Biostatistics Lecture Notes

Survival Analysis

Ming-Chieh Shih

April 19, 2023

1 Mathematical Formulation of Survival Data

Survival analysis, or time-to-event analysis, is a branch of statistical modelling that focuses on the time T needed for an event (or a set of events) to occur, eg. the time from cancer diagnosis until death, or the time from the manufacture a bulb to its failure. This time will always be non-negative, so we have $T \geq 0$. In the following, we assume that T is also continuous if not specified otherwise.

Since T is a random variable, it can be characterized by its cumulative density function (cdf) and probability density function (pdf):

$$\text{Cumulative density function: } F(t) = \mathbb{P}[T \leq t] \tag{1}$$

$$\text{Probability density function: } f(t) = \frac{\partial}{\partial t} F(t) \tag{2}$$

$$= \lim_{\Delta t \to 0^+} \frac{F(t+\Delta t) - F(t)}{\Delta t} \tag{3}$$

$$= \lim_{\Delta t \to 0^+} \frac{\mathbb{P}[T \leq t+\Delta t] - \mathbb{P}[T \leq t]}{\Delta t} \tag{4}$$

$$= \lim_{\Delta t \to 0^+} \frac{\mathbb{P}[T \in (t, t+\Delta t]]}{\Delta t} \tag{5}$$

We first look at the cdf, $F(t)$. By definition, $F(t)$ will be non-decreasing, right-continuous and since T is non-negative,

$$F(0) = \mathbb{P}(T = 0) = 0; \qquad F(\infty) = 1 \tag{6}$$

The meaning of $F(t)$ is *the probability of only surviving up to time t*. In survival analysis, we do not usually work with $F(t)$, but works with the survival function, which means *the probability of surviving past time t*:

$$\text{Survival function: } S(t) := \mathbb{P}[T > t] = 1 - F(t) \tag{7}$$

which should be non-increasing, right-continuous and since T is non-negative,

$$S(0) = 1 - \mathbb{P}(T = 0) = 1; \qquad S(\infty) = 0 \tag{8}$$

In addition, although the definition of $f(t)$ show that it represents the density of events at time t, often it is more intuitive to consider the *conditional* density of events defined as follows:

$$\text{Hazard function: } h(t) := \lim_{\Delta t \to 0^+} \frac{\mathbb{P}[T \in [t, t+\Delta t) \mid T \geq t]}{\Delta t} \tag{9}$$

$$= \lim_{\Delta t \to 0^+} \frac{\mathbb{P}[T \in [t, t+\Delta t) \ \cap \ T \geq t]}{\mathbb{P}(T \geq t)\Delta t} \tag{10}$$

$$= \left[\lim_{\Delta t \to 0^+} \frac{\mathbb{P}[T \in [t, t+\Delta t)]}{\Delta t}\right] \frac{1}{S(t^-)} \tag{11}$$

$$= \frac{f(t)}{S(t)} \tag{12}$$

where in the last equality we use the assumption that T is continuous. Note that the meaning of $h(t)$ is the density of events *among the subjects alive right before time t*, which has three main advantages compared to $f(t)$:

1. **Explainability:** $f(t)$ looks at the density of events among the whole population, which may not fit common intuition. To spell this out, suppose we are modelling the survival of the general population, with time measured in years. Then $f(100)$ would be much smaller than $f(70)$, since it is much more likely for a person to have a lifespan of 70 years old than 100 years old. However, $h(100)$ would be much larger than $h(70)$, since it is far more likely that a 100-year-old person dies within the next year than a 70-year-old person. Therefore, when we are saying that older people are more likely to decease, we are actually talking about hazards but not plain probability density.

2. **Stable as modelling target:** In biomedical studies, most of the $h(t)$ varies relatively slowly in t, while $f(t)$ can have large fluctuations. Therefore, it is easier to find a concise model with respect to $h(t)$ than to $f(t)$. As an example, if the event is occurring homogeneously and randomly, $h(t)$ will be just a constant, while $f(t)$ will be related to exponential functions.

3. **Simplifying estimation:** Non-parametric estimation in survival analysis often relies on calculating the probability of event among subjects at risk (i.e. alive before a certain time point), which corresponds to the definition of hazard functions. In addition, the cumulative hazard function, defined as

$$\text{Cumulative hazard function: } H(t) := \int_0^t h(s)\, ds \tag{13}$$

, has good analytical properties and is extensively used in the theoretical development of survival analysis.

Based on these newly defined functions, we can derive the following identities:

$$h(t) = -\frac{d}{dt}\log S(t) \tag{14}$$
$$S(t) = e^{-H(t)} \tag{15}$$
$$H(T) \sim Exp(1) \tag{16}$$

To prove Equation (14), we have

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1-F(t)} = -\frac{-f(t)}{1-F(t)} = -\frac{\frac{d}{dt}(1-F(t))}{1-F(t)} = -\frac{d}{dt}\log(1-F(t)) = -\frac{d}{dt}\log S(t) \tag{17}$$

Integrating both sides above from 0 to t, we can then prove Equation (15)

$$H(t) = \int_0^t h(s)ds = \int_0^t -\frac{d}{ds}\log S(s)ds = -\log S(t) + \log S(0) = -\log S(t) \tag{18}$$
$$\Rightarrow \quad S(t) = e^{-H(t)} \tag{19}$$

For Equation (16), we have

$$\mathbb{P}[H(T) > s] = \mathbb{P}[-\log S(T) > s] \tag{20}$$
$$= \mathbb{P}[S(T) < e^{-s}] \tag{21}$$
$$= \mathbb{P}[T > S^{-1}(e^{-s})] \tag{22}$$
$$= S(S^{-1}(e^{-s})) = e^{-s} \tag{23}$$

which is the survival function of exponential distribution with rate 1, as we will later show.

2 Common Distributions For Survival Time Modelling

Exponential distribution

An exponential distribution with rate parameter $\lambda > 0$ has the cdf and pdf

$$F(t) = 1 - e^{-\lambda t} \tag{24}$$
$$f(t) = \lambda e^{-\lambda t} \tag{25}$$

The survival, hazard and cumulative hazard functions are thus

$$S(t) = e^{-\lambda t} \tag{26}$$
$$h(t) = \lambda \tag{27}$$
$$H(t) = \lambda t \tag{28}$$

Therefore, the exponential distribution assumes *constant hazards*, which is the case when the event randomly occurs homogeneously across time at rate λ. This leads to the *Memoryless property* of exponential distributions, where for all $t_0 \geq 0$

$$\mathbb{P}[T > t] = \mathbb{P}[T > t + t_0 | T > t_0] \tag{29}$$

In addition, since λ is the rate of event occurrence, we can see $Exp(\lambda)$ as "fastforwarding" the time of X by λ times. Therefore, the event occurrence time would be shrinked by a factor of $\frac{1}{\lambda}$ compared to $Exp(1)$, so we have

$$Exp(\lambda) \sim \frac{1}{\lambda} Exp(1) \tag{30}$$

or more generally

$$Exp(k\lambda) \sim \frac{1}{k} Exp(\lambda) \tag{31}$$

Previously we have shown that exponential distribution belongs to the exponential family. Therefore, with properly assigned link function, GLMs can also be used in modelling exponentially distribution survival time.

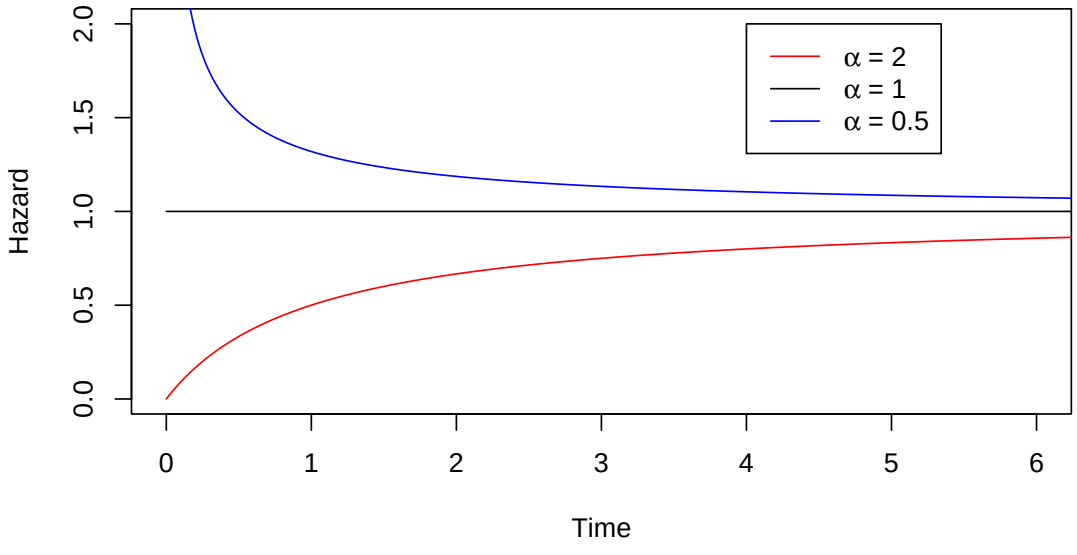
Gamma distribution

Exponential distributions are in essence Gamma distributions with shape parameter as 1. Therefore, to relax the assumption of constant hazards, we may model the survival time with Gamma distributions of shape parameter α, which has the following pdf

$$f(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \, t^{\alpha - 1} \, e^{-\lambda t} \tag{32}$$

Its cdf does not have a close form, so its survival function, hazard function and cumulative hazard function all do not have closed forms. We can still numerically plot the hazard function (with rate parameter 1) and see that when $0 < \alpha < 1$, the hazards is monotonically decreasing, and when $\alpha > 1$, the hazards in monotonically increasing. When $\alpha = 1$, Gamma distribution is exactly the exponential distribution and thus has constant hazards. Also notice in the plot that the hazards converges to $1 = \lambda$, no matter what the value of α is. This may not correspond to some biomedical survival outcomes where we expect the hazards to grow indefinitely with time.

Since Gamma distributions with known shape parameter also belongs to the exponential family, so the GLM framework still applies.



