Modelling in survival analysis

Christos T Nakas, Ph.D.
Adapted from the course notes of Prof. Yiannoutsos, Indiana University

February 2, 2022

Introduction

Survival Analysis typically focuses on **time to event** data. In the most general sense, it consists of techniques for positive-valued random variables, such as

- time to death
- time to onset (or relapse) of a disease
- length of stay in a hospital
- duration of a strike

Some useful references

- Collett: *Modelling Survival Data in Medical Research*
- Cox and Oakes: *Analysis of Survival Data*
- Kleinbaum: *Survival Analysis: A self-learning text*
- Klein & Moeschberger: *Survival Analysis: Techniques for censored and truncated data*

Definitions and notation

Failure time random variables are always **non-negative**.

That is, if we denote the failure time by T, then $T \geq 0$.

T can either be **discrete** (taking a finite set of values, e.g. $a_1, a_2, \ldots, a_n$) or **continuous** (defined on $(0, \infty)$).

A random variable X is called a **censored failure time random variable** if $X = \min(T, U)$, where U is a non-negative censoring variable.

Defining a time random variable

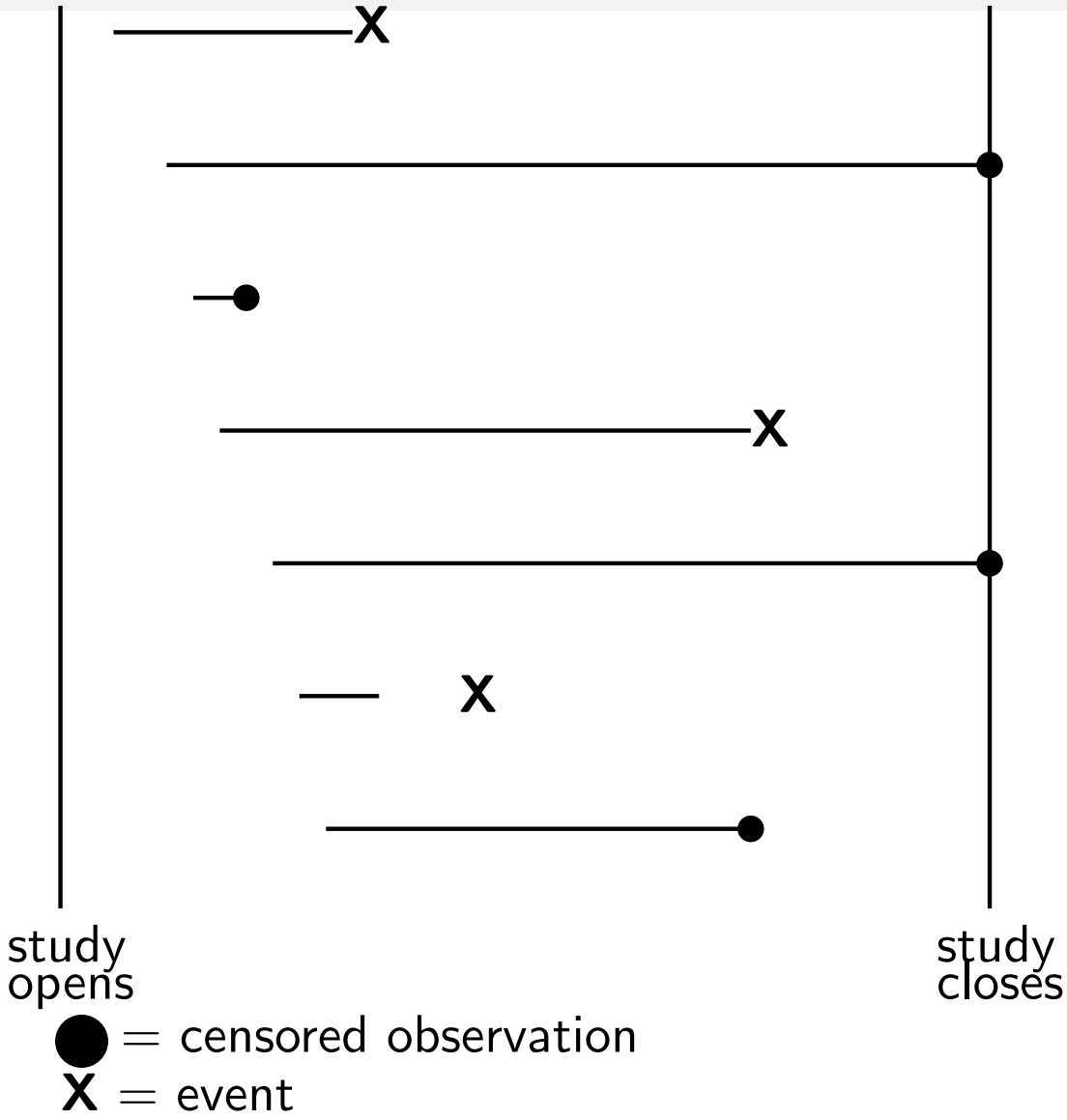
In order to define a failure time random variable, we need:

(1) an unambiguous **time origin**
(e.g. randomization to clinical trial, purchase of car)

(2) a **time scale**
(e.g. real time (days, years), mileage of a car)

(3) definition of the **event**
(e.g. death, need a new car transmission)

Illustration of survival data



The illustration of survival data on the previous page shows several features which are typically encountered in analysis of survival data:

- individuals do not all enter the study at the same time
- when the study ends, some individuals still haven't had the event yet
- other individuals drop out or get lost in the middle of the study, and all we know about them is the last time they were still "free" of the event

The first feature is referred to as **"staggered entry"**
The last two features relate to **"censoring"** of the failure time events.

Types of censoring

Right-censoring

We have right censoring when only the r.v. $X_i = \min(T_i, U_i)$ is observed due to

- loss to follow-up
- drop-out
- study termination

We call this right-censoring because the true unobserved event is to the right of our censoring time; i.e., all we know is that the event has not happened at the end of follow-up.

Failure events

In addition to observing X_i, we also get to see the **failure indicator**:

$$\delta_i = \begin{cases} 1 & \text{if} \quad T_i \leq U_i \\ 0 & \text{if} \quad T_i > U_i \end{cases}$$

Some software packages instead assume we have a **censoring indicator**:

$$c_i = \begin{cases} 0 & \text{if} \quad T_i \leq U_i \\ 1 & \text{if} \quad T_i > U_i \end{cases}$$

Right-censoring is the most common type of censoring assumption we will deal with in survival analysis.

Left-censored data

Left censoring

We have left censoring when we can only observe $Y_i = \max(T_i, U_i)$ and the failure indicators:

$$\epsilon_i = \begin{cases} 1 & \text{if} \quad U_i \leq T_i \\ 0 & \text{if} \quad U_i > T_i \end{cases}$$

e.g. In studies of time to HIV seroconversion, some of the enrolled subjects have already seroconverted at entry into the study - they are left-censored.

Interval-censored data

Interval censoring

We have interval censoring when we observe (L_i, R_i) where $T_i \in (L_i, R_i)$

ex #1: Time to prostate cancer, observe longitudinal PSA measurements

ex #2: Time to undetectable viral load in AIDS studies, based on measurements of viral load taken at each clinic visit

Independent versus informative censoring

We say censoring is **independent** (non-informative) if U_i is independent of T_i.

- **ex.1** If U_i is the planned end of the study (say, 2 years after the study opens), then it is usually independent of the event times
- **ex.2** If U_i is the time that a patient drops out of the study because they've gotten much sicker and/or had to discontinue taking the study treatment, then U_i and T_i are probably not independent

What does "independent censoring" mean?

In plain language, independent censoring simply means

An individual censored at U should be representative of all subjects who survive to U.

This means that censoring at U *could* depend on prognostic characteristics measured at baseline, but that among all those with the same baseline characteristics, the probability of censoring prior to or at time U should be the same.

Informative censoring

Censoring is considered **informative** if the distribution of U_i contains any information about the parameters characterizing the distribution of T_i.

Types of censoring times

Suppose we have a sample of observations on n people:

$$(T_1, U_1), (T_2, U_2), ..., (T_n, U_n)$$

There are three main types of censoring times:

- **Type I:** All the U_i's are the same
  e.g. animal studies, all animals sacrificed after 2 years
- **Type II:** $U_i = T_{(r)}$, the time of the rth failure.
  e.g. animal studies, stop when 4/6 have tumors
- **Random:** the U_i's are random variables, δ_i's are failure indicators:

$$\delta_i = \begin{cases} 1 & \text{if} \quad T_i \leq U_i \\ 0 & \text{if} \quad T_i > U_i \end{cases}$$

Example A. Fecundability

Women who had recently given birth were asked to recall how long it took them to become pregnant, and whether or not they smoked during that time. The outcome of interest is time to pregnancy (in menstrual cycles).

Cycle	Smokers	Non-smokers
1	29	198
2	16	107
3	17	55
4	4	38
5	3	18
6	9	22
7	4	7
8	5	9
9	1	5
10	1	3
11	1	6
12	3	6
12+	7	12

Example A. Duration of nursing home stay

Morris et al., *Case Studies in Biometry*, Ch 12

The National Center for Health Services Research studied 36 for-profit nursing homes to assess the effects of different financial incentives on length of stay.

"Treated" nursing homes received higher per diems for Medicaid patients, and bonuses for improving a patient's health and sending them home.

The study included 1601 patients admitted between May 1, 1981 and April 30, 1982.

Example B: MAC Prevention Clinical Trial

ACTG 196 was a randomized clinical trial to study the effects of combination regimens on prevention of MAC (*mycobacterium avium complex*), one of the most common OIs in AIDS patients.
The **treatment regimens** were:

- clarithromycin (new)
- rifabutin (standard)
- clarithromycin plus rifabutin

Other characteristics of trial:

- Patients enrolled between April 1993 and February 1994
- Follow-up ended August 1995
- In February 1994, rifabutin dosage was reduced from 3 pills/day (450mg) to 2 pills/day (300mg) due to concern over **uveitis**[1]

The main intent-to-treat analysis compared the 3 treatment arms without adjusting for this change in dosage.

[1] *Uveitis* is an adverse experience resulting in inflammation of the uveal tract in the eyes (about 3-4% of patients reported uveitis).

More Definitions and Notation

There are several equivalent ways to characterize the probability distribution of a survival random variable. Some of these are familiar; others are special to survival analysis. We will focus on the following terms:

- The density function $f(t)$
- The survivor function $S(t)$
- The hazard function $\lambda(t)$
- The cumulative hazard function $\Lambda(t)$

The survival density function

Suppose that T takes values in $a_1, a_2, \ldots, a_n$. Then, the survival density function (or Probability Mass Function) for discrete r.v.'s is defined as:

$$\begin{aligned} f(t) &= Pr(T = t) \\ &= \begin{cases} f_j & \text{if} \quad t = a_j, j = 1, 2, \ldots, n \\ 0 & \text{if} \quad t \neq a_j, j = 1, 2, \ldots, n \end{cases} \end{aligned}$$

The survival density function for continuous r.v.'s

The survival density function for continuous r.v.'s is defined as follows:

$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leq T \leq t + \Delta t)$$

The Survivorship Function

The survivorship function is as follows: $S(t) = P(T \geq t)$.

In other settings, the cumulative distribution function, $F(t) = P(T \leq t)$, is of interest.

In survival analysis, our interest tends to focus on the survival function, $S(t)$.

Survivorship function for continuous and discrete r.v.'s

For a continuous random variable:

$$S(t) = \int_t^{\infty} f(u)du$$

For a discrete random variable:

$$\begin{aligned} S(t) &= \sum_{u \geq t} f(u) \\ &= \sum_{a_j \geq t} f(a_j) = \sum_{a_j \geq t} f_j \end{aligned}$$

Notes

- From the definition of $S(t)$ for a continuous variable, $S(t) = 1 - F(t)$ as long as $f(t)$ is absolutely continuous
- For a discrete variable, we have to decide what to do if an event occurs exactly at time t; i.e., does that become part of $F(t)$ or $S(t)$?
- To get around this problem, several books define $S(t) = Pr(T > t)$, or else define $F(t) = Pr(T < t)$ (eg. Collett)

The Hazard function

The Hazard Function $\lambda(t)$.

Sometimes called an *instantaneous failure rate*, the *force of mortality*, or the *age-specific failure rate*.

The hazard function for continuous r.v.'s

The hazard function is defined as follows for continuous random variables:

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leq T < t + \Delta t | T \geq t) \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{Pr([t \leq T < t + \Delta t] \bigcap [T \geq t])}{Pr(T \geq t)} \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{Pr(t \leq T < t + \Delta t)}{Pr(T \geq t)} \\
&= \frac{f(t)}{S(t)}
\end{aligned}$$

The hazard function for discrete random variables

The hazard function is defined as follows for discrete random variables:

$$\begin{aligned}
\lambda(a_j) \equiv \lambda_j &= Pr(T = a_j | T \geq a_j) \\
&= \frac{P(T = a_j)}{P(T \geq a_j)} \\
&= \frac{f(a_j)}{S(a_j)} \\
&= \frac{f(t)}{\sum_{k:a_k \geq a_j} f(a_k)}
\end{aligned}$$

The Cumulative Hazard function

The Cumulative Hazard function $\Lambda(t)$ is defined as follows:

- **Continuous random variables:**

$$\Lambda(t) = \int_0^t \lambda(u)du$$

- **Discrete random variables:**

$$\Lambda(t) = \sum_{k:a_k<t} \lambda_k$$

Comment

The cumulative hazard does not have a very intuitive interpretation.

However, it turns out to be very useful for certain graphical assessments:

- consistency with certain parametric models
- evaluation of proportional hazards assumption for Cox models

Relationship between $S(t)$ and $\lambda(t)$

We've already shown that, for a continuous r.v.

$$\lambda(t) = \frac{f(t)}{S(t)}$$

For a left-continuous survivor function $S(t)$, we can show:

$$f(t) = -S'(t) \quad \text{or} \quad S'(t) = -f(t)$$

We can use this relationship to show that:

$$\begin{aligned} -\frac{d}{dt}[\log S(t)] &= -\left(\frac{1}{S(t)}\right) S'(t) \\ &= -\frac{-f(t)}{S(t)} = \frac{f(t)}{S(t)} \end{aligned}$$

So another way to write $\lambda(t)$ is as follows:

$$\lambda(t) = -\frac{d}{dt}[\log S(t)]$$

Relationship between $S(t)$ and $\Lambda(t)$: Continuous case

In the case of continuous time we have

$$\begin{aligned}
\Lambda(t) &= \int_0^t \lambda(u)du \\
&= \int_0^t \frac{f(u)}{S(u)}du \\
&= \int_0^t -\frac{d}{du}\log S(u)du \\
&= -\log S(t) + \log S(0) \\
&\Rightarrow S(t) = e^{-\Lambda(t)}
\end{aligned}$$

Relationship between $S(t)$ and $\Lambda(t)$: Discrete case

Suppose that $a_j < t \leq a_{j+1}$. Then

$$\begin{aligned} S(t) &= P(T \geq a_1, T \geq a_2, \ldots, T \geq a_{j+1}) \\ &= P(T \geq a_1)P(T \geq a_2 | T \geq a_1) \cdots P(T \geq a_{j+1} | T \geq a_j) \\ &= (1-\lambda_1) \times \cdots \times (1-\lambda_j) \\ &= \prod_{k:a_k<t} (1-\lambda_k) \end{aligned}$$

Cox defines $\Lambda(t) = \sum_{k:a_k<t} \log(1-\lambda_k)$ so that $S(t) = e^{-\Lambda(t)}$ in the discrete case, as well.

Some hazard shapes seen in applications

Hazard shapes seen in application are as follows:

- **increasing**

  e.g. aging after 65

- **decreasing**

  e.g. survival after surgery

- **bathtub**

  e.g. age-specific mortality

- **constant**

  e.g. survival of patients with advanced chronic disease

Estimating the survival or hazard function

We can estimate the survival (or hazard) function in two ways:

- by specifying a parametric model for $\lambda(t)$ based on a particular density function $f(t)$
- by developing an empirical estimate of the survival function (i.e., non-parametric estimation)

Under no censoring

If no censoring:

The empirical estimate of the survival function, $\tilde{S}(t)$, is the proportion of individuals with event times greater than t.

Estimation under censoring

With censoring:

If there are censored observations, then $\tilde{S}(t)$ is not a good estimate of the true $S(t)$, so other non-parametric methods must be used to account for censoring (life-table methods, Kaplan-Meier estimator)

Measuring Central Tendency in Survival

- **Mean survival** - call this μ

$$\begin{aligned} \mu &= \int_0^\infty uf(u)du \quad \text{for continuous } T \\ &= \sum_{j=1}^{n} a_j f_j \quad \text{for discrete } T \end{aligned}$$

- **Median survival** - call this τ, is defined by

$$S(\tau) = 0.5$$

Similarly, any other percentile could be defined.

In practice, we don't usually hit the median survival at exactly one of the failure times. In this case, the estimated median survival is the *smallest* time τ such that

$$\hat{S}(\tau) \leq 0.5$$

Some Parametric Survival Distributions

The **Exponential** distribution (1 parameter)

$$
\begin{aligned}
f(t) &= \lambda e^{-\lambda t} \text{ for } t \geq 0 \\
S(t) &= \int_t^{\infty} f(u)du = e^{-\lambda t} \\
\lambda(t) &= \frac{f(t)}{S(t)} \\
&= \lambda \qquad \text{constant hazard!} \\
\Lambda(t) &= \int_0^t \lambda(u)\, du \\
&= \int_0^t \lambda\, du = \lambda t
\end{aligned}
$$

Check: Does $S(t) = e^{-\Lambda(t)}$?

Characteristics of the exponential distribution

median:

solve $0.5 = S(\tau) = e^{-\lambda\tau}$:

$$\Rightarrow \tau = \frac{-\log(0.5)}{\lambda}$$

mean:

$$\int_0^\infty u\lambda e^{-\lambda u} du = \frac{1}{\lambda}$$

Weibull distribution (2 parameters)

Generalizes exponential:

$$
\begin{aligned}
S(t) &= e^{-\lambda t^{\kappa}} \\
f(t) &= \frac{-d}{dt} S(t) = \kappa \lambda t^{\kappa-1} e^{-\lambda t^{\kappa}} \\
\lambda(t) &= \kappa \lambda t^{\kappa-1} \\
\Lambda(t) &= \int_0^t \lambda(u) du = \lambda t^{\kappa}
\end{aligned}
$$

λ - the *scale* parameter
κ - the *shape* parameter

Weibull distribution (cont'd)

The Weibull distribution is convenient because of simple forms. It includes several hazard shapes:

$\kappa = 1 \rightarrow$ constant hazard

$0 < \kappa < 1 \rightarrow$ decreasing hazard

$\kappa > 1 \rightarrow$ increasing hazard

The Rayleigh distribution

The Rayleigh distribution is another 2-parameter generalization of exponential:

$$\lambda(t) = \lambda_0 + \lambda_1 t$$

The Compound Exponential distribution

This is given by the following definition: $T \sim \exp(\lambda)$, $\lambda \sim g$

$$f(t) = \int_0^\infty \lambda e^{-\lambda t} g(\lambda) d\lambda$$

The log-normal and log-logistic distributions

Possible distributions for T obtained by specifying for $\log T$ any convenient family of distributions, e.g.

$\log T \sim$ normal (non-monotone hazard)

$\log T \sim$ logistic

When to use each one

Why use one versus another?

- technical convenience for estimation and inference
- explicit simple forms for $f(t), S(t)$, and $\lambda(t)$.
- qualitative shape of hazard function

One can usually distinguish between a one-parameter model (like the exponential) and two-parameter (like Weibull or log-Normal) in terms of the adequacy of fit to a dataset.

Without a lot of data, it may be hard to distinguish between the fits of various 2-parameter models (i.e., Weibull vs log-normal)

Estimating the Survival Function

Methods for estimating a survivorship function

$$S(t) = Pr(T \geq t)$$

without resorting to parametric methods:

(1) **Kaplan-Meier**

(2) **Life-table** (Actuarial Estimator)

(3) **Cumulative hazard estimator**

The Kaplan-Meier estimator

The Kaplan-Meier (or KM) estimator is probably the most popular approach. It can be justified from several perspectives:

- product limit estimator
- likelihood justification
- redistribute to the right estimator

We will start with an intuitive motivation based on conditional probabilities, then review some of the other justifications.

Motivation

First, consider an example where there is no censoring.

The following are times of remission (weeks) for 21 leukemia patients receiving control treatment (Table 1.1 of Cox & Oakes):

1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

How would we estimate S(10), the probability that an individual survives to time 10 or later?

What about $\tilde{S}(8)$? Is it $\frac{12}{21}$ or $\frac{8}{21}$?

Let's construct a table of $\tilde{S}(t)$:

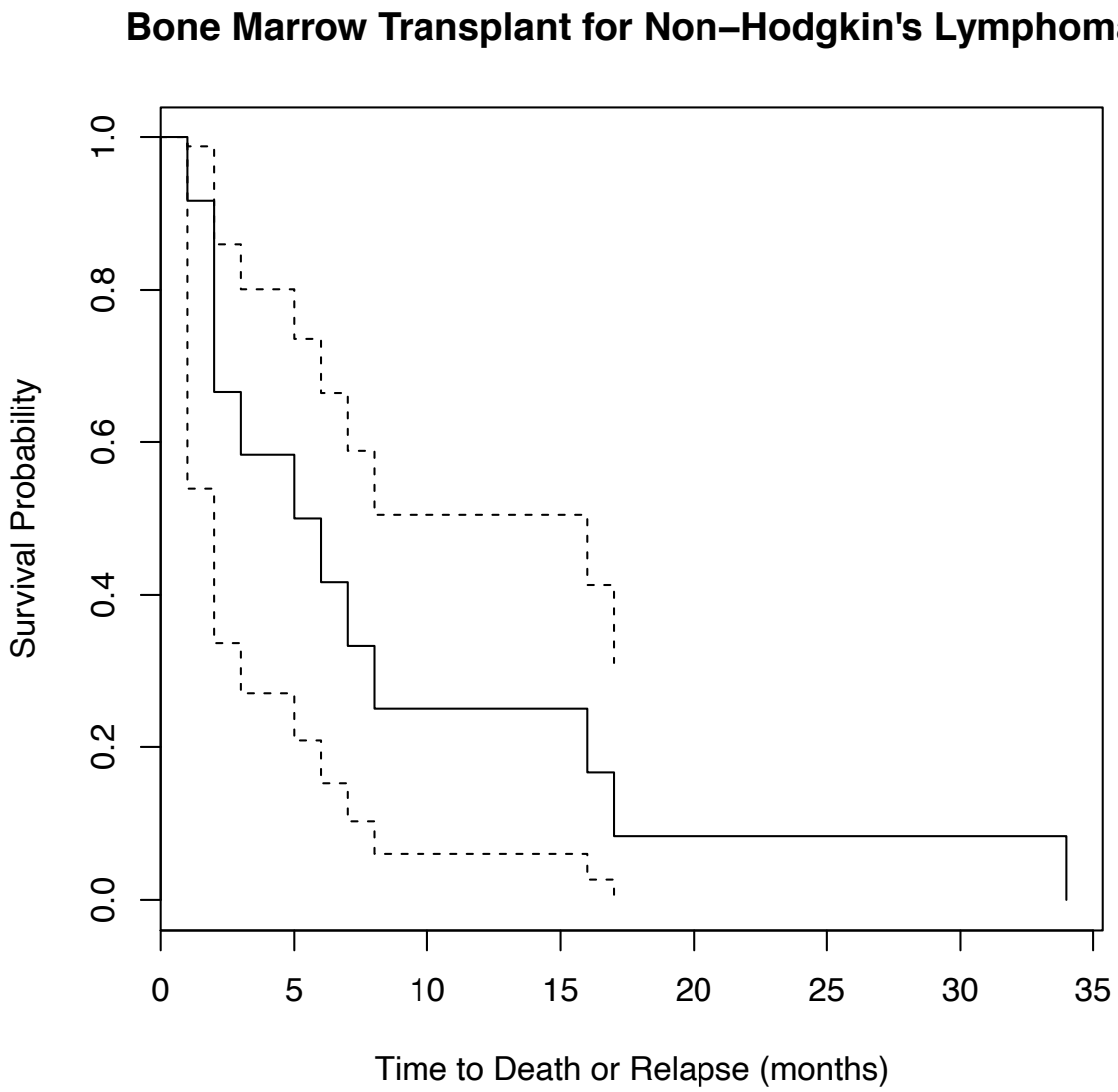
Values of t	$\hat{S}(t)$
$t \leq 1$	$21/21$=1.000
$1 < t \leq 2$	$19/21$=0.905
$2 < t \leq 3$	$17/21$=0.809
$3 < t \leq 4$	
$4 < t \leq 5$	
$5 < t \leq 8$	
$8 < t \leq 11$	
$11 < t \leq 12$	
$12 < t \leq 15$	
$15 < t \leq 17$	
$17 < t \leq 22$	
$22 < t \leq 23$	

The Empirical Survival Function

When there is no censoring, the general formula is:

$$\tilde{S}(t) = \frac{\#\ individuals\ with\ T \geq t}{total\ sample\ size}$$

Example for the leukemia data (control arm)



When there is censoring

Consider the treated group from Table 1.1 of Cox and Oakes:

$$6^+, 6, 6, 6, 7, 9^+, 10^+, 10, 11^+, 13, 16, 17^+$$
$$19^+, 20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 35^+$$

[Note: times with $^+$ are right censored]
We know S(6)= 21/21, because everyone survived at least until time 6 or greater. But, we can't say S(7) = 17/21, because we don't know the status of the person who was censored at time 6.

In a 1958 paper in the *Journal of the American Statistical Association*, Kaplan and Meier proposed a way to nonparametrically estimate S(t), even in the presence of censoring. The method is based on the ideas of **conditional probability**.

A quick review of conditional probability

Suppose A and B are two events.

