592)\exp(-0.0307\times 10)}{\exp(0)} \\ &= 3.187\times 0.7357 = 2.345,\end{aligned}$$
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% 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 5.2, 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.

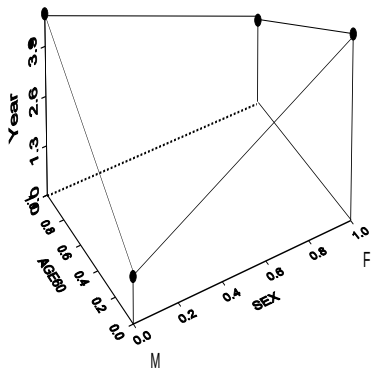


Figure 5.2 *Interaction between* SEX *and* AGE60.

In Figure 5.3, 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, KPS.PRE. = 80 and obtain the summary of the adjusted quantiles and means 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 5.3 displays both ordinal and disordinal interactions. The survival

curve for females who are younger than 60 years never steps down 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.

If one sets the covariate KPS.PRE. equal to different values, one can study its relationship to the interaction as well as its effect on the various estimated quantiles of the survival distribution. However, this is tedious. 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 nonparametric methodology is presented in Chapter 8 of our book.

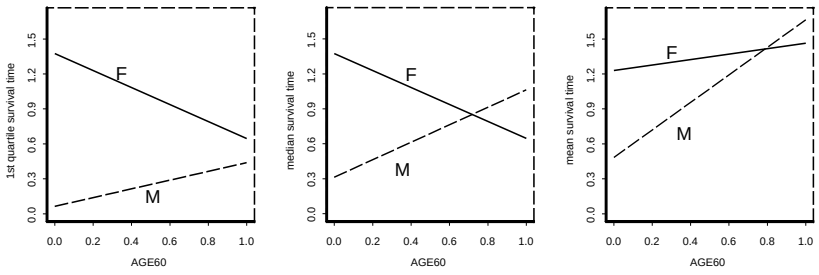


Figure 5.3 *Interaction between* SEX *and* AGE60 *adjusted for* KPS.PRE. *and* GROUP *via* `coxph` *and then evaluated at* GROUP = 1 *and* KPS.PRE. = 80.

5.2 Stratified Cox PH regression

We stratify on a categorical variable such as group, gender, and exposure still fitting the other covariates. We do this to obtain nonparametric estimated survival curves for the different levels having adjusted for the other covariates. We then plot the curves to view the estimate of the categorical effect, after adjusting for the effects of the other covariates. If the curves cross or are nonproportional, this implies the existence of the interaction effect unexplained in the model. Then look for appropriate interaction term(s) to include in the model, or stay with the stratified model. If the curves are proportional, this indicates that the interaction effect is well explained by the model you have identified and it supports the Cox PH model. Then use the Cox PH model without the stratification. The disadvantage when we stratify, and the PH assumption is satisfied, is that we cannot obtain an estimated coefficient of the categorical variable effect.

We now apply this procedure to our final model for CNS data. In the following S program we first stratify on the GROUP variable still fitting KPS.PRE., SEX, AGE60, and SEX:AGE60 as covariates. Next, we repeat this procedure for SEX. Again, the disadvantage here is that we cannot obtain an estimated coefficient of the group and sex effects, respectively.

```
> attach(cns2)
> cns2.coxint7 <- coxph(Surv(B3TODEATH,STATUS)~strata(GROUP)
       +KPS.PRE.+SEX+AGE60+SEX:AGE60)
> cns2.coxint7
             coef exp(coef) se(coef)     z      p
 KPS.PRE. -0.0326     0.968   0.0108 -3.03 0.0025
      SEX -2.2028     0.110   0.7195 -3.06 0.0022
    AGE60 -1.1278     0.324   0.4778 -2.36 0.0180
SEX:AGE60  2.2576     9.560   0.8785  2.57 0.0100

Likelihood ratio test=20.3  on 4 df, p=0.000433  n= 58

> cns2.coxint8 <- coxph(Surv(B3TODEATH,STATUS)~strata(SEX)
       +KPS.PRE.+GROUP+AGE60+SEX:AGE60)
> cns2.coxint8
            coef exp(coef) se(coef)     z      p
 KPS.PRE. -0.033     0.968   0.0104 -3.19 0.0014
    GROUP  1.178     3.247   0.3829  3.08 0.0021
    AGE60 -0.994     0.370   0.4552 -2.18 0.0290
SEX:AGE60  2.244     9.427   0.8791  2.55 0.0110

Likelihood ratio test=27  on 4 df, p=0.0000199  n= 58
```

```
 # The following gives plots of survival curves resulting from
 # stratified Cox PH models to detect any pattern.
 # Figure 5.4: upper part.
> par(mfrow=c(2,2))
> survfit.int7 <- survfit(cns2.coxint7)
> plot(survfit.int7,col=1,lty=3:4,lwd=2,cex=3,label=c(10,10,7),
   xlab="Survival time in years from first BBBD",
     ylab="Percent alive",yscale=100)
> legend(3.0,.92,c("group=0","group=1"),lty=3:4,lwd=2)
> survfit.int8 <- survfit(cns2.coxint8)
> plot(survfit.int8,col=1,lty=3:4,lwd=2,cex=3,label=c(10,10,7),
   xlab="Survival time in years from first BBBD",
     ylab="Percent alive",yscale=100)
> legend(3.8,.6,c("male","female"),lty=3:4,lwd=2)
```

For the Weibull regression model, recall (4.5) the log of the cumulative hazard function is linear in $\log(t)$. In general, when we look at the Cox PH model as well as the Weibull model, the plot of $H(t)$ against t on a log-log scale can be very informative. In the `plot` function, the optional function "`fun="cloglog"`" takes the `survfit` object and plots $H(t)$ against t on a log-log scale.

The following S code plots cumulative hazard functions against t, on a log-log scale, resulting from stratified Cox PH models to detect a nonproportional hazards trend for the SEX and GROUP variables.

```
 # Figure 5.4: lower part.
> plot(survfit.int7,fun="cloglog",col=1,lty=3:4,label=c(10,10,7),
   lwd=2,xlab="time in years from first BBBD",
     ylab="log-log cumulative hazard")
> legend(0.05,.8,c("group=0","group=1"),lwd=2,lty=3:4)
> plot(survfit.int8,fun="cloglog",col=1,lty=3:4,label=c(10,10,7),
   lwd=2,xlab="time in years from first BBBD",
     ylab="log-log cumulative hazard")
> legend(0.05,.8,c("male","female"),lwd=2,lty=3:4)
> detach()
```

Discussion

- Figure 5.4 shows clear differences between the two groups and between the males and females, respectively. Further, for both GROUP and SEX, the two curves are proportional. This supports the Cox PH model.
- Stratification doesn't change the p-values of the variables in the model cns2.coxint6. The estimated coefficients are very close as well. That is, the model cns2.coxint6 explains all the interaction among the covariates.

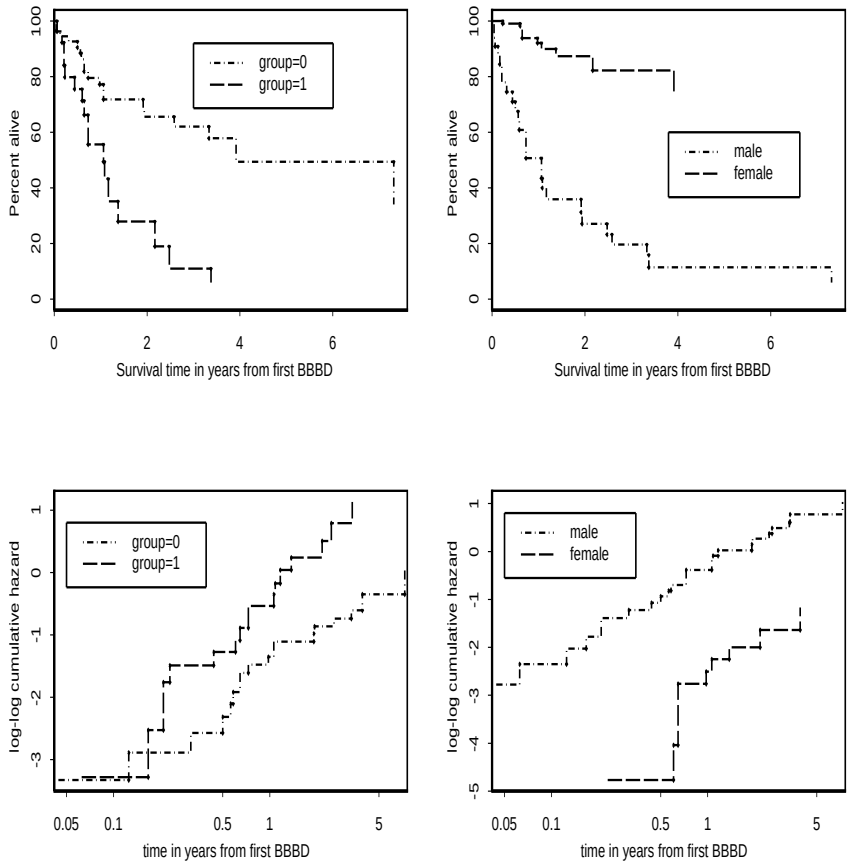


Figure 5.4 *Stratified survivor and log-log cumulative hazards plots to check for PH assumption.*

Remarks:

The Cox PH model formula says that the hazard at time t is the product of two quantities $h_0(t)$, an unspecified baseline hazard function, and $\exp(\sum_{j=1}^{m} \beta_j x^{(j)})$. The key features of the PH assumption are that

1 $h_0(t)$ is a function of t, but does not involve the covariates $x^{(j)}$.

2 $\exp(\sum_{j=1}^{m} \beta_j x^{(j)})$ involves the covariates $x^{(j)}$, but does not involve t.

These two key features imply the HR must then be constant with respect to

time t. We now provide an example of a situation where the PH assumption is violated.

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? To answer this note that when a patient undergoes serious surgery, as when removing a cancerous tumor, there is usually a high risk for complications from surgery or perhaps even death early in the recovery process, and once the patient gets past this early critical period, the benefits of surgery, if any, can be observed.

Thus, in a study that compares surgery to no surgery, we might expect to see hazard functions for each group that appear in Figure 5.5. Notice that these two functions cross at about three days, and that prior to three days the hazard for the surgery group is higher than the hazard for the no surgery group. Whereas, after three days, we have the reverse. For example, looking at the graph more closely, we can see that at two days, when $t = 2$, the HR of no surgery $(E = 1)$ to surgery $(E = 0)$ patients yields a value less than one. In contrast, at $t = 5$ days, the HR is greater than one. Thus, if the description of the hazard function for each group is accurate, the hazard ratio is not constant over time as HR is some number less than one before three days and greater than one after three days. Hence, the PH assumption is violated as the HR does vary with time. **The general rule is that if the hazard functions**

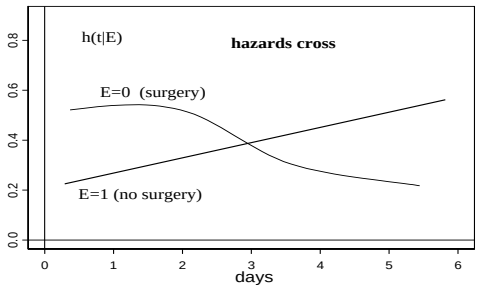


Figure 5.5 *Hazards crossing over time.*

cross over time, the PH assumption is violated. If the Cox PH model is inappropriate, there are several options available for the analysis:

- 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. Or, if there are other covariates in the model, use a Cox model stratified on E.
- 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 of our book.
- use the **censored regression quantile** approach, presented in Chapter 8 of our book, which allows crossover effects. This approach is nonparametric and is free of the PH assumption for its validity.