Weibull distribution

Weibull distribution is another extension of exponential distribution that aims to relax the assumption of constand hazards by defining (letting $p > 0$)

$$Weibull(p, \lambda) \sim [Exp(\lambda^p)]^{1/p} \tag{33}$$

Based on this definition, we have the following cdf and pdf,

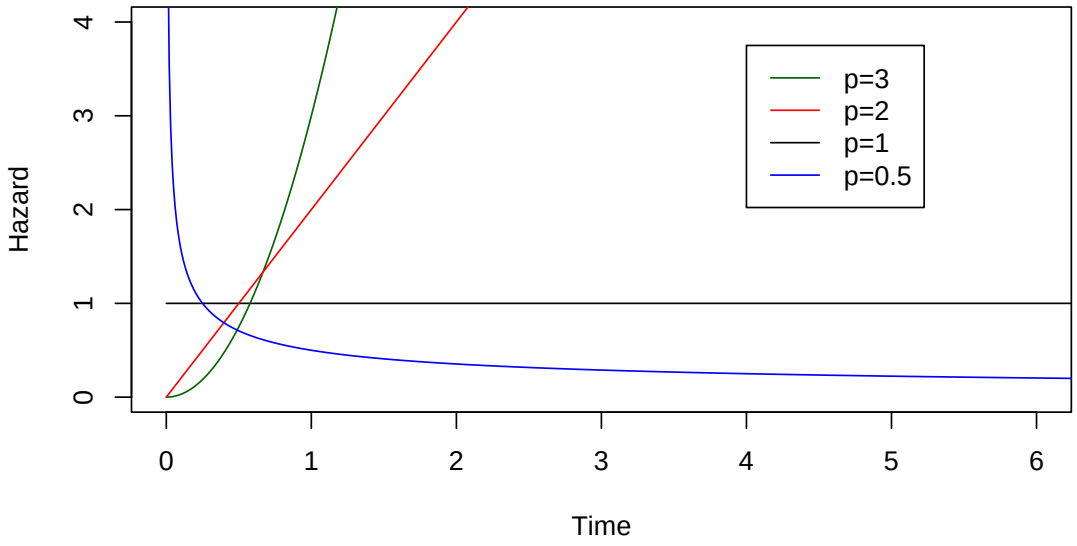
$$F(t) = 1 - e^{-(\lambda t)^p} \tag{34}$$
$$f(t) = p\lambda^p t^{p-1} e^{-(\lambda t)^p} \tag{35}$$

Therefore we have,

$$h(t) = p\lambda^p t^{p-1} \tag{36}$$
$$H(t) = (\lambda t)^p \tag{37}$$

Therefore, under Weibull distribution, the hazards is proportional to a power of time, t^{p-1}. When $0 < p < 1$, the hazards monotonically decreases and goes to zero; when $p > 1$, the hazards monotonically increases and goes to infinity. When $p = 1$, Weibull distribution become exponential distribution with constant hazards. Also, with this parameterization, Weibull distribution has a similar identity as exponential distribution, where

$$Weibull(p, k\lambda) \sim \frac{1}{k} Weibull(p, \lambda) \tag{38}$$



Generalized Gamma distribution

Generalized Gamma distribution unifies the generalization done by Gamma distribution and Weibull distribution, so that we have the pdf,

$$f(t) = \frac{\lambda^\alpha}{\Gamma(\alpha/p)} p t^{\alpha-1} e^{-(\lambda t)^p} \tag{39}$$

When $p = 1$, the distribution becomes a Gamma distribution with shape parameter α. When $\alpha = 1$, the distribution becomes a Weibull distribution with shape parameter p. When both parameters are 1, the distribution becomes an exponential. Under this parameterization, we may use likelihood ratio test to test the adequacy of these models.

Lognormal distribution

In previous distributions, the hazard function is always monotonic, and the shape parameters determine if the function is monotonically increasing or decreasing. In the case where the hazards is not monotonic, we may try the lognormal distribution, which is defined as

$$LN(t) \sim \exp[N(\mu, \sigma^2)] \tag{40}$$

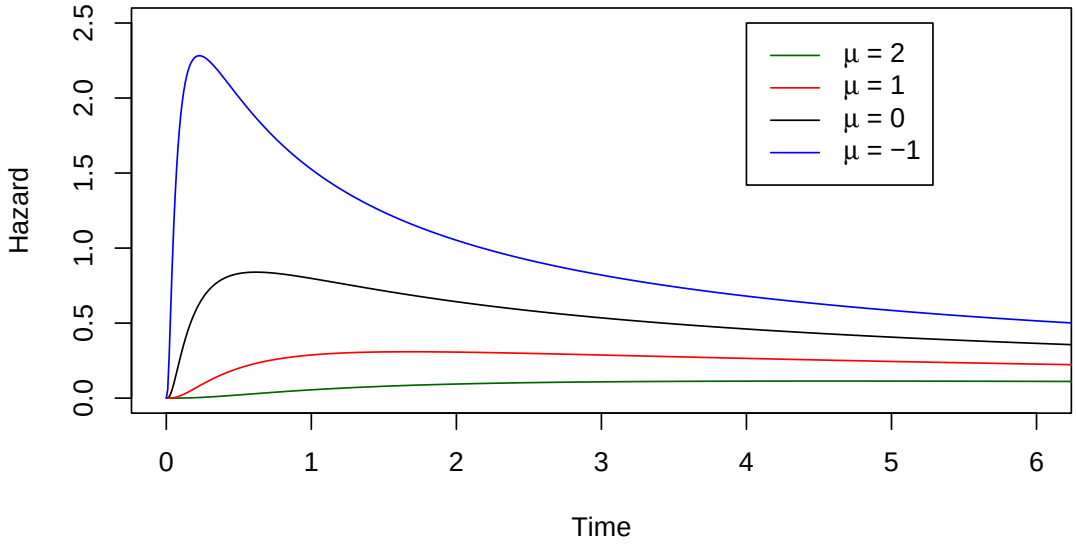
which implies the following pdf, cdf and hazard functions

$$f(t) = \frac{1}{\sqrt{2\pi\sigma t}} \exp\Big(- \frac{(\log t - \mu)^2}{2\sigma^2} \Big) \tag{41}$$

$$F(t) = \Phi\Big(\frac{\log t - \mu}{\sigma} \Big) \tag{42}$$

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{43}$$

As shown in the figure below, lognormal distribution assumes a hazard function that rises up initially and then drops back to zero as time progresses. A smaller μ indicates a more aggressive change in hazards, while σ changes the scale of the distribution.



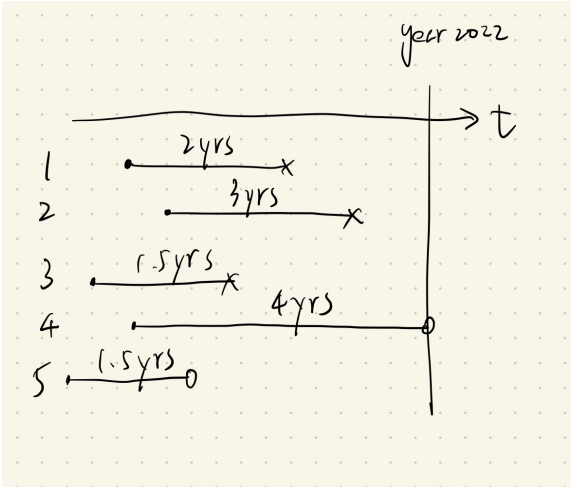
3 Mathematical Formulation of Censoring

If survival analyses were just dealing with non-negative outcomes, it would not be too complicated a problem since we can always try to transform the outcome or use the GLM framework we have just learned. In fact, if we are willing to assume that the outcome is distributed as lognormal, we can just transform the survival time with logarithm and proceed on with linear regression. However, another common feature in survival data is the presence of *censoring*, which prevents us from observing the true event time T, but can only know its bounds.

For example, suppose we have a cohort study that aims to determine the survival of stage 2 colon cancer patients after surgery. The study follows the patients until the end of year 2022 and records the date of their death. However, after the study ends, there will be still patients that are alive. To illustrate this, we have a graph showing five patients in the cohort, where the filled dots are the time each patient was recruited into the cohort, and the Xs are the time each patient died.

For patients 1, 2 and 3, we know their exact survival times as 2, 3 and 1.5 years. However, for patient 4, the study ended before they died, so we only know that their survival time was greater than 4 years, or $T_4 > 4$. For patient 5, after following for 1.5 years, they dropped out of the study for certain reasons, so they we could not know if they died between their drop-out date and the end of 2022. Therefore, we also just know that $T_5 > 1.5$. This kind of censoring is called *right censoring*, which means that we only have a *lower bound* of the true event time. In contrast, if we only have a *upper bound* of the true event time, then this is called *left censoring*. If we know the true event time is *within an interval*, then this is called *interval censoring*. Most of the time, survival data in the biomedical field will present with right censoring, so in the following we will discuss it in depth.

The most common formulation of right censoring is to assume that there are two different "events", out event of interest and the "censoring event." Each subject i would thus have two event times: T_i, the event time for our event of interest, and C_i, the event time for the censoring event. Whenever the censoring event occurs before our event of interest, we can only observe the



censoring time, as well as the fact that the subject has censored. Therefore, what we actually observe is the following pair of variables

$$(Y_i, \Delta_i) = (\min(T_i, C_i), \mathbf{1}(T_i \leq C_i)) \tag{44}$$

where Y_i is the time until end of follow-up for subject i, and Δ_i is the indicator of if the event of interest was observed. If Δ_i was observed to be 0, then subject i has censored at time Y_i.

More often than not, the subjects can be assumed to be independent, so that T_i are mutually independent, and C_i are mutually independent among subjects. However, whether T_i and C_i are independent is then an important assumption. When T_i and C_i are independent, then the censoring is *non-informative*, and there are plethora of methods to infer T without assuming the distribution of C. However, when T_i and C_i are dependent, eg. T_i is the death time of cancer-specific causes and C_i includes the death time due to stroke, then the censoring is mostly *informative*, and it will be impossible to determine the distribution of T based on (Y, Δ) alone. We will talk about non-informative censoring more when we try to construct the likelihood of the data in the next section.

4 Parametric Inference of Right-Censored Survival Data

In the previous section, we formulated right censorship in survival data by first defining the variable pair (T_i, C_i) as the event times for our event of interest (which we shorthand as the *event time*) and censoring event (which we term as the *censoring time*). Since we can only observe exactly one of T_i and C_i, the observable pair of variables we can actually observe is

$$(Y_i, \Delta_i) = (\min(T_i, C_i), \mathbf{1}(T_i \leq C_i)) \tag{45}$$

where Y_i is either the event time or censoring time, whichever comes first, and Δ_i is an indicator for Y_i being the event time.

To use maximum likelihood estimation (MLE) to do inference, we first construct the likelihood for (Y_i, Δ_i). We make the assumption of T and C being independent, and define the following functions:

- $f_{T,C}(t, c)$ as the joint probability density function (pdf) for T and C.
- $f_T(t; \theta_T)$, $S_T(t; \theta_T)$ and $h_T(t; \theta_T)$ as the marginal pdf, survival function and hazard function for T, where θ_T is the parameter vector for the distribution of T.
- $f_C(c; \theta_C)$ and $S_C(c; \theta_C)$ as the marginal pdf and survival function for C, where θ_T is the parameter vector for the distribution of C.