Then,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Multiplication law of probability: can be obtained from the above relationship, by multiplying both sides by $P(B)$:

$$P(A \cap B) \quad = \quad P(A|B)\, P(B)$$

Extension to more than 2 events

Suppose $A_1, A_2 ... A_k$ are k different events. Then, the probability of all k events happening together can be written as a product of conditional probabilities:

$$\begin{aligned} P(A_1 \cap A_2 ... \cap A_k) &= P(A_k | A_{k-1} \cap ... \cap A_1) \times \\ &\quad \times P(A_{k-1} | A_{k-2} \cap ... \cap A_1) \\ &\quad ... \\ &\quad \times P(A_2 | A_1) \\ &\quad \times P(A_1) \end{aligned}$$

Bringing it all together

Now, let's apply these ideas to estimate $S(t)$:

Suppose $a_k < t \leq a_{k+1}$. Then

$$\begin{aligned} S(t) &= P(T \geq a_{k+1}) \\ &= P(T \geq a_1, T \geq a_2, \ldots, T \geq a_{k+1}) \\ &= P(T \geq a_1) \times \prod_{j=1}^{k} P(T \geq a_{j+1} | T \geq a_j) \\ &= \prod_{j=1}^{k} [1 - P(T = a_j | T \geq a_j)] \\ &= \prod_{j=1}^{k} [1 - \lambda_j] \end{aligned}$$

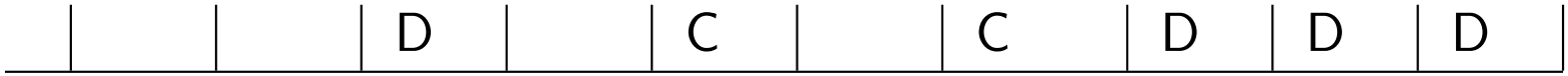
So,

$$\begin{aligned}\hat{S}(t) &\cong \prod_{j=1}^{k}\left(1-\frac{d_j}{r_j}\right)\\ &= \prod_{j:a_j<t}\left(1-\frac{d_j}{r_j}\right)\end{aligned}$$

d_j is the number of deaths at a_j
r_j is the number at risk at a_j

Intuition behind the Kaplan-Meier Estimator

Think of dividing the observed timespan of the study into a series of fine intervals so that there is a separate interval for each time of death or censoring:



Using the law of conditional probability,

$$Pr(T \geq t) \quad = \quad \prod_j Pr(\text{survive } j\text{-th interval } I_j \mid \text{survived to start of } I_j)$$

where the product is taken over all the intervals including or preceding time t.

What can happen

These are possibilities for each interval:

(1) **No events (death or censoring)** - conditional probability of surviving the interval is 1

(2) **Censoring** - assume they survive to the end of the interval, so that the conditional probability of surviving the interval is 1

(3) **Death, but no censoring** - conditional probability of *not* surviving the interval is # deaths (d) divided by # 'at risk' (r) at the beginning of the interval. So the conditional probability of surviving the interval is $1 - (d/r)$.

(4) **Tied deaths and censoring** - assume censorings last to the end of the interval, so that conditional probability of surviving the interval is still $1 - (d/r)$

General Formula for jth interval

It turns out we can write a general formula for the conditional probability of surviving the j-th interval that holds for all 4 cases:

$$1 - \frac{d_j}{r_j}$$

We could use the same approach by grouping the event times into intervals (say, one interval for each month), and then counting up the number of deaths (events) in each to estimate the probability of surviving the interval (this is called the *lifetable estimate*).

However, the assumption that those censored last until the end of the interval wouldn't be quite accurate, so we would end up with a cruder approximation.

The Kaplan-Meier - product-limit - estimator

As the intervals get finer and finer, the approximations made in estimating the probabilities of getting through each interval become smaller and smaller, so that the estimator converges to the true $S(t)$.

This intuition clarifies why an alternative name for the KM is the product limit estimator.

The Kaplan-Meier estimator of the survivorship function (or survival probability) $S(t) = Pr(T \geq t)$ is:

$$\hat{S}(t) = \prod_{j:\tau_j<t} \frac{r_j - d_j}{r_j} = \prod_{j:\tau_j<t} \left(1 - \frac{d_j}{r_j}\right)$$

where,

- $\tau_1, ...\tau_K$ are the K distinct death times observed in the sample
- d_j is the number of deaths at τ_j
- r_j is the number of individuals "at risk" right before the j-th death time (everyone dead or censored at or after that time).
- c_j is the number of censored observations between the j-th and $(j+1)$-st death times. Censorings tied at τ_j are included in c_j

Note: two useful formulas are:

$$(1) \quad r_j = r_{j-1} - d_{j-1} - c_{j-1}$$

$$(2) \quad r_j = \sum_{l \geq j} (c_l + d_l)$$

The Cox and Oakes example

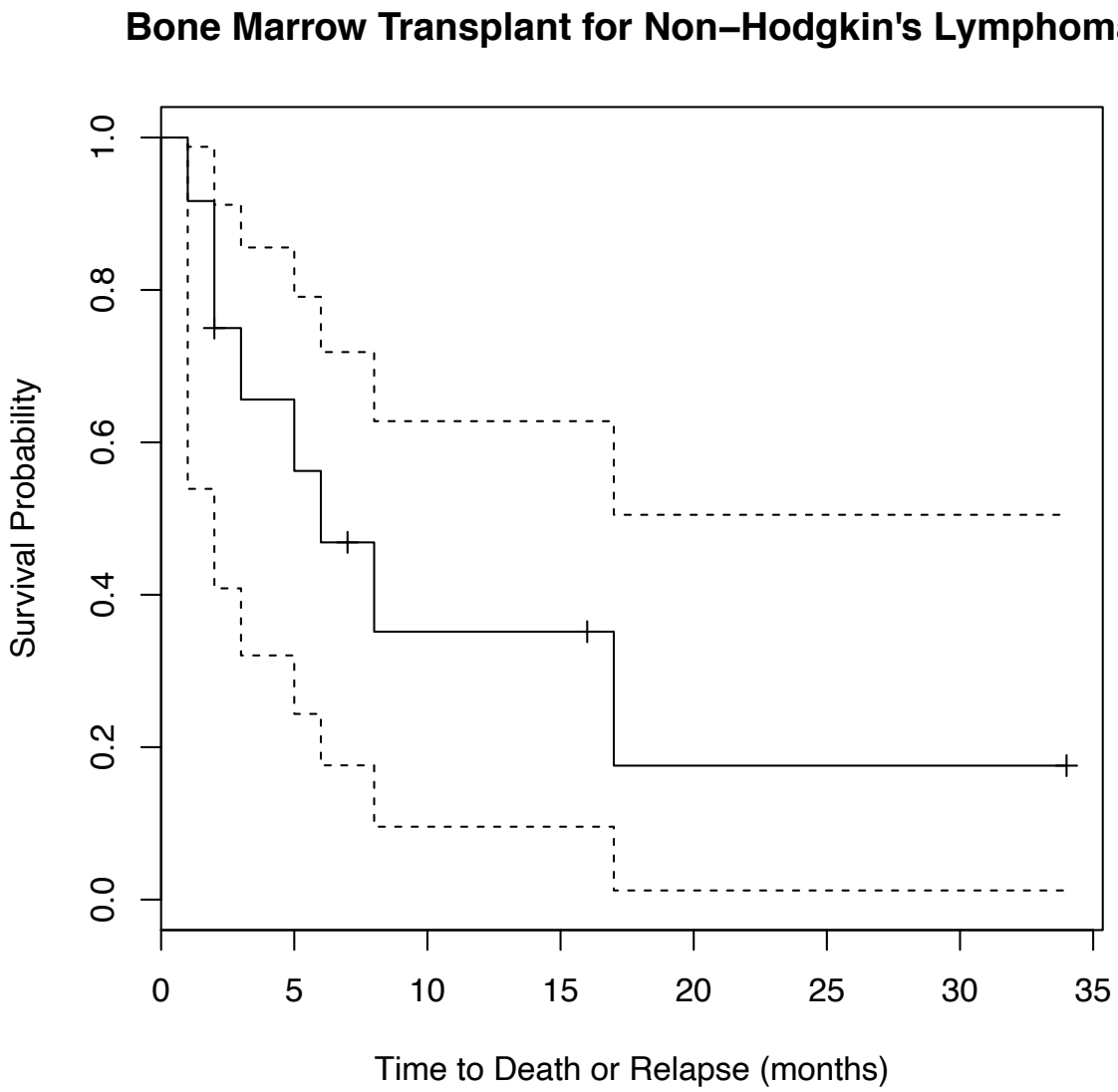
Make a table with a row for every death or censoring time:

τ_j	d_j	c_j	r_j	$1-(d_j/r_j)$	$\hat{S}(\tau_j^+)$
6	3	1	21	$\frac{18}{21}=0.857$	
7	1	0	17		
9	0	1	16		
10					
11					
13					
16					
17					
19					
20					
22					
23					

Note that:

- $\hat{S}(t^+)$ only changes at death (failure) times
- $\hat{S}(t^+)$ is 1 up to the first death time
- $\hat{S}(t^+)$ only goes to 0 if the last event is a death

KM plot for treated leukemia patients



Statistical software

Note: most statistical software packages summarize the KM survival function at τ_j^+, i.e., *just after* the time of the j-th failure.

In other words, they provide $\hat{S}(\tau_j^+)$.

When there is no censoring, the empirical survival estimate would then be:

$$\tilde{S}(t^+) = \frac{\#\ individuals\ with\ T > t}{total\ sample\ size}$$

Greenwood's formula (cont'd)

Since $\hat{S}(t) = \exp[\log[\hat{S}(t)]]$, (by B),

$$var(\hat{S}(t)) \quad = \quad [\hat{S}(t)]^2 var\left[\log[\hat{S}(t)]\right]$$

Greenwood's Formula:

$$var(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:\tau_j<t} \frac{d_j}{(r_j - d_j)r_j}$$

Back to confidence intervals

For a 95% confidence interval, we could use

$$\hat{S}(t) \pm z_{1-\alpha/2}\, se[\hat{S}(t)]$$

where $se[\hat{S}(t)]$ is calculated using Greenwood's formula.
Problem: This approach can yield values > 1 or < 0.
Better approach: Get a 95% confidence interval for

$$L(t) = \log(-\log(S(t)))$$

Since this quantity is unrestricted, the confidence interval will be in the right range when we transform back.

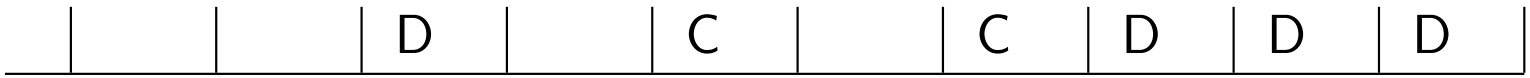
R output for treated leukemia patients

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	12	1	0.917	0.0798	0.5390	0.988
2	11	2	0.750	0.1250	0.4084	0.912
3	8	1	0.656	0.1402	0.3204	0.856
5	7	1	0.562	0.1482	0.2437	0.791
6	6	1	0.469	0.1503	0.1762	0.718
7	5	0	0.469	0.1503	0.1762	0.718
8	4	1	0.352	0.1517	0.0956	0.628
16	3	0	0.352	0.1517	0.0956	0.628
17	2	1	0.176	0.1456	0.0120	0.505
34	1	0	0.176	0.1456	0.0120	0.505

Estimating the cumulative hazard: The Nelson-Aalen estimator

Suppose we want to estimate $\Lambda(t) = \int_0^t \lambda(u)du$, the cumulative hazard at time t.

Just as we did for the KM, think of dividing the observed timespan of the study into a series of fine intervals so that there is only one event per interval:



The cumulative hazard $\Lambda(t)$ can then be approximated by a sum:

$$\hat{\Lambda}(t) = \sum_j \lambda_j \Delta$$

where the sum is over intervals, λ_j is the value of the hazard in the j-th interval and Δ is the width of each interval. Since $\hat{\lambda}\Delta$ is approximately the probability of dying in the interval, we can further approximate by

$$\hat{\Lambda}(t) = \sum_j d_j / r_j$$

It follows that $\Lambda(t)$ will change only at death times, and hence we write the Nelson-Aalen estimator as:

$$\hat{\Lambda}_{NA}(t) = \sum_{j:\tau_j < t} d_j / r_j$$

The Fleming-Harrington (FH) estimator

			D		C		C	D	D	D
r_j	n	n	n	n-1	n-1	n-2	n-2	n-3	n-4	
d_j	0	0	1	0	0	0	0	1	1	
c_j	0	0	0	0	1	0	1	0	0	
$\hat{\lambda}(t_j)$	0	0	1/n	0	0	0	0	$\frac{1}{n-3}$	$\frac{1}{n-4}$	
$\hat{\Lambda}(t_j)$	0	0	1/n	1/n	1/n	1/n	1/n			

Once we have $\hat{\Lambda}_{NA}(t)$, we can also find another estimator of $S(t)$ (Fleming-Harrington):

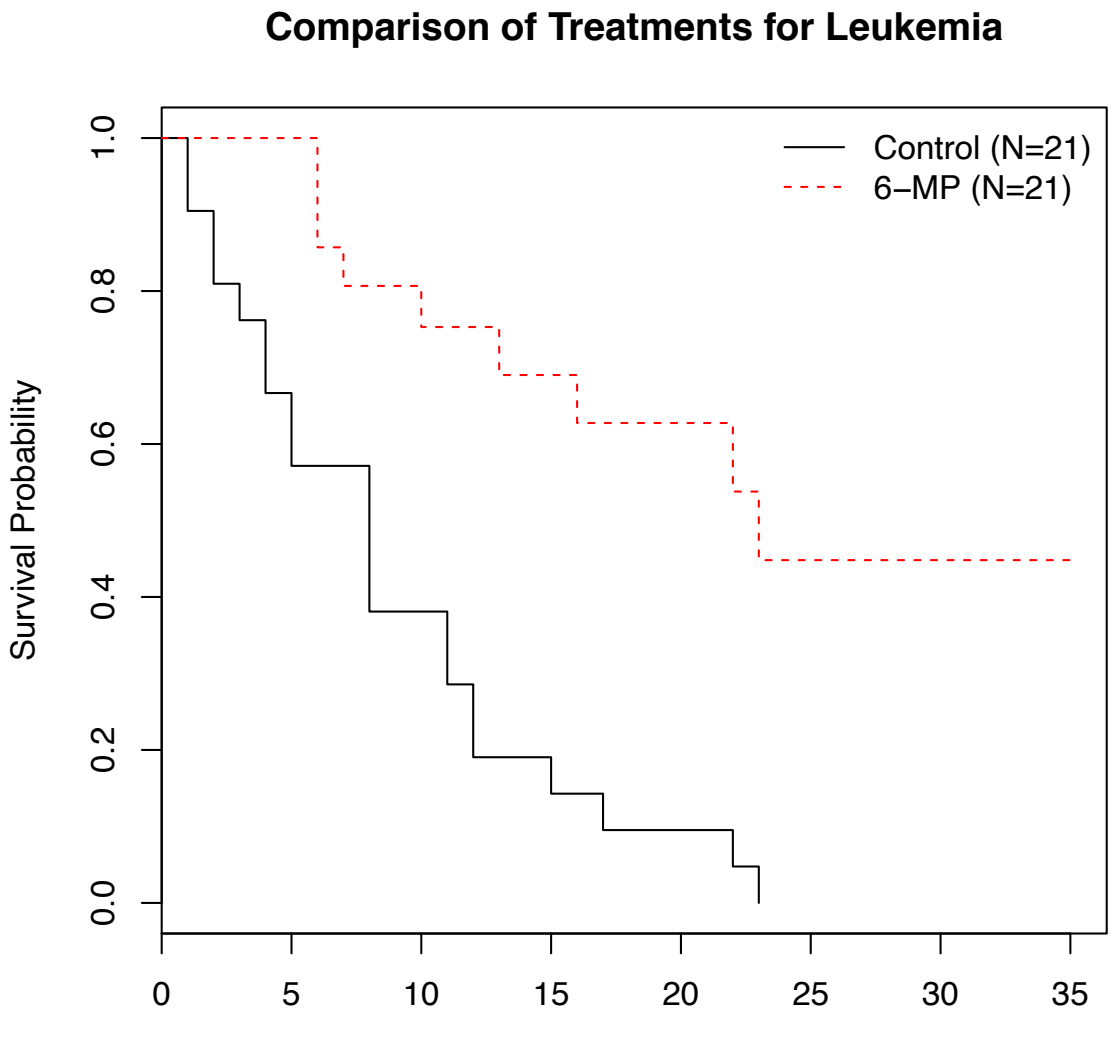
$$\hat{S}_{FH}(t) = \exp(-\hat{\Lambda}_{NA}(t))$$

In general, this estimator of the survival function will be close to the Kaplan-Meier estimator, $\hat{S}_{KM}(t)$ We can also go the other way ... we can take the Kaplan-Meier estimate of $S(t)$, and use it to calculate an alternative estimate of the cumulative hazard function:

$$\hat{\Lambda}_{KM}(t) \quad = \quad -\log \hat{S}_{KM}(t)$$

Comparison of Survival Curves

Now we want to compare the survival estimates between two groups. Example: Time to remission of leukemia patients



How can we form a basis for comparison?

At a specific point in time, we could see whether the confidence intervals for the survival curves overlap.

However, the confidence intervals we have been calculating are **"pointwise"** $\Rightarrow$ they correspond to a confidence interval for $\hat{S}(t^*)$ at a single point in time, t^*.

In other words, we can't say that the true survival function $S(t)$ is contained between the pointwise confidence intervals with 95% probability.

(**Aside:** if you're interested, the issue of confidence **bands** for the estimated survival function are discussed in Section 4.4 of Klein and Moeschberger)

Looking at whether the confidence intervals for $\hat{S}(t^*)$ overlap between the 6MP and placebo groups would only focus on comparing the two treatment groups at a single point in time, t^*.

Should we base our overall comparison of $\hat{S}(t)$ on:

- the furthest distance between the two curves?
- the median survival for each group?
- the average hazard? (for exponential distributions, this would be like comparing the mean event times)
- adding differences between the two survival estimates over time?

$$\sum_j \left[\hat{S}(t_{jA}) - \hat{S}(t_{jB})\right]$$

- a weighted sum of differences, where the weights reflect the number at risk at each time?
- a rank-based test? i.e., we could rank all of the event times, and then see whether the sum of ranks for one group was less than the other.

Nonparametric comparisons of groups

All of these are pretty reasonable options, and we'll see that there have been several proposals for how to compare the survival of two groups. For the moment, we are sticking to nonparametric comparisons.

Why nonparametric?

- **fairly robust**
- **efficient relative to parametric tests**
- **often simple and intuitive**

Before continuing the description of the two-sample comparison, I'm going to try to put this in a general framework to give a perspective of where we're heading in this class.

General Framework for Survival Analysis

We observe $(X_i, \delta_i, \mathbf{Z}_i)$ for individual i, where

- X_i is a censored failure time random variable
- δ_i is the failure/censoring indicator
- $\mathbf{Z}_i$ represents a set of covariates

Note that $\mathbf{Z}_i$ might be a scalar (a single covariate, say treatment or gender) or may be a $(p \times 1)$ vector (representing several different covariates).

Types of covariates

These covariates might be:

- continuous
- discrete
- time-varying (more later)

If $\mathbf{Z}_i$ is a scalar and is binary, then we are comparing the survival of two groups, like in the leukemia example.

More generally though, it is useful to build a model that characterizes the relationship between survival and all of the covariates of interest.

Comparing two survival distributions

We'll proceed as follows:

- Two group comparisons
- Multigroup and stratified comparisons - stratified logrank
- Failure time regression models
  - Cox proportional hazards model
  - Accelerated failure time model

Two sample tests

- Mantel-Haenszel logrank test
- Peto & Peto's version of the logrank test
- Gehan's Generalized Wilcoxon
- Peto & Peto's and Prentice's generalized Wilcoxon
- Tarone-Ware and Fleming-Harrington classes
- Cox's F-test (non-parametric version)

The Mantel-Haenszel Logrank test

The logrank test is the most well known and widely used.

It also has an intuitive appeal, building on standard methods for binary data. (Later we will see that it can also be obtained as the score test from a partial likelihood from the Cox Proportional Hazards model.)

First consider the following (2×2) table classifying those with and without the event of interest in a two group setting:

	Event		
Group	Yes	No	Total
0	d_0	$n_0 - d_0$	n_0
1	d_1	$n_1 - d_1$	n_1
Total	d	$n - d$	n

The M-H test (con'd)

If the margins of this table are considered fixed, then d_0 follows a *hypergeometric* distribution. Under the null hypothesis of no association between the event and group, it follows that

$$E(d_0) = \frac{n_0 d}{n}$$

$$Var(d_0) = \frac{n_0\, n_1\, d(n-d)}{n^2(n-1)}$$

Therefore, under H_0:

$$\chi^2_{MH} = \frac{[d_0 - n_0\, d/n]^2}{\frac{n_0\, n_1\, d(n-d)}{n^2(n-1)}} \sim \chi^2_1$$

This is the Mantel-Haenszel statistic and is approximately equivalent to the Pearson χ^2 test for equality of the two groups given by:

$$\chi^2_p = \sum \frac{(o-e)^2}{e}$$

Note: recall that the Pearson χ^2 test was derived for the case where only the row margins were fixed, and thus the variance above was replaced by:

$$Var(d_0) = \frac{n_0\, n_1\, d(n-d)}{n^3}$$

Toxicity in a clinical trial with two treatments

	Toxicity		
Group	Yes	No	Total
0	8	42	50
1	2	48	50
Total	10	90	100

$$\chi^2_p = 4.00 \quad (p = 0.046)$$

$$\chi^2_{MH} = 3.96 \quad (p = 0.047)$$

Introduction

Now we will explore the relationship between survival and explanatory variables via modelling. In this class, we consider two broad classes of regression models:

Proportional Hazards (PH) models

$$\lambda(t; \mathbf{Z}) = \lambda_{\mathbf{0}}(\mathbf{t})\boldsymbol{\Psi}(\mathbf{Z})$$

Most commonly, we write the second term as: $\Psi(\mathbf{Z}) = \mathbf{e}^{\beta \mathbf{Z}}$

Suppose $Z = 1$ for treated subjects and $Z = 0$ for untreated subjects. Then this model says that the hazard is increased by a factor of e^{β} for treated subjects versus untreated subjects (c^{β} might be < 1).

This is an example of a **semi-parametric** model.

Accelerated Failure Time (AFT) models

These types of models are as follows:

$$\log(T) = \mu + \beta \mathbf{Z} + \sigma \mathbf{w}$$

where w is an "error distribution". Typically, we place a **parametric** assumption on w:

- exponential, Weibull, Gamma
- lognormal

Covariates

In general, $\mathbf{Z}$ is a *vector* of covariates of interest.
$\mathbf{Z}$ may include:

- continuous factors (eg, age, blood pressure)
- discrete factors (gender, marital status)
- possible interactions (age by sex interaction)

Just as in standard linear regression, if we have a discrete covariate A with a levels, then we will need to include $(a-1)$ dummy variables $(U_1, U_2, \ldots, U_a)$ such that $U_j = 1$ if $A = j$. Then

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_2 U_2 + \beta_3 U_3 + \cdots + \beta_a U_a)$$

(In the above model, the subgroup with $A = 1$ or $U_1 = 1$ is the reference group.)

Interactions

Two factors, A and B, interact if the hazard of death depends on the combination of levels of A and B.

We follow the principle of hierarchical models, and only include interactions if all of the associated main effects are also included.

The example just given was based on a proportional hazards model, but the description of the types of covariates we might want to include in our model applies to both the AFT and PH model.

Introduction

We'll start out by focusing on the Cox PH model, and address some of the following questions:

- What does the term $\lambda_0(t)$ mean?
- What's "proportional" about the PH model?
- How do we estimate the parameters in the model?
- How do we interpret the estimated values?
- How can we construct tests of whether the covariates have a significant effect on the distribution of survival times?
- How do these tests compare to the logrank test or the Wilcoxon test?

The Cox Proportional Hazards model

$$\lambda(t; \mathbf{Z}) = \lambda_{\mathbf{0}}(\mathbf{t}) \exp(\beta \mathbf{Z})$$

This is the most common model used for survival data. Why?

- flexible choice of covariates
- fairly easy to fit
- standard software exists

Why do we call it proportional hazards?

Think of the first example, where $Z = 1$ for treated and $Z = 0$ for control. Then if we think of $\lambda_1(t)$ as the hazard rate for the treated group, and $\lambda_0(t)$ as the hazard for control, then we can write:

$$\begin{aligned} \lambda_1(t) &= \lambda(t; Z = 1) = \lambda_0(t) \exp(\beta Z) \\ &= \lambda_0(t) \exp(\beta) \end{aligned}$$

This implies that the ratio of the two hazards is a constant, ϕ, which does NOT depend on time, t. In other words, the hazards of the two groups remain proportional over time.

$$\phi = \frac{\lambda_1(t)}{\lambda_0(t)} = e^{\beta}$$

ϕ is referred to as the **hazard ratio**. **What is the interpretation of β here?**

The Baseline Hazard Function

In the example of comparing two treatment groups, $\lambda_0(t)$ is the hazard rate for the control group.

In general, $\lambda_0(t)$ is called the **baseline hazard function**, and reflects the underlying hazard for subjects with all covariates $Z_1, ..., Z_p$ equal to 0 (i.e., the "reference group").

The general form is:

$$\lambda(t; \mathbf{Z}) = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_p Z_p)$$

So when we substitute all of the Z_j's equal to 0, we get:

$$\begin{aligned} \lambda(t, \mathbf{Z} = \mathbf{0}) &= \lambda_0(t) \exp(\beta_1 * 0 + \beta_2 * 0 + \cdots + \beta_p * 0) \\ &= \lambda_0(t) \end{aligned}$$

The baseline hazard function (cont'd)

In the general case, we think of the i-th individual having a set of covariates $\mathbf{Z_i} = (\mathbf{Z_{1i}}, \mathbf{Z_{2i}}, ..., \mathbf{Z_{pi}})$, and we model their hazard rate as some multiple of the baseline hazard rate:

$$\lambda_i(t, \mathbf{Z_i}) \quad = \quad \lambda_0(t) \exp(\beta_1 Z_{1i} + \cdots + \beta_p Z_{pi})$$

This means we can write the log of the hazard ratio for the i-th individual to the reference group as:

$$\log\left(\frac{\lambda_i(t)}{\lambda_0(t)}\right) \quad = \quad \beta_1 Z_{1i} + \beta_2 Z_{2i} + \cdots + \beta_p Z_{pi}$$

The Cox Proportional Hazards model is a linear model for the log of the hazard ratio

Advantages of the Cox PH model

One of the biggest advantages of the framework of the Cox PH model is that we can estimate the parameters β which reflect the effects of treatment and other covariates without having to make any assumptions about the form of $\lambda_0(t)$.

In other words, we don't have to assume that $\lambda_0(t)$ follows an exponential model, or a Weibull model, or any other particular parametric model.

That's what makes the model *semi-parametric*.

Likelihood Estimation for the PH Model

Kalbfleisch and Prentice derive a likelihood involving only β and $\mathbf{Z}$ (not $\lambda_0(t)$) based on the marginal distribution of the ranks of the observed failure times (in the absence of censoring).

Cox (1972) derived the same likelihood, and generalized it for censoring, using the idea of a **partial likelihood**.

Suppose we observe $(X_i, \delta_i, \mathbf{Z}_i)$ for individual i, where

- X_i is a censored failure time random variable
- δ_i is the failure/censoring indicator (1=fail,0=censor)
- $\mathbf{Z}_i$ represents a set of covariates

The covariates may be continuous, discrete, or time-varying.