CHAPTER 6

Model Checking: Data Diagnostics

Objectives of this chapter:

After studying Chapter 6, the student should:

1 Know and understand the definition of **model deviance**:

(a) likelihood of fitted model

(b) likelihood of saturated model

(c) deviance residual.

2 Be familiar with the term **hierarchical models**.

3 Know the definition of **partial deviance**, its relationship to the likelihood ratio test statistic, and how we use it to reduce models and test for overall model adequacy.

4 Know how to interpret the measure **dfbeta**.

5 Know that the S function `survReg` along with companion function `resid` provides the `deviance` residuals, `dfbeta`, and `dfbetas`.

6 Be familiar with Cox's **partial likelihood** function.

7 Be familiar with and how to use the following residuals to assess the various proportional hazards model assumptions:

(a) *Cox-Snell residuals*

(b) *Martingale residuals*

(c) *Deviance residuals*

(d) *Schoenfeld residuals*

(e) *Scaled Schoenfeld residuals*

(f) *dfbetas.*

8 Be familiar with the S functions `coxph` and `cox.zph` and which residuals these functions provide.

6.1 Basic graphical methods

When searching for a parametric model that fits the data well, we use graphical displays to check the model's appropriateness; that is, the goodness of fit. Miller (1981, page 164) points out that "the human eye can distinguish well between a straight line and a curve." We quote Miller's basic principle as it should guide the method of plotting.

> **Basic principle:**
>
> Select the scales of the coordinate axes so that if the model holds, a plot of the data resembles a straight line, and if the model fails, a plot resembles a curve.

The construction of the Q-Q plot (page 55) for those log-transformed distributions, which are members of the location and scale family of distributions, follows this basic principle. The linear relationships summarized in Table 3.1, page 55, guided this construction. Some authors, including Miller, prefer to plot the uncensored points (y_i, z_i), $i = 1, \cdots, r \leq n$. This plot is commonly called a **probability plot**. We prefer the convention of placing the log data y_i on the vertical axis and the standard quantiles z_i on the horizontal axis; hence, the Q-Q plot.

The S function `survReg` only fits models for log-time distributions belonging to the location and scale family. For this reason we have ignored the gamma model until now. A Q-Q plot is still an effective graphical device for nonmembers of the location and scale family. For these cases, we plot the ordered uncensored times t_i against the corresponding quantiles q_i from the distribution of interest. If the model is appropriate, the points should lie very close to the 45^o-line through the origin $(0, 0)$. We compute and plot the quantiles based on the K-M estimates against the quantiles based on the parametric assumptions. That is, for each uncensored t_i, compute $\widehat{p}_i = 1 - \widehat{S}(t_i)$, where $\widehat{S}(t_i)$ denotes the K-M estimate of survival probability at time t_i. Then, with this set of probabilities, compute the corresponding quantiles q_i from the assumed distribution with MLE's used for the parameter values. Finally, plot the pairs (q_i, t_i). Note that $\widehat{p}_i = 1 - \widehat{S}(t_i) = 1 - \widehat{S}_{\text{model}}(q_i)$. To compute the MLE's for the unknown parameters in S, the two functions available are `nlmin` and `nlminb`. As these functions find a local minimum, we use these functions to minimize $(-1)\times$the log-likelihood function. For our example, we draw the Q-Q plot for the AML data fit to a gamma model. In this problem, we must use `nlminb` since the gamma has bound-constrained parameters; that is, $k > 0$ and $\lambda > 0$, corresponding to shape and scale, respectively. The function `qq.gamma` gives the Q-Q plot for data fit to a gamma. See Figure 6.1.

```
> attach(aml)
 # Q-Q plot for maintained group
> weeks.1 <- weeks[group==1]
> status.1 <- status[group==1]
```

```
> weeks1 <- list(weeks.1)
> status1 <- list(status.1)
> qq.gamma(Surv(weeks.1,status.1),weeks1,status1)
 # The 2nd and 3rd arguments must be list objects.
    shape      rate
 1.268666 0.0223737  #MLE's
 # Q-Q plot for nonmaintained group
> weeks.0 <- weeks[group == 0]
> status.0 <- status[group == 0]
> weeks0 <- list(weeks.0)
> status0 <- list(status.0)
> qq.gamma(Surv(weeks.0,status.0),weeks0,status0)
    shape       rate
 1.987217 0.08799075  # MLE'S
> detach()
```

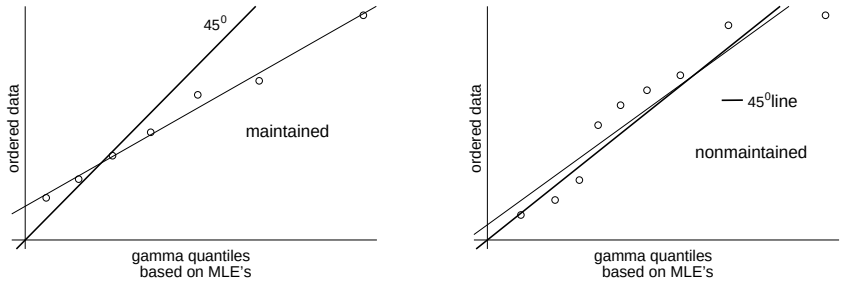


Figure 6.1 *Q-Q plot for AML data fit to gamma model. MLE's used for parameter values. Points are fit to least squares line.*

It's important to draw the 45^o-line. For without the comparison, the least squares line fitted only to uncensored times would have led us to believe the gamma model fit the maintained group well. But this is quite the contrary. The fit is very poor in the upper tail. The estimated gamma quantiles q_i are markedly larger than their corresponding sample quantiles t_i. One reason for this over-fit is the MLE's are greatly influenced by the presence of the one extreme value 161+. It is clear from the previous Weibull, log-logistic, and log-normal Q-Q plots (Figure 3.13, page 77), the log-logistic is a much better choice to model the AML maintained group. Notice the gamma Q-Q plot for this group has a similar pattern to the Weibull Q-Q plot. In contrast, the gamma seems to fit the nonmaintained group quite well. There are no extreme values in this group.

For the two-sample problem, let $x = 1$ and $x = 0$ represent the two groups. To check the validity of the Cox PH model, recall from Chapter 4.3 that

$h(t|1) = \exp(\beta)h(t|0)$, where $\exp(\beta)$ is constant with respect to time. This implies $S(t|1) = (S(t|0))^{\exp(\beta)}$ or $\log S(t|1) = \exp(\beta) \log S(t|0)$. These graphs are displayed in Figure 6.2. The plots of the empirical quantities constructed

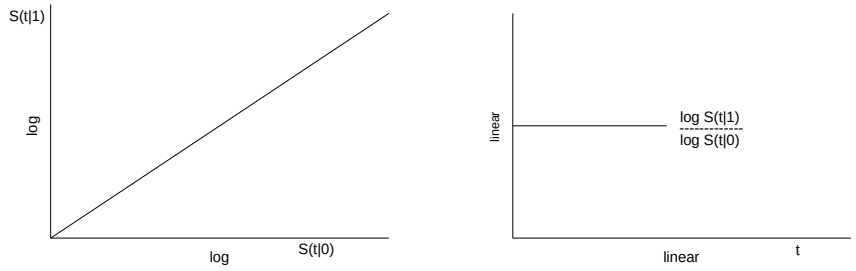


Figure 6.2 *Graph of cumulative hazards ratio.*

with the K-M estimate for each group should reflect the foregoing relationships if the PH assumption is satisfied.

Equivalently, we can plot the kernel estimates of hazard (2.11) for each group on the same plot. To validate the PH assumption a plot of the ratio of smoothed hazards should be roughly constant over the follow-up time. See Figure 2.7, page 43. It is clear the AML data violate the PH assumption.

To check for a shift by translation, calculate the K-M estimate of survival for each group separately and plot. The curves should be vertically parallel. For example, as the log-gamma is a location family, this plot is useful. An example is displayed in Figure 6.3.

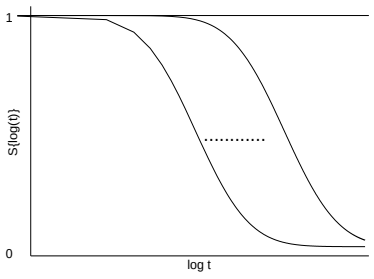


Figure 6.3 *A graph to check for a shift by translation.*

6.2 Weibull regression model

In this section we continue to work with the Motorette data first presented and analyzed in Chapter 4.6, page 92. There AIC selects the Weibull model as the best model and the Q-Q plot supports this. We now consider model diagnostics. We delay the S code until all relevant new definitions are presented.

Recall from expressions (4.1) and (4.4) the Weibull regression model has hazard and survivor functions

$$h(t|\underline{x}) = h_0(t) \cdot \exp(\underline{x}'\underline{\beta}) = \alpha \cdot (\widetilde{\lambda})^{\alpha} \cdot t^{\alpha-1}, \quad \text{where} \quad \widetilde{\lambda} = \lambda \cdot \left(\exp(\underline{x}'\underline{\beta})\right)^{\frac{1}{\alpha}},$$

and

$$S(t|\underline{x}) = \exp\left(-(\widetilde{\lambda}t)^{\alpha}\right).$$

The log of the cumulative hazard (4.5) is

$$\log\left(H(t|\underline{x})\right) = \log\left(-\log\left(S(t|\underline{x})\right)\right) = \alpha\log(\lambda) + \underline{x}'\underline{\beta} + \alpha\log(t).$$

Expression (4.3) tells us

$$Y = \log(T) = \underline{x}'\underline{\beta}^* + \beta_0^* + \sigma \cdot Z,$$

where $Z \sim$ standard extreme value distribution.

Graphical checks of overall model adequacy

We see that $\log(t_p)$ is not only linear in z_p, but also in each $x^{(j)}$, $j = 1, \ldots, m$. Further, the above linear model says $(Y - \beta_0^* - \underline{x}'\underline{\beta}^*)/\sigma = Z$. Define the ith residual e_i to be

$$e_i = \frac{y_i - \widehat{y}_i}{\widehat{\sigma}},$$

where $\widehat{y}_i = \widehat{\beta}_0^* + \underline{x}'\widehat{\underline{\beta}}^*$ is the ith estimated linear predictor. Under the Weibull model, the set of uncensored residuals should behave roughly like a set of iid standard extreme value variates. Let $e_1, e_2, \ldots, e_r$, $r \leq n$, represent the ordered uncensored residuals. We draw a Q-Q plot (page 63) of the points (z_i, e_i), $i = 1, \ldots, r \leq n$. In the recipe given on page 63, replace the sample quantile y_i with e_i and proceed to obtain the corresponding parametric quantile z_i. If the model under study (here it is the Weibull) is appropriate, the points (z_i, e_i) should lie very close to the 45^o-line through the origin. Figure **??** displays the Q-Q plot. Lastly, draw m scatter plots of the points $(x_i^{(j)}, y_i)$, $i = 1, \ldots, r \leq n$ and $j = 1, \ldots, m$. Each plot should display a straight line pattern. If not, perhaps transforming those $x_i^{(j)}$'s could improve the fit. See Figure 6.5.

The function `qq.reg.resid.s` (`qq.reg.resid.r` for R) draws a Q-Q plot of the e_i residuals. It has six arguments. They are:

```
data = data.frame
time = survival time variable name in data.frame
```

```
  status = name of status variable in data.frame
  fit = a survReg object
  quantile = "qweibull" or "qnorm" or "qlogis"
  xlab = "type your label" E.g., "standard extreme value quantiles"
```

S code for Q-Q plot of (z_i, e_i) after fitting the Motorette data to a Weibull regression model:

```
> fit.weib <- survReg(Surv(time,status) ~ x,dist="weibull",
              data=motorette)
> qq.reg.resid.s(motorette,motorette$time,motorette$status,fit.weib,
              "qweibull","standard extreme value quantiles")
              # Produces Figure 6.4
```

The Q-Q plot is also very useful for detecting overall adequacy of the final reduced regression model; that is, goodness-of-fit. As the single covariate x in the Motorette data has three distinct levels, we draw two Q-Q plots. In Figure 6.8, each group is fit to its own Weibull. The lines have different slopes and intercepts. In Figure 6.9, we fit a regression model with covariate x. The lines have same slope, but different intercepts. These plots can reveal additional information masked in Figures 6.4 and 6.5.