The event $Y_i = y_i, \Delta_i = 1$ is equivalent to $T_i = y_i, C_i \geq y_i$. Therefore we have

$$f_{Y,\Delta}(y_i, \delta_i = 1) = \int_{y_i}^{\infty} f_{T,C}(y_i, c)\ dc \tag{46}$$

$$= \int_{y_i}^{\infty} f_T(y_i; \theta_T) f_C(c; \theta_C)\ dc \tag{47}$$

$$= f_T(y_i; \theta_T) \int_{y_i}^{\infty} f_C(c; \theta_C)\ dc \tag{48}$$

$$= f_T(y_i; \theta_T) S_C(y_i; \theta_C) \tag{49}$$

where Equation (47) is due to independence between T and C. Similarly, the event $Y_i = y_i, \Delta_i = 0$ is equivalent to $C_i = y_i, T_i > y_i$, so we have

$$f_{Y,\Delta}(y_i, \delta_i = 0) = \int_{y_i}^{\infty} f_{T,C}(t, y_i)\ dt \tag{50}$$

$$= \int_{y_i}^{\infty} f_T(t; \theta_T) f_C(y_i; \theta_C)\ dt \tag{51}$$

$$= f_C(y_i; \theta_C) \int_{y_i}^{\infty} f_T(t; \theta_T)\ dt \tag{52}$$

$$= f_C(y_i; \theta_C) S_T(y_i; \theta_T) \tag{53}$$

Combining Equations (49) and (53), we have

$$f_{Y,\Delta}(y_i, \delta_i) := L_i(\theta_T, \theta_C; y_i, \delta_i) = \Big[f_T(y_i; \theta_T)^{\delta_i} S_T(y_i; \theta_T)^{1-\delta_i} \Big] \Big[f_C(y_i; \theta_C)^{1-\delta_i} S_C(y_i; \theta_C)^{\delta_i} \Big] \tag{54}$$

Therefore, the likelihood for the full data set is

$$L_{T,C}(\theta_T, \theta_C) = \Big[\prod_i f_T(y_i; \theta_T)^{\delta_i} S_T(y_i; \theta_T)^{1-\delta_i} \Big] \Big[\prod_i f_C(y_i; \theta_C)^{1-\delta_i} S_C(y_i; \theta_C)^{\delta_i} \Big] \tag{55}$$

Alternatively, using the fact that hazard function $h(t) = f(t)/S(t)$, we have

$$L_{T,C}(\theta_T, \theta_C) = \Big[\prod_i h_T(y_i; \theta_T)^{\delta_i} S_T(y_i; \theta_T) \Big] \Big[\prod_i f_C(y_i; \theta_C)^{1-\delta_i} S_C(y_i; \theta_C)^{\delta_i} \Big] \tag{56}$$

Suppose we are only interested in the inference of θ_T. If θ_T and θ_C do not share the same parameters, the second product term does not contribute to the likelihood-based inference of θ_T, since it is only a constant relative to θ_T in the log-likelihood. Therefore, our inference can be based on only the first product term

$$L(\theta_T) = \prod_i h_T(y_i; \theta_T)^{\delta_i} S_T(y_i; \theta_T) \tag{57}$$

Or, for the log-likelihood

$$\ell(\theta_T) = \sum_i \delta_i \log h_T(y_i; \theta_T) + \log S_T(y_i; \theta_T) \tag{58}$$

Now let's use the above likelihood for inference of exponentially distributed survival time, which is frequently used in epidemiology to obtain the *incidence (rate)* of an event of interest. Suppose the time to event $T_1, T_2, ...T_n$ are *i.i.d.* distributed as $\text{Exp}(\lambda)$, which are subject to non-informative right censoring so that for each subject we observe the pair of random variables (Y_i, Δ_i). For brevity, we denote $R = \sum_i \Delta_i$ as the total number of events observed, and $M = \sum_i Y_i$ as the total follow-up person-time. Since T_i follow *i.i.d.* exponential distributions, we have

$$\log h_T(y_i; \lambda) = \log \lambda \tag{59}$$

$$\log S_T(y_i; \lambda) = \log e^{-\lambda y_i} = -y_i \lambda \tag{60}$$

Therefore, the score function with respect to λ would be

$$U(\lambda) = \frac{\partial}{\partial \lambda} \ell(\lambda) = \frac{\partial}{\partial \lambda} \Big[\sum_i \Delta_i \log \lambda - Y_i \lambda \Big] \tag{61}$$

$$= \frac{\partial}{\partial \lambda} [R \log \lambda - M\lambda] \tag{62}$$

$$= \frac{R}{\lambda} - M \tag{63}$$

The MLE for λ can be obtained by letting $U(\hat{\lambda}) = 0$, so we have

$$\hat{\lambda} = \frac{R}{M} \tag{64}$$

which is the number of events divided by the total follow-up person-time. To calculate the asymptotic variance of $\hat{\lambda}$, we first derive the negative second derivative of the log-likelihood:

$$-\frac{d^2}{d\lambda^2}\ell(\lambda) = -\frac{d}{d\lambda}\left(\frac{R}{\lambda} - M\right) = \frac{R}{\lambda^2} \tag{65}$$

Therefore, we have the Fisher information and observed information

$$\mathcal{I}(\lambda) = \frac{\mathbb{E}[R]}{\lambda^2} = \frac{n\mathbb{P}[T \leq C]}{\lambda^2} := \frac{np}{\lambda^2} \tag{66}$$

$$I(\hat{\lambda}) = \frac{R}{\hat{\lambda}^2} = \frac{M^2}{R} \tag{67}$$

where in Equation (66), R is a sum of *i.i.d.* Bernoulli variables Δ_i, so its expectation value would be np where p is the probability of $\Delta_i = 1$. From the theory of MLE, we would have the asymptotic distribution

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{d} N(0, \lambda^2/p) \tag{68}$$

or, with slight abuse of notation

$$\hat{\lambda} \xrightarrow{d} N(\lambda, \lambda^2/np) \approx N(\lambda, R/M^2) \tag{69}$$

However, this asymptotic distribution is not ideal since (1) λ is a positive number, while normal distribution can take negative values (2) its variance has terms of unknown parameter λ that needs to be approximated with R/M. We can therefore try basing our inference on $\log\lambda$ instead of λ.

One way to infer with respect to $\log\lambda$ is to let $\theta = \log\lambda$, express the log-likelihood with θ, and redo the above all again. Another way is to tweak our current result with respected to λ a little bit, which we will now do. First, since MLEs are *functional invariant*, the MLE of $\log\lambda$ would be $\log\hat{\lambda} = \log(R/M)$. Then, to obtain the asymptotic distribution of $\log\hat{\lambda}$, we use the **delta method**, which states

Theorem 1 (Univariate delta method) *Suppose a sequence of one-dimensional random variables $\hat{\theta}_n$ satisfies $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2)$, and $g(.)$ is a function so that $g'(\theta)$ exists, then*

$$\sqrt{n}[g(\hat{\theta}_n) - g(\theta)] \xrightarrow{d} N(0, \sigma^2(g'(\theta))^2)$$

Therefore, we can let $\theta_n = \hat{\lambda}$, $\theta = \lambda$, $\sigma^2 = \lambda^2/p$ and $g(\lambda) = \log\lambda$ so that $g'(\lambda) = 1/\lambda$. From delta method we yield

$$\sqrt{n}(\log\hat{\lambda} - \log\lambda) \xrightarrow{d} N\left(0, \frac{\lambda^2}{p}\left(\frac{1}{\lambda}\right)^2 = \frac{1}{p}\right) \tag{70}$$

Or, with slight abuse of notation

$$\log\hat{\lambda} \xrightarrow{d} N(\log\lambda, 1/np) \approx N(\log\lambda, 1/R) \tag{71}$$

Therefore, hypothesis testing for $H_0 : \lambda = \lambda_0$ can be based on the Z statistic

$$Z = \sqrt{R}(\log\hat{\lambda} - \log\lambda_0) \tag{72}$$

which should follow a standard normal distribution under H_0. To construct a confidence interval for λ, since we have

$$\mathbb{P}[\log(\hat{\lambda}) - Z_{1-\alpha/2}/\sqrt{R} < \log\lambda < \log(\hat{\lambda}) + Z_{1-\alpha/2}/\sqrt{R}] \approx 1 - \alpha \tag{73}$$

We can construct a confidence interval with asymptotic coverage $(1-\alpha)$ by

$$\left(\hat{\lambda}e^{-Z_{1-\alpha/2}/\sqrt{R}}, \hat{\lambda}e^{Z_{1-\alpha/2}/\sqrt{R}}\right) \tag{74}$$

Let's see how this compares to the results produced by the survival regression command `phreg` provided in package `eha`. We use the Primary Biliary Cholangitis (`pbc`) data built in the `survival`

`package`. The variable `time` is the follow-up time for each subject. The variable `status` takes value 0 if the subject is censored, 1 if the subject underwent liver transplantation, and 2 if the subject died. Suppose we are interested in the composite event of transplantation and death, so that status taking value 1 or 2 are both seen as events of interest. We now try to infer on the incidence of event:

```
library(survival)
library(eha)
vcov.phreg <- function(mod) mod$var
data(pbc)

pbc <- pbc[!is.na(pbc$trt), ]

R <- sum(pbc$status >= 1)
M <- sum(pbc$time)

Z <- qnorm(1-0.05/2)

# Use phreg to perform exponential regression
mod <- phreg(Surv(time, status >= 1) ~ 1,
             dist = "weibull", shape = 1, data = pbc)
mod

## Call:
## phreg(formula = Surv(time, status >= 1) ~ 1, data = pbc, dist = "weibull",
##     shape = 1)
##
## Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
## log(scale)                     8.377               0.083     0.000
##
##  Shape is fixed at  1
##
## Events                    144
## Total time at risk        625985
## Max. log. likelihood      -1350.3

# Estimates for lambda
R/M

## [1] 0.0002300375

exp(-coef(mod))

##   log(scale)
## 0.0002300375

# Confidence interval for lambda
c(R/M * exp(-Z/sqrt(R)), R/M * exp(Z / sqrt(R)))

## [1] 0.0001953733 0.0002708520

exp(rev(-confint(mod)))

## [1] 0.0001953733 0.0002708520
```

We can see that the exponential regression carried out by `phreg` produced the same results as our theoretical derivation.

In the previous demonstration we inferred the incidence rate of an event of interest, i.e. we only had one group in our data. Suppose now we would like to compare the incidence of two subgroups and infer the ratio of their incidence rate, i.e. incidence rate ratio (IRR). We can calculate the number of events and follow-up person-time for the two subgroups separately, say R_1, R_2 and

M_1, M_2. Then we have:

$$\log \hat{\lambda}_1 = \log(R_1/M_1) \xrightarrow{d} N(\log \lambda_1, 1/R_1) \tag{75}$$

$$\log \hat{\lambda}_2 = \log(R_2/M_2) \xrightarrow{d} N(\log \lambda_2, 1/R_2) \tag{76}$$

where λ_1 and λ_2 are the incidence rate for the two subgroups. Since the two subgroups are independent, we have

$$\log \hat{\lambda}_1 - \log \hat{\lambda}_2 \xrightarrow{d} N(\log \lambda_1 - \log \lambda_2, 1/R_1 + 1/R_2) \tag{77}$$

Or, written in the form of IRR:

$$\log(\hat{\lambda}_1/\hat{\lambda}_2) \xrightarrow{d} N(\log(\lambda_1/\lambda_2), 1/R_1 + 1/R_2) \tag{78}$$

Therefore, hypothesis testing for $H_0 : \lambda_1 = \lambda_2$ can be based on the Z statistic

$$Z = \sqrt{R} \log(\hat{\lambda}_1/\hat{\lambda}_2) \tag{79}$$

which follows a standard normal distribution under H_0. To construct a confidence interval for the IRR, since we have

$$\mathbb{P}[\log(\hat{\lambda}_1/\hat{\lambda}_2) - Z_{1-\alpha/2}\sqrt{1/R_1 + 1/R_2} < \log(\lambda_1/\lambda_2) < \log(\hat{\lambda}_1/\hat{\lambda}_2) + Z_{1-\alpha/2}\sqrt{1/R_1 + 1/R_2}] \approx 1 - \alpha \tag{80}$$

We can construct a confidence interval with asymptotic coverage $(1 - \alpha)$ by

$$\left((\hat{\lambda}_1/\hat{\lambda}_2)e^{-Z_{1-\alpha/2}\sqrt{1/R_1+1/R_2}}, (\hat{\lambda}_1/\hat{\lambda}_2)e^{Z_{1-\alpha/2}\sqrt{1/R_1+1/R_2}}\right) \tag{81}$$