Suppose there are K distinct failure (or death) times, and let $\tau_1, ...\tau_K$ represent the K ordered, distinct death times.

For now, assume there are no tied death times.

Let $\mathcal{R}(t) = \{i : x_i \geq t\}$ denote the set of individuals who are "at risk" for failure at time t.

More about risk sets:

- I will refer to $\mathcal{R}(\tau_j)$ as the risk set at the jth failure time
- I will refer to $\mathcal{R}(X_i)$ as the risk set at the failure time of individual i
- There will still be r_j individuals in $\mathcal{R}(\tau_j)$.
- r_j is a number, while $\mathcal{R}(\tau_j)$ identifies the actual subjects at risk

What is the partial likelihood?

Intuitively, it is a product over the set of observed death times of the conditional probabilities of seeing the observed deaths, given the set of individuals at risk at those times.

At each death time τ_j, the contribution to the likelihood is:

$$\begin{aligned} L_j(\beta) &= Pr(\text{individual j fails}|\text{1 failure from } \mathcal{R}(\tau_j)) \\ &= \frac{Pr(\text{individual } j \text{ fails}|\text{ at risk at } \tau_j)}{\sum_{\ell \in \mathcal{R}(\tau_j)} Pr(\text{individual } \ell \text{ fails}|\text{ at risk at } \tau_j)} \\ &= \frac{\lambda(\tau_j; \mathbf{Z}_j)}{\sum_{\ell \in \mathcal{R}(\tau_j)} \lambda(\tau_j; \mathbf{Z}_\ell)} \end{aligned}$$

Under the PH assumption, $\lambda(t; \mathbf{Z}) = \lambda_0(t)e^{\beta \mathbf{Z}}$, so we get:

$$L^{partial}(\beta) = \prod_{j=1}^{K} \frac{\lambda_0(\tau_j)e^{\beta \mathbf{Z}_j}}{\sum_{\ell \in \mathcal{R}(\tau_j)} \lambda_0(\tau_j)e^{\beta \mathbf{Z}_\ell}}$$

Another derivation

In general, the likelihood contributions for censored data fall into two categories:

- **Individual is censored at X_i:**

$$L_i(\beta) = \mathbf{S(X_i)} = \exp[-\int_{\mathbf{0}}^{\mathbf{X_i}} \lambda_\mathbf{i}(\mathbf{u})\mathbf{du}]$$

- **Individual fails at X_i:**

$$L_i(\beta) = \mathbf{S(X_i)}\lambda_\mathbf{i}(\mathbf{X_i}) = \lambda_\mathbf{i}(\mathbf{X_i}) \exp[-\int_{\mathbf{0}}^{\mathbf{X_i}} \lambda_\mathbf{i}(\mathbf{u})\mathbf{du}]$$

Thus, everyone contributes $S(X_i)$ to the likelihood, and only those who fail contribute $\lambda_i(X_i)$.

This means we get a total likelihood of:

$$L(\beta) \quad = \quad \prod_{i=1}^{n} \lambda_i(X_i)^{\delta_i} \ \exp[-\int_0^{X_i} \lambda_i(u)du]$$

The above likelihood holds for all censored survival data, with general hazard function $\lambda(t)$. In other words, we haven't used the Cox PH assumption at all yet.

Now, let's multiply and divide by the term $\left[\sum_{j\in\mathcal{R}(X_i)} \lambda_i(X_i)\right]^{\delta_i}$:

$$L(\beta) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \left[\frac{\lambda_{\mathbf{i}}(\mathbf{X_i})}{\sum_{\mathbf{j}\in\mathcal{R}(\mathbf{X_i})} \lambda_{\mathbf{i}}(\mathbf{X_i})}\right]^{\delta_{\mathbf{i}}} \left[\sum_{\mathbf{j}\in\mathcal{R}(\mathbf{X_i})} \lambda_{\mathbf{i}}(\mathbf{X_i})\right]^{\delta_{\mathbf{i}}} \exp[-\int_{\mathbf{0}}^{\mathbf{X_i}} \lambda_{\mathbf{i}}(\mathbf{u})\mathbf{du}]$$

Cox (1972) argued that the first term in this product contained almost all of the information about β, while the second two terms contained the information about $\lambda_0(t)$, i.e., the baseline hazard.

If we just focus on the first term, then under the Cox PH assumption:

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{n} \left[\frac{\lambda_i(X_i)}{\sum_{j \in \mathcal{R}(X_i)} \lambda_i(X_i)} \right]^{\delta_i} \\
&= \prod_{i=1}^{n} \left[\frac{\lambda_0(X_i) \exp(\beta \mathbf{z_i})}{\sum_{j \in \mathcal{R}(X_i)} \lambda_0(X_i) \exp(\beta \mathbf{z_j})} \right]^{\delta_i} \\
&= \prod_{i=1}^{n} \left[\frac{\exp(\beta \mathbf{z_i})}{\sum_{j \in \mathcal{R}(X_i)} \exp(\beta \mathbf{z_j})} \right]^{\delta_i}
\end{aligned}
$$

This is the partial likelihood defined by Cox. Note that it does not depend on the underlying hazard function $\lambda_0(\cdot)$. Cox recommends treating this as an ordinary likelihood for making inferences about β in the presence of the nuisance parameter $\lambda_0(\cdot)$.

A simple example

Consider the following small data set:

individual	X_i	δ_i	Z_i
1	9	1	4
2	8	0	5
3	6	1	7
4	10	1	3

Now let's compile the pieces that go into the partial likelihood contributions at each failure time:

j	ordered failure time X_i	$\mathcal{R}(X_i)$	i_j	Likelihood contribution $\left[e^{\beta Z_i} / \sum_{j \in \mathcal{R}(X_i)} e^{\beta Z_j}\right]^{\delta_i}$
1	6	{1,2,3,4}	3	$e^{7\beta}/[e^{4\beta} + e^{5\beta} + e^{7\beta} + e^{3\beta}]$
2	8	{1,2,4}	2	1
3	9	{1,4}	1	$e^{4\beta}/[e^{4\beta} + e^{3\beta}]$
4	10	{4}	4	$e^{3\beta}/e^{3\beta} = 1$

Notes on the partial likelihood

$$
\begin{aligned}
L(\beta) &= \prod_{j=1}^{n} \left[\frac{e^{\beta \mathbf{z_j}}}{\sum_{\ell \in \mathcal{R}(X_j)} e^{\beta \mathbf{z}_\ell}} \right]^{\delta_j} \\
&= \prod_{j=1}^{K} \frac{e^{\beta \mathbf{Z_j}}}{\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta \mathbf{Z}_\ell}}
\end{aligned}
$$

where the product is over the K death (or failure) times.

- contributions only at the death times
- the partial likelihood is NOT a product of independent terms, but of conditional probabilities
- There are other choices besides $\Psi(\mathbf{z}) = \mathbf{e}^{\beta \mathbf{z}}$, but this is the most common and the one for which software is generally available.

Partial Likelihood inference

Inference can be conducted by treating the partial likelihood as though it satisfied all the regular likelihood properties (take the more advanced failure time course to see why!!) The **log-partial likelihood** is

$$\begin{aligned}
\ell(\beta) &= \log\left[\prod_{j=1}^{n}\frac{e^{\beta \mathbf{z_j}}}{\sum_{\ell\in\mathcal{R}(X_j)} e^{\beta \mathbf{z}_\ell}}\right]^{\delta_j} \\
&= \log\left[\prod_{j=1}^{K}\frac{e^{\beta \mathbf{z_j}}}{\sum_{\ell\in\mathcal{R}(\tau_j)} e^{\beta \mathbf{z}_\ell}}\right] \\
&= \sum_{j=1}^{K}\left[\beta \mathbf{z_j} - \log[\sum_{\ell\in\mathcal{R}(\tau_\mathbf{j})} \mathbf{e}^{\beta \mathbf{z}_\ell}]\right] = \sum_{j=1}^{K} l_j(\beta)
\end{aligned}$$

where l_j is the log-partial likelihood contribution at the j-th ordered death time.

Partial Likelihood inference (cont'd)

Suppose there is only one covariate (β is one-dimensional).

The **partial likelihood score equations** are:

$$U(\beta) = \frac{\partial}{\partial \beta}\ell(\beta) = \sum_{j=1}^{n} \delta_j \left[Z_j - \frac{\sum_{\ell \in \mathcal{R}(X_j)} Z_\ell e^{\beta Z_\ell}}{\sum_{\ell \in \mathcal{R}(X_j)} e^{\beta Z_\ell}} \right]$$

We can express $U(\beta)$ intuitively as a sum of "observed" minus "expected" values:

$$U(\beta) = \frac{\partial}{\partial \beta}\ell(\beta) = \sum_{j=1}^{n} \delta_j (Z_j - \bar{Z}_j)$$

where $\bar{Z}_j$ is the "weighted average" of the covariate Z over all the individuals in the risk set at time τ_j. Note that β is involved through the term $\bar{Z}_j$.

The maximum partial likelihood estimators can be found by solving $U(\beta) = 0$.

Inference from the partial likelihood (cont'd)

Like standard likelihood theory, it can be shown (not easily) that

$$\frac{(\widehat{\beta} - \beta)}{se(\hat{\beta})} \sim N(0, 1)$$

The variance of $\hat{\beta}$ can be obtained by inverting the second derivative of the partial likelihood,

$$var(\hat{\beta}) \sim \left[-\frac{\partial^2}{\partial \beta^2} \ell(\beta) \right]^{-1}$$

From the above expression for $U(\beta)$, we have:

$$\frac{\partial^2}{\partial \beta^2} \ell(\beta) = \sum_{j=1}^{n} \delta_j \left[-\frac{\sum_{\ell \in \mathcal{R}(\tau_j)} (Z_j - \bar{Z}_j)^2 e^{\beta Z_\ell}}{\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta Z_\ell}} \right]$$

Note: The true variance of $\hat{\beta}$ is a function of the unknown β. We calculate the "observed" information by substituting the partial likelihood estimate of β into the above variance formula.

Simple Example for 2-group comparison: (no ties)

Group 0: $4^+, 7, 8^+, 9, 10^+ \implies Z_i = 0$
Group 1: $3, 5, 5^+, 6, 8^+ \implies Z_i = 1$

	ordered failure	Number at risk		Likelihood contribution
j	time X_i	Group 0	Group 1	$\left[e^{\beta Z_i} / \sum_{j \in \mathcal{R}(X_i)} e^{\beta Z_j}\right]^{\delta_i}$
1	3	5	5	$e^{\beta}/[5 + 5e^{\beta}]$
2	5	4	4	$e^{\beta}/[4 + 4e^{\beta}]$
3	6	4	2	$e^{\beta}/[4 + 2e^{\beta}]$
4	7	4	1	$e^{0}/[4 + 1e^{\beta}] = 1/[4 + e^{\beta}]$
5	9	2	0	$e^{0}/[2 + 0] = 1/2$

Again, we take the product over the likelihood contributions, then maximize to get the partial MLE for β.

What does β represent in this case?

Notes

- The "observed" information matrix is generally used because in practice, people find it has better properties. Also, the "expected" is very hard to calculate.
- There is a nice analogy with the score and information matrices from more standard regression problems, except that here we are summing over observed death times, rather than individuals.
- Newton Raphson is used by many of the computer packages to solve the partial likelihood equations.

Fitting Cox PH model with R

R uses the "coxph" command.

```
coxph(formula, data=, weights, subset,
      na.action, init, control,
      ties=c("efron","breslow","exact"),
      singular.ok=TRUE, robust=FALSE,
      model=FALSE, x=FALSE, y=TRUE, tt, method, ...)
```

Example Leukemia Data

```
Call:
coxph(formula = Surv(weeks, remiss) ~ trt, data = leukemia, ties = "breslow")

  n= 42, number of events= 30

       coef exp(coef) se(coef)      z Pr(>|z|)
trt -1.5092    0.2211   0.4096 -3.685 0.000229 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    exp(coef) exp(-coef) lower .95 upper .95
trt    0.2211      4.523   0.09907    0.4934

Concordance= 0.69  (se = 0.053 )
Rsquare= 0.304   (max possible= 0.989 )
Likelihood ratio test= 15.21  on 1 df,   p=9.615e-05
Wald test            = 13.58  on 1 df,   p=0.0002288
Score (logrank) test = 15.93  on 1 df,   p=6.571e-05
```

More Notes:

Here are some comments:

- The Cox Proportional hazards model has the advantage over a simple logrank test of giving us an estimate of the "risk ratio" (i.e., $\phi = \lambda_1(t)/\lambda_0(t)$). This is more informative than just a test statistic, and we can also form confidence intervals for the risk ratio.
- In this case, $\hat{\phi} = 0.221$, which can be interpreted to mean that the hazard for relapse among patients treated with 6-MP is less than 25% of that for placebo patients.

Adjustments for ties

The proportional hazards model assumes a continuous hazard – ties are not possible. There are four proposed modifications to the likelihood to adjust for ties.

(1) **Cox's (1972) modification:** "discrete" method

(2) **Peto-Breslow method**

(3) **Efron's (1977) method**

(4) **Exact method (Kalbfleisch and Prentice)**

(5) **Exact marginal method**

Some notation

$\tau_1, ...\tau_K$	the K ordered, distinct death times
d_j	the number of failures at τ_j
H_j	the "history" of the entire data set, up to the j-th death or failure time, including the time of the failure, but not the identities of the d_j who fail there.
$i_{j1}, ...i_{jd_j}$	the identities of the d_j individuals who fail at τ_j

Cox's (1972) modification: The "discrete" method

Cox's method assumes that if there are tied failure times, they truly happened at the same time. It is based on a discrete likelihood.

The **partial likelihood** is:

$$
\begin{aligned}
L(\beta) &= \prod_{j=1}^{K} Pr(i_{j1}, \ldots i_{jd_j} \text{ fail} \mid d_j \text{ fail at } \tau_j, \text{ from } \mathcal{R}) \\
&= \prod_{j=1}^{K} \frac{Pr(i_{j1}, \ldots i_{jd_j} \text{ fail} \mid \text{ in } \mathcal{R}(\tau_j))}{\sum_{\ell \in s(j,d_j)} Pr(\ell_1, \ldots . \ell_{d_j} \text{ fail} \mid \text{ in } \mathcal{R}(\tau_j))} \\
&= \prod_{j=1}^{K} \frac{\exp(\beta \mathbf{z_{i_{j1}}}) \cdots \exp(\beta \mathbf{z_{i_{jd_j}}})}{\sum_{\ell \in s(j,d_j)} \exp(\beta \mathbf{z_{\ell_1}}) \cdots \exp(\beta \mathbf{z_{\ell_{d_j}}})} \\
&= \prod_{j=1}^{K} \frac{\exp(\beta \mathbf{S_j})}{\sum_{\ell \in s(j,d_j)} \exp(\beta \mathbf{S_{j\ell}})}
\end{aligned}
$$

In the previous formula

- $s(j, d_j)$ is the set of all possible sets of d_j individuals that can possibly be drawn from the risk set at time τ_j
- S_j is the sum of the Z's for all the d_j individuals who fail at τ_j
- $S_{j\ell}$ is the sum of the Z's for all the d_j individuals in the ℓ-th set drawn out of $s(j, d_j)$

Simple Example (with ties)

What does this all mean??!!

Let's modify our previous simple example to include ties.

Group 0: $4^+, 6, 8^+, 9, 10^+ \implies Z_i = 0$

Group 1: $3, 5, 5^+, 6, 8^+ \implies Z_i = 1$

j	Ordered failure time X_i	Number at risk Group 0	Number at risk Group 1	Lik. Contribution $e^{\beta S_j} / \sum_{\ell \in s(j,d_j)} e^{\beta S_{j\ell}}$
1	3	5	5	$e^{\beta}/[5 + 5e^{\beta}]$
2	5	4	4	$e^{\beta}/[4 + 4e^{\beta}]$
3	6	4	2	$e^{\beta}/[6 + 8e^{\beta} + e^{2\beta}]$
4	9	2	0	$e^{0}/2 = 1/2$

Comments

The tie occurs at $t = 6$, when $\mathcal{R}(\tau_j) = \{Z = 0 : (6, 8^+, 9, 10^+), Z = 1 : (6, 8^+)\}$. Of the $\binom{6}{2} = 15$ possible pairs of subjects at risk at t=6, there are 6 pairs formed where both are from group 0 ($S_j = 0$), 8 pairs formed with one in each group ($S_j = 1$), and 1 pairs formed with both in group 1 ($S_j = 2$).

Problem: With numbers of ties, the denominator can have many many terms and be difficult to calculate.

The Breslow method: (default)

Breslow and Peto suggested replacing the term $\sum_{\ell \in s(j,d_j)} e^{\beta S_{j\ell}}$ in the denominator by the term $\left(\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta Z_\ell}\right)^{d_j}$, so that the following modified partial likelihood would be used:

$$L(\beta) = \prod_{j=1}^{K} \frac{e^{\beta S_j}}{\sum_{\ell \in s(j,d_j)} e^{\beta S_{j\ell}}} \approx \prod_{j=1}^{K} \frac{e^{\beta S_j}}{\left(\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta Z_\ell}\right)^{d_j}}$$

Justification

Suppose individuals 1 and 2 fail from $\{1, 2, 3, 4\}$ at time τ_j. Let $\phi(i)$ be the hazard ratio for individual i (compared to baseline).

$$\begin{aligned}
\frac{e^{\beta S_j}}{\sum_{\ell \in s(j,d_j)} e^{\beta S_{j\ell}}} &= \frac{\phi(1)}{\phi(1)+\phi(2)+\phi(3)+\phi(4)} \times \frac{\phi(2)}{\phi(2)+\phi(3)+\phi(4)} \\
&\quad + \frac{\phi(2)}{\phi(1)+\phi(2)+\phi(3)+\phi(4)} \times \frac{\phi(1)}{\phi(1)+\phi(3)+\phi(4)} \\
&\approx \frac{2\phi(1)\phi(2)}{[\phi(1)+\phi(2)+\phi(3)+\phi(4)]^2}
\end{aligned}$$

The Peto (Breslow) approximation will break down when the number of ties are relative to the size of the risk sets, and then tends to yield estimates of β which are biased toward 0.

Efron's (1977) method

Efron suggested an even closer approximation to the discrete likelihood:

$$L(\beta) = \prod_{j=1}^{K} \frac{e^{\beta S_j}}{\left(\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta Z_\ell} + \frac{j-1}{d_j} \sum_{\ell \in \mathcal{D}(\tau_j)} e^{\beta Z_\ell} \right)^{d_j}}$$

Like the Breslow approximation, Efron's method will yield estimates of β which are biased toward 0 when there are many ties.

However, (1995) Allison recommends the Efron approximation since it is much faster than the exact methods and tends to yield much closer estimates than the default Breslow approach.

Exact method (Kalbfleisch and Prentice)

The "discrete" option that we discussed in (1) is an exact method based on a discrete likelihood (assuming that tied events truly ARE tied).

This second exact method is based on the continuous likelihood, under the assumption that if there are tied events, that is due to the imprecise nature of our measurement, and that there must be some *true* ordering.

All possible orderings of the tied events are calculated, and the probabilites of each are summed.

Example

Here is an example with 2 tied events (1,2) from risk set (1,2,3,4):

$$\frac{e^{\beta S_j}}{\sum_{\ell \in s(j,d_j)} e^{\beta S_{j\ell}}} = \frac{e^{\beta S_1}}{e^{\beta S_1} + e^{\beta S_2} + e^{\beta S_3} + e^{\beta S_4}} \times \frac{e^{\beta S_2}}{e^{\beta S_2} + e^{\beta S_3} + e^{\beta S_4}}$$

$$+ \frac{e^{\beta S_2}}{e^{\beta S_1} + e^{\beta S_2} + e^{\beta S_3} + e^{\beta S_4}} \times \frac{e^{\beta S_1}}{e^{\beta S_1} + e^{\beta S_3} + e^{\beta S_4}}$$

Bottom Line

Implications of Ties (See Allison (1995), p.127-137):

(1) **When there are no ties**, all four options give *exactly* the same results.

(2) **When there are only a few ties**, it won't make much difference which method is used. However, since the exact methods won't take much extra computing time, you might as well use one of them.

(3) **When there are many ties** (relative to the number at risk), the Breslow option (default) performs poorly (Farewell & Prentice, 1980; Hsieh, 1995). Both of the approximate methods, Breslow and Efron, yield coefficients that are attenuated (biased toward 0).

Implication of ties (cont'd)

(4) **The choice of which exact method to use** should be based on substantive grounds - are the tied event times truly tied? ...or are they the result of imprecise measurement?

(5) **Computing time of exact methods** is much longer than that of the approximate methods. However, in most cases it will still be less than 30 seconds even for the exact methods.

(6) **Best approximate method** - the Efron approximation nearly always works better than the Breslow method, with no increase in computing time, so use this option if exact methods are too computer-intensive.

R Commands for PH Model with Ties

R offers four options for adjustments with tied data:

- `breslow` (default)
- `efron`
- `exactp`
- `exactm` - an exact marginal likelihood calculation

Fecundability data example

```
Call:
coxph(formula = Surv(cycle, censor) ~ smoker, data = fecund)

  n= 586, number of events= 567

          coef exp(coef) se(coef)      z Pr(>|z|)
smoker -0.3878    0.6786   0.1140 -3.401 0.000671 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

       exp(coef) exp(-coef) lower .95 upper .95
smoker    0.6786      1.474    0.5427    0.8485

Concordance= 0.537  (se = 0.014 )
Rsquare= 0.021   (max possible= 1 )
Likelihood ratio test= 12.57  on 1 df,   p=0.000392
Wald test            = 11.57  on 1 df,   p=0.0006712
Score (logrank) test = 11.71  on 1 df,   p=0.0006218
```

More on the Cox PH model

I. Confidence intervals and hypothesis tests
   - Two methods for confidence intervals
   - Wald tests and likelihood ratio tests
   - Interpretation of parameter estimates
   - An example with real data from an AIDS clinical trial

II. Predicted survival under proportional hazards

III. Predicted medians and P-year survival

Constructing Confidence intervals and tests for the Hazard Ratio

Many software packages provide estimates of β, but the hazard ratio (i.e., $\exp(\beta)$) is usually the parameter of interest.

We can use the delta method to get standard errors for $\exp(\hat{\beta})$:

$$Var(\exp(\hat{\beta})) = \exp(2\hat{\beta})Var(\hat{\beta})$$

Constructing confidence intervals for $\exp(\beta)$

We have two options: (assuming that β is a scalar):

I. Using $se(\exp\hat{\beta})$ obtained above via the delta method as $se(\exp\hat{\beta}) = \sqrt{[Var(\exp(\hat{\beta}))]}$, calculate the endpoints as:

$$[L, U] = [e^{\hat{\beta}} - 1.96\, se(e^{\hat{\beta}}), e^{\hat{\beta}} + 1.96\, se(e^{\hat{\beta}})]$$

II. Form a confidence interval for $\hat{\beta}$, and then exponentiate the endpoints.

$$[L, U] = [e^{\hat{\beta} - 1.96 se(\hat{\beta})}, e^{\hat{\beta} + 1.96 se(\hat{\beta})}]$$

Method II is preferable since $\hat{\beta}$ converges to a normal distribution more quickly than $\exp(\hat{\beta})$.

Hypothesis Tests:

For each covariate of interest, the null hypothesis is

$$H_o : \beta_j = 0$$

A Wald test[2] of the above hypothesis is constructed as:

$$Z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \qquad \text{or} \qquad \chi^2 = \left[\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}\right]^2$$

[2] The first follows a normal distribution, and the second follows a χ^2 with 1 df.

The test for $\beta_j = 0$ assumes that all other terms in the model are fixed. If we have a factor A with a levels, then we would need to construct a χ^2 test with $(a-1)$ df, using a test statistic based on a quadratic form:

$$\chi^2_{(a-1)} = \widehat{\boldsymbol{\beta}}'_A Var(\widehat{\boldsymbol{\beta}}_A)^{-1} \widehat{\boldsymbol{\beta}}_A$$

where $\boldsymbol{\beta}_A = (\beta_2, ..., \beta_a)'$ are the $(a-1)$ coefficients corresponding to $Z_2, ..., Z_a$ (or $Z_1, ..., Z_{a-1}$, depending on the reference group).

Comparing nested models $\Rightarrow$ Likelihood Ratio Tests

Suppose there are $(p+q)$ explanatory variables measured:

$$Z_1, \ldots, Z_p, Z_{p+1}, \ldots, Z_{p+q}$$

and proportional hazards are assumed.

Consider the following models:

- **Model 1:** (contains only the first p covariates)

$$\frac{\lambda_i(t, \mathbf{Z})}{\lambda_0(t)} = \exp(\beta_1 Z_1 + \cdots + \beta_p Z_p)$$

- **Model 2:** (contains all $(p+q)$ covariates)

$$\frac{\lambda_i(t, \mathbf{Z})}{\lambda_0(t)} = \exp(\beta_1 Z_1 + \cdots + \beta_{p+q} Z_{p+q})$$

Constructing the likelihood-ratio test

These are *nested* models. For such nested models, we can construct a **likelihood ratio** test of

$$H_0 : \beta_{p+1} = \cdots = \beta_{p+q} = 0$$

as:

$$\chi^2_{LR} = -2\left[\log(\hat{L}(1)) - \log(\hat{L}(2))\right]$$

Under H_o, this test statistic is approximately distributed as χ^2 with q df.