The `survReg` procedure in S gives the MLE's

$$\widehat{\beta}_0^*, \ \underline{\widehat{\beta}}^*, \ \widehat{\sigma}, \ \text{and} \ \widehat{\widetilde{\mu}} = \widehat{\beta}_0^* + \underline{x}'\underline{\widehat{\beta}}^*. \tag{6.1}$$

For the Weibull parameters we have

$$\widehat{\lambda} = \exp(-\widehat{\beta}_0^*), \ \underline{\widehat{\beta}} = -\widehat{\alpha}\,\underline{\widehat{\beta}}^*, \ \widehat{\alpha} = 1/\widehat{\sigma}, \ \text{and} \ \widehat{\widetilde{\lambda}} = \exp(-\widehat{\widetilde{\mu}}). \tag{6.2}$$

Note that `survReg` provides the **fitted times** $\widehat{T}_i$. So,

$$\widehat{Y}_i = \log(\widehat{T}_i) = \widehat{\widetilde{\mu}}_i \ . \tag{6.3}$$

Also recall (page 50) the p.d.f. of $Y_i = \log(T_i)$ and the corresponding survivor function evaluated at these estimates are

$$f(y_i|\widehat{\widetilde{\mu}}_i, \widehat{\sigma}) \ = \ \frac{1}{\widehat{\sigma}} \exp\left(\frac{y_i - \widehat{\widetilde{\mu}}_i}{\widehat{\sigma}} - \exp\left(\frac{y_i - \widehat{\widetilde{\mu}}_i}{\widehat{\sigma}}\right)\right) \tag{6.4}$$

$$S(y_i|\widehat{\widetilde{\mu}}_i, \widehat{\sigma}) \ = \ \exp\left(- \exp\left(\frac{y_i - \widehat{\widetilde{\mu}}_i}{\widehat{\sigma}}\right)\right). \tag{6.5}$$

Deviance, deviance residual, and graphical checks for outliers

We now consider a measure useful in detecting outliers. Define the **model deviance** as

$$\mathcal{D} \ = \ -2 \times (\text{log-likelihood of the fitted model} - \text{log-likelihood of the saturated model})$$

$$= \quad -2 \times \left(\sum_{i=1}^{n} \Big(\log(\widehat{L}_i) - \log(\widehat{L}_{si}) \Big) \right), \tag{6.6}$$

where $\widehat{L}_i$ denotes the ith individual's likelihood evaluated at the MLE's, and $\widehat{L}_{si}$ denotes the ith factor in the saturated likelihood evaluated at the MLE of θ_i. A saturated model in the regression setting without censoring is one with n parameters that fit the n observations perfectly. But in the presence of censored data, one needs to be careful. In view of (1.13), the factors of the likelihood corresponding to censored observations entail maximizing the survival probability. Let $\theta_1, \ldots, \theta_n$ denote the n parameters. This entails that for uncensored observations we obtain the MLE's with **no** constraints; whereas for censored observations, maximizing a survival probability imposes a constraint on the θ_i's fit to these censored y_i's. According to Klein & Moeschberger (1997, page 359), in computing the deviance the nuisance parameters are held fixed between the fitted and the saturated model. In the Weibull regression model, the only nuisance parameter is the σ and is held fixed at the MLE value obtained in the fitted model. The measure $\mathcal{D}$ can be used as a goodness of fit criterion. The larger the model deviance, the poorer the fit and vice versa. For an approximate size-α test, compare the calculated $\mathcal{D}$ value to the χ^2_α critical value with $n - m - 1$ degrees of freedom.

Under the random (right) censoring model and under the assumption that censoring time has no connection with the survival time, recall the **likelihood function** of the sample (1.13) is

$$L = L(\beta_0^*; \underline{\beta}^*; \sigma) = L(\underline{\mu}; \sigma) = \prod_{i=1}^{n} L_i(\widetilde{\mu}_i; \sigma),$$

where

$$\begin{aligned} L_i(\widetilde{\mu}_i; \sigma) &= \big(f(y_i|\widetilde{\mu}_i, \sigma)\big)^{\delta_i} \big(S(y_i|\widetilde{\mu}_i, \sigma)\big)^{1-\delta_i} \quad \text{and} \\ \delta_i &= \begin{cases} 1 & \text{if } y_i \text{ is uncensored} \\ 0 & \text{if } y_i \text{ is censored.} \end{cases} \end{aligned}$$

In preparation to define the deviance residual, we first define two types of residuals which are the parametric analogues to those defined and discussed in some detail in Section 6.3.

Cox-Snell residual

The ith *Cox-Snell residual* is defined as

$$r_{C\,i} = \widehat{H}_0(t_i) \times \exp(\underline{x}_i' \widehat{\underline{\beta}}), \tag{6.7}$$

where $\widehat{H}_0(t_i)$ and $\widehat{\underline{\beta}}$ are the MLE's of the baseline cumulative hazard function and coefficient vector, respectively. As these residuals are always nonnegative, their plot is difficult to interpret. These are not residuals in the sense of linear

models because they are not the difference between the observed and fitted values. Their interpretation is discussed in Section 6.3.

Martingale residual

The ith *martingale residual* is defined as

$$\widehat{M}_i = \delta_i - r_{C\,i}. \tag{6.8}$$

The $\widehat{M}_i$ take values in $(-\infty, 1]$ and are always negative for censored observations. In large samples, the martingale residuals are uncorrelated and have expected value equal to zero. But they are not symmetrically distributed about zero.

Deviance residual

The ith *deviance residual*, denoted by D_i, is the square root of the ith term of the deviance, augmented by the sign of the $\widehat{M}_i$:

$$D_i = \text{sign}(\widehat{M}_i) \times \sqrt{-2 \times \left(\log\left(\widehat{L}_i(\widehat{\widetilde{\mu}}_i, \widehat{\sigma})\right) - \log(\widehat{L}_{si})\right)}. \tag{6.9}$$

These residuals are expected to be symmetrically distributed about zero. Hence, their plot is easier to interpret. But we caution these do not necessarily sum to zero. The model deviance then is

$$\mathcal{D} = \sum_{i=1}^{n} {D_i}^2 = \text{the sum of the squared deviance residuals.}$$

When there is light to moderate censoring, the D_i should look like an iid normal sample. Therefore, the deviance residuals are useful in detecting outliers. To obtain the D_i, use `> resid(fit,type="deviance")` where `fit` is a `survReg` object. A plot of the D_i against the fitted log-times is given in Figure 6.6.

There are three plots constructed with D_i that are very useful in helping to detect outliers. One is the normal probability plot. Here we plot the kth ordered D_i against its normal score $Z((k-.375)/(n+.25))$ where $Z(A)$ denotes the Ath quantile of the standard normal distribution. Outliers will be points that fall substantially away from a straight line. The second graph plots the D_i against the estimated *risk scores* $\sum_{j=1}^{m} \widehat{\beta}_j^* x_i^{(j)}$. This plot should look like a scatter of random noise about zero. Outliers will have large absolute deviations and will sit apart from the point cloud. The third graph plots D_i against its observation (index) number. Again, we look for points that are set apart with large absolute value. See Figure 6.10.

For the interested reader, the following is the expression for the ***i*th deviance residual (6.9) under the extreme value model**, which corresponds to

fitting the Weibull regression model.

$$D_i = \text{sign}(\widehat{M}_i) \times \sqrt{-2 \times \left\{ \widehat{M}_i + \delta_i \log(\delta_i - \widehat{M}_i) \right\}}, \tag{6.10}$$

where $\widehat{M}_i$ is defined in expression (6.8) and

$$r_{C\,i} = \left(\hat{\lambda} t_i\right)^{\widehat{\alpha}} \times \exp\left(\underline{x}_i' \widehat{\underline{\beta}}\right) = \exp\left(\frac{y_i - \widehat{\widetilde{\mu}}_i}{\widehat{\sigma}}\right), \tag{6.11}$$

which follows from expression (4.6). The derivation of this expression is given in our book. This now matches the definition of deviance residual to be presented in Section 6.3.3.

Partial deviance

We now consider hierarchical (nested) models. Let R denote the reduced model and F denote the full model which consists of additional covariates added to the reduced model. Partial deviance is a measure useful for model building. We define **partial deviance** as

$$\begin{aligned} \mathcal{PD} &= \text{Deviance (additional covariates} \mid \text{covariates in the reduced model)} \\ &= \mathcal{D}(R) - \mathcal{D}(F) = -2 \log\left(\widehat{L}(R)\right) + 2 \log\left(\widehat{L}(F)\right) \qquad (6.12) \\ &= -2 \log\left(\frac{\widehat{L}(R)}{\widehat{L}(F)}\right). \end{aligned}$$

We see that the partial deviance is equivalent to the LRT statistic. Hence, the LRT checks to see if there is significant partial deviance. We reject when $\mathcal{PD}$ is "large." If the partial deviance is large, this indicates that the additional covariates improve the fit. If the partial deviance is small, it indicates they don't improve the fit and the smaller model (R) is just as adequate. Hence, drop the additional covariates and continue with the reduced model. Partial deviance is analogous to the extra sum of squares, SSR(additional covariates|covariates in R)=SSE(R)−SSE(F), for ordinary linear regression models. In fact, when the $\log(T_i)$'s are normal and no censoring is present, partial deviance simplifies to

$$n \log\left(\frac{\text{MSE}(R)}{\text{MSE}(F)}\right) + (P_F - P_R),$$

where P_F and P_R are the number of parameters in the full and reduced models, respectively. The argument of the log function can be easily expressed as a function of the classic F test statistic to test a reduced model against the full model. The $\mathcal{PD}$ simplifies to an increasing function of the classic F statistic, which has in its numerator the extra sum of squares SSE(R)−SSE(F). WHY!

dfbeta

dfbeta is a useful measure to assess the influence of each point on the estimated coefficients $\widehat{\beta}_j$'s. This measure is analogous to that used in regular

linear regression. Large values suggest we inspect corresponding data points. The measure **dfbetas** is dfbeta divided by the s.e.$(\widehat{\beta}_j)$. We obtain these quantities via the companion function `resid` where `fit` is a `survReg` object.
`> resid(fit, type="dfbeta").`
See Figure 6.7 for a plot of the dfbeta for each observation's influence on the coefficient of the x variable. See Section 6.3.6 for a more detailed discussion of the dfbeta measure.

Motorette example: Is the Weibull regression model appropriate?

