Survival analysis notes

Rob Trangucci

March 10, 2025

Contents

1 Notation **4**

2 Introduction **5**

2.1 Independent censoring . . . 6
2.2 Mean time to failure . . . 6
2.3 Survival function . . . 7
2.3.1 Properties of the survival function . . . 7
2.4 Hazard function . . . 9
2.4.1 Properties of the hazard function . . . 10
2.5 Density function for survival time . . . 10
2.6 Cumulative hazard function . . . 10
2.7 Discrete survival time . . . 11
2.7.1 Connection between discrete and continuous survival functions . . . 12
2.8 Mean residual lifetime . . . 13
2.9 Examples . . . 14

3 Censoring and truncation **16**

3.1 Right censoring . . . 16
3.1.1 Type I censoring . . . 17
3.1.2 Generalized type I censoring . . . 17
3.1.3 Type II censoring . . . 19
3.1.4 Generalized Type II censoring . . . 19
3.1.5 Independent censoring . . . 19
3.2 Noninformative censoring . . . 19
3.2.1 Reasons for informative censoring . . . 22
3.3 Truncation . . . 22
3.4 Likelihood construction . . . 24

4 Nonparametric estimator of survival function **27**
4.1 Derivation of Nelson-Aalen and Kaplan-Meier estimators . . . 27
4.1.1 Kaplan-Meier estimator standard error . . . 30
4.2 Confidence intervals . . . 32
4.2.1 Handling ties in the Nelson-Aalen estimator . . . 34
4.2.2 Handling ties in the Kaplan-Meier estimator . . . 34
4.3 Nonparametric tests . . . 34
4.4 More on log-rank tests . . . 38

5 Parametric and nonparametric regression models **40**
5.1 Asymptotic interlude . . . 45
5.1.1 Asymptotic confidence intervals . . . 48
5.1.2 Asymptotic tests . . . 49
5.1.3 Tests in terms of observed information . . . 51
5.1.4 Composite tests . . . 52
5.2 More on parametric regression models . . . 56
5.3 Weibull regression . . . 56
5.3.1 Parametric proportional hazards models . . . 57
5.3.2 Testing for proportional hazards . . . 58
5.3.3 Accelerated failure time formulation . . . 59
5.4 AFT models . . . 60
5.4.1 Model checking in AFT models . . . 61
5.4.2 Cox-Snell residuals . . . 62
5.4.3 Influence of data points in likelihood equations . . . 63
5.5 Cox Proportional Hazards Model . . . 71
5.5.1 Cox model likelihood derivation and the Breslow estimator without ties 71
5.5.2 Alternative view of the Cox model . . . 73
5.5.3 Interpretation of the Cox regression model . . . 75
5.5.4 Score function of the Cox model . . . 76
5.5.5 Model checking in the Cox model . . . 78
5.5.6 Stratified Cox model . . . 84
5.5.7 Including time dependent variables in the Cox model . . . 85
5.5.8 Solving the problem with time-varying covariates . . . 87
5.6 Frailty and unobserved variation . . . 88
5.6.1 Marginal survival and hazard functions . . . 89
5.6.2 Comparison between two risk groups with frailty . . . 92

5.6.3 False protectivity 94
5.6.4 Frailty and influence functions 96
5.6.5 Cox PH with omitted variables 98

6 Appendix 99

6.1 Map between Weibull parameterizations 99
6.2 Summary of tests 100
6.2.1 Log-rank test 100
6.2.2 Likelihood based tests 101
6.2.3 Comments and 103

Chapter 1

Notation

Notation	Description
C_i	Random variable representing the time to censoring
$T_i = \min(X_i, C_i)$	Observable event time
$\delta_i = \mathbb{1}(T_i = X_i)$	Indicator variable equal to one if event time is a failure time
$P_\theta(X \le t)$	Distribution function of X indexed by parameters θ
$S_X(t;\theta)$	Survival function for random variable X evaluated at t, parameters θ
$f_X(t;\theta)$	Density function for random variable X evaluated at t, parameters θ
$\lambda_X(t;\theta)$	Hazard function for random variable X evaluated at t, parameters θ

Table 1.1: List of notation used throughout the notes

Chapter 2

Introduction

This introduction is based in part on Klein, Moeschberger, et al. 2003, and in part on O. Aalen et al. 2008 plus Fleming and Harrington 2005.

Survival analysis is the modeling and analysis of time-to-event data; this means we will be studying how to model **nonnegative** random variables (time will always be measured in such a way so that the observations are nonnegative). Think about a clinical trial for a new COVID vaccine and how you might model the length of time between study entry and infection in each arm of the trial. Let X_i be the time from trial entry to infection for the i-th participant. These sorts of trials are typically run until a prespecified number of people have become infected. Let n be the total number of participants in the trial and let r be the prespecified number of infections. Let T_i be the observed infection time for the i-th participant. This means that for r participants, $T_i = X_i$, but for $n - r$ participants we know only that the time-to-infection is larger than the observed time. Let C_i denote the time from study entry for participant i to study end. Then $T_i = \min(X_i, C_i)$, and let $\delta_i = \mathbb{1}(T_i = X_i)$. The density of T_i is related to the joint probability for X_i and C_i, which is indexed by a possibly infinite dimensional parameter θ: $P_\theta(X_i > t, C_i > c)$. When $\delta_i = 1$, and $T_i = X_i$, the likelihood of the observation is

$$\left(-\frac{\partial}{\partial u} P_\theta(X_i > u, C_i > t)\right)\bigg|_{u=t},$$

while the likelihood for $\delta_i = 0$ is

$$\left(-\frac{\partial}{\partial u} P_\theta(X_i > t, C_i > u)\right)\bigg|_{u=t},$$

Then $T_i = C_i$ for the other $n - r$ participants. Under the null hypothesis that the vaccine has no effect, the population distribution function for all n participants for X_i, C_i is $P_\theta(X_1 > x, C_1 > c)$ (i.e. the distribution for survival times in the treatment group and the placebo group is the same). Then the joint density for the observed infection times is as follows:

$$f_{T_1,\dots,T_n}(t_1,\dots,t_n;\theta) = n!\prod_{i=1}^{r}\left(-\frac{\partial}{\partial u}P_\theta(X_1 > u, C_1 > t_{(i)})\right)\Big|_{u=t_{(i)}}\prod_{i=r+1}^{n}\left(\left(-\frac{\partial}{\partial u}P_\theta(X_1 > t_{(i)}, C_1 > u)\right)\Big|_{u=t_{(i)}}\right),$$

where $t_{(i)}$ is the i-th order statistic of the set $\{t_1,\dots,t_n\}$. Note that this is different from most other data analysis where missing observations are not expected to occur with much frequency. On the contrary, in survival analysis, missingness, both *truncation* and *censoring* are expected to occur with nearly every dataset, so much of our time will be spent ensuring our methods work when data arise with these peculiarities.

2.1 Independent censoring

Now suppose that $X_1 \perp\!\!\!\perp C_1$, and that θ partitions into η and ϕ, such that

$$P_\theta(X_1 > x, C_1 > c) = P_\eta(X_1 > x)P_\phi(C_1 > c).$$

Then we can rewrite the joint observational density for T_i as:

$$\begin{aligned} f_{T_1,\dots,T_n}(t_1,\dots,t_n;\theta) &= n!\left(\prod_{i=1}^{r} f_{X_1}(t_{(i)};\eta)\right)\prod_{i=r+1}^{n}P_\eta(X_1 > t_{(i)}) \\ &\quad\times\left(\prod_{i=1}^{r}P_\phi(C_1 > t_{(i)})\right)\prod_{i=r+1}^{n}f_C(t_{(i)};\phi). \end{aligned}$$

If we are only interested about inference about η, the parameters that govern the distribution of the true time-to-infection random variables, we can ignore the the distribution for the censoring random variables C_1, and maximize the likelihood because, in η:

$$f_{T_1,\dots,T_n}(t_1,\dots,t_n;\eta) \propto \left(\prod_{i=1}^{r} f_{X_1}(t_{(i)};\eta)\right)\prod_{i=r+1}^{n}P_\eta(X_1 > t_{(i)})$$

We will talk in more detail about censoring in the coming lectures.

2.2 Mean time to failure

O. Aalen et al. 2008 notes that we cannot even compute a simple mean in this situation, so something like a t-test will be useless. As an aside, let's try to compute a mean from the data above. Let $\bar{T} = \frac{1}{n}\sum_{i=1}^{n}T_i$. We can show that $\lim_{n\to\infty}\bar{T} \le \mathbb{E}[X_i]$ with probability 1.

Proof. Let $T_i = X_i\mathbb{1}(X_i \le C_i) + C_i\mathbb{1}(X_i > C_i)$. Then by the SLLN $\bar{T} \overset{\text{a.s.}}{\to} \mathbb{E}[T_i]$.

$$\begin{aligned} \mathbb{E}[T_i] &= \mathbb{E}[X_i\mathbb{1}(X_i \le C_i)] + \mathbb{E}[C_i\mathbb{1}(X_i > C_i)] \\ &\le \mathbb{E}[X_i\mathbb{1}(X_i \le C_i)] + \mathbb{E}[X_i\mathbb{1}(X_i > C_i)] = \mathbb{E}[X_i] \end{aligned}$$

□

2.3 Survival function

How can we compute the mean time to infection then? One way to estimate the mean time to infection is to first estimate the function $S_{X_i}(t;\theta) = P_\theta(X_i > t)$, which is also known as the *survival function*. Recall this fact about non-negative random variables $X_i \geq 0$ w.p. 1:

$$\mathbb{E}[X_i] = \int_0^\infty P_\theta(X_i > t)dt$$

This follows from an application of Fubini's theorem applied to the integral:

$$\begin{aligned}
\mathbb{E}[X_i] &= \int_0^\infty u dP_{X_i}(u;\theta) \\
&= \int_0^\infty \int_0^\infty \mathbb{1}(0 \leq t \leq u)\, dt\, dP_{X_i}(u;\theta) \\
&= \int_0^\infty \int_0^\infty \mathbb{1}(0 \leq t \leq u)\, dP_{X_i}(u;\theta) dt \\
&= \int_0^\infty P_\theta(X_i > t)dt
\end{aligned}$$

2.3.1 Properties of the survival function

Let $F_{X_i}(t;\theta) = P_\theta(X_i \leq t)$. Then because the survival function is defined as $S_{X_i}(t;\theta) = 1 - F_{X_i}(t;\theta)$ (also known as the complementary CDF) the survival function inherits its properties from the CDF. The survival function:

1. $S_{X_i}(t;\theta)$ is a nonincreasing function
2. $S_{X_i}(0;\theta) = 1$
3. $\lim_{t\to\infty} S_{X_i}(t;\theta) = 0$
4. Has lefthand limits:
$$\lim_{s\nearrow t} S_{X_i}(s;\theta) = S_{X_i}(t-;\theta).$$
5. Is right continuous:
$$\lim_{s\searrow t} S_{X_i}(s;\theta) = S_{X_i}(t;\theta).$$

An example of a discrete survival function is shown in Figure 2.1.

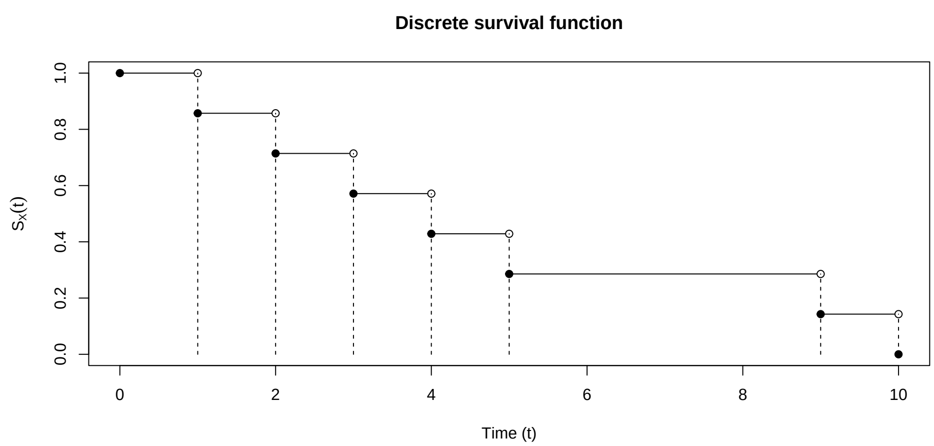

Figure 2.1: Example plot of a survival function for a discrete survival time, bounded between $[0, 10]$

2.4 Hazard function

Another way to characterize the random variable X_i is the *hazard function*, which is typically denoted as $\lambda(t)$ or $h(t)$ and is defined as

$$\begin{aligned}\lambda_{X_i}(t) &= \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \mathbb{P}_\theta\left(t \le X_i < t + \Delta t \mid X_i \ge t\right) \\ &= \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \frac{\mathbb{P}_\theta\left(t \le X_i < t + \Delta t\right)}{\mathbb{P}_\theta\left(X_i \ge t\right)}\end{aligned}$$

First, note that we can define $\mathbb{P}_\theta\left(X_i \ge t\right)$ in terms of the survival function as:

$$\mathbb{P}_\theta\left(X_i \ge t\right) = \lim_{s \nearrow t} S_{X_i}(s; \theta).$$

Using the notation introduced in Section 2.3.1, we can write this as

$$\mathbb{P}_\theta\left(X_i \ge t\right) = S_{X_i}(t-; \theta).$$

Of course, when X_i is absolutely continuous,$S_{X_i}(t-; \theta) = S_{X_i}(t; \theta)$, but when X_i is discrete, or mixed discrete and continuous, as noted above, it is not true in general that the survival function is left-continuous.

A few things to note about $\lambda_{X_i}(t; \theta)$: when X_i is an absolutely continuous random variable, which occurs when we're considering survival in continuous time, we can write this in terms of the probability density function $f_{X_i}(t; \theta)$ and the cumulative distribution function $F_{X_i}(t; \theta)$:

$$\begin{aligned}\lambda_{X_i}(t) &= \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \frac{\mathbb{P}_\theta\left(t \le X_i < t + \Delta t\right)}{\mathbb{P}_\theta\left(X_i \ge t\right)} \\ &= \lim_{\Delta t \searrow 0} \frac{F_{X_i}(t + \Delta t; \theta) - F_{X_i}(t; \theta)}{\Delta t} \times \frac{1}{1 - F_{X_i}(t; \theta)} \\ &= \frac{f_{X_i}(t; \theta)}{1 - F_{X_i}(t; \theta)}.\end{aligned}$$

Let's examine how the survival function and the hazard function fit together.

$$\lambda_{X_i}(t) = \frac{f_{X_i}(t; \theta)}{S_{X_i}(t-; \theta)}.$$

Note that we can write the hazard function in terms of the survival function instead of the density, when X_i is absolutely continuous:

$$\begin{aligned}\lambda_{X_i}(t) &= \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \frac{\mathbb{P}_\theta\left(t \le X_i < t + \Delta t\right)}{\mathbb{P}_\theta\left(X_i \ge t\right)} \\ &= \lim_{\Delta t \searrow 0} \frac{S_{X_i}(t; \theta) - S_{X_i}(t + \Delta t; \theta)}{\Delta t} \times \frac{1}{S_{X_i}(t; \theta)} \\ &= -\frac{d}{dt} S_{X_i}(t; \theta) / S_{X_i}(t; \theta).\end{aligned}$$

This implies that

$$\lambda_{X_i}(t) = -\frac{d}{dt}\log S_{X_i}(t;\theta).$$

If we integrate both sides, we get another important identity in survival analysis:

$$\int_0^u \frac{d}{dt}\log S_{X_i}(t;\theta)dt = -\int_0^u \lambda_{X_i}(t)dt \tag{2.1}$$

$$\log S_{X_i}(u;\theta) - \log S_{X_i}(0;\theta) = -\int_0^u \lambda_{X_i}(t)dt \quad \text{note } S_{X_i}(0;\theta) = 1 \tag{2.2}$$

$$S_{X_i}(u;\theta) = \exp\left(-\int_0^u \lambda_{X_i}(t)dt\right) \tag{2.3}$$

2.4.1 Properties of the hazard function

The relationship $S_{X_i}(u;\theta) = \exp\left(-\int_0^u \lambda_{X_i}(t)dt\right)$ and the properties of the survival function reveal the following facts about the hazard function and highlight its differences with a probability density.

1. $\lim_{t\to\infty} S_{X_i}(t;\theta) = 0$ implies that $\lim_{t\to\infty}\int_0^t \lambda_X(u)du = \infty$
2. Given that $S_{X_i}(t;\theta)$ is a nonincreasing function, $\lambda_X(t) \geq 0$ for all t.

So unlike a probability density function, $\lambda_X(t)$ isn't integrable over the support of the random variable.

2.5 Density function for survival time

Given that we have $S_{X_i}(t;\theta)$ and $\lambda(t) = \frac{f_{X_i}(t;\theta)}{S_{X_i}(t-;\theta)}$, we can recover the density, $f_{X_i}(t;\theta)$ easily:

$$f_{X_i}(t;\theta) = \lambda_{X_i}(t)S_{X_i}(t-;\theta)$$

2.6 Cumulative hazard function

One final important quantity that describes a survival distribution is that of *cumulative hazard*, which we'll denote as $\Lambda_{X_i}(t)$, though it is also denoted as $H(t)$ in Klein, Moeschberger, et al. 2003. This is defined as you might expect:

$$\Lambda_{X_i}(t) = \int_0^t \lambda_{X_i}(u)du.$$

It has the important property that for any absolutely continuous failure time X_i with a given cumulative hazard function, the random variable $Y_i = \Lambda_{X_i}(X_i)$ is exponentially distributed with rate 1. The derivation is straightforward. Remember that $P(X_i > t) = \exp\left(-\Lambda_{X_i}(t)\right)$

$$
\begin{aligned}
P(\Lambda_{X_i}(X_i) > t) &= P(X_i > \Lambda_{X_i}^{-1}(t)) \\
&= \exp\left(-\Lambda_{X_i}(\Lambda_{X_i}^{-1}(t))\right) \\
&= \exp\left(-t\right)
\end{aligned}
$$

2.7 Discrete survival time

We've been working with continuous survival times until now. If X_i is a discrete random variable with support on $\{t_1, t_2, \dots\}$, we lose some of the tidyness of the previous derivations. We can define the distribution of X_i in terms of the survival function, $P_\theta(X_i > t)$. First let $p_j = P_\theta(X_i = t_j)$, so

$$
S_{X_i}(t; \theta) = P_\theta(X_i > t) = \sum_{j | t_j > t} p_j
$$

We can also define the hazard function for a discrete random variable:

$$
\lambda_{X_i}(t_j) = \frac{p_j}{S_{X_i}(t_{j-1}; \theta)} = \frac{p_j}{p_j + p_{j+1} + \dots}
$$

Note that $p_j = S_{X_i}(t_{j-1}; \theta) - S_{X_i}(t_j; \theta)$, then

$$
\lambda_{X_i}(t_j) = 1 - \frac{S_{X_i}(t_j; \theta)}{S_{X_i}(t_{j-1}; \theta)}.
$$

If we let $t_0 = 0$ then $S_{X_i}(t_0; \theta) = 1$. This allows us to write the survival function in a sort of telescoping product:

$$
\begin{aligned}
P_\theta(X_i > t_j) &= P_\theta(X_i > t_0) \frac{P_\theta(X_i > t_1)}{P_\theta(X_i > t_0)} \frac{P_\theta(X_i > t_2)}{P_\theta(X_i > t_1)} \dots \frac{P_\theta(X_i > t_j)}{P_\theta(X_i > t_{j-1})} \\
&= 1 \frac{S_{X_i}(t_1; \theta)}{S_{X_i}(t_0; \theta)} \frac{S_{X_i}(t_2; \theta)}{S_{X_i}(t_1; \theta)} \dots \frac{S_{X_i}(t_j; \theta)}{S_{X_i}(t_{j-1}; \theta)}
\end{aligned}
$$

This yields another way to write $S_{X_i}(t; \theta)$:

$$
S_{X_i}(t; \theta) = \prod_{j | t_j \leq t} (1 - \lambda_{X_i}(t_j)). \tag{2.4}
$$

It turns out that we can write the survival function for continuous random variables in the same way.

2.7.1 Connection between discrete and continuous survival functions

Recall the definition of the hazard function:

$$\lambda_{X_i}(t) = \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \mathbb{P}_\theta \left(t \leq X < t + \Delta t \mid X \geq t \right)$$

Note that $\lambda_{X_i}(t)\, \Delta t$ is approximately $\mathbb{P}_\theta \left(t \leq X < t + \Delta t \mid X \geq t \right)$. Let $\mathcal{T}$ be a partition of $(0, \infty)$ with partition size Δt, $t_0 = 0$:

$$\mathcal{T} = \bigcup_{j=0}^{\infty} [t_j, t_j + \Delta t).$$

Then we can use Equation (2.4) to represent the survival function:

$$S_{X_i}(t; \theta) = \prod_{j \mid t_j + \Delta t \leq t} (1 - \lambda_{X_i}(t_j)\Delta t). \tag{2.5}$$

We can show that as the partition of the time domain gets finer and finer, we will recover $S_{X_i}(t; \theta) = \exp(-\int_0^t \lambda_{X_i}(u) du)$

$$S_{X_i}(t; \theta) = \prod_{j \in \mathcal{T} \mid t_j + \Delta t \leq t} (1 - \lambda_{X_i}(t_j)\Delta t) \tag{2.6}$$

$$\log S_{X_i}(t; \theta) = \sum_{j \in \mathcal{T} \mid t_j + \Delta t \leq t} \log(1 - \lambda_{X_i}(t_j)\Delta t) \tag{2.7}$$

We use the Taylor expansion of $\log(1 - \lambda_{X_i}(t_j)\Delta t)$ for small $\lambda_{X_i}(t_j)\Delta t$, assuming that $\lambda_{X_i}(t)$ is sufficiently well-behaved for all t.

$$\log(1 - \lambda_{X_i}(t_j)\Delta t) \approxeq -\lambda_{X_i}(t_j)\Delta t.$$

Then

$$\log S_{X_i}(t; \theta) \approxeq \sum_{j \in \mathcal{T} \mid t_j + \Delta t \leq t} -\lambda_{X_i}(t_j)\Delta t \tag{2.8}$$

As

$$\lim_{\Delta t \searrow 0} \sum_{j \in \mathcal{T} \mid t_j + \Delta t \leq t} -\lambda_{X_i}(t_j)\Delta t = -\int_0^t \lambda_{X_i}(u) du.$$

So, $S_{X_i}(t; \theta) = \exp(-\int_0^t \lambda_{X_i}(u) du)$, or

$$S_{X_i}(t; \theta) = \exp(-\lambda_{X_i}(t)) \tag{2.9}$$

2.8 Mean residual lifetime

We also might be interested in the *mean residual lifetime* (mrl for short), or the expected lifetime given survival up to a certain point:

$$\mathbb{E}\left[X_i - x \mid X_i > x\right].$$

We can compute this for an absolutely continuous random variable by using the survival function:

$$\frac{\int_x^\infty (u-x) f_{X_i}(u;\eta) du}{S_{X_i}(x;\eta)} = \frac{\int_x^\infty S_{X_i}(u;\eta) du}{S_{X_i}(x;\eta)}$$

To derive the mrl in terms of the survival function, note that we can use Fubini again on the numerator (Exercise 1), or we can use integration by parts:

$$\begin{aligned}\int_x^\infty (u-x) f_{X_i}(u) du &= -\int_x^\infty (u-x) \frac{d}{du} S_{X_i}(u) du \\ &= -(u-x) S_{X_i}(u)\Big|_{u=x}^{\infty} + \int_x^\infty S_{X_i}(u) du\end{aligned}$$

and use the fact that $\lim_{u\to\infty} S_{X_i}(u) = 0$. We also need the following:

$$\lim_{u\to\infty} uP(X_i > u) = 0. \tag{2.10}$$

This is a pretty weak condition, random variables with second moments satisfy this condition (Exercise 2), as do random variables with only first moments. It turns out that under this condition we'll have a weak law of large numbers (see §7.1 in Resnick 2019).

Suppose we assume that $\mathbb{E}[X] \leq \infty$. Then we can write:

$$\mathbb{E}[X] = \mathbb{E}[X \mathbb{1}(X \leq n)] + \mathbb{E}[X \mathbb{1}(X > n)]$$

Note that if we define $X_n = X\mathbb{1}(X \leq n)$ then

$$X_1(\omega) \leq X_2(\omega) \leq \cdots \leq X_k(\omega) \leq \ldots.$$

By the Monotone Convergence Theorem (MCT), $\mathbb{E}[X_n] \to \mathbb{E}[X]$. Then

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X_n] + \mathbb{E}[X\mathbb{1}(X > n)] \\ &\geq \mathbb{E}[X_n] + \mathbb{E}[n\mathbb{1}(X > n)] \\ &= \mathbb{E}[X_n] + nP(X > n)\end{aligned}$$

This leads to the system of inequqalities:

$$\mathbb{E}[X] - \mathbb{E}[X_n] \geq nP(X > n) \geq 0.$$

By the MCT $\mathbb{E}[X] - \mathbb{E}[X_n] \to 0$ so

$$\lim_{n\to\infty} nP(X_i > n) = 0.$$

However, there are random variables for which $\mathbb{E}[X_i]$ does not exist, but do satisfy Equation (2.10) (see the end of §7.1 in Resnick 2019).

2.9 Examples

The first example we'll run through is for an exponentially distributed survival time:

$$X_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda).$$

The survival function is $S_X(t) = e^{-\lambda t}$. We can read off from this that $\Lambda(t) = \lambda t$. What's the hazard function? Let's plot the hazard function. What does this imply about the exponential distribution (memorylessness)? The mean lifetime is $\frac{1}{\lambda}$. The mean residual lifetime is:

$$\begin{aligned}\frac{\int_t^\infty e^{-\lambda u} du}{e^{-\lambda t}} &= \frac{1}{\lambda}\frac{e^{-\lambda t} du}{e^{-\lambda t}} \\ &= \frac{1}{\lambda}.\end{aligned}$$

This is a consequence of the memoryless property of the exponential distribution.

Another parametric distribution for survival times is the Weibull.

$$X_i \overset{\text{iid}}{\sim} \text{Weibull}(\gamma, \alpha).$$

The survival function:

$$S_X(t) = \exp(-\gamma t^\alpha).$$

Again, we have that $\Lambda(t) = \gamma t^\alpha$, so we can take the derivative with respect to t to get the hazard:

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

This is more flexible than the exponential distribution, though note that for $\alpha = 1$, $X_i \sim \text{Exponential}(\gamma)$, so the Weibull family contains the exponential family as a special case. The α parameter allows for the hazard rate to have more flexibility than the exponential. If $\alpha > 1$, the hazard rate is increasing in t. This corresponds to an aging process, whereby the longer something has survived, the higher the rate of failure. If $\alpha < 1$, the hazard rate is decreasing in t. This might correspond to something like the hazard for SIDS, which is quite high for children before 1 year old, but drops off rapidly after 1. Let's compute the mean lifetime, $\mathbb{E}[X] = \int_0^\infty S_X(t)dt$, using a v-sub, $v = t^\alpha$, so $v^{\frac{1}{\alpha}} = t \to \frac{1}{\alpha}v^{\frac{1}{\alpha}-1}dv = dt$:

$$\begin{aligned}\int_0^\infty \exp(-\gamma t^\alpha)dt &= \frac{1}{\alpha}\int_0^\infty v^{\frac{1}{\alpha}-1}\exp(-\gamma v)dv \\ &= \frac{1}{\alpha}\frac{1}{\gamma^{\frac{1}{\alpha}}}\Gamma(\frac{1}{\alpha}) \\ &= \frac{\Gamma(\frac{1}{\alpha}+1)}{\gamma^{\frac{1}{\alpha}}}\end{aligned}$$

The mean residual lifetime is a bit more involved. Let $v = \gamma u^\alpha$ so $\left(\frac{v}{\gamma}\right)^{1/\alpha} = u \to \gamma^{-1/\alpha}\frac{1}{\alpha}v^{\frac{1}{\alpha}-1}dv = du$:

$$\begin{aligned}\int_t^\infty \exp(-\gamma u^\alpha)du &= \gamma^{-1/\alpha}\frac{1}{\alpha}\int_{\gamma t^\alpha}^\infty v^{\frac{1}{\alpha}-1}\exp(-v)dv \\ &= \gamma^{-1/\alpha}\frac{1}{\alpha}\Gamma(\frac{1}{\alpha}, \gamma t^\alpha),\end{aligned}$$

where $\Gamma(\frac{1}{\alpha}, \gamma t^\alpha)$ is the upper incomplete Gamma function.

Chapter 3

Censoring and truncation

Now let's delve into more detail about censoring, and how the likelihood can be built up from the hazard function and the survival function. Klein, Moeschberger, et al. 2003 define censoring as imprecise knowledge about an event time. If we observe a failure or an event exactly, the observation is not censored, but if we know only that an observation occurred within a range of values, we say the observation is censored. Let X_i, as usual, be our failure time, which is not completely observed. Instead if:

- $X_i \in [U, \infty)$, the observation is *right censored*
- $X_i \in [0, V)$, the observation is *left censored*
- $X_i \in [U, V)$, the observation is *interval censored*

3.1 Right censoring

Right censoring occurs when a survival time is known to be larger than a given value. This is the most common censoring scenario in survival analysis.

Recall our definition in Chapter 2:

- Let X_i be the time to failure, or time to event for individual i.
- Let C_i be the time to censoring. It may be helpful to think about C_i as the time to investigator measurement.
- Let $\delta_i = \mathbb{1}(X_i \leq C_i)$.
- Let $T_i = \min(X_i, C_i)$.

Given our definitions in Section 3.1, when an observation is censored, or when a measurement is taken of the survival time before the event has happened, $\delta_i = 0$ and $T_i = C_i$.

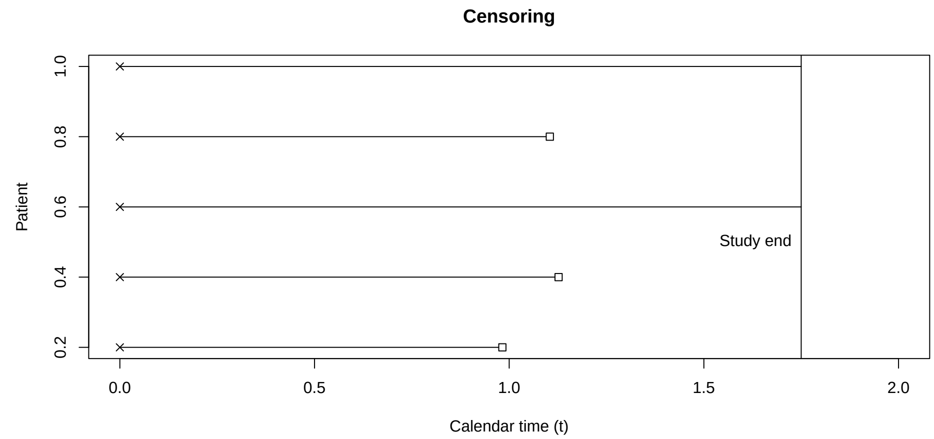

Figure 3.1: Example of Type I censoring.

3.1.1 Type I censoring

The simplest censoring scenario is one in which all individuals have the same, nonrandom censoring time. Imagine a study is designed to follow 5 startups that are spun out of a tech incubator to study how long it takes a company to land its first contract. This information will be used for designing investments 2 years from the study date, so the study has a length of 1.75 years. We can say that all observations will have to have occurred, or not, by 1.75 years.

Figure 3.1 shows a potential result of the study, where 2 out of the 5 companies have not landed a contract. In this case,

- For all individuals such that $\delta_i = 0 \implies X_i > C$
- $\delta_i = 1 \implies T_i = X_i$.

3.1.2 Generalized type I censoring

A more general scenario, which is closer to most examples in clinical trials, is when each individual has a different study entry time and the investigator has a preset study end time. This is called generalized Type I censoring. These study entry times are typically assumed to be independent of the survival time. This is shown in Figure 3.2. When study entry is independent from survival time, the analysis proceeds as shown in Figure 3.3. For generalized type I censoring,

- For all individuals such that $\delta_i = 0 \implies X_i > C_i$

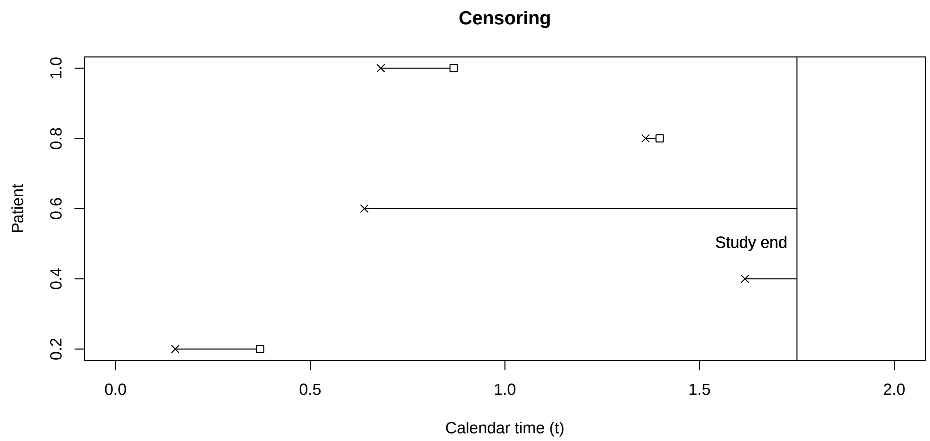

Figure 3.2: Example of generalized Type I censoring, where each individual has a separate study entry time.