Some examples using the R `coxph` command: The likelihood ratio test

Model 1:

```
coxph(formula = Surv(mactime, macstat) ~ karnof + rif + clari,
    data = mac)

  n= 1177, number of events= 121

           coef exp(coef) se(coef)      z Pr(>|z|)
karnof -0.04485   0.95614  0.01064 -4.217 2.48e-05 ***
rif     0.87197   2.39161  0.23694  3.680 0.000233 ***
clari   0.27557   1.31728  0.25801  1.068 0.285509
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

       exp(coef) exp(-coef) lower .95 upper .95
karnof    0.9561     1.0459    0.9364    0.9763
rif       2.3916     0.4181    1.5032    3.8051
clari     1.3173     0.7591    0.7944    2.1842

Concordance= 0.649  (se = 0.028 )
Rsquare= 0.027   (max possible= 0.73 )
Likelihood ratio test= 32.02  on 3 df,   p=5.193e-07
Wald test            = 32.29  on 3 df,   p=4.548e-07
Score (logrank) test = 33.16  on 3 df,   p=2.977e-07
```

Likelihood-ratio example (cont'd)

Model 2:

```
coxph(formula = Surv(mactime, macstat) ~ karnof + rif + clari + cd4, data = mac)

  n= 1177, number of events= 121

            coef exp(coef)  se(coef)      z Pr(>|z|)
karnof -0.036874  0.963798  0.010665 -3.457 0.000546 ***
rif     0.879749  2.410294  0.237092  3.711 0.000207 ***
clari   0.252345  1.287041  0.258337  0.977 0.328664
cd4    -0.018360  0.981807  0.003684 -4.984 6.23e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

       exp(coef) exp(-coef) lower .95 upper .95
karnof    0.9638     1.0376    0.9439    0.9842
rif       2.4103     0.4149    1.5145    3.8360
clari     1.2870     0.7770    0.7757    2.1354
cd4       0.9818     1.0185    0.9747    0.9889

Concordance= 0.716  (se = 0.028 )
Rsquare= 0.053   (max possible= 0.73 )
Likelihood ratio test= 63.77  on 4 df,   p=4.682e-13
Wald test            = 55.59  on 4 df,   p=2.449e-11
Score (logrank) test = 56.22  on 4 df,   p=1.806e-11
```

The likelihood ratio test of significance for the CD4 count

The **likelihood ratio** test for the effect of CD4 is

$$\begin{aligned} \chi^2_{LR} &= -2\left[\log(\hat{L}(1)) - \log(\hat{L}(2))\right] \\ &= -2\left[-754.4910 - (-738.6162))\right] = 31.7496 \end{aligned}$$

This is compared to a chi-square statistic with 1 degree of freedom, resulting in a p-value which is virtually zero. The R code and results are as follows:

```
Analysis of Deviance Table
 Cox model: response is  Surv(mactime, macstat)
 Model 1: ~ karnof + rif + clari
 Model 2: ~ karnof + rif + clari + cd4
   loglik Chisq Df P(>|Chi|)
1 -754.49
2 -738.62 31.75  1 1.754e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Estimates of the hazard ratio

The above output produces the estimated hazard ratio along with 95% confidence intervals by forming a CI for the log HR (beta), and then exponentiating the bounds)

We can also compute the hazard ratio ourselves, by exponentiating the coefficients:

$$HR_{cd4} = \exp(-0.01835) = 0.98$$

... or with R,

```
exp(confint(fit.mac1))
             2.5 %    97.5 %
karnof  0.9438599 0.9841573
rif     1.5144570 3.8360410
clari   0.7757030 2.1354482
cd4     0.9747439 0.9889221
```

Why is this HR so close to 1, and yet still significant?

What is the interpretation of this HR?

The interpretation of this hazard ratio is the change in the hazard *for each additional CD4 cell/μl*.

This is the reason that, although the hazard ratio is small, it is still significant (as it is associated with a single-cell difference).

Note here that this is a very strong structural assumption of this model because it assumes a *linear* relationship between the hazard ratio and CD4 cell count, which is unlikely to be correct throughout the CD4 spectrum.

Comparison of treatment effect for MAC

In the mac study, there were three treatment arms (rif, clari, and the rif+clari combination). Because we have only included the `rif` and `clari` effects in the model, the combination therapy is the "reference" group.

We can conduct an overall test of treatment using the `wald.test` command in R (part of the `aod` package):

```
Wald test:
----------


Chi-squared test:
X2 = 17.0, df = 2, P(> X2) = 2e-04
```

for a 2 df Wald chi-square test of whether both treatment coefficients are equal to 0. This `wald.test` command can be used to conduct many different tests based on the Wald test.

Testing the difference between treatment arms

We can also test whether there is a difference between the `rif` and `clari` treatment arms:

```
Call:
coxph(formula = Surv(mactime, macstat) ~ karnof + rif + I(rif +
    clari) + cd4, data = mac)

  n= 1177, number of events= 121

                    coef exp(coef)  se(coef)      z Pr(>|z|)
karnof         -0.036874  0.963798  0.010665 -3.457 0.000546 ***
rif             0.627403  1.872742  0.211970  2.960 0.003078 **
I(rif + clari)  0.252345  1.287041  0.258337  0.977 0.328664
cd4            -0.018360  0.981807  0.003684 -4.984 6.23e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Here the test of the difference between the two arms is $\chi^2 = 2.960$ with p-value $p = 0.0031$.

Predicting survival through the Cox model

The major drawback of the Cox model is that it does not provide estimates for $\lambda_0(t)$ the baseline hazard.

The Cox PH model says that $\lambda_i(t, \mathbf{Z}) = \lambda_0(t) \exp(\beta \mathbf{Z})$. What does this imply about the survival function, $S_z(t)$, for the i-th individual with covariates $\mathbf{Z}_i$?

For the baseline (reference) group, we have:

$$S_0(t) \quad = \quad e^{-\int_0^t \lambda_0(u)du} = e^{-\Lambda_0(t)}$$

This is by the definition of a survival function (see intro notes).

Without estimating $\lambda_0(t)$ we cannot estimate $S(t; \mathbf{Z})$

For the i-th patient with covariates $\mathbf{Z}_i$, we have:

$$\begin{aligned} S_i(t) &= e^{-\int_0^t \lambda_i(u)du} = e^{-\Lambda_i(t)} \\ &= e^{-\int_0^t \lambda_0(u)\exp(\boldsymbol{\beta}\mathbf{Z}_i)du} \\ &= e^{-\exp(\boldsymbol{\beta}\mathbf{Z}_i)\int_0^t \lambda_0(u)du} \\ &= \left[e^{-\int_0^t \lambda_0(u)du}\right]^{\exp(\boldsymbol{\beta}\mathbf{Z}_i)} \\ &= [S_0(t)]^{\exp(\boldsymbol{\beta}\mathbf{Z}_i)} \end{aligned}$$

(This uses the mathematical relationship $[e^b]^a = e^{ab}$)

Thus, if we cannot estimate $\lambda_0(t)$ we cannot estimate $S_0(t)$ and, consequently, we cannot estimate $S(t; \mathbf{Z})$.

Estimating $S_0(t)$ through Kaplan Meier

We could use the KM estimator, but there are a few disadvantages of that approach:

- It would only use the survival times for observations contained in the reference group, and not all the rest of the survival times.
- It would tend to be somewhat choppy, since it would reflect the smaller sample size of the reference group.
- It's possible that there are no subjects in the dataset who are in the "reference" group

  For example, say covariates are `health` and `gender`; there is no one of `health`==0 (patients of perfect health) in our dataset.

Taking advantage of the Cox model itself

Instead, we will use a baseline hazard estimator which takes advantage of the proportional hazards assumption to get a smoother estimate.

$$\hat{S}_i(t) \quad = \quad [\hat{S}_0(t)]^{\exp(\widehat{\boldsymbol{\beta}}\mathbf{Z}_i)}$$

Using the above formula, we substitute $\widehat{\boldsymbol{\beta}}$ based on fitting the Cox PH model, and calculate $\hat{S}_0(t)$ by one of the following approaches:

- Breslow estimator (Stata, R)
- Kalbfleisch/Prentice estimator (SAS)

The Breslow Estimator

The Breslow estimator is as follows:

$$\hat{S}_0(t) = \exp^{-\hat{\Lambda}_0(t)}$$

where $\hat{\Lambda}_0(t)$ is the estimated cumulative baseline hazard:

$$\hat{\Lambda}(t) = \sum_{j:\tau_j<t} \left(\frac{d_j}{\sum_{k \in \mathcal{R}(\tau_j)} \exp(\beta_1 Z_{1k} + \ldots \beta_p Z_{pk})} \right)$$

The Breslow Estimator: further motivation

The Breslow estimator is based on extending the concept of the Nelson-Aalen estimator to the proportional hazards model.

Recall that for a single sample with no covariates, the **Nelson-Aalen Estimator** of the cumulative hazard is:

$$\hat{\Lambda}(t) = \sum_{j:\tau_j<t} \frac{d_j}{r_j}$$

where d_j and r_j are the number of deaths and the number at risk, respectively, at the j-th death time.

When there are covariates and assuming the PH model above, one can generalize this to estimate the cumulative baseline hazard by adjusting the denominator:

$$\hat{\Lambda}(t) = \sum_{j:\tau_j<t} \left(\frac{d_j}{\sum_{k\in\mathcal{R}(\tau_j)} \exp(\beta_1 Z_{1k} + \ldots \beta_p Z_{pk})} \right)$$

The Kalbfleisch/Prentice Estimator

This estimator is as follows:

$$\hat{S}_0(t) = \prod_{j:\tau_j<t} \hat{\alpha}_j$$

where $\hat{\alpha}_j, j = 1, ...d$ are the MLE's obtained by assuming that $S(t; Z)$ satisfies

$$S(t; Z) = [S_0(t)]^{e^{\beta Z}} = \left[\prod_{j:\tau_j<t} \alpha_j\right]^{e^{\beta Z}} = \prod_{j:\tau_j<t} \alpha_j^{e^{\beta Z}}$$

Using R to Predict Survival

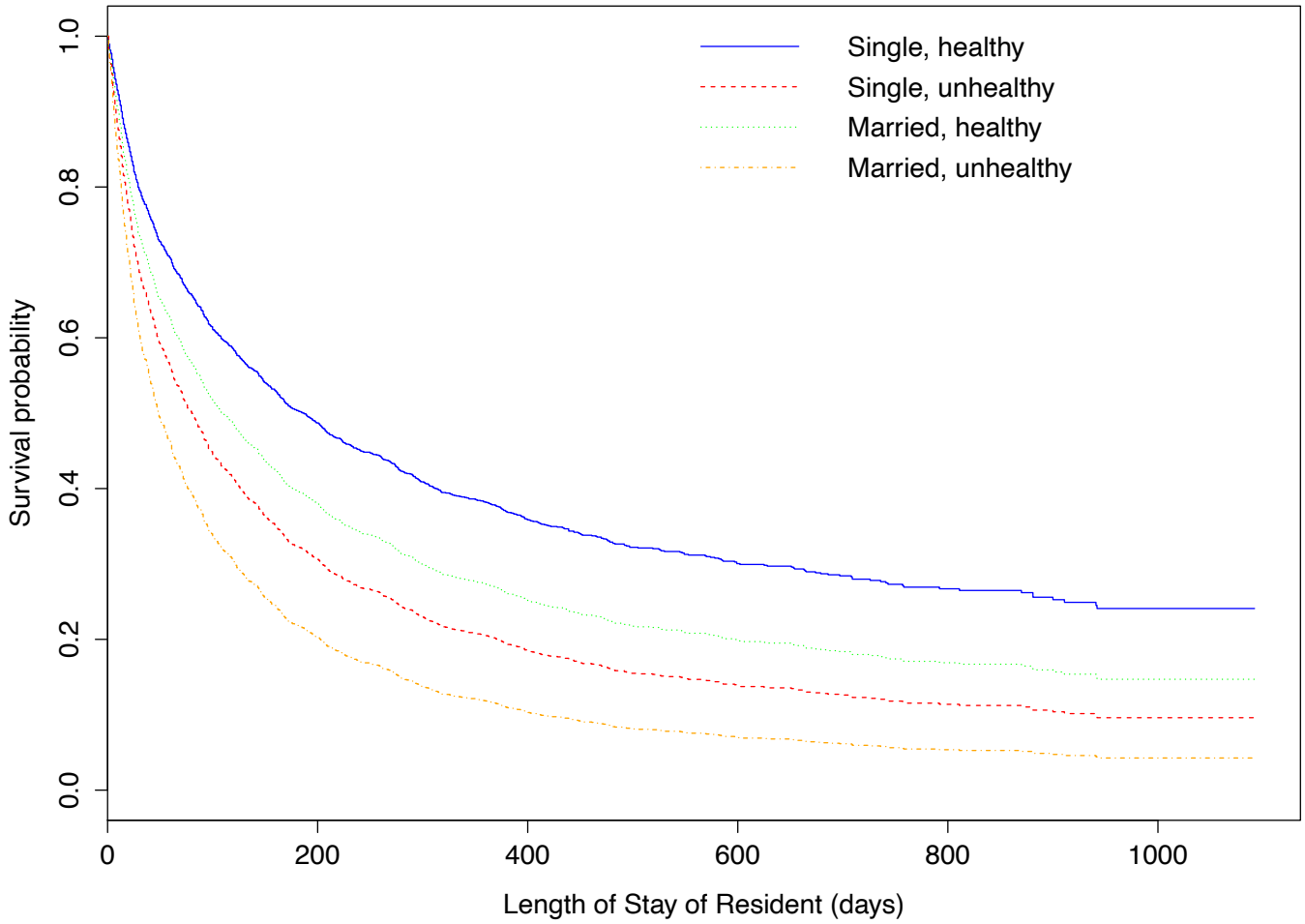
Consider predicted survival for the groups of men and women with the best (`health==2`) and worst health (`health==5`). The R command `survfit` calculates the predicted survival values.

```
              1          2         3          4
  [1,] 0.9896288 0.98298849 0.9860573 0.97715684    1
  [2,] 0.9815838 0.96987163 0.9752767 0.95963670    2
  [3,] 0.9773101 0.96293165 0.9695622 0.95040011    3
  [4,] 0.9692168 0.94984305 0.9587642 0.93304299    4
  [5,] 0.9587077 0.93295246 0.9447896 0.91076637    5
  [6,] 0.9515173 0.92146452 0.9352587 0.89569475    6
  [7,] 0.9428590 0.90770541 0.9238150 0.87772909    7
   .     .          .          .          .          .
   .     .          .          .          .          .
   .     .          .          .          .          .
[592,] 0.2408962 0.09607878 0.1470458 0.04263921 1088
[593,] 0.2408962 0.09607878 0.1470458 0.04263921 1091
[594,] 0.2408962 0.09607878 0.1470458 0.04263921 1092
```

In the output above, 1-4 represent the four groups and the last column the 594 distinct failure times when the survival is estimated.

Pictorial representation of predicted survival

We can get a visual picture of what the proportional hazards assumption implies by looking at these four subgroups.



Predicted Medians

Suppose we want to find the predicted median survival for an individual with a specified combination of covariates (e.g., a single male with health status 0 - note that none such individual exists in the data!).

There are three possible approaches:

(1) Calculate the median from the subset of individuals with the specified covariate combination (using KM approach)

(2) Generate predicted survival curves for each combination of

(3) Generate the predicted survival curve from the estimated baseline hazard. This is done as follows:
We want the estimated median (M) for an individual with covariates $\mathbf{Z}_i$. We know

$$S(M; Z) = [S_0(M)]^{e^{\beta Z_i}} = 0.5$$

Hence, M satisfies (multiplying both sides by $e^{-\beta Z_i}$):

$$S_0(M) = [0.5]^{e^{-\beta Z_i}}$$

covariates, and obtain the medians directly.

Finding the median through the Cox regression model

Another approach would be to use the Cox model itself and find the median for each group. Consider the following output:

```
              1          2         3          4
 [47,] 0.7345588 0.60187880 0.6600326 0.50471265   47
 [48,] 0.7303122 0.59616290 0.6548988 0.49826795   48
   .     .          .          .          .          .
   .     .          .          .          .          .
   .     .          .          .          .          .
 [78,] 0.6583264 0.50256716 0.5694792 0.39588599   78
 [79,] 0.6561186 0.49979634 0.5669086 0.39294923   80
   .     .          .          .          .          .
   .     .          .          .          .          .
   .     .          .          .          .          .
[107,] 0.5977249 0.42871652 0.5000270 0.31960180  109
[108,] 0.5965826 0.42736892 0.4987404 0.31824952  110
   .     .          .          .          .          .
   .     .          .          .          .          .
   .     .          .          .          .          .
[171,] 0.5003829 0.31997626 0.3935705 0.21552407  185
[172,] 0.4991766 0.31870771 0.3922932 0.21437409  187
   .     .          .          .          .          .
   .     .          .          .          .          .
   .     .          .          .          .          .
```

Using the Cox model to define the median

Recall that previously we defined the median as the *smallest* value of t for which $\hat{S}(t) \leq 0.5$, so the medians from above would be 185, 80, 109, and 48 days for single healthy, single unhealthy, married healthy, and married unhealthy, respectively.

The following R output summarizes the previous output as follows:

```
     n events median 0.95LCL 0.95UCL
1 1591   1269    187     155     227
2 1591   1269     80      65      97
3 1591   1269    110      86     149
4 1591   1269     48      39      66
```

We note that, unlike the case of the Kaplan-Meier approach, the entire sample was used to estimate these medians; even nursing home patients who were not part of this analyses (e.g., men with medium health). This was accomplished because of the proportionality of the hazards assumed by the Cox model.

Estimating P-year survival

Suppose we want to find the P-year survival rate for an individual with a specified combination of covariates, $\hat{S}(P; \mathbf{Z}_i)$

For an individual with $\mathbf{Z}_i = 0$, the P-year survival can be obtained from the baseline survivorship function, $\hat{S}_0(P)$

For individuals with $\mathbf{Z}_i \neq 0$, it can be obtained as:

$$\hat{S}(P; \mathbf{Z}_i) = [\hat{S}_0(P)]^{e^{\widehat{\boldsymbol{\beta}}\mathbf{Z}_i}}$$

Notes

The following comments are important:

- Although I say "P-year" survival, the units of time in a particular dataset may be days, weeks, or months. The answer here will be in the same units of time as the original data.
- If $\widehat{\boldsymbol{\beta}}\mathbf{Z}_i$ is positive, then the P-year survival rate for the i-th individual will be lower than for a baseline individual.

  Why is this true?

Estimating P-year survival with R

R has the command `predict`, which has the option `"expected"`, which lists the expected number of events by time t for a given set of covariates. Note that this is the *estimated cumulative hazard* $\hat{\Lambda}(t; \mathbf{Z})$!

Thus, the estimated survival for time t (or, more relevantly for P-year survival) is,

$$\hat{S}(P; \mathbf{Z}) = \exp\left[-\hat{\Lambda}(t; \mathbf{Z})\right]$$

One and two-year survival in the nursing home example

To estimate P-year survival for the four groups in the nursing home example is given as follows:

```
              1         2         3          4
  [1,] 0.9896288 0.98298849 0.9860573 0.97715684    1
  [2,] 0.9815838 0.96987163 0.9752767 0.95963670    2
     .         .         .         .          .     .
     .         .         .         .          .     .
     .         .         .         .          .     .
[286,] 0.3796947 0.20316145 0.2713845 0.11689747  364
[287,] 0.3790374 0.20258294 0.2707520 0.11644939  365
[288,] 0.3783757 0.20200128 0.2701156 0.11599931  366
     .         .         .         .          .     .
     .         .         .         .          .     .
     .         .         .         .          .     .
```

So that the one-year "survival" (remaining in the nursing home) is 37.9%, 20.2%, 27.0% and 11.6% for single healthy, single unhealthy, married healthy and married unhealthy individuals respectively.

Estimating P-year survival through the $\hat{\Lambda}(P; \mathbf{Z})$

We can use R and the command `predict` to calculate the estimated cumulative hazard at $t = P$ and the estimated survival $\hat{S}(P; \mathbf{Z})$:

```
newdata3 <- data.frame(married = c(0,0,1,1),
                       health = c(2,5,2,5),
                       fail= c(1,1,1,1), los=365)

predict(fit.cox, newdata=newdata3, type="expected")
[1] 0.9701205 1.5966059 1.3065521 2.1502986

exp(-predict(fit.cox, newdata=newdata3, type="expected"))
[1] 0.3790374 0.2025829 0.2707520 0.1164494
```

Model Selection in Survival Analysis

Suppose we have a censored survival time that we want to model as a function of a (possibly) set of covariates. Two important questions are:

- How to decide which covariates to use
- How to decide if the final model fits well

To address these topics, we'll consider a new example:

Survival of Atlantic Halibut - Smith et al

Obs #	*Survival Time* (min)	*Censoring* Indicator	*Tow Duration* (min.)	Diff in *Depth*	*Length* of Fish (cm)	*Handling* Time (min.)	Total *log(catch)* ln(weight)
100	353.0	1	30	15	39	5	5.685
109	111.0	1	100	5	44	29	8.690
113	64.0	0	100	10	53	4	5.323
116	500.0	1	100	10	44	4	5.323
⋮							

Process of Model Selection

Collett (Section 3.6) has an excellent discussion of various approaches for model selection. In practice, model selection proceeds through a combination of

- knowledge of the science
- trial and error, common sense
- automatic variable selection procedures
  - forward selection
  - backward selection
  - stepwise seletion

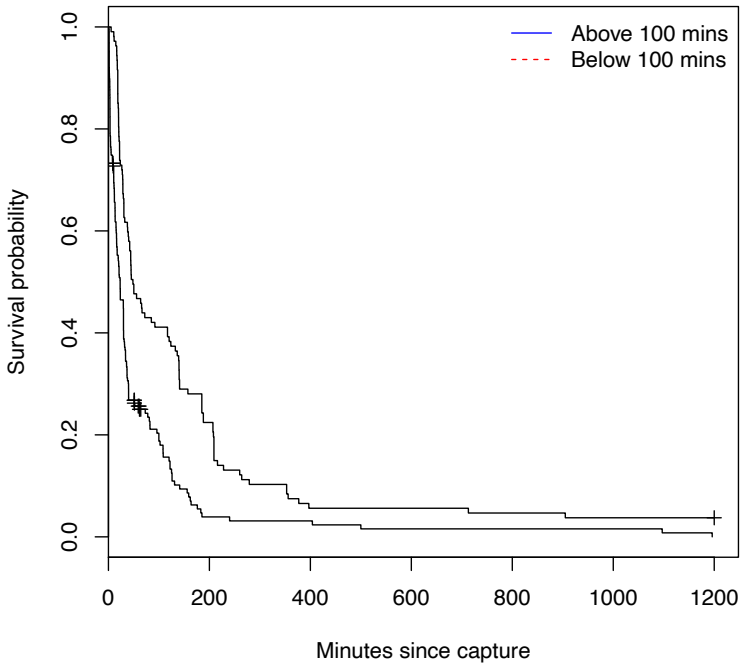
Many advocate the approach of first doing a univariate analysis to "screen" out potentially significant variables for consideration in the multivariate model (see Collett).

Let's start with this approach!

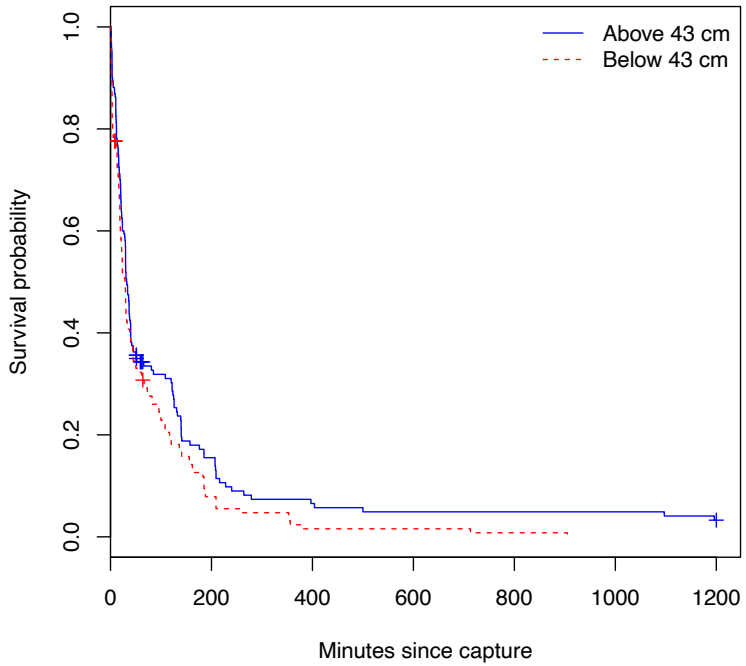
Univariate KM plots of Atlantic Halibut survival

Note: Continuous variables have been dichotomized at the median.

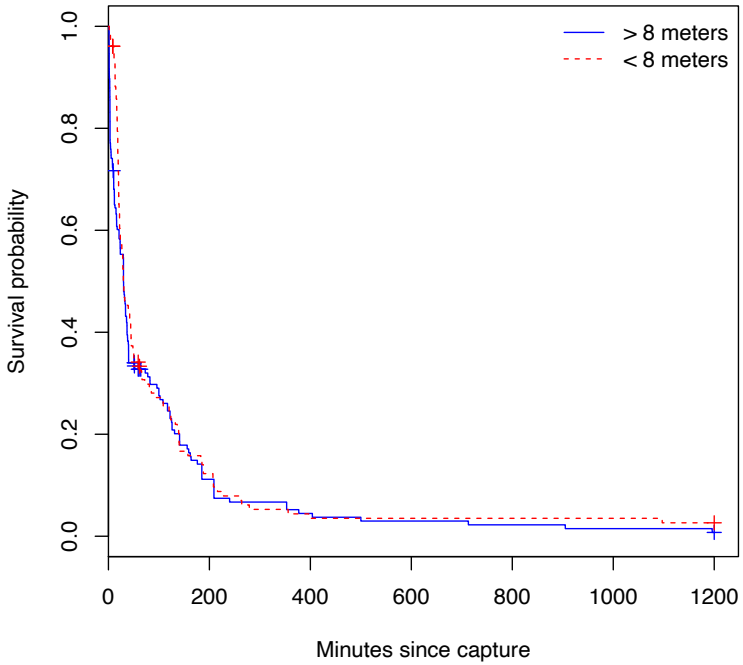
Tow duration (mins)



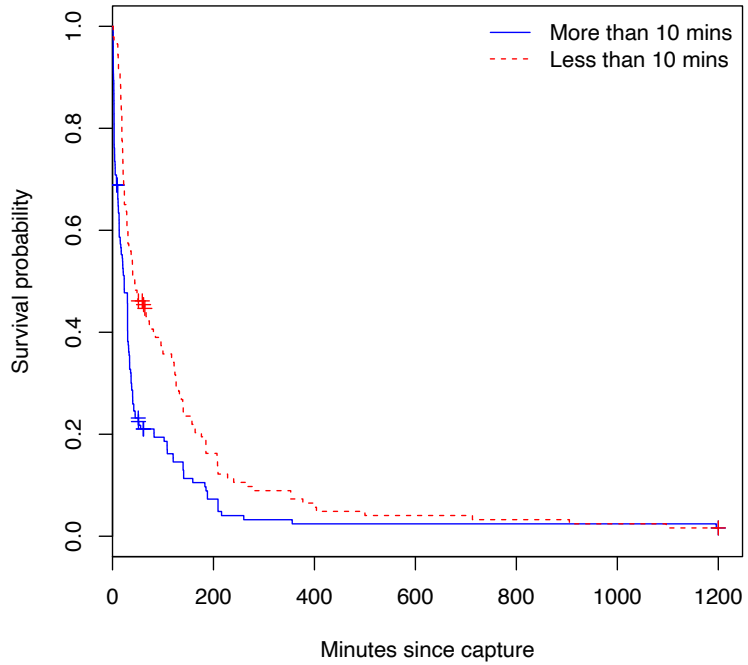
length (cm)



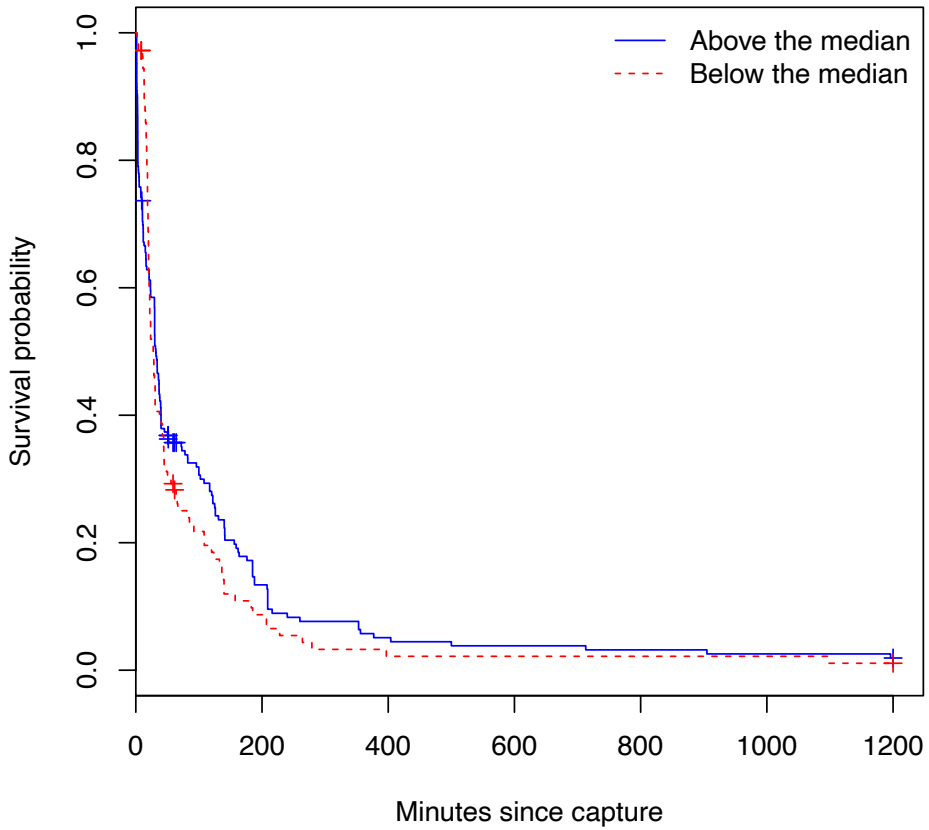
Depth (meters)



handling (mins)



log-catch



Which covariates look like they might be important?