Figure 6.4:

```
> attach(motorette)
# See page 126.
```

Figure 6.5:

```
> plot.logt.x(time,status,x) # Plot of log(t) against x.

 # Now the Weibull regression fit:
> motor.fit <- survReg(Surv(time,status) ~ x,dist="weibull")
> dresid <- resid(motor.fit,type="deviance")
> riskscore <- log(fitted(motor.fit)) - coef(motor.fit)[1]
```

Figure 6.6:

```
> plot(log(fitted(motor.fit)),dresid)
> mtext("Deviance Residuals vs log Fitted Values (muhat)",
          3,-1.5)
> abline(h=0)
```

Figure 6.10:

```
> index <- seq(1:30)
> par(mfrow=c(2,2))
> plot(riskscore,dresid,ylab="deviance residuals")
> abline(h=0)
> qqnorm.default(dresid,datax=F,plot=T,
                  ylab="deviance residuals")
> qqline(dresid)
> plot(index,dresid,ylab="deviance residual")
> abline(h=0)
```

Figure 6.7:

```
 # We plot dfbeta to assess influence of each point on the
 # estimated coefficient.
> dfbeta <- resid(motor.fit,type="dfbeta")
> plot(index,dfbeta[,1],type="h",ylab="Scaled change in
        coefficient",xlab="Observation")
```

Figure 6.8:

```
> xln <- levels(factor(x))
> ts.1 <- Surv(time[as.factor(x)==xln[1]],
                 status[as.factor(x)==xln[1]])
> ts.2 <- Surv(time[as.factor(x)==xln[2]],
                 status[as.factor(x)==xln[2]])
> ts.3 <- Surv(time[as.factor(x)==xln[3]],
                 status[as.factor(x)==xln[3]])
> qq.weibull(list(ts.1,ts.2,ts.3))
```

Figure 6.9:

```
> xln <- levels(factor(x))
> ts.1 <- Surv(time[as.factor(x)==xln[1]],
                 status[as.factor(x)==xln[1]])
> ts.2 <- Surv(time[as.factor(x)==xln[2]],
                 status[as.factor(x)==xln[2]])
> ts.3 <- Surv(time[as.factor(x)==xln[3]],
                 status[as.factor(x)==xln[3]])
> qq.weibreg(list(ts.1,ts.2,ts.3),motor.fit)
```

We compute the log-likelihood of saturated model, partial deviance, and then compare to the output from the `anova` function.

```
> summary(motor.fit)
             Value Std. Error     z         p
(Intercept) -11.89      1.966 -6.05 1.45e-009
          x   9.04      0.906  9.98 1.94e-023
 Log(scale)  -1.02      0.220 -4.63 3.72e-006

Scale= 0.361

Loglik(model)= -144.3   Loglik(intercept only)= -155.7
    Chisq= 22.67 on 1 degrees of freedom, p= 1.9e-006
 # Chisq=22.67 is the LRT value for testing the
 # significance of the x variable.
> loglikR <- motor.fit$loglik[1]
> loglikR     # Model has only intercept.
[1] -155.6817
> loglikF <- motor.fit$loglik[2]
> loglikF     # Model includes the covariate x.
[1] -144.3449
> ModelDev <- sum(resid(motor.fit,type="deviance")^2)
> ModelDev
```

```
[1] 46.5183  # Full model deviance
> loglikSat <- loglikF + ModelDeviance/2
> loglikSat
[1] -121.0858
> nullDev <- - 2*(loglikR - loglikSat)
> nullDev  # Reduced Model (only intercept)
[1] 69.19193
> PartialDev <- nullDev - ModelDev
> PartialDev
[1] 22.67363 # which equals the LRT value.
    # The following ANOVA output provides Deviance
    # which is really the partial deviance. This is
    # easily seen.
> anova(motor.fit)
 Analysis of Deviance Table Response: Surv(time,status)
 Terms added sequentially (first to last)
       Df  Deviance Resid.  Df    -2*LL       Pr(Chi)
 NULL                        2   311.3634
   x   -1  22.67363          3   288.6898   1.919847e-006
> detach()
```

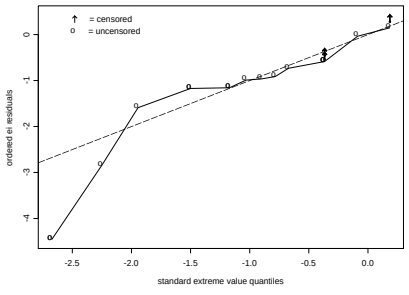


Figure 6.4 *Q-Q plot for the e_i residuals. Dashed line is the 45^o-line.*

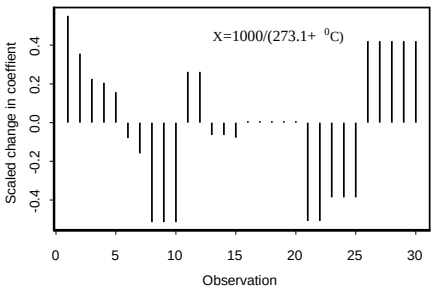


Figure 6.7 *The dfbeta plot helps assess each point's influence on $\widehat{\beta}$.*

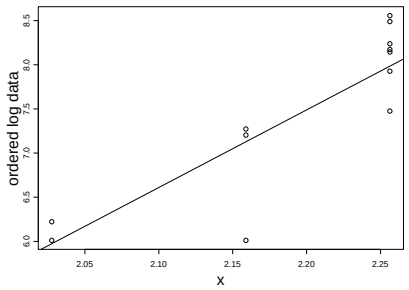


Figure 6.5 *Log(t) against x. Least squares line.*

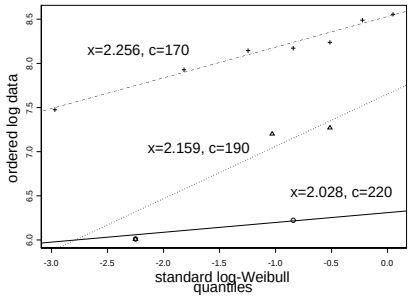


Figure 6.8 *Q-Q plot. Different intercepts and slopes.*

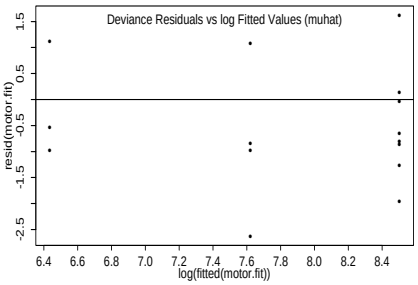


Figure 6.6 *Deviance residuals against fitted log-times.*

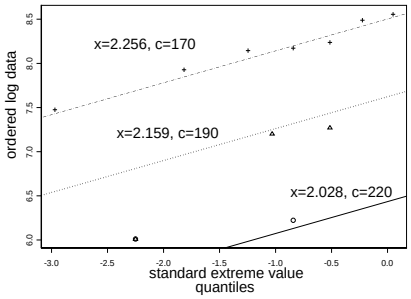


Figure 6.9 *Q-Q plot for model $y = \beta_0^* + \beta_1^* x + \text{error}$. Each line based on MLE's. Lines have same slope, but different intercepts.*

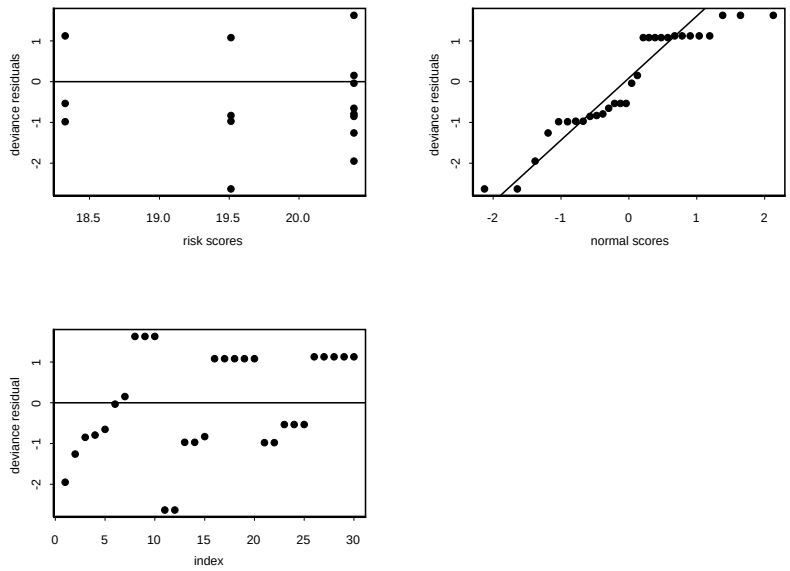


Figure 6.10 *Motorette data: deviance residuals against risk scores, normal scores, and index.*

Results:

- In Figure 6.8, each group is fit separately. The graphs suggest the Weibull model gives an adequate description of each group.
- Figure 6.9 supports the Weibull regression model describes well the role temperature plays in the acceleration of failure of the motorettes.
- Figure 6.5 displays a straight line. Figure 6.7 shows no influential points. Both Figure 6.6 and Figure 6.10 (deviance residuals vs. risk scores) display a random scatter about zero except for a possible outlier whose deviance residual value is -2.634, which, incidentally, represents the two extreme cases detected by the deviance residual vs. index plot. These two cases correspond to the possible outlier revealed in the Q-Q plot displayed in Figure 6.4.
- The plot of deviance residuals against their normal scores in Figure 6.10 suggests one potential outlier. But this is somewhat misleading. The three upper right points correspond to cases with the same deviance residual value of 1.626052, but with different normal scores. This occurs because the S function `qqnorm` assigns these residuals their distinct ranks $k = 28, 29,$ and 30. Hence, their normal scores ($Z((k-.375)/(n+.25))$) are 1.361, 1.61,

and 2.04, respectively. However, if we follow the convention of assigning the average rank to tied observations, then each of these three tied deviance residuals now has the normal score value of 1.61. In this case, the three points are now the single point in the middle and there are no apparent outliers in this plot.

- The LRT per the `anova` function, with a p-value of 1.9×10^{-6}, provides strong evidence the Weibull model with the predictor variable x is adequate. Equivalently, the p-value of 1.94×10^{-23} for the estimated coefficient of x provides this strong evidence.

6.3 Cox proportional hazards model

Recall from Chapter 4.3 that this model has hazard function

$$\begin{aligned} h(t|\underline{x}) = h_0(t) \cdot \exp(\underline{x}'\underline{\beta}) &= h_0(t) \cdot \exp(\beta_1 x^{(1)} + \cdots + \beta_m x^{(m)}) \\ &= h_0(t) \cdot \exp(\beta_1 x^{(1)}) \times \exp(\beta_2 x^{(2)}) \times \cdots \times \exp(\beta_m x^{(m)}), \end{aligned}$$

where at two different points $\underline{x}_1, \underline{x}_2$, the proportion

$$\frac{h(t|\underline{x}_1)}{h(t|\underline{x}_2)} = \frac{\exp(\underline{x}_1'\underline{\beta})}{\exp(\underline{x}_2'\underline{\beta})} = \exp\left((\underline{x}_1' - \underline{x}_2')\underline{\beta}\right),$$

called the hazards ratio (HR), is constant with respect to time t.

As the baseline hazard function is not specified in the Cox model, the likelihood function cannot be fully specified. To see this, recall that

$$f(\cdot) = h(\cdot) \times S(\cdot).$$

The hazard function $h(\cdot)$ depends on the baseline hazard $h_0(\cdot)$. Hence, so does the p.d.f. Cox (1975) defines a likelihood based on conditional probabilities which are free of the baseline hazard. His estimate is obtained from maximizing this likelihood. In this way he avoids having to specify $h_0(\cdot)$ at all. We derive this likelihood heuristically. Let t^* denote a time at which a death has occurred. Let $\mathcal{R}(t^*)$ be the risk set at time t^*; that is, the indices of individuals who are alive and not censored just before t^*. First,

$$\begin{aligned} &P\{\text{one death in } [t^*, t^* + \triangle t^*) \mid \mathcal{R}(t^*)\} \\ &\quad = \sum_{l \in \mathcal{R}(t^*)} P\{T_l \in [t^*, t^* + \triangle t^*) \mid T_l \geq t^*\} \\ &\quad \approx \sum_{l \in \mathcal{R}(t^*)} h(t^*|\underline{x}_l)\triangle t^* \\ &\quad = \sum_{l \in \mathcal{R}(t^*)} h_0(t^*) \cdot \exp(\underline{x}_l'\underline{\beta})\triangle t^*. \end{aligned}$$

Thus, if we let $P\{\text{one death at } t^* \mid \mathcal{R}(t^*)\}$ denote the

$$\sum_{l \in \mathcal{R}(t^*)} P(T_l = t^* | T_l \geq t^*) \, ,$$

then we have

$$P\{\text{one death at } t^* \mid \mathcal{R}(t^*)\} = \sum_{l \in \mathcal{R}(t^*)} h_0(t^*) \cdot \exp(\underline{x}_l' \underline{\beta}).$$