Let us demonstrate this with the `pbc` dataset, where `trt` takes value 1 if the subject received D-penicillamine and 2 if the subject received placebo. We would like to estimate and infer the IRR of event between these two subgroups:

```
# Calculate the number of events for each group based on treatment status
R1 <- sum(pbc$status[pbc$trt == 1] >= 1)
M1 <- sum(pbc$time[pbc$trt == 1])
R2 <- sum(pbc$status[pbc$trt == 2] >= 1)
M2 <- sum(pbc$time[pbc$trt == 2])

# Use phreg to perform exponential regression
mod2 <- phreg(Surv(time, status >= 1) ~ (trt == 1),
              dist = "weibull", shape = 1, data = pbc)
mod2

## Call:
## phreg(formula = Surv(time, status >= 1) ~ (trt == 1), data = pbc,
##     dist = "weibull", shape = 1)
##
## Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
## trt == 1
##            FALSE    0.491     0         1           (reference)
##             TRUE    0.509     0.048     1.050     0.167     0.772
##
## log(scale)                    8.402               0.120     0.000
##
##  Shape is fixed at  1
##
## Events                    144
## Total time at risk        625985
## Max. log. likelihood      -1350.3
## LR test statistic         0.08
## Degrees of freedom        1
## Overall p-value           0.771685
```

```r
# Estimates for incidence rate ratio
(R1/M1)/(R2/M2)

## [1] 1.04958

exp(coef(mod2))

## trt == 1TRUE   log(scale)
##      1.04958   4456.76812

# Confidence interval for incidence rate ratio
c((R1/M1)/(R2/M2) * exp(-Z*sqrt(1/R1+1/R2)),
  (R1/M1)/(R2/M2) * exp(Z*sqrt(1/R1+1/R2)))

## [1] 0.7568769 1.4554783

exp(confint(mod2))

##                       2.5 %      97.5 %
## trt == 1TRUE     0.7568769    1.455478
## log(scale)    3520.0385082 5642.774076
```

We can see that we still arrive at results identical to that produced by `phreg`. In addition, the 95% confidence interval of the IRR includes 1, so we *cannot* reject the null hypothesis that the two subgroups have the same IRR.

5 Regression Methods for Parametric Survival Analysis

In addition to modelling the distribution of survival time over the whole population or over a handful of subgroups, often times it is of our interest how covariates are related to the survival, eg. how survival changes as the body mass index (BMI) of a cancer patient increases. There are two commonly used models that allows us to regress survival on covariates, which we elaborate below:

5.1 Proportional hazards model

In proportional hazards regression, it is assumed that a function of covariates X acts *multiplicatively* on the *hazard function*, i.e. most generally

$$h(t|X) = h_0(t)g(X) \tag{82}$$

where $g(X)$ is non-negative since hazard functions are non-negative. The reason why this framework is termed as *proportional* hazards regression is that the hazard function is *proportional to* a certain function of the covariates. Note that under this framework, the hazard ratio between two subjects with covariate status X_1 and X_2 would be

$$\frac{h_1(t)}{h_2(t)} = \frac{h(t|X_1)}{h(t|X_2)} = \frac{h_0(t)g(X_1)}{h_0(t)g(X_2)} = \frac{g(X_1)}{g(X_2)} \tag{83}$$

so that the hazard ratio *does not depend on time* and only depends on the covariates. Based on Equation (82), we have

$$\begin{aligned} S(t|X) = \exp(-H(t|X)) &= \exp\Big(-\int_0^t h(s|X)ds\Big) && (84)\\ &= \exp\Big(-\int_0^t h_0(s)g(X)ds\Big) && (85)\\ &= \exp\Big(-\int_0^t h_0(s)ds \cdot g(X)\Big) && (86)\\ &= \exp(-H_0(t)g(X)) = S_0(t)^{g(X)} && (87) \end{aligned}$$

which can aid the likelihood construction as we have elaborated in the previous section.

More often than not, for simplicity, we would want $g(X)$ be related to $X\beta$, the linear combination of elements in X. Since $g(X)$ must be non-negative, we can let $g(X) = \exp(X\beta)$, which from Equations (82) and (87) we have

$$h(t|X) = h_0(t)\exp(X\beta) \tag{88}$$

$$S(t|X) = S_0(t)^{\exp(X\beta)} \tag{89}$$

Under this circumstance, suppose we can write X as $(X_1, X_2, ..., X_p)^\top$ and β as $(\beta_1, \beta_2, ..., \beta_p)^\top$ (note that we do not include an intercept here, why?), then for two subjects with covariates $(x_1, x_2, ..., x_j + 1, ..., x_p)^\top$ and $(x_1, x_2, ..., x_j, ..., x_p)^\top$, their hazard ratio would be, from Equation (83)

$$\frac{\exp(x_1\beta_1 + x_2\beta_2 + \cdots + (x_j+1)\beta_j + \cdots + x_p\beta_p)}{\exp(x_1\beta_1 + x_2\beta_2 + \cdots + x_j\beta_j + \cdots + x_p\beta_p)} = \frac{\exp((x_j+1)\beta_j)}{\exp(x_j\beta_j)} = \exp(\beta_j) \tag{90}$$

Therefore, $\exp(\beta_j)$ is the hazard ratio when covariates other than x_j are held constant and x_j is incremented by 1, and β_j is *log hazard ratio* under the increment.

Although $h_0(t)$ can be set as any hazard function, it is not guaranteed that under the proportional hazards setting, $h(t|X)$ will be of the same class of $h_0(t)$. That is, some survival time distributions are not closed under the proportional hazards framework. Weibull distribution is one of the distributions that is closed under proportional hazards. To see that, suppose $h_0(t)$ is set as the hazard function for Weibull(p, λ_0), then we have, from Equation (88)

$$h_0(t) = p\lambda_0^p t^{p-1} \tag{91}$$

$$h(t|X) = h_0(t)\exp(X\beta) = p\lambda_0^p t^{p-1}\exp(X\beta) = p(\lambda_0\exp(X\beta/p))^p t^{p-1} \tag{92}$$

Therefore, the survival distribution for a subject of covariate X is Weibull$(p, \lambda_0\exp(X\beta/p))$. Or, equivalently, we are fixing the shape parameter as p and modelling the rate parameter λ as

$$\log\lambda = \log\lambda_0 + X\beta/p \tag{93}$$

However, if $h_0(t)$ is set as the hazard function for LN(μ, σ^2), then we have

$$S_0(t) = 1 - \Phi\Big(\frac{\log t - \mu}{\sigma}\Big) \tag{94}$$

$$S(t|X) = S_0(t)^{\exp(X\beta)} = \Big[1 - \Phi\Big(\frac{\log t - \mu}{\sigma}\Big)\Big]^{\exp(X\beta)} \tag{95}$$

which cannot be reduced back to the structural form for survival functions of log-normal survivals.

We now demonstrate how proportional hazards regression can be carried out in real data using the `pbc` dataset. Suppose in addition to the effect of treatment (`trt`), we would also want to investigate the effect of the following variables:

- `age`: Age in years
- `sex`: Sex, M or F
- `edema`: The subject's edema status, 0 if no edema, 0.5 if untreated or successfully treated, 1 if edema despite diuretics
- `albumin`: Serum albumin level in g/dL
- `ascites`: If the subject had ascites
- `ast`: Serum AST level in U/mL
- `bili`: Serum bilirubin level in mg/dL
- `protime`: Prothrombin time, a test for blood coagulation, in seconds

We can then fit a proportional hazards regression using Weibull distribution as baseline survival distribution. Since the values for `ascites` does not have numerical meanings, we use it as factor. Using the `phreg` routine, we get the following results:

```r
# Fit a proportional hazards regression with Weibull distribution
mod3 <- phreg(Surv(time, status >= 1) ~ (trt == 1) + age + sex + factor(edema) +
                albumin + ascites + ast + bili + protime, dist = "weibull", data = pbc)
mod3
```

```
## Call:
## phreg(formula = Surv(time, status >= 1) ~ (trt == 1) + age +
##     sex + factor(edema) + albumin + ascites + ast + bili + protime,
##     data = pbc, dist = "weibull")
##
## Covariate           W.mean      Coef Exp(Coef)  se(Coef)    Wald p
## trt == 1
##            FALSE    0.491     0         1           (reference)
##             TRUE    0.509     0.125     1.133     0.177     0.482
## age                49.230     0.017     1.018     0.009     0.055
## sex
##                m    0.112     0         1           (reference)
##                f    0.888    -0.507     0.602     0.235     0.031
## factor(edema)
##                0    0.906     0         1           (reference)
##              0.5    0.074     0.153     1.165     0.279     0.584
##                1    0.020     0.985     2.679     0.310     0.001
## albumin            3.622    -0.953     0.385     0.238     0.000
## ascites            0.031     0.438     1.549     0.294     0.136
## ast              115.421     0.004     1.004     0.001     0.007
## bili               2.139     0.096     1.101     0.017     0.000
## protime           10.631     0.240     1.272     0.078     0.002
##
## log(scale)                   8.562               0.859     0.000
## log(shape)                   0.458               0.068     0.000
##
## Events                    144
## Total time at risk        625985
## Max. log. likelihood      -1263.7
## LR test statistic         169.21
## Degrees of freedom        10
## Overall p-value           0
```

```r
# Generate confidence intervals for the exponentiated regression coefficients
exp(confint(mod3))
```

```
##                         2.5 %       97.5 %
## trt == 1TRUE        0.8002563 1.603554e+00
## age                 0.9996441 1.035895e+00
## sexf                0.3798091 9.552078e-01
## factor(edema)0.5    0.6738792 2.014642e+00
## factor(edema)1      1.4603445 4.914099e+00
## albumin             0.2416183 6.147120e-01
## ascites             0.8713826 2.755130e+00
## ast                 1.0010523 1.006719e+00
## bili                1.0643606 1.138746e+00
## protime             1.0916515 1.481312e+00
## log(scale)        970.6573958 2.818368e+04
## log(shape)          1.3832177 1.808502e+00
```

From the results for the regression, we see that a bunch of variables may influence our composite event of interest. For example, for the variable `sex`, we see that after controlling for other covariates, the hazard ratio of female compared to male is `0.602` with 95% confidence interval `(0.380, 0.955)`. This implies taken all other risk factors into consideration, females are *less* risky than males with regard to our event of interest. For the variable `ast`, the estimated hazard ratio is `1.004` with 95% confidence interval `(1.001, 1.007)`, which implies that after controlling for

other covariates, each 1 U/mL increase of AST level in the serum will result in a 0.4% increase of hazards. Since an 1 U/mL increase in AST is not relevant clinically, we may raise the exponentiated coefficient and confidence interval bounds to the 10^{th} power, arriving at a hazard ratio of 1.041 with 95% confidence interval $(1.010, 1.069)$. This implies that each 10 U/mL increase of AST level in the serum with result in a 4.1% increase in hazards, which is clinically more interpretable.