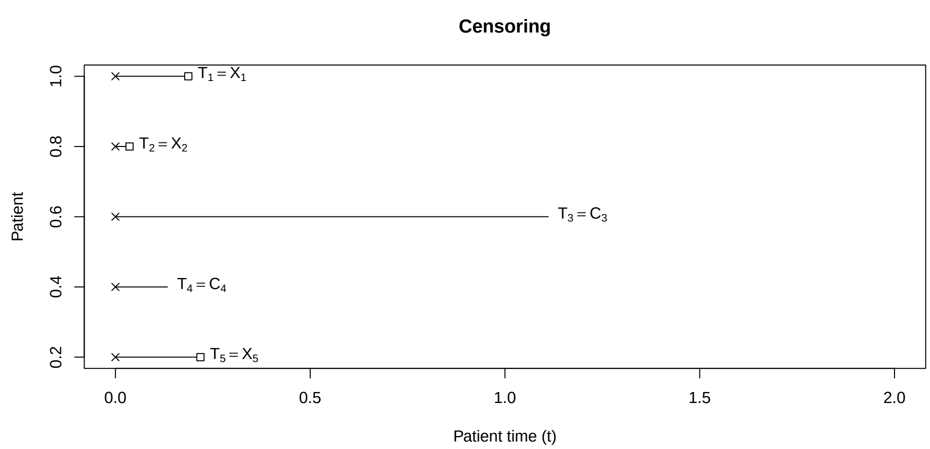

Figure 3.3: Example of generalized Type I censoring, viewed in patient time.

- $\delta_i = 1 \implies T_i = X_i$.

This is different from Type I censoring in that each individual has a different censoring time.

3.1.3 Type II censoring

Type II censoring occurs when all units have the same study entry time, but researchers design the study to end when $r < n$ units fail out of n total units under observation.

- For the first r, lucky or unlucky participants, $\delta_i = 1 \implies T_i = X_{(i)}$ or the i^{th} order statistic.
- For the remaining $n - r$ individuals, $\delta_i = 0 \implies X_i > X_{(r)}$.

3.1.4 Generalized Type II censoring

You may be wondering, what happens when units have differing start times but we want to end the trial after the r-th failure? It turns out that this was not a solved problem until Rühl et al. 2023, which was quite surprising to me.

3.1.5 Independent censoring

A third type of censoring, helpfully called independent censoring, takes $X_i \perp\!\!\!\perp C_i$, and thus conclusions similar to those of generalized type I censoring can be drawn:

- For all individuals such that $\delta_i = 0 \implies X_i > C_i$
- $\delta_i = 1 \implies T_i = X_i$.

3.2 Noninformative censoring

All of the previous censoring scenarios can be summarized as noninformative censoring. Let the parameters indexing the censoring distribution be ϕ, while the parameters indexing the failure time distribution are η. Noninformative censoring is defined as the following equality:

$$\lambda_{X_i}(t) = \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \mathbb{P}_{\eta,\phi} \left(t \leq X_i < t + \Delta t \mid X_i \geq t, C_i \geq t \right) \tag{3.1}$$

Note that this implies the following:

$$\mathbb{P}_{\eta} \left(t \leq X_i < t + \Delta t \mid X_i \geq t \right) = \mathbb{P}_{\eta,\phi} \left(t \leq X_i < t + \Delta t \mid X_i \geq t, C_i \geq t \right). \tag{3.2}$$

which is equivalent to writing that failure

For independent censoring, Equation (3.2) holds, given that $P_{\eta,\phi}(X_i > t, C_i > c) = P_\eta(X_i > t)P_\phi(C_i > c)$ and that η and ϕ are *variationally independent*.

This means that the parameter space $\Omega_{\eta,\phi}$ is the Cartesian product of the parameter space for η and ϕ.

Definition 3.2.1. Variational independence Let $\eta \in \Omega_\eta$ and $\phi \in \Omega_\phi$. The joint space is denoted as $\Omega_{\eta,\phi}$. If $\Omega_{\eta,\phi} = \Omega_\eta \times \Omega_\phi$, η and ϕ are variationally independent. In other words, the range for η does not change given a value for ϕ.

Under independent censoring, the observable hazard for uncensored failure times is as follows:

$$\frac{-\frac{\partial}{\partial u} P_{\eta,\phi}(X_i > u, C_i > t-) \mid_{u=t}}{P_{\eta,\phi}(X_i > t-, C_i > t-)} = \frac{-\frac{d}{du} S_{X_i}(u; \eta)}{S_{X_i}(t-; \eta)} \tag{3.3}$$

Here's an example that demonstrates the nonidentifiability of the joint distribution for censoring and failure time:

Example 3.2.1. Dependent failure and censoring
Let $P_{\theta,\alpha,\mu}(X_i > x, C_i > c) = \exp(-\alpha x - \mu c - \theta x c)$. We can find the marginal survival functions just by evaluating $P_{\theta,\alpha,\mu}(X_i > x, C_i > 0)$ and vice-versa, which yields:

$$P_\alpha(X_i > x) = \exp(-\alpha x)$$
$$P_\mu(C_i > c) = \exp(-\mu c)$$

Both of these distributions have constant hazards. However, the observable hazard is the following:

$$\frac{-\frac{\partial}{\partial u} P_{\theta,\alpha,\mu}(X_i > u, C_i > t-) \mid_{u=t}}{P_{\theta,\alpha,\mu}(X_i > t-, C_i > t-)} = \alpha + \theta t$$
$$\frac{-\frac{\partial}{\partial u} P_{\theta,\alpha,\mu}(X_i > t-, C_i > u) \mid_{u=t}}{P_{\theta,\alpha,\mu}(X_i > t-, C_i > t-)} = \mu + \theta t$$

This leads to an observable survival function:

$$S_{X_i}(x; \alpha, \theta) = \exp(-\alpha x - \theta x^2/2)$$
$$S_{C_i}(c; \mu, \theta) = \exp(-\mu c - \theta c^2/2)$$

If we mistakenly assume that the failure time and the censoring time are independent we'll get the following joint distribution:

$$S_{X_i}(x; \alpha, \theta) S_{C_i}(c; \mu, \theta) \neq \exp(-\alpha x - \mu c - \theta x c).$$

However, if we calculate the true observable survival function $P_{\theta,\alpha,\mu}(X_i > x, C_i > X_i-)$ we get:

$$\int_x^\infty -\frac{\partial}{\partial u} P_{\theta,\alpha,\mu}(X_i > u, C_i > t-) \mid_{u=t} dt = \int_x^\infty (\alpha + \theta t) \exp(-\alpha t - \mu t - \theta t^2) dt$$

while the observable survival function implied the by the erroneously assumed independent distributons is:

$$\begin{aligned}\int_x^\infty -\frac{\partial}{\partial u} S_{X_i}(u;\alpha,\theta) S_{C_i}(t-;\mu,\theta) \mid_{u=t} dt &= \int_x^\infty (-\frac{d}{dt} \exp(-\alpha t - \theta t^2/2)) \exp(-\mu t - \theta t^2/2) \\ &= \int_x^\infty (\alpha + \theta t) \exp(-\alpha t - \mu t - \theta t^2) du\end{aligned}$$

Thus, two different joint densities lead to the same observable survival functions, so the joint distribution is nonidentifiable.

Here is an example showing that we may have dependent censoring and failure times, but still end up with noninformative censoring:

Example 3.2.2. Dependent failure and censoring can be noninformative
Let Y_1, Y_2 and Y_{12} be exponentially distributed with rates $\alpha_1, \alpha_2, \alpha_{12}$, respectively. Let $X = Y_1 \wedge Y_{12}$ and $C = Y_2 \wedge Y_{12}$. The survival function $P_{\alpha_1,\alpha_2,\alpha_{12}}(X > x, C > c) = P(Y_1 > x, Y_2 > c, Y_{12} > x \vee c) = e^{-\alpha_1 x - \alpha_2 c - \alpha_{12} x \vee c}$ Then marginally X is exponential with rate $\alpha_1 + \alpha_{12}$, which is also equal to its hazard function. In order for noninformative censoring to hold, we need to check Equation (3.1), or that

$$\alpha_1 + \alpha_{12} = \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \mathbb{P}_{\alpha_1,\alpha_2,\alpha_{12}} \left(t \le X < t + \Delta t \mid X \ge t, C \ge t\right) \tag{3.4}$$

Because $t + \Delta t \vee t = t + \Delta t$ as $\Delta t > 0$,

$$\lim_{\Delta t \searrow 0} \frac{e^{-\alpha_1 t - \alpha_2 t - \alpha_{12} t} - e^{-(\alpha_1 + \alpha_{12})(t + \Delta t) - \alpha_2 t}}{\Delta t} \tag{3.5}$$

which just equals $e^{-\alpha t} - \frac{d}{ds} e^{-(\alpha_1 + \alpha_{12})s} \mid_{s=t}$ or $(\alpha_1 + \alpha_{12}) e^{-\alpha_1 t - \alpha_2 t - \alpha_{12} t}$. Then

$$\lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \mathbb{P}_{\alpha_1,\alpha_2,\alpha_{12}} \left(t \le X < t + \Delta t \mid X \ge t, C \ge t\right) = \frac{(\alpha_1 + \alpha_{12}) e^{-\alpha_1 t - \alpha_2 t - \alpha_{12} t}}{e^{-\alpha_1 t - \alpha_2 t - \alpha_{12} t}} \tag{3.6}$$

$$= \alpha_1 + \alpha_{12} \tag{3.7}$$

So in this case, while X and C are dependent, we still have noninformative censoring.

The benefit of noninformative censoring is that we can ignore the censoring random variables when constructing the likelihood for the survival random variables.

3.2.1 Reasons for informative censoring

A simple hypothetical situation with informative censoring would be one in which sick patients are lost to follow-up.

3.3 Truncation

While censoring can be seen as partial information about an observation, truncation deals with exact observations of selected units. The simplest example of truncation is when measurements are made using an instrument with a lower limit of detection. Imagine using a microscope to measure the diameter of cells on a plate that has a lower limit of detection of 5 microns. If interest lies in inferring the population mean diameter of the cells, one must take into account the fact that only cells with diameters of greater than 5 micros can be seen with the microscope.

Failure to take truncation into account can be a source of bias in inference.

$$\begin{aligned}
\mathbb{E}\left[X_i\right] &= \mathbb{E}\left[X_i \mid X_i \geq V\right] P(X_i \geq V) + \mathbb{E}\left[X_i \mid X_i < V\right] P(X_i < V) \\
&= \mathbb{E}\left[X_i \mid X_i \geq V\right] + P(X_i < V)(\mathbb{E}\left[X_i \mid X_i < V\right] - \mathbb{E}\left[X_i \mid X_i \geq V\right]) \\
&\leq \mathbb{E}\left[X_i \mid X_i \geq V\right]
\end{aligned}$$

The last line follows because $(\mathbb{E}\left[X_i \mid X_i < V\right] - \mathbb{E}\left[X_i \mid X_i \geq V\right]) \leq 0$. Using an estimator for $\mathbb{E}\left[X_i \mid X_i \geq V\right]$ when the target of inference in $\mathbb{E}\left[X_i\right]$ would result in positive bias. Of course, when the estimator instead estimates $\mathbb{E}\left[X_i \mid X_i < V\right]$ the bias would be negative. Depending on the value of V and the distribution of X_i, the bias can be severe.

For example, suppose a researcher is interested in learning about the impact of medication refills on the lifespans of patients. The researcher has access to a database in which they select patients who refilled medications at least once. The researcher subsequently selects a control group that is perfectly matched to the medication refill group, and upon analyzing the data, the analyst discovers that refilling prescription medication leads to longer lifespans. What is wrong with this analysis?

The observations in this example can be said to be left-truncated, because the researcher conditions the observations in the treatment group on having a lifespan long enough to fill a medication.

Formally, we say that the density for a truncated observation is conditioned on the probability of the observation lying in the truncated region.

- If a researcher selects $\mathbb{1}\left(X_i \geq V\right)$ we say the data are left-truncated, and $f_{X_i}(x;\eta) = \frac{-\frac{d}{dx} S_{X_i}(x;\eta)}{S_{X_i}(v;\eta)}$

- If a researcher selects $\mathbb{1}\left(X_i \leq U\right)$ we say the data are right-truncated, and $f_{X_i}(x;\,\eta) = \frac{-\frac{d}{dx}S_{X_i}(x;\eta)}{1-S_{X_i}(u;\eta)}$
- If a researcher selects $\mathbb{1}\left(V \leq X_i \leq U\right)$ we say the data are interval-truncated, and $f_{X_i}(x;\,\eta) = \frac{-\frac{d}{dx}S_{X_i}(x;\eta)}{S_{X_i}(v;\eta)-S_{X_i}(u;\eta)}$

3.4 Likelihood construction

We now turn to how to construct likelihoods in each of the prior scenarios, under censored or truncated data. As a reminder:

- Let X_i be the time to failure, or time to event for individual i.
- Let C_i be the time to censoring. It may be helpful to think about C_i as the time to investigator measurement.
- Let $\delta_i = \mathbb{1}(X_i \le C_i)$.
- Let $T_i = \min(X_i, C_i)$.

When $\delta_i = 1$, we observe $T_i = X_i$; this is the event that $\{X_i = T_i, C_i \ge X_i\}$. When $\delta_i = 0$, we observe $T_i = C_i$; this is the event that $\{C_i = T_i, C_i < X_i\}$. Let the joint distribution of X_i, C_i be written as $P_\theta(X > x, C > c)$, and further let $\theta = (\eta, \phi)$ such that $P_\theta(X > x, C > c) = P_\eta(X > x)P_\phi(C > c \mid X > x)$. We showed in Chapter 2 that the likelihood corresponding to the random variables T_i, δ_i, $f_{T_i,\delta_i}(t, \delta; \theta)$, can be written in terms of partial derivatives of the joint density function when X_i and C_i are absolutely continuous random variables.

$$f_{T_i,\delta_i}(t,\delta;\theta) = \left(-\frac{\partial}{\partial u}P_\theta(X \ge u, C \ge t)\,\big|_{u=t}\right)^{\delta}\left(-\frac{\partial}{\partial u}P_\theta(X \ge t, C \ge u)\,\big|_{u=t}\right)^{1-\delta}$$

Let's rewrite the partial derivatives in terms of their limits:

$$f_{T_i,\delta_i}(t,\delta;\theta) = \left(\lim_{\Delta t \searrow 0}\frac{1}{\Delta t}P_\theta(t \le X < t+\Delta t, C \ge t)\right)^{\delta}\left(\lim_{\Delta t \searrow 0}\frac{1}{\Delta t}P_\theta(X \ge t, t \le C < t+\Delta t)\right)^{1-\delta}$$

We can factorize the distribution function:

$$\begin{aligned} f_{T_i,\delta_i}(t,\delta;\theta) = &\left(\lim_{\Delta t \searrow 0}\frac{1}{\Delta t}P_\eta(t \le X < t+\Delta t \mid X \ge t)P_\theta(C \ge t \mid t \le X < t+\Delta t)P_\eta(X \ge t)\right)^{\delta} \\ &\times\left(\lim_{\Delta t \searrow 0}\frac{1}{\Delta t}P_\phi(t \le C < t+\Delta t \mid X \ge t)P_\eta(X \ge t)\right)^{1-\delta} \end{aligned}$$

and rearranging and subbing in $\lambda_\eta(t)$ for the hazard function:

$$f_{T_i,\delta_i}(t,\delta;\theta) = (\lambda_\eta(t))^{\delta}\,P_\eta(X \ge t)P_\theta(C \ge t \mid t \le X < t+\Delta t)^{\delta}\left(\lim_{\Delta t \searrow 0}\frac{1}{\Delta t}P_\theta(t \le C < t+\Delta t \mid X \ge t)\right)^{1-\delta}$$

Assuming that $P_\theta(C_i \ge t \mid X_i = x) = P_\phi(C_i \ge t)$ leads to the following

$$f_{T_i,\delta_i}(t,\delta;\theta) = (\lambda_\eta(t))^{\delta}\,P_\eta(X \ge t)P_\phi(C \ge t)^{\delta}\left(\lim_{\Delta t \searrow 0}\frac{1}{\Delta t}P_\phi(t \le C < t+\Delta t)\right)^{1-\delta}$$

This means that we can factorize the joint density:

$$f_{T_i,\delta_i}(t,\delta;\theta) = f_{X_i,\delta_i}(t,\delta;\eta) f_{C_i,\delta_i}(t,\delta;\phi).$$

Thus, noninformative censoring and parameter separability yield a separable joint density. This means that when we want to do maximum likelihood for survival data, we can *ignore* the model for the censoring times, $f_{C_i,\delta_i}(t,\delta;\phi)$, and focus on only the model for the failure times:

$$f_{X_i,\delta_i}(t,\delta;\eta) = \lambda_{X_i}(t)^{\delta} P_{\eta}(X \geq t).$$

We can write this expression fully in terms of the hazard function by recalling Equation (2.3):

$$f_{X_i,\delta_i}(t,\delta;\eta) = \lambda_{X_i}(t)^{\delta} \exp\left(-\int_0^t \lambda_{X_i}(u)du\right). \tag{3.8}$$

Example 3.4.1. MLE for exponential survival time Let $X_i \overset{\text{iid}}{\sim} \text{Exp}(\alpha)$ and assume we have independent censoring ($X_i \perp\!\!\!\perp C_i$), the parameters for the censoring process are separable from α, and that C_i are iid such that $\mathbb{E}[C_i] < \infty$. Then our observed data are $T_i = \min(X_i, C_i)$ and $\delta_i = \mathbb{1}(X_i \leq C_i)$. According to Equation (4.1) we can write the likelihood as

$$\begin{aligned} f_{\alpha}(t_1,\ldots,t_n,\delta_1,\ldots,\delta_n) &= \prod_{i=1}^{n} \alpha^{\delta_i} \exp(-\textstyle\sum_{i=1}^{n} \int_0^{t_i} \alpha du) \\ &= \alpha^{\sum_{i=1}^{n} \delta_i} \exp(-\alpha \textstyle\sum_{i=1}^{n} t_i) \end{aligned}$$

The log-likelihood is

$$\log(f_{\alpha}(t_1,\ldots,t_n,\delta_1,\ldots,\delta_n)) = \log(\alpha)\sum_{i=1}^{n}\delta_i - \alpha\sum_{i=1}^{n} t_i$$

which has the maximizer

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n} t_i}.$$

Let's show that this converges a.s. to α as $n \to \infty$. We can rewrite $\frac{\sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n} t_i}$ as

$$\frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq C_i)}{\frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}(X_i \leq C_i) + C_i \mathbb{1}(X_i > C_i)}$$

The top and bottom expressions converge a.s. by Kolmogorov's Strong Law of Large Numbers to

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq C_i) \overset{\text{a.s.}}{\to} \mathbb{E}_{(X_i,C_i)}[\mathbb{1}(X_i \leq C_i)]$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}(X_i \leq C_i) + C_i \mathbb{1}(X_i > C_i) \overset{\text{a.s.}}{\to} \mathbb{E}_{(X_i,C_i)}[X_i \mathbb{1}(X_i \leq C_i) + C_i \mathbb{1}(X_i > C_i)]$$

We can evaluate the top expression using the tower property of expectation:

$$\begin{aligned}\mathbb{E}_{(X_i,C_i)}\left[\mathbb{1}\left(X_i \le C_i\right)\right] &= \mathbb{E}_{C_i}\left[\mathbb{E}_{X_i|C_i}\left[\mathbb{1}\left(X_i \le c\right) \mid C_i = c\right]\right] \\ &= \mathbb{E}_{C_i}\left[1 - e^{-\alpha C_i}\right]\end{aligned}$$

where the second line follows from the independent censoring condition. The bottom expression becomes:

$$\begin{aligned}\mathbb{E}_{(X_i,C_i)}\left[X_i\mathbb{1}\left(X_i \le C_i\right) + C_i\mathbb{1}\left(X_i > C_i\right)\right] &= \mathbb{E}_{C_i}\left[\mathbb{E}_{X_i|C_i}\left[X_i\mathbb{1}\left(X_i \le c\right) \mid C_i = c\right]\right] \\ &\quad + \mathbb{E}_{C_i}\left[\mathbb{E}_{X_i|C_i}\left[c\mathbb{1}\left(X_i > c\right) \mid C_i = c\right]\right] \\ &= \mathbb{E}_{C_i}\left[\frac{1}{\alpha}(1-(1+\alpha C_i)e^{-\alpha C_i})\right] + \mathbb{E}_{C_i}\left[C_i e^{-\alpha C_i}\right] \\ &= \frac{1}{\alpha}\mathbb{E}_{C_i}\left[1 - e^{-\alpha C_i}\right]\end{aligned}$$

Thus

$$\frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n t_i} \overset{\text{a.s.}}{\to} \alpha$$

To show that $\int_0^c x\alpha e^{-\alpha x}dx = \frac{1}{\alpha}(1-(1+\alpha c)e^{-\alpha c})$, we can use the trick of differentiating under the integral sign.

$$\begin{aligned}\alpha\int_0^c xe^{-\alpha x}dx &= \alpha\int_0^c -\frac{d}{d\alpha}e^{-\alpha x}dx \\ &= \alpha\left(-\frac{d}{d\alpha}\right)\int_0^c e^{-\alpha x}dx \\ &= \alpha\left(-\frac{d}{d\alpha}\right)\frac{1}{\alpha}(1-e^{-\alpha c}) \\ &= \alpha\left(\frac{1-(1+\alpha c)e^{-\alpha c}}{\alpha^2}\right)\end{aligned}$$

Chapter 4

Nonparametric estimator of survival function

4.1 Derivation of Nelson-Aalen and Kaplan-Meier estimators

When we have $(X_i, C_i) \overset{\text{iid}}{\sim} F$ such that noninformative censoring and parameter separability hold, we showed in Equation (3.8) that we can write the likelihood for the survival process:

$$f_\eta(t_1, \ldots, t_n, \delta_1, \ldots, \delta_n) = \prod_{i=1}^{n} \lambda_\eta(t_i)^{\delta_i} \exp\left(-\int_0^{t_i} \lambda_\eta(u) du\right).$$

This can again be simplified by collecting terms inside the exponential:

$$f_\eta(t_1, \ldots, t_n, \delta_1, \ldots, \delta_n) = \left(\prod_{i=1}^{n} \lambda_\eta(t_i)^{\delta_i}\right) \exp\left(-\sum_{i=1}^{n} \int_0^{t_i} \lambda_\eta(u) du\right). \tag{4.1}$$

Let's make a slight change to how we write the survival function. Define the indicator function $Y(u)$ to be

$$Y_i(u) = \mathbb{1}\left(t_i \geq u\right).$$

This function is left-continuous, with right-hand limits, an example of which is shown in Figure 4.1:

This allows us to rewrite our likelihood as follows:

$$\begin{aligned} f_\eta(t_1, \ldots, t_n, \delta_1, \ldots, \delta_n) &= \left(\prod_{i=1}^{n} \lambda_\eta(t_i)^{\delta_i}\right) \exp\left(-\sum_{i=1}^{n} \int_0^{\infty} Y_i(u) \lambda_\eta(u) du\right) & (4.2) \\ &= \left(\prod_{i=1}^{n} \lambda_\eta(t_i)^{\delta_i}\right) \exp\left(-\int_0^{\infty} \lambda_\eta(u) \sum_{i=1}^{n} Y_i(u) du\right) & (4.3) \end{aligned}$$

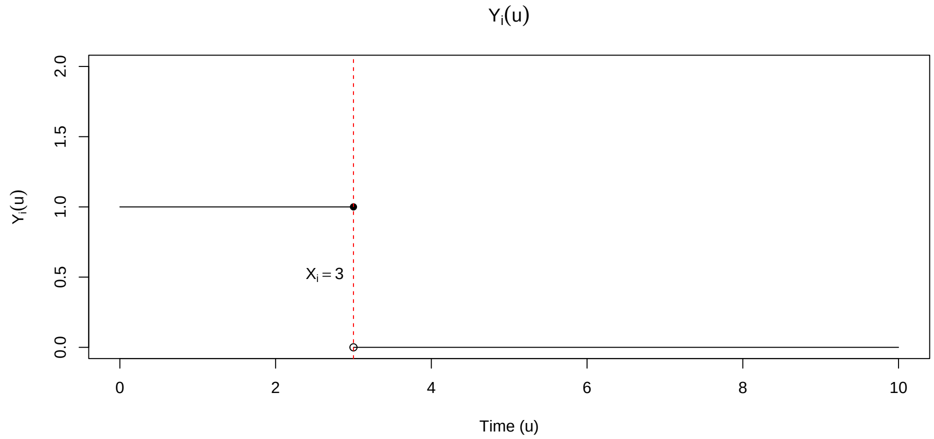

Figure 4.1: Example plot of an at-risk function

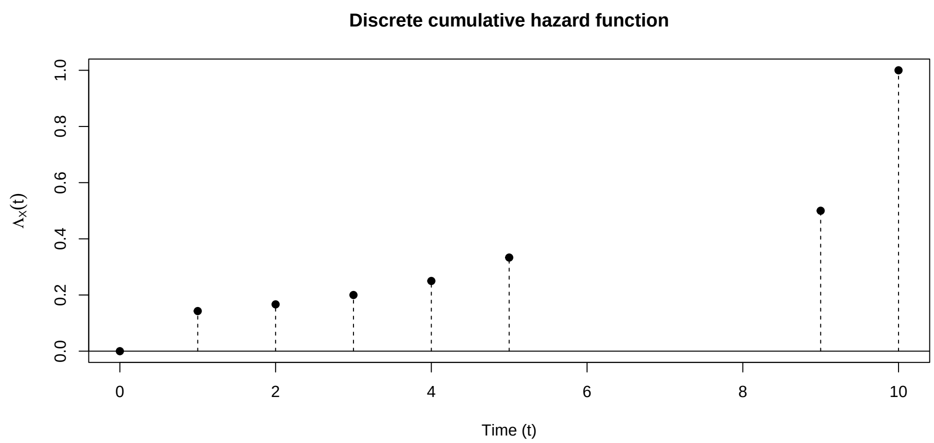

Figure 4.2: Example plot of a discrete hazard function

For notational convenience, we'll define the function $\overline{Y}(u)$ as:

$$\overline{Y}(u) = \sum_{i=1}^{n} Y_i(u).$$

Then our likelihood is:

$$f_\eta(t_1, \ldots, t_n, \delta_1, \ldots, \delta_n) = \left(\prod_{i=1}^{n} \lambda_\eta(t_i)^{\delta_i}\right) \exp\left(-\int_0^\infty \lambda_\eta(u)\overline{Y}(u)du\right) \tag{4.4}$$

We can consider a nonparametric model for the hazard, estimating λ at each t_i as a separate parameter. An example of this is shown in Figure 4.2, which corresponds to the discrete survival function in Figure 2.1. In order to evaluate the integral

$$\int_0^\infty \lambda(u)\overline{Y}(u)du,$$

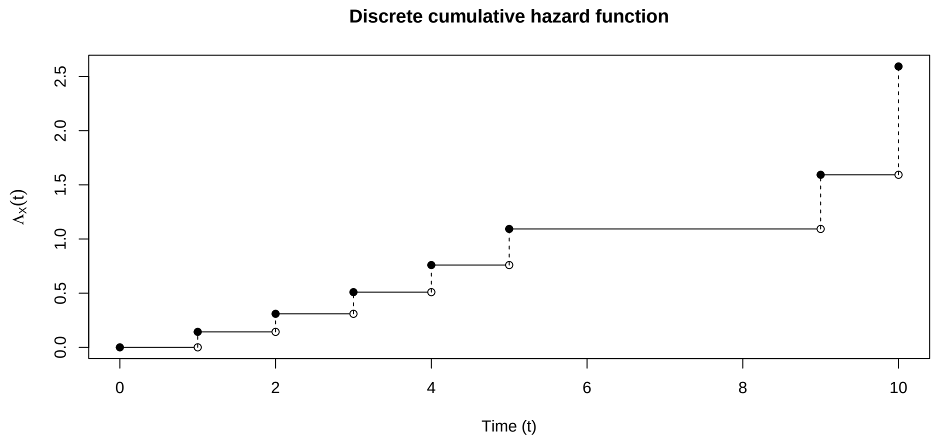

Figure 4.3: Example plot of a discrete cumulative hazard function

note that we can rewrite $\lambda(t_i)$ as

$$\lambda(t_i) = \Lambda(t_i) - \Lambda(t_i-),$$

where $\Lambda(t)$ is the cumulative hazard function. We'll write as $\lambda(u)$ as $d\Lambda(u)$. Finally, recall that because $S(t)$ is right-continuous, $\Lambda(t)$ is also right-continuous. We'll also need a bit of integration theory from Lebesgue-Stieltjies integrals. Suppose that G is a right-continuous, monotone function on $[0, \infty)$ with countably many discontinuities at a_i, and let $dG(a_i) = G(a_i) - G(a_i-)$. Then for a measurable function F on $[0, \infty)$, the integral over a set B

$$\int_B F(x) dG(x) = \sum_{i|a_i \in B} F(a_i) dG(a_i).$$

Using these results, the integral can be evaluated to

$$\int_0^\infty \left(\overline{Y}(u)\right) d\Lambda(u) du = \sum_{j=1}^n \lambda(t_j)\overline{Y}(t_j)$$

Let's take the log of the expression to get a log-likelihood:

$$\log f_\eta(t_1, \ldots, t_n, \delta_1, \ldots, \delta_n) = \sum_{i=1}^n \delta_i \log(\lambda_\eta(t_i)) - \sum_{j=1}^n \lambda_\eta(t_j)\overline{Y}(t_j) \tag{4.5}$$

Taking the gradient with respect to $\lambda_\eta(t_i)$ gives

$$\nabla \log f_\eta(t_1, \ldots, t_n, \delta_1, \ldots, \delta_n) = \frac{\delta_i}{\lambda_\eta(t_i)} - \overline{Y}(t_i). \tag{4.6}$$

Note that the Hessian is also diagonal, which implies asymptotic independence of $\lambda(t_i)$. This is solved at

$$\hat{\lambda}_\eta(t_i) = \frac{\delta_i}{\overline{Y}(t_i)} \tag{4.7}$$

This gives an expression for $\Lambda(t)$:

$$\Lambda^{\mathrm{NA}}(t) = \sum_{i|\delta_i=1,t_i\le t} \frac{1}{\overline{Y}(t_i)} \tag{4.8}$$

This also gives an expression for $S(t)$:

$$S^{\mathrm{KM}}(t) = \prod_{i|\delta_i=1,t_i\le t} (1 - \frac{1}{\overline{Y}(t_i)}). \tag{4.9}$$

This is also known as the **Kaplan-Meier estimator**. An alternative expression is:

$$S^{\mathrm{NA}}(t) = \exp\left(-\textstyle\sum_{i|\delta_i=1,t_i\le t} \frac{1}{\overline{Y}(t_i)}\right) \tag{4.10}$$

We can show that the cumulative hazard as implied by Equation (4.9) is asymptotically equivalent to Equation (4.8). Given Equation (2.9)

$$\Lambda^{\mathrm{KM}} = -\log\left(\prod_{i|\delta_i=1,t_i\le t} (1 - \frac{1}{\overline{Y}(t_i)})\right) \tag{4.11}$$

$$= -\sum_{i|\delta_i=1,t_i\le t} \log(1 - \frac{1}{\overline{Y}(t_i)}) \tag{4.12}$$

$$\approx \sum_{i|\delta_i=1,t_i\le t} \frac{1}{\overline{Y}(t_i)} \tag{4.13}$$

where the last line follows from the Taylor approximation of $\log(1-x) \approx -x$ when $x \approx 0$.

4.1.1 Kaplan-Meier estimator standard error

In order to get the standard errors for the Kaplan-Meier estimator, we'll use a Taylor expansion:

$$\log(S^{\mathrm{KM}}(t)) \approx \log(S(t)) + \frac{1}{S(t)}(S^{\mathrm{KM}}(t) - S(t)) \tag{4.14}$$

which leads to

$$\mathrm{Var}\left(\log(S^{\mathrm{KM}}(t))\right) = \frac{1}{S(t)^2}\mathrm{Var}\left(S^{\mathrm{KM}}(t)\right)$$

or

$$\mathrm{Var}\left(S^{\mathrm{KM}}(t)\right) = \mathrm{Var}\left(\log(S^{\mathrm{KM}}(t))\right) S(t)^2.$$

We use the plug-in estimator for $S(t)$ here, so we get:

$$\mathrm{Var}\left(S^{\mathrm{KM}}(t)\right) = \mathrm{Var}\left(\log(S^{\mathrm{KM}}(t))\right) (S^{\mathrm{KM}}(t))^2.$$

Now we need an expression for $\mathrm{Var}\left(\log(S^{\mathrm{KM}}(t))\right)$. First we write the log of the KM estimator:

$$\log \hat{S}^{\mathrm{KM}}(t) = \sum_{i|t_i \leq t} \log(1 - \hat{\lambda}(t_i)). \tag{4.15}$$

First we find the Taylor expansion for each term, which is justified by the fact that $(1-\hat{\lambda}(t_i)) \approx (1 - \lambda(t_i))$ for large samples:

$$\log(1 - \hat{\lambda}(t_i)) \approx \log(1 - \lambda(t_i)) - \frac{1}{1 - \lambda(t_i)}(\hat{\lambda}(t_i) - \lambda(t_i)) \tag{4.16}$$

Then

$$\mathrm{Var}\left(\log(1 - \hat{\lambda}(t_i))\right) \approx \frac{1}{(1 - \lambda(t_i))^2}\mathrm{Var}\left(\hat{\lambda}(t_i)\right)$$

We can estimate the $\mathrm{Var}\left(\hat{\lambda}(t_i)\right)$ as:

$$\mathrm{Var}\left(\hat{\lambda}(t_i)\right) = \mathrm{Var}\left(\delta_i\right)/\bar{Y}^2(t_i)$$

Treating δ_i as a binomial random variable with $\bar{Y}(t_i)$ number of trials:

$$\delta_i \sim \mathrm{Binomial}(\bar{Y}(t_i), p_i)$$

The variance of δ_i is $p_i(1 - p_i)\bar{Y}(t_i)$. Using $\hat{\lambda}(t_i)$ as a plug-in estimator for p_i as $\hat{\lambda}(t_i)$, this gives:

$$\mathrm{Var}\left(\delta_i\right) = \hat{\lambda}(t_i)(1 - \hat{\lambda}(t_i))\bar{Y}(t_i)$$

Putting this together with the $\bar{Y}^2(t_i)$ in the denominator gives the following estimate for the variance of $\hat{\lambda}(t_i)$:

$$\frac{\hat{\lambda}(t_i)(1 - \hat{\lambda}(t_i))}{\bar{Y}(t_i)}.$$

Finally using the plug-in estimator for $(1-\lambda(t_i))^2$ in the denominator of the Taylor expansion formula gives:

$$\begin{aligned}
\mathrm{Var}\left(\log(1 - \hat{\lambda}(t_i))\right) &\approx \frac{1}{(1 - \hat{\lambda}(t_i))^2}\frac{\hat{\lambda}(t_i)(1 - \hat{\lambda}(t_i))}{\bar{Y}(t_i)} && (4.17)\\
&\approx \frac{\hat{\lambda}(t_i)}{\bar{Y}(t_i)(1 - \hat{\lambda}(t_i))} && (4.18)\\
&\approx \frac{\delta_i}{\bar{Y}(t_i)(\bar{Y}(t_i) - \hat{\lambda}(t_i))} && (4.19)
\end{aligned}$$

Putting this all together along with the fact that $\lambda(t_i) \overset{\text{asymp}}{\perp\!\!\!\perp} \lambda(t_j)$, yields what is known as **Greenwood's formula**:

$$\mathrm{Var}\left(S^{\mathrm{KM}}(t)\right) = (S^{\mathrm{KM}}(t))^2 \sum_{i|\delta_i = 1, t_i \leq t} \frac{\delta_i}{\overline{Y}(t_i)(\overline{Y}(t_i) - \delta_i)}.$$

4.2 Confidence intervals

If we wanted to construct asymptotic, point-wise confidence intervals for the KM estimator, we can go about it in several ways. The most straightforward way to compute confidence intervals is to directly use the estimated survival function at t_0 and the standard error estimator from Greenwood's formula. Let $\hat{\sigma}(t)$ be

$$\sqrt{\sum_{i|d_i=1,t_i\leq t} \frac{d_i}{\overline{Y}(t_i)(\overline{Y}(t_i)-d_i)}}.$$

Then our confidence interval, C^{KM}, is

$$C^{\mathrm{KM}} = \left(\hat{S}^{\mathrm{KM}}(t_0) - z_{1-\alpha/2}\hat{\sigma}\hat{S}^{\mathrm{KM}}(t_0), \hat{S}^{\mathrm{KM}}(t_0) + z_{1-\alpha/2}\hat{\sigma}\hat{S}^{\mathrm{KM}}(t_0)\right)$$

The issue with this confidence interval is that it is not guaranteed to be greater than zero or less than 1, so we may have nonsensical results for upper and lower bounds. A solution is to build a confidence set for a suitably transformed Kaplan Meier estimator, and use the inverse transformation to enforce the natural $[0,1]$ bounds. One option is to use the logit transformation, another is to use the log-log transformation.