Collett's Model Selection Approach

This approach assumes that all variables are considered to be on an equal footing, and there is no *a priori* reason to include any specific variables (like treatment).

(1) Fit a univariate model for each covariate, and identify the predictors significant at some level p_1, say 0.20.

(2) Fit a multivariate model with all significant univariate predictors, and use *backward* selection to eliminate non-significant variables at some level p_2, say 0.10.

(3) Starting with final step (2) model, consider each of the non-significant variables from step (1) using *forward* selection, with significance level p_3, say 0.10.

(4) Do final pruning of main-effects model (omit variables that are non-significant, add any that are significant), using *stepwise* regression with significance level p_4. At this stage, you may also consider adding interactions between any of the main effects currently in the model, under the hierarchical principle.

Collett recommends using a likelihood ratio test for all variable inclusion/exclusion decisions.

Model selection in R

R fits various models applying the AIC criteria described by Collett:

$$\text{minimize AIC} = -2 \log(\hat{L}) + (k * q)$$

where q is the number of unknown parameters in the model and α is typically between 2 and 6 (they suggest $k = 3$).

The model is then chosen which minimizes the AIC (similar to maximizing log-likelihood, but with a penalty for number of variables in the model)

R command for Forward selection

Forward Selection

We use the "forward" option. Each step is carried out so that the Akaike Information Criterion (Aic) is minimized. The R command is as follows:

```
fit.stepForw = stepAIC(fitmin,scope = list(lower = formula(fitmin),
                                           upper = formula(fitmax)),
                       direction = "forward",k = 3,trace = 1)
summary(fit.stepForw)
```

Where `fitmin` is the empty Cox model (i.e., a model without covariates) and `fitmax` is the saturated model (i.e., the Cox model with all possible covariates in it).

Output

The results are given in the following output:

```
Call:
coxph(formula = Surv(survtime, censor) ~ handling + logcatch +
    towdur + length, data = halibut)

  n= 294, number of events= 273

              coef exp(coef)  se(coef)      z Pr(>|z|)
handling  0.054947  1.056485  0.009870  5.567 2.59e-08 ***
logcatch -0.183373  0.832458  0.050982 -3.597 0.000322 ***
towdur    0.007744  1.007774  0.002017  3.839 0.000123 ***
length   -0.036960  0.963715  0.010028 -3.686 0.000228 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

          exp(coef) exp(-coef) lower .95 upper .95
handling     1.0565     0.9465    1.0362    1.0771
logcatch     0.8325     1.2013    0.7533    0.9199
towdur       1.0078     0.9923    1.0038    1.0118
length       0.9637     1.0377    0.9450    0.9828

Concordance= 0.683  (se = 0.02 )
Rsquare= 0.25   (max possible= 1 )
Likelihood ratio test= 84.5   on 4 df,   p=0
Wald test            = 90.71  on 4 df,   p=0
Score (logrank) test = 94.51  on 4 df,   p=0
```

R cmmand for Backward selection

Backward Selection We use the "backward" option. Each step is carried out so that the Akaike Information Criterion (Aic) is minimized. The R command is as follows:

```
fit.backward = stepAIC(fitmax,scope = list(lower = formula(fitmin),
                                           upper = formula(fitmax)),
                       direction = "backward", k = 3,trace = 1)
summary(fit.backward)
```

Output

```
Call:
coxph(formula = Surv(survtime, censor) ~ towdur + length + handling +
    logcatch, data = halibut)

  n= 294, number of events= 273

              coef exp(coef)  se(coef)      z Pr(>|z|)
towdur    0.007744  1.007774  0.002017  3.839 0.000123 ***
length   -0.036960  0.963715  0.010028 -3.686 0.000228 ***
handling  0.054947  1.056485  0.009870  5.567 2.59e-08 ***
logcatch -0.183373  0.832458  0.050982 -3.597 0.000322 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

          exp(coef) exp(-coef) lower .95 upper .95
towdur       1.0078     0.9923    1.0038    1.0118
length       0.9637     1.0377    0.9450    0.9828
handling     1.0565     0.9465    1.0362    1.0771
logcatch     0.8325     1.2013    0.7533    0.9199

Concordance= 0.683  (se = 0.02 )
Rsquare= 0.25   (max possible= 1 )
Likelihood ratio test= 84.5   on 4 df,   p=0
Wald test            = 90.71  on 4 df,   p=0
Score (logrank) test = 94.51  on 4 df,   p=0
```

R command for Stepwise selection

The stepwise procedure has a forward and a backward version.

The only difference is whether one starts with the empty or the saturated model. In both cases the model is refit and all variables that are not to be included in the model are discarded.

Forward and Backward stepwise selection

Forward stepwise Selection

We use the "both" option. Each step is carried out so that the Akaike Information Criterion (Aic) is minimized. The R command is as follows:

```
fit.stepBack = stepAIC(fitmin, scope = list(lower = formula(fitmin),
                                            upper = formula(fitmax)),
                       direction = "both", k = 3, trace = 1)
summary(fit.stepBack)
```

Backward stepwise Selection

```
fit.stepBack = stepAIC(fitmax, scope = list(lower = formula(fitmin),
                                            upper = formula(fitmax)),
                       direction = "both", k = 3, trace = 1)
summary(fit.stepBack)
```

Where `fitmin` is the empty Cox model (i.e., a model without covariates) and `fitmax` is the saturated model (i.e., the Cox model with all possible covariates in it).

Output: Backward stepwise selection

The results are given in the following output:

```
Call:
coxph(formula = Surv(survtime, censor) ~ towdur + length + handling +
    logcatch, data = halibut)

  n= 294, number of events= 273

              coef exp(coef)  se(coef)      z Pr(>|z|)
towdur    0.007744  1.007774  0.002017  3.839 0.000123 ***
length   -0.036960  0.963715  0.010028 -3.686 0.000228 ***
handling  0.054947  1.056485  0.009870  5.567 2.59e-08 ***
logcatch -0.183373  0.832458  0.050982 -3.597 0.000322 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

          exp(coef) exp(-coef) lower .95 upper .95
towdur       1.0078     0.9923    1.0038    1.0118
length       0.9637     1.0377    0.9450    0.9828
handling     1.0565     0.9465    1.0362    1.0771
logcatch     0.8325     1.2013    0.7533    0.9199

Concordance= 0.683  (se = 0.02 )
Rsquare= 0.25   (max possible= 1 )
Likelihood ratio test= 84.5   on 4 df,   p=0
Wald test            = 90.71  on 4 df,   p=0
Score (logrank) test = 94.51  on 4 df,   p=0
```

Output: Forward stepwise selection

```
Call:
coxph(formula = Surv(survtime, censor) ~ handling + logcatch +
    towdur + length, data = halibut)

  n= 294, number of events= 273

              coef exp(coef)  se(coef)      z Pr(>|z|)
handling  0.054947  1.056485  0.009870  5.567 2.59e-08 ***
logcatch -0.183373  0.832458  0.050982 -3.597 0.000322 ***
towdur    0.007744  1.007774  0.002017  3.839 0.000123 ***
length   -0.036960  0.963715  0.010028 -3.686 0.000228 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

         exp(coef) exp(-coef) lower .95 upper .95
handling    1.0565     0.9465    1.0362    1.0771
logcatch    0.8325     1.2013    0.7533    0.9199
towdur      1.0078     0.9923    1.0038    1.0118
length      0.9637     1.0377    0.9450    0.9828

Concordance= 0.683  (se = 0.02 )
Rsquare= 0.25   (max possible= 1 )
Likelihood ratio test= 84.5   on 4 df,   p=0
Wald test            = 90.71  on 4 df,   p=0
Score (logrank) test = 94.51  on 4 df,   p=0
```

Notes

- When the halibut data was analyzed with the forward, backward and stepwise options, the same final model was reached. However, this will not always be the case.
- Variables can be forced into the model.

Assessing overall model fit

How do we know if the model fits well?

- Always look at univariate plots (Kaplan-Meiers) Construct a Kaplan-Meier survival plot for each of the important predictors, like the ones shown at the beginning of these notes.
- Check proportionality assumption
- **Check residuals!**
  - **(a)** generalized (Cox-Snell)
  - **(b)** martingale
  - **(c)** deviance
  - **(d)** Schoenfeld
  - **(e)** weighted Schoenfeld

Using residuals to test model fit

Residuals for survival data are slightly different than for other types of models, due to the censoring.

Before we start talking about residuals, we need an important basic result:

Inverse CDF:
If T_i (the survival time for the i-th individual) has survivorship function $S_i(t)$, then the transformed random variable $S_i(T_i)$ (i.e., the survival function evaluated at the actual survival time T_i) should be from a uniform distribution on $[0, 1]$, and hence $-\log[S_i(T_i)]$ should be from a unit exponential distribution.

More mathematically,

$$\begin{aligned} \text{If} \quad T_i &\sim S_i(t) \\ \text{then} \quad S_i(T_i) &\sim \mathit{Uniform}[0,1] \\ \text{and} \quad -\log S_i(T_i) &\sim \mathit{Exponential}(1) \end{aligned}$$

The fact that $S(T_i) \sim U[0,1]$ results from the definition of the survival function, $S(x) = P(T > x) = 1 - P(T \leq x)$.

Now consider that $P(S(T) \leq x) = P(T \geq S^{-1}(x))$ by the fact that $g^{-1}g(x) = x$, where $g^{-1}(x)$ is the unique inverse of x, and because $S(\cdot)$ is non-increasing (so we reverse the sign of the inequality).

Thus, $P(S(T) \leq x) = S(S^{-1}(x)) = x$ which is the definition of a uniformly-distributed random variable.

The fact that $-\log S_i(T_i) \sim \mathit{Exponential}(1)$ results from the Inverse CDF Theorem.

Generalized (Cox-Snell) Residuals

The implication of the last result is that if the model is correct, the estimated cumulative hazard for each individual at the time of their death or censoring should be like a censored sample from a unit exponential. This quantity is called the *generalized* or *Cox-Snell* residual.

Here is how the generalized residual might be used. Suppose we fit a PH model:

$$S(t;Z) = [S_0(t)]^{\exp(\beta Z)}$$

or, in terms of hazards:

$$\begin{aligned} \lambda(t;Z) &= \lambda_0(t)\exp(\beta Z) \\ &= \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_k Z_k) \end{aligned}$$

After fitting, we have:

- $\hat{\beta}_1, \ldots, \hat{\beta}_k$
- $\hat{S}_0(t)$

So, for each person with covariates $\mathbf{Z}_i$, we can get

$$\hat{S}(t; \mathbf{Z}_i) = [\hat{S}_0(t)]^{\exp(\beta \mathbf{Z}_i)}$$

This gives a predicted survival probability at each time t in the dataset (see notes from the previous lecture).

Then we can calculate

$$\hat{\Lambda}_i = -\log[\hat{S}(T_i; Z_i)]$$

In other words, first we find the predicted survival probability at the actual survival time for an individual, then log-transform it.

Based on the properties of a unit exponential model

- plotting $-\log(\hat{S}(t))$ vs t should yield a straight line
- plotting $\log[-\log S(t)]$ vs $\log(t)$ should yield a straight line through the origin with slope=1.

To convince yourself of this, start with $S(t) = e^{-\lambda t}$ and calculate $\log[-\log S(t)]$. What do you get for the slope and intercept?

(Note: this does not necessarily mean that the underlying distribution of the original survival times is exponential!)

Obtaining residuals from R

- Fit a Cox PH model with `coxph` comand
- Use the `residuals` command
- Use the `type` option to specify the type of the residual you want to obtain. R provides the following types of residuals:
  - martingale residuals
  - deviance residuals
  - score residuals
  - schoenfeld and scaled Schoenfeld residuals
  - dfbeta, dfbetas
  - partial residuals

Obtaining Cox-Snell residuals from R

We obtain Cox-Snell residuals from R as follows:

- Use the `predict` command with the `csnell` option
- Define a survival dataset using the Cox-Snell residuals as the "pseudo" failure times
- Calculate the estimated KM survival
- Take the $\log[-\log(S(t))]$ based on the above
- Generate the log of the Cox-Snell residuals
- Graph $\log[-\log S(t)]$ vs $\log(t)$

Cox-Snell residuals in the halibut data ample

- Fit the Cox model:

```
fit = coxph( Surv(survtime,censor) ~ towdur + handling + length
> + logcatch, data = halibut)
summary(fit)
```

- Define the residuals

```
halibut$csres = halibut$censor - residuals(fit,type = "martingale")
```

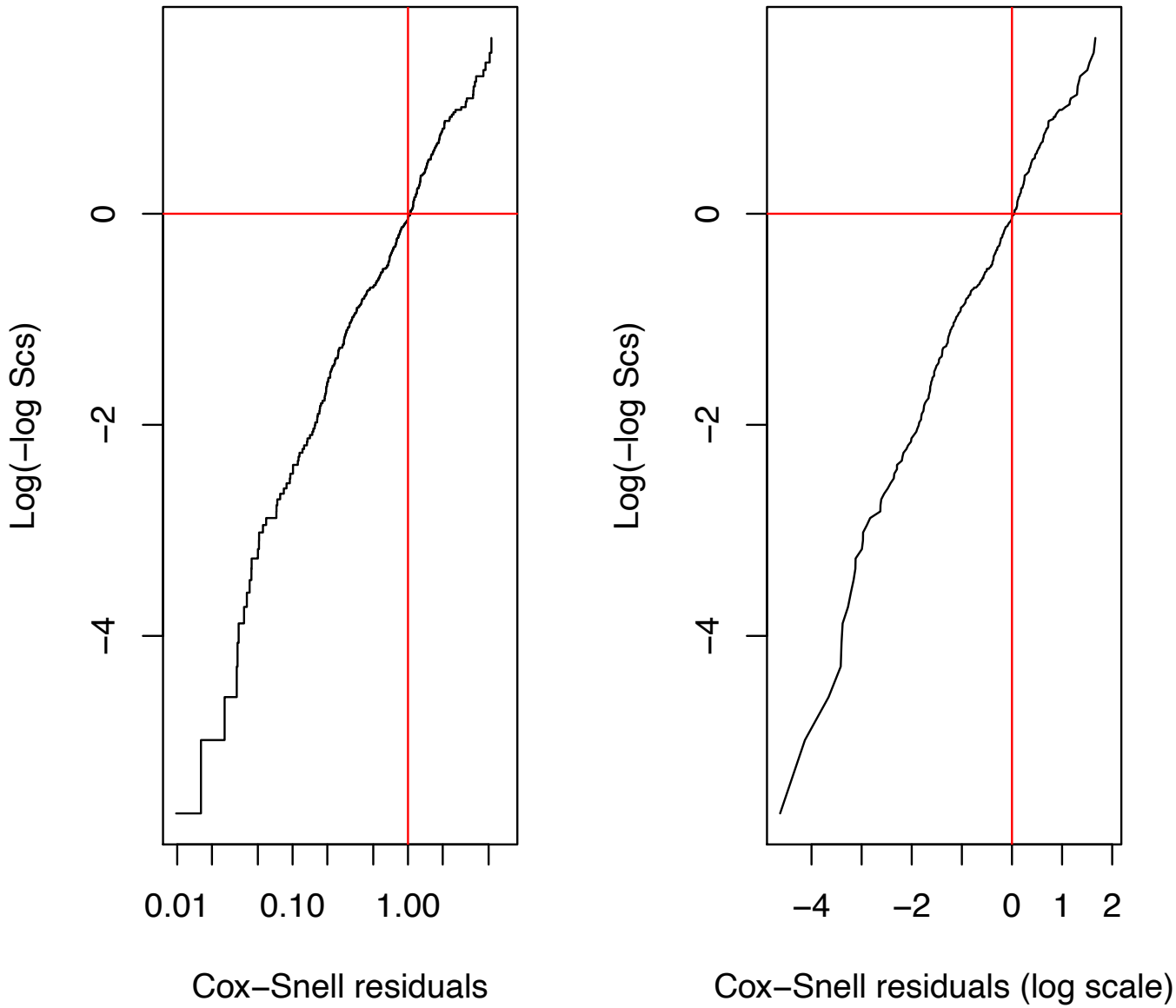
- Assess the fit

```
fitcs = survfit(Surv(csres,censor) ~ 1,data = halibut)
```

- Plot the results

```
plot(fitcs,fun = "cloglog",conf.int = F,mark.time = F,
     xlab = "Cox-Snell residuals (log scale)",ylab = "Log(-log Scs)")
abline(h = 0,col = "red")
abline(v = 1,col = "red") # Due to log scale!
```

Cox-Snell residuals in the halibut data example



Notes

The fit from the Cox-Snell residuals is fairly good, which suggests that the model fits adequately.

Nevertheless, Allison states that the "Cox-Snell residuals... are not very informative for Cox models estimated by partial likelihood."

Martingale Residuals

Martingale residuals are defined for the i-th individual as:

$$r_i = \delta_i - \hat{\Lambda}(T_i)$$

Properties of the Martingale residuals

The Martingale residuals have the following properties:

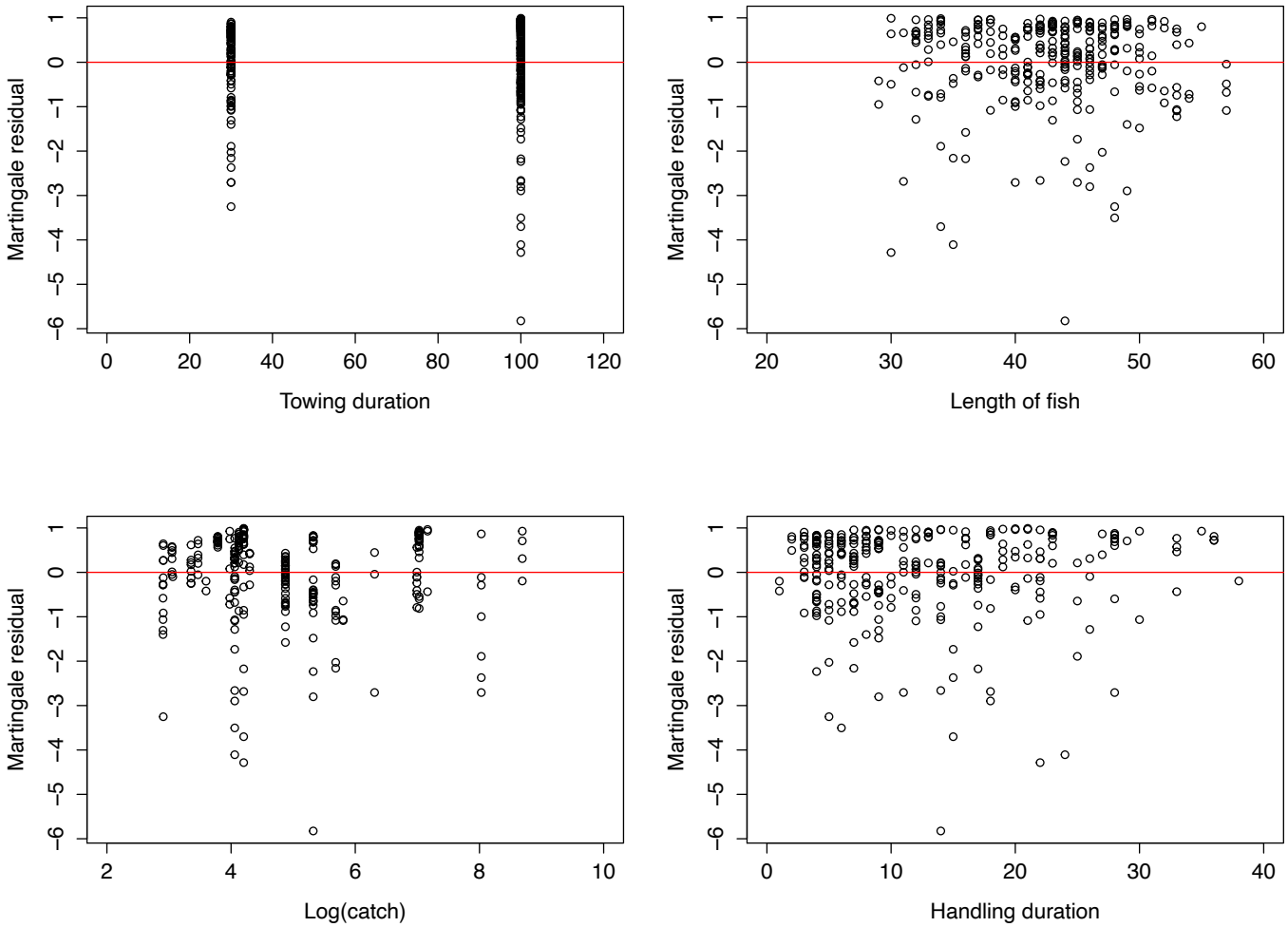
- r_i's have mean 0
- range of r_i's is between $-\infty$ and 1
- approximately uncorrelated (in samples)
- **Interpretation:** - the residual r_i can be viewed as the difference between the observed number of deaths (0 or 1) for subject i between time 0 and T_i, and the expected numbers based on the fitted model.

Obtaining martingale residuals with R

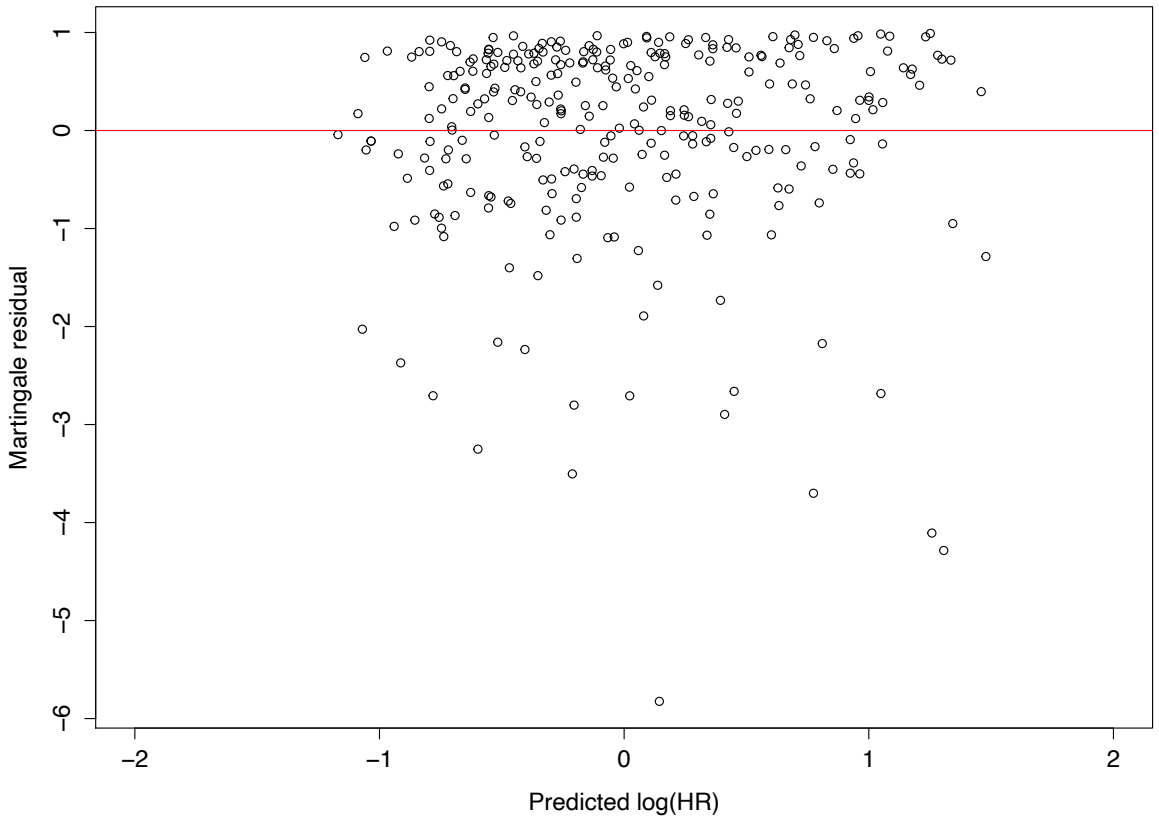
Martingale residuals are obtained through the `residuals` command (using `type="martingale"`).