Also, we see that the estimate of `log(shape)`, which is $\log p$ in the parameterization of Weibull distribution, is estimated as `0.458` and said to be significantly away from 0 according to the Wald p-value. Since a $\log p$ of 0 corresponds to the exponential distribution, this implies that simplifying the baseline survival distribution into an exponential distribution would not be a good idea, since its fit would significantly worse than the current model. Alternatively, we can use likelihood ratio test to carry out the comparison between Weibull distribution and exponential distribution by fitting a exponential distribution-based proportional hazards regression, and comparing the log-likelihood between the two models. The degrees of freedom for the test would be 1, since we are only constraining p to be 1 in the exponential distribution-based model.

```
# Fit a proportional hazards model with Exponential distribution
mod3_exp <- phreg(Surv(time, status >= 1) ~ (trt == 1) + age + sex + factor(edema) +
                    albumin + ascites + ast + bili + protime, dist = "weibull",
                  shape = 1, data = pbc)
mod3_exp
```

```
## Call:
## phreg(formula = Surv(time, status >= 1) ~ (trt == 1) + age +
##     sex + factor(edema) + albumin + ascites + ast + bili + protime,
##     data = pbc, dist = "weibull", shape = 1)
##
## Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
## trt == 1
##            FALSE    0.491     0         1           (reference)
##             TRUE    0.509     0.081     1.084     0.174     0.644
## age                49.230     0.013     1.013     0.009     0.159
## sex
##                m    0.112     0         1           (reference)
##                f    0.888    -0.523     0.593     0.233     0.025
## factor(edema)
##                0    0.906     0         1           (reference)
##              0.5    0.074     0.218     1.243     0.273     0.425
##                1    0.020     0.655     1.925     0.312     0.036
## albumin            3.622    -0.736     0.479     0.231     0.001
## ascites            0.031     0.350     1.419     0.296     0.237
## ast              115.421     0.004     1.004     0.001     0.014
## bili               2.139     0.070     1.073     0.017     0.000
## protime           10.631     0.244     1.277     0.076     0.001
##
## log(scale)                   9.440               1.332     0.000
##
##  Shape is fixed at  1
##
## Events                    144
## Total time at risk        625985
## Max. log. likelihood      -1282.1
## LR test statistic         136.43
## Degrees of freedom        10
## Overall p-value           0
```

```
# Do a likelihood ratio test
1-pchisq(2*(logLik(mod3) - logLik(mod3_exp)), 1)
```

```
## 'log Lik.' 1.241104e-09 (df=12)
```

5.2 Accelerated failure time model

Apart from proportional hazards regression, another common approach in parametric survival regression is the **accelerated failure time (AFT) model**. As its name suggests, the AFT model assumes that the covariates serves as speed knobs for time, so that the time for some subjects flows faster than others.

Mathematically, AFT models assume that given a subject with covariate value X (let call them subject X), their time would accelerated by a factor of $g(X)$ (which should of course be positive) compared to a subject with reference covariate value. Now suppose the survival function of the reference subject is $S_0(.)$, then at time t, the survival probability for the subject X is already $S_0(g(X)t)$, since their time flows $g(X)$ times faster. Therefore, we have the following relation for AFT models:

$$S(t|X) = S_0(g(X)t)$$

Based on Section 5.2, we have

$$f(t|X) = -\frac{\partial}{\partial t}S(t|X) = g(X)\Big(-\frac{\partial}{\partial(g(X)t)}S_0(g(X)t)\Big) = g(X)f_0(g(X)t) \tag{96}$$

$$h(t|X) = \frac{f(t|X)}{S(t|X)} = \frac{g(X)f_0(g(X)t)}{S_0(g(X)t)} = g(X)\frac{f_0(g(X)t)}{S_0(g(X)t)} = g(X)h(g(X)t) \tag{97}$$

where $f_0(.)$ and $h_0(.)$ are the probability density function and hazard function for the reference subject. An alternative, commonly seen definition for AFT models is based on the relationship between random variables for survival time. Suppose we denote the survival time for a reference subject is T_0, then we can define the survival time for "subject X" as:

$$\log T = -g(X) + \log T_0 \tag{98}$$

which implies,

$$T = \frac{T_0}{g(X)} \tag{99}$$

This definition is equivalent to our definition based on survival functions, since by definition,

$$S(t|X) = \mathbb{P}(T > t) = \mathbb{P}\Big(\frac{T_0}{g(X)} > t\Big) = \mathbb{P}(T_0 > g(X)t) = S_0(g(X)t) \tag{100}$$

The alternative definition for AFT models in Equation (98) is sometimes enticing, since we can strip the mean out of $\log T_0$ so that $\log T_0 = \mu + \varepsilon$ where the mean of the error ε is zero, and arrive at the following expression

$$\log T = \mu - g(X) + \varepsilon \tag{101}$$

which implies that if we treat $\log T$ as the outcome, we can use linear or non-linear regression to fit the model, with the conditional mean of $\log T$ modelled to $\mu - g(X)$.

Just as in proportional hazards models, often we would want $g(X)$ to be related to $X\beta$, the linear combination of elements in X. Since $g(X)$ must be positive, an intuitive choice is to let $g(X) = \exp(X\beta)$, so from Section 5.2 and eqs. (96) and (97), we have

$$S(t|X) = S_0(\exp(X\beta)t) \tag{102}$$

$$f(t|X) = \exp(X\beta)f_0(\exp(X\beta)t) \tag{103}$$

$$h(t|X) = \exp(X\beta)h_0(\exp(X\beta)t) \tag{104}$$

which can be plugged into the likelihood functions in parametric inference. Also, using the form of $g(X)$, we can express the mean survival time (μ_T) and median survival time (m_T) as

$$\mu_T = \mathbb{E}[T] = \mathbb{E}\Big[\frac{T_0}{\exp(X\beta)}\Big] = \mathbb{E}[T_0]\exp(-X\beta) \tag{105}$$

$$S(m_T) = S_0(\exp(X\beta)m_T) = 1/2 \quad \Rightarrow \quad m_T = S_0^{-1}(1/2)\exp(-X\beta) \tag{106}$$

Under this circumstance, suppose we can write X as $(X_1, X_2, ..., X_p)^\top$ and β as $(\beta_1, \beta_2, ..., \beta_p)^\top$, then for two subjects with covariates $(x_1, x_2, ..., x_j+1, ..., x_p)^\top$ and $(x_1, x_2, ..., x_j, ..., x_p)^\top$, the ratio of their mean survival time would be, from Equation (105)

$$\frac{\mathbb{E}[T_0]\exp(-x_1\beta_1 - x_2\beta_2 \cdots - (x_j+1)\beta_j \cdots - x_p\beta_p)}{\mathbb{E}[T_0]\exp(-x_1\beta_1 - x_2\beta_2 \cdots - x_j\beta_j \cdots - x_p\beta_p)} = \frac{-\exp((x_j+1)\beta_j)}{-\exp(x_j\beta_j)} = \exp(-\beta_j) \tag{107}$$

Similarly, the ratio of their medial survival time would also be, from Equation (106)

$$\frac{S_0^{-1}(1/2)\exp(-x_1\beta_1 - x_2\beta_2 \cdots - (x_j+1)\beta_j \cdots - x_p\beta_p)}{S_0^{-1}(1/2)\exp(-x_1\beta_1 - x_2\beta_2 \cdots - x_j\beta_j \cdots - x_p\beta_p)} = \frac{-\exp((x_j+1)\beta_j)}{-\exp(x_j\beta_j)} = \exp(-\beta_j) \quad (108)$$

Therefore, when covariates other than x_j is held constant and x_j is incremented by 1, then both the mean and median survival time is *shortened* by a ratio of $\exp(-\beta_j)$, which is considered easier to interpret than hazard ratios by some.

In contrast to proportional hazards models, most of the common parametric survival distributions are closed under AFT models. That is, given that $S_0(t)$ belongs to some family of distributions, then most of the time $S(t|X)$ also belongs to the same family of distributions. For example, suppose $S_0(t)$ is set as the survival function for $\text{LN}(\mu, \sigma^2)$, then we have, from Section 5.2

$$S(t|X) = S_0(\exp(X\beta)t) = 1 - \Phi\Big(\frac{\log(\exp(X\beta)t) - \mu}{\sigma}\Big) = 1 - \Phi\Big(\frac{\log t - (\mu - X\beta)}{\sigma}\Big) \quad (109)$$

which is the survival function for $\text{LN}(\mu - X\beta, \sigma^2)$. However, note that even if a family of distribution can be modelled with both the proportional hazards model and the AFT model, the resulting models may not be structurally equivalent. In fact, it can be show that the Weibull family is *the only* family where a proportional hazards model is also an AFT model. In particular, previouly we have shown that hazard function for the proportional hazards model with $\text{Weibull}(p, \lambda_0)$ as baseline survival distribution is

$$h_{\text{PH}}(t|X) = p\lambda_0^p t^{p-1}\exp(X\beta) \quad (110)$$

And from Equation (97), the hazard function for the AFT model with $\text{Weibull}(p, \lambda_0)$ as baseline survival distribution is

$$\begin{aligned} h_{\text{AFT}}(t|X) &= p\lambda_0^p(\exp(X\gamma)t)^{p-1}\exp(X\gamma) && (111)\\ &= p\lambda_0^p t^{p-1}[\exp(X\gamma)]^p && (112)\\ &= p\lambda_0^p t^{p-1}\exp(X(p\gamma)) && (113) \end{aligned}$$

Comparing Equations (110) and (113), we yield $\beta = p\gamma$. That is, if both the proportional hazards model and the AFT model are fit with Weibull distribution of shape parameter p, then the regression coefficients of the proportional hazards model will be exactly p times the coefficients for the AFT model. We can verify this with the `pbc` dataset:

```
vcov.aftreg <- function(mod) mod$var
# Fit a Weibull PH model and an AFT model with the same set of covariates
mod_ph <- phreg(Surv(time, status >= 1) ~ (trt == 1) + age + sex,
                dist = "weibull", data = pbc)
mod_aft <- aftreg(Surv(time, status >= 1) ~ (trt == 1) + age + sex,
                  dist = "weibull", data = pbc)

# List the estimated regression coefficients
coef(mod_ph)

## trt == 1TRUE          age         sexf   log(scale)   log(shape)
##  -0.01719679   0.02247571  -0.42743581   8.93451222   0.16112738

coef(mod_aft)

## trt == 1TRUE          age         sexf   log(scale)   log(shape)
##  -0.01463537   0.01913018  -0.36383970   8.93446040   0.16112780

# List the ratios of coefficients between the PH model and the AFT model
round(coef(mod_ph) / coef(mod_aft), 3)

## trt == 1TRUE          age         sexf   log(scale)   log(shape)
##        1.175        1.175        1.175        1.000        1.000

# The ratios should be exactly the shape parameter, p
round(exp(coef(mod_ph)["log(shape)"]), 3)
```

```
## log(shape)
##      1.175

# List the exponentiated negative regression coefficients for the AFT model
exp(-coef(mod_aft))

## trt == 1TRUE          age         sexf   log(scale)   log(shape)
##  1.014742988  0.981051637  1.438843552  0.000131769  0.851183280

# List the confidence intervals
exp(-confint(mod_aft))[, c(2,1)]

##                     97.5 %        2.5 %
## trt == 1TRUE 7.656988e-01 1.3447890457
## age          9.674268e-01 0.9948683128
## sexf         9.896715e-01 2.0918766971
## log(scale)   5.753935e-05 0.0003017598
## log(shape)   7.385023e-01 0.9810571280
```

The ratios of regression coefficients between the proportional hazards model and the AFT model are all `1.175`, which is exactly the estimated shape parameter p and confirms our theoretical derivation. The interpretation for the exponentiated negative coefficients for the AFT model, as we have elaborated earlier, is the ratio of survival time shortening for 1 unit increment of the corresponding covariate. For example, the exponentiated negative coefficient for `age` is `0.981`, meaning that as the age increase by 1 year, the patient's expected survival time would be shortened by a factor of `0.981`. From its 95% confidence interval `(0.967, 0.995)`, this factor is significantly away from 1, meaning that age has a significant influence on event occurence after controlling for treatment and sex.