We'll walk through the log-log transformation:

Note that we have the following result:

$$\operatorname{Var}\left(\log(\hat{S}^{\mathrm{KM}}(t))\right) = \frac{1}{S(t)^2}\operatorname{Var}\left(\hat{S}^{\mathrm{KM}}(t)\right) \tag{4.20}$$

$$= \sum_{i|d_i=1,t_i\leq t} \frac{d_i}{\overline{Y}(t_i)(\overline{Y}(t_i)-d_i)}. \tag{4.21}$$

Then

$$\log(-\log(\hat{S}^{\mathrm{KM}}(t))) \approx \log(-\log(S(t))) - \frac{1}{\log(S(t))}(\log(S^{\mathrm{KM}}(t)) - \log(S(t)))$$

So

$$\operatorname{Var}\left(\log(-\log(\hat{S}^{\mathrm{KM}}(t)))\right) \approx \frac{1}{\log(S(t))^2}\operatorname{Var}\left(\log(\hat{S}^{\mathrm{KM}}(t)\right)$$

or

$$\mathrm{SE}(\log(-\log(\hat{S}^{\mathrm{KM}}(t)))) \approx \frac{1}{|\log(S(t))|}\hat{\sigma}(t)$$

We don't know $S(t)$, so we'll plug-in KM estimator for $S(t)$:

$$\mathrm{SE}(\log(-\log(\hat{S}^{\mathrm{KM}}(t)))) \approx \frac{1}{|\log(S^{\mathrm{KM}}(t))|}\hat{\sigma}(t)$$

Let $u = \log(-\log S(t))$, $\hat{u} = \log(-\log(\hat{S}^{\mathrm{KM}}(t)))$, and $\hat{\sigma}_u = \mathrm{SE}(\log(-\log(\hat{S}^{\mathrm{KM}}(t))))$. Then

$$\hat{S}^{\mathrm{KM}}(t) = \exp(-e^{\hat{u}}).$$

Note that $\exp(-e^u)$ is a monotone decreasing function of its input, u. This means that for a set $[a, b]$

$$a \le u \le b \implies \exp(-e^a) \ge \exp(-e^u) \ge \exp(-e^b).$$

We'll take it as a given that asymptotically,

$$\frac{\hat{u} - u}{\hat{\sigma}_u} \xrightarrow{d} \mathcal{N}(0,1).$$

Then we can derive an alternative asymptotic confidence interval for the Kaplan-Meier estimator of survival at time t_0 by transforming a confidence interval for u. Let $z_{1-\alpha/2}$ be the $1 - \alpha/2$ quantile of a standard normal distribution with CDF Φ, or

$$z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2).$$

$$\begin{aligned}
P\left(-z_{1-\alpha/2} \le \frac{\hat{u} - u}{\hat{\sigma}_u} \le z_{1-\alpha/2}\right) &= P\left(\hat{u} - \hat{\sigma}_u z_{1-\alpha/2} \le u \le \hat{u} + \hat{\sigma}_u z_{1-\alpha/2}\right) \\
&= P\left(\exp(-e^{\hat{u}-\hat{\sigma}_u z_{1-\alpha/2}}) \ge \exp(-e^u) \ge \exp(-e^{\hat{u}+\hat{\sigma}_u z_{1-\alpha/2}})\right) \\
&= P\left(\exp(-e^{\hat{u}} e^{-\hat{\sigma}_u z_{1-\alpha/2}}) \ge \exp(-e^u) \ge \exp(-e^{\hat{u}} e^{\hat{\sigma}_u z_{1-\alpha/2}})\right) \\
&= P\left(\exp(-e^{\hat{u}})^{e^{-\hat{\sigma}_u z_{1-\alpha/2}}} \ge \exp(-e^u) \ge \exp(-e^{\hat{u}})^{e^{\hat{\sigma}_u z_{1-\alpha/2}}}\right) \\
&= P\Big((\hat{S}^{\mathrm{KM}}(t))^{e^{-\mathrm{SE}(\log(-\log(\hat{S}^{\mathrm{KM}}(t))))z_{1-\alpha/2}}} \ge S(t) \\
&\qquad \ge (\hat{S}^{\mathrm{KM}}(t))^{e^{\mathrm{SE}(\log(-\log(\hat{S}^{\mathrm{KM}}(t))))z_{1-\alpha/2}}}\Big).
\end{aligned}$$

So

$$P\left(S(t) \in \left((\hat{S}^{\mathrm{KM}}(t))^{e^{\mathrm{SE}(\log(-\log(\hat{S}^{\mathrm{KM}}(t))))z_{1-\alpha/2}}}, (\hat{S}^{\mathrm{KM}}(t))^{e^{-\mathrm{SE}(\log(-\log(\hat{S}^{\mathrm{KM}}(t))))z_{1-\alpha/2}}}\right)\right) \overset{\text{asympt.}}{=} 1 - \alpha \tag{4.22}$$

4.2.1 Handling ties in the Nelson-Aalen estimator

We had assumed that no two events could occur at the same time, but for most real datasets this isn't realistic. A distinction must be made between a) assuming that ties are present in the data because, despite the true events happening in continuous time and thus no two events exactly coincide, the data have been rounded such that this exact ordering of events is lost, or b) that the true events happen in discrete time, and so there are truly events that co-occur.

In the continuous time scenario, O. Aalen et al. 2008 suggests using a modified estimator for hazard at time t_i when there are multiple $\delta_i = 1$. Let d_i be the number of events observed at time t_i. Then the proposed estimator for $\hat{\lambda}(t_i)$ is:

$$\hat{\lambda}(t_i) = \sum_{j=0}^{d_i-1} \frac{1}{\overline{Y}(t_i) - j} \tag{4.23}$$

In discrete time the proposal is to use:

$$\hat{\lambda}(t_i) = \frac{d_i}{\overline{Y}(t_i)} \tag{4.24}$$

4.2.2 Handling ties in the Kaplan-Meier estimator

It turns out, after some algebra, that using either Equation (4.23) or Equation (4.24) results in the following tie-corrected estimator for the KM estimator:

$$\hat{S}^{\mathrm{KM}}(t) = \prod_{i|d_i \geq 1, t_i \leq t} \left(1 - \frac{d_i}{\overline{Y}(t_i)}\right) \tag{4.25}$$

Greenwood's formula is then

$$\mathrm{Var}\left(S^{\mathrm{KM}}(t)\right) = (S^{\mathrm{KM}}(t))^2 \sum_{i|d_i \geq 1, t_i \leq t} \frac{d_i}{\overline{Y}(t_i)(\overline{Y}(t_i) - d_i)}.$$

This is the more commonly known form.

4.3 Nonparametric tests

Now that we've derived the nonparametric estimator for the cumulative hazard function, $\Lambda^{\mathrm{NA}}(t) = \sum_{i|\delta_i=1, t_i \leq t} \frac{1}{\overline{Y}(t_i)}$, we may be interested in testing the hypothesis that two populations have different cumulative hazard functions.

Intuitively it would make sense to compare the difference between the two cumulative hazard functions up to some τ:

$$\Lambda_1^{\mathrm{NA}}(\tau) - \Lambda_2^{\mathrm{NA}}(\tau) \tag{4.26}$$

and if this difference were large relative to the standard error under the null hypothesis, reject the null in favor of the alternative.

Let's formalize this a bit more. If the null hypothesis is:

$$H_0 : \lambda_1(t) = \lambda_2(t) \forall t \in [0, \tau]$$

then we can represent this common hazard function at $\lambda(t)$. Under the null, the nonparametric estimator combines all of the event times into one dataset and estimates $\hat{\lambda}(t)$. Let $\overline{Y}(t) = \overline{Y}_1(t) + \overline{Y}_1(t)$ be the total population at risk between the two samples. Let there be n_1 and n_2 samples in each respective study set. Let $t_1 \leq t_2 \leq< \cdots < t_{n_1+n_2}$ be the total combined set of event times. Let d_i be the total number of failures occurring at time t_i, and let d_{ij} be the total number of failures occurring at time t_i for sample j. Note that this could be zero.

Then the Nelson-Aalen estimator, assuming discrete time ties, for the cumulative hazard under the null is

$$\hat{\Lambda}^{\text{NA}}(\tau) = \sum_{i=1|t_i \leq \tau}^{n_1+n_2} \frac{d_i}{\overline{Y}(t_i)} \tag{4.27}$$

Then let the Nelson-Aalen estimator for the j-th cumulative hazard be

$$\hat{\Lambda}^{\text{NA}}(\tau) = \sum_{i=1|t_i \leq \tau}^{n_1+n_2} \frac{d_{ij}}{\overline{Y}_j(t_i)} \tag{4.28}$$

Given the common index over t_i we can compare the two sums more easily:

$$Z_j(\tau) = \sum_{i=1|t_i \leq \tau}^{n_1+n_2} \left(\frac{d_{ij}}{\overline{Y}_j(t_i)} - \frac{d_i}{\overline{Y}(t_i)} \right). \tag{4.29}$$

We can weight the comparisons differently by adding a weighting factor that is a function of t and j:

$$Z_j(\tau) = \sum_{i=1|t_i \leq \tau}^{n_1+n_2} W_j(t_i) \left(\frac{d_{ij}}{\overline{Y}_j(t_i)} - \frac{d_i}{\overline{Y}(t_i)} \right). \tag{4.30}$$

Crucially, this weighting function has the property that $W_j(t_i) = 0$ when $\overline{Y}_j(t_i) = 0$, because the hazard rate estimator $\hat{\lambda}_j(t_i)$ is not defined in this case. Let's rewrite the statistic $Z(\tau)$ a bit differently to elucidate the statistical properties, assuming that $W_j(t_i) = W(t_i)\overline{Y}_j(t_i)$, which satisfies the requirement that $W_j(t_i) = 0$ when $\overline{Y}_j(t_i) = 0$:

$$Z_j(\tau) = \sum_{i=1|t_i \leq \tau}^{n_1+n_2} W(t_i)\overline{Y}_j(t_i) \left(\frac{d_{ij}}{\overline{Y}_j(t_i)} - \frac{d_i}{\overline{Y}(t_i)} \right) \tag{4.31}$$

$$= \sum_{i=1|t_i \leq \tau}^{n_1+n_2} W(t_i) \left(d_{ij} - d_i \frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)} \right) \tag{4.32}$$

Now, conditional on $d_i, \overline{Y}_j(t_i), \overline{Y}(t_i)$, d_{ij} are distributed as hypergeometric random variables. Recall the definition of a hypergeometric random variable: It defines the distribution of successes (in our case this is failures) in a sample size of n from a finite population of size N where the total number of successes is K, with mean $n\frac{K}{N}$. The analogy to our scenario is d_{ij} is the number of failures in a samples of size $\overline{Y}_j(t_i)$ in a population size $\overline{Y}(t_i)$ where the total number of failures is d_i.

$$p(d_{ij} = k \mid d_i, \overline{Y}_j(t_i), \overline{Y}(t_i)) = \frac{\binom{d_i}{k}\binom{\overline{Y}(t_i)-d_i}{\overline{Y}_j(t_i)-k}}{\binom{\overline{Y}(t_i)}{\overline{Y}_j(t_i)}}.$$

Then

$$\mathbb{E}\left[d_{ij} \mid d_i, \overline{Y}_j(t_i), \overline{Y}(t_i)\right] = d_i\frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}$$

For notational convenience, let's call $A_{ij} = d_{ij} - d_i\frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}$. Under the null hypothesis, the mean of $Z_j(\tau)$ is zero, because $\mathbb{E}\left[A_{ij} \mid d_i, \overline{Y}_j(t_i), \overline{Y}(t_i)\right] = 0$ so

$$\mathbb{E}\left[A_{ij}\right] = \mathbb{E}_{d_i,\overline{Y}_j(t_i),\overline{Y}(t_i)}\left[\mathbb{E}\left[A_{ij} \mid d_i, \overline{Y}_j(t_i), \overline{Y}(t_i)\right]\right] = 0.$$

We can also compute the variance using our result.

$$\mathrm{Var}\left(Z_j(\tau)\right) = \sum_i W(t_i)^2\mathrm{Var}\left(A_{ij}\right) + 2\sum_{i<k} W(t_i)W(t_j)\mathrm{Cov}\left(A_{ij}, A_{kj}\right)$$

Given the hypergeometric distribution, we can read off the variance as

$$\mathrm{Var}\left(A_{ij}\right) = d_i\frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\left(1 - \frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\right)\frac{\overline{Y}(t_i) - d_i}{\overline{Y}(t_i) - 1}$$

Now let's compute $\mathrm{Cov}\left(A_{ij}, A_{kj}\right)$, noting that $i < k$. We know that $\mathbb{E}\left[A_{ij}\right] = 0$, so we just need to compute $\mathbb{E}\left[A_{ij}A_{kj}\right]$. We can use the tower property of expectation. First we need to define something called the history, or the *filtration*, of the process. A filtration is an increasing family of σ-algebras, $\{\mathcal{F}_l, 0 \le l < \infty\}$ such that $\mathcal{F}_l \subset \mathcal{F}_m$ for all $l < m$. This is a way of formalizing the idea that as time progresses, information about events accrues. If an event $E \in \mathcal{F}_l$ then $\mathbb{E}\left[\mathbb{1}(E) \mid \mathcal{F}_l\right] = \mathbb{1}(E)$, because we're conditioning on the full set of information, and E is part of that information. It's analogous to saying for two random variables X, Y, $\mathbb{E}\left[XY \mid X\right] = X\mathbb{E}\left[Y \mid X\right]$. Taking this approach below, we show that the covariance is zero. Let $\mathcal{F}_k$ be the collection of information just before t_k, which means, more formally that it is

$$\mathcal{F}_k = \sigma\{d_{i1}, d_{i2}, \overline{Y}_1(t_i), \overline{Y}_2(t_i), i < k\} \tag{4.33}$$

Then

$$\mathbb{E}\left[A_{ij}A_{kj}\right] = \mathbb{E}\left[\mathbb{E}\left[A_{ij}A_{kj} \mid \mathcal{F}_k\right]\right] \tag{4.34}$$
$$= \mathbb{E}\left[A_{ij}\mathbb{E}\left[A_{kj} \mid \mathcal{F}_k\right]\right] \tag{4.35}$$
$$= 0. \tag{4.36}$$

Where the last line follows because

$$\mathbb{E}\left[A_{kj} \mid \mathcal{F}_k\right] = \mathbb{E}_{d_i,\overline{Y}_j(t_i),\overline{Y}(t_i)}\left[\mathbb{E}\left[A_{kj} \mid \mathcal{F}_k, d_i, \overline{Y}_j(t_i), \overline{Y}(t_i)\right]\right] \tag{4.37}$$
$$= 0 \tag{4.38}$$

as we showed above. Thus,

$$\mathrm{Var}\left(Z_j(\tau)\right) = \sum_i W(t_i)^2 \mathrm{Var}\left(A_{ij}\right) \tag{4.39}$$
$$= \sum_i W(t_i)^2 \left(d_i \frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\left(1 - \frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\right)\frac{\overline{Y}(t_i) - d_i}{\overline{Y}(t_i) - 1}\right) \tag{4.40}$$

Note that, due to A_{ij} being mean zero, we have that

$$\begin{aligned}\mathrm{Var}\left(A_{ij}\right) &= \mathbb{E}_{d_i,\overline{Y}_j(t_i),\overline{Y}(t_i)}\left[\mathrm{Var}\left(A_{ij} \mid d_i, \overline{Y}_j(t_i), \overline{Y}(t_i)\right)\right] + \mathrm{Var}\left(\mathbb{E}\left[A_{ij} \mid d_i, \overline{Y}_j(t_i), \overline{Y}(t_i)\right]\right) \\ &= \mathbb{E}_{d_i,\overline{Y}_j(t_i),\overline{Y}(t_i)}\left[\mathrm{Var}\left(A_{ij} \mid d_i, \overline{Y}_j(t_i), \overline{Y}(t_i)\right)\right]\end{aligned}$$

Then

$$\mathrm{Var}\left(A_{ij}\right) = \mathbb{E}_{d_i,\overline{Y}_j(t_i),\overline{Y}(t_i)}\left[\mathrm{Var}\left(A_{ij} \mid d_i, \overline{Y}_j(t_i), \overline{Y}(t_i)\right)\right]$$

This means that we can construct an unbiased estimator for $\mathrm{Var}\left(Z_j(\tau)\right)$ by the following:

$$\mathrm{Var}\left(\hat{Z}_j(\tau)\right) = \sum_i W(t_i)^2 \mathrm{Var}\left(A_{ij} \mid d_i, \overline{Y}(t_i), \overline{Y}_j(t_i)\right)$$

and

$$\begin{aligned}\mathbb{E}\left[\mathrm{Var}\left(\hat{Z}_j(\tau)\right)\right] &= \sum_i W(t_i)^2 \mathbb{E}\left[\mathrm{Var}\left(A_{ij} \mid d_i, \overline{Y}(t_i), \overline{Y}_j(t_i)\right)\right] \\ &= \sum_i W(t_i)^2 \mathrm{Var}\left(A_{ij}\right) \\ &= \mathrm{Var}\left(Z_j(\tau)\right)\end{aligned}$$

We won't go into the details yet, but it turns out that

$$\frac{Z_j(\tau)}{\sqrt{\mathrm{Var}\left(\hat{Z}_j(\tau)\right)}} \overset{\text{asympt.}}{\sim} \mathrm{Normal}(0, 1)$$

One could use this result to define a rejection region that is calibrated under the null.

4.4 More on log-rank tests

I motivated the log-rank test by stating that we wanted to compare estimates of the hazard function. Let's do a quick derivation to show why this is the case: We start with the weighted log-rank test as we have derived it:

$$Z_j(\tau) = \sum_{i=1|t_i\leq\tau}^{n_1+n_2} W(t_i)\left(d_{ij} - d_i\frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\right) \tag{4.41}$$

We can express this in terms of hazard estimators $\hat{\lambda}_j(t_i) = \frac{d_{ij}}{\overline{Y}_j(t_i)}$: Let's let $j \in \{1,2\}$. Then

$$\begin{aligned}
\sum_{i=1|t_i\leq\tau}^{n_1+n_2} W(t_i)\left(d_{ij} - d_i\frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\right) &= \sum_{i=1|t_i\leq\tau}^{n_1+n_2} W(t_i)\left(\frac{d_{ij}\overline{Y}(t_i) - d_i\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\right)\\
&= \sum_{i=1|t_i\leq\tau}^{n_1+n_2} W(t_i)\left(\frac{d_{ij}\overline{Y}(t_i) - (d_{ij}+d_{ij'})\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\right)\\
&= \sum_{i=1|t_i\leq\tau}^{n_1+n_2} W(t_i)\left(\frac{d_{ij}\overline{Y}_{j'}(t_i) - d_{ij'}\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\right)\\
&= \sum_{i=1|t_i\leq\tau}^{n_1+n_2} W(t_i)\frac{\overline{Y}_{j'}(t_i)\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\left(\frac{d_{ij}}{\overline{Y}_j(t_i)} - \frac{d_{ij'}}{\overline{Y}_{j'}(t_i)}\right)
\end{aligned}$$

Thus we can see that $Z_1(\tau) = -Z_2(\tau)$. Let's rewrite this in terms of integrals over the positive reals

$$\begin{aligned}
\sum_{i=1|t_i\leq\tau}^{n_1+n_2} W(t_i)\frac{\overline{Y}_{j'}(t_i)\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\left(\frac{d_{ij}}{\overline{Y}_j(t_i)} - \frac{d_{ij'}}{\overline{Y}_{j'}(t_i)}\right) &= \int_0^\infty W(u)\frac{\overline{Y}_{j'}(u)\overline{Y}_j(u)}{\overline{Y}(u)}\left(d\hat{\Lambda}_1(u) - d\hat{\Lambda}_2(u)\right)\\
&= \int_0^\infty W(u)\frac{\overline{Y}_{j'}(u)\overline{Y}_j(u)}{\overline{Y}(u)}d\left(\hat{\Lambda}_1(u) - \hat{\Lambda}_2(u)\right)
\end{aligned}$$

A more general Lebesgue-Stieltjies theory will show that the integral above is well-defined. More on this later...

Let's say we're going to test multiple groups for equality of hazard rates. Then we will write the log-rank statistic like so, with $n = \sum_{j=1}^J n_j$:

$$Z_j(\tau) = \sum_{i=1|t_i\leq\tau}^{n} W(t_i)\left(d_{ij} - d_i\frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\right) \tag{4.42}$$

The variance of $Z_j(\tau)$ is as was derived. We can show, and I mentioned, that $d_{i1},\dots,d_{iJ} \mid d_i, \overline{Y}_1(t_i),\dots,\overline{Y}_J(t_i)$ is multivariate hypergeometric distributed. That means we can derive the variance and the covariance for these random variables. I'll spare the details here. Given the result that in the two-group test, $Z_1(\tau) = -Z_2(\tau)$, we might expect the $Z_j(\tau)$ to be linearly dependent. This is indeed the case, which we can see from the fact that the sum of all $Z_j(\tau)$ is zero. Then we might ask how do we construct a test statistic from a degenerate random variable. The answer is that we choose $J-1$ of the statistics, and it doesn't matter which statistics we choose. Given the covariance matrix Σ, we can construct a quadratic form:

$$\chi^2 = (Z_1(\tau), Z_2(\tau), \dots, Z_{J-1}(\tau))\Sigma^{-1}(Z_1(\tau), Z_2(\tau), \dots, Z_{J-1}(\tau))^T \tag{4.43}$$

which, under H_0, is asymptotically distributed χ^2 with $J-1$ degrees of freedom.

Let $\mathbf{Z}(\tau) = (Z_1(\tau), Z_2(\tau), \dots, Z_J(\tau))^T$ and let $\Sigma = \mathrm{Cov}(\mathbf{Z}(\tau))$. To show why it doesn't matter which groups we choose, imagine we have two matrices $A \in \mathbb{R}^{J-1\times J}$ and $B \in \mathbb{R}^{J-1\times J}$ which, when left multiplying the vector$\mathbf{Z}(\tau)$ select subsets of the $J-1$ groups. An example of A for $J=3$ might be:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{4.44}$$

Let both A and B be rank $J-1$. We define χ^2_A to be

$$\chi^2_A = (A\mathbf{Z}(\tau))^T(A\Sigma A^T)^{-1}A\mathbf{Z}(\tau) \tag{4.45}$$

$$\chi^2_B = (B\mathbf{Z}(\tau))^T(B\Sigma B^T)^{-1}B\mathbf{Z}(\tau) \tag{4.46}$$

As A and B are full-row-rank there exists an invertible matrix C such that $B = CA$. Then

$$\chi^2_B = (CA\mathbf{Z}(\tau))^T(CA\Sigma A^T C^T)^{-1}CA\mathbf{Z}(\tau) \tag{4.47}$$

$$= \mathbf{Z}(\tau))^T A^T C^T (C^T)^{-1}(A\Sigma A^T)^{-1}C^{-1}CA\mathbf{Z}(\tau) \tag{4.48}$$

$$= \mathbf{Z}(\tau))^T A^T (A\Sigma A^T)^{-1}A\mathbf{Z}(\tau) \tag{4.49}$$

$$= (A\mathbf{Z}(\tau))^T(A\Sigma A^T)^{-1}A\mathbf{Z}(\tau) \tag{4.50}$$

$$= \chi^2_A \tag{4.51}$$

Chapter 5

Parametric and nonparametric regression models

This chapter combines content from O. Aalen et al. 2008, Klein, Moeschberger, et al. 2003, Harrell et al. 2001, Collett 1994, and Keener 2010.

Thus far we have dealt exclusively with simple univariate estimation. More often than not, we will also have covariates associated with our failure time observations. Let the observed failure data, be, as usual X_i is time to failure, C_i is time to censoring, $T_i = \min(X_i, C_i)$, is the observed event time, and $\delta = \mathbb{1}(X_i \leq C_i)$ is the censoring indicator. Suppose we also have covariates for each individual i $\mathbf{z}_i \in \mathbb{R}^k$. These could be age, sex at birth, comorbidities. Over a short enough timespan, these covariates can be considered fixed over time. Other covariates, like blood pressure, or time since last colonoscopy, would be time varying covariates, which we'll denote as $\mathbf{z}(x)_i$.

Much of our study has been on the hazard function $\lambda(t)$. We'll consider this parameterized by a vector of parameters $\boldsymbol{\theta}$, so we'll write $\lambda(t \mid \boldsymbol{\theta})$ for the hazard function. In order to incorporate covariates into the hazard rate, we'll work with relative risk regression, or

$$\lambda_i(t) = \lambda_0(t \mid \boldsymbol{\theta}) r(\boldsymbol{\beta}, \mathbf{z}_i)$$

where r is a function $\mathbb{R} \to \mathbb{R}^+$. Note that this assumes that all individuals share a common baseline hazard, $\lambda_0(t \mid \boldsymbol{\theta})$, and have time-invariant, individual relative risk contributions $r(\boldsymbol{\beta}, \boldsymbol{z}_i)$. A common choice is that $r(\boldsymbol{\beta}, \boldsymbol{z}_i) \equiv \exp(\boldsymbol{z}_i^T \boldsymbol{\beta})$.

The function is called the relative risk function because when we compare the hazard rates for two individuals i and j, the common baseline hazard drops out of the comparison:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \exp(\boldsymbol{z}_i)^T \boldsymbol{\beta}) / \exp(\boldsymbol{z}_j^T \boldsymbol{\beta}).$$

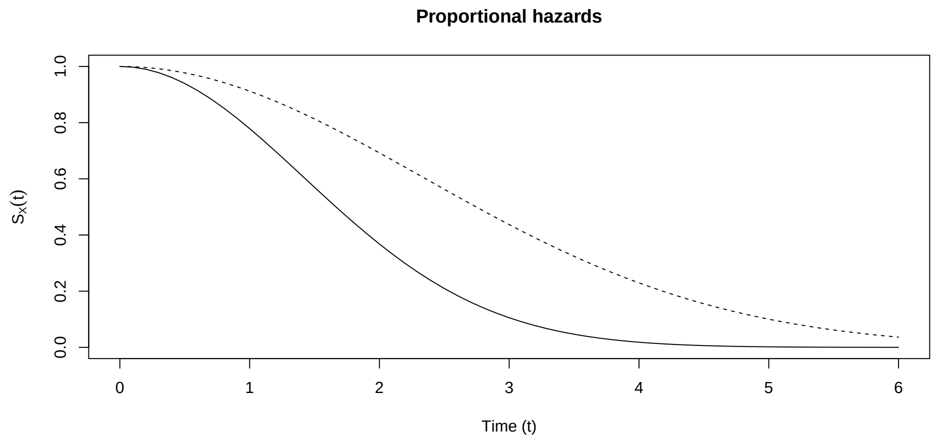

Figure 5.1: Example of survival functions with proportional hazards

Of course, the above holds with general $r(\boldsymbol{\beta}, \boldsymbol{z}_i)$. Let's see what this implies for the survival function for i vs. j:

$$\begin{aligned}
S_i(t) &= \exp\left(-\int_0^t e^{\mathbf{z}_i^T \boldsymbol{\beta}} \lambda_0(u \mid \boldsymbol{\theta}) du\right) \\
&= \exp\left(-\int_0^t \lambda_0(u \mid \boldsymbol{\theta}) du\right)^{e^{\mathbf{z}_i^T \boldsymbol{\beta}}} \\
&= \left(\exp\left(-\int_0^t \lambda_0(u \mid \boldsymbol{\theta}) du\right)^{e^{\mathbf{z}_j^T \boldsymbol{\beta}}}\right)^{\frac{e^{\mathbf{z}_i^T \boldsymbol{\beta}}}{e^{\mathbf{z}_j^T \boldsymbol{\beta}}}} \\
&= \left(\exp\left(-\int_0^t \lambda_0(u \mid \boldsymbol{\theta}) du\right)^{e^{\mathbf{z}_j^T \boldsymbol{\beta}}}\right)^{e^{(\mathbf{z}_i^T - \mathbf{z}_j^T)\boldsymbol{\beta}}} \\
&= S_j(t)^{e^{(\mathbf{z}_i^T - \mathbf{z}_j^T)\boldsymbol{\beta}}}
\end{aligned}$$

What this means is that the survival curves never cross. To see why, note that $S_i(0) = S_j(0) = 1$, and WLOG, suppose $(\mathbf{z}_i^T - \mathbf{z}_j^T)\boldsymbol{\beta} \leq 0$. Then $S_i(t) \geq S_j(t)$ for all t. See Figure 5.1 for a demonstration of proportional hazards. See Figure 5.1 for a demonstration of proportional hazards and Figure 5.2 for a demonstration of nonproportional hazards.

Proportional hazards (or relative risk) models assume that the survival functions never cross, which is a strong assumption.

Let's do a simple example.

Example 5.0.1. Simple exponential regression The following example is adapted from Collett 1994. Suppose we have individuals grouped into two groups, groups 1 and 2, and let $\mathbf{z}_i$

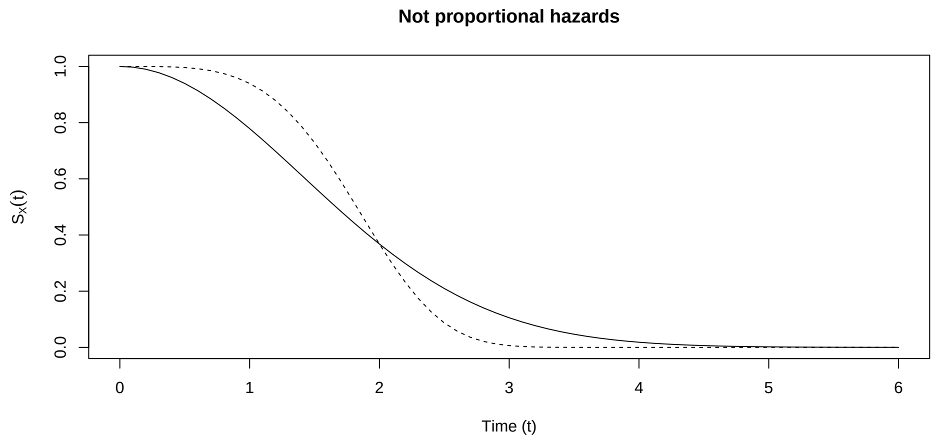

Figure 5.2: Example of survival functions that do not adhere to proportional hazards

equal 1 for those in group 2 and 0 for those in group 1. Suppose further we have noninformative censoring, parameter separability, and exponentially distributed survival times with common baseline hazard of λ, so we have observed the following dataset:

$$\{(t_i, \delta_i, z_i), i = 1, \ldots, n\}$$

Then the hazard rate for group 1 is λ, while the hazard in group 2 is λe^{β}. Let $n_1 = \sum_i (1 - z_i)$ and $n_2 = sum_i z_i$. Then the likelihood contribution for the individuals for whom $z_i = 0$ is

$$\prod_{i|z_i=0} \lambda^{\delta_i} e^{-\lambda t_i}$$

and the likelihood contribution for individuals in group 2 is

$$\prod_{i|z_i=1} (\lambda e^{\beta})^{\delta_i} e^{-\lambda e^{\beta} t_i}$$

We can simplify this. Let $r_1 = \sum_i (1 - z_i)\delta_i$, and let $r_2 = \sum_i z_i \delta_i$. Let $T_1 = \sum_i (1 - z_i) t_i$, and $T_2 = \sum_i z_i t_i$. Then the joint likelihood may be written:

$$\lambda^{r_1} e^{-\lambda T_1} (\lambda e^{\beta})^{r_2} e^{-\lambda e^{\beta} T_2} = \lambda^{r_1 + r_2} e^{-\lambda T_1} e^{r_2 \beta} e^{-\lambda e^{\beta} T_2}.$$

Let $\ell(\lambda, \beta)$ be the log-likelihood function. Then the score equations are

$$\frac{\partial}{\partial \lambda} \ell(\lambda, \beta) : \frac{r_1 + r_2}{\lambda} - T_1 - e^{\beta} T_2$$
$$\frac{\partial}{\partial \beta} \ell(\lambda, \beta) : r_2 - \lambda e^{\beta} T_2$$

solving these for the unknowns is

$$\frac{r_1 + r_2}{T_1 + e^{\beta} T_2} = \lambda$$
$$\frac{r_2}{\lambda T_2} = e^{\beta}$$

which simplifies to

$$\hat{\lambda} = \frac{r_1}{T_1}$$
$$\hat{e^{\beta}} = \frac{T_1/r_1}{T_2/r_2} = \frac{r_2}{T_2}\frac{T_1}{r_1}$$

These estimates make sense: The first is the reciprocal of the average survival time for those in Group 1, and the second is the ratio of the average survival times in each group.