Now, let $t_{(1)}, \ldots, t_{(r)}$ denote the $r \leq n$ distinct ordered (uncensored) death times, so that $t_{(j)}$ is the jth ordered death time. Let $\underline{x}_{(j)}$ denote the vector of covariates associated with the individual who dies at $t_{(j)}$. Then, for each j, we have

$$\begin{aligned} L_j(\underline{\beta}) &= P\{\text{individual with } \underline{x}_{(j)} \text{ dies at } t_{(j)} \mid \text{one death in } \mathcal{R}(t_{(j)}) \text{ at } t_{(j)}\} \\ &= \frac{P\{\text{individual with } \underline{x}_{(j)} \text{ dies at } t_{(j)} \mid \text{individual in } \mathcal{R}(t_{(j)})\}}{P\{\text{one death at } t_{(j)} \mid \mathcal{R}(t_{(j)})\}} \\ &= \frac{h_0(t_{(j)}) \cdot \exp(\underline{x}_{(j)}' \underline{\beta})}{\sum_{l \in \mathcal{R}(t_{(j)})} h_0(t_{(j)}) \cdot \exp(\underline{x}_l' \underline{\beta})} \\ &= \frac{\exp(\underline{x}_{(j)}' \underline{\beta})}{\sum_{l \in \mathcal{R}(t_{(j)})} \exp(\underline{x}_l' \underline{\beta})}. \end{aligned}$$

The product of these over the r uncensored death times yields what Cox refers to as the partial likelihood. The **partial likelihood function**, denoted by $L_c(\underline{\beta})$, is thus defined to be

$$L_c(\underline{\beta}) = \prod_{j=1}^{r} L_j(\underline{\beta}) = \prod_{j=1}^{r} \frac{\exp(\underline{x}_{(j)}' \underline{\beta})}{\sum_{l \in \mathcal{R}(t_{(j)})} \exp(\underline{x}_l' \underline{\beta})}. \tag{6.13}$$

Recall that in the random censoring model we observe the times $y_1, \ldots, y_n$ along with the associated $\delta_1, \ldots, \delta_n$ where $\delta_i = 1$ if the y_i is uncensored (i.e., the actual death time was observed) and $\delta_i = 0$ if y_i is censored. We can now give an equivalent expression for the partial likelihood function in terms of all n observed times:

$$L_c(\underline{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\underline{x}_i' \underline{\beta})}{\sum_{l \in \mathcal{R}(y_i)} \exp(\underline{x}_l' \underline{\beta})} \right)^{\delta_i}. \tag{6.14}$$

Remarks:

1 Cox's estimates maximize the log-partial likelihood.

2 To analyze the effect of covariates, there is no need to estimate the nuisance parameter $h_0(t)$, the baseline hazard function.

3 Cox argues that most of the relevant information about the coefficients $\underline{\beta}$ for regression with censored data is contained in this partial likelihood.

4 This partial likelihood is not a true likelihood in that it does not integrate out to 1 over $\{0,1\}^n \times \Re_+{}^n$.

5 Censored individuals do not contribute to the numerator of each factor. But they do enter into the summation over the risk sets at death times that occur before a censored time.

6 Furthermore, this partial likelihood depends only on the ranking of the death times, since this determines the risk set at each death time. Consequently, inference about the effect of the explanatory variables on the hazard function depends only on the rank order of the death times! Here we see why this is often referred to as nonparametric. It only depends on the rank order! Look at the partial likelihood. There is no visible $t_{(j)}$ in the estimate for $\underline{\beta}$. It is a function of the $\underline{x}_{(j)}$'s which are determined by the rank order of the death times. So, the estimates are a function of the rank order of the death times.

We now present data diagnostic methods. We delay the examples and all S code until all relevant definitions and methods are presented.

6.3.1 Cox-Snell residuals for assessing the overall fit of a PH model

Recall from (1.6) the relationship

$$H(t) = -\log\left(S(t)\right) = -\log\left(1 - F(t)\right),$$

where F denotes the true d.f. of the survival time T and H denotes the true cumulative hazard rate. Also recall that regardless of the form of F, the random variable $F(T)$ is distributed uniformly on the unit interval (0,1). Hence, the random variable $H(T)$ is distributed exponentially with hazard rate $\lambda = 1$. WHY! Let $\underline{x}_i$ denote the i-th individual's covariate vector. Then for a given $\underline{x}_i$, $H(t|\underline{x}_i)$ denotes the true cumulative hazard rate for an individual with covariate vector $\underline{x}_i$. It then follows

$$H(T_i|\underline{x}_i) \sim \exp(\lambda = 1).$$

Hence, if the Cox PH model is correct, then for a given $\underline{x}_i$, it follows

$$H(T_i|\underline{x}_i) = H_0(T_i) \times \exp\left(\sum_{j=1}^{m} \beta_j x_i^{(j)}\right) \sim \exp(\lambda = 1). \tag{6.15}$$

The *Cox-Snell residuals* (Cox and Snell, 1968) are defined as

$$r_{C\,i} = \widehat{H}_0(Y_i) \times \exp\left(\sum_{j=1}^{m} \widehat{\beta}_j x_i^{(j)}\right), i = 1, \ldots, n, \tag{6.16}$$

where $Y_i = \min(T_i, C_i)$. The $\widehat{\beta}_j$'s are the *maximum partial likelihood estimates*, the estimates obtained from maximizing Cox's partial likelihood (6.14). The $\widehat{H}_0(t)$ is an empirical estimate of the cumulative hazard at time t. Typically

this is either the Breslow or Nelson-Aalen estimate (page 29). S offers both with Nelson-Aalen as the default. For the definition of Breslow estimator, see Klein & Moeschberger (1997, page 237). If the final PH model is correct and the $\widehat{\beta}_j$'s are close to the true values of the β_j's, the $r_{C\,i}$'s should resemble a censored sample from a unit exponential distribution. Let $H_E(t)$ denote the cumulative hazard rate of the unit exponential. Then $H_E(t) = t$. Let $\widehat{H}_{r_C}(t)$ denote a consistent estimator of the cumulative hazard rate of the $r_{C\,i}$'s. Then $\widehat{H}_{r_C}(t)$ should be close to $H_E(t) = t$. Thus, for each uncensored $r_{C\,i}$, $\widehat{H}_{r_C}(r_{C\,i}) \approx r_{C\,i}$. To check whether the $r_{C\,i}$'s resemble a censored sample from a unit exponential, the plot of $\widehat{H}_{r_C}(r_{C\,i})$ against $r_{C\,i}$ should be a 45^o-line through the origin. See Figure 6.11.

Remarks:

1 The Cox-Snell residuals are most useful for examining the overall fit of a model. A shortcoming is they do not indicate the type of departure from the model detected when the estimated cumulative hazard plot is not linear.

2 Ideally, the plot of $\widehat{H}_{r_C}(r_{C\,i})$ against $r_{C\,i}$ should include a confidence band so that significance can be addressed. Unfortunately, the $r_{C\,i}$ are not exactly a censored sample from a distribution. So this plot is generally used only as a rough diagnostic. A formal test of adequacy of the Cox PH model is given in Section 6.3.5.

3 The closeness of the distribution of the $r_{C\,i}$'s to the unit exponential depends heavily on the assumption that, when $\underline{\beta}$ and H_0 are replaced by their estimates, the probability integral transform $F(T)$ still yields uniform (0,1) distributed variates. This approximation is somewhat suspect for small samples. Furthermore, departures from the unit exponential distribution may be partly due to the uncertainty in estimating the parameters $\underline{\beta}$ and H_0. This uncertainty is largest in the right-hand tail of the distribution and for small samples.

6.3.2 Martingale residuals for identifying the best functional form of a covariate

The martingale residual is a slight modification of the Cox-Snell residual. When the data are subject to right censoring and all covariates are time-independent (fixed at the start of the study), then the *martingale residuals*, denoted by $\widehat{M}_i$, are defined to be

$$\widehat{M}_i = \delta_i - \widehat{H}_0(Y_i) \times \exp\left(\sum_{j=1}^{m} \widehat{\beta}_j x_i^{(j)}\right) = \delta_i - r_{C\,i}, i = 1, \ldots, n, \qquad (6.17)$$

where $r_{C\,i}$ is the Cox-Snell residual.

These residuals are used to examine the best functional form for a given

covariate using the assumed Cox PH model for the remaining covariates. Let the covariate vector $\underline{x}$ be partitioned into a $\underline{x}_*$ for which we know the functional form, and a single continuous covariate $x^{(1)}$ for which we are unsure of what functional form to use. We assume $x^{(1)}$ is independent of $\underline{x}_*$. Let $g(\cdot)$ denote the best function of $x^{(1)}$ to explain its effect on survival. The Cox PH model is then,

$$H(t|\underline{x}_*, x^{(1)}) = H_0(t) \times \exp\left(\underline{x}'_* \underline{\beta}_*\right) \times \exp\left(g(x^{(1)})\right) \ , \qquad (6.18)$$

where $\underline{\beta}_*$ is an $m-1$ dimensional coefficient vector. To find $g(\cdot)$, we fit a Cox PH model to the data based on $\underline{x}_*$ and compute the martingale residuals, $\widehat{M}_i$, $i = 1, \ldots, n$. These residuals are plotted against the values $x_i^{(1)}$, $i = 1, \ldots, n$. A smoothed fit of the scatter plot is typically used. The smooth-fitted curve gives some indication of the function $g(\cdot)$. If the plot is linear, then no transformation of $x^{(1)}$ is needed. If there appears to be a threshold, then a discretized version of the covariate is indicated. The S function `coxph` provides martingale residuals as default and the S function `scatter.smooth` displays a smoothed fit of the scatter plot of the martingale residuals versus the covariate $x^{(1)}$. See Figure 6.12.

Remarks:

1 Cox-Snell residuals can be easily obtained from martingale residuals.

2 It is common practice in many medical studies to discretize continuous covariates. The martingale residuals are useful for determining possible cut points for such variables. In Chapter 6.3.8 of our book we present a cut point analysis with bootstrap validation conducted for the variable KPS.PRE. in the CNS data.

3 The martingale residual for a subject is the difference between the observed and the expected number of deaths for the individual. This is so because we assume that no subjects can have more than one death and the second factor in expression (6.17) is the estimated cumulative hazard of death for the individual over the interval $(0, y_i)$.

4 The martingale residuals sum to zero; that is, $\sum_{i=1}^{n} \widehat{M}_i = 0$. For "large" n, the $\widehat{M}_i$'s are an uncorrelated sample from a population with mean zero. However, they are not symmetric around zero because the martingale residuals take values between $-\infty$ and 1.

5 For the more general definition of the martingale residuals which includes time-dependent covariates, see Klein & Moeschberger (1997, pages 333 and 334). On page 337 under *Theoretical Notes* these authors further explain why a smoothed plot of the martingale residuals versus a covariate should reveal the correct functional form for including $x^{(1)}$ in a Cox PH model.

6.3.3 Deviance residuals to detect possible outliers

These residuals were defined and discussed in great detail in the previous section on diagnostic methods for parametric models. Except for a slight modification in the definition of deviance, all plots and interpretations carry over. What's different here is that we no longer have a likelihood. We are working with a partial likelihood. However, we may still define deviance analogously, using the partial likelihood. All tests and their large sample distributions still apply. The deviance residual is used to obtain a residual that is more symmetrically shaped than a martingale residual as the martingale residual can be highly skewed. The *deviance residual* (Therneau, Grambsch, and Fleming, 1990) is defined by

$$D_i = \text{sign}(\widehat{M}_i) \times \sqrt{-2 \times \left(\widehat{M}_i + \delta_i \log(\delta_i - \widehat{M}_i)\right)}\,, \tag{6.19}$$

where $\widehat{M}_i$ is the martingale residual defined in Subsection 6.3.2. The log function inflates martingale residuals close to one, while the square root contracts the large negative martingale residuals. In all plots, potential outliers correspond to large absolute valued deviance residuals. See Figure 6.13.