6 Nonparametric Inference of Right-Censored Survival Data

6.1 Kaplan-Meier estimator and Greenwood's formula

Previously we have shown that as long as the censoring is non-informative, we can write the likelihood of the data $\{(y_i, \delta_i)\}_{i=1}^n$ as:

$$L(\theta_T) = \prod_{i=1}^{n} f_T(y_i;\theta_T)^{\delta_i} S_T(y_i;\theta_T)^{1-\delta_i} \tag{114}$$

where θ_T is the vector of parameters for the survival distribution we assumed for the event. Now suppose instead we do not want to make parametric assumptions, so that now we have the likelihood

$$L(S_T) = \prod_{i=1}^{n} f_T(y_i)^{\delta_i} S_T(y_i)^{1-\delta_i} \tag{115}$$

and our goal is to find a survival distribution function S_T among *all possible survival functions* that maximize the likelihood. Since we are obtaining our estimator non-parametrically, the MLE for this problem, $\hat{S}_T$, is called the *non-parametric maximum likelihood estimator (NPMLE)*.

Intuitively, to maximize the likelihood, $\hat{S}_T$ would be a step function only dropping at timepoints where events occur. To see this, first we define some notations: let $v_1, v_2, ..., v_m$ be distinct ordered timepoints from the set of observations $y_1, y_2, ..., y_n$, and suppose at time v_j, d_j subjects experienced events and c_j subjects were censored. Also by convention we let $v_0 = 0, d_0 = 0, c_0 = 0$ and $v_{m+1} = \infty, d_{m+1} = 0, c_{m+1} = 0$. With the new notations, the likeihood $L(S_T)$ can be rewritten as

$$L(S_T) = \prod_{j=0}^{m+1} f_T(v_j)^{d_j} S_T(v_j)^{c_j} \tag{116}$$

Now suppose S_T is *any* viable survival function. Given $k, l \in \{0, 1, 2, ..., m\}$ where $k < l$, we now try to improve S_T to $\tilde{S}_T$ within $t \in [v_k, v_l)$ by maximizing $L(\tilde{S}_T)$. Since $\tilde{S}_T$ is held equal to S_T

outside of $[v_k, v_l)$, we only have to maximize the part of the likelihood related to the value of $\tilde{S}_T$ within $[v_k, v_l)$, which is

$$L^{kl}(\tilde{S}_T) = \Big[\prod_{j=k}^{l-1} \tilde{f}_T(v_j)^{d_j} \tilde{S}_T(v_j)^{c_j}\Big] \tilde{f}_T(v_l)^{d_l} \tag{117}$$

First let us set k, l so that there are events at both ends of $[v_k, v_l)$ and no events within $[v_k, v_l)$, i.e. $(d_k \neq 0, d_l \neq 0) \wedge (d_j = 0, \forall k < j < l)$. Under this setting, we have the partial likelihood

$$L^{kl}(\tilde{S}_T) = \tilde{f}_T(v_k)^{d_k} \Big[\prod_{j=k}^{l-1} \tilde{S}_T(v_j)^{c_j}\Big] \tilde{f}_T(v_l)^{d_l} \tag{118}$$

This partial likelihood implies that $\tilde{S}_T$ cannot drop down within the eventless interval (v_k, v_l), i.e. $\tilde{S}_T(t) = \tilde{S}_T(v_k),\ \forall t \in (v_k, v_l)$. To see this, suppose the condition is violated so that $\tilde{S}_T(v_l^-) < \tilde{S}_T(v_k)$. We can thus define another survival function $\tilde{S}_T$ by

$$\tilde{S}^*_T(t) = \tilde{S}_T(v_k), \quad t \in (v_k, v_l) \tag{119}$$
$$\tilde{S}^*_T(t) = \tilde{S}_T(t), \quad \text{otherwise} \tag{120}$$

Based on the definition, we have the following equalities and inequalities, where $k < j < l$:

$$\tilde{S}^*_T(v_k) = \tilde{S}_T(v_k) \tag{121}$$
$$\tilde{f}^*_T(v_k) = \tilde{S}^*_T(v_k^-) - \tilde{S}^*_T(v_k) = \tilde{S}_T(v_k^-) - \tilde{S}_T(v_k) = \tilde{f}_T(v_k) \tag{122}$$
$$\tilde{S}^*_T(v_j) = \tilde{S}_T(v_k) \geq \tilde{S}_T(v_j) \tag{123}$$
$$\tilde{f}^*_T(v_l) = \tilde{S}^*_T(v_l^-) - \tilde{S}^*_T(v_l) = \tilde{S}_T(v_k) - \tilde{S}_T(v_l) > \tilde{S}_T(v_l^-) - \tilde{S}_T(v_l) = \tilde{f}_T(v_l) \tag{124}$$

Therefore, we have

$$L^{kl}(\tilde{S}^*_T) = \tilde{f}^*_T(v_k)^{d_k} \Big[\prod_{j=k}^{l-1} \tilde{S}^*_T(v_j)^{c_j}\Big] \tilde{f}^*_T(v_l)^{d_l} > \tilde{f}_T(v_k)^{d_k} \Big[\prod_{j=k}^{l-1} \tilde{S}_T(v_j)^{c_j}\Big] \tilde{f}_T(v_l)^{d_l} = L^{kl}(\tilde{S}_T) \tag{125}$$

which contradicts with the fact that $\tilde{S}_T$ maximizes the partial likelihood L^{kl}.

Second, let us set $k = 0$ and let l be the first timepoint with event occurrence, i.e. $(d_l \neq 0) \wedge (d_j = 0,\ \forall j < l)$, then we have the partial likelihood

$$L^{0l}(\tilde{S}_T) = \Big[\prod_{j=1}^{l-1} \tilde{S}_T(v_j)^{c_j}\Big] \tilde{f}_T(v_l)^{d_l} \tag{126}$$

This partial likelihood implies that $\tilde{S}_T$ should remain at 1 until the first event occurs, i.e. $\tilde{S}_T(t) = 1, \forall t < v_l$. To see this, suppose the condition is violated so that $\tilde{S}_T(v_l^-) < 1$. We can thus define another survival function $\tilde{S}^*_T$ by

$$\tilde{S}^*_T(t) = 1, \quad t < v_l \tag{127}$$
$$\tilde{S}^*_T(t) = \tilde{S}_T(t), \quad \text{otherwise} \tag{128}$$

Based on the definition, we have the following equalities and inequalities, where $0 < j < l$:

$$\tilde{S}^*_T(v_j) = 1 \geq \tilde{S}_T(v_j) \tag{129}$$
$$\tilde{f}^*_T(v_l) = \tilde{S}^*_T(v_l^-) - \tilde{S}^*_T(v_l) = 1 - \tilde{S}_T(v_l) > \tilde{S}_T(v_l^-) - \tilde{S}_T(v_l) = \tilde{f}_T(v_l) \tag{130}$$

Therefore, we have

$$L^{0l}(\tilde{S}^*_T) = \Big[\prod_{j=1}^{l-1} \tilde{S}^*_T(v_j)^{c_j}\Big] \tilde{f}^*_T(v_l)^{d_l} > \Big[\prod_{j=1}^{l-1} \tilde{S}_T(v_j)^{c_j}\Big] \tilde{f}_T(v_l)^{d_l} = L^{01}(\tilde{S}_T) \tag{131}$$

which contradicts with the fact that $\tilde{S}_T$ maximizes the partial likelihood L^{0l}.

At last, we set $l = m + 1$ and let k be the last timepoint with event occurrene, i.e. $(d_k \neq 0) \wedge (d_j = 0,\ \forall j > k)$, then we have the partial likelihood:

$$L^{k(m+1)}(\tilde{S}_T) = \tilde{f}_T(v_k)^{d_k}\Big[\prod_{j=k}^{m} \tilde{S}_T(v_j)^{c_j}\Big] \tag{132}$$

This partial likelihood implies that $\tilde{S}_T$ should not drop after the last event until the end of follow-up, i.e. $\tilde{S}_T(t) = \tilde{S}_T(v_k),\ \forall t, v_k < t \leq v_m$. To see this, suppose the condition is violated so that $\tilde{S}_T(v_m) < \tilde{S}_T(v_k)$ where $v_k < v_m$. We thus may define another survival function $\tilde{S}^*_T$ by

$$\tilde{S}^*_T(t) = \tilde{S}_T(v_k), \quad v_k < t \leq v_m \tag{133}$$

$$\tilde{S}^*_T(t) = \tilde{S}_T(t), \quad \text{otherwise} \tag{134}$$

Based on the definition, we have the following equalities and inequalities, where $k < j < m$:

$$\tilde{S}^*_T(v_k) = \tilde{S}_T(v_k) \tag{135}$$

$$\tilde{f}^*_T(v_k) = \tilde{S}^*_T(v_k^-) - \tilde{S}^*_T(v_k) = \tilde{S}_T(v_k^-) - \tilde{S}_T(v_k) = \tilde{f}_T(v_k) \tag{136}$$

$$\tilde{S}^*_T(v_j) = \tilde{S}_T(v_k) \geq \tilde{S}_T(v_j) \tag{137}$$

$$\tilde{S}^*_T(v_m) = \tilde{S}_T(v_k) > \tilde{S}_T(v_m) \tag{138}$$

$$\tag{139}$$

Therefore, we have

$$L^{k(m+1)}(\tilde{S}^*_T) = \tilde{f}^*_T(v_k)^{d_k}\Big[\prod_{j=k}^{m} \tilde{S}^*_T(v_j)^{c_j}\Big] > \tilde{f}_T(v_k)^{d_k}\Big[\prod_{j=k}^{m} \tilde{S}_T(v_j)^{c_j}\Big] = L^{k(m+1)}(\tilde{S}_T) \tag{140}$$

which contradicts with the fact that $\tilde{S}_T$ maximizes the partial likelihood $L^{k(m+1)}$.

From the three observations above, we see the NPMLE for S_T is a step function that can only drop at timepoints with event occurence (A technicality here is that when there are subjects censored at v_m, i.e. at the longest time of follow-up, the survival function *can* drop at $t > v_m$ without affecting the likelihood. In fact, in this case the NPMLE for S_T is only identified up until v_m, and any valid survival function after v_m does not influence the likelihood). Therefore, we can reduce our NPMLE problem to a MLE problem with finite number of parameters. Let $t_1, t_2 ..., t_q$ be the ordered timepoints where events occured, then we can reparameterize S_T as

$$S_T(t) = \prod_{j,\ t_j \leq t} (1 - h_j) \tag{141}$$

where $0 \leq h_j \leq 1$, so that the survival function only drops at timepoints where events occured.