We can show using Example 3.4.1 that both of these estimators converge a.s. to the true values. $\frac{r_2}{T_2} \overset{\text{a.s.}}{\to} \lambda e^{\beta}$, $\frac{T_1}{r_1} \overset{\text{a.s.}}{\to} \frac{1}{\lambda}$

Let's find the asymptotic variance of the estimand β

$$\frac{\partial}{\partial \lambda}\left(\frac{\partial}{\partial \lambda}\ell(\lambda, \psi)\right) = -\frac{r_1 + r_2}{\lambda^2} \tag{5.1}$$
$$\frac{\partial}{\partial \beta}\left(\frac{\partial}{\partial \lambda}\ell(\lambda, \psi)\right) = -e^{\beta} T_2 \tag{5.2}$$
$$\frac{\partial}{\partial \beta}\left(\frac{\partial}{\partial \beta}\ell(\lambda, \psi)\right) = -\lambda e^{\beta} T_2 \tag{5.3}$$

Then the observed information matrix is

$$\begin{bmatrix} \frac{r_1+r_2}{\lambda^2} & e^{\beta} T_2 \\ e^{\beta} T_2 & \lambda e^{\beta} T_2 \end{bmatrix} \tag{5.4}$$

which has the inverse:

$$\frac{1}{\frac{(r_1+r_2)e^{\beta}T_2}{\lambda} - e^{2\beta}T_2^2} \begin{bmatrix} \lambda e^{\beta} T_2 & -e^{\beta} T_2 \\ -e^{\beta} T_2 & \frac{r_1+r_2}{\lambda^2} \end{bmatrix} \tag{5.5}$$

So the plug-in standard error for β is

$$\sqrt{\frac{\frac{r_1+r_2}{\lambda^2}}{\frac{(r_1+r_2)e^{\beta}T_2}{\lambda} - e^{2\beta}T_2^2}}$$

Plugging in the MLEs gives

$$\sqrt{\frac{\frac{r_1+r_2}{(r_1/T_1)^2}}{\frac{(r_1+r_2)\frac{T_1 r_2}{r_1}}{r_1/T_1} - (\frac{T_1 r_2}{r_1})^2}} = \sqrt{\frac{r_1 + r_2}{r_1 r_2}}$$

We can use this expression to generate an asymptotic confidence interval for β:

$$P(\beta \in C^{\beta}) = P\left(\beta \in \left(e^{\hat{\beta}} - z_{1-\alpha/2}\sqrt{\frac{r_1 + r_2}{r_1 r_2}}, e^{\hat{\beta}} + z_{1-\alpha/2}\sqrt{\frac{r_1 + r_2}{r_1 r_2}}\right)\right)$$

In the preceding example, we shied away from using the Fisher information because T_2 was not easily accessible. But we can use the results from Example 3.4.1 to derive an exact expression for the asymptotic sampling variance for the MLE.

Example 5.0.2. Continued example This is an expansion of the example in Collett 1994.

$$\frac{\partial}{\partial\lambda}\left(\frac{\partial}{\partial\lambda}\ell(\lambda,\psi)\right) = -\frac{r_1 + r_2}{\lambda^2} \tag{5.6}$$

$$\frac{\partial}{\partial\beta}\left(\frac{\partial}{\partial\lambda}\ell(\lambda,\psi)\right) = -e^{\beta}T_2 \tag{5.7}$$

$$\frac{\partial}{\partial\beta}\left(\frac{\partial}{\partial\beta}\ell(\lambda,\psi)\right) = -\lambda e^{\beta}T_2 \tag{5.8}$$

Using the results of Example 3.4.1, we know that

$$\mathbb{E}\left[r_1\right] = n_1\mathbb{E}_{C_i}\left[1 - e^{-\lambda C_i}\right], \ \mathbb{E}\left[r_2\right] = n_2\mathbb{E}_{C_i}\left[1 - e^{-\lambda e^{\beta}C_i}\right], \text{and}\ \mathbb{E}\left[T_2\right] = n_2\frac{1}{\lambda e^{\beta}}\mathbb{E}_{C_i}\left[1 - e^{-\lambda e^{\beta}C_i}\right]$$

Then the Fisher information is

$$\begin{bmatrix} \frac{n_1\mathbb{E}_{C_i}\left[1-e^{-\lambda C_i}\right]+n_2\mathbb{E}_{C_i}\left[1-e^{-\lambda e^{\beta}C_i}\right]}{\lambda^2} & \frac{1}{\lambda}n_2\mathbb{E}_{C_i}\left[1 - e^{-\lambda e^{\beta}C_i}\right] \\ \frac{1}{\lambda}n_2\mathbb{E}_{C_i}\left[1 - e^{-\lambda e^{\beta}C_i}\right] & n_2\mathbb{E}_{C_i}\left[1 - e^{-\lambda e^{\beta}C_i}\right] \end{bmatrix} \tag{5.9}$$

Let $\mathbb{E}\left[r_{i1}\right] = \mathbb{E}_{C_i}\left[1 - e^{-\lambda C_i}\right]$ and $\mathbb{E}\left[r_{i2}\right] = \mathbb{E}_{C_i}\left[1 - e^{-\lambda e^{\beta}C_i}\right]$. We know the asymptotic variance of the MLE is the inverse of the Fisher information matrix. The inverse is:

$$\frac{\lambda^2}{n_1 n_2\mathbb{E}\left[r_{i1}\right]\mathbb{E}\left[r_{i2}\right]}\begin{bmatrix} n_2\mathbb{E}\left[r_{i2}\right] & -n_2\mathbb{E}\left[r_{i2}\right]/\lambda \\ -n_2\mathbb{E}\left[r_{i2}\right]/\lambda & \frac{n_1\mathbb{E}[r_{i1}]+n_2\mathbb{E}[r_{i2}]}{\lambda^2} \end{bmatrix} = \begin{bmatrix} \frac{\lambda^2}{n_1\mathbb{E}[r_{i1}]} & -\frac{\lambda}{n_1\mathbb{E}[r_{i1}]} \\ -\frac{\lambda}{n_1\mathbb{E}[r_{i1}]} & \frac{n_1\mathbb{E}[r_{i1}]+n_2\mathbb{E}[r_{i2}]}{n_1 n_2\mathbb{E}[r_{i1}]\mathbb{E}[r_{i2}]} \end{bmatrix} \tag{5.10}$$

So the asymptotic standard error for β is

$$\sqrt{\frac{n_1\mathbb{E}_{C_i}\left[1 - e^{-\lambda C_i}\right] + n_2\mathbb{E}_{C_i}\left[1 - e^{-\lambda e^{\beta}C_i}\right]}{n_1 n_2\mathbb{E}_{C_i}\left[1 - e^{-\lambda C_i}\right]\mathbb{E}_{C_i}\left[1 - e^{-\lambda e^{\beta}C_i}\right]}}$$

5.1 Asymptotic interlude

As you've already no doubt gathered, many of the results for inference and hypothesis testing in survival analysis rely on asymptotic normality of the MLE. Before we get too much further into the quarter, I thought it would be a good idea to review the asymptotic results for maximum likelihood. This outline of results is from Keener 2010.

Let $X_i, i = 1, 2, \ldots$ be distributed *i.i.d.* with density f_θ where $\theta \in \mathbb{R}^p$. We suppose that the support of X_i does not depend on θ, and that our MLE's are consistent for θ. This is pretty mild, and only requires that likelihood ratios are integrable and our model is identifiable.

Given these conditions, we can expand each dimension of the gradient of the log-likelihood evaluated at the MLE $\ell(\hat{\theta})$ around the true parameter value $\theta^\dagger$ in a one-term Taylor expansion:

$$(\nabla_\theta \ell(\theta) \mid_{\theta=\hat{\theta}_n})_j = (\nabla_\theta \ell(\theta) \mid_{\theta=\theta^\dagger})_j + (\nabla_\theta (\nabla_\theta \ell(\theta))_j) \mid_{\theta=\tilde{\theta}_n^j} (\hat{\theta}_n - \theta^\dagger)$$

where $\tilde{\theta}_n^j$ is a point on the chord between $\hat{\theta}_n$ and $\theta^\dagger$ and may depend on the coordinate j. Noting that $(\nabla_\theta \ell(\theta) \mid_{\theta=\hat{\theta}_n})_j = 0$ for all j, we get the set of p linear equations:

$$(\nabla_\theta \ell(\theta) \mid_{\theta=\theta^\dagger})_j = -(\nabla_\theta (\nabla_\theta \ell(\theta))_j) \mid_{\theta=\tilde{\theta}_n^j} (\hat{\theta}_n - \theta^\dagger)$$

Multiplying both sides by $n^{-1/2}$ gives:

$$n^{-1/2}(\nabla_\theta \ell(\theta) \mid_{\theta=\theta^\dagger})_j = -n^{-1}(\nabla_\theta (\nabla_\theta \ell(\theta))_j) \mid_{\theta=\tilde{\theta}_n^j} n^{-1/2}(\hat{\theta}_n - \theta^\dagger)$$

We can write all p one-term Taylor expansions in matrix form by concatenating all of our equations together. Let H be a $p \times p$ matrix where the j^{th} row is

$$H_{[j,:]} = (\nabla_\theta (\nabla_\theta \ell(\theta))_j) \mid_{\theta=\tilde{\theta}_n^j}$$

Then the equations in matrix form are:

$$\frac{1}{\sqrt{n}} \nabla_\theta \ell(\theta) \mid_{\theta=\theta^\dagger} = -\frac{1}{n} H \sqrt{n}(\hat{\theta}_n - \theta^\dagger)$$

Writing out the expressions,

$$\frac{1}{\sqrt{n}} \nabla_\theta \ell(\theta) \mid_{\theta=\theta^\dagger}, \quad \frac{1}{n} H$$

as explicit sums gives:

$$\frac{1}{\sqrt{n}} \nabla_\theta \ell(\theta) \mid_{\theta=\theta^\dagger} = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} (\nabla_\theta \log f_\theta(X_i)) \mid_{\theta=\theta^\dagger} \tag{5.11}$$

$$-\frac{1}{n} H_{[j,:]} = -\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta (\nabla_\theta \log f_\theta(X_i)) \mid_{\theta=\tilde{\theta}_n^j} . \tag{5.12}$$

Given the structure of these terms, Equation (5.11) will be amenable to a multivariate version of the CLT, while Equation (5.12) will be amenable to a weak law of large numbers. We'll take the following multivariate CLT as given:

Theorem 5.1.1. Multivariate CLT, (Keener 2010) Let $X_1, X_2, \ldots$ be i.i.d random vectors in $\mathbb{R}^k$ with a common mean $\mathbb{E}[X_i] = \mu$ and common covariance matrix $\Sigma = \mathbb{E}[(X_i - \mu)(X_i - \mu)^T]$. If $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$, then

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \text{Normal}(0, \Sigma)$$

By the multivariate central limit (MCLT) theorem, Equation (5.11) converges in distribution to

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}(\nabla_\theta \log f_\theta(X_i))\mid_{\theta=\theta^\dagger}\xrightarrow{d}\mathcal{N}(\mathbb{E}\left[(\nabla_\theta \log f_\theta(X_i))\mid_{\theta=\theta^\dagger}\right],\mathbb{E}\left[(\nabla_\theta \log f_\theta(X_i))\mid_{\theta=\theta^\dagger}(\nabla_\theta \log f_\theta(X_i))\mid^T_{\theta=\theta^\dagger}\right]).$$

Note that $\mathcal{I}(\theta^\dagger)=\mathbb{E}\left[(\nabla_\theta \log f_\theta(X_i))\mid_{\theta=\theta^\dagger}(\nabla_\theta \log f_\theta(X_i))\mid^T_{\theta=\theta^\dagger}\right]$. We'll also take the fact that

$$-\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta(\nabla_\theta \log f_\theta(X_i))\mid_{\theta=\tilde{\theta}_n^j}\xrightarrow{p}-\mathbb{E}\left[\nabla_\theta^2 \log f_\theta(X_1)\right]_{[j,:]} \tag{5.13}$$

We'll also need a lemma about the solutions to random linear equations:

Lemma 5.1.2. Lemma 5.2 in Lehmann and Casella 1998 Suppose there are a set of p equations, $j=1,\dots,p$:

$$\sum_{k=1}^{p}A_{jkn}Y_{kn}=T_{jn}.$$

Let $T_{1n},\dots,T_{pn}$ converge in distribution to $T_1,\dots,T_p$. Furthermore, suppose that for each j,k, $A_{jkn}\xrightarrow{p}a_{jk}$ such that the matrix A with $(j,k)^{\text{th}}$ element a_{jk} is nonsingular. Then if the distribution of $T_1,\dots,T_p$ has a disitribution with repsect to the Lebesgue measure over $\mathbb{R}^p$, $Y_{1n},\dots,Y_{pn}$ tend in probability to $A^{-1}T$.

Thus by Lemma 5.1.2 we have that the solution, $\sqrt{n}(\hat{\theta}_n-\theta^\dagger)$ converges in probability to

$$\sqrt{n}(\hat{\theta}_n-\theta^\dagger)\xrightarrow{p}(-\mathbb{E}\left[\nabla_\theta^2 \log f_\theta(X_1)\right]_{[j,:]})^{-1}\mathcal{I}(\theta^\dagger)^{1/2}\mathcal{Z}$$

where $\mathcal{Z}\sim\text{Normal}(0,I_p)$, or

$$\sqrt{n}(\hat{\theta}_n-\theta^\dagger)\xrightarrow{d}\mathcal{N}\left(0,\mathbb{E}\left[-(\nabla_\theta^2 \log f_\theta(X_1))\mid_{\theta=\theta^\dagger}\right]^{-1}\mathcal{I}(\theta^\dagger)\left(\mathbb{E}\left[-(\nabla_\theta^2 \log f_\theta(X_1))\mid_{\theta=\theta^\dagger}\right]^{-1}\right)^T\right)$$

Assuming that the Fisher information is invertible

$$\begin{aligned}\mathbb{E}\left[-(\nabla_\theta^2 \log f_\theta(X_1))\mid_{\theta=\theta^\dagger}\right]^{-1}\mathbb{E}\left[(\nabla_\theta \log f_\theta(X_i))\mid_{\theta=\theta^\dagger}(\nabla_\theta \log f_\theta(X_i))\mid^T_{\theta=\theta^\dagger}\right]\left(\mathbb{E}\left[-(\nabla_\theta^2 \log f_\theta(X_1))\mid_{\theta=\theta^\dagger}\right]^{-1}\right)^T\\ =\mathcal{I}(\theta^\dagger)^{-1}\mathcal{I}(\theta^\dagger)\mathcal{I}(\theta^\dagger)^{-1}\\ =\mathcal{I}(\theta^\dagger)^{-1}\end{aligned}$$

Putting this all together shows that

$$\sqrt{n}(\hat{\theta}_n-\theta^\dagger)\xrightarrow{d}\mathcal{N}(0,\mathcal{I}(\theta^\dagger)^{-1})$$

Estimators of variance-covariance matrix

In the previous section, we encountered several consistent estimators of the variance covariance matrix:

$$-\frac{1}{n}\nabla^2_\theta \ell(\theta)\mid_{\theta=\theta^\dagger} \xrightarrow{p} \mathcal{I}(\theta^\dagger)$$
$$\frac{1}{n}\sum_{i=1}^n (\nabla_\theta \log f_\theta(X_i))\mid_{\theta=\theta^\dagger} (\nabla_\theta \log f_\theta(X_i))\mid^T_{\theta=\theta^\dagger} \xrightarrow{p} \mathcal{I}(\theta^\dagger)$$

These expressions assume that our inferential model matches the data generating model. In the event our inferential model is different than the true data generating model, it can be shown that the scaled MLE converges asymptotically to

$$\sqrt{n}(\hat{\theta}_n - \theta^\dagger) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[-(\nabla^2_\theta \log f_\theta(X_1))\mid_{\theta=\theta^\dagger}\right]^{-1} \mathbb{E}\left[(\nabla_\theta \log f_\theta(X_i))\mid_{\theta=\theta^\dagger} (\nabla_\theta \log f_\theta(X_i))\mid^T_{\theta=\theta^\dagger}\right] \mathbb{E}\left[-(\nabla^2_\theta \log f_\theta(X_1))\mid_{\theta=\theta^\dagger}\right]^{-1}\right)$$

where the key difference is that $\theta^\dagger$ is no longer the parameter for the true data generating process, but is instead the parameter the minimizes the KL divergence between the assumed inferential model and the true distribution generating the data.

Thus, the following sandwich estimator for the variance covariance matrix is often preferred over either of the above expressions:

$$\hat{\Sigma}_R = \left(-\frac{1}{n}\nabla^2_\theta \ell(\theta)\mid_{\theta=\hat{\theta}}\right)^{-1} \frac{1}{n}\sum_{i=1}^n (\nabla_\theta \log f_\theta(X_i))\mid_{\theta=\hat{\theta}} (\nabla_\theta \log f_\theta(X_i))\mid^T_{\theta=\hat{\theta}} \left(-\frac{1}{n}\nabla^2_\theta \ell(\theta)\mid_{\theta=\hat{\theta}}\right)^{-1} \tag{5.14}$$
$$\xrightarrow{p} \mathrm{Var}\left(\sqrt{n}(\hat{\theta}_n - \theta^\dagger)\right) \tag{5.15}$$

where $\hat{\theta}$ is the MLE.

5.1.1 Asymptotic confidence intervals

For the most part, we'll be concerned with univariate confidence intervals, but in multivariate models like the Weibull distribution we'll need to compute the full inverse of the Fisher information. WLOG, let the index of the parameter of interest be 1, so the asymptotic variance of our MLE for the parameter of interest is $\sigma_1^2(\theta^\dagger) = \mathcal{I}(\theta^\dagger)^{-1}_{1,1}$. We can also define

$$\sigma_1^2(\hat{\theta}) = \mathcal{I}(\hat{\theta})^{-1}_{1,1}.$$

I'll also ditch the n subscript and just let $\hat{\theta}$ be our MLE based on n observations. By **??**,

$$\frac{\sigma_1^2(\hat{\theta})}{\sigma_1^2(\theta^\dagger)} \xrightarrow{p} 1.$$

This allows us to use a plug-in estimator for $\mathcal{I}(\theta^\dagger)^{-1}$, $\mathcal{I}(\hat{\theta})^{-1}$.

$$\begin{aligned}\frac{\sqrt{n}(\hat{\theta}_1 - \theta_1^\dagger)}{\sigma_1(\hat{\theta})} &= \frac{\sigma_1(\theta^\dagger)}{\sigma_1(\hat{\theta})}\frac{\sqrt{n}(\hat{\theta}_1 - \theta_1^\dagger)}{\sigma_1(\theta^\dagger)} \\ &\xrightarrow{d} \mathcal{N}(0,1)\end{aligned}$$

Using **??**, we can create an asymptotic confidence interval by noting that:

$$P\left(\frac{\sqrt{n}(\hat{\theta}_1 - \theta_1^\dagger)}{\sigma_1(\hat{\theta})} \le x\right) = \Phi(x),$$

where $\Phi(x)$ is the CDF a normal distribution with zero mean and unit variance.

Then

$$P\left(\frac{\sqrt{n}(\hat{\theta}_1 - \theta_1^\dagger)}{\sigma_1(\hat{\theta}_1)} \in (-z_{1-\alpha/2}, z_{1-\alpha/2})\right) = P\left(\theta_1^\dagger \in \left(\hat{\theta}_1 - z_{1-\alpha/2}\frac{\sigma_1(\hat{\theta}_1)}{\sqrt{n}}, \hat{\theta}_1 + z_{1-\alpha/2}\frac{\sigma_1(\hat{\theta}_1)}{\sqrt{n}}\right)\right)$$

5.1.2 Asymptotic tests

Wald test

The Wald test is derived directly from the asymptotic distribution of the MLE. Under the null hypothesis $\theta^\dagger = \theta_0$, the test statistic:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$$

so

$$n(\hat{\theta}_n - \theta_0)^T \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0) \sim \chi^2(p)$$

This follows from the simple fact that if a random vector in $\mathbb{R}^n$, Z, is distributed multivariate normal, or $Z \sim \mathcal{N}(0,\Sigma)$, then $\Sigma^{-1/2}Z \sim \mathcal{N}(0,I)$, so $Z^T\Sigma^{-1/2}\Sigma^{-1/2}Z = \sum_{i=1}^n X_i^2$ where $X_i \sim \mathcal{N}(0,1)$.

Rao's score test

In our proof of the asymptotic distribution of the MLE, we used the fact that

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^n(\nabla_\theta \log f_\theta(X_i)) \mid_{\theta=\theta^\dagger} \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta^\dagger)).$$

This idea can be used to derive the Rao's Score test, which uses the fact that under $H_0 : \theta \in \Theta_0$, the gradient evaluated at the restricted MLE (i.e. the MLE restricted to the parameter space Θ_0) is nearly zero, and we can recover a similar limiting distribution. As above let

$$\frac{1}{\sqrt{n}}\nabla_\theta \ell(\theta) \mid_{\theta=\theta^\dagger} = \sqrt{n}\frac{1}{n}\sum_{i=1}^n(\nabla_\theta \log f_\theta(X_i)) \mid_{\theta=\theta^\dagger}$$

Assuming that under the null distribution the restrited MLE $\hat{\theta}_0$ is consistent for $\theta^\dagger \in \Theta_0$, then

$$\frac{1}{\sqrt{n}} \nabla_\theta \ell(\theta) \mid_{\theta=\hat{\theta}_0} \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta^\dagger))$$

The Score test statistic is:

$$T_S = \left(\frac{1}{\sqrt{n}} \nabla_\theta \ell(\theta) \mid_{\theta=\hat{\theta}_0} \right)^T \mathcal{I}(\hat{\theta}_0)^{-1} \frac{1}{\sqrt{n}} \nabla_\theta \ell(\theta) \mid_{\theta=\hat{\theta}_0}$$

This test statistic is distribution $\chi^2(p)$ under H_0.

Likelihood ratio test

The LRT comes from a two-term asymptotic expansion of the log-likelihood, as opposed to the one term expansion:

$$\begin{aligned}
-\ell(\theta_0) &= -\ell(\hat{\theta}) - \nabla_\theta \ell(\theta) \mid_{\theta=\hat{\theta}} (\hat{\theta} - \theta_0) - \frac{1}{2} (\hat{\theta} - \theta_0)^T \nabla_\theta^2 \ell(\theta) \mid_{\theta=\tilde{\theta}} (\hat{\theta} - \theta_0) \\
\ell(\hat{\theta}) - \ell(\theta_0) &= -\frac{1}{2} (\hat{\theta} - \theta_0)^T \nabla_\theta^2 \ell(\theta) \mid_{\theta=\tilde{\theta}} (\hat{\theta} - \theta_0) \\
&= \frac{1}{2} (\sqrt{n}(\hat{\theta} - \theta_0))^T \frac{-\nabla_\theta^2 \ell(\theta) \mid_{\theta=\tilde{\theta}}}{n} (\sqrt{n}(\hat{\theta} - \theta_0))
\end{aligned}$$

As before,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$$

and

$$-\frac{\nabla_\theta^2 \ell(\theta) \mid_{\theta=\tilde{\theta}}}{n} \xrightarrow{p} \mathcal{I}(\theta_0)$$

so

$$2(\ell(\hat{\theta}) - \ell(\theta_0)) \xrightarrow{d} \chi^2(p)$$

For all of the prior example, a convenient estimator for the Fisher information is the average of the *observed information.* The observed information is just the negative of the matrix of second derivatives of the log-likelihood:

$$-\nabla_\theta^2 \ell(\theta) = -\begin{bmatrix} \frac{\partial^2}{\partial^2\theta_1}\ell(\theta) & \frac{\partial^2}{\partial\theta_1\partial\theta_2}\ell(\theta) & \dots & \frac{\partial^2}{\partial\theta_1\partial\theta_p}\ell(\theta) \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1}\ell(\theta) & \frac{\partial^2}{\partial^2\theta_2}\ell(\theta) & \dots & \frac{\partial^2}{\partial\theta_2\partial\theta_p}\ell(\theta) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\theta_p\partial\theta_1}\ell(\theta) & \frac{\partial^2}{\partial\theta_p\partial\theta_2}\ell(\theta) & \dots & \frac{\partial^2}{\partial\theta_p\partial\theta_p}\ell(\theta) \end{bmatrix} \tag{5.16}$$

This is often denoted as

$$i(\theta) \equiv -\nabla_\theta^2 \ell(\theta).$$

Replacing $\ell(\theta) = \sum_i \log f_\theta(X_i)$ and using the fact that derivatives are linear operators:

$$i(\theta) = -\sum_i \begin{bmatrix} \frac{\partial^2}{\partial^2\theta_1}\log f_\theta(X_i) & \frac{\partial^2}{\partial\theta_1\partial\theta_2}\log f_\theta(X_i) & \dots & \frac{\partial^2}{\partial\theta_1\partial\theta_p}\log f_\theta(X_i) \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1}\log f_\theta(X_i) & \frac{\partial^2}{\partial^2\theta_2}\log f_\theta(X_i) & \dots & \frac{\partial^2}{\partial\theta_2\partial\theta_p}\log f_\theta(X_i) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\theta_p\partial\theta_1}\log f_\theta(X_i) & \frac{\partial^2}{\partial\theta_p\partial\theta_2}\log f_\theta(X_i) & \dots & \frac{\partial^2}{\partial\theta_p\partial\theta_p}\log f_\theta(X_i) \end{bmatrix} \tag{5.17}$$

we can see that the natural estimator of $\mathcal{I}(\theta)$ is the average observed information, which does indeed converge in probability to the Fisher information

$$\frac{1}{n} i(\theta) \xrightarrow{p} \mathcal{I}(\theta).$$

Of course, typically we won't know θ (unless we're evaluating $i(\theta)$ at θ_0), so we use the plug-in estimator, or $i(\hat{\theta}_n)$ which still converges in probability to the Fisher information:

$$\frac{1}{n} i(\hat{\theta}_n) \xrightarrow{p} \mathcal{I}(\theta).$$

5.1.3 Tests in terms of observed information

When we use observed information in place of the Fisher information, the Wald and Score tests look a bit different:

Wald test with the observed information

$$n(\hat{\theta}_n - \theta_0)^T \frac{1}{n} i(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) = (\hat{\theta}_n - \theta_0)^T i(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \overset{\text{asympt.}}{\sim} \chi^2(p)$$

Score test with the observed information

$$\begin{aligned}
T_S &= \left(\frac{1}{\sqrt{n}}\nabla_\theta \ell(\theta)\mid_{\theta=\hat{\theta}_0}\right)^T (\frac{1}{n}i(\hat{\theta}_0))^{-1}\frac{1}{\sqrt{n}}\nabla_\theta \ell(\theta)\mid_{\theta=\hat{\theta}_0} \\
&= \left(\frac{1}{\sqrt{n}}\nabla_\theta \ell(\theta)\mid_{\theta=\hat{\theta}_0}\right)^T n(i(\hat{\theta}_0))^{-1}\frac{1}{\sqrt{n}}\nabla_\theta \ell(\theta)\mid_{\theta=\hat{\theta}_0} \\
&= \left(\nabla_\theta \ell(\theta)\mid_{\theta=\hat{\theta}_0}\right)^T i(\hat{\theta}_0)^{-1}\nabla_\theta \ell(\theta)\mid_{\theta=\hat{\theta}_0}
\end{aligned}$$

5.1.4 Composite tests

This section is an expansion of Appendix B in Klein, Moeschberger, et al. 2003.

We can modify all of our tests to accommodate testing a subset of the parameters. Typically we'll have a subset of our parameter vector, let's call it ψ, that we're interested in, and we have another subset, ϕ, that are nuisance parameters. In the Example 5.0.1, we'll likely be interested in testing if $\beta \neq 0$, and thus we won't care about testing λ.

Let's let $\theta = (\psi, \phi)$, and let $\theta \in \mathbb{R}^p$ so $\psi \in \mathbb{R}^k$, $k < p$, $\phi \in \mathbb{R}^{p-k}$. Our null hypothesis will be:

$$H_0 : \psi = \psi_0.$$

Let $\hat{\phi}(\psi_0)$ be the MLE for the nuisance parameter with ψ fixed under the null hypothesis. We'll also partition the information matrix into a 2 by 2 block matrix:

$$\mathcal{I}(\psi,\phi) = \begin{bmatrix} \mathbb{E}\left[-\nabla^2_\psi \log f_\theta(X_1)\right] & \mathbb{E}\left[-\nabla^2_{\psi,\phi} \log f_\theta(X_1)\right] \\ \mathbb{E}\left[-\nabla^2_{\psi,\phi} \log f_\theta(X_1)\right] & \mathbb{E}\left[-\nabla^2_\phi \log f_\theta(X_1)\right] \end{bmatrix} = \begin{bmatrix} \mathcal{I}_{\psi,\psi} & \mathcal{I}_{\psi,\phi} \\ \mathcal{I}^T_{\psi,\phi} & \mathcal{I}_{\phi,\phi} \end{bmatrix}$$

The inverse can also be partitioned into a 2 by 2 block matrix:

$$\mathcal{I}(\psi,\phi)^{-1} = \begin{bmatrix} \mathcal{I}^{\psi,\psi} & \mathcal{I}^{\psi,\phi} \\ (\mathcal{I}^{\psi,\phi})^T & \mathcal{I}^{\phi,\phi} \end{bmatrix}$$

The expression for $\mathcal{I}^{\psi,\psi}$ can be found from the block matrix inversion formula:

$$\begin{aligned}
\mathcal{I}^{\psi,\psi} &= \mathcal{I}^{-1}_{\psi,\psi} + \mathcal{I}^{-1}_{\psi,\psi}\mathcal{I}_{\psi,\phi}\left(\mathcal{I}_{\phi,\phi} - \mathcal{I}^T_{\psi,\phi}\mathcal{I}^{-1}_{\psi,\psi}\mathcal{I}_{\psi,\phi}\right)^{-1}\mathcal{I}^T_{\psi,\phi}\mathcal{I}^{-1}_{\psi,\psi} \qquad (5.18) \\
&= \left(\mathcal{I}_{\psi,\psi} - \mathcal{I}_{\psi,\phi}\mathcal{I}^{-1}_{\phi,\phi}\mathcal{I}^T_{\psi,\phi}\right)^{-1} \qquad (5.19)
\end{aligned}$$

All of these results hold for the observed information, $i(\psi,\phi)$.

Composite Wald test

Again using normal distribution theory, we can derive the Wald test with the observed information:

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{\psi,\psi}).$$

The Wald test statistic is then:

$$T_W = \sqrt{n}(\hat{\psi}_n - \psi_0)^T \left(\mathcal{I}^{\psi,\psi} \mid_{\psi=\psi_0,\phi=\phi_0}\right)^{-1} (\hat{\psi}_n - \psi_0)\sqrt{n}$$

Using the appropriate transformation for the observed information in place of the Fisher information, we get

$$T_W = (\hat{\psi}_n - \psi_0)^T \left(i^{\psi,\psi} \mid_{\psi=\hat{\psi}_n,\phi=\hat{\pi}_n}\right)^{-1} (\hat{\psi}_n - \psi_0) \xrightarrow{d} \chi^2_k \tag{5.20}$$

Composite Score test

The composite score test is a bit more complicated. The joint asymptotic distribution of the score is:

$$\sqrt{n}\frac{1}{n}\nabla_{(\psi,\phi)}\ell(\psi,\phi) \mid_{\psi=\psi_0,\phi=\hat{\phi}(\psi_0)} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} \mathcal{I}_{\psi,\psi} & \mathcal{I}_{\psi,\phi} \\ \mathcal{I}^T_{\psi,\phi} & \mathcal{I}_{\phi,\phi} \end{bmatrix}\right)$$

But when we have a nuisance parameter, under the null distribution we solve the score equations

$$\nabla_\phi \ell(\psi_0, \phi) = 0,$$

leading to an MLE for ϕ, $\hat{\phi}(\psi_0)$, that is dependent on ψ_0. This means the distribution for $\sqrt{n}\frac{1}{n}\nabla_\psi \ell(\psi,\phi) \mid_{\psi=\psi_0,\phi=\hat{\phi}(\psi_0)}$ needs to condition on the score equations for ψ being zero. If the score equations are asymptotically normally distributed, then the score equations for ψ are conditionally normal. Recall that if vectors X, Y are multivariate normal with marginal variance covariance matrices Σ_X, Σ_Y and $\Sigma_{X,Y}$ is the covariance matrix of X with Y, then $X \mid Y$ is multivariate normal with parameters

$$\mathbb{E}[X] + \Sigma_{X,Y}\Sigma_Y^{-1}(Y - \mathbb{E}[Y]), \quad \Sigma_X - \Sigma_{X,Y}\Sigma_Y^{-1}\Sigma^T_{X,Y}.$$

In our case, the marginal mean of the score equations are zero, and $Y \equiv \nabla_\phi \ell(\psi_0, \hat{\phi}(\psi_0))$ is zero, so the conditional distribution of the score of ψ is

$$\sqrt{n}\frac{1}{n}\nabla_\psi \ell(\theta) \mid_{\psi=\psi_0,\phi=\hat{\phi}(\psi_0)} \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\psi,\psi} - \mathcal{I}_{\psi,\phi}\mathcal{I}^{-1}_{\phi,\phi}\mathcal{I}^T_{\phi,\psi}).$$

The test statistic is then

$$n^{-1/2}\nabla_\psi \ell(\theta) \mid_{\psi=\psi_0,\phi=\hat{\phi}(\psi_0)} \left(\mathcal{I}_{\psi,\psi} - \mathcal{I}_{\psi,\phi}\mathcal{I}^{-1}_{\phi,\phi}\mathcal{I}^T_{\phi,\psi}\right)^{-1} n^{-1/2}\nabla_\psi \ell(\theta) \mid_{\psi=\psi_0,\phi=\hat{\phi}(\psi_0)}$$

as we showed in Equation (5.19), the inverse matrix is the same as $\mathcal{I}^{\psi,\psi}$, so, subbing in our observed information matrix again, we get the final

$$T_S = \nabla_\psi \ell(\theta) \mid_{\psi=\psi_0,\phi=\hat{\phi}(\psi_0)} i(\psi_0, \hat{\phi}(\psi_0))^{\psi,\psi} \nabla_\psi \ell(\theta) \mid_{\psi=\psi_0,\phi=\hat{\phi}(\psi_0)}$$

which is asymptotically distributed as χ^2_k.