Remarks:

1 Therneau, Grambsch, and Fleming (1990) note "When censoring is minimal, less than 25% or so, these residuals are symmetric around zero. For censoring greater than 40%, a large bolus of points with residuals near zero distorts the normal approximation but the transform is still helpful in symmetrizing the set of residuals." Obviously, deviance residuals do not necessarily sum to zero.

2 Type `resid(fit,type="deviance")`, where `fit` is the `coxph` object, to obtain these residuals.

6.3.4 Schoenfeld residuals to examine fit and detect outlying covariate values

The kth *Schoenfeld residual* (Schoenfeld, 1982) defined for the kth subject on the jth explanatory variable $x^{(j)}$ is given by

$$r_{s_{jk}} = \delta_k\{x_k^{(j)} - a_k^{(j)}\}, \tag{6.20}$$

where δ_k is the kth subject's censoring indicator, $x_k^{(j)}$ is the value of the jth explanatory variable on the kth individual in the study,

$$a_k^{(j)} = \frac{\sum_{m \in \mathcal{R}(y_k)} \exp(\underline{x}'_m \underline{\widehat{\beta}}) x_m^{(j)}}{\sum_{m \in \mathcal{R}(y_k)} \exp(\underline{x}'_m \underline{\widehat{\beta}})},$$

and $\mathcal{R}(y_k)$ is the risk set at time y_k. The MLE $\widehat{\underline{\beta}}$ is obtained from maximizing the Cox's partial likelihood function $L_c(\underline{\beta})$ (6.14). Note that nonzero residuals only arise from uncensored observations.

We see this residual is just the difference between $x_k^{(j)}$ and a weighted average of the values of explanatory variables over individuals at risk at time y_k. The **weight** used for the mth individual in the risk set at y_k is

$$\frac{\exp(\underline{x}'_m\widehat{\underline{\beta}})}{\sum_{m\in\mathcal{R}(y_k)}\exp(\underline{x}'_m\widehat{\underline{\beta}})},$$

which is the contribution from this individual to the maximized partial likelihood (6.14). Further, since the MLE of $\underline{\beta}$, $\widehat{\underline{\beta}}$, is such that

$$\frac{\partial \log\big(L_c(\underline{\beta})\big)}{\partial\beta_j}\bigg|_{\widehat{\underline{\beta}}} = 0,$$

the Schoenfeld residuals for each predictor $x^{(j)}$ must sum to zero. These residuals also have the property that in large samples the expected value of $r_{s_{jk}}$ is zero and they are uncorrelated with each other. Furthermore, suppose y_k is a small failure time relative to the others. Then its risk set is huge. Hence, in general not only do subjects in the risk set have a wide range of covariate values, but also the weight assigned to each covariate value associated with the risk set is small. Therefore, individuals with large covariate values who die at early failure times would have large positive Schoenfeld residuals. This can be most easily seen if we rewrite $r_{s_{jk}}$ (6.20) as

$$x_k^{(j)}\Big(1-\frac{\exp(\underline{x}'_k\widehat{\underline{\beta}})}{\sum_{m\in\mathcal{R}(y_k)}\exp(\underline{x}'_m\widehat{\underline{\beta}})}\Big)-\sum_{l\in\mathcal{R}(y_k);\,l\neq k}\Big(x_l^{(j)}\frac{\exp(\underline{x}'_l\widehat{\underline{\beta}})}{\sum_{m\in\mathcal{R}(y_k)}\exp(\underline{x}'_m\widehat{\underline{\beta}})}\Big). \tag{6.21}$$

It is clear from expression (6.21) that the first term is large and the second term is small relative to the first term. Similarly, the individuals with small covariate values who die at early failure times would have large negative Schoenfeld residuals. WHY! Therefore, a few relatively large absolute valued residuals at early failure times may not cause specific concern. Thus, these residuals are helpful in detecting outlying covariate values for early failure times. However, if the PH assumption is satisfied, large Schoenfeld residuals are not expected to appear at late failure times. WHY! Therefore, we should check the residuals at late failure times. See Figure 6.14.

Remarks:

1 Schoenfeld calls these residuals the partial residuals as these residuals are obtained from maximizing the partial likelihood function. Collett (1994, page 155), among others, calls these residuals the score residuals as the first derivative of the log-partial likelihood can be considered as the efficient score.

2 Use `coxph.detail` to obtain the detailed `coxph` object. This includes ranked observed times along with a corresponding censoring status vector and covariate information.

3 Type `resid(fit,type="schoenfeld")`, where `fit` is the `coxph` object, to obtain these residuals. `coxph` does not output the value of Schoenfeld residual for subjects whose observed survival time is censored as these are zeros.

4 If the assumption of proportional hazards holds, a plot of these residuals against ordered death times should look like a tied down random walk. Otherwise, the plot will show too large residuals at some times.

6.3.5 Grambsch and Therneau's test for PH assumption

As an alternative to proportional hazards, Grambsch and Therneau (1994) consider time-varying coefficients $\underline{\beta}(t) = \underline{\beta} + \underline{\theta} g(t)$, where $g(t)$ is a predictable process (a postulated smooth function). Given $g(t)$, they develop a score test for $H_0 : \underline{\theta} = \underline{0}$ based on a generalized least squares estimator of $\underline{\theta}$. Defining *scaled Schoenfeld residuals* by the product of the inverse of the estimated variance-covariance matrix of the kth Schoenfeld residual and the kth Schoenfeld residual, they show the kth scaled Schoenfeld residual has approximately mean $\underline{\theta} g(t_k)$ and the kth Schoenfeld residual has an easily computable variance-covariance matrix. Motivated by these results, they also develop a graphical method. They show by Monte Carlo simulation studies that a smoothed scatter plot of $\widehat{\underline{\beta}}(t_k)$, the kth scaled Schoenfeld residual plus $\widehat{\underline{\beta}}$ (the maximum partial likelihood estimate of $\underline{\beta}$), versus t_k reveals the functional form of $\beta(t)$. Under H_0, we expect to see a constant function over time. Both of these can be easily done with the S functions `cox.zph` and `plot`. See Figure 6.15.

Remarks:

1 The function $g(t)$ has to be specified. The default in the S function `cox.zph` is K-M(t). The options are $g(t) = t$ and $g(t) = \log(t)$ as well as a function of one's own choice.

2 `plot(out)`, where `out` is the `cox.zph` object, gives a plot for each covariate. Each plot is of a component of $\widehat{\beta}(t)$ versus t together with a spline smooth and ± 2 s.e. pointwise confidence bands for the spline smooth.

3 A couple of useful plots for detecting violations of the PH assumption are recommended:

(a) A plot of log-cumulative hazard rates against time is useful when x is a group variable. For example, if there are two treatment groups, plot both curves on the same graph and compare them. If the curves are parallel

over time, it supports the PH assumption. If they cross, this is a blatant violation.

(b) A plot of differences in log-cumulative hazard rates against time is also useful. This plot displays the differences between the two curves in the previous graph. If the PH assumption is met, this plot is roughly constant over time. Otherwise, the violation will be glaring. This plot follows Miller's basic principle discussed here on page 122.

6.3.6 dfbetas to assess influence of each observation

Here we want to check the influence of each observation on the estimate $\widehat{\underline{\beta}}$ of the $\underline{\beta}$. Let $\widehat{\underline{\beta}}_{(k)}$ denote the estimated vector of coefficients computed on the sample with the kth observation deleted. Then we check which components of the vector $\widehat{\underline{\beta}} - \widehat{\underline{\beta}}_{(k)}$ have unduly large absolute values. Do this for each of the n observations. One might find this measure similar to dfbetas in the linear regression. This involves fitting $n+1$ Cox regression models. Obviously, this is computationally expensive unless the sample size is small. Fortunately, there exists an approximation based on the Cox PH model fit obtained from the whole data that can be used to circumvent this computational expense. The kth *dfbeta* is defined as

$$\text{dfbeta}_k = I(\widehat{\underline{\beta}})^{-1}(r^*_{s_{1k}}, \ldots, r^*_{s_{mk}})', \tag{6.22}$$

where $I(\widehat{\underline{\beta}})^{-1}$ is the inverse of the observed Fisher information matrix, and for $j = 1, \ldots, m$,

$$r^*_{s_{jk}} = \delta_k\{x_k^{(j)} - a_k^{(j)}\} - \exp(\underline{x}_k'\widehat{\underline{\beta}}) \sum_{t_i \le y_k} \frac{\{x_k^{(j)} - a_i^{(j)}\}}{\sum_{l \in \mathcal{R}(t_i)} \exp(\underline{x}_l'\widehat{\underline{\beta}})}.$$

Note that the first component is the kth Schoenfeld residual and the second component measures the combined effect over all the risk sets that include the kth subject. This expression, proposed by Cain and Lange (1984), well approximates the difference $\widehat{\underline{\beta}} - \widehat{\underline{\beta}}_{(k)}$ for $k = 1, \ldots, n$. The authors note that the above two components in general have opposite signs. The second component increases in absolute magnitude with t_k, as it is the sum of an increasing number of terms. Thus, for early death times, the first component dominates, while for later death times, the second is usually of greater magnitude. This means that for patients who die late, the fact that the patient lived a long time, and thus was included in many risk sets, has more effect upon $\widehat{\underline{\beta}}$ than does the fact that the patient died rather than was censored. Plots of these quantities against the case number (index) or against their respective covariate $x_k^{(j)}$ are used to gauge the influence of the kth observation on the jth coefficient. See Figure 6.16.

Remarks:

1 The S function `resid(fit,type="dfbetas")` computes dfbeta divided by the s.e.'s for the components of $\widehat{\underline{\beta}}$, where `fit` is the `coxph` object.

2 Collett (1994) calls these standardized delta-beta's.

3 There are a number of alternate expressions to expression (6.22). For example, see pages 359 through 365 in Klein & Moeschberger (1997).

4 This measure is analogous to the measures of influence for ordinary linear regression developed by Belsley *et al.* (1980) and Cook and Weisberg (1982).

6.3.7 CNS lymphoma example: checking the adequacy of the PH model

We apply some model checking techniques on the final reduced model cns2.coxint6, page 110.

```
 # Cox-Snell residuals for overall fit of a model are not
 # provided directly by coxph object. You can derive them
 # from the martingale residuals which are the default
 # residuals.
```

Figure 6.11:

```
> attach(cns2)
> rc <- abs(STATUS - cns2.coxint6$residuals) # Cox-Snell
                                             # residuals!
> km.rc <- survfit(Surv(rc,STATUS) ~ 1)
> summary.km.rc <- summary(km.rc)
> rcu <- summary.km.rc$time # Cox-Snell residuals of
                            # uncensored points.
> surv.rc <- summary.km.rc$surv
> plot(rcu,-log(surv.rc),type="p",pch=".",
xlab="Cox-Snell residual rc",ylab="Cumulative hazard on rc")
> abline(a=0,b=1); abline(v=0); abline(h=0)

 # The martingale residual plot to check functional form of
 # covariate follows.
```

Figure 6.12:

```
> fit <- coxph(Surv(B3TODEATH,STATUS) ~ GROUP+SEX+AGE60+
                SEX:AGE60)
> scatter.smooth(cns2$KPS.PRE.,resid(fit),type="p",pch=".",
     xlab="KPS.PRE.",ylab="Martingale residual")
```

```
 # The deviance residual plots to detect outliers follow:
```

Figure 6.13:

```
> dresid <- resid(cns2.coxint6,type="deviance") # deviance
                                                # residual
> plot(dresid,type="p",pch=".")
> abline(h=0)
> plot(B3TODEATH,dresid,type="p",pch=".")
> abline(h=0)
> plot(GROUP,dresid,type="p",pch=".")
> abline(h=0)
> plot(SEX,dresid,type="p",pch=".")
> abline(h=0)
> plot(AGE60,dresid,type="p",pch=".")
> abline(h=0)
> plot(KPS.PRE.,dresid,type="p",pch=".")
> abline(h=0)

 # Schoenfeld residuals to examine fit and detect outlying
 # covariate values
```