Once the martingale residual is created, you can plot it versus the predicted log HR (i.e., $\beta \mathbf{Z}_i$), or any of the individual covariates.

Martingale residuals in the halibut data example



Martingale residuals against the predicted log hazard ratio



Deviance Residuals

One problem with the martingale residuals is that they tend to be asymmetric.

A solution is to use **deviance residuals**. For person i, these are defined as a function of the martingale residuals (r_i):

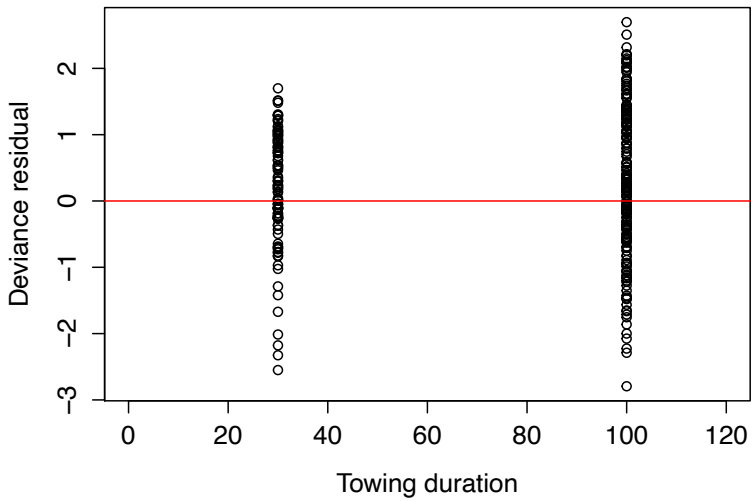
$$\hat{D}_i = \text{sign}(\hat{r}_i)\sqrt{-2[\hat{r}_i + \delta_i log(\delta_i - \hat{r}_i)]}$$

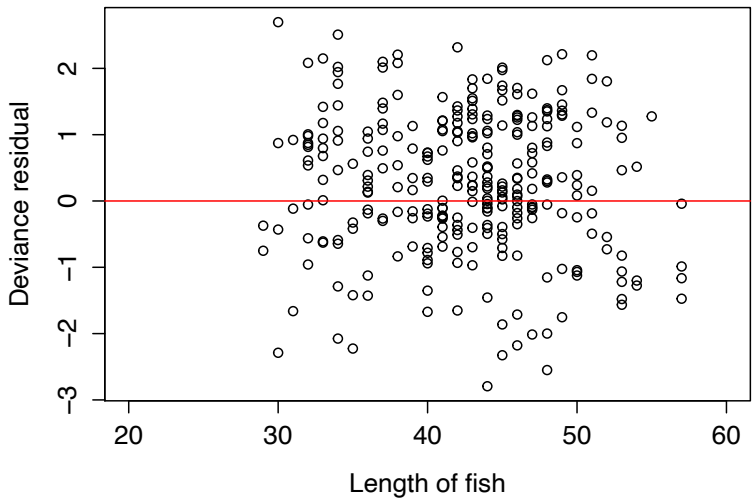
Obtaining deviance residuals in R

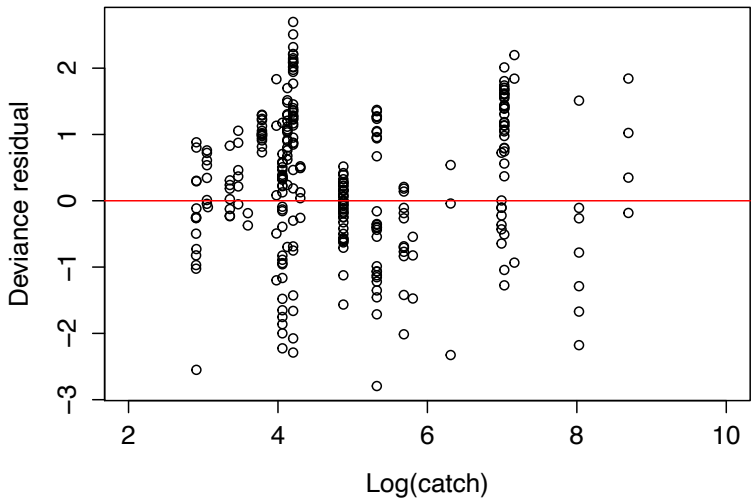
In R, the deviance residuals are generated using `residuals` command with the `type="deviance"` option:
The deviance residuals can then be plotted versus the predicted log(HR) or the individual covariates, as shown for the Martingale residuals.

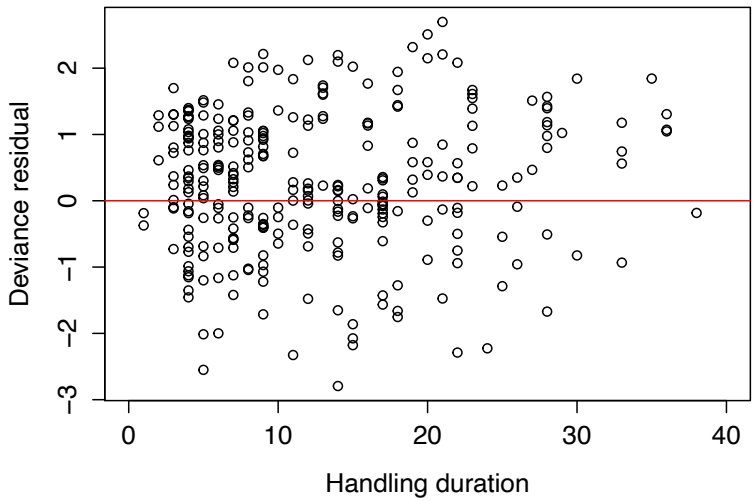
Deviance residuals behave much like residuals from OLS regression (i.e., mean=0, s.d.=1). They are negative for observations with survival times that are smaller than expected.

Deviance Residuals

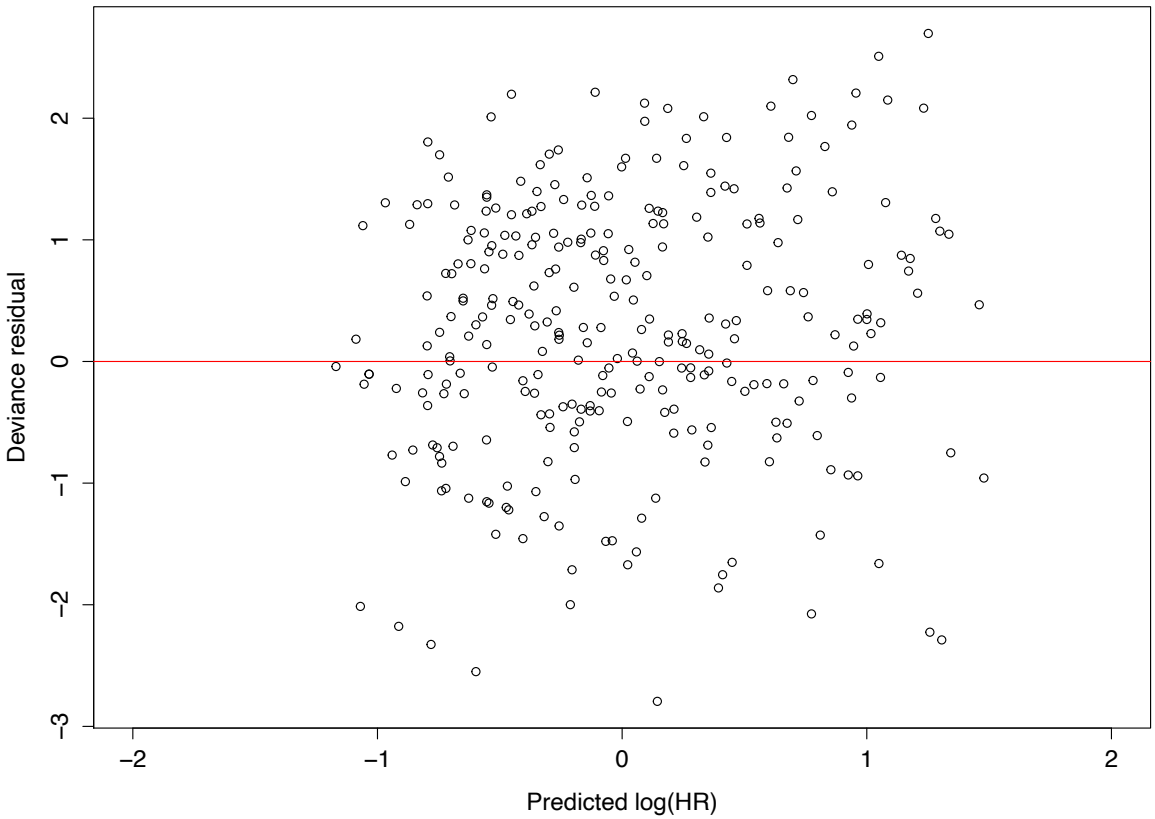








Deviance residuals versus predicted log hazard ratio



Schoenfeld residuals

These are defined at each observed failure time as:

$$r_{ij}^{s} = Z_{ij}(t_i) - \bar{Z}_j(t_i)$$

where

$$\bar{Z}_j(t_i) = \frac{\sum_{\ell \in \mathcal{R}(\tau_j)} Z_\ell e^{\beta Z_\ell}}{\sum_{\ell \in \mathcal{R}(\tau_j)} e^{\beta Z_\ell}}$$

is the average value of the covariates at each failure time.

Notes

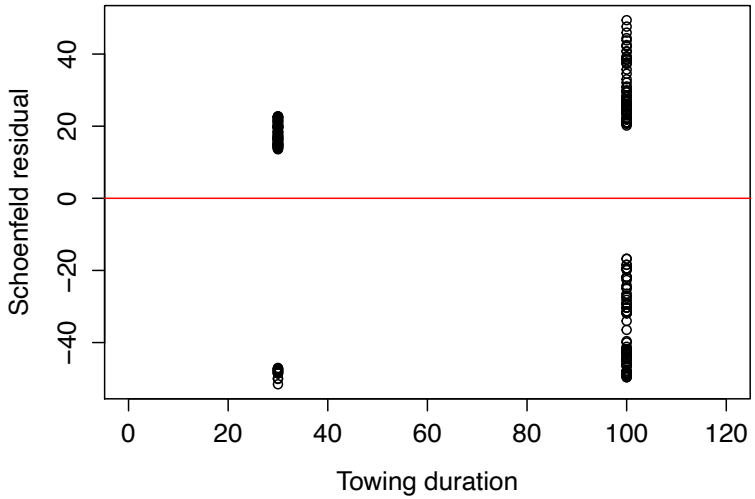
The following are some comments about the Schoenfeld residuals:

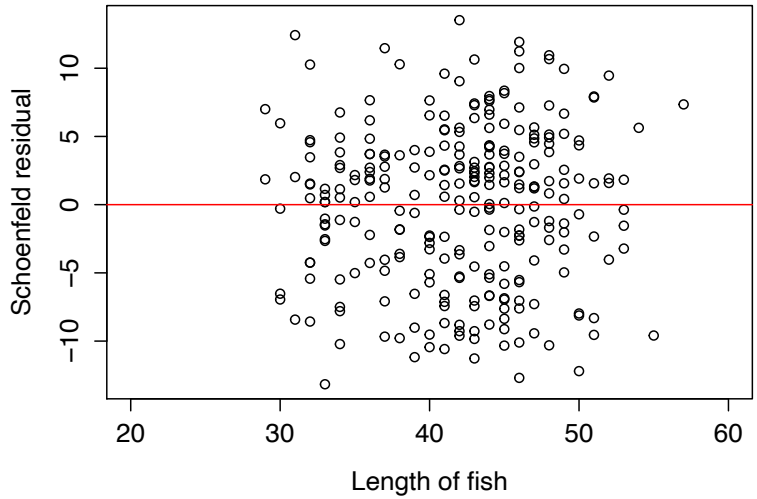
- represent the difference between the observed covariate and the average over the risk set at that time
- calculated for each covariate
- not defined for censored failure times.
- useful for assessing time trend or lack or proportionality, based on plotting versus event time
- sum to zero, have expected value zero, and are uncorrelated (in samples)

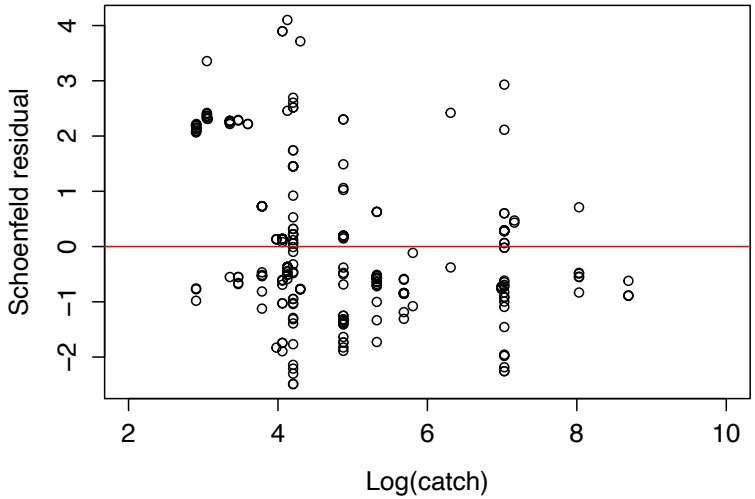
‘

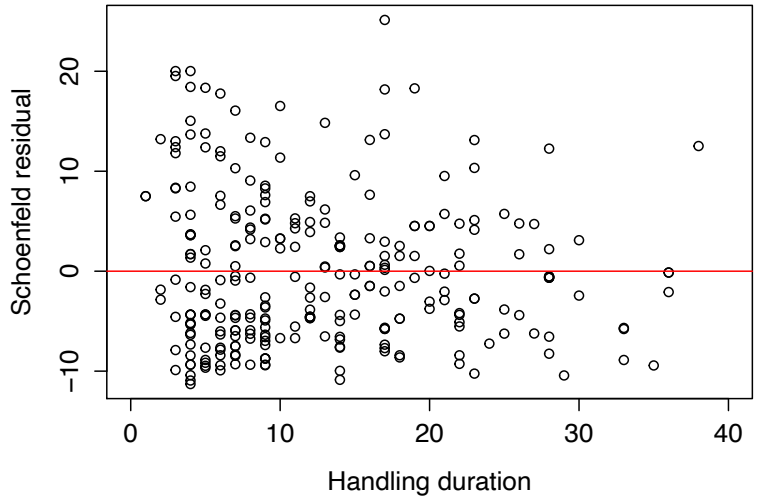
In R, the Schoenfeld residuals are generated in the `residuals` command itself, using the `type="schoenfeld"` option.

Schoenfeld residuals in the halibut data example









Weighted Schoenfeld Residuals

These are actually used more often than the previous unweighted version, because they are more like the typical OLS residuals (i.e., symmetric around 0).

They are defined as:

$$r_{ij}^{w} \quad = \quad n\widehat{V}\ r_{ij}^{s}$$

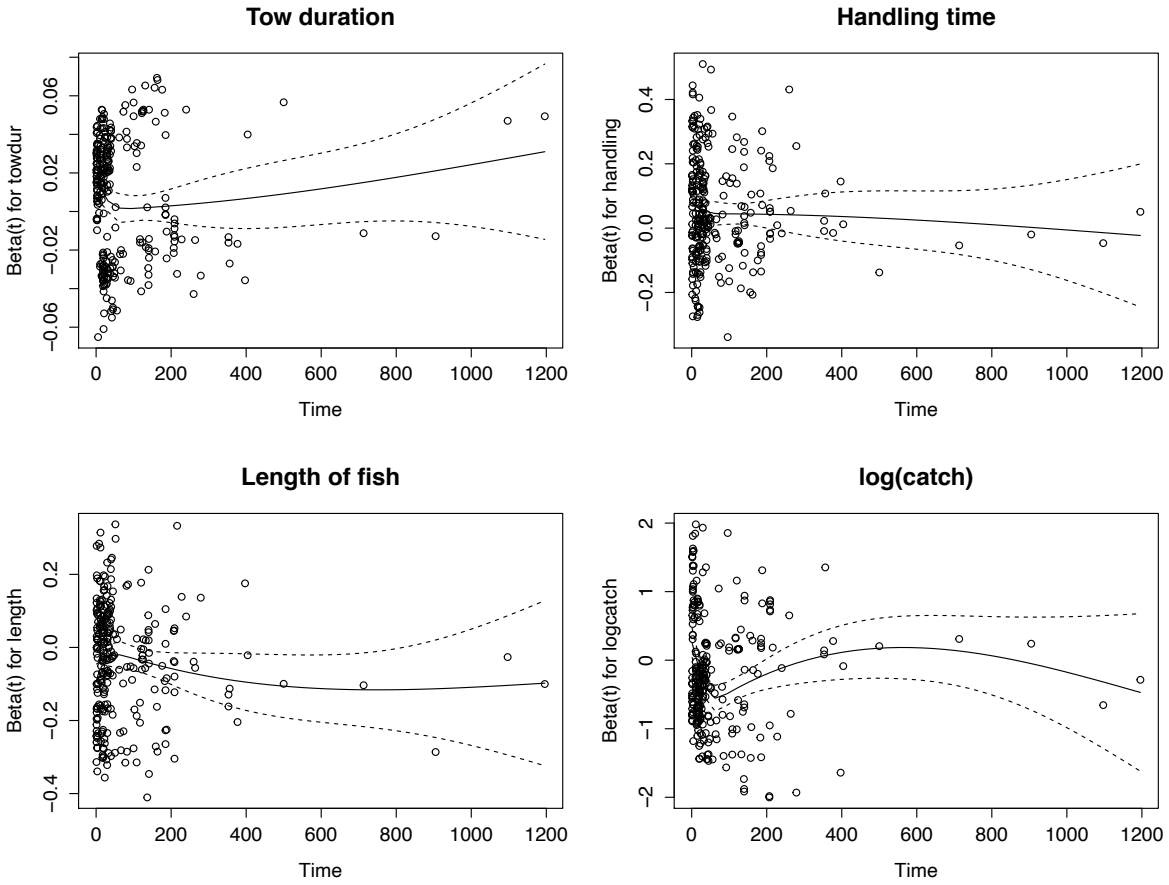
where $\widehat{V}$ is the estimated variance of $\hat{\boldsymbol{\beta}}$. The weighted residuals can be used in the same way as the unweighted ones to assess time trends and lack of proportionality.

Obtaining the weighted Schoenfeld residuals from R

In R, use the command `cox.zph` with the option.

```
cox.zph(fit,transform = "identity")
```

This results in the following plots:



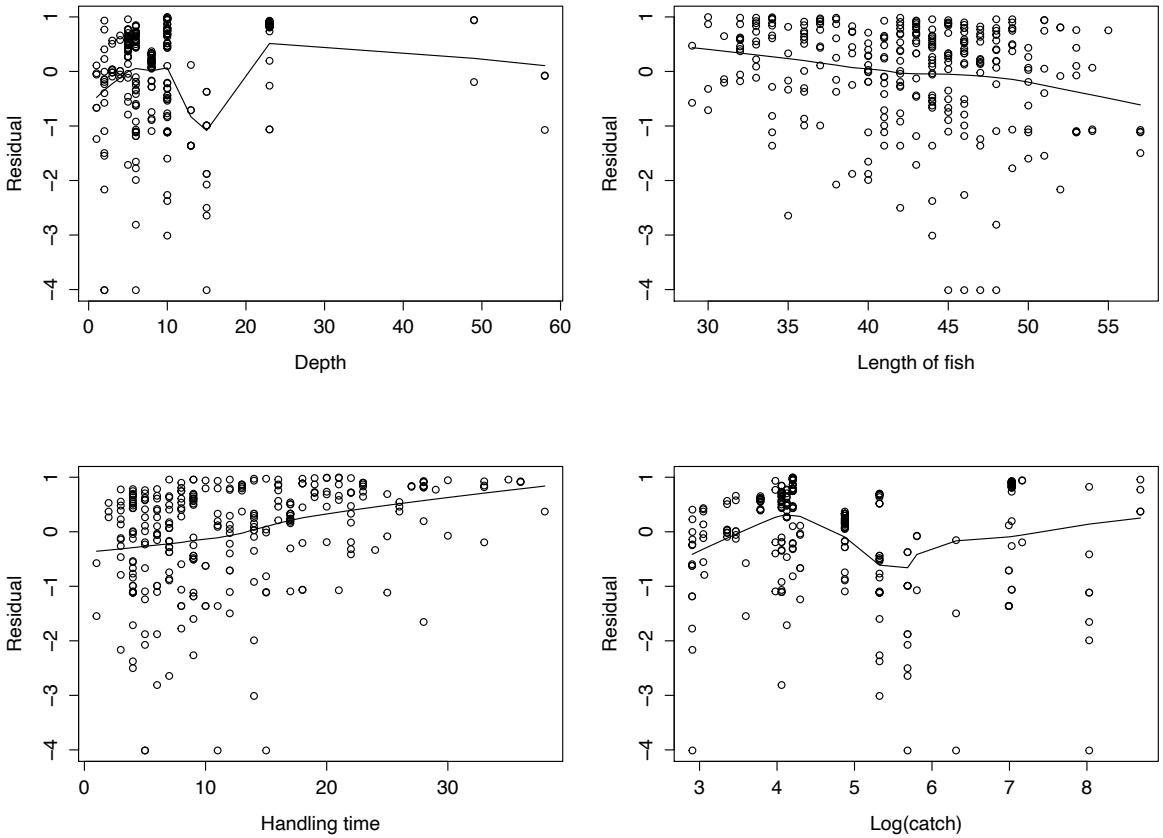
Using Residual plots to explore relationships

If you calculate martingale or deviance residuals without any covariates in the model and then plot against covariates, you obtain a graphical impression of the relationship between the covariate and the hazard.

In R, it is easy to do this (also possible in stata using the "estimate" option)

Residual plots

We plot the martingale residuals versus each of the covariates. This produces the following plots:



Can we improve the model?

The plots appear to have some structure, which indicate that we could be leaving something out. It is always a good idea to check for interactions.

In this case, there are several important interactions. I used a backward selection model forcing all main effects to be included, and considering all pairwise interactions.

Backward elimination with pairwise interactions

The results from this analysis are as follows:

```
Call:
coxph(formula = Surv(survtime, censor) ~ towdur + depth + length +
    handling + logcatch + towdepth + lengthdepth + handlingdepth +
    towlength + towhandling, data = halibut)

  n= 294, number of events= 273

                   coef  exp(coef)   se(coef)      z Pr(>|z|)
towdur        -0.0756235  0.9271652  0.0174010 -4.346 1.39e-05 ***
depth          0.1249947  1.1331424  0.0639359  1.955 0.050583 .
length        -0.0775731  0.9253594  0.0255046 -3.042 0.002354 **
handling       0.0045787  1.0045892  0.0321858  0.142 0.886876
logcatch      -0.2241951  0.7991592  0.0715613 -3.133 0.001731 **
towdepth       0.0029236  1.0029279  0.0004994  5.854 4.79e-09 ***
lengthdepth   -0.0060456  0.9939727  0.0013568 -4.456 8.36e-06 ***
handlingdepth -0.0041262  0.9958823  0.0011834 -3.487 0.000489 ***
towlength      0.0011826  1.0011833  0.0003540  3.340 0.000836 ***
towhandling    0.0011146  1.0011152  0.0003556  3.134 0.001723 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Interpretation: Handling alone doesn't seem to affect survival, unless it is combined with a longer towing duration or shallower trawling depths.

Assessing the PH Assumption

So far, we've been considering the following Cox PH model:

$$\lambda(t, \mathbf{Z}) = \lambda_0(t) \ \exp(\beta \mathbf{Z}) = \lambda_0(t) \ \exp\left(\sum \beta_j Z_j\right)$$

where β_j is the parameter for the the j-th covariate (Z_j).

Important features of this model:

(1) the baseline hazard depends on t, but not on the covariates $Z_1, ..., Z_p$

(2) the hazard ratio, i.e., $\exp(\beta \mathbf{Z})$, depends on the covariates $\mathbf{Z} = (Z_1, ..., Z_p)$, but not on time t.

Assumption (2) is what led us to call this a proportional hazards model. That's because we could take the ratio of the hazards for two individuals with covariates $\mathbf{Z}_i$ and $\mathbf{Z}_{i'}$, and write it as a constant in terms of the covariates.

Hazard Ratio:

$$\begin{aligned}
\frac{\lambda(t, \mathbf{Z}_i)}{\lambda(t, \mathbf{Z}_{i'})} &= \frac{\lambda_0(t)\exp(\beta \mathbf{Z}_i)}{\lambda_0(t)\exp(\beta \mathbf{Z}_{i'})} \\
&= \frac{\exp(\beta \mathbf{Z}_i)}{\exp(\beta \mathbf{Z}_{i'})} \\
&= \exp[\beta(\mathbf{Z}_i - \mathbf{Z}_{i'})] \\
&= \exp[\sum \beta_j (Z_{ij} - Z_{i'j})] \; = \; \theta
\end{aligned}$$

In the last formula, Z_{ij} is the value of the j-th covariate for the i-th individual. For example, Z_{42} might be the value of GENDER (0 or 1) for the the 4-th person.

We can also write the hazard for the i-th person as a constant times the hazard for the i'-th person:

$$\lambda(t, \mathbf{Z}_i) \quad = \quad \theta \; \lambda(t, \mathbf{Z}_{i'})$$

Thus, the HR between two types of individuals is constant (i.e., $=\theta$) over time. These are mathematical ways of stating the proportional hazards assumption.

Ways to check the PH assumption

There are several options for checking the assumption of proportional hazards:

I. **Graphical**
   - **(a)** Plots of survival estimates for two subgroups
   - **(b)** Plots of $\log[-\log(\hat{S})]$ vs $\log(t)$ for two subgroups
   - **(c)** Plots of weighted Schoenfeld residuals vs time
   - **(d)** Plots of observed survival probabilities versus expected under PH model

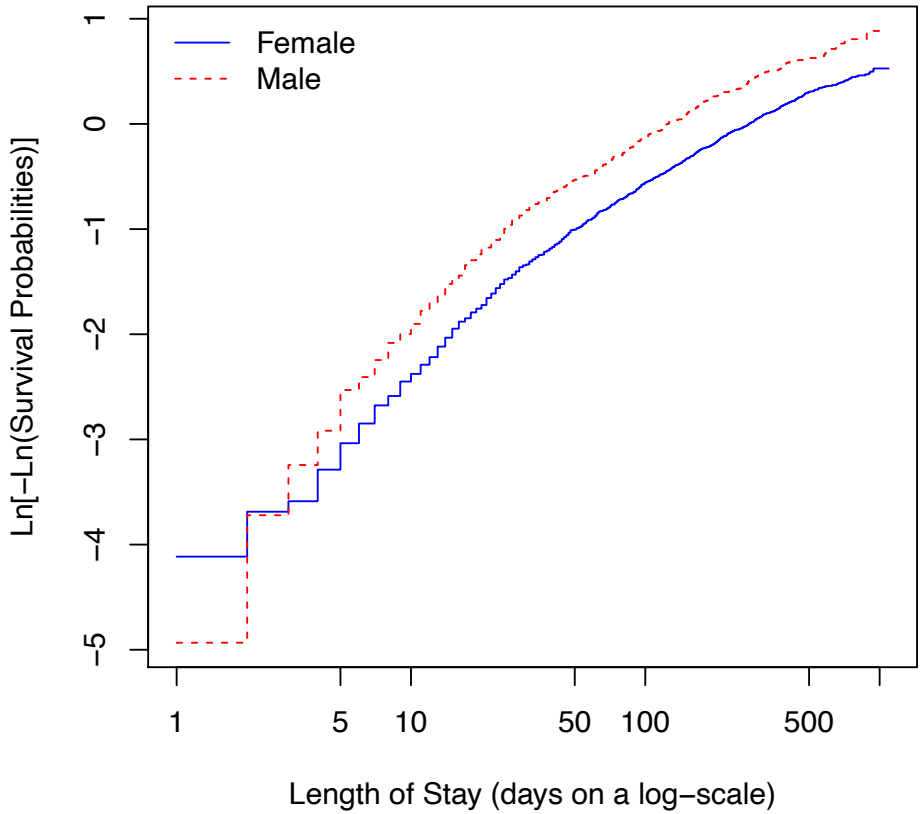
II. **Use of goodness of fit tests** - we can construct a goodness-of-fit test based on comparing the observed survival probability (from `sts list`) with the expected (from `stcox`) under the assumption of proportional hazards - see Kleinbaum ch.4

III. **Including interaction terms between a covariate and** t (time-dependent covariates)

Example: Nursing Home - gender

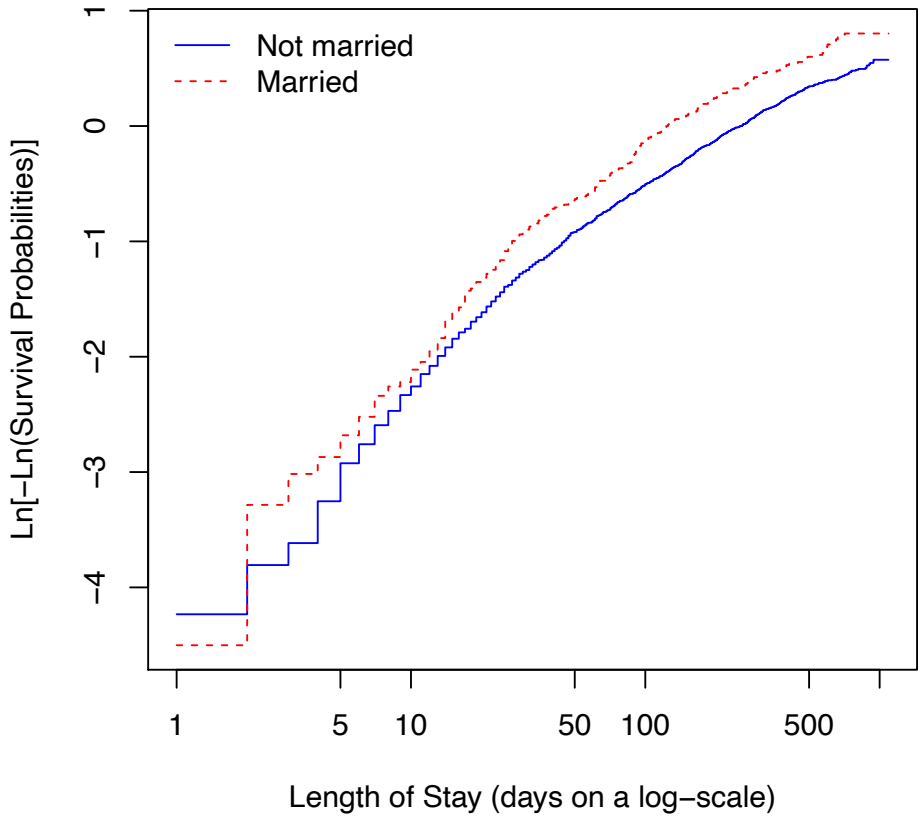
```
fitKMgen = survfit( Surv(los,fail) ~ gender,data = nurshome)