Composite likelihood ratio test

The composite likelihood ratio test is similar to the likelihood ratio test:

$$T_{LR} = 2(\ell(\hat{\psi}, \hat{\phi}) - \ell(\psi_0, \hat{\phi}(\psi_0)))$$

and this is again asymptotically distributed as χ^2_k

Example 5.1.1. Continued relative risk example Suppose we are interested in testing the hypothesis $H_0 : \beta = 0$ vs $H_a : \beta \neq 0$.

Recall the definitions of r_1, r_2, T_1, T_2:

$$\begin{aligned} r_1 &= \sum_{i=1}^{n}(1 - z_i)\delta_i & T_1 &= \sum_{i=1}^{n}(1 - z_i)t_i \\ r_2 &= \sum_{i=1}^{n} z_i\delta_i & T_2 &= \sum_{i=1}^{n} z_i t_i \end{aligned}$$

We showed in Example 5.0.1 that the log-likelihood was:

$$\ell(\lambda, \beta) = (r_1 + r_2)\log\lambda - \lambda T_1 + r_2\beta - \lambda e^{\beta} T_2 \tag{5.21}$$

The score equations are

$$\begin{aligned} \frac{\partial}{\partial\lambda}\ell(\lambda, \beta) &: \frac{r_1 + r_2}{\lambda} - T_1 - e^{\beta} T_2 \\ \frac{\partial}{\partial\beta}\ell(\lambda, \beta) &: r_2 - \lambda e^{\beta} T_2 \end{aligned}$$

and the matrix of second derivatives of the log-likelihood with respect to λ, β, also known as the *observed information*, is

$$\nabla^2_{\lambda,\beta}\ell(\lambda, \beta) = \begin{bmatrix} \frac{r_1+r_2}{\lambda^2} & e^{\beta} T_2 \\ e^{\beta} T_2 & \lambda e^{\beta} T_2 \end{bmatrix} \tag{5.22}$$

The unrestricted MLE, (i.e. the MLE under the alternative hypothesis), is:

$$\begin{aligned} \hat{\lambda} &= \frac{r_1}{T_1} \\ e^{\hat{\beta}} &= \frac{r_2}{T_2}\frac{T_1}{r_1} \end{aligned}$$

Under the null hypothesis that $\beta = 0$, we have the restricted likelihood:

$$\ell(\lambda, \beta = 0) = (r_1 + r_2)\log\lambda - \lambda T_1 - \lambda T_2 \tag{5.23}$$

which can be differentiated with respect to λ, set to zero, and solved for λ:

$$\hat{\lambda}_0 = \frac{r_1 + r_2}{T_1 + T_2} \tag{5.24}$$

The inverse of the observed information evaluated at the unrestricted MLE was shown to be

$$\frac{r_1 + r_2}{r_1 r_2} \tag{5.25}$$

The inverse of the observed information is:

$$\hat{\mathcal{I}}^{-1}(\lambda, \beta) = \frac{1}{\frac{(r_1+r_2)e^{\beta}T_2}{\lambda} - e^{2\beta}T_2^2} \begin{bmatrix} \lambda e^{\beta} T_2 & -e^{\beta} T_2 \\ -e^{\beta} T_2 & \frac{r_1+r_2}{\lambda^2} \end{bmatrix} \tag{5.26}$$

which when the 2, 2 element is evaluated at the $\hat{\lambda}_0$, or

$$\hat{\mathcal{I}}^{-1}(\hat{\lambda}_0, 0)_{2,2} = \frac{(T_1 + T_2)^2}{(r_1 + r_2)T_1 T_2}$$

Now for the test statistics:

- **Likelihood ratio test**: After some algebra, we get

$$T_{LR} = 2r_1\left(\log\left(\frac{r_1}{T_1}\right) - \log\left(\frac{r_1 + r_2}{T_1 + T_2}\right)\right) + 2r_2\left(\log\left(\frac{r_2}{T_2}\right) - \log\left(\frac{r_1 + r_2}{T_1 + T_2}\right)\right)$$

- **Wald test**: The test statistic is:

$$T_W = \left(\log\frac{r_2/T_2}{r_1/T_1}\right)^2 \frac{r_1 r_2}{r_1 + r_2}.$$

- **Score test** The starting test statistic is:

$$T_S = \left(r_2 - (r_1 + r_2)\frac{T_2}{T_1 + T_2}\right)^2 \frac{(T_1 + T_2)^2}{(r_1 + r_2)T_1 T_2}.$$

  This is sort of interesting because it looks a bit like the log-rank statistic! $\frac{T_2}{T_1+T_2}$ is a bit like the proportion of time at risk the second group experienced, and the expected total failures in the second group is this proportion multiplied by the total failures in both groups. It's not too hard to see why you might want to reject the null that $\beta = 0$ if this statistic were large. This simplifies to

$$T_S = \frac{(T_1 r_2 - T_2 r_1)^2}{(r_1 + r_2)T_1 T_2}.$$

For an observed dataset of $r_1 = 10, r_2 = 12, T_1 = 25, T_2 = 27$, they all yield values around 0.06, which is far below the critical value of 3.84, which is the 95$^{\text{th}}$ quantile from a χ^2_1.

5.2 More on parametric regression models

Information is from Collett 1994, Harrell et al. 2001, O. O. Aalen 1988, O. Aalen et al. 2008.

5.3 Weibull regression

A common parametric proportional hazards model is the Weibull, which we encountered way back in lecture 2. The baseline hazard has functional form:

$$\lambda_0(t \mid \alpha, \gamma) = \gamma \alpha t^{\alpha - 1}.$$

so the full regression model has the form

$$\lambda_i(t \mid \alpha, \gamma, \boldsymbol{\beta}) = \gamma \alpha t^{\alpha - 1} \exp(\mathbf{z}_i^T \boldsymbol{\beta}),$$

with survival function:

$$S(t) = \exp(-\gamma t^{\alpha} \exp(\mathbf{z}_i^T \boldsymbol{\beta}))$$

The interesting thing about the Weibull is that it isn't just a parametric model for survival time; it can be justified using extreme value theory as the minimum of iid nonnegative random variables. Aalen writes in O. O. Aalen 1988:

> Hence, if cancer may result from one of the first cells to undergo malignant transformation, then the time to appearance of cancer might very well follow a Weibull distribution, when lime is measured from an appropriate point. This principle has more general validity. An individual is subject to the risk of several different causes of death and the one which first causes fatality determines the life time. Hence the life time might be supposed to follow an extreme distribution for each individual.

Model fit check

For any survival model the following identity holds:

$$S^{-1}(S(t)) = t.$$

Thus an effective model check is to use a nonparametric estimate of the survival function, either $\hat{S}^{\mathrm{KM}}(t)$ or $\hat{S}^{\mathrm{NA}}(t)$, apply the parametric form of S_θ^{-1} to the nomnparamtetric survival function estimate, and to plot this function against t. The graph should be roughly linear in t.

Example 5.3.1. Weibull model check Assuming $X_i \sim \text{Weibull}(\gamma, \alpha)$, the survival function is

$$S(t) = \exp(-\gamma t^{\alpha}).$$

The inverse function is found as follows:

$$\begin{aligned} p &= \exp(-\gamma t^{\alpha}) \\ -\log p &= \gamma t^{\alpha} \\ (\frac{-\log p}{\gamma})^{1/\alpha} &= t \end{aligned}$$

Then we can check the following plot: Under noninformative sampling with observed data $(t_i, d_i), i = 1, \ldots, n$, $\hat{S}^{\text{KM}}(t) = \prod_{i|t_i \le t}(1 - \frac{d_i}{\overline{Y}(t)})$ is the nonparametric estimator of the survival function. a plot of

$$(\frac{-\log \hat{S}^{\text{KM}}(t)}{\gamma})^{1/\alpha} \text{v.s.} t$$

should be roughly linear.

Another implication in the Weibull distribution case case is the following:

$$S(t) = \exp(-\gamma t^{\alpha}) \implies \log(-\log p) = \log(\gamma) + \alpha \log(t).$$

This leads to an alternative way to do a model check:

$$\log(-\log \hat{S}^{\text{KM}}(t)) \text{v.s. } \log(t)$$

should be roughly linear with slope α.

5.3.1 Parametric proportional hazards models

Recall our definition of proportional hazards employing an exponential function with $\mathbf{z}_i \in \mathbb{R}^k$:

$$\lambda(t \mid \mathbf{z}_i) = \lambda_0(t \mid \boldsymbol{\theta}) \exp(\boldsymbol{\beta}^T \mathbf{z}_i) \tag{5.27}$$

This implies the following properties for our model:

$$\log \lambda(t \mid \mathbf{z}_i) = \log \lambda_0(t \mid \boldsymbol{\theta}) + \boldsymbol{\beta}^T \mathbf{z}_i \tag{5.28}$$

$$\log \Lambda(t \mid \mathbf{z}_i) = \log \Lambda_0(t \mid \boldsymbol{\theta}) + \boldsymbol{\beta}^T \mathbf{z}_i \tag{5.29}$$

This means that the predictors act linearly on the log scale for both the hazard ratio and the cumulative hazard, and that the effect of the predictors is constant over time.

The interpretation of coefficients is as the change in the log hazard, or log cumulative hazard:

$$\beta_j = \log\lambda(t \mid z_1, \dots z_{j-1}, z_j + 1, z_{j+1}, \dots z_k) - \log\lambda(t \mid z_1, \dots z_{j-1}, z_j, z_{j+1}, \dots z_k).$$

Alternatively, we have

$$e^{\beta_j} = \frac{\lambda(t \mid z_1, \dots z_{j-1}, z_j + 1, z_{j+1}, \dots z_k)}{\lambda(t \mid z_1, \dots z_{j-1}, z_j, z_{j+1}, \dots z_k)}.$$

Increasing z_j by 1 has the effect of increasing the hazard of an event by e^{β_j}.

As discussed previously and shown in Figure 5.1, When we have a single categorical predictor, we can assess the validity of proportional hazards by plotting the $\log(-\log)$ of the KM estimate of survival within each subgroup, and determining if the lines are roughly linear in $\log t$ and if they are parallel. If they are not parallel, but are straight, this may be an indication that one could fit separate the groups with separate shape, or α, parameters.

5.3.2 Testing for proportional hazards

Following Collett 1994, in the Weibull model we may test the proportional hazards assumption by fitting a more flexible model and using a composite likelihood ratio test. Suppose we have patients categorized into 3 age groups, and we use dummy coding for our design matrix:

Group	**Predictors**
Youngest group	$\mathbf{z}_i = (0,0)^T$
Middle group	$\mathbf{z}_i = (1,0)^T$
Oldest group	$\mathbf{z}_i = (0,1)^T$

and we want to test whether fitting the following proportional hazards Weibull regression model:

$$X_i \sim \text{Weibull}(\gamma e^{\boldsymbol{\beta}^T \mathbf{z}_i}, \alpha)$$

is sufficient. An alternative model that allows for hazards that are not proportional is

$$X_i \sim \text{Weibull}(\gamma e^{\boldsymbol{\beta}^T \mathbf{z}_i}, \alpha e^{\boldsymbol{\theta}^T \mathbf{z}_i})$$

Note that this alternative model is equivalent to fitting separate Weibull models to each group. Then the null hypothesis we'd like to test is whether $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = 0$. We can use the composite likelihood ratio test to determine whether the data contradict this null hypothesis. The test statistic would be distributed as χ^2_2 given the constraints in the null hypothesis.

The test statistic in the case where we fit separate models to each subgroup is

$$2(\ell_1(\hat{\psi}_1, \hat{\phi}_1) + \ell_2(\hat{\psi}_2, \hat{\phi}_2) + \ell_3(\hat{\psi}_3, \hat{\phi}_3) - \ell(\psi_0, \hat{\phi}(\psi_0)))$$

where $\ell_j(\hat{\psi}_j, \hat{\phi}_j), j = 1, 2, 3$ is the log-likelihood from the fitted Weibull model to each age group.

5.3.3 Accelerated failure time formulation

There is an alternative way to specify the Weibull model, wherein we model the log of the survival times as being a linear function of covariates.

$$\log(X_i) = \mu + \mathbf{z}_i^T \boldsymbol{\eta} + \sigma \epsilon_i$$

Let ϵ_i be Gumbel distributed with a probability density function

$$f(\epsilon) = \exp(\epsilon - e^{\epsilon})$$

If we let $\nu = e^{\epsilon}$, then we can compute the density over ν. $\epsilon(\nu) = \log(\nu)$. $f(\nu) = f(\epsilon(\nu)) \frac{d}{d\nu} \epsilon(\nu)$

$$\exp(\log(\nu) - e^{\log(\nu)})/\nu = e^{-\nu} \tag{5.30}$$

This shows that $e^{\epsilon} \sim$ Exponential(1). Now we can write the survival function of X_i:

$$\begin{aligned}
S(t) &= P(X_i > t) \\
&= P(\log(X_i) > \log(t)) \\
&= P(\mu + \mathbf{z}_i^T \boldsymbol{\eta} + \sigma \epsilon_i > \log(t)) \\
&= P(\epsilon_i > (\log(t) - \mu - \mathbf{z}_i^T \boldsymbol{\eta})/\sigma) \\
&= P(e^{\epsilon_i} > \exp(\log(t) - \mu - \mathbf{z}_i^T \boldsymbol{\eta})^{1/\sigma}) \\
&= \exp\left(-\exp(\log(t) - \mu - \mathbf{z}_i^T \boldsymbol{\eta})^{1/\sigma}\right) \\
&= \exp\left(-e^{-\mu/\sigma} t^{1/\sigma} \exp(\mathbf{z}_i^T(-\boldsymbol{\eta}/\sigma))\right)
\end{aligned}$$

Recall the survival function of a Weibull with hazard function $\gamma \alpha t^{\alpha-1} \exp(\mathbf{z}_i^T \boldsymbol{\beta})$:

$$S(t) = \exp(-\gamma t^{\alpha} \exp(\mathbf{z}_i^T \boldsymbol{\beta})).$$

Then there are the following correspondences between our parameters for the log-linear model and the original proportional hazards model:

$$\begin{aligned}
\alpha &= \frac{1}{\sigma} \\
\gamma &= e^{-\mu/\sigma} \\
\boldsymbol{\beta} &= -\boldsymbol{\eta}/\sigma
\end{aligned}$$

In general, the correspondence between the model for the log-failure time and the proportional hazards will not hold, but it does in the Weibull model.

5.4 AFT models

This information is from Chapter 12 in Klein, Moeschberger, et al. 2003. Generally, AFT models are specified by modeling the survival function as follows:

$$\begin{aligned} S(t \mid \mathbf{z}) &= S_0(t \exp(\boldsymbol{\theta}^T \mathbf{z})) \\ &= P(X_i > t \exp(\boldsymbol{\theta}^T \mathbf{z})) \\ &= P(\frac{X_i}{\exp(\boldsymbol{\theta}^T \mathbf{z})} > t) \end{aligned}$$

where S_0 is the survival function for an individual with $\mathbf{z} = \mathbf{0}$. Thus, we take a population model for S, S_0, and for an individual with covariates $\mathbf{z}$, and $\exp(\boldsymbol{\theta}^T \mathbf{z}) > 1$, survival time is shrunk towards zero. We might also say that for an individual with $\exp(\boldsymbol{\theta}^T \mathbf{z})$, their probability of survival at time t is as if they were an individual with a survival function evaluated at $t_0 = t \exp(\boldsymbol{\theta}^T \mathbf{z})$. Recall that the survival function and the hazard function are related via the following equation:

$$-\frac{\partial}{\partial t} \log(S(t)) = \lambda(t).$$

Note that when $S(t) = S(g(t))$ for a known differentiable function $g(t)$, the following will hold:

$$-\frac{\partial}{\partial t} \log S(g(t)) = -\left(\frac{\partial}{\partial g} \log(S(g))\right) |_{g=g(t)} \frac{\partial}{\partial t} g(t) \implies -\frac{\partial}{\partial t} \log S(g(t)) = \lambda(g(t)) \frac{\partial}{\partial t} g(t) \tag{5.31}$$

When we use an AFT model for X_i, this implies the following about the hazard rate, using the result in Equation (5.31):

$$-\frac{\partial}{\partial t} \log S(t \mid \mathbf{z}) = \exp(\boldsymbol{\theta}^T \mathbf{z}) \lambda_0(t \exp(\boldsymbol{\theta}^T \mathbf{z})) \tag{5.32}$$

Of course, sometimes this corresponds to a proportional hazards model, as in the Weibull case, but most times it does not.

For the Weibull, recall that $\lambda_0(t) = \gamma \alpha t^{\alpha-1}$ so writing the hazard as above would lead to:

$$\exp(\boldsymbol{\theta}^T \mathbf{z}) \lambda_0(t \exp(\boldsymbol{\theta}^T \mathbf{z})) = \exp(\boldsymbol{\theta}^T \mathbf{z}) \gamma \alpha (t \exp(\boldsymbol{\theta}^T \mathbf{z}))^{\alpha-1} \tag{5.33}$$

$$= \exp(\boldsymbol{\theta}^T \mathbf{z})^{\alpha} \gamma \alpha t^{\alpha-1} \tag{5.34}$$

This formulation allows us to write $\log(X_i)$ as a linear model:

$$\log(X_i) = \mu + \mathbf{z}_i^T \boldsymbol{\eta} + \sigma \epsilon_i.$$

Note that $-\boldsymbol{\theta} = \boldsymbol{\eta}$. The distribution of ϵ_i is a modeling choice. We saw that the extreme value distribution is equivalent to the Weibull proportional hazards regression. Any distribution over $\mathbb{R}$ will work, though common choices are normally distributed ϵ_i, leading to $X_i \sim \text{LogNormal}$, and log-logistic distributed ϵ_i.

The log-logistic model uses the following density for ϵ_i:

$$f_\epsilon(x) = \frac{e^x}{(1 + e^x)^2}, \tag{5.35}$$

which leads to survival function of:

$$S(t) = \frac{1}{1 + \lambda t^\alpha} \tag{5.36}$$

$$\Lambda(t) = -\log(S(t)) \tag{5.37}$$

$$= \log(1 + \lambda t^\alpha) \tag{5.38}$$

The log-logistic model has the unique property that the odds of survival for an individual at time t are proportional to the odds of survival for the base population:

$$\frac{S(t \mid \mathbf{z})}{1 - S(t \mid \mathbf{z})} = \exp(\boldsymbol{\beta}^T \mathbf{z}) \frac{S_0(t)}{1 - S_0(t)}$$

where $\boldsymbol{\beta} = -\boldsymbol{\gamma}\sigma$.

Of course, we can't just fit these models to the log of the observed failure times because we have censoring. Thus we'll need to do numerical maximum likelihood as we did for other survival models.

5.4.1 Model checking in AFT models

The relationships that held for the Weibull regressions can be ported to other AFT models. Klein, Moeschberger, et al. 2003 suggest checking a function of the cumulative hazard against a function of t to assess adequacy of model fit. We can use the (tie-corrected) Nelson-Aalen estimator of the cumulative hazard function:

$$\hat{\Lambda}^{\text{NA}}(t) = \sum_{i \mid t_i \leq t} \frac{d_i}{\overline{Y}(t)}$$

and examine transformations thereof against appropriate transformations of t.

For the log-logistic model, $\Lambda(t) = \log(1 + \lambda t^\alpha)$. This implies that

$$\log(\exp(\hat{\Lambda}^{\text{NA}}(t)) - 1) \approx \log \lambda + \alpha \log t$$

We can compute similar expressions for the Weibull and the log-normal model.

5.4.2 Cox-Snell residuals

Recall from Section 2.6 that the following relationship holds: When $X_i \sim F$ with cumulative hazard function $\Lambda(t)$

$$\Lambda(X_i) \sim \text{Exp}(1).$$

We can use this idea to generate graphical checks for our models.

Continuing with the log-logistic model, we could graphically assess whether the following Cox-Snell residual, denoted r_i^C:

$$r_i^C = \log(1 + e^{\mathbf{z}_i^T \hat{\boldsymbol{\theta}}} \hat{\lambda} t_i^{\hat{\alpha}})$$

is exponentially distributed with unit rate. The issue with plotting these residuals directly against the quantiles of an exponential distribution is that for the censored observations, $\Lambda(C_i)$ won't be exponentially distributed. But we can use the properties of the cumulative hazard function to our advantage, namely that it is nondecreasing in t. Thus for censored observations where $t_i = c_i$, this implies that $x_i \geq t_i$. Thus, $\Lambda(t_i) \leq \Lambda(x_i)$, so we can say that when $\delta_i = 0$, $\Lambda(x_i)$ is censored at $\Lambda(t_i)$.

The solution is to use the Kaplan-Meier estimator again! We can form the censored cumulative hazard sample:

$$\{(\tilde{t}_i = \min(\Lambda(x_i), \Lambda(c_i)), \delta_i = \mathbb{1}(x_i \leq c_i)), i = 1, \dots, n\} = \tag{5.39}$$

$$\{(\tilde{t}_i = \Lambda(t_i), \delta_i = \mathbb{1}(x_i \leq c_i)), i = 1, \dots, n\} \tag{5.40}$$

where the second line follows from the nondecreasing characteristic of $\Lambda(t)$.

Then we can fit the Kaplan Meier estimator to the dataset $(\tilde{t}_i, \delta_i)$ observations to infer the non-censored distribution of $\Lambda(x_i)$. The procedure is as outlined below:

1. Fit a parametric survival model to $\{(t_i, \delta_i, \mathbf{z}_i), i = 1, \dots, n\}$

2. Calculate the Cox-Snell residuals using the estimated survival model: $\{(\tilde{t}_i = \hat{\Lambda}(t_i), \delta_i = \mathbb{1}(x_i \leq c_i)), i = 1, \dots, n\}$

3. Fit a Kaplan-Meier estimator to the datatset Equation (5.39)

4. Plot the $\log(-\log(\hat{S}^{\text{KM}}(t)))$ vs. $\log t$ to see whether a line with zero intercept and slope 1 fits in the confidence intervals

5.4.3 Influence of data points in likelihood equations

The material in this section is from Collett 1994, Cain and Lange 1984, and Broderick et al. 2023. Like in linear regression, we'd like to determine if some of our data points are influencing our conclusions; armed with this information, perhaps we can expand the model to incorporate these outliers, or perhaps there is a data processing error that we can rectify and re-run our analysis.

One idea is to determine whether omitting one data point appreciably changes our estimate of our parameter of interest. The simplest way to do this is to refit the data n-times, where each time we omit one data point. For small datasets, this is reasonable, but when we have large n, or a very complex model, it may be infeasible to refit the model n times.

Instead, we can cleverly use Taylor expansions to approximate the effect of small perturbations in the data on the estimated coefficient. If these small perturbations induce large changes in our estimated coefficients, then it stands to reason that the datapoints that have been perturbed are influential to our estimates.

Let's make things more concrete. Suppose we have a model with a parameter vector, $\boldsymbol{\theta} \in \mathbb{R}^k$, and a maximum likelihood estimate thereof $\hat{\boldsymbol{\theta}}$. We'd like to understand how $\hat{\boldsymbol{\theta}}$ changes if we drop one datapoint. Let the index of this datapoint be j. We can formalize the idea of dropping a datapoint by examining the score equations. Recall our typical problem setup: We have n observations, each of which is a triplet of the time to failure or the time to censoring, t_i, an indicator δ_i that t_i is the time to failure, and $\mathbf{z}_i \in \mathbb{R}^k$, the covariate vector associated with each unit. Let our likelihood for each observation be $f_{\boldsymbol{\theta}}(t_i, \delta_i, \mathbf{z}_i)$ Let

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(t_i, \delta_i, \mathbf{z}_i).$$

The score equations are defined as usual:

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_i, \delta_i, \mathbf{z}_i) \tag{5.41}$$

and $\hat{\boldsymbol{\theta}}$ is the solution to the set of equations $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) \,|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}$

We can introduce the variables w_i into the equation above, as well as the collection of the w_i into the vector $\mathbf{w}$:

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \mathbf{w}) = \sum_{i=1}^{n} w_i \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_i, \delta_i, \mathbf{z}_i) \tag{5.42}$$

The vector $\hat{\boldsymbol{\theta}}(\mathbf{w})$ solves the equations

$$\sum_{i=1}^{n} w_i \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_i, \delta_i, \mathbf{z}_i) \,|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{w})} = \mathbf{0} \tag{5.43}$$

Note that our original MLE, $\hat{\boldsymbol{\theta}} \equiv \hat{\boldsymbol{\theta}}(\mathbf{1})$. Deleting the j^{th} datapoint amounts to setting $w_j = 0$.

Then the idea is to approximate $\hat{\boldsymbol{\theta}}(\mathbf{w})$ near the vector $\mathbf{1}$. For the m^{th} element of $\hat{\boldsymbol{\theta}}(\mathbf{w})$, $\hat{\boldsymbol{\theta}}(\mathbf{w})_m$, this is:

$$\hat{\boldsymbol{\theta}}(\mathbf{w})_m \approx \hat{\boldsymbol{\theta}}(\mathbf{1})_m + \sum_{i=1}^{n}(w_i - 1)\left(\frac{\partial}{\partial w_i}\hat{\boldsymbol{\theta}}(\mathbf{w})_m\right)\Bigg|_{\mathbf{w}=\mathbf{1}} \tag{5.44}$$

When all but one of these w_i is equal to 1, namely $w_j = 0$, let

$$\mathbf{w}_{(j)} = (\mathbf{1}_{j-1}^T, 0, \mathbf{1}_{n-j}^T)^T.$$

Then we get

$$\hat{\boldsymbol{\theta}}(\mathbf{w}_{(j)})_m \approx \hat{\boldsymbol{\theta}}(\mathbf{1})_m - \left(\frac{\partial}{\partial w_j}\hat{\boldsymbol{\theta}}(\mathbf{w})_m\right)\Bigg|_{\mathbf{w}=\mathbf{1}} \tag{5.45}$$

Thus for the whole vector $\hat{\boldsymbol{\theta}}(\mathbf{w}_{(j)})$ we get, as in Cain and Lange 1984,

$$\hat{\boldsymbol{\theta}}(\mathbf{w}_{(j)}) \approx \hat{\boldsymbol{\theta}}(\mathbf{1}) - \left(\frac{\partial}{\partial w_j}\hat{\boldsymbol{\theta}}(\mathbf{w})\right)\Bigg|_{\mathbf{w}=\mathbf{1}} \tag{5.46}$$

where

$$\frac{\partial}{\partial w_j}\hat{\boldsymbol{\theta}}(\mathbf{w}) = (\frac{\partial}{\partial w_j}\hat{\boldsymbol{\theta}}(\mathbf{w})_1, \ldots, \frac{\partial}{\partial w_j}\hat{\boldsymbol{\theta}}(\mathbf{w})_k)^T.$$

The question remains how to calculate $\frac{\partial}{\partial w_j}\hat{\boldsymbol{\theta}}(\mathbf{w})$ evaluated at $\mathbf{w} = 1$?

Let the vector $\mathbf{U}(\boldsymbol{\theta}, \mathbf{w})$ be defined as

$$\mathbf{U}(\boldsymbol{\theta}, \mathbf{w}) = \sum_{i=1}^{n} w_i \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_i, \delta_i, \mathbf{z}_i).$$

Note that the score equations are a function of the parameter vector and the vector of weights. The MLE given a set of weights $\mathbf{w}$, $\hat{\boldsymbol{\theta}}(\mathbf{w})$ solves the system of equations:

$$\mathbf{U}(\hat{\boldsymbol{\theta}}(\mathbf{w}), \mathbf{w}) = \mathbf{0}.$$

Then the implicit function theorem (more detail here?) allows us to differentiate the expression above with respect to w_j and solve for the derivative of interest, $\frac{\partial}{\partial w_j}\hat{\boldsymbol{\theta}}(\mathbf{w})$.

Recalling the chain rule for multivariate functions: Let $v(x(t), y(t))$ and calculate $\frac{\partial}{\partial t}v(x(t), y(t))$:

$$\frac{\partial}{\partial t}v(x(t), y(t)) = \frac{\partial v(x,y)}{\partial x}\bigg|_{x=x(t),y=y(t)} \frac{\partial x(u)}{\partial u}\bigg|_{u=t} + \frac{\partial v(x,y)}{\partial y}\bigg|_{x=x(t),y=y(t)} \frac{\partial y(u)}{\partial u}\bigg|_{u=t}.$$

We can differentiate the expression for the score function:

$$\frac{\partial}{\partial w_j}\mathbf{U}(\hat{\boldsymbol{\theta}}(\mathbf{w}), \mathbf{w}) = \frac{\partial}{\partial w_j}\mathbf{0} \implies \frac{\partial \mathbf{U}(\boldsymbol{\theta}, \mathbf{w})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}} \frac{\hat{\boldsymbol{\theta}}(\mathbf{w})}{\partial w_j}\bigg|_{\mathbf{w}=\mathbf{1}} + \frac{\mathbf{U}(\boldsymbol{\theta}, \mathbf{w})}{\partial w_j}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}} = \mathbf{0}$$

Assuming that

$$\left.\frac{\partial \mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}$$

is invertible, which is equivalent to requiring that the observed information matrix evaluated at the MLE with the complete data be invertible, we can solve the equation for the quantity of interest, $\frac{\partial}{\partial w_j}\hat{\boldsymbol{\theta}}(\mathbf{w})$ evaluated at $\mathbf{w} = 1$.

$$\left.\frac{\hat{\boldsymbol{\theta}}(\mathbf{w})}{\partial w_j}\right|_{\mathbf{w}=\mathbf{1}} = \left(-\left.\frac{\partial \mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}\right)^{-1} \left.\frac{\mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial w_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}$$

Now we need to evaluate $\left.\frac{\mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial w_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}$.

$$\begin{aligned}\frac{\partial}{\partial w_j}\mathbf{U}(\boldsymbol{\theta},\mathbf{w}) &= \frac{\partial}{\partial w_j}\left(\sum_{i=1}^{n} w_i \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_i,\delta_i,\mathbf{z}_i)\right)\\ &= \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_j,\delta_j,\mathbf{z}_j)\end{aligned}$$

so

$$\left.\frac{\mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial w_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}} = \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_j,\delta_j,\mathbf{z}_j)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}\,.$$

Finally, we get the general equation for the sensitivity of the MLE to the deletion of the j^{th} data point:

$$\hat{\boldsymbol{\theta}}(\mathbf{1}) - \hat{\boldsymbol{\theta}}(\mathbf{w}_{(j)}) \approx \left(-\left.\frac{\partial \mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}\right)^{-1} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_j,\delta_j,\mathbf{z}_j)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}} \tag{5.47}$$

This makes a good bit of sense; if the gradient of the log-likelihood function at a point lies along a direction of large uncertainty, this datapoint will have a large influence on the MLE.

The expression in Equation (5.47) also makes sense when viewed through the lens of the limiting distribution for the MLE. Note that

$$-\left.\frac{\partial \mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}} = -\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})}$$

Recall that from a previous lecture we have that a Taylor expansion for $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$

$$\sqrt{n}(\hat{\boldsymbol{\theta}}(\mathbf{1}) - \boldsymbol{\theta}^{\dagger}) = \left(-\frac{1}{n}\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}(\mathbf{1})}\right)^{-1}\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\dagger}}$$

Using the Taylor expansion formula with remainders yields

$$\sqrt{n}(\hat{\boldsymbol{\theta}}(\mathbf{1}) - \boldsymbol{\theta}^{\dagger}) = \left(-\frac{1}{n}\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\dagger}}\right)^{-1}\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\dagger}} + o_p(1)$$

The plug-in estimator for the right-hand side at $\boldsymbol{\theta}^\dagger = \hat{\boldsymbol{\theta}}(\mathbf{1})$ yields:

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}}(\mathbf{1}) - \boldsymbol{\theta}^\dagger) &= \left(-\frac{1}{n}\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})}\right)^{-1}\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})} + o_p(1)\\ &= \left(-\frac{1}{n}\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}}\log(f_{\boldsymbol{\theta}}(t_i,\delta_i,\mathbf{z}_i))\mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})} + o_p(1)\end{aligned}$$

Dividing each side by $\sqrt{n}$ yields

$$\hat{\boldsymbol{\theta}}(\mathbf{1}) - \boldsymbol{\theta}^\dagger = \left(-\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})}\right)^{-1}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}}\log(f_{\boldsymbol{\theta}}(t_i,\delta_i,\mathbf{z}_i))\mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})} + o_p(1/\sqrt{n})$$

Thus, asymptotically, each observation $(t_i, \delta_i, \mathbf{z}_i)$ perturbs the deviation between the MLE and the true value by approximately:

$$\left(-\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})}\right)^{-1}\nabla_{\boldsymbol{\theta}}\log(f_{\boldsymbol{\theta}}(t_i,\delta_i,\mathbf{z}_i))\mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})}.$$

Linear regression

In the case of the linear regression model with normally distributed errors and known variance, we have the following results:

$$\log f_{\boldsymbol{\theta}}(t_j, \delta_j, \mathbf{z}_j)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}} = \mathbf{z}_i(y_i - \hat{\boldsymbol{\beta}}\mathbf{z}_i)$$

and

$$\left(-\left.\frac{\partial\mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}\right)^{-1} = -(\mathbf{Z}^T\mathbf{Z})^{-1}.$$

Assuming that $\mathbf{Z}$ is full column rank, we can decompose the variance covariance matrix as:

$$-(\mathbf{Z}^T\mathbf{Z})^{-1} = -\mathbf{Q}A\mathbf{Q}^T$$

where $\mathbf{Q}$ is a matrix with the orthonormal eigenvectors of $\mathbf{Z}^T\mathbf{Z}$ as columns. Thus the influence of the j^{th} datapoint is

$$-\mathbf{Q}A\mathbf{Q}^T\mathbf{z}_i(y_i - \hat{\boldsymbol{\beta}}\mathbf{z}_i).$$

If $\mathbf{z}_i$ lies in a direction of large uncertainty for the variance-covariance matrix (i.e. the vector is aligned with the eigenvector associated with a large eigenvalue), *and* there is a large fitted residual, the datapoint will have a lot of influence on at least one of the coefficients.