Figure 6.14:

```
> detail <- coxph.detail(cns2.coxint6) # detailed coxph object
> time <- detail$y[,2] # ordered times including censored ones
> status <- detail$y[,3] # censoring status
> sch <- resid(cns2.coxint6,type="schoenfeld") # Schoenfeld
                                               # residuals
> plot(time[status==1],sch[,1],xlab="Ordered survival time",
   ylab="Schoenfeld residual for KPS.PRE.") # time[status==1]
         # is the ordered uncensored times and sch[,1] is the
         # Schoenfeld resid's for KPS.PRE.

 # The scaled Schoenfeld residuals and the Grambsch and
 # Therneau's test for time-varying coefficients to assess
 # PH assumption follow:
```

Figure 6.15:

```
> PH.test <- cox.zph(cns2.coxint6)
> PH.test
             rho chisq     p
 KPS.PRE.  0.0301 0.025 0.874
    GROUP  0.1662 1.080 0.299
      SEX  0.0608 0.103 0.748
    AGE60 -0.0548 0.114 0.736
```

```
SEX:AGE60  0.0872 0.260 0.610
   GLOBAL      NA 2.942 0.709

> par(mfrow=c(3,2)); plot(PH.test)

 # The dfbetas is approximately the change in the
 # coefficients scaled by their standard error. This
 # assists in detecting influential observations on
 # the estimated beta coefficients.
```

Figure 6.16:

```
> par(mfrow=c(3,2))
> bresid <- resid(cns2.coxint6,type="dfbetas")
> index <- seq(1:58)
> plot(index,bresid[,1],type="h",ylab="scaled change in coef",
         xlab="observation")
> plot(index,bresid[,2],type="h",ylab="scaled change in coef",
         xlab="observation")
> plot(index,bresid[,3],type="h",ylab="scaled change in coef",
         xlab="observation")
> plot(index,bresid[,4],type="h",ylab="scaled change in coef",
         xlab="observation")
> plot(index,bresid[,5],type="h",ylab="scaled change in coef",
         xlab="observation")

 # For the sake of comparison, we consider the scaled
 # Schoenfeld residuals and the test for time-varying
 # coefficients for the main effects model cns2.cox3.
```

Figure 6.17:

```
> PHmain.test <- cox.zph(cns2.cox3)
> PHmain.test
           rho  chisq     p
KPS.PRE. 0.0479 0.0671 0.796
   GROUP 0.1694 1.1484 0.284
     SEX 0.2390 1.9500 0.163
  GLOBAL     NA 3.1882 0.364

> par(mfrow=c(2,2)); plot(PHmain.test)
> detach()
```

Results:

- We see from the Cox-Snell residual plot, Figure 6.11, that the final model

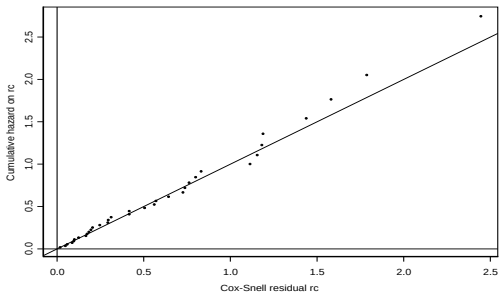


Figure 6.11 *Cox-Snell residuals to assess overall model fit.*

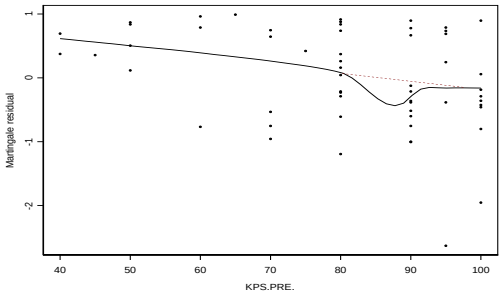


Figure 6.12 *Martingale residuals to look for best functional form of the continuous covariate* KPS.PRE.

gives a reasonable fit to the data. Overall the residuals fall on a straight line with an intercept zero and a slope one. Further, there are no large departures from the straight line and no large variation at the right-hand tail.

- In the plot of the Martingale residuals, Figure 6.12, there appears to be a bump for KPS.PRE. between 80 and 90. However, the lines before and after the bump nearly coincide. Therefore, a linear form seems appropriate for KPS.PRE. There are occasions where a discretized, perhaps dichotomized, version of a continuous variable is more appropriate and informative. See Chapter 6.3.8 of our book for an extensive cut point analysis conducted in the next subsection.
- The deviance residual plot, Figure 6.13, shows a slight tendency for larger survival times to have negative residuals. This suggests that the model overestimates the chance of dying at large times. However, there is only one possible outlier at the earliest time and this may not cause concern about the adequacy of the model. All the other plots in the same figure

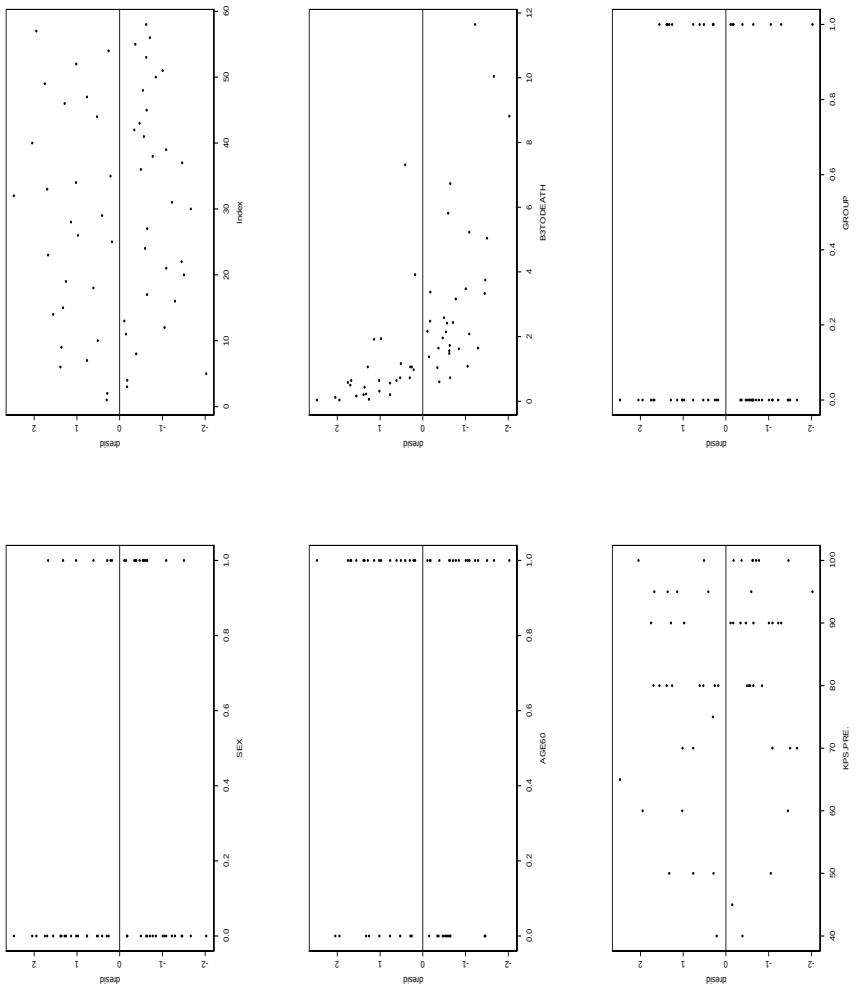


Figure 6.13 *Deviance residuals to check for outliers.*

show that the residuals are symmetric around zero and there is at most one possible outlier.

- In Figure 6.14, the subjects with the largest absolute valued Schoenfeld residuals for
  KPS.PRE. are 40, 8, 35, and 11. These subjects have very early failure times .125, .604, .979, and 1.375 years and are the patients who have either the largest or the smallest KPS.PRE. values. Thus, these residuals do not cause specific concern. The plots for the other covariates are not shown here. But

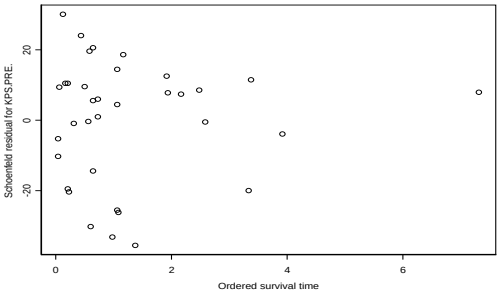


Figure 6.14 *Schoenfeld residuals for* KPS.PRE. *against ordered survival times.*

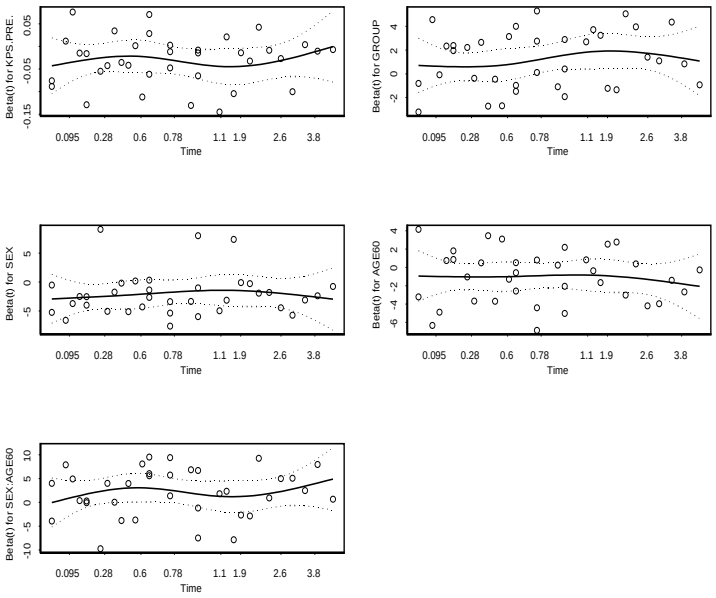


Figure 6.15 *Diagnostic plots of the constancy of the coefficients in cns2.coxint6. Each plot is of a component of* $\widehat{\beta}(t)$ *against ordered time. A spline smoother is shown, together with* ± 2 *standard deviation bands.*

all of them show no large residuals. Therefore, the PH assumption seems to be appropriate.

- The results from the test for constancy of the coefficients based on scaled Schoenfeld residuals indicate the PH assumption is satisfied by all five covariates in the model with all p-values being at least 0.299. Figure 6.15 also supports that the PH assumption is satisfied for all the covariates in the model.
- The plot of the dfbetas, Figure 6.16, shows that most of the changes in the

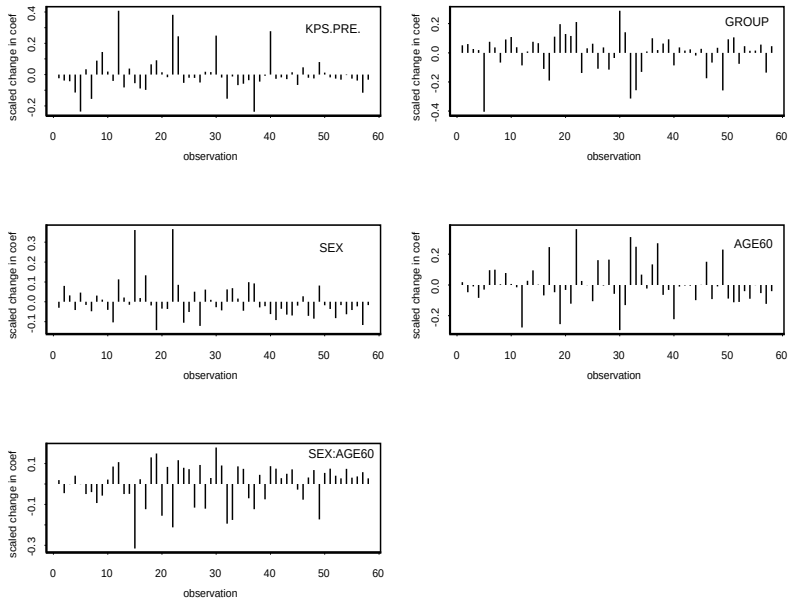


Figure 6.16 *The dfbetas to detect influential observations on the five estimated coefficients corresponding to the predictors.*

regression coefficients are less than .3 s.e.'s of the coefficients and all others are less than .4 s.e.'s. Therefore, we conclude that there are no influential subjects.