To find the MLE for the parameterization in Equation (141) heuristically, note that the survival function can also be decomposed as follows, denoting r as the largest integer satisfying $t_r \leq t$:

$$S_T(t) = \mathbb{P}(T > t) \tag{142}$$

$$= \mathbb{P}(T > t | T > t_r)\mathbb{P}(T > t_r) \tag{143}$$

$$= \mathbb{P}(T > t | T > t_r)\mathbb{P}(T > t_r | T > t_{r-1})\mathbb{P}(T > t_{r-1}) \tag{144}$$

$$\vdots \tag{145}$$

$$= \mathbb{P}(T > t | T > t_r) \prod_{j=1}^{r} \mathbb{P}(T > t_j | T > t_{j-1}) \tag{146}$$

The estimated $\mathbb{P}(T > t | T > t_r)$ would be 1 since empirically there are no events within $(t_r, t]$. Comparing Equations (141) and (146), we have the following relation

$$h_j = 1 - \mathbb{P}(T > t_j | T > t_{j-1}) \tag{147}$$

$$= \mathbb{P}(T \leq t_j | T > t_{j-1}) \tag{148}$$

$$= \mathbb{P}(T \leq t_j | T \geq t_j, T > t_{j-1})\mathbb{P}(T \geq t_j, t_{j-1}) \tag{149}$$

$$= \mathbb{P}(T \leq t_j | T \geq t_j) \tag{150}$$

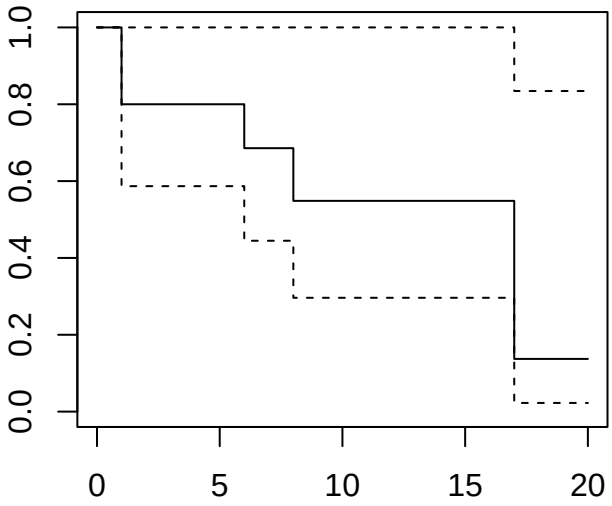
$$= \mathbb{P}(T = t_j | T \geq t_j) \tag{151}$$

where in Equation (150), since empirically there are no events within $(t_{j-1}, t_j]$, $T > t_{j-1}$ is equivalent to $T \geq t_j$. In turn, $h_j := \mathbb{P}(T = t_j | T \geq t_j)$ can be estimated empirically by $\frac{d_j}{R_j}$, where R_j is the number of subjects at risk (still in the study) right before t_j, and d_j is the number of subjects experiencing event at time t_j. This is called the **Kaplan-Meier estimator**. To see how Kaplan-Meier estimators are calculated, we suppose that ten patients are followed up for events, and their observed event or censor times are (in month) $1, 1, 4^+, 6, 6^+, 8, 17, 17, 17, 20^+$ (t^+ implies censored at time t). We would then have the following life table.

j	t_j	R_j	d_j	$\hat{h}_j$	$\hat{S}_j$
1	1	10	2	2/10	$1 \times (1 - 2/10) = 0.8$
2	6	7	1	1/7	$0.8 \times (1 - 1/7) = 0.69$
3	8	5	1	1/5	$0.69 \times (1 - 1/5) = 0.55$
4	17	4	3	3/4	$0.55 \times (1 - 3/4) = 0.14$

This estimated curve can be plotted using R as follows:

```
survtime <- c(1,1,4,6,6,8,17,17,17,20)
event <- c(1,1,0,1,0,1,1,1,1,0)
km <- survfit(Surv(survtime, event) ~ 1)
plot(km)
```



The dotted line in the figure is the pointwise 95% confidence interval for the curve. To derive the formulae for the confidence intervals, first note the Kaplan-Meier estimator is

$$\hat{S}_T(t) = \prod_{j,\, t_j \leq t} (1 - \hat{h}_j) \tag{152}$$

Working with products is more difficult than working with sums, so we take the natural logarithm on both sides and yield

$$\log(\hat{S}_T(t)) = \sum_{j,\, t_j \leq t} \log(1 - \hat{h}_j) \tag{153}$$

Now assuming $\hat{h}_j$ are approximately mutually independent and using the delta method, we have:

$$var[\log(\hat{S}_T(t))] = var[\sum_{j,\ t_j \le t} \log(1-\hat{h}_j)] \tag{154}$$

$$\approx \sum_{j,\ t_j \le t} var[\log(1-\hat{h}_j)] \tag{155}$$

$$\approx \sum_{j,\ t_j \le t} var[\hat{h}_j] \frac{1}{(1-\hat{h}_j)^2} \tag{156}$$

Now since $\hat{h}_j$ is an empirical proportion with R_j participants, we can approximate its variance by $\frac{h_j(1-h_j)}{R_j} \approx \frac{\hat{h}_j(1-\hat{h}_j)}{R_j}$. So we have

$$var[\log(\hat{S}_T(t))] \approx \sum_{j,\ t_j < t} \frac{\hat{h}_j(1-\hat{h}_j)}{R_j} \frac{1}{(1-\hat{h}_j)^2} = \sum_{j,\ t_j < t} \frac{d_j}{R_j(R_j-d_j)} \tag{157}$$

Therefore, using the delta method again, we have the **Greenwood's formula**

$$var[\hat{S}_T(t)] \approx var[\log \hat{S}_T(t)](\hat{S}_T(t))^2 \approx (\hat{S}_T(t))^2 \sum_{j,\ t_j < t} \frac{d_j}{R_j(R_j-d_j)} \tag{158}$$

and the (pointwise) asymptotmic $(1-\alpha)$ confidence interval for $S_T(t)$ can be constructed as

$$\left(\hat{S}_T(t) \pm Z_{1-\alpha/2} \hat{S}_T(t) \sqrt{\sum_{j,\ t_j < t} \frac{d_j}{R_j(R_j-d_j)}} \right) \tag{159}$$

Eminent students may notice that Equation (159) can produce a confidence interval that is outside of $[0,1]$ (especially when the number of events is small), which is unreasonable since survival functions should be bound between 0 and 1. A commonly used strategy is to construct the confidence interval for $\log(-\log(S_T(t)))$ first since the function $\log(-\log(.))$ is bijective and maps $[0,1]$ to the real number line (and $\pm\infty$). From Equation (157) using delta method again, we have

$$var[\log(-\log(\hat{S}_T(t)))] \approx var[\log(\hat{S}_T(t))] \Big[\frac{-1}{\log(\hat{S}_T(t))}\Big]^2 \approx \frac{1}{[\log(\hat{S}_T(t))]^2} \sum_{j,\ t_j < t} \frac{d_j}{R_j(R_j-d_j)} \tag{160}$$

So the (pointwise) asymptotic $(1-\alpha)$ confidence interval for $\log(-\log(S_T(t)))$ is

$$\left(\log(-\log(\hat{S}_T(t))) \mp Z_{1-\alpha/2} \frac{1}{\log(\hat{S}_T(t))} \sqrt{\sum_{j,\ t_j < t} \frac{d_j}{R_j(R_j-d_j)}} \right) \tag{161}$$

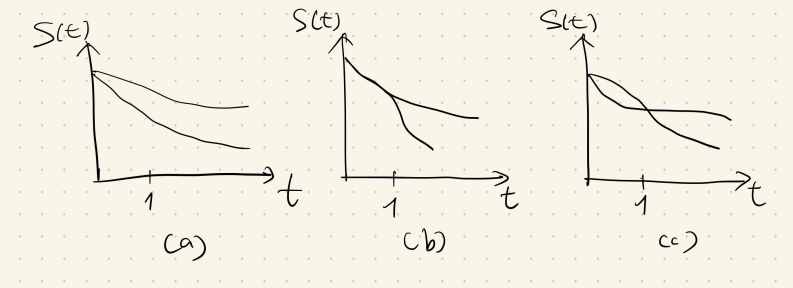
which, after tranformation back to $S_T(t)$, has the following confidence interval

$$\left([\hat{S}_T(t)]^{\exp\left(\pm Z_{1-\alpha/2} \frac{1}{\log(\hat{S}_T(t))} \sqrt{\sum_{j,\ t_j < t} \frac{d_j}{R_j(R_j-d_j)}} \right)} \right) \tag{162}$$

6.2 Log-rank test and its derivatives

Now we have shown how to estimate survival curves non-parametrically, the next question would be how to compare the survival curves between two (or more) groups. That is, we would like to test for the null hypothesis $H_0 : S_0(t) = S_1(t)$, where $S_0(t)$ and $S_1(t)$ are the survival functions for group 0 and group 1. In parametric survival analyses, if the two survival curves belong to the same parametric family, then testing for equivalence of survival curves equates to setting their parameters to be identical, which can be tested using likelihood ratio tests. However, in non-parameteric survival analyses, although we've shown that the Kaplan-Meier estimator is a non-parametric MLE, since the number of parameters for the whole survival function is in principle infinite, regular theories for tests against MLE may not work.

One simple alternative is to test for equality of survival function at one timepoint or a finite set of timepoints. For example, we may opt to test if $S_0(1) = S_1(1)$, which may be done by



comparing $\hat{S}_0(1)$ and $\hat{S}_1(1)$, using Greenwood's formula to quantify their uncertainties. However, this approach discards information from other timepoints, and can be of low power if S_0 and S_1 are similar near $t = 1$ but vastly different at other timepoints (e.g. scenarios (b) and (c) in the following figure). In the following, we will introduce a test that truly tests the equivalence of the whole survival curve, which is the **log-rank test**.

We first define some additional notations. Let $t_1, t_2, ..., t_q$ still be the ordered timepoints with event occurrence. Denote R_{0j} and R_{1j} as the number of subjects at risk for group 0 and group 1 right before t_j, and write $R_j = R_{0j} + R_{1j}$ as the total number of at-risk subjects. Denote d_{0j} and d_{1j} as the number of events occurring at t_j in group 0 and group 1, and write $d_j = d_{0j} + d_{1j}$ as the total number of events. Then for each timepoint t_j, we may construct the following table:

$t = t_j$	Event	Non-event	At risk
Group 0	d_{0j}	$R_{0j} - d_{0j}$	R_{0j}
Group 1	d_{1j}	$R_{1j} - d_{1j}$	R_{1j}
Total	d_j	$R_j - d_j$	R_j

When the marginals $R_{0j}, R_{1j}, d_j, (R_j - d_j)$ treated as fixed, the cells in the table only have one degree of freedom, and let's set our cell of interest to be d_{1j}. Under the null hypothesis of $H_0 : S_0(t) = S_1(t)$, which is equivalent to $h_0(t) = h_1(t)$ and roughly means that "the subjects in both groups have the same probability of experiencing events", we can see this table as drawing d_j balls from a pool of R_{0j} "Group 0 balls" and R_{1j} "Group 1 balls". d_{1j} is then the number of drawn "Group 1 balls". Therefore, the conditional distribution for d_{1j} is a *hypergeometric distribution* with parameters $N = R_j, K = R_{1j}$ and $n = d_j$. Consequently, we can define

$$O_j := d_{1j} \tag{163}$$

$$E_j := \mathbb{E}[O_j | R_j, R_{1j}, d_j] \tag{164}$$

$$= d_j \frac{R_{1j}}{R_j} \tag{165}$$

$$V_j := var[O_j | R_j, R_{1j}, d_j] \tag{166}$$

$$= \frac{R_{1j}(R_j - R_{1j}) d_j (R_j - d_j)}{R_j^2 (R_j - 1)} \tag{167}$$

Now we can define the log-rank statistic by

$$O := \sum_j O_j, \quad E := \sum_j E_j, \quad V := \sum_j V_j \tag{168}$$

$$Z := \frac{O - E}{\sqrt{V}} \tag{169}$$

which is approximately normally distributed under the null hypothesis.