Compare this to the exact calculation of the influence on $\hat{\beta}$ from removing one point in a linear regression model:

$$\hat{\beta}(\mathbf{w}) - \hat{\beta}(\mathbf{1}) = \frac{-\mathbf{Q}A\mathbf{Q}^T\mathbf{z}_i(y_i - \hat{\boldsymbol{\beta}}\mathbf{z}_i)}{1 - \mathbf{z}_i^T\mathbf{Q}A\mathbf{Q}^T\mathbf{z}_i}$$

Another thing to note is that if the information matrix is block diagonal, then the gradients corresponding to one parameter block can't influence the MLE of the opposing parameter block.

For notational ease:

$$\Delta\hat{\boldsymbol{\theta}}_j = \hat{\boldsymbol{\theta}}(\mathbf{1}) - \hat{\boldsymbol{\theta}}(\mathbf{w}_{(j)}).$$

Collett 1994, citing Hall et al. 1982, suggests standardizing the sensitivity to the control for the inverse of the variance-covariance of the estimated $\hat{\boldsymbol{\theta}}$, namely:

$$(\Delta\hat{\boldsymbol{\theta}}_j)^T i(\hat{\boldsymbol{\theta}})(\Delta\hat{\boldsymbol{\theta}}_j).$$

Recall that

$$i(\hat{\boldsymbol{\theta}}) = -\left.\frac{\partial \mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}$$

This leads to the tidy expression:

$$\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_j,\delta_j,\mathbf{z}_j)\big|^T_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}} \left(-\left.\frac{\partial \mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}\right)^{-1} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_j,\delta_j,\mathbf{z}_j)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}. \tag{5.48}$$

Alternatively, we can use the sandwich estimator for the asymptotic variance covariance matrix, which is defined in Equation (5.14) as:

$$\hat{\Sigma}_R = \left(-\frac{1}{n}\left.\frac{\partial \mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \left(\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_i,\delta_i,\mathbf{z}_i)\left(\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_i,\delta_i,\mathbf{z}_i)\right)^T\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})}\right)\left(-\frac{1}{n}\left.\frac{\partial \mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}\right)^{-1}$$

Note that this is variance/covariance matrix for $\sqrt{n}(\hat{\theta}_n - \theta^{\dagger})$. We instead want to get a sense for the variance/covariance matrix for $\hat{\theta}_n$, so we divide the expression by $\sqrt{n}$, leading to a variance estimate that is scaled by n^{-1}. Using the statistic

$$(\Delta\hat{\boldsymbol{\theta}}_j)^T (n^{-1}\hat{\Sigma}_R)^{-1}(\Delta\hat{\boldsymbol{\theta}}_j)$$

and noting the following equality:

$$(n^{-1}\hat{\Sigma}_R)^{-1} = \left(-\left.\frac{\partial \mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}\right)\left(\sum_{i=1}^n \left(\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_i,\delta_i,\mathbf{z}_i)\left(\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_i,\delta_i,\mathbf{z}_i)\right)^T\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})}\right)^{-1}\left(-\left.\frac{\partial \mathbf{U}(\boldsymbol{\theta},\mathbf{w})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}\right)$$

yields:

$$\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_j,\delta_j,\mathbf{z}_j)\big|^T_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}\left(\sum_{i=1}^n \left(\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_i,\delta_i,\mathbf{z}_i)\left(\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_i,\delta_i,\mathbf{z}_i)\right)^T\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})}\right)^{-1} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(t_j,\delta_j,\mathbf{z}_j)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1}),\mathbf{w}=\mathbf{1}}.$$

Let's do an example where we can analytically calculate the influence score for a single observation on the parameter vector:

Example 5.4.1. Influence of datapoints in exponential regression model The inverse of the observed information is:

$$\hat{\mathcal{I}}^{-1}(\hat{\lambda}, \hat{\beta}) = \frac{1}{\frac{(r_1+r_2)e^{\hat{\beta}}T_2}{\hat{\lambda}} - e^{2\hat{\beta}}T_2^2} \begin{bmatrix} \hat{\lambda}e^{\hat{\beta}}T_2 & -e^{\hat{\beta}}T_2 \\ -e^{\hat{\beta}}T_2 & \frac{r_1+r_2}{\hat{\lambda}^2} \end{bmatrix} \tag{5.49}$$

which simplifies to:

$$\begin{bmatrix} \frac{r_1}{T_1^2} & -\frac{1}{T_1} \\ -\frac{1}{T_1} & \frac{r_1+r_2}{r_1 r_2} \end{bmatrix} \tag{5.50}$$

and the score equations are:

$$\frac{\partial}{\partial \lambda}\ell(\lambda, \beta) = \frac{\delta_i}{\lambda} - e^{z_i \beta} t_i \tag{5.51}$$

$$\frac{\partial}{\partial \beta}\ell(\lambda, \beta) = \delta_i z_i - z_i \lambda e^{z_i \beta} t_i \tag{5.52}$$

which we evaluate at the MLE:

$$\hat{\lambda} = \frac{r_1}{T_1}$$

$$e^{\hat{\beta}} = \frac{r_2}{T_2}\frac{T_1}{r_1}$$

to yield

$$\frac{\partial}{\partial \lambda}\ell(\lambda, \beta) = \frac{\delta_i T_1}{r_1} - \left(\frac{r_2}{T_2}\frac{T_1}{r_1}\right)^{z_i} t_i \tag{5.53}$$

$$\frac{\partial}{\partial \beta}\ell(\lambda, \beta) = \delta_i z_i - \left(\frac{r_2}{T_2}\right) z_i t_i \tag{5.54}$$

For an individual with $z_i = 0$, this gives the sensitivities:

$$\begin{bmatrix} \frac{\delta_i - t_i \frac{r_1}{T_1}}{T_1} \\ \frac{t_i}{T_1} - \frac{\delta_i}{r_1} \end{bmatrix} = \begin{bmatrix} \frac{\delta_i - t_i \frac{r_1}{T_1}}{T_1} \\ \frac{t_i \frac{r_1}{T_1} - \delta_i}{r_1} \end{bmatrix} \tag{5.55}$$

These expressions make sense. At a mathematical level, they agree with the total derivatives for each function: $\frac{r_1}{T_1}$ and $-\log(r_1/T_1)$. Our expression for the sensitivity of the MLE to the omission of one datapoint is in terms of the difference between the MLE of the full model and the MLE of the leave-one-observation-out model:

$$\hat{\boldsymbol{\theta}}(\mathbf{1}) - \hat{\boldsymbol{\theta}}(\mathbf{w}_{(j)})$$

This means that the change in total time at risk for a group $j = 1, 2$, or T_j, is positive, as is the change in total failures for each group:

$$T_j - (T_j)_{(i)} = t_i \tag{5.56}$$

$$r_j - (r_j)_{(i)} = \delta_i \tag{5.57}$$

Then the expression the

$$\mathrm{d}\frac{r_1}{T_1} = \frac{\partial}{\partial r_1}\frac{r_1}{T_1}\mathrm{d}r_1 + \frac{\partial}{\partial T_1}\frac{r_1}{T_1}\mathrm{d}T_1 \tag{5.58}$$

$$= \frac{\mathrm{d}r_1}{T_1} - \mathrm{d}T_1\frac{r_1}{T_1^2} \tag{5.59}$$

$$\approx \frac{\delta_i - t_i\frac{r_1}{T_1}}{T_1} \tag{5.60}$$

and

$$\mathrm{d} - \log(r_1/T_1) = -\frac{\partial}{\partial r_1}\log(r_1/T_1)\mathrm{d}r_1 - \frac{\partial}{\partial T_1}\log(r_1/T_1)\mathrm{d}T_1 \tag{5.61}$$

$$= \frac{\mathrm{d}T_1}{T_1} - \frac{\mathrm{d}r_1}{r_1} \tag{5.62}$$

$$\approx \frac{t_i\frac{r_1}{T_1} - \delta_i}{r_1} \tag{5.63}$$

It helps to think about the units of the parameter estimates. λ measures the rate of failures per unit time, while β measures the log of the relative rates of failure. Thus β is unitless. Remember that

$$\delta_i - t_i\frac{r_j}{T_j}$$

is the residual for an individual i in group j. It compares the observed failure to the expected failure rate, which in the exponential model is just the estimated rate of failure times the time at risk for i, or t_i. When one removes an individual from group 1 the estimate for the rate of failure in group 1 declines by the residual expected failure per unit time. At the same time, the log relative rate of failure must increase by the residual failure per unit failure because the estimator for β is $\log(r_2/T_2) - \log(\hat{\lambda})$. Thus any change in $\hat{\lambda}$ has an opposite change for $\hat{\beta}$.

For an individual with $z_i = 1$, the sensitivities are:

$$\begin{bmatrix} 0 \\ \frac{\delta_i}{r_2} - \frac{t_i}{T_2} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{\delta_i - t_i\frac{r_2}{T_2}}{r_2} \end{bmatrix} \tag{5.64}$$

Again, this makes sense; $\hat{\lambda} = \frac{r_1}{T_1}$, so omitting an individual in group 2 can't change the MLE for λ. Finally, given $\hat{\beta} = \log(r_2/T_2) - \log(\hat{\lambda})$, omitting a datapoint will decrease the failure rate estimate within group 2 by the residual scaled by the failure rate. Note the total derivative of $\log(r_2/T_2)$, as above, is:

$$\mathrm{d}\log(r_2/T_2) = \frac{\mathrm{d}r_2}{r_2} - \frac{\mathrm{d}T_2}{T_2} \tag{5.65}$$

We can also calculate the scaled total deviation. For $z_i = 0$ we have:

$$\frac{\left(t_i \frac{r_1}{T_1} - \delta_i\right)^2}{r_1} \tag{5.66}$$

and for $z_i = 1$ we have

$$\frac{\left(t_i \frac{r_2}{T_2} - \delta_i\right)^2}{r_2} \tag{5.67}$$

5.5 Cox Proportional Hazards Model

The Cox proportional hazards model is one of the most widely-used statistical models. It is a semiparametric model for the hazard ratio, which means we will avoid specifying a parameteric form for the baseline, time-varying hazard rate, while specifying a parametric model for the influence of covariates on the hazard rate:

$$\lambda(t \mid \mathbf{z}) = \lambda_0(t)\exp(\mathbf{z}^T\boldsymbol{\beta})$$

In the following section we'll derive the likelihood for the model in a similar way to our NPMLE derivation of the hazard rate.

5.5.1 Cox model likelihood derivation and the Breslow estimator without ties

Suppose we have the standard survival-analysis triplet of observable random variables for each unit i under study:

$$\{(T_i = \min(X_i, C_i), \Delta_i = \mathbb{1}\left(X_i \leq C_i\right), \mathbf{z}_i), i = 1, \dots, n\}$$

where X_i is the absolutely continuous time to failure for unit i, C_i is the absolutely continuous time to censoring, and $\mathbf{z}_i$ is a length-p vector of time-invariant covariates that we are conditioning on.

As stated above, we assume that the hazard function for the distribution of X_i is:

$$\lambda_i(t) = \lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta})$$

and we'll leave the function $\lambda_0(t)$ unspecified. As in Chapter 4, we'll derive an estimator for $\lambda_0(t)$ at the event times $\{t_i, i = 1, \dots, n\}$. Let $\Lambda(t)$ be the right-continuous-with-left-hand-limits (cádlág for short, in French) cumulative hazard function with mass points at t_i. Note that $\lambda(t_i) = \Lambda(t_i) - \Lambda(t_i-)$, and we define the integral

$$\int_B F(t)\mathrm{d}\Lambda(t) = \sum_{i \mid t_i \in B} F(t_i)\lambda(t_i).$$

This is a nonparametric estimator because as the data grow, so too does the dimension of the parameter space. This is akin to the Theoretical note in Klein, Moeschberger, et al. 2003, and an exercise in O. O. Aalen 1988.

The joint likelihood for the model for set of observed data $\{(t_i, \delta_i, \mathbf{z}_i), i = 1, \dots, n\}$, which is assumed to have no ties in event times, is:

$$L(\lambda_0, \boldsymbol{\beta}) = \prod_{i=1}^{n}\left(\lambda_0(t_i)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\right)^{\delta_i}\exp(-\exp(\mathbf{z}_i^T\boldsymbol{\beta})\textstyle\int_0^{t_i}\lambda_0(u)du). \tag{5.68}$$

Recall the definition of $Y_i(u)$ from Chapter 4:

$$Y_i(u) = \mathbb{1}(t_i \geq u).$$

The function is left continuous with a jump from 1 to 0 at t_i. Then we can rewrite the model as

$$L(\lambda_0, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\lambda_0(t_i)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\right)^{\delta_i} \exp(-\exp(\mathbf{z}_i^T\boldsymbol{\beta}) \textstyle\int_0^\infty Y_i(u)\lambda_0(u)du) \tag{5.69}$$

$$= \left(\prod_{i=1}^{n} \left(\lambda_0(t_i)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\right)^{\delta_i}\right) \exp(-\textstyle\int_0^\infty \sum_{i=1}^n \exp(\mathbf{z}_i^T\boldsymbol{\beta})Y_i(u)\lambda_0(u)du) \tag{5.70}$$

$$= \left(\prod_{i=1}^{n} \left(\lambda_0(t_i)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\right)^{\delta_i}\right) \exp(-\textstyle\sum_{j=1}^n \left(\sum_{i=1}^n \exp(\mathbf{z}_i^T\boldsymbol{\beta})Y_i(t_j)\right)\lambda_0(t_j)) \tag{5.71}$$

Let's fix a value of $\boldsymbol{\beta}$ and compute the NPMLE for $\lambda_0(t_j)$. The log-likelihood for λ_0 is

$$\ell(\lambda_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \log\left(\lambda_0(t_i)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\right) - \textstyle\sum_{j=1}^n \left(\sum_{i=1}^n \exp(\mathbf{z}_i^T\boldsymbol{\beta})Y_i(t_j)\right)\lambda_0(t_j)$$

Differentiating with respect to $\lambda_0(t_j)$, provided $\delta_j = 1$, gives

$$\frac{\partial}{\partial\lambda_0(t_j)}\ell(\lambda_0, \boldsymbol{\beta}) = \frac{1}{\lambda_0(t_j)} - \sum_{i=1}^{n} \exp(\mathbf{z}_i^T\boldsymbol{\beta})Y_i(t_j)$$

We note that the second derivative with respect to $\lambda_0(t_j)$ is strictly negative for all positive $\lambda_0(t_j)$ Setting this expression equal to zero and solving for $\lambda_0\hat{(t_j)}$ will give the unique NPMLE

$$\hat{\lambda}_0(t_j) = \frac{1}{\sum_{i=1}^n \exp(\mathbf{z}_i^T\boldsymbol{\beta})Y_i(t_j)} \tag{5.72}$$

The denominator can be simplified if we define the set $R(t_j) : \{i \mid Y_i(t_j) = 1, i = 1, \ldots, n\}$. This is called the **risk set** at time t_j.

$$\hat{\lambda}_0(t_j) = \frac{1}{\sum_{i\in R(t_j)} \exp(\mathbf{z}_i^T\boldsymbol{\beta})} \tag{5.73}$$

Let's substitute this NPMLE into the likelihood in Equation (5.68). Recognize that by definition of δ_j and $\lambda_0(t_j)$ only jumping at event-of-interest times, the estimator is equivalently defined as

$$\hat{\lambda}_0(t_j) = \frac{\delta_j}{\sum_{i\in R(t_j)} \exp(\mathbf{z}_i^T\boldsymbol{\beta})} \tag{5.74}$$

Subbing this back into Equation (5.71) gives

$$L(\boldsymbol{\beta}) = \left(\prod_{i=1}^{n}\left(\frac{\exp(\mathbf{z}_i^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)}\exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right)^{\delta_i}\right)\exp\left(-\sum_{j=1}^{n}\delta_j\frac{\left(\sum_{i=1}^n \exp(\mathbf{z}_i^T\boldsymbol{\beta})Y_i(t_j)\right)}{\sum_{i\in R(t_j)}\exp(\mathbf{z}_i^T\boldsymbol{\beta})}\right) \tag{5.75}$$

$$= \left(\prod_{i=1}^{n}\left(\frac{\exp(\mathbf{z}_i^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)}\exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right)^{\delta_i}\right)\exp\left(-\sum_{j=1}^{n}\delta_j\right) \tag{5.76}$$

Then we get the final term for the Cox model partial likelihood:

$$L(\boldsymbol{\beta}) \propto \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{z}_i^T \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta})} \right)^{\delta_i} \tag{5.77}$$

$$= \left(\prod_{i|\delta_i=1} \frac{\exp(\mathbf{z}_i^T \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta})} \right) \tag{5.78}$$

where we note that maximizing Equation (5.77) will maximize Equation (5.76).

Let $\hat{\boldsymbol{\beta}}$ be the MLE of the expression $L(\boldsymbol{\beta})$. Then the estimator for the cumulative hazard is defined as:

$$\hat{\Lambda}_0(t) = \sum_{t_i|\delta_i=1, t_i \le t} \frac{1}{\sum_{j \in R(t_i)} \exp(\mathbf{z}_j^T \hat{\boldsymbol{\beta}})} \tag{5.79}$$

This is known as the **Breslow estimator** for the cumulative hazard.

5.5.2 Alternative view of the Cox model

The final form of the partial likelihood for the Cox model is:

$$L(\boldsymbol{\beta}) \propto \left(\prod_{i|\delta_i=1} \frac{\exp(\mathbf{z}_i^T \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta})} \right) \tag{5.80}$$

If we multiply top and bottom by $\lambda_0(t_i)$ (essentially we'll be multiplying by 1 a bunch of times), we get:

$$L(\boldsymbol{\beta}) \propto \left(\prod_{i|\delta_i=1} \frac{\lambda_0(t_i) \exp(\mathbf{z}_i^T \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \lambda_0(t_i) \exp(\mathbf{z}_j^T \boldsymbol{\beta})} \right) \tag{5.81}$$

If we write out the hazard function explicitly we get the following probability distribution:

$$L(\boldsymbol{\beta}) \propto \left(\prod_{i|\delta_i=1} \frac{\lim_{dt \searrow 0} P(t_i \le X_i < t_i + dt \mid X_i \ge t_i)}{\sum_{j \in R(t_i)} \lim_{dt \searrow 0} P(t_i \le X_j < t_i + dt \mid X_j \ge t_i)} \right) \tag{5.82}$$

Under the assumption of independent event times, we can interpret this probability function as the probability the i^{th} participant surviving to time t_i and failing just after t_i conditional on exactly 1 death occurring just after time t_i among those surviving to time t_i. Finally, under noninformative censoring, and noting that this is

$$L(\boldsymbol{\beta}) \propto \left(\prod_{i|\delta_i=1} \frac{\lim_{dt \searrow 0} P(t_i \le X_i < t_i + dt \mid X_i \ge t_i, C_i \ge t_i)}{\sum_{j=1}^{n} \lim_{dt \searrow 0} P(t_i \le X_j < t_i + dt \mid X_j \ge t_i, C_i \ge t_i)} \right) \tag{5.83}$$

Where the sum in the denominator follows because if $C_i < t_i$ or $X_i < t_i$, then $P(t_i \le X_j < t_i + dt \mid X_j \ge t_i, C_i \ge t_i) = 0$.

The no ties assumption is not problematic with absolutely continuous data. In practice there will be ties in the data.

Data with ties

The simplest approximation when there are ties in the data is to consider the times at which the ties occurred to be distinct, but mismeasured. We define $\{\tau_j, j = 1, \dots r = \sum_i \delta_i\}$ to be the distinct times at which failures occur. We can expand the dataset with

$$\{(\tau_j, d_j = \sum_{i|t_i=\tau_j} \delta_i, \mathbf{s}_j = \sum_{i|t_i=\tau_j} \mathbf{z}_i), j = 1, \dots, r\}.$$

We still need access to the risk set function $R(t) = \sum_i Y_i(t)$ and $\mathbf{z}_i$. The partial likelihood can be written in terms of the new dataset:

$$L(\boldsymbol{\beta}) \propto \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{z}_i^T \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta})} \right)^{\delta_i} \tag{5.84}$$

$$= \prod_{j=1}^{r} \prod_{i|t_i=\tau_j} \left(\frac{\exp(\mathbf{z}_i^T \boldsymbol{\beta})}{\sum_{k \in R(t_i)} \exp(\mathbf{z}_k^T \boldsymbol{\beta})} \right) \tag{5.85}$$

$$= \prod_{j=1}^{r} \frac{\exp(\mathbf{s}_j^T \boldsymbol{\beta})}{\left(\sum_{k \in R(\tau_j)} \exp(\mathbf{z}_k^T \boldsymbol{\beta})\right)^{d_j}} \tag{5.86}$$

This is not quite right because we've ignored the fact that when failure time is continuous, all true failure times are ordered in time. Let τ_j be a time interval in which several units failed, and let the set of units that fail at τ_j be $D(\tau)$, defined as:

$$D(\tau_j) = \{i \mid t_i = \tau_j, i = 1, \dots, n\}.$$

The proper way to handle ties would be to integrate over all possible permutations of failure times.

Exact handling of ties This section follows Kalbfleisch and Prentice 2002. Let the indices of the set of units failing at time τ_j be $\{1, \dots, d_j\}$. Let Q_j be the set of permutations of $D(\tau_j)$ and let $P = (p_1, \dots, p_{d_j})$ be an element of this permutation. Finally, let the extended at risk set be $R(\tau_j, P, m) = R(\tau_j) \setminus \{p_1, \dots, p_{m-1}\}$ where we let p_0 be the empty set. For a single term in the likelihood at time τ_j, we need to integrate the term

$$\prod_{i|\delta_i=1, t_i=\tau_j} \frac{\exp(\mathbf{z}_i^T \boldsymbol{\beta})}{\sum_{k \in R(\tau_j)} \exp(\mathbf{z}_k^T \boldsymbol{\beta})} \tag{5.87}$$

over all permutations of $D(\tau_j)$. We have no extra information to weight the orderings, so we give them all equal weight as $\frac{1}{d_j!}$. This integral is

$$\frac{\exp(\mathbf{s}_j^T \boldsymbol{\beta})}{d_j!} \sum_{P \in Q_j} \prod_{m=1}^{d_j} \frac{1}{\sum_{k \in R(\tau_j, P, m)} \exp(\mathbf{z}_k^T \boldsymbol{\beta})} \tag{5.88}$$

Not surprisingly, this calculation is prohibitively computationally intensive. The part that is computationally intensive is the sum.

Let's look at at a simple example to see one way we might approximate this integral:

Example 5.5.1. Suppose we have $d_j = 2$ for some τ_j. Let the risk set at τ_j be $R(\tau_j)$, and let i and m be the indices such that $t_i = t_m = \tau_j$. Let $c = \sum_{k\in R(\tau_j)} \exp(\mathbf{z}_k^T\beta)$, $a = \exp(\mathbf{z}_i^T\beta)$, $b = \exp(\mathbf{z}_m^T\beta)$. If $t_i \le t_m$, the terms in the Cox model should be:

$$\frac{a}{c} \times \frac{b}{c-a}$$

and if $t_m < t_i$, we would instead have

$$\frac{b}{c} \times \frac{a}{c-b}$$

Because we have no information about which event is more probable, we weight them equally with $\frac{1}{2}$:

$$\frac{1}{2}\frac{a}{c} \times \frac{b}{c-a} + \frac{1}{2}\frac{b}{c} \times \frac{a}{c-b} = \frac{ab}{2}\left(\frac{1}{c} \times \frac{1}{c-a} + \frac{1}{c} \times \frac{1}{c-b}\right)$$

We could approximate this expression instead by:

$$\frac{ab}{2c}\left(\frac{1}{c-a} + \frac{1}{c-b}\right) \approx \frac{ab}{2c}\left(\frac{2}{c-(a+b)/2}\right)$$

If we note that we have $d_j!$ terms in the sum, we can approximate the sum by Efron's approximation:

$$\frac{d_j!}{\prod_{m=0}^{d_j-1}\left(\sum_{k\in R(\tau_j)} \exp(\mathbf{z}_k^T\boldsymbol{\beta}) - \frac{m}{d_j}\sum_{k\in D(\tau_j)} \exp(\mathbf{z}_k^T\boldsymbol{\beta})\right)} \tag{5.89}$$

The intuition for this method is that you approximate the decrement in the risk set by the average of the relative risk of failure of the failed units at time τ_j.

5.5.3 Interpretation of the Cox regression model

The key idea for the Cox regression model is that we can maximize the partial likelihood without worrying about specifying any form for the baseline hazard rate $\lambda_0(t)$. Informally, this allows us to use standard asymptotic tests and confidence intervals for $\boldsymbol{\beta}$ without worrying about the infinite dimensional (i.e. unknown function) baseline hazard rate. Thus we can use all the asymptotic likelihood techniques we developed for parametric models for the Cox model.

5.5.4 Score function of the Cox model

Given that we'll use maximum likelihood to fit the Cox model, the score equations for the Cox model will be of importance to us. As shown in the previous section, the partial likelihood is:

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{z}_i^T \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta})} \right)^{\delta_i} \tag{5.90}$$

The log-partial likelihood is thus

$$\log PL(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left(\mathbf{z}_i^T \boldsymbol{\beta} - \log \left(\sum_{j \in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta}) \right) \right) \tag{5.91}$$

The gradient of this function with respect to the parameter vector $\boldsymbol{\beta}$ is

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log PL(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left(\mathbf{z}_i - \frac{\sum_{j \in R(t_i)} \mathbf{z}_j \exp(\mathbf{z}_j^T \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta})} \right) \tag{5.92}$$

This can be seen to be the difference between the covariate value for individual i who fails at time t_i and the weighted average covariate value for individuals in the risk set at t_i.

Example 5.5.2. Simple two-group Cox regression example Let's assume that we observe failure time data from two groups, 1 and 2, that contain no tied event times. The observed data is $\{(t_i, \delta_i, z_i), i = 1, \dots, n\}$. Each observation i is paired with a scalar value z_i taking the value 0 when i is in group 1, and 1 otherwise. The hazard model takes the form:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta z_i).$$

with $\lambda_0(t)$ left unspecified.

The partial likelihood for β is

$$\prod_{i=1}^{n} \left(\frac{\exp(\beta z_i)}{\sum_{k \in R(t_i)} \exp(\beta z_k)} \right)^{\delta_i} \tag{5.93}$$

Equation (5.93) simplifies if we create an alternative dataset by generating $\{\tau_j = t_i \mid \delta_i = 1\}$, and $\{(\tau_j, \delta_{2j}), j = 1, \dots r = \sum_i \delta_i\}$ the times of observed failures:

$$\prod_{j=1}^{r} \frac{\exp(\beta \delta_{2j})}{\sum_{k \in R(\tau_j)} \exp(\beta z_k)} \tag{5.94}$$

Let's write out the log-likelihood and the score equation for β:

$$\sum_{j=1}^{r} \beta \delta_{2j} - \sum_{j=1}^{r} \log \left(\sum_{k \in R(\tau_j)} \exp(\beta z_k) \right) \tag{5.95}$$

This can again be simplified if we keep track of the number of individuals at risk in each group. Let these variables be, as before, denoted $\overline{Y}_1(\tau_j), \overline{Y}_2(\tau_j)$. As a reminder, we define these variables as:

$$\overline{Y}_1(\tau_j) = \sum_{i=1}^{n} (1 - z_i) \mathbb{1} \left(t_i \geq \tau_j\right) \tag{5.96}$$

$$\overline{Y}_2(\tau_j) = \sum_{i=1}^{n} z_i \mathbb{1} \left(t_i \geq \tau_j\right) \tag{5.97}$$

Because $z_k = 0$ for all $\overline{Y}_1(\tau_j)$ we get

$$\sum_{j=1}^{r} \beta\delta_{2j} - \sum_{j=1}^{r} \log\left(\overline{Y}_1(\tau_j) + \overline{Y}_2(\tau_j)e^{\beta}\right) \tag{5.98}$$

One more simplification is that $\sum_{j=1}^{r} \delta_{2j} = \sum_{i=1}^{n} \delta_i z_i$, which we'll call r_2, or the total failures in group 2.

$$r_2\beta - \sum_{j=1}^{r} \log\left(\overline{Y}_1(\tau_j) + \overline{Y}_2(\tau_j)e^{\beta}\right) \tag{5.99}$$

Taking the gradient with respect to β gives:

$$r_2 - \sum_{j=1}^{r} \frac{\overline{Y}_2(\tau_j)e^{\beta}}{\overline{Y}_1(\tau_j) + \overline{Y}_2(\tau_j)e^{\beta}} \tag{5.100}$$

Setting this equation equal to zero leads to an equation we cannot explicitly solve in terms of β:

$$\sum_{j=1}^{r} \frac{\overline{Y}_2(\tau_j)e^{\beta}}{\overline{Y}_1(\tau_j) + \overline{Y}_2(\tau_j)e^{\beta}} = r_2 \tag{5.101}$$

An alternative is to use the score test. Here is the benefit of the score test in examples like these, where we don't have to maximize the log-likelihood under the alternative in order to test the hypothesis that the rates of failure are different between the two groups. Take a look at Section 5.1 to remind yourself about what the score test entails.

Here is the score equation for β

$$\sum_{j=1}^{r} \left(\delta_{2j} - \frac{\overline{Y}_2(\tau_j)e^{\beta}}{\overline{Y}_1(\tau_j) + \overline{Y}_2(\tau_j)e^{\beta}}\right) \tag{5.102}$$

The second derivative of the log-likelihood with respect to β is

$$-\sum_{j=1}^{r} \frac{\overline{Y}_1(\tau_j)\overline{Y}_2(\tau_j)e^{\beta}}{(\overline{Y}_1(\tau_j) + \overline{Y}_2(\tau_j)e^{\beta})^2} \tag{5.103}$$

The score test statistic is formed by evaluating the score and the observed information at $\beta = 0$:

$$\begin{aligned}
\frac{\partial}{\partial \beta}\ell(\beta)\mid_{\beta=0} &= \sum_{j=1}^{r}\left(\delta_{2j} - \frac{\overline{Y}_2(\tau_j)}{\overline{Y}_1(\tau_j) + \overline{Y}_2(\tau_j)}\right) \\
-\frac{\partial^2}{\partial^2 \beta}\ell(\beta)\mid_{\beta=0} &= \sum_{j=1}^{r}\frac{\overline{Y}_1(\tau_j)\overline{Y}_2(\tau_j)}{(\overline{Y}_1(\tau_j) + \overline{Y}_2(\tau_j))^2}
\end{aligned} \tag{5.104}$$

These are exactly the numerator and the denominator of the log-rank test statistic when there are no ties present. Remember, the log-rank test, using data collected through time point τ is

$$\frac{Z_j(\tau)}{\sqrt{\mathrm{Var}(Z_j(\tau))}}$$

With expressions for numerator and denominator below:

$$Z_j(\tau) = \sum_{i=1|t_i\leq\tau}^{n_1+n_2} W(t_i)\left(d_{ij} - d_i\frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\right)\mathrm{Var}\left(Z_j(\tau)\right) \quad = \sum_i W(t_i)^2\left(d_i\frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\left(1 - \frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\right)\frac{\overline{Y}(t_i) - d_i}{\overline{Y}(t_i) - 1}\right) \tag{5.105}$$

In our case here, d_i is always equal to 1 because we have no ties. With two groups $1 - \frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}$ is $\frac{\overline{Y}_1(t_i)}{\overline{Y}(t_i)}$, so the numerator and denominator simplify to equal the equations in Equation (5.104).

The duality between the Cox model and the log-rank test sheds some light as to the power of the log-rank test. Namely, the log-rank test tends to be powerful against the alternative proportional hazards hypotheses.

5.5.5 Model checking in the Cox model

We can use a lot of the same ideas we've used for parametric models for model checking in the Cox model. Remember the Cox-Snell residuals we defined using the estimated cumulative hazard function $\hat{\Lambda}_i(t)$:

$$e_i = \hat{\Lambda}_i(t_i) \overset{\text{approx}}{\sim} \text{Exponential}(1) \tag{5.106}$$

We can also use this function to define what's called a martingale residual:

$$e_i^M = \delta_i - \hat{\Lambda}_i(t_i)$$

These residuals are much closer to the residuals in linear regression models in that they sum to zero for any fitted model, and are zero in expectation in large samples and approximately uncorrelated. Exercise: show that this is true for the Cox model. A downside of the martingale residuals is that we are required to estimate the cumulative hazard function, which may not be of interest.