plot(fitKMgen, mark.time = F, fun = "cloglog",
     xlab = "Length of Stay (days on a log-scale)", ylab = "Ln[-Ln(Survival Probabilities)]",
     lty = 1:2,col = c("blue","red"))
legend("topleft",lty = 1:2,col = c("blue","red"),bty="n",legend = c("Female","Male"))
```



Example: Nursing Home - marital status

A similar situation is the case with respect to the effect of marital status;



This is equivalent to comparing plots of the log cumulative hazard, $\log(\hat{\Lambda}(t))$, between the covariate levels, since

$$\Lambda(t) = \int_0^t \lambda(u; Z)du = -\log[S(t)]$$

How do we interpret the above?

A strategy of assuming that PH holds unless there is very strong evidence to counter this assumption:

- estimated survival curves are fairly separated, then cross
- estimated log cumulative hazard curves cross, or look very unparallel over time
- weighted Schoenfeld residuals clearly increase or decrease over time (you could fit a OLS regression line and see if the slope is significant)
- test for time $\times$ covariate interaction term is significant (this relates to time-dependent covariates)

What if the PH model does not hold?

If PH doesn't exactly hold for a particular covariate but we fit the PH model anyway, then what we are getting is sort of an average HR, averaged over the event times.

In most cases, this is not such a bad estimate. Allison claims that too much emphasis is put on testing the PH assumption, and not enough to other important aspects of the model.

Implications of proportional hazards

Consider a PH model with a single covariate, Z:

$$\lambda(t; Z) = \lambda_0(t)e^{\beta Z}$$

What does this imply for the relation between the survivorship functions at various values of Z?

Under PH,

$$\log[-\log[S(t; Z)]] = \log[-\log[S_0(t)]] + \beta Z$$

Implications for the cumulative hazard

In general, we have the following relationship:

$$\begin{aligned}
\Lambda_i(t) &= \int_0^t \lambda_i(u)du \\
&= \int_0^t \lambda_0(u)\exp(\beta \mathbf{Z}_i)du \\
&= \exp(\beta \mathbf{Z}_i)\int_0^t \lambda_0(u)du \\
&= \exp(\beta \mathbf{Z}_i)\,\Lambda_0(t)
\end{aligned}$$

This means that the ratio of the cumulative hazards is the same as the ratio of hazard rates:

$$\frac{\Lambda_i(t)}{\Lambda_0(t)} = \exp(\beta \mathbf{Z}_i) = \exp(\beta_1 Z_{1i} + \cdots + \beta_p Z_{pi})$$

Implications for the linear predictor

Using the above relationship, we can show that:

$$\begin{aligned} \beta \mathbf{Z}_i &= \log\left(\frac{\Lambda_i(t)}{\Lambda_0(t)}\right) \\ &= \log \Lambda_i(t) - \log \Lambda_0(t) \\ &= \log[-\log S_i(t)] - \log[-\log S_0(t)] \\ \text{so} \quad \log[-\log S_i(t)] &= \log[-\log S_0(t)] + \beta \mathbf{Z}_i \end{aligned}$$

What is the bottom line?

Thus, to assess if the hazards are actually proportional to each other over time

- calculate Kaplan Meier Curves for various levels of Z
- compute $\log[-\log(\hat{S}(t;Z))]$ (i.e., log cumulative hazard)
- plot vs log-time to see if they are parallel (lines or curves)

Note: If Z is continuous, break into categories.

Assessing proportionality with several covariates

If there are enough data and you only have a couple of covariates, create a new covariate that takes a different value for every combination of covariate values.

If there are too many covariates (or not enough data) for this, then there is a way to test proportionality for each variable, one at a time, using the stratification option.

What if proportional hazards fails?

If the PH assumption fails, we can do any of the following:

- do a stratified analysis
- include a time-varying covariate to allow changing hazard ratios over time
- include interactions with time

The second these two options relate to time-dependent covariates, which is something we will cover later in this course.

We will focus on the first alternative, and then the second two options will be briefly described.

Stratified Analyses

Suppose:

- we are happy with the proportionality assumption on Z_1
- proportionality simply does not hold between various levels of a second variable Z_2.

If Z_2 is discrete (with a levels) and there are enough data, fit the following **stratified model**:

$$\lambda(t; Z_1, Z_2) = \lambda_{Z_2}(t) e^{\beta Z_1}$$

For example, a new treatment might lead to a 50% decrease in hazard of death versus the standard treatment, but the hazard for standard treatment might be different for each hospital.

A stratified model can be useful both for primary analysis and for checking the PH assumption.

Assessing PH assumption for several covariates

Suppose we have several covariates ($\mathbf{Z} = Z_1,\ Z_2,\ \ldots\ Z_p$), and we want to know if the following PH model holds:

$$\lambda(t; \mathbf{Z}) = \lambda_0(t)\ e^{\beta_1 Z_1 + \ldots + \beta_p Z_p}$$

To start, we fit a model which stratifies by Z_k:

$$\lambda(t; \mathbf{Z}) = \lambda_{0Z_k}(t)\ e^{\beta_1 Z_1 + \ldots + \beta_{k-1} Z_{k-1} + \beta_{k+1} Z_{k+1} + \ldots + \beta_p Z_p}$$

Since we can estimate the survival function for any subgroup, we can use this to estimate the baseline survival function, $S_{0Z_k}(t)$, for each level of Z_k.

Then we compute $-\log S(t)$ for each level of Z_k, controlling for the other covariates in the model, and graphically check whether the log cumulative hazards are parallel across strata levels.

Time-dependent covariates

In many situations it is useful to consider covariates that change over time. These are called "time-dependent" covariates. Such are of two kinds:

1. Internal variables
   These are related to each patient and are measurable while the patient is under observation
2. External variables
   These are variables that do not depend on the physical observation of the patient such as
   1. Variables such as age that are known once the birth date or age at enrollment to the study is known
   2. Variables that are independent of any individual like levels of pollution or temperature

Time-updated covariates and the Cox model

These time-updated or dependent variables can be entered into the Cox model in direct extension of the simpler non-time-updated case

$$\lambda_i(t; \mathbf{Z}_i) = \lambda_0(t) \exp \sum_{j=1}^{n} \beta_j Z_{ij}(t)$$

where $\lambda_0(t)$ is the baseline hazard associated with all covariates being equal to zero *during all time points* t. So the Cox model is generalized as

$$\sum_{i=1}^{n} \delta_i \left\{ \sum_{j=1}^{p} \beta_j Z_{ij}(t_i) - \log \sum_{l \in R(t_i)} \exp \left(\sum_{j=1}^{p} \beta_j Z_{jl}(t_i) \right) \right\}$$

this means that we will need to have all the variable (especially internal ones) available at each event time. It is important to understand that this is no longer a proportional hazards model.

Checking the PH assumption via time-updated covariates

As noted earlier, we can check the PH assumption by introducing an interaction between the effect and time

$$\lambda(t; Z) = \lambda_0(t) \ \exp\{\beta_1 Z_1 + \cdots + \beta_p Z_p + \gamma Z_1 * t\}$$

... or carrying out a stratified analysis

$$\lambda(t; Z) = \lambda_{0,Z_1}(t) \ \exp\{\beta_2 Z_2 + \cdots + \beta_p Z_p\}$$

where $\lambda_{0,Z_i}(t)$ are baseline hazards over all levels of Z_1.

Notice that these are different analyses: In the first case we impose a *linear* change (increase or decrease depending on the sign of γ) on the hazard ratio over time, while, in the second case, the change of the baseline hazard over time in the various levels of Z_1 can be arbitrary.

We will focus here on the first case.

Parametric survival analysis

So far, we have focused primarily on nonparametric and semi-parametric approaches to survival analysis, with heavy emphasis on the Cox proportional hazards model:

$$\lambda(t, \mathbf{Z}) \quad = \quad \lambda_0(t) \ \exp(\boldsymbol{\beta}\mathbf{Z})$$

We used the following estimating approach:

- We estimated $\lambda_0(t)$ nonparametrically, using the Kaplan-Meier estimator, or using the Kalbfleisch/Prentice estimator under the PH assumption
- We estimated $\boldsymbol{\beta}$ by assuming a linear model between the log HR and covariates, under the PH model

Both estimates were based on maximum likelihood theory.

Reading: for parametric models see Collett.

Reasons for considering a parametric approach

There are several reasons why we should consider some alternative approaches based on parametric models:

- The assumption of proportional hazards might not be appropriate (based on major departures)
- If a parametric model actually holds, then we would probably gain efficiency
- We may want to handle non-standard situations like
  - interval censoring
  - incorporating population mortality
- We may want to make some connections with other familiar approaches (e.g. use of the Poisson likelihood)
- We may want to obtain some estimates for use in designing a future survival study.

A simple start: Exponential Regression

- **Observed data:** $(X_i, \delta_i, \mathbf{Z}_i)$ for individual i,
  $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, ..., Z_{ip})$ represents a set of p covariates.
- **Right censoring:** Assume that $X_i = \min(T_i, U_i)$
- **Survival distribution:** Assume T_i follows an exponential distribution with a parameter λ that depends on $\mathbf{Z}_i$, say $\lambda_i = \Psi(\mathbf{Z}_i)$. Then we can write:

$$T_i \sim exponential(\Psi(\mathbf{Z}_i))$$

Review

First, let's review some facts about the exponential distribution (from our first survival lecture):

$$
\begin{aligned}
f(t) &= \lambda e^{-\lambda t} && \text{for } t \geq 0 \\
S(t) &= P(T \geq t) = \int_t^{\infty} f(u)du = e^{-\lambda t} \\
F(t) &= P(T < t) = 1 - e^{-\lambda t} \\
\lambda(t) &= \frac{f(t)}{S(t)} = \lambda && \text{constant hazard!} \\
\Lambda(t) &= \int_0^t \lambda(u)\, du = \int_0^t \lambda\, du = \lambda t
\end{aligned}
$$

Modeling the hazard in exponential regression

Now, we say that λ is a constant *over time* t, but we want to let it depend on the covariate values, so we are setting

$$\lambda_i \quad = \quad \Psi(\mathbf{Z}_i)$$

The hazard rate would therefore be the same for any two individuals with the same covariate values.
Although there are many possible choices for Ψ, one simple and natural choice is:

$$\Psi(\mathbf{Z}_i) = \exp[\beta_0 + Z_{i1}\beta_1 + Z_{i2}\beta_2 + ... + Z_{ip}\beta_p]$$

WHY?

- ensures a positive hazard
- for an individual with $\mathbf{Z} = \mathbf{0}$, the hazard is e^{β_0}.

The model is called **exponential regression** because of the natural generalization from regular linear regression

Exponential regression for the 2-sample case

- Assume we have only a single covariate $\mathbf{Z} = Z$, i.e., $(p = 1)$.
  Hazard Rate:

$$\Psi(\mathbf{Z}_i) = \exp(\beta_0 + Z_i\beta_1)$$

- Define:
  $Z_i = 0$ if individual i is in group 0
  $Z_i = 1$ if individual i is in group 1
- **What is the hazard for group 0?**
- **What is the hazard for group 1?**
- **What is the hazard ratio of group 1 to group 0?**
- **What is the interpretation of β_1?**

Likelihood for Exponential Model

Under the assumption of right censored data, each person has one of two possible contributions to the likelihood:

(a) they have an **event** at X_i $(\delta_i = 1) \Rightarrow$ contribution is

$$L_i = \underbrace{S(X_i)}_{\text{survive to } X_i} \cdot \underbrace{\lambda(X_i)}_{\text{fail at } X_i} = e^{-\lambda X_i}\ \lambda$$

(b) they are **censored** at X_i $(\delta_i = 0) \Rightarrow$ contribution is

$$L_i = \underbrace{S(X_i)}_{\text{survive to } X_i} = e^{-\lambda X_i}$$

The likelihood for the exponential model (cont'd)

The **likelihood** is the product over all of the individuals:

$$
\begin{aligned}
\mathcal{L} &= \prod_i L_i \\
&= \prod_i \underbrace{\left(\lambda e^{-\lambda x_i}\right)^{\delta_i}}_{\text{events}} \underbrace{\left(e^{-\lambda x_i}\right)^{(1-\delta_i)}}_{\text{censorings}} \\
&= \prod_i \lambda^{\delta_i} \left(e^{-\lambda x_i}\right)
\end{aligned}
$$

Maximum Likelihood for Exponential

How do we use the likelihood?

- first take the log
- then take the partial derivative with respect to β
- then set to zero and solve for $\widehat{\beta}$
- this gives us the **maximum likelihood estimators**

Likelihood equations

The log-likelihood is:

$$\begin{aligned}
\log \mathcal{L} &= \log \left[\prod_i \lambda^{\delta_i} \quad \left(e^{-\lambda X_i} \right) \right] \\
&= \sum_i \left[\delta_i \log(\lambda) - \lambda X_i \right] \\
&= \sum_i \left[\delta_i \log(\lambda) \right] - \sum_i \lambda X_i
\end{aligned}$$

For the case of exponential regression, we now substitute the hazard $\lambda = \Psi(\mathbf{Z}_i)$ in the above log-likelihood:

$$\log \mathcal{L} = \sum_i \left[\delta_i \log(\Psi(\mathbf{Z}_i)) \right] - \sum_i \Psi(\mathbf{Z}_i) X_i \tag{1}$$

General Form of Log-likelihood for Right Censored Data

In general, whenever we have right censored data, the likelihood and corresponding log likelihood will have the following forms:

$$\begin{aligned} \mathcal{L} &= \prod_i [\lambda_i(X_i)]^{\delta_i} \; S_i(X_i) \\ \log \mathcal{L} &= \sum_i [\delta_i \log (\lambda_i(X_i))] - \sum_i \Lambda_i(X_i) \end{aligned}$$

where

- $\lambda_i(X_i)$ is the hazard for the individual i who fails at X_i
- $\Lambda_i(X_i)$ is the cumulative hazard for an individual at their failure or censoring time

For example, see the derivation of the likelihood for a Cox model on p.11-18 of Lecture 4 notes. We started with the likelihood above, then substituted the specific forms for $\lambda(X_i)$ under the PH assumption.

Consider our model for the hazard rate:

$$\lambda = \Psi(\mathbf{Z}_i) = \exp[\beta_0 + Z_{i1}\beta_1 + Z_{i2}\beta_2 + ... + Z_{ip}\beta_p]$$

We can write this using vector notation, as follows:

$$\begin{aligned} \text{Let} \quad \mathbf{Z}_i &= (1, Z_{i1}, ...Z_{ip})^T \\ \text{and} \quad \boldsymbol{\beta} &= (\beta_0, \beta_1, ...\beta_p) \end{aligned}$$

(Since β_0 is the intercept (i.e., the log hazard rate for the baseline group), we put a "1" as the first term in the vector $\mathbf{Z}_i$.) Then, we can write the hazard as:

$$\Psi(\mathbf{Z}_i) = \exp[\boldsymbol{\beta}\mathbf{Z}_i]$$

Now we can substitute $\Psi(\mathbf{Z}_i) = \exp[\boldsymbol{\beta}\mathbf{Z}_i]$ in the log-likelihood shown in (1):

$$\log \mathcal{L} = \sum_{i=1}^{n} \delta_i(\boldsymbol{\beta}\mathbf{Z}_i) - \sum_{i=1}^{n} X_i \exp(\boldsymbol{\beta}\mathbf{Z}_i)$$

Score Equations

Taking the derivative with respect to β_0, the score equation is:

$$\frac{\partial \log \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^{n} [\delta_i - X_i \exp(\boldsymbol{\beta} \mathbf{Z}_i)]$$

For β_k, $k = 1, \ldots p$, the equations are:

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial \beta_k} &= \sum_{i=1}^{n} [\delta_i Z_{ik} - X_i Z_{ik} \exp(\boldsymbol{\beta} \mathbf{Z}_i)] \\ &= \sum_{i=1}^{n} Z_{ik} [\delta_i - X_i \exp(\boldsymbol{\beta} \mathbf{Z}_i)] \end{aligned}$$

To find the MLE's, we set the above equations to 0 and solve (simultaneously). The equations above imply that the MLE's are obtained by setting the weighted number of failures ($\sum_i Z_{ik} \delta_i$) equal to the weighted cumulative hazard ($\sum_i Z_{ik} \Lambda(X_i)$).

Variance of the MLE

To find the variance of the MLE's, we need to take the second derivatives:

$$-\frac{\partial^2 \log \mathcal{L}}{\partial \beta_k \partial \beta_j} = \sum_{i=1}^{n} Z_{ik} Z_{ij} X_i \exp(\boldsymbol{\beta} \mathbf{Z}_i)$$

Some algebra (see Cox and Oakes section 6.2) reveals that

$$Var(\widehat{\boldsymbol{\beta}}) = I(\boldsymbol{\beta})^{-1} = \left[\mathbf{Z}(\mathbf{I} - \Pi)\mathbf{Z}^T\right]^{-1}$$

where

- $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ is a $(p+1) \times n$ matrix (p covariates plus the "1" for the intercept β_0)
- $\Pi = diag(\pi_1, \ldots, \pi_n)$ (this means that Π is a diagonal matrix, with the terms $\pi_1, \ldots, \pi_n$ on the diagonal)
- π_i is the probability that the i-th person is censored, so $(1 - \pi_i)$ is the probability that they failed.
- **Note:** The information $I(\boldsymbol{\beta})$ (inverse of the variance) is proportional to the number of failures, not the sample size. This will be important when we talk about study design.

The Single Sample Problem ($Z_i = 1$ for everyone)

First, what is the MLE of β_0?

We set $\frac{\partial \log \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^{n} [\delta_i - X_i \exp(\beta_0 Z_i)]$ equal to 0 and solve:

$$\begin{aligned} \Rightarrow \quad \sum_{i=1}^{n} \delta_i &= \sum_{i=1}^{n} [X_i \exp(\beta_0)] \\ d &= \exp(\beta_0) \sum_{i=1}^{n} X_i \\ \exp(\widehat{\beta_0}) &= \frac{d}{\sum_{i=1}^{n} X_i} \\ \hat{\lambda} &= \frac{d}{t} \end{aligned}$$

where d is the total number of deaths (or events), and $t = \sum X_i$ is the total person-time contributed by all individuals.

MLE estmate for β

If d/t is the MLE for λ, what does this imply about the MLE of β_0?

Using the previous formula $Var(\hat{\beta}) = \left[\mathbf{Z}(\mathbf{I} - \Pi)\mathbf{Z}^T\right]^{-1}$,
what is the variance of $\widehat{\beta}_0$?:

With some matrix algebra, you can show that it is:

$$Var(\widehat{\beta}_0) = \frac{1}{\sum_{i=1}^{n}(1-\pi_i)} = \frac{1}{d}$$

What about $\hat{\lambda} = e^{\hat{\beta}_0}$?

By the delta method,

$$\begin{aligned} Var(\hat{\lambda}) &= \hat{\lambda}^2 \; Var(\widehat{\beta}_0) \\ &= ? \end{aligned}$$

The Two-Sample Problem:

	Z_i	Subjects	Events	Follow-up
Group 0:	$Z_i = 0$	n_0	d_0	$t_0 = \sum_{i=1}^{n_0} X_i$
Group 1:	$Z_i = 1$	n_1	d_1	$t_1 = \sum_{i=1}^{n_1} X_i$

The log-likelihood

$$\log \mathcal{L} = \sum_{i=1}^{n} \delta_i(\beta_0 + \beta_1 Z_i) - \sum_{i=1}^{n} X_i \exp(\beta_0 + \beta_1 Z_i)$$

$$\text{so} \quad \frac{\partial \log \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^{n} [\delta_i - X_i \exp(\beta_0 + \beta_1 Z_i)]$$

$$= (d_0 + d_1) - (t_0 e^{\beta_0} + t_1 e^{\beta_0 + \beta_1})$$

$$\frac{\partial \log \mathcal{L}}{\partial \beta_1} = \sum_{i=1}^{n} Z_i [\delta_i - X_i \exp(\beta_0 + \beta_1 Z_i)]$$

$$= d_1 - t_1 e^{\beta_0 + \beta_1}$$

This implies:

$$\hat{\lambda}_1 = e^{\hat{\beta}_0+\hat{\beta}_1} = ?$$

$$\hat{\lambda}_0 = e^{\hat{\beta}_0} = ?$$

$$\hat{\beta}_0 = ?$$

$$\hat{\beta}_1 = ?$$

The maximum likelihood estimates (MLE's) of the hazard rates under the exponential model are the number of events divided by the person-years of follow-up!

(this result will be relied on heavily when we discuss study design)

Regression: Means and Medians

Mean Survival Time

For the exponential distribution, $E(T) = 1/\lambda$.

- **Control Group:**

$$\overline{T}_0 \;=\; 1/\hat{\lambda}_0 \;=\; 1/\exp(\hat{\beta}_0)$$

- **Treatment Group:**

$$\overline{T}_1 \;=\; 1/\hat{\lambda}_1 \;=\; 1/\exp(\hat{\beta}_0 + \hat{\beta}_1)$$

Means and medians (cont'd)

Median Survival Time

This is the value M at which $S(t) = e^{-\lambda t} = 0.5$, so $M = \text{median} = \frac{-\log(0.5)}{\lambda}$

- **Control Group:**

$$\hat{M}_0 = \frac{-\log(0.5)}{\hat{\lambda}_0} = \frac{-\log(0.5)}{\exp(\hat{\beta}_0)}$$

- **Treatment Group:**

$$\hat{M}_1 = \frac{-\log(0.5)}{\hat{\lambda}_1} = \frac{-\log(0.5)}{\exp(\hat{\beta}_0 + \hat{\beta}_1)}$$

Exponential Regression: Variance Estimates and Test Statistics

We can also calculate the variances of the MLE's as simple functions of the number of failures:

$$var(\hat{\beta}_0) = \frac{1}{d_0}$$

$$var(\hat{\beta}_1) = \frac{1}{d_0} + \frac{1}{d_1}$$

Inference

So our test statistics are formed as:

For testing $H_o : \beta_0 = 0$:

$$\begin{aligned} \chi^2_w &= \frac{\left(\hat{\beta}_0\right)^2}{var(\hat{\beta}_0)} \\ &= \frac{[\log(d_0/t_0)]^2}{1/d_0} \end{aligned}$$

For testing $H_o : \beta_1 = 0$:

$$\begin{aligned} \chi^2_w &= \frac{\left(\hat{\beta}_1\right)^2}{var(\hat{\beta}_1)} \\ &= \frac{\left[\log(\frac{d_1/t_1}{d_0/t_0})\right]^2}{\frac{1}{d_0} + \frac{1}{d_1}} \end{aligned}$$

How would we form confidence intervals for the hazard ratio?

The Likelihood Ratio test statistic

This is an alternative to the Wald test. It is based on 2 times the log of the ratio of the likelihoods under the null and alternative. We reject H_0 if 2 log(LR) $> \chi^2_{1,0.05}$, where

$$LR = \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)} = \frac{\mathcal{L}(\widehat{\lambda}_0, \widehat{\lambda}_1)}{\mathcal{L}(\widehat{\lambda})}$$

The Likelihood Ratio test statistic (cont'd)

For a sample of n independent exponential random variables with parameter λ, the Likelihood is:

$$\begin{aligned} L &= \prod_{i=1}^{n} [\lambda^{\delta_i} \exp(-\lambda x_i)] \\ &= \lambda^d \exp(-\lambda \sum x_i) \\ &= \lambda^d \exp(-\lambda n \bar{x}) \end{aligned}$$

where d is the number of deaths or failures. The log-likelihood is

$$\ell = d \log(\lambda) - \lambda n \bar{x}$$

and the MLE is

$$\widehat{\lambda} = d/(n\bar{x})$$

2-Sample Case: LR test calculations

Data:

Group 0: d_0 failures among the n_0 females
mean failure time is $\bar{x}_0 = (\sum_i^{n_0} X_i)/n_0$

Group 1: d_1 failures among the n_1 males
mean failure time is $\bar{x}_1 = (\sum_i^{n_1} X_i)/n_1$

Under the **alternative hypothesis**:

$$\begin{aligned} \mathcal{L} &= \lambda_1^{d_1} \exp(-\lambda_1 n_1 \bar{x}_1) \times \lambda_0^{d_0} \exp(-\lambda_0 n_0 \bar{x}_0) \\ \log(\mathcal{L}) &= d_1 \log(\lambda_1) - \lambda_1 n_1 \bar{x}_1 + d_0 \log(\lambda_0) - \lambda_0 n_0 \bar{x}_0 \end{aligned}$$

The MLE's are:

$$\begin{aligned} \widehat{\lambda}_1 &= d_1/(n_1 \bar{x}_1) \qquad \text{for males} \\ \widehat{\lambda}_0 &= d_0/(n_0 \bar{x}_0) \qquad \text{for females} \end{aligned}$$

MLEs under the null hypothesis

$$\begin{aligned} \mathcal{L} &= \lambda^{d_1+d_0} \exp[-\lambda(n_1\bar{x}_1 + n_0\bar{x}_0)] \\ \log(\mathcal{L}) &= (d_1+d_0)\log(\lambda) - \lambda[n_1\bar{x}_1 + n_0\bar{x}_0] \end{aligned}$$

The corresponding MLE is

$$\widehat{\lambda} = (d_1+d_0)/[n_1\bar{x}_1 + n_0\bar{x}_0]$$

Constructing the LR test

A likelihood ratio test can be constructed by taking twice the difference of the log-likelihoods under the alternative and the null hypotheses:

$$-2\left[(d_0+d_1)\log\left(\frac{d_0+d_1}{t_0+t_1}\right)-d_1\log[d_1/t_1]-d_0\log[d_0/t_0]\right]$$

Exponential Regression in R