- For the sake of comparison, we consider the main effects model cns2.cox3, page 112, as well. Although the results from the test for constancy of the coefficients indicate that the PH assumption is satisfied by all three covariates in the model with all p-values being at least 0.16, Figure 6.17 gives some mild evidence that the PH assumption may be violated for the GROUP and SEX variables. This results from the fact that in this model there are no interaction effect terms when there is a significant interaction effect between SEX and AGE60 as evidenced by the model cns2.coxint6. This again tells us how important it is to consider interaction effects in modelling.

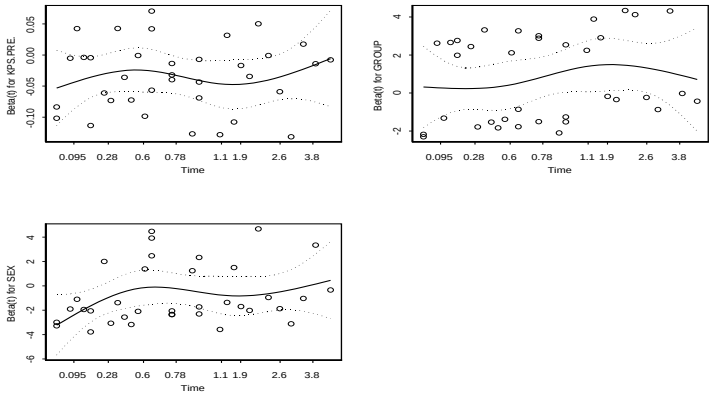


Figure 6.17 *Diagnostic plots of the constancy of the coefficients in cns2.cox3. Each plot is of a component of* $\widehat{\beta}(t)$ *against ordered time. A spline smoother is shown, together with* ± 2 *standard deviation bands.*

References

Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.*, **6**, 701 – 726.

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley-Interscience.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, **AC-19**, 716 – 723.

Andersen, P.K and Gill, R.R. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.*, **10**, 1100 – 1120.

Babu, G.J. and Feigelson, E.D. (1996). *Astrostatistics*, London: Chapman & Hall.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Bickel, P.J. and Doksum, K.A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics, Vol.I, 2nd Edition*. New Jersey: Prentice-Hall, Inc.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.*, **2**, 437 – 453.

Cain, K.C. and Lange, N.T. (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics*, **40**, 493 – 499.

Caplehorn, J., et al. (1991). Methadone dosage and retention of patients in maintenance treatment. *Med. J. Australia*, **154**, 195 – 199.

Collett, D. (1999). *Modelling Survival Data in Medical Research*. London: Chapman & Hall/CRC.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman & Hall.

Cox, D.R. (1959). The analysis of exponentially distributed life-times with two types of failure. *J.R. Statist. Soc.*, **B**, **21**, 411 – 421.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *J.R. Statist. Soc.*, **B**, **34**, 187 – 220.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, **62**, 269 – 276.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman & Hall.

Cox, D.R. and Snell, E.J. (1968). A general definition of residuals (with discussion). *J.R. Statist. Soc.*, **A**, **30**, 248 – 275.

Dahlborg, S.A., Henner, W. D., Crossen, J.R., Tableman, M., Petrillo, A., Braziel, R. and Neuwelt, E.A. (1996). Non-AIDS primary CNS lymphoma: the first example of a durable response in a primary brain tumor using enhanced chemotherapy delivery without cognitive loss and without radiotherapy. *The Cancer Journal from Scientific American*, **2**, 166 – 174.

Davison, A.C. and Hinkley, D.V. (1999). *Bootstrap Methods and their Application*. London: Cambridge University Press.

DeGroot, M.H. (1986). *Probability and Statistics, 2nd Edition.* New York: Addison-Wesley.

Edmunson, J.H., Fleming, T.R., Decker, D.G., Malkasian, G.D., Jefferies, J.A., Webb, M.J., and Kvols, L.K. (1979). Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma vs. minimal residual disease. *Cancer Treatment Reports*, **63**, 241−47.

Efron, B. (1967). The two sample problem with censored data. *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **4**. New York: Prentice-Hall, 831 − 853.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1 − 26.

Efron, B. (1998). R. A. Fisher in the 21st Century. *Statist. Sci.*, **13**, 95 − 122.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *Amer. Statist.*, **37**, 36 − 48.

Efron, B. and Petrosian, V. (1992). A simple test of independence for truncated data with applications to red shift surveys, *Astrophys. J.*, **399**, 345 − 352.

Efron, B. and Tibshirani (1993). *An Introduction to the Bootstrap.* London: Chapman & Hall.

Embury, S.H., Elias, L., Heller, P.H., Hood, C.E., Greenberg, P.L., and Schrier, S.L. (1977). Remission maintenance therapy in acute myelogenous leukemia. *Western Journal of Medicine*, **126**, 267 − 272.

Finkelstein, D.M., Moore, D.F., and Schoenfeld, D.A. (1993). A proportional hazards model for truncated AIDS data. *Biometrics*, **49**, 731 − 740.

Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Galton, F. (1889). *Natural Inheritance.* London: Macmillan.

Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, **52**, 203 − 223.

Gooley, T.A., Leisenring, W., Crowley, J., and Storer, B.E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statist. Med.*, **18**, 695 − 706.

Gooley, T.A., Leisenring, W., Crowley, J.C., and Storer, B.E. (2000). Why the Kaplan-Meier fails and the cumulative incidence function succeeds when estimating failure probabilities in the presence of competing risks. Editor: J.C. Crowley. *Handbook of Statistics in Clinical Oncology.* New York: Marcel Dekker, Inc., 513 − 523.

Grambsch, P. and Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515 − 526.

Gray, R.J. (2002). `cmprsk` library, competing risks library for S-PLUS. http://biowww.dfci.harvard.edu/~gray/.

Gray, R.J. (2002). `cmprsk.zip`, competing risks R library. http://www.r-project.org/~CRAN/.

Greenwood, M. (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects*, **33**, 1 − 26, London: Her Majesty's Stationery Office.

Gutenbrunner, C., Jurečková, J., Koenker, R., and Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *J. Nonparametric Statist.*, **2**, 307−331.

Hoel, D.G. and Walburg, H.E., Jr. (1972). Statistical analysis of survival experiments. *J. Natl. Cancer Inst.*, **49**, 361 − 372.

Hogg, R.V. and Craig, A.T. (1995). *Introduction to Mathematical Statistics, 5th Edition*. New Jersey: Prentice Hall.

Hosmer, D.W. Jr. and Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: Wiley.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**, 457 − 481.

Keiding, N. (1992). Independent delayed entry, in *Survival Analysis: State of the Art*, J.P. Klein and P. Goel, eds. Boston: Kluwer Academic Publishers, 309 − 326.

Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.

Kleinbaum, D.G. (1995). *Survival Analysis: A Self-Learning Text*. New York: Springer.

Koenker, R. (1994). Confidence intervals for regression quantiles, in *Asymptotic Statistics: Proc. 5th Prague Symposium*, editors: P. Mandl and M. Hušková. Heidelberg: Physica-Verlag.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33−50.

Koenker, R. and d'Orey, V. (1987). Computing regression quantiles. *Appl. Statist.*, **36**, 383 − 393.

Koenker, R., and Geling, O. (2001). Reappraising Medfly longevity: a quantile regression survival analysis. *J. Amer. Statist. Assoc.*, **96**, 458 − 468.

Koenker, R. and Machado, J. (1999). Goodness of fit and related inference procedures for quantile regression. *J. Amer. Statist. Assoc.*, **94**, 1296 − 1310.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: Wiley.

Lee, E.T. (1992). *Statistical Methods for Survival Data Analysis, 2nd Edition*. New York: John Wiley & Sons.

Leiderman, P.H., Babu, D., Kagia, J., Kraemer, H.C., and Leiderman, G.F. (1973). African infant precocity and some social influences during the first year. *Nature*, **242**, 247 − 249.

Lenneborg, C.E. (2000). *Data Analysis by Resampling: Concepts and Applications*. Pacific Grove: Duxbury.

MathSoft (1999). *S-PLUS 2000-Guide to Statistics*. Seattle, WA: MathSoft, Inc.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. National Cancer Institute*, **22**, 719 − 322.

Miller, R.G. (1981). *Survival Analysis*. New York: Wiley.

Morrell, C.H. (1999). Simpson's Paradox: an example from a longitudinal study in South Africa. *J. Statist. Ed.*, **7**, **3**.

Nelson, W.B. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**, 945 − 965.

Nelson, W.B. and Hahn, G.B. (1972). Linear estimation of regression relationships from censored data, part 1−simple methods and their applications (with discussion). *Technometrics*, **14**, 247 − 276.

Peterson, A.V. (1975). *Nonparametric Estimation in the Competing Risks Problem*. Ph.D.thesis, Department of Statistics, Stanford University.

Peto, R. (1973). Empirical survival curves for interval censored data. *Appl. Statist.*, **22**, 86 − 91.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *J.R. Statist. Soc.*, **A**, **135**, 185 − 198.

Pike, M.C. (1966). A method of analysis of certain class of experiments in carcinogenesis. *Biometrics*, **22**, 142 − 161.

Portnoy, S. (1991a). Asymptotic behavior of the number of regression quantile breakpoints. *SIAM J. Sci. Stat. Comp.*, **12**, 867 − 883.

Portnoy, S. (1991b). Behavior of regression quantiles in non-stationary, dependent cases. *J. Multivar. Anal.*, **38**, 100 − 113.

Portnoy, S. (2003). Censored regression quantiles. *J. Amer. Statist. Assoc.*, to appear.

Portnoy, S. and Koenker, R. (1997). The Gaussian Hare and the Laplacian Tortoise: computability of squared-error vs. absolute-error estimators. *Statist. Sci.*, **12**, 279 − 300.

Prentice, R.L. and Marek, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics*, **35**, 861 − 867.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, **11**, 453−466.

Reid, N. (1994). A conversation with Sir David Cox, *Statist. Sci.*, **9**, 439 − 455.

Ross, S.M. (2000). *Introduction to Probability Models, 7th Edition.* Orlando: Academic Press, Inc.

Schoenfeld, D.A. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239 − 241.

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *J.R. Statist. Soc.*, **B**, **13**, 238 − 241.

Smith, P.J. (2002). *Analysis of Failure and Survival Data.* Boca Raton: Chapman & Hall/CRC.

Therneau, T.M., Grambsch, P.M., and Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika*, **69**, 239 − 241.

Tsai, W.Y., Jewell, N.P., and Wang, M.C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, **74**, 883 − 886.

Tsai, W.Y. (1990). The assumption of independence of truncation time and failure Time. *Biometrika*, **77**, 169 − 177.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proc. Natl. Acad. Sci.*, **72**, 20 − 22.

Tsuang, M.T. and Woolson, R.F. (1977). Mortality in patients with schizophrenia, mania and depression. *British Journal of Psychiatry*, **130**, 162 − 166.

van der Vaart, A.W. (1998). *Asymptotic Statistics.* London: Cambridge University Press.

Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S, 4th Edition.* New York: Springer-Verlag, Inc.

Woolson, R.F. (1981). Rank tests and a one-sample log rank test for comparing observed survival data to a standard population. *Biometrics*, **37**, 687 − 696.