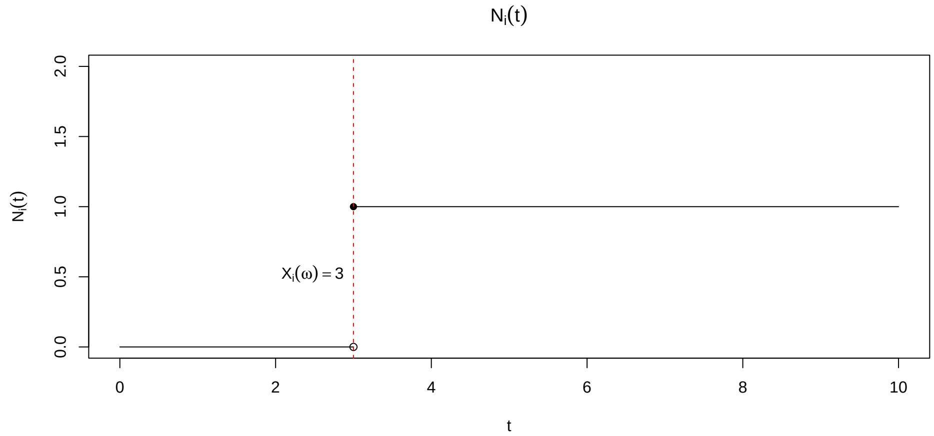

Figure 5.3: Example plot of $N_i(t) = \mathbb{1}\,(3 \le t)$

Martingale residuals

We've defined δ_i as the indicator that the i^{th} participant has experienced a failure over the course of the study. This is defined as $\delta_i = \mathbb{1}\,(X_i \le C_i)$ where C_i is the time of censoring. We can imagine defining this variable for every time point t after the i^{th} participant enters the study. Let $N_i(t) = \mathbb{1}\,(X_i \le t, X_i \le C_i)$. This equals δ_i when t equals the study end point. On the event that $X_i \le C_i$, $N_i(t)$ is a step-function of time, equalling 0 for all $t \ge X_i$, and then jumping to 1 at $t + \epsilon$ for all $\epsilon > 0$. A plot of this function is shown in Figure 5.3. This variable can be seen to be a function of t, and is technically called a stochastic process. One way to think about this is that $N_i(t)$ is a random function of time. To see this note that for every draw of X_i, $N_i(t)$ is a different step function, jumping up to 1 at X_i.

Of course, a natural quantity that arises from our definition is

$$\mathbb{E}\left[N_i(t)\right] = \mathbb{E}\left[\mathbb{1}\,(X_i \le t, X_i \le C_i)\right] = P(X_i \le t, X_i \le C_i).$$

Let $G_{C_i}(x)$ be the survival function of C_i, or $P(C_i > x)$ evaluated at x, and let $S_{X_i}(x)$ be the survival function of X_i. Writing down the expression for the expectation of $N_i(t)$ in terms of the hazard ratio for X_i, $\lambda_i(x)$, will show that the martingale residuals have mean zero: We'll assume for ease of exposition that $X_i \perp\!\!\!\perp C_i$. Remember that $\lambda_i(x) = f_{X_i}(x)/S_{X_i}(x-)$, and that we've defined $T_i = \min(X_i, C_i)$. Let the density function for T_i be $f_{T_i}(y)$.

$$\begin{aligned}
\mathbb{E}\left[N_i(t)\right] &= P(X_i \le t, X_i \le C_i) \\
&= \int_0^t G_{C_i}(x-) f_{X_i}(x) dx \\
&= \int_0^t G_{C_i}(x-) f_{X_i}(x) \frac{S_{X_i}(x-)}{S_{X_i}(x-)} dx \\
&= \int_0^t G_{C_i}(x-) S_{X_i}(x-) \frac{f_{X_i}(x)}{S_{X_i}(x-)} dx \\
&= \int_0^t P(C_i \ge x, X_i \ge x) \lambda_i(x) dx \\
&= \int_0^t \mathbb{E}\left[\mathbb{1}\left(C_i \ge x, X_i \ge x\right)\right] \lambda_i(x) dx \\
&= \int_0^t \mathbb{E}\left[\mathbb{1}\left(T_i \ge x\right)\right] \lambda_i(x) dx \quad (T_i = \min(X_i, C_i)) \\
&= \int_0^\infty \int_0^\infty \mathbb{1}\left(t \ge x\right) \mathbb{1}\left(y \ge x\right) f_{T_i}(y) dy \lambda_i(x) dx \\
&= \int_0^\infty \int_0^\infty \mathbb{1}\left(t \ge x\right) \mathbb{1}\left(y \ge x\right) \lambda_i(x) dx f_{T_i}(y) dy \text{ (Fubini)} \\
&= \int_0^\infty \int_0^t \mathbb{1}\left(y \ge x\right) \lambda_i(x) dx f_{T_i}(y) dy \\
&= \mathbb{E}\left[\int_0^t \mathbb{1}\left(T_i \ge x\right) \lambda_i(x) dx\right]
\end{aligned}$$

It immediately follows that

$$\mathbb{E}\left[N_i(t) - \int_0^t \lambda_i(x) \mathbb{1}\left(T_i \ge x\right) dx\right] = 0.$$

Schoenfeld residuals

Instead, we can use the score equations above to generate residuals, called **Schoenfeld residuals**. Let z_{ik} be the k^{th} component of the vector $\mathbf{z}_i$, and let

$$\hat{a}_{ik} = \frac{\sum_{j \in R(t_i)} z_{jk} \exp(\mathbf{z}_j^T \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta})}$$

$$e_{ik}^S = \delta_i \left(z_{ik} - \hat{a}_{ik}\right) \tag{5.107}$$

These give some sense of how much the i^{th} observation is contributing to the score equations for β_k at the MLE for $\boldsymbol{\beta}$.

This residual highlights the conditional probability view of the Cox model. In this interpretation, the i^{th} participant will be selected for failure at time t_i with probability:

$$\frac{\exp(\mathbf{z}_i \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta})} \tag{5.108}$$

This means, conditional on the set of observed values $\mathbf{z}_j, j \in R(t_i)$, that $\mathbf{z}_i$ is a random variable with a mean:

$$\mathbb{E}\left[\mathbf{z}_i \mid R(t_i)\right] = \frac{\sum_{j\in R(t_i)} \mathbf{z}_j \exp(\mathbf{z}_j^T \boldsymbol{\beta})}{\sum_{j\in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta})} \tag{5.109}$$

and variance:

$$\operatorname{Var}\left(\mathbf{z}_i \mid R(t_i)\right) = \frac{\sum_{j\in R(t_i)} \mathbf{z}_j \mathbf{z}_j^T \exp(\mathbf{z}_j^T \boldsymbol{\beta})}{\sum_{j\in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta})} - \mathbb{E}\left[\mathbf{z}_i \mid R(t_i)\right]^2 \tag{5.110}$$

Testing proportional hazards

This view also can be used to show that these residuals can be used to determine whether the proportional hazards assumption is valid. For the rest of this section, for exposition purposes, let's assume we have one covariate, z_i. Suppose we are worried that our data may be generated by a model with a hazard ratio defined as

$$\lambda_i(t) = \lambda_0(t) \exp(\beta z_i + g(t)\theta z_i)$$

for some function $g(t)$. Let the null hypothesis be that $g(t) = 0$, and the alternative be that $g(t) \neq 0$. If we write the true (unobservable) Schoenfeld residuals for this model we get:

$$\epsilon_i^S = z_i - \mathbb{E}_{H_0}\left[z_i \mid R(t_i)\right]$$

where

$$\mathbb{E}_{H_0}\left[z_i \mid R(t_i)\right] = \frac{\sum_{j\in R(t_i)} z_j \exp(z_j \beta)}{\sum_{j\in R(t_i)} \exp(z_j \beta)}$$

which we can expand with

$$\epsilon_i^S = z_i - \frac{\sum_{j\in R(t_i)} z_j \exp(\beta z_j + g(t)\theta z_j)}{\sum_{j\in R(t_i)} \exp(\beta z_j + g(t)\theta z_j)} + \frac{\sum_{j\in R(t_i)} z_j \exp(\beta z_j + g(t)\theta z_j)}{\sum_{j\in R(t_i)} \exp(\beta z_j + g(t)\theta z_j)} - \mathbb{E}_{H_0}\left[z_i \mid R(t_i)\right]$$

Under this formulation

$$\mathbb{E}\left[z_i - \frac{\sum_{j\in R(t_i)} z_j \exp(\beta z_j + g(t)\theta z_j)}{\sum_{j\in R(t_i)} \exp(\beta z_j + g(t)\theta z_j)} \mid R(t_i)\right] = 0$$

by definition of the conditional expectation of z_i. Let's do a one-term Taylor expansion of the true conditional mean under the alternative model about $g(t) = 0$:

$$\begin{aligned}
\frac{\sum_{j\in R(t_i)} z_j \exp(\beta z_j + g(t)\theta z_j)}{\sum_{j\in R(t_i)} \exp(\beta z_j + g(t)\theta z_j)} &\approx \frac{\sum_{j\in R(t_i)} z_j \exp(\beta z_j)}{\sum_{j\in R(t_i)} \exp(\beta z_j)} + \frac{\partial}{\partial g(t)} \frac{\sum_{j\in R(t_i)} z_j \exp(\beta z_j + g(t)\theta z_j)}{\sum_{j\in R(t_i)} \exp(\beta z_j + g(t)\theta z_j)}\Bigg|_{g(t)=0} g(t) \\
&= \frac{\sum_{j\in R(t_i)} z_j \exp(\beta z_j)}{\sum_{j\in R(t_i)} \exp(\beta z_j)} \\
&+ g(t) \; \frac{\sum_{j\in R(t_i)} \theta z_j^2 \exp(\beta z_j + g(t)\theta z_j)}{\sum_{j\in R(t_i)} \exp(\beta z_j + g(t)\theta z_j)} - g(t)\theta \frac{(\sum_{j\in R(t_i)} z_j \exp(\beta z_j + g(t)\theta z_j))^2}{(\sum_{j\in R(t_i)} \exp(\beta z_j + g(t)\theta z_j))^2}\Bigg| \\
&= \mathbb{E}_{H_0}\left[z_i \mid R(t_i)\right] + \operatorname{Var}_{H_0}\left(z_i \mid R(t_i)\right) \theta g(t)
\end{aligned}$$

Plugging this in above gives:

$$\epsilon_i^S \approx z_i - \frac{\sum_{j\in R(t_i)} z_j \exp(\beta z_j + g(t)\theta z_j)}{\sum_{j\in R(t_i)} \exp(\beta z_j + g(t)\theta z_j)} + \mathrm{Var}_{H_0}\left(z_i \mid R(t_i)\right)\theta g(t)$$

Taking conditional expecations gives:

$$\mathbb{E}\left[\epsilon_i^S \mid R(t_i)\right] \approx \mathrm{Var}_{H_0}\left(z_i \mid R(t_i)\right)\theta g(t)$$

Using the plug-in estimator e_i^S for $\mathbb{E}\left[\epsilon_i^S \mid R(t_i)\right]$ gives that

$$e_i^S \approx \mathrm{Var}_{H_0}\left(z_i \mid R(t_i)\right)\theta g(t)$$

under the alternative. Thus, plotting e_i^S against t gives a sense for whether there is evidence against the proportional hazards assumption.

Letting $e_i^S = z_i - \mathbb{E}_{H_0}\left[z_i \mid R(t_i)\right]$ gives

Influence function for Cox PH model

Of course, we have a tool that will give another approximation of how much an individual observation contributes to the score equations, the influence function we derived in Section 5.4.3. To make this more precise, we'll need to define the weighted score equations:

$$\nabla_\beta p\ell(\boldsymbol{\beta}, \mathbf{w}) = \sum_{i=1}^n w_i\delta_i\left(\mathbf{z}_i - \frac{\sum_{j\in R(t_i)} \mathbf{z}_j w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right) \tag{5.111}$$

The key thing to note is that the weight vector shows up in two places, because omitting an observation omits the observation from the risk set, against which other observations are measured. We can see this by rewriting the score equations for an observation indexed by k.

$$\nabla_\beta p\ell(\boldsymbol{\beta}, \mathbf{w}) = w_k\delta_k\left(\mathbf{z}_k - \frac{\sum_{j\in R(t_k)} \mathbf{z}_j w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_k)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right) + \sum_{i\neq k}^n w_i\delta_i\left(\mathbf{z}_i - \frac{\sum_{j\in R(t_i)} \mathbf{z}_j w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right) \tag{5.112}$$

Then we take the gradient of Equation (5.112) with respect to w_k. Note that w_k will show up for all risk sets prior to *and including* t_k. Then we can write the gradient as:

$$\begin{aligned}
\frac{\partial}{\partial w_k}\nabla_\beta p\ell(\boldsymbol{\beta}, \mathbf{w}) &= \delta_k\left(\mathbf{z}_k - \frac{\sum_{j\in R(t_k)} \mathbf{z}_j w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_k)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right) - w_k\delta_k\frac{\partial}{\partial w_k}\left(\frac{\sum_{j\in R(t_k)} \mathbf{z}_j w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_k)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right) \\
&\quad - \sum_{i\mid t_i<t_k} w_i\delta_i\frac{\partial}{\partial w_k}\left(\frac{\sum_{j\in R(t_i)} \mathbf{z}_j w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right) \\
&= \delta_k\left(\mathbf{z}_k - \frac{\sum_{j\in R(t_k)} \mathbf{z}_j w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_k)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right) - \sum_{i\mid t_i\leq t_k} w_i\delta_i\frac{\partial}{\partial w_k}\left(\frac{\sum_{j\in R(t_i)} \mathbf{z}_j w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right)
\end{aligned}$$

The gradient of the second term is

$$\frac{\partial}{\partial w_k}\left(\frac{\sum_{j\in R(t_i)} \mathbf{z}_j w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right) = \frac{\mathbf{z}_k \exp(\mathbf{z}_k^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})} \tag{5.113}$$

$$- \exp(\mathbf{z}_k^T\boldsymbol{\beta})\frac{\sum_{j\in R(t_i)} \mathbf{z}_j w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{(\sum_{j\in R(t_i)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta}))^2} \tag{5.114}$$

Putting this all back together gives us

$$\begin{aligned}\frac{\partial}{\partial w_k}\nabla_\beta p\ell(\boldsymbol{\beta},\mathbf{w}) = \delta_k\left(\mathbf{z}_k - \frac{\sum_{j\in R(t_k)} \mathbf{z}_j w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_k)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right)\\ - \sum_{i|t_i\le t_k} w_i\delta_i \frac{\exp(\mathbf{z}_k^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\left(\mathbf{z}_k - \frac{\sum_{j\in R(t_i)} \mathbf{z}_j w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)} w_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right)\end{aligned}$$

Evaluating this term at $w_i = 1 \forall i$ gives us

$$\begin{aligned}\frac{\partial}{\partial w_k}\nabla_\beta p\ell(\boldsymbol{\beta},\mathbf{w})\bigg|_{\mathbf{w}=1} = \delta_k\left(\mathbf{z}_k - \frac{\sum_{j\in R(t_k)} \mathbf{z}_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_k)} \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right)\\ - \sum_{i|t_i\le t_k} \delta_i \frac{\exp(\mathbf{z}_k^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)} \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\left(\mathbf{z}_k - \frac{\sum_{j\in R(t_i)} \mathbf{z}_j \exp(\mathbf{z}_j^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)} \exp(\mathbf{z}_j^T\boldsymbol{\beta})}\right)\end{aligned}$$

If we look at this for the m^{th} element of $\boldsymbol{\beta}$, we can rewrite it in terms of the Schoenfeld residuals and the terms $\hat{a}_{im}$:

$$\left(\frac{\partial}{\partial w_k}\nabla_\beta p\ell(\boldsymbol{\beta},\mathbf{w})\bigg|_{\mathbf{w}=1}\right)_m = e_{im}^S - \sum_{i|t_i\le t_k} \delta_i \frac{\exp(\mathbf{z}_k^T\boldsymbol{\beta})}{\sum_{j\in R(t_i)} \exp(\mathbf{z}_j^T\boldsymbol{\beta})}(z_{km} - \hat{a}_{im})$$

This shows that the impact of leaving out one observation on the score equation is a) the direct effect of the observed failure for the k^{th} unit had on the likelihood and b) the indirect effect of being excluded from the risk set; this impact occurs even if the k^{th} patient is not observed to fail. The second term also increases in magnitude as the time at risk increases, so for patients at risk for a long time, this term typically outweighs the first term.

5.5.6 Stratified Cox model

Sometimes the Cox model won't be sufficiently flexible for our modeling needs. The issue is that despite the baseline time-varying hazard being an unknown unspecified function, it is assumed to describe the baseline hazard for all individuals in the population. This often won't hold for heterogeneous populations. The solution is to use a stratified Cox model, which allows for the baseline hazard to depend on a known stratum. For example, let's say that the time course of a disease is known to vary by treatment. Then we would want a model that could replicate that pattern. For patient i in treatment group j, let the function $j[i] : \mathbb{Z}^+ \to \mathbb{Z}^+$ be the map between patient index and treatment group (AKA stratum) membership. This is just a fancy way to say that there is a vector with n elements where each element represents the stratum of the i^{th} individual. Let there be J treatment groups (i.e strata). Let the hazard ratio be defined as:

$$\lambda_{ij}(t) = \lambda_{j[i]}(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta}).$$

We can use the results from our derivation of the Cox partial likelihood to aid in our derviation of the stratified Cox model's likelihood.

The joint likelihood for the observed data $\{(t_i, \delta_i, \mathbf{z_i}, j[i]), i = 1, \ldots, n\}$ is

$$L(\{\lambda_j(t)\}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\lambda_{j[i]}(t_i)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\right)^{\delta_i} \exp\left(-\exp(\mathbf{z}_i^T\boldsymbol{\beta}) \int_0^\infty Y_i(u)\lambda_{j[i]}(u)du\right) \tag{5.115}$$

We can rewrite this as the product over j and an inner product over the units i such that $j[i] = j$:

$$L(\{\lambda_j(t)\}, \boldsymbol{\beta}) = \prod_{j=1}^{J} \prod_{i|j[i]=j} \left(\lambda_j(t_i)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\right)^{\delta_i} \exp\left(-\exp(\mathbf{z}_i^T\boldsymbol{\beta}) \int_0^\infty Y_i(u)\lambda_j(u)du\right) \tag{5.116}$$

Let each $\Lambda_j(t)$ be unspecified right-continuous non-decreasing step functions that jump at the collection of times $\{t_i \mid j[i] = j\}$.:

$$L(\{\lambda_j(t)\}, \boldsymbol{\beta}) = \left(\prod_{j=1}^{J} \prod_{i|j[i]=j} \left(\lambda_j(t_i)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\right)^{\delta_i}\right) \exp(-\sum_{j=1}^{J} \int_0^\infty \textstyle\sum_{i|j[i]=j} \exp(\mathbf{z}_i^T\boldsymbol{\beta}) Y_i(u)\lambda_j(u)du) \tag{5.117}$$

$$= \left(\prod_{j=1}^{J} \prod_{i|j[i]=j} \left(\lambda_j(t_i)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\right)^{\delta_i}\right) \exp(-\textstyle\sum_{j=1}^{J} \sum_{k|j[k]=j} \left(\sum_{i|j[i]=j} \exp(\mathbf{z}_i^T\boldsymbol{\beta}) Y_i(t_k)\right) \lambda_j(t_k)) \tag{5.118}$$

Solving the score equations for Equation (5.118) gives an expression for $\lambda_j(t_k)$:

$$\hat{\lambda}_j(t_k) = \frac{1}{\sum_{i|R(t_k), j[i]=j} \exp(\mathbf{z}_i^T\boldsymbol{\beta})} \tag{5.119}$$

Plugging this back into the equations above gives

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{J} \prod_{i|j[i]=j} \left(\frac{\exp(\mathbf{z}_i^T \boldsymbol{\beta})}{\sum_{k|R(t_i),j[k]=j} \exp(\mathbf{z}_k^T \boldsymbol{\beta})} \right)^{\delta_i} \exp\left(-\sum_{i=1}^{n} \delta_i \right) \tag{5.120}$$

Thus the likelihood is stratified by j, but additive over j. The likelihood enforces that information about β is shared across all units, but the risk set to which each failure is compared is limited to individuals within the same stratum.

5.5.7 Including time dependent variables in the Cox model

Let's say we now have a variable $W_i(t)$ that changes with t and we wish to include this variable in our Cox model. There is no mathematical reason we cannot include this variable in our model. The hazard ratio for an individual is now:

$$\lambda_i(t) = \lambda_0(t) \exp(\mathbf{z}_i^T \boldsymbol{\beta} + \gamma W_i(t))$$

where γ is the coefficient for the time-varying covariate. The first thing to note is that this is no longer a proportional hazards model, because $\lambda_i(t)/\lambda_0(t)$ is not constant in time. Instead:

$$\lambda_i(t)/\lambda_0(t) = \exp(\mathbf{z}_i^T \boldsymbol{\beta} + \gamma W_i(t)).$$

Now we construct the partial likelihood.

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{z}_i^T \boldsymbol{\beta} + \gamma W_i(t_i))}{\sum_{j\in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta} + \gamma W_j(t_i))} \right)^{\delta_i} \tag{5.121}$$

The complication arises in the denominator, where we need to know the values of W_j at time t_i.

This isn't an issue for variables that are exogenous, or determined outside of the patient's survival process. One example is $W_i(t)$ is the dose of a medicine that is administered as part of clinical trial. This variable is controlled by investigators and is hypothetically known at every time point for every patient.

But for variables that are related to the patients' survival processes, we might not know these variables at every point. Let's say we're running an influenza vaccine efficacy trial and we are measuring antibody levels at regular visits after administration of vaccines. Our primary outcome is influenza infection. Suppose that patient i becomes infected on day t_i. What values should we use for $W_i(t_i)$? What about $W_j(t_i)$?

Let's say that we have $W_i(t_i - 1)$ for the i^{th} participant, and we have $W_j(t_i - 1)$ and $W_j(t_i + 1)$ for participants $j \in R(t_i)$. The value for the $W_i(t_i)$ should be the $W_i(t_i - 1)$

because we have no other choices. As for participants $j \in R(t_i)$, we should use the $W_j(t_i - 1)$, though it may be tempting to use $W_j(t_i + 1)$. Using values in the future will bias our estimates of γ because this does not accurately represent the relative risk for unit i *at time* t_i. In fact, using values in the past will *also* bias our estimates of γ, similarly because this does not accurately represent the relative risk for unit i *at time* t_i either. Solutions for this

We can still use the Breslow estimator for cumulative baseline hazard:

$$\hat{\Lambda}_0(t) = \sum_{i|t_i \le t, \delta_i = 1} \frac{1}{\sum_{j \in R(t_i} \exp(\mathbf{z}_j^T \boldsymbol{\beta} + \gamma W_j(t_i))} \tag{5.122}$$

Representing time-varying covariates in survival data

The way we represent time-varying covariates in a dataset is somewhat different than what we're used to with simpler survival data. The reason for this is that every individual has different numbers of time-varying covariate values depending on their time to failure or censoring. Thus a datatset in wide format might look like: An alternative is called the

id	time	status	age	atime1	atime2	atime3	a1	a2	a3
1	100	0	45	0	60	90	3.8	3.4	2.9
2	80	1	65	0	60	NA	2.8	2.4	NA

Table 5.1: Caption

counting process representation. The data now look like: These sort of data can be fitted

id	obs	time	age	a	start	stop	status
1	1	0	45	3.8	0	60	0
1	2	60	45	3.4	60	90	0
1	3	90	45	2.9	90	100	0
2	1	0	65	2.8	0	60	0
2	2	60	65	2.4	60	80	1

Table 5.2: Caption

in the survival package using the following command: `coxph(Surv(start, stop, status) age + a, data = data)`.

5.5.8 Solving the problem with time-varying covariates

This discussion follows Tsiatis and Davidian 2004. Imagine we had the following set of random variables for each participant in a clinical trial:

$$\{(T_i = \min(X_i, C_i), \mathbf{z}_i, W_i(u) \mid 0 \leq u \leq T_i), i = 1, \ldots, n\}.$$

If we observed this process for each individual, we could fit a Cox PH model as above:

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{z}_i^T \boldsymbol{\beta} + \gamma W_i(t_i))}{\sum_{j \in R(t_i)} \exp(\mathbf{z}_j^T \boldsymbol{\beta} + \gamma W_j(t_i))} \right)^{\delta_i} \tag{5.123}$$

where we know $W_i(u)$ exactly for each individual for their entire at risk period.

This is obviously not realistic; typically participants will have intermittent measurements of biomarkers. An example might be CD4 counts, a measure of the concentration of T-cells in the blood, which is an important measurement for those with HIV. A low CD4 count can be an indication that an individual is at risk for AIDS; in fact a CD4 count of below 200 is one of the diagnostic criteria for AIDS.

Furthermore, we typically will not observe any underlying "true" values of $W_i(u)$, but instead we'll observe some noisy proxy for that variable, say $\tilde{W}_i(u)$. The solution to this is to model $W_i(u)$ as an unknown parameter that we learn about via observations of $\tilde{W}_i(u)$ at different time points:

$$\begin{aligned}
\tilde{W}_i(u_j) &= W_i(u_j) + \varepsilon_i(u_j) \\
\varepsilon_i(u_j) &\sim F \\
W_i(u_j) &= \alpha_{i0} + \alpha_{i1} u_j \\
(\alpha_{i0}, \alpha_{i1}) &\sim G
\end{aligned}$$

Then the hazard ratio for X_i would be:

$$\lambda_i(t) = \lambda_0(t) \exp(\gamma(\alpha_{i0} + \alpha_{i1} t) + \mathbf{z}_i^T \boldsymbol{\beta})$$

The key limitation for survival analysis is that we need access to $W_i(u)$ for all u prior to censoring or failure. This is only possible via a model for $W_i(u)$. One could imagine a nonparametric model taking the place of the simple linear model employed above. Despite its simplicity, the model above is investigated in Tsiatis and Davidian 2004.

5.6 Frailty and unobserved variation

Sometimes we want to account for extra variability that occurs at the patient level. This is typically handled via a positive, time-invariant random variable, ξ_i, that multiplies the hazard function:

$$\lambda_i(t) \mid \xi_i = \xi_i \lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}}$$

Then the survival function, conditional on the random variable ξ_i is

$$S_i(t \mid \xi_i) = \exp(-\xi_i \int_0^t \lambda_0(u) e^{\mathbf{z}_i^T \boldsymbol{\beta}} du) \tag{5.124}$$

The population survival function, or that which marginalizes over the population density of ξ_i is

$$\mathbb{E}_{\xi_i}\left[S_i(t \mid \xi_i)\right] = \mathbb{E}_{\xi_i}\left[\exp(-\xi_i \int_0^t \lambda_0(u) e^{\mathbf{z}_i^T \boldsymbol{\beta}} du)\right] \tag{5.125}$$

If we assume that the frailty term is gamma distributed, we can recover an analytic form for the survival function. Let $\Lambda_0(t)$ be the shared cumulative hazard so that the individual cumulative hazard is $w_i \Lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}}$.

$$\begin{aligned}
S_i(t) &= \frac{k^\theta}{\Gamma(\theta)} \int_0^\infty \exp(-w\Lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}}) w^{\theta-1} \exp(-kw) dw \\
&= \frac{k^\theta}{\Gamma(\theta)} \int_0^\infty w^{\theta-1} \exp(-(k + \Lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}}) w) dw \\
&= \frac{k^\theta}{\Gamma(\theta)} \frac{\Gamma(\theta)}{(k + \Lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}})^\theta} \\
&= \left(\frac{k}{k + \Lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}}}\right)^\theta
\end{aligned}$$

In fact, the gamma frailty model allows for analytic forms for the density as well:

$$\begin{aligned}
f_i(t, \delta) &= \frac{k^\theta}{\Gamma(\theta)} \int_0^\infty (w\lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}})^\delta \exp(-w\Lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}}) w^{\theta-1} \exp(-kw) dw \\
&= (\lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}})^\delta \frac{k^\theta}{\Gamma(\theta)} \int_0^\infty w^{\theta+\delta-1} \exp(-(k + \Lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}}) w) dw \\
&= (\lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}})^\delta \frac{k^\theta}{\Gamma(\theta)} \frac{\Gamma(\theta + \delta)}{(k + \Lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}})^{\theta+\delta}} \\
&= (\theta\lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}})^\delta \frac{k^\theta}{(k + \Lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}})^{\theta+\delta}} \qquad \text{nb: } \Gamma(\theta + 1) = \theta\Gamma(\theta) \\
&= \left(\frac{\theta\lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}}}{k + \Lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}}}\right)^\delta \left(\frac{k}{k + \Lambda_0(t) e^{\mathbf{z}_i^T \boldsymbol{\beta}}}\right)^\theta
\end{aligned}$$

This shows that the hazard rate is

$$\frac{\theta\lambda_0(t)e^{\mathbf{z}_i^T\boldsymbol{\beta}}}{k+\Lambda_0(t)e^{\mathbf{z}_i^T\boldsymbol{\beta}}}.$$

One thing to note is that, while it is a proportional hazards model conditional on ξ_i, after marginalizing over the frailty distribution it is no longer a proportional hazard model.

5.6.1 Marginal survival and hazard functions

Something to note is that the marginal survival function when frailty is present is calculated as

$$\mathbb{E}_\xi\left[e^{-\xi\Lambda(t)}\right].$$

This is the Laplace transform of ξ:

$$\mathcal{L}(c)=\mathbb{E}_\xi\left[e^{-\xi c}\right].$$

evaluated at $\Lambda(t)$, or $\mathcal{L}(\Lambda(t))$. Thus, if we know the Laplace transform for ξ, we'll easily know the marginal survival function.

The relationship between the population hazard function and the population survival function can be derived by considering the density of the failure time X_i marginalizing over ξ_i:

$$\begin{aligned}
f_{X_i}(t) &= \mathbb{E}_{\xi_i}\left[f_{X_i|\xi_i}(t)\right] && (5.126)\\
&= \mathbb{E}_{\xi_i}\left[\xi_i\lambda_0(t)S(t\mid\xi_i)\right] && (5.127)\\
&= \mathbb{E}_{\xi_i}\left[\xi_i\lambda_0(t)\mathbb{E}\left[\mathbb{1}\left(X_i>t\right)\mid\xi_i\right]\right] && (5.128)\\
&= \mathbb{E}_{\xi_i}\left[\mathbb{E}\left[\xi_i\lambda_0(t)\mathbb{1}\left(X_i>t\right)\mid\xi_i\right]\right] && (5.129)\\
&= \mathbb{E}_{\xi_i,X_i}\left[\xi_i\lambda_0(t)\mathbb{1}\left(X_i>t\right)\right] && (5.130)\\
&= \mathbb{E}_{\xi_i}\left[\xi_i\lambda_0(t)\mid X_i>t\right]\mathbb{E}_{\xi_i,X_i}\left[\mathbb{1}\left(X_i>t\right)\right] && (5.131)
\end{aligned}$$

where the last line follows from

$$\mathbb{E}\left[\xi_i\mid\mathbb{1}\left(A\right)\right]=\frac{\mathbb{E}\left[\xi_i\mathbb{1}\left(A\right)\right]}{\mathbb{E}\left[\mathbb{1}\left(A\right)\right]}.$$

Thus this shows that the population hazard, denoted $\mu(t)$, is

$$\begin{aligned}
\mu(t) &= \frac{f_{X_i}(t)}{\mathbb{E}_{\xi_i,X_i}\left[\mathbb{1}\left(X_i>t\right)\right]} && (5.132)\\
&= \mathbb{E}_{\xi_i}\left[\xi_i\lambda_0(t)\mid X_i>t\right] && (5.133)
\end{aligned}$$

This shows another way to get to the population hazard rate:

$$
\begin{aligned}
-\frac{\partial}{\partial t}\log(S(t)) &= -\frac{\partial}{\partial t}\log\mathcal{L}(\Lambda(t)) \\
&= \frac{-\frac{\partial}{\partial t}\mathcal{L}(\Lambda(t))}{\mathcal{L}(\Lambda(t))} \\
&= -\lambda(t)\frac{\mathcal{L}(\Lambda(t))'}{\mathcal{L}(\Lambda(t))}
\end{aligned}
$$

The Laplace transform of a gamma distribution is

$$
\mathcal{L}(c) = \left(\frac{k}{k+c}\right)^{\theta}
$$

Given the structure of the proportional frailty model, namely $\lambda_i(t) = \xi_i\lambda_0(t)$, it is natural to enforce the constraint that $\mathbb{E}[\xi] = 1$. This would mean that $k = \theta$. The most common way to paramterize this model is in terms of the variance, which in this case is $\frac{\theta}{k^2} = \frac{1}{\theta}$. Let $\nu = \frac{1}{\theta}$. Then

$$
\begin{aligned}
\mathcal{L}(c) &= \left(\frac{k}{k+c}\right)^{\theta} \\
&= \left(\frac{1}{1+c/k}\right)^{\theta} \\
&= \left(\frac{1}{1+c/\theta}\right)^{\theta} \\
&= (1+\nu c)^{-\frac{1}{\nu}}
\end{aligned}
$$

With a baseline hazard rate of $\lambda_0(t)$, we get the following form for the population survival function:

$$
\begin{aligned}
S(t) &= \mathcal{L}(\Lambda(t)) \\
&= (1+\nu\Lambda_0(t))^{-\frac{1}{\nu}}
\end{aligned}
$$

Using the fact that

$$
\mathcal{L}'(c) = -(1+\nu c)^{-\frac{1}{\nu}-1}
$$

$$
\begin{aligned}
\mu(t) &= -\lambda(t)\frac{\mathcal{L}(\Lambda(t))'}{\mathcal{L}(\Lambda(t))} \\
&= -\lambda_0(t)\frac{-(1+\nu\Lambda_0(t))^{-\frac{1}{\nu}-1}}{(1+\nu\Lambda_0(t))^{-\frac{1}{\nu}}} \\
&= \frac{\lambda_0(t)}{1+\nu\Lambda_0(t)} \qquad (5.134)
\end{aligned}
$$

As ν or, equivalently the variance, increases, the population hazard decreases as time increases.