An immediate observation for Z is that when the hazards of group 1 is always greater than group 0, then O, which is the total number of events that occurred in group 1, would be greater than expected, leading to a large positive Z. In contrasts, when the hazards of group 1 is always lesser than group 0, we would get a negative Z with large magnitude. However, in the case where the hazards of the two groups are intertwined like scenario (c) in the figure above, O may not be too different than E since O_j is sometimes larger than and sometime smaller than E_j. In fact, the log-rank test can be reformulated as testing if $\theta = 1$ in the hypothesis $h_1(t) = \theta h_0(t)$. Therefore, it has the best power when the hazards are truly proportional like scenario (a), sub-optimal power when the hazards of one group is still greater than the other across time but not proportional

like scenario (b), and has **low power if the hazard functions of the two groups cross** like scenario (c).

From the discussion above, we see that the vanilla log-rank test statistic treats each contingency table as equally important, so it may have sub-optimal power when then difference in hazards is greater in some time intervals and lesser in other. In light of this, based on prior knowledge for when the hazards difference would be greater, we can construct **weighted log-rank test** statistics as follows

$$Z(\boldsymbol{w}) = \frac{\sum_j w_j(O_j - E_j)}{\sqrt{\sum_j w_j^2 V_j}} \tag{170}$$

where the weight w_j is set to be larger at timepoints that is expected have larger hazards difference. The null distribution of this statistic is still approximately normal. Some common settings for the weights are as follows:

Name	w_j
Log-rank	1
Wilcoxon	R_j
Tarone-Ware	$\sqrt{R_j}$
Peto-Prentice	$\hat{S}(t_j)$
Fleming-Harrington	$[\hat{S}(t_{j-1})]^p[1-\hat{S}(t_{j-1})]^q$

where $\hat{S}(t)$ is the Kaplan-Meier estimator for both groups combined. Notice taht Wilcoxon, Taron-Ware and Peto-Prentice all place larger weights at earlier timepoints, while for Fleming-Harrington we may place larger weights at later timepoints by setting a larger q.

Another extension to the log-rank test is to account for between-group difference in characteristics. For example, suppose we want to know if the treated group has better survival than the untreated group, but we observe that the treated group has younger patients, who naturally have better survival even without the treatment. In this case, we may split the data into the $L = 2$ strata consisting of young patients ($l = 0$) and old patients ($l = 1$). For each stratum $l \in \{0, 1\}$ we yield $O^{(l)}, E^{(l)}$ as the observed and expected number of events, and $V^{(l)}$ as the variance for $O^{(l)}$. We can then construct the **stratified log-rank test** statistic as

$$Z_{\text{strat}} = \frac{\sum_l (O^{(l)} - E^{(l)})}{\sqrt{\sum_l V^{(l)}}} \tag{171}$$

Note that now with each strata, all patients are of the same age group, so the difference between $O^{(l)}$ and $E^{(l)}$ cannot stem from difference in age distribution between groups, and can thus be better ascribed to the effect of the treatment.

The last extension to log-rank test we'll talk about is the case where there are more than two groups, say, $G + 1$ groups labeled $\{0, 1, 2, ..., G\}$. In this case, the null hypothesis we are testing would be $H_0 : S_0(t) = S_1(t) = S_2(t) = ... = S_G(t)$, and the contingency table for timepoint t_j would shape as follows

$t = t_j$	Event	Non-event	At risk
Group 0	d_{0j}	$R_{0j} - d_{0j}$	R_{0j}
Group 1	d_{1j}	$R_{1j} - d_{1j}$	R_{1j}
Group 2	d_{2j}	$R_{2j} - d_{2j}$	R_{2j}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Group G	d_{Gj}	$R_{Gj} - d_{Gj}$	R_{Gj}
Total	d_j	$R_j - d_j$	R_j

Now the degrees of freedom for the cells after fixing the marginals would be G, and we may set $d_{1j}, d_{2j}, ..., d_{Gj}$ as the cells of interest, whose condition joint distribution would be a *multivariate hypergeometric distribution* under the null hypothesis. In light of this, we can define the follows

$$O_j := (d_{1j}, d_{2j}, ..., d_{Gj})^\top \tag{172}$$

$$E_j := \mathbb{E}[O_j | R_j, d_j, R_{1j}, R_{2j}, ..., R_{Gj}] \tag{173}$$

$$= \frac{d_j}{R_j}(R_{1j}, R_{2j}, ..., R_{Gj})^\top \tag{174}$$

$$V_j := var[O_j | R_j, d_j, R_{1j}, R_{2j}, ..., R_{Gj}] \tag{175}$$

where for the detail of the matrix V_j:

$$V_{j(kk)} = \frac{R_{kj}(R_j - R_{kj})d_j(R_j - d_j)}{R_j^2(R_j - 1)} \tag{176}$$

$$V_{j(kl)} = -\frac{R_{kj}R_{lj}d_j(R_j - d_j)}{R_j^2(R_j - 1)} \tag{177}$$

Now we can define the log-rank statistic by

$$O := \sum_j O_j, \quad E := \sum_j E_j, \quad V := \sum_j V_j \tag{178}$$

$$Z := (O - E)^\top V^{-1}(O - E) \tag{179}$$

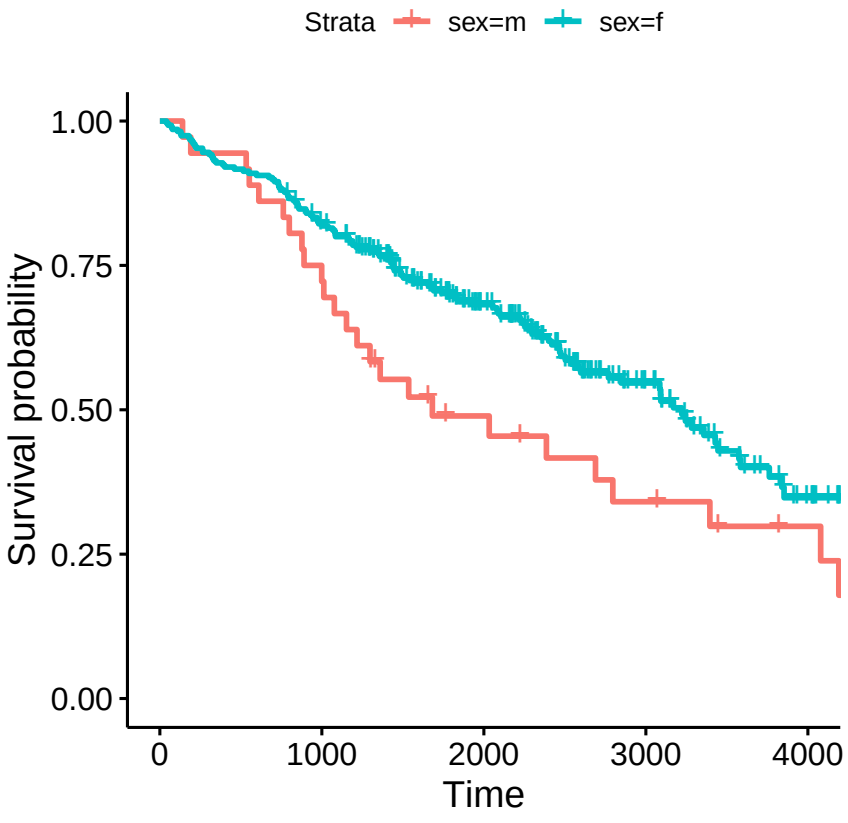
which is approximately distributed as χ^2_G under the null hypothesis. Note that this test, like ANOVA, tests if the survival function of *any* group deviates from others. To clarify which group is different from which, we will still need post-hoc tests.

We will show how log-rank tests work using the `pbc` dataset again. Suppose we are to test if the survival of male patients (`sex = 'm'`) are different from female patients (`sex = 'f'`). We can first plot the Kaplan-Meier curve of the two groups using the package `survminer` (Note that we have called the `pbc` dataset and removed those with `trt == NA` before):

```
library(survminer)

## Loading required package:  ggplot2
## Loading required package:  ggpubr
##
## Attaching package:  'survminer'
## The following object is masked from 'package:survival':
##
##     myeloma

pbc_km <- survfit(Surv(time, status >= 1) ~ sex, data = pbc)
ggsurvplot(pbc_km)
```



From the plot, we can see that the survival of male and female patients seems to be different, with females less likely to experience event than males. To formally show this, we can use the log-rank test:

```
survdiff(Surv(time, status >= 1) ~ sex, data = pbc)

## Call:
## survdiff(formula = Surv(time, status >= 1) ~ sex, data = pbc)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=m  36       25     16.6     4.248      4.85
## sex=f 276      119    127.4     0.554      4.85
##
##  Chisq= 4.9  on 1 degrees of freedom, p= 0.03
```

which shows that the survival of male patients is indeed significantly worse than female patients. However, it still may be the case that in this cohort, male patients are older and thus lead to its worse survival, while gender does not actually affect survival directly. To address this concern, we can stratify the population by age, using the median age as cut-off point:

```
median(pbc$age)

## [1] 49.79466

survdiff(Surv(time, status >= 1) ~ sex + strata(age >= 50), data = pbc)

## Call:
## survdiff(formula = Surv(time, status >= 1) ~ sex + strata(age >=
##     50), data = pbc)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=m  36       25     16.8     3.993      4.66
## sex=f 276      119    127.2     0.528      4.66
##
##  Chisq= 4.7  on 1 degrees of freedom, p= 0.03
```

From the stratified log-rank test, after accounting for the influence of age, the survival of male patients is still significantly worse than female patients (which can be inferred by noticing the observed number of events for male patients is more than expected).

For categorical variables with more than two levels, we can use the same syntax to carry out the log-rank test, though behind the scene the test statistic is now constructed using quadratic forms and compared against chi-square distributions. For example, if we are interested in the survival difference between different `edema` groups:

```
survdiff(Surv(time, status >= 1) ~ edema, data = pbc)

## Call:
## survdiff(formula = Surv(time, status >= 1) ~ edema, data = pbc)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## edema=0   263      106   130.79      4.70     51.57
## edema=0.5  29       19    10.53      6.82      7.37
## edema=1    20       19     2.68     99.45    102.40
##
##  Chisq= 112  on 2 degrees of freedom, p= <2e-16
```

which shows that patients with different `edema` status do not have the same survival overall. Looking into the observed and expected number of events, patients with `edema = 0` seems to be less hazardous than others since they have fewer observed events than expected. However, to formally test this, we'll need another log-rank test with appropriate multiple comparison correction.