```
Call:
survreg(formula = Surv(losyr, fail) ~ gender, data = nurshome,
    dist = "exp")
             Value Std. Error     z        p
(Intercept) -0.0578     0.0333 -1.73 8.28e-02
gender      -0.5162     0.0619 -8.34 7.62e-17

Scale fixed at 1

Exponential distribution
Loglik(model)= -1006.3   Loglik(intercept only)= -1038.4
Chisq= 64.2 on 1 degrees of freedom, p= 1.1e-15
Number of Newton-Raphson Iterations: 5
n= 1591
```

Since $Z = 8.337$, the chi-square test is $Z^2 = 69.51$.

The Weibull regression model

At the beginning of the course, we saw that the survivorship function for a Weibull random variable is:

$$S(t) \quad = \quad \exp[-\lambda(t^{\kappa})]$$

and the hazard function is:

$$\lambda(t) \quad = \quad \kappa\,\lambda\,t^{(\kappa-1)}$$

The Weibull regression model assumes that for someone with covariates $\mathbf{Z}_i$, the survivorship function is

$$S(t;\mathbf{Z}_i) = \exp[-\Psi(\mathbf{Z}_i)(t^{\kappa})]$$

where $\Psi(\mathbf{Z}_i)$ is defined as in exponential regression to be:

$$\Psi(\mathbf{Z}_i) = \exp[\beta_0 + Z_{i1}\beta_1 + Z_{i2}\beta_2 + \ldots Z_{ip}\beta_p]$$

For the 2-sample problem, we have:

$$\Psi(\mathbf{Z}_i) = \exp[\beta_0 + Z_{i1}\beta_1]$$

Weibull MLEs for the 2-sample problem:

Log-likelihood:

$$\log \mathcal{L} = \sum_{i=1}^{n} \delta_i \log \left[\kappa \exp(\beta_0 + \beta_1 Z_i) X_i^{\kappa-1}\right] - \sum_{i=1}^{n} X_i^{\kappa} \exp(\beta_0 + \beta_1 Z_i)$$

$$\Rightarrow \quad \exp(\hat{\beta}_0) = d_0/t_0\kappa \qquad \exp(\hat{\beta}_0 + \hat{\beta}_1) = d_1/t_1\kappa$$

where

$$t_{j\kappa} = \sum_{i=1}^{n_j} X_i^{\hat{\kappa}} \text{ among } n_j \text{ subjects}$$

$$\hat{\lambda}_0(t) = \hat{\kappa}\ \exp(\hat{\beta}_0)\ t^{\hat{\kappa}-1} \quad \hat{\lambda}_1(t) = \hat{\kappa}\ \exp(\hat{\beta}_0 + \hat{\beta}_1)\ t^{\hat{\kappa}-1}$$

$$\begin{aligned} \widehat{HR} &= \hat{\lambda}_1(t)/\hat{\lambda}_0(t) = \exp(\hat{\beta}_1) \\ &= \exp\left(\frac{d_1/t_1\kappa}{d_0/t_0\kappa}\right) \end{aligned}$$

Weibull Regression: Means and Medians

Mean Survival Time

For the Weibull distribution, $E(T) = \lambda^{(-1/\kappa)}\Gamma[(1/\kappa) + 1]$.

- **Control Group:**

$$\overline{T}_0 = \hat{\lambda}_0^{(-1/\hat{\kappa})}\ \Gamma[(1/\hat{\kappa}) + 1]$$

- **Treatment Group:**

$$\overline{T}_1 = \hat{\lambda}_1^{(-1/\hat{\kappa})}\ \Gamma[(1/\hat{\kappa}) + 1]$$

Median Survival Time

For the Weibull distribution, $M = \text{median} = \left[\frac{-\log(0.5)}{\lambda}\right]^{1/\kappa}$

- **Control Group:**

$$\hat{M}_0 = \left[\frac{-\log(0.5)}{\hat{\lambda}_0}\right]^{1/\hat{\kappa}}$$

- **Treatment Group:**

$$\hat{M}_1 = \left[\frac{-\log(0.5)}{\hat{\lambda}_1}\right]^{1/\hat{\kappa}}$$

where $\hat{\lambda}_0 = \exp(\hat{\beta}_0)$ and $\hat{\lambda}_1 = \exp(\hat{\beta}_0 + \hat{\beta}_1)$.

The Gamma function

Note: the symbol Γ is the "gamma" function. If x is an integer, then

$$\Gamma(x) \quad = \quad (x-1)!$$

The Weibull regression model is very easy to fit:

- In STATA: Just specify `dist(weibull)` instead of `dist(exp)` within the `streg` command
- In R: we use the `survreg` command with the `dist="exp"` option.

Fitting the Weibull model in R

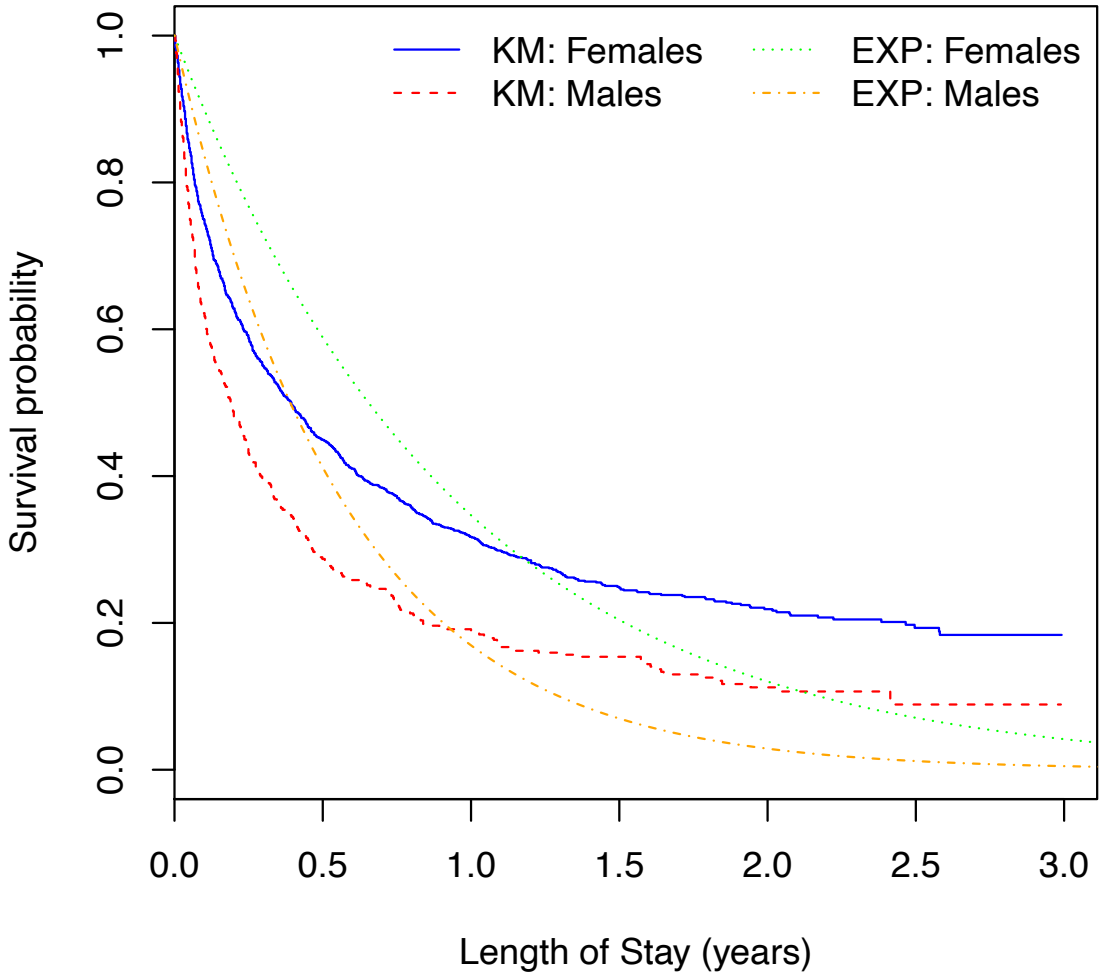
```
Call:
survreg(formula = Surv(losyr, fail) ~ gender, data = nurshome,
    dist = "weibull")
              Value Std. Error     z        p
(Intercept) -0.143     0.0542 -2.65 8.13e-03
gender      -0.673     0.1011 -6.66 2.67e-11
Log(scale)   0.487     0.0232 20.99 8.94e-98

Scale= 1.63

Weibull distribution
Loglik(model)= -731.1   Loglik(intercept only)= -751.9
Chisq= 41.73 on 1 degrees of freedom, p= 1e-10
Number of Newton-Raphson Iterations: 5
n= 1591
```

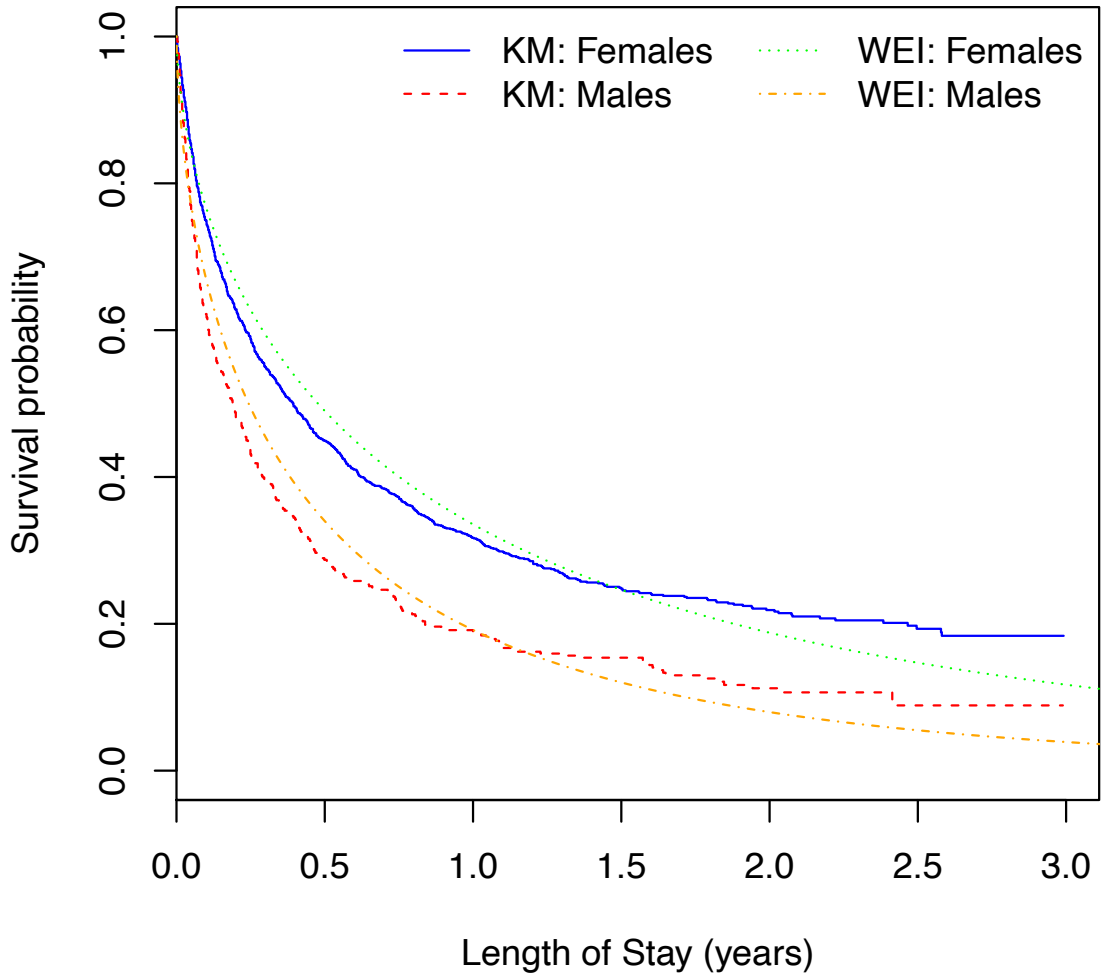
Comparison of Exponential with Kaplan-Meier

We can see how well the Exponential model fits by comparing the survival estimates for males and females under the exponential model, i.e., $P(T \geq t) = e^{(-\hat{\lambda}_z t)}$, to the Kaplan-Meier survival estimates:



Comparison of Weibull with Kaplan-Meier

We can see how well the Weibull model fits by comparing the survival estimates, $P(T \geq t) = e^{(-\hat{\lambda}_z t^{\hat{\kappa}})}$, to the Kaplan-Meier survival estimates.



Other useful plots for evaluating fit

- $-\log(\hat{S}(t))$ vs t
- $\log[-\log(\hat{S}(t))]$ vs $\log(t)$

Why are these useful?
If T is exponential, then $S(t) = \exp(-\lambda t))$

$$\begin{aligned} \text{so} \quad \log(S(t)) &= -\lambda t \\ \text{and} \quad \Lambda(t) &= \lambda\, t \end{aligned}$$

a straight line in t with slope λ and intercept=0

If T is Weibull, then $S(t) = \exp(-(\lambda t)^{\kappa})$

$$\begin{aligned} \text{so} \quad \log(S(t)) &= -\lambda t^{\kappa} \\ \text{then} \quad \Lambda(t) &= \lambda t^{\kappa} \\ \text{and} \quad \log(-\log(S(t))) &= \log(\lambda) + \kappa * \log(t) \end{aligned}$$

a straight line in $\log(t)$ with slope κ and intercept $\log(\lambda)$.

Goodness of fit plots

- We can calculate our estimated $\Lambda(t)$ and plot it versus t, and if it seems to form a straight line, then the exponential distribution is probably appropriate for our dataset.
- We can plot $\log \hat{\Lambda}(t)$ versus $\log(t)$, and if it seems to form a straight line, then the Weibull distribution is probably appropriate for our dataset.

Comparison of methods for the two-sample problem

Data:

	Z_i	Subjects	Events	Follow-up
Group 0:	$Z_i = 0$	n_0	d_0	$t_0 = \sum_{i=1}^{n_0} X_i$
Group 1:	$Z_i = 1$	n_1	d_1	$t_1 = \sum_{i=1}^{n_1} X_i$

In General:

$$\lambda_z(t) \quad = \quad \lambda(t, Z = z) \qquad \text{for } z = 0 \text{ or } 1.$$

The hazard rate depends on the value of the covariate Z. In this case, we are assuming that we only have a single covariate, and it is binary ($Z = 1$ or $Z = 0$)

Models

Exponential Regression:

$$\begin{aligned} \lambda_z(t) &= \exp(\beta_0 + \beta_1 Z) \\ \Rightarrow \lambda_0 &= \exp(\beta_0) \\ \lambda_1 &= \exp(\beta_0 + \beta_1) \\ HR &= \exp(\beta_1) \end{aligned}$$

Weibull Regression:

$$\begin{aligned} \lambda_z(t) &= \kappa \ \exp(\beta_0 + \beta_1 Z) \ t^{\kappa-1} \\ \Rightarrow \lambda_0 &= \kappa \ \exp(\beta_0) \ t^{\kappa-1} \\ \lambda_1 &= \kappa \ \exp(\beta_0 + \beta_1) \ t^{\kappa-1} \\ HR &= \exp(\beta_1) \end{aligned}$$

Models (cont'd)

Proportional Hazards Model:

$$\lambda_z(t) = \lambda_0(t)\exp(\beta_1)$$

$$\Rightarrow \lambda_0 = \lambda_0(t) \qquad \text{KM?}$$

$$\lambda_1 = \lambda_0(t)\exp(\beta_1)$$

$$HR = \exp(\beta_1)$$

Remarks

We make the following remarks:

- Exponential model is a special case of the Weibull model with $\kappa = 1$ (note: Collett uses γ instead of κ)
- Exponential and Weibull models are both special cases of the Cox PH model.
- If either the exponential model or the Weibull model is valid, then these models will tend to be more efficient than PH (smaller s.e.'s of estimates). This is because they assume a particular form for $\lambda_0(t)$, rather than estimating it at every death time.

The Accelerated Failure Time Model

The general form of an accelerated failure time (AFT) model is:

$$\log(T_i) = \boldsymbol{\beta}_{AFT}\mathbf{Z}_i + \sigma\epsilon$$

where

- $\log(T_i)$ is the log of a survival time
- $\boldsymbol{\beta}_{AFT}$ is the vector of AFT model parameters corresponding to the covariate vector $\mathbf{Z}_i$
- ϵ is a random "error" term
- σ is a scale factor

In other words, we can model the log-survival times as a linear function of the covariates!

By choosing different distributions for ϵ, we can obtain different parametric distributions:

- Exponential
- Weibull
- Gamma
- Log-logistic
- Normal
- Lognormal

We can compare the predicted survival under any of these parametric distributions to the KM estimated survival to see which one seems to fit best.
Once we decide on a certain class of model (say, Gamma), we can evaluate the contributions of covariates by finding the MLE's, and constructing Wald, Score, or LR tests of the covariate effects.

We can motivate the AFT model by first demonstrating the following two relationships:

1. For the Exponential Model:

If the failure times $T_i = T(\mathbf{Z}_i)$ follow an exponential distribution, i.e., $S_i(t) = e^{-\lambda_i t}$ with $\lambda_i = exp(\beta \mathbf{Z}_i)$, then

$$\log(T_i) \quad = \quad -\beta \mathbf{Z}_i + \epsilon$$

where ϵ follows an extreme value distribution (which just means that e^{ϵ} follows a unit exponential distribution).

2. For the Weibull Model:

If the failure times $T_i = T(\mathbf{Z}_i)$ follow a Weibull distribution, i.e., $S_i(t) = e^{\lambda_i t^{\kappa}}$ with $\lambda_i = exp(\beta \mathbf{Z}_i)$, then

$$\log(T_i) \quad = \quad -\sigma\beta\mathbf{Z}_i + \sigma\epsilon$$

where ϵ again follows an extreme value distribution, and $\sigma = 1/\kappa$. In other words, both the Exponential and Weibull model can be written in the form of a log-linear model for the survival times, if we choose the right distribution for ϵ.

The log-linear form for the exponential can be derived by:

(1) Creating a new variable $T_0 = T_Z \times \exp(\beta \mathbf{Z}_i)$

(2) Taking the log of T_Z, yielding $\log(T_Z) = \log\left(\frac{T_0}{\exp(\beta \mathbf{Z}_i)}\right)$

Step (1): For an exponential model, recall that:

$$S_i(t) = Pr(T_Z \geq t) = e^{-\lambda t}, \quad \text{with } \lambda = \exp(\beta \mathbf{Z}_i)$$

It follows that $T_0 \sim exp(1)$:

$$\begin{aligned} S_0(t) = Pr(T_0 \geq t) &= Pr(T_Z \cdot \exp(\beta \mathbf{Z}) \geq t) \\ &= Pr(T_Z \geq t \exp(-\beta \mathbf{Z})) \\ &= \exp\left[-\lambda\ t\ \exp(-\beta \mathbf{Z})\right] \\ &= \exp\left[-\exp(\beta \mathbf{Z})\ t\ \exp(-\beta \mathbf{Z})\right] \\ &= \exp(-t) \end{aligned}$$

Step (2): Now take the log of the survival time:

$$\begin{aligned} \log(T_Z) &= \log\left(\frac{T_0}{\exp(\beta \mathbf{Z}_i)}\right) \\ &= \log(T_0) - \log\left(\exp(\beta \mathbf{Z}_i)\right) \end{aligned}$$

Relationship between Exponential and Weibull

If T_Z has a Weibull distribution, i.e., $S(t) = e^{-\lambda t^{\kappa}}$
with $\lambda = \exp(\boldsymbol{\beta}\mathbf{Z}_i)$, then you can show that the new variable

$$T_Z^* = T_Z^{\kappa}$$

follows an exponential distribution with parameter $\exp(\boldsymbol{\beta}\mathbf{Z}_i)$. Based on the previous page, we can therefore write:

$$\log(T^*) = -\boldsymbol{\beta}\mathbf{Z} + \epsilon$$

(where ϵ has an extreme value distribution.)
But since $\log(T^*) = \log(T^{\kappa}) = \kappa \times \log(T)$, we can write:

$$\begin{aligned} \log(T) &= \log(T^*)/\kappa \\ &= (1/\kappa)(-\boldsymbol{\beta}\mathbf{Z}_i + \epsilon) \\ &= -\sigma\boldsymbol{\beta}\mathbf{Z}_i + \sigma\epsilon \end{aligned}$$

where $\sigma = 1/\kappa$.

This motivates the following general definition of the **Accelerated Failure Time Model** by:

$$\log(T_i) = \boldsymbol{\beta}_{AFT}\mathbf{Z}_i + \sigma\epsilon$$

where ϵ is a random "error" term, σ is a scale factor, Y is the log of a survival random variable, and

$$\boldsymbol{\beta}_{AFT} = -\sigma\boldsymbol{\beta}_e$$

where $\boldsymbol{\beta}_e$ came from the hazard $\lambda = \exp(\boldsymbol{\beta}\mathbf{Z})$.

The defining feature of an AFT model is:

$$S(t; \mathbf{Z}) = S_i(t) = S_0(\phi\ t)$$

That is, the effect of covariates is to accelerate (stretch) or decelerate (shrink) the time-scale.

Effect of AFT on hazard:

$$\lambda_i(t) \quad = \quad \phi\ \lambda_0(\phi t)$$

One way to interpret the AFT model is via its effect on median survival times. If $S_i(t) = 0.5$, then $S_0(\phi t) = 0.5$. This means:

$$M_i = \phi M_0$$

Interpretation:

- For $\phi < 1$, there is an acceleration of the endpoint (if $M_0 = 2$yrs in control and $\phi = 0.5$, then $M_i = 1$yr.
- For $\phi > 1$, there is a stretching or delay in endpoint
- In general, the lifetime of individual i is ϕ times what they would have experienced in the reference group

Since ϕ must be positive and a function of the covariates, we model $\phi = \exp(\boldsymbol{\beta}\mathbf{Z}_i)$.

When does Proportional hazards = AFT?

According to the proportional hazards model:

$$S(t) = S_0(t)^{\exp(\boldsymbol{\beta}\mathbf{Z}_i)}$$

and according to the accelerated failure time model:

$$S(t) = S_0(t \exp(\boldsymbol{\beta}\mathbf{Z}_i))$$

Say $T_i \sim Weibull(\lambda, \kappa)$. Then $\lambda(t) = \lambda\kappa t^{(\kappa-1)}$
Under the AFT model:

$$\begin{aligned}
\lambda_i(t) &= \phi\ \lambda_0(\phi t) \\
&= e^{\boldsymbol{\beta}\mathbf{Z}_i}\ \lambda_0(e^{\boldsymbol{\beta}\mathbf{Z}_i} t) \\
&= e^{\boldsymbol{\beta}\mathbf{Z}_i}\ \lambda_0\kappa \left(e^{\boldsymbol{\beta}\mathbf{Z}_i} t\right)^{(\kappa-1)} \\
&= \left(e^{\boldsymbol{\beta}\mathbf{Z}_i}\right)^{\kappa}\ \lambda_0\kappa t^{(\kappa-1)} \\
&= \left(e^{\boldsymbol{\beta}\mathbf{Z}_i}\right)^{\kappa}\ \lambda_0(t)
\end{aligned}$$

Special cases of AFT models

- Exponential regression: $\sigma = 1$, ϵ following the extreme value distribution.
- Weibull regression: σ arbitrary, ϵ following the extreme value distribution.
- Lognormal regression: σ arbitrary, ϵ following the normal distribution.