This is related to the selection effect in survivors. Let's say we want to understand the distribution of frailty for people who survive past a certain time point. We can use Laplace transforms to do so. Remember, $S(t) = \mathcal{L}(\Lambda(t))$. If we calculate $S(t \mid X > s)$ and we recognize the functional form as corresponding to the Laplace transform of a random variable, $\mathbb{E}_{\xi}\left[e^{-\xi\Lambda(t)}\right]$ then we can say that ξ is distributed according to the distribution corresponding to the Laplace transform.

Thus, we want the survival function for people surviving past a time s, for $t > s$:

$$P(X > t \mid X > s) = \frac{P(X > t)}{P(X > s)} \tag{5.135}$$

$$= \frac{\mathbb{E}\left[e^{-\xi\Lambda(t)}\right]}{\mathbb{E}\left[e^{-\xi\Lambda(s)}\right]} \tag{5.136}$$

For the Gamma distribution we get

$$\frac{\mathbb{E}\left[e^{-\xi\Lambda(t)}\right]}{\mathbb{E}\left[e^{-\xi\Lambda(s)}\right]} = (k + \Lambda(t))^{-\theta}(k + \Lambda(s))^{\theta} \tag{5.137}$$

$$= \left(\frac{k + \Lambda(s)}{k + \Lambda(t)}\right)^{\theta} \tag{5.138}$$

$$= \left(\frac{k + \Lambda(s)}{k + \Lambda(s) + (\Lambda(t) - \Lambda(s))}\right)^{\theta} \tag{5.139}$$

This conditional survival function is the Laplace transform of a Gamma random variable with shape θ and rate $k + \Lambda(s)$, evaluated at $\Lambda(t) - \Lambda(s)$. What does this show us? We know that the mean is no longer 1, comparing this to the Laplace transform for a Gamma random variable parameterized with k and θ:

$$\mathcal{L}(c) = \left(\frac{k}{k + c}\right)^{\theta}.$$

This variable has a mean of

$$\frac{\theta}{k}.$$

Thus for the survivors, the expected frailty is

$$\frac{\theta}{k + \Lambda(s)}$$

This is declining as s increases. In the case where $k = \theta = \nu^{-1}$, we get

$$\frac{\nu^{-1}}{\nu^{-1} + \Lambda(s)} = \frac{1}{1 + \nu\Lambda(s)}$$

5.6.2 Comparison between two risk groups with frailty

This selection effect can have consequences when comparing a group at high risk for an event and a group at low risk for an event. Suppose that the individual hazard rate for high risk group is a multiple of the hazard rate in the low-risk group $\lambda_{0H}(t) = r\lambda_0(t)$ for $r > 1$. The individual hazard rates have frailties attached to them, so they are:

$$\lambda_{iH}(t) = \xi_{iH} r \lambda_0(t)$$
$$\lambda_{iL}(t) = \xi_{iL} \lambda_0(t)$$

Let $\mu_H(t)$ and $\mu_L(t)$ be the population hazard rates for the high and low risk individuals, respectively. If ξ_{iH} and ξ_{iL} have the same density, namely a gamma with mean 1 and a variance equal to ν, then we can use the results from Equation (5.134) to derive the population relative risk:

$$\begin{aligned}\frac{\mu_H(t)}{\mu_L(t)} &= \frac{r\lambda_0(t)}{1 + r\nu\Lambda_0(t)} \left(\frac{\lambda_0(t)}{1 + \nu\Lambda_0(t)} \right)^{-1} \\ &= r \frac{1 + \nu\Lambda_0(t)}{1 + r\nu\Lambda_0(t)}\end{aligned}$$

This gives rise to a population hazard ratio comparison for the two groups, or relative risk, that is declining in time. In this example the relative risk is always above 1, so we would still be "right" directionally about individual relative risk from using the population relative risk comparison, namely that the high risk group has higher risk than the lower group. This is not always the case.

Time-varying individual relative risk

Suppose instead that the high-risk group is such that some exposure imparts has increased this group's relative risk compared to a baseline hazard at small times but this relative risk declines to 1 as time increases. The function $r(t) = 1 + e^{-t}$ satisfies this basic trajectory. If we assume that the low-risk group has a constant individual level hazard the individual hazard rates are written:

$$\lambda_{iH}(t) = \xi_{iH}(1 + e^{-t})$$
$$\lambda_{iL}(t) = \xi_{iL}$$

The cumulative hazard rates are:

$$\Lambda_{iH}(t) = \xi_{iH}(t + 1 - e^{-t})$$
$$\Lambda_{iL}(t) = \xi_{iL} t$$

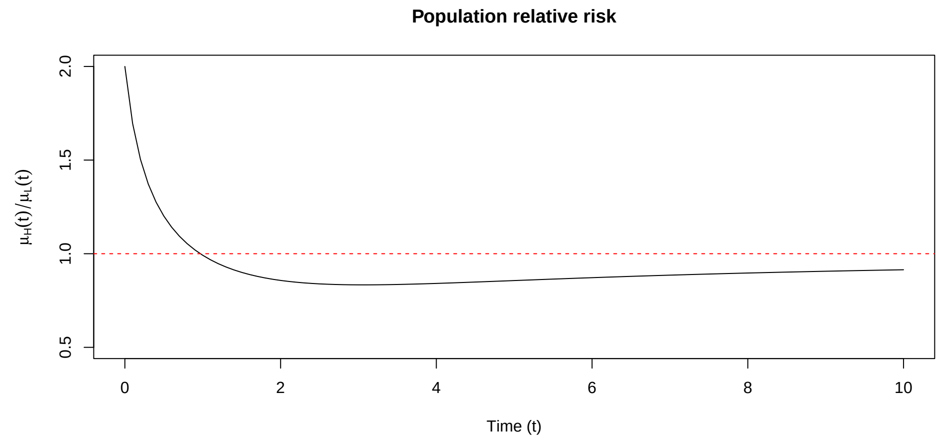

Figure 5.4: Plot of $\frac{\mu_H(t)}{\mu_L(t)}$ versus time for $\nu = 1.5$. Plot shows how population relative risk can begin above 1 and dip below 1 even when individual relative risk always remains above 1.

This results in a population relative risk as:

$$\frac{\mu_H(t)}{\mu_L(t)} = (1 + e^{-t})\frac{1 + \nu t}{1 + \nu t + \nu(1 - e^{-t})}$$

The odd thing is that the population relative risk starts at 2 and declines to below 1 as time increases, as can be seen in Figure 5.4. This results in a paradox, namely that individual relative risk is always greater than 1 but that population relative risk does not adhere to this relationship.

Discontinuing treatment

Suppose that we have two treatment groups, a placebo group, and a group that receives an experimental drug. For a treatment that works, we assume that the individual baseline hazard in the placebo group is $\xi_i\lambda_0(t)$, while the hazard in the treatment group is $\xi_i r\lambda_0(t)$ for $r < 1$. Thus, the treatment decreases the risk of an event. Suppose treatment is discontinued at time t_1, and we compare the hazard ratios of the placebo and treatment groups at time $t > 1$. The cumulative hazards are below, where $\Lambda_{iP}(t)$ is the cumulative hazard for the placebo group, and $\Lambda_{iT}(t)$ is the cumulative hazard in the treatment group.

$$\begin{aligned}
\Lambda_{iP}(t) &= \xi_i\Lambda_0(t) \\
\Lambda_{iT}(t) &= \xi_i t(\Lambda_0(t) - \Lambda_0(t_1) + r\Lambda_0(t_1)) \\
&= \xi_i t(\Lambda_0(t) + \Lambda_0(t_1)(r - 1))
\end{aligned}$$

Assuming a gamma distributed frailty with variance ν yields the following ratio of the population hazard rates at time t:

$$\begin{aligned}
\frac{\mu_P(t)}{\mu_T(t)} &= \frac{\lambda_0(t)}{1+\nu\Lambda_0(t)}\left(\frac{\lambda_0(t)}{1+\nu\left(\Lambda_0(t)+\Lambda_0(t_1)(r-1)\right)}\right)^{-1} \\
&= \frac{1+\nu\Lambda_0(t)+\nu\Lambda_0(t_1)(r-1)}{1+\nu\Lambda_0(t)} \\
&= 1-(1-r)\frac{\nu\Lambda_0(t_1)}{1+\nu\Lambda_0(t)}
\end{aligned}$$

Thus, it will appear as if the treatment had a negative effect, namely that after discontinuing treatment the hazard increased for those in the treatment groups, though it was only the case that the treatment prevented frailer patients from failing earlier, while in the placebo group the frailest members failed earliest. Then the frailer members in the treatment group fail in the treatment group when the treatment is stopped.

5.6.3 False protectivity

When there are competing risks, one may observe that a covariate implies a protective effect on an event of interest even when there is no true relationship between the factors. To see why, let's consider the following model. Suppose there are two events that may censor one another, B, and C, with associated time-to-event variables X_{iB} and X_{iB}. Further suppose that the individual level hazard rates are $\xi_B\lambda_B(t)$ and $\xi_C\lambda_C(t)$, and that given (ξ_B, ξ_C) the events are independent of one another. At time t, the survival function is as follows:

$$\begin{aligned}
P(X_{iB} > t, X_{iC} > t) &= \mathbb{E}\left[P(X_{iB} > t, X_{iC} > t \mid \xi_B, \xi_C)\right] \\
&= \mathbb{E}\left[P(X_{iB} > t \mid \xi_B)P(X_{iC} > t) \mid \xi_C)\right] \\
&= \mathbb{E}\left[\exp\left(-\left(\xi_B\Lambda_B(t) + \xi_C\Lambda_C(t)\right)\right)\right]
\end{aligned}$$

If ξ_B and ξ_C are independent, we can take the expectation of each of these functions separately, and any covariate the influences B won't appear to influence C on a population level, and vice versa. However, consider the following model for ξ_B and ξ_C:

$$\begin{aligned}
\xi_B &= W_1 + W_3 \\
\xi_C &= W_2 + W_3 \\
W_1, W_2, W_3 &\overset{\text{iid}}{\sim} \text{Gamma}(\nu)
\end{aligned}$$

Then ξ_B, ξ_C are dependent, but using the results of our Laplace transforms earlier, we can compute the survival probability:

$$\mathbb{E}\left[\exp\left(-\left(\xi_B \Lambda_B(t) + \xi_C \Lambda_C(t)\right)\right)\right] = \mathbb{E}\left[e^{-W_1 \Lambda_B(t)}\right]\mathbb{E}\left[e^{-W_2 \Lambda_C(t)}\right]\mathbb{E}\left[e^{-W_3(\Lambda_B(t)+\Lambda_C(t))}\right] \quad (5.140)$$

$$= \mathcal{L}(\Lambda_B(t))\mathcal{L}(\Lambda_C(t))\mathcal{L}(\Lambda_B(t) + \Lambda_C(t)) \quad (5.141)$$

If we want the density of time to failure at t from B, we can compute:

$$-\frac{\partial}{\partial u} P(X_{iB} > u, X_{iC} > t) \mid_{u=t}$$

This can be done directly from the Laplace transforms

$$-\frac{\partial}{\partial u}\mathcal{L}(\Lambda_B(u))\mathcal{L}(\Lambda_C(t))\mathcal{L}(\Lambda_B(u) + \Lambda_C(t)) \mid_{u=t} \quad (5.142)$$

$$= -\lambda_B(t)\mathcal{L}(\Lambda_C(t))\left(\mathcal{L}'(\Lambda_B(t))\mathcal{L}(\Lambda_B(h) + \Lambda_C(t)) + \mathcal{L}(\Lambda_B(t))\mathcal{L}'(\Lambda_B(h) + \Lambda_C(t))\right) \quad (5.143)$$

In order to get the population hazard ratio, we divide the expression by $\mathcal{L}(\Lambda_B(t))\mathcal{L}(\Lambda_C(t))\mathcal{L}(\Lambda_B(t)+\Lambda_C(t))$ to get:

$$\mathbb{E}\left[\lambda_{iB}(t)\right] = -\lambda_B(t)\left(\frac{\mathcal{L}'(\Lambda_B(t))}{\mathcal{L}(\Lambda_B(t))} + \frac{\mathcal{L}'(\Lambda_B(h) + \Lambda_C(t))}{\mathcal{L}(\Lambda_B(h) + \Lambda_C(t))}\right) \quad (5.144)$$

We finally use our gamma assumption from above to get the following expression:

$$\mathbb{E}\left[\lambda_{iB}(t)\right] = -\lambda_B(t)\left((1 + \nu\Lambda_B(t))^{-\frac{1}{\nu}} + (1 + \nu\left(\Lambda_B(t) + \Lambda_C(t)\right))^{-\frac{1}{\nu}}\right) \quad (5.145)$$

Consider comparing population rates of occurence of event B for two levels of a binary covariate Z that positively influences the rate of occurence of event C but not of B. That is $\lambda_C(t, z) = \lambda_C(t)e^{\beta z}$ for $\beta > 0$ and $\lambda_B(t)$ does not depend on z. The population ratio is as follows:

$$\frac{\left(\mathbb{E}\left[\lambda_{iB}(t)\right]\right)_{Z=1}}{\left(\mathbb{E}\left[\lambda_{iB}(t)\right]\right)_{Z=0}} = \frac{(1 + \nu\Lambda_B(t))^{-\frac{1}{\nu}} + (1 + \nu\left(\Lambda_B(t) + e^{\beta}\Lambda_C(t)\right))^{-\frac{1}{\nu}}}{(1 + \nu\Lambda_B(t))^{-\frac{1}{\nu}} + (1 + \nu\left(\Lambda_B(t) + \Lambda_C(t)\right))^{-\frac{1}{\nu}}} \quad (5.146)$$

$$\leq 1 \quad (5.147)$$

by virtue of the fact that

$$(1 + \nu\left(\Lambda_B(t) + e^{\beta}\Lambda_C(t)\right))^{-\frac{1}{\nu}} < (1 + \nu\left(\Lambda_B(t) + \Lambda_C(t)\right))^{-\frac{1}{\nu}}$$

This may seem like an academic exercise, but the real-world example of this sort of false protectivity can have dire implications if it is not correctly ascertained. Di Serio 1997 gives the example of the potential false protectivity in bone marrow transplant recipients for leukemia. In this example, there are two events that can occur after transplant: death from complications related to the transplant, such as the body rejecting the transplant (or the transplant rejecting the recipient), and relapse from leukemia. The covariate of interest is called HLA disparity, or a measure of genetic mismatch in bone marrow transplant donors and respective recipients. HLA disparity is negative in that it increases the risk of a transplant being rejected. It has been shown, however, that there is a counterintuitive protective effect from HLA disparity against the relapse of leukemia. If this is a true effect, it would have policy implications on who is selected to be a donor for a leukemia patients.

Instead, it is not possible to rule out false protectivity, namely that HLA disparity increases the risk of death from complications of transplant, and thus appears to reduce the risk of leukemia relapse in survivors. Di Serio 1997 posits that it may be related to the health of the bone marrow after transplantation, which is unobservable. If the marrow is in a healthy state, it may be more likely to result in complications after transplant if there is an HLA disparity. On the contrary, if the marrow were in an unhealthy state under HLA disparity, there might not be complications post transplant, but instead the relapse from leukemia could be more likely.

5.6.4 Frailty and influence functions

Remember back to Section 5.4.3 where we derived an expression for the influence function:

$$\left(-\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})}\right)^{-1} \nabla_{\boldsymbol{\theta}} \log(f_{\boldsymbol{\theta}}(t_i, \delta_i, \mathbf{z}_i)) \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{1})} .$$

For a survival model with a frailty term, we have the following marginal likelihood:

$$f_i(t,\delta) = (\exp(\mathbf{z}_i^T\boldsymbol{\beta})\lambda_0(t))^{\delta} \int_0^{\infty} w^{\delta} \exp(-w\Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta}))f(w)dw \tag{5.148}$$

Taking logs of this expression gives us

$$\log f_i(t,\delta) = \delta(\mathbf{z}_i^T\boldsymbol{\beta} + \log\lambda_0(t)) + \log\left(\int_0^{\infty} w^{\delta} \exp(-w\Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta}))f(w)dw\right) \tag{5.149}$$

Taking gradients of both sides with respect to $\boldsymbol{\beta}$ gives

$$\nabla_{\boldsymbol{\beta}} \log f_i(t,\delta) = \mathbf{z}_i\left(\delta - \frac{\int_0^{\infty} w^{1+\delta}\Lambda_0(t)\exp(-w\Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta}) + \mathbf{z}_i^T\boldsymbol{\beta})f(w)dw}{\int_0^{\infty} w^{\delta}\exp(-w\Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta}))f(w)dw}\right) \tag{5.150}$$

simplifying a bit to

$$\nabla_{\boldsymbol{\beta}} \log f_i(t,\delta) = \mathbf{z}_i\left(\delta - \Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\frac{\int_0^{\infty} w^{1+\delta}\exp(-w\Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta}))f(w)dw}{\int_0^{\infty} w^{\delta}\exp(-w\Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta}))f(w)dw}\right) \tag{5.151}$$

Compare this to the non-frailty version of the model:

$$f_i(t,\delta) = (\exp(\mathbf{z}_i^T\boldsymbol{\beta})\lambda_0(t))^\delta \exp(-\Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta})) \tag{5.152}$$

$$\log f_i(t,\delta) = \delta(\mathbf{z}_i^T\boldsymbol{\beta} + \log\lambda_0(t)) - \Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta}) \tag{5.153}$$

with gradients

$$\nabla_{\boldsymbol{\beta}} \log f_i(t,\delta) = \mathbf{z}_i\left(\delta - \Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\right) \tag{5.154}$$

If $\xi \sim \text{Gamma}(\theta,k)$, we get

$$\begin{aligned}\int_0^\infty w^\delta \exp(-w\Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta}))f(w)dw &= \frac{k^\theta}{\Gamma(\theta)}\int_0^\infty \exp(-w(\Lambda_0(t)e^{\mathbf{z}_i^T\boldsymbol{\beta}} + k))w^{\theta+\delta-1}dw \\ &= \frac{k^\theta}{\Gamma(\theta)}\frac{\Gamma(\theta+\delta)}{(k+\Lambda_0(t)e^{\mathbf{z}_i^T\boldsymbol{\beta}})^{\theta+\delta}} \\ &= \left(\frac{\theta}{k}\right)^\delta\left(\frac{k}{k+\Lambda_0(t)e^{\mathbf{z}_i^T\boldsymbol{\beta}}}\right)^{\theta+\delta}\end{aligned}$$

and, noting that: $\Gamma(\theta+1+\delta) = (\theta+\delta)\Gamma(\theta+\delta) = (\theta+\delta)\theta^\delta\Gamma(\theta)$

$$\begin{aligned}\int_0^\infty w^{\delta+1} \exp(-w\Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta}))f(w)dw &= \frac{k^\theta}{\Gamma(\theta)}\int_0^\infty \exp(-w(\Lambda_0(t)e^{\mathbf{z}_i^T\boldsymbol{\beta}} + k))w^{\theta+1+\delta-1}dw \\ &= \frac{k^\theta}{\Gamma(\theta)}\frac{\Gamma(\theta+1+\delta)}{(k+\Lambda_0(t)e^{\mathbf{z}_i^T\boldsymbol{\beta}})^{\theta}} \\ &= \frac{(\theta+\delta)\theta^\delta}{k^{1+\delta}}\left(\frac{k}{k+\Lambda_0(t)e^{\mathbf{z}_i^T\boldsymbol{\beta}}}\right)^{\theta+1+\delta}\end{aligned}$$

The ratio of these two expressions gives:

$$\frac{\theta+\delta}{k+\Lambda_0(t)e^{\mathbf{z}_i^T\boldsymbol{\beta}}} \tag{5.155}$$

Incorporating this into the expression above and substituting $k = \theta = \nu^{-1}$ we get:

$$\nabla_{\boldsymbol{\beta}} \log f_i(t,\delta) = \mathbf{z}_i\left(\delta - \Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\frac{1+\nu\delta}{1+\nu\Lambda_0(t)e^{\mathbf{z}_i^T\boldsymbol{\beta}}}\right) \tag{5.156}$$

We can see that when $\delta = 1$ the term:

$$\frac{1+\nu\delta}{1+\nu\Lambda_0(t)e^{\mathbf{z}_i^T\boldsymbol{\beta}}}$$

shrinks the cumulative hazard term $\Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta})$ towards zero if the cumulative hazard term is greater than 1 and away from zero otherwise. When $\delta = 0$ the term shrinks the $\Lambda_0(t)\exp(\mathbf{z}_i^T\boldsymbol{\beta})$ towards zero. This should act to shrink extreme residuals towards zero.

5.6.5 Cox PH with omitted variables

This is an important point when thinking about the Cox proportional hazards model and omitted variable bias. We can think of this problem similarly to frailty. Suppose the true hazard function is the following:

$$\lambda_i(t) = \lambda_0(t)\exp(z_{i1}\beta_1 + z_{i2}\beta_2) \tag{5.157}$$

but that we don't observe z_{i2}. Then we can think of the observed model as a frailty model:

$$\lambda_i(t) = \xi_i\lambda_0(t)\exp(z_{i1}\beta_1) \tag{5.158}$$

where $\xi_i = e^{z_{i2}\beta_2}$. If we suppose that z_{i2} has a population distribution, then the marginal hazard function is what we'll be inferring when fitting a Cox model:

$$\mathbb{E}\left[\lambda_i(t) \mid z_{i1}\right] = \mathbb{E}\left[e^{z_{i2}\beta_2} \mid z_{i1}\right]\lambda_0(t)\exp(z_{i1}\beta_1) \tag{5.159}$$

Assuming that $\beta_2 \neq 0$, we can use our results from above

$$\mathbb{E}\left[\lambda_i(t) \mid z_{i1}\right] = -\lambda_0(t)\exp(z_{i1}\beta_1)\frac{\mathcal{L}_{e^{z_{i2}\beta_2}|z_{i1}}(\Lambda(t))'}{\mathcal{L}_{e^{z_{i2}\beta_2}|z_{i1}}(\Lambda(t))}$$

This will not typically be a proportional hazards model anymore. Suppose, for argument's sake, that $e^{z_{i2}\beta_2} \mid z_{i1}$ was gamma distributed with a rate parameter depending on z_{i1}. Then

$$\mathbb{E}\left[\lambda_i(t) \mid z_{i1}\right] = \frac{\theta\lambda_0(t)e^{z_{i1}\beta_1}}{k(z_{i1}) + \Lambda_0(t)e^{z_{i1}\beta_1}}.$$

and

$$\frac{\mathbb{E}\left[\lambda_i(t) \mid z_{i1}\right]}{\mathbb{E}\left[\lambda_j(t) \mid z_{j1}\right]} = e^{(z_{i1}-z_{j1})\beta_1}\frac{k(z_{j1}) + \Lambda_0(t)e^{z_{j1}\beta_1}}{k(z_{i1}) + \Lambda_0(t)e^{z_{i1}\beta_1}}$$

This is clearly not a PH model, so any PH model we use will result in biased inferences and bad coverage, though you can create scenarios in which the bias isn't too bad, and coverage can be corrected by using the sandwich covariance estimator.

Chapter 6

Appendix

6.1 Map between Weibull parameterizations

Our course notes (and Klein, Moeschberger, et al. 2003) define the Weibull hazard as:

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

Base `R` defines the Weibull parameterization for `rweibull(n, shape=`α`, scale=`σ`)` as

$$\lambda(t) = \frac{\alpha}{\sigma}\left(\frac{t}{\sigma}\right)^{\alpha-1}$$

The `survival` package parameterizes the Weibull, with `intercept=`μ`, scale = `τ, as

$$\lambda(t) = \frac{1}{\tau e^{\mu/\tau}} t^{1/\tau-1}$$

Thus, we can see that the following identities hold:

$$\gamma = \frac{1}{\sigma^\alpha} \implies \sigma = \frac{1}{\gamma^{1/\alpha}}$$
$$\gamma = e^{-\mu/\tau} \implies \mu = -\tau\log(\gamma)$$

This also implies that regression coefficients from `survreg` are interpreted differently from the typical interpretation from a proportional hazards model. The proportional hazards Weibull model is typically written

$$\gamma e^{\boldsymbol{\beta}^T\mathbf{z}_i}\alpha t^{\alpha-1}$$

But `survreg` parameterizes the model as

$$\frac{1}{\tau e^{(\mu+\boldsymbol{\theta}^T\mathbf{z}_i)/\tau}} t^{1/\tau-1}$$

This means that:

$$\boldsymbol{\beta} = -\boldsymbol{\theta}/\tau$$
$$\gamma = e^{-\mu/\tau}$$

Thus, a positive coefficient in the parametric hazard which indicates that the variable increases hazard, all else being equal, will be negative in `survreg`'s coefficient results and vice versa.

6.2 Summary of tests

We've covered many tests throughout the course, and we'll review the tests, the assumptions and the hypotheses for these.

6.2.1 Log-rank test

- Descrption: The log-rank test is used to compare survival time distributions of individuals i split between two groups, $j = 1, 2$, via a test of the hazard ratios.
- Assumptions:
  1. Times to failure X_i are indepdently distributed, with censoring C_i that is independent of X_i, and, if random, is distributed with parameters that are variationally independent of the parameters governing X_i.
  2. Observed times $T_i = \min(X_i, C_i)$, with $\delta_i = \mathbb{1}(T_i = X_i)$.
  3. With ties: $d_i = \sum_{k=1}^{n} \delta_k \mathbb{1}(t_i = t_k)$.
- Null hypothesis:
  $$H_0 : \lambda_1(t) = \lambda_2(t) \forall t \in [0, \tau]$$
  where τ is a prespecified time.
- Notation: Let g_i be the categorical variable for each individual denoting group membership, $g_i \in \{1, 2\}$.

$$\bar{Y}(t_i) = \sum_{k=1}^{n} \mathbb{1}(t_k \geq t_i), \tag{6.1}$$
$$\bar{Y}_j(t_i) = \sum_{k=1}^{n} \mathbb{1}(t_k \geq t_i) \mathbb{1}(g_i = j), \tag{6.2}$$
$$d_{ij} = \sum_{k=1}^{n} \delta_k \mathbb{1}(t_i = t_k) \mathbb{1}(g_i = j). \tag{6.3}$$

- Test Statistic:

$$Z_j(\tau) = \sum_{i=1|t_i\leq\tau}^{n_1+n_2} \left(d_{ij} - d_i \frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)} \right) \tag{6.4}$$

$$\mathrm{Var}\left(Z_j(\tau)\right) = \sum_i \left(d_i \frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)} \left(1 - \frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\right) \frac{\overline{Y}(t_i) - d_i}{\overline{Y}(t_i) - 1} \right) \tag{6.5}$$

- Asymptotic distribution of the test statistic:

$$\left(\frac{Z_j(\tau)}{\sqrt{\mathrm{Var}\left(Z_j(\tau)\right)}} \right)^2 \xrightarrow{d} \chi^2(1)$$

Extension of log-rank test

There are various extensions of the log-rank test that add a positive weight function $W(t_i)$ to the test statistic in order to boost the power of the test statistic under different alternative hypotheses:

$$Z_j(\tau) = \sum_{i=1|t_i\leq\tau}^{n_1+n_2} W(t_i) \left(d_{ij} - d_i \frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)} \right) \tag{6.6}$$

$$\mathrm{Var}\left(Z_j(\tau)\right) = \sum_i W(t_i)^2 \left(d_i \frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)} \left(1 - \frac{\overline{Y}_j(t_i)}{\overline{Y}(t_i)}\right) \frac{\overline{Y}(t_i) - d_i}{\overline{Y}(t_i) - 1} \right) \tag{6.7}$$

Some examples of these are given below, with the test name followed by the weight function:

1. Gehan: $W(t_i) = \bar{Y}(t_i)$

2. Peto-Peto: $W(t_i) = S^{KM}(t_i)$, where $S^{KM}(t_i)$ is the Kaplan-Meier estimator derived from the pooled sample.

3. Fleming-Harrington: $W(t_i) = (S^{KM}(t_i))^p(1 - S^{KM}(t_i))^q$ for $p, q \geq 0$, many of which $p + q = 1$.

6.2.2 Likelihood based tests

For parametric and semiparametric models indexed by a parameter $\theta \in \Theta$ we studied likelihood ratio tests, Wald tests and likelihood ratio tests.

All of these tests were derived under the following assumptions:

1. $X_i \overset{\text{iid}}{\sim} f_X(x \mid \theta^\dagger)$

2. $\{x \mid f_X(x \mid \theta) > 0\}$ does not depend on θ.

3. $\theta^\dagger$ is an interior point of $\omega \subseteq \Theta$, of dimension p

4. The Fisher information for every θ exists, and is equal to the variance of the score and the expected negative matrix of second derivatives

5. In ω the density admits all third derivatives

6. The Fisher information for every θ is positive definite and $\|I(\theta)\|_\infty < \infty$

7. All third derivatives are bounded by a finitely integrable function of X, $M(X)$,

Under these assumptions, we can derive the following tests under the null hypothesis:

$$H_0 : X_i \overset{\text{iid}}{\sim} f_X(x \mid \theta_0)$$

For the Wald test and the Score tests we'll use the observed information, which is defined as:

$$i(\theta) \equiv -\nabla_\theta^2 \sum_{i=1}^n \log f_X(x_i \mid \theta).$$

LRT

$$T_{LRT} = 2(\ell(\hat{\theta}) - \ell(\theta_0))$$

Wald test with the observed information

$$T_W = (\hat{\theta}_n - \theta_0)^T i(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)$$

Score test with the observed information

$$T_S = \left(\nabla_\theta \ell(\theta) \mid_{\theta=\hat{\theta}_0}\right)^T i(\hat{\theta}_0)^{-1} \nabla_\theta \ell(\theta) \mid_{\theta=\hat{\theta}_0}$$

All tests are asymptotically $\chi^2(p)$.

Composite tests

We'll also give the composite versions of the tests, where we have nuisance parameters ϕ and parameters of interest ψ such that $\theta = (\psi, \phi)$, where ψ is dimension k and θ is dimension p.

Composite LRT

$$T_{LRT} = 2(\ell(\hat{\psi}, \hat{\phi}) - \ell(\psi_0, \hat{\phi}(\psi_0)))$$

Composite wald test with the observed information

$$T_W = (\hat{\psi}_n - \psi_0)^T \left(i^{\psi,\psi} \mid_{\psi=\hat{\psi}_n, \phi=\hat{\phi}_n} \right)^{-1} (\hat{\psi}_n - \psi_0)$$

where

$$i^{\psi,\psi} \mid_{\psi=\hat{\psi}_n, \phi=\hat{\phi}_n}$$

is the block of the inverse observed information matrix, that corresponds to the parameters ψ.

Composite score test with the observed information

$$T_S = \nabla_\psi \ell(\theta) \mid_{\psi=\psi_0, \phi=\hat{\phi}(\psi_0)} i(\psi_0, \hat{\phi}(\psi_0))^{\psi,\psi} \nabla_\psi \ell(\theta) \mid_{\psi=\psi_0, \phi=\hat{\phi}(\psi_0)}$$

$$i(\psi_0, \hat{\phi}(\psi_0))^{\psi,\psi} = i^{\psi,\psi} \mid_{\psi=\psi_0, \phi=\hat{\phi}_n(\psi_0)}$$

is the block of the inverse information matrix corresponding to the parameters ψ evaluated at ψ_0 and $\hat{\phi}_n(\psi_0)$.

All of these are asymptotically $\chi^2(k)$.

6.2.3 Comments and

There is a duality between the score test under the null hypothesis $\beta = 0$ for a Cox PH model fitted to two groups and the log-rank test.

Bibliography

[1] John P Klein, Melvin L Moeschberger, et al. *Survival analysis: techniques for censored and truncated data*. Vol. 1230. Springer, 2003.

[2] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

[3] Thomas R Fleming and David P Harrington. "Counting Processes and Survival Analysis". In: *Wiley Series in Probability and Statistics* (2005).

[4] Sidney Resnick. *A probability path*. Springer, 2019.

[5] Jasmin Rühl, Jan Beyersmann, and Sarah Friedrich. "General independent censoring in event-driven trials with staggered entry". en. In: *Biometrics* 79.3 (2023), pp. 1737–1748. ISSN: 0006-341X, 1541-0420. DOI: 10.1111/biom.13710.

[6] Frank E Harrell et al. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Vol. 608. Springer, 2001.

[7] David Collett. *Modelling survival data in medical research*. Chapman & Hall, 1994.

[8] Robert W. Keener. *Theoretical Statistics*. en. Springer Texts in Statistics. New York, NY: Springer New York, 2010. ISBN: 978-0-387-93838-7 978-0-387-93839-4. DOI: 10.1007/978-0-387-93839-4.

[9] E. L. Lehmann and George Casella. *Theory of point estimation*. en. 2nd ed. Springer texts in statistics. New York: Springer, 1998. ISBN: 978-0-387-98502-2.

[10] Odd O. Aalen. "Heterogeneity in survival analysis". en. In: *Statistics in Medicine* 7.11 (Nov. 1988), pp. 1121–1137. ISSN: 02776715, 10970258. DOI: 10.1002/sim.4780071105.

[11] Kevin C. Cain and Nicholas T. Lange. "Approximate Case Influence for the Proportional Hazards Regression Model with Censored Data". en. In: *Biometrics* 40.2 (June 1984), p. 493. ISSN: 0006341X. DOI: 10.2307/2531402.

[12] Tamara Broderick, Ryan Giordano, and Rachael Meager. *An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?* en. arXiv:2011.14999 [econ, stat]. July 2023.

[13] Gaineford J Hall, William H Rogers, and Daryl Pregibon. *Outliers matter in survival analysis*. Vol. 6761. Rand Corporation, 1982.

[14] J. D. Kalbfleisch and Ross L. Prentice. *The statistical analysis of failure time data*. en. 2nd ed. Wiley series in probability and statistics. Hoboken, N.J: J. Wiley, 2002. ISBN: 978-0-471-36357-6.

[15] Anastasios A. Tsiatis and Marie Davidian. "Joint Modeling of Longitudinal and time-to-event data: an overview". In: *Statistica Sinica* 14 (2004), pp. 809–834.

[16] Clelia Di Serio. "The protective impact of a covariate on competing failures with an example from a bone marrow transplantation study". In: *Lifetime data analysis* 3 (1997), pp. 99–122.