Biostat 537: Survival Analaysis

TA Session 2

Ethan Ashby

January 12, 2024

Review of Last Week

1. Survival data is almost always subject to incompleteness – right censoring is the most common but other forms abound.
2. Survival analysis methods must account for censoring to (a) make efficient use of the available data and (b) avoid bias due to informative censoring.
3. The *independent censoring assumption* says survival information from participants in *any subgroup* censored at time t can be recovered from those in the same subgroup who remained at risk at time t.
4. The survivor function and hazard function are distinct but related quantities central to survival analysis techniques

$$h(t) = \frac{-\frac{d}{dt}S(t)}{S(t)} \qquad S(t) = \exp\left(-\int_0^t h(u)du\right)$$

Presentation Overview

Characteristics of Parametric Survival Models

Parametric models *fully specify* the shape of the distribution of the survival times.

Pros

1. Allows analytical calculation of quantities of interest: survivor function, hazard, mean survival times, etc.
2. Very efficient inference from data when model is correct.

Cons

1. May lead to very bad estimates if our model is incorrect!

Parametric Survival Distributions

Dist	Density f	Hazard h	Survivor S	Notes
Exponential(λ)	$f(t) = \lambda e^{-\lambda t}$,	$h(t) = \lambda$	$S(t) = e^{-\lambda t}$	
Weibull(α, λ)	$f(t) = \alpha\lambda(\lambda t)^{\alpha-1} e^{-(\lambda t)^{\alpha}}$	$h(t) = \alpha\lambda^{\alpha} t^{\alpha-1}$	$S(t) = e^{-(\lambda t)^{\alpha}}$	$p > 1 \uparrow$ $p < 1 \downarrow$ $p = 1$, Exp
Gamma(λ, β)	$f(t) = \frac{\lambda^{\beta} t^{\beta-1} e^{-\lambda t}}{\Gamma(\beta)}$	No closed form	No closed form	$\beta > 1 \uparrow$ $\beta < 1 \downarrow$ $\beta = 1$, Exp
Gen-Gamma(λ, β, p)	$f(t) = \frac{p\lambda^{p\beta} t^{p\beta-1} e^{-(\lambda t)^{p}}}{\Gamma(\beta)}$	No closed form	No closed form	$p = 1$, Gamma $\beta = 1$, Weibull $\beta = p = 1$, Exp

Some notable properties

1. *Memoryless property of the exponential distribution*: probability of failure depends only on the time increment

$$P(T > s + t | T > s) = P(T > t)$$

2. Piecewise-exponential can be a simple approach to approximate more complex hazards.
3. Weibull distributions are often a good starting point for parametric survival modeling in practice.
4. Weibull hazards are especially useful in *regression modelling* of survival data, as they can be viewed as proportional hazards and accelerated failure time (AFT) models.

Roadmap

Worked Example: Weibull Distribution

The Weibull(α, λ) has the following hazard function

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

The *cumulative hazard* takes the following form

$$H(t) := \int_0^t h(u)du = [(\lambda u)^{\alpha}]_0^t = (\lambda t)^{\alpha}$$

The *survivor function* takes the form

$$S(t) := \exp(-H(t)) = \exp\left(-(\lambda t)^{\alpha}\right)$$

Worked Example: Weibull Distribution

The *median survival time* is calculated by setting the survival function equal to 1/2 and solving for t!

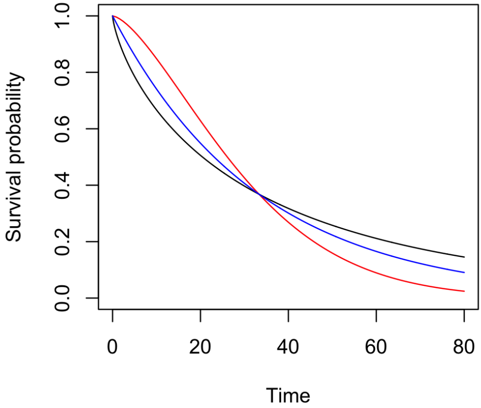
$$
\begin{aligned}
S(t) &= \exp\left(-(\lambda t)^{\alpha}\right) = 1/2 \\
\implies -(\lambda t)^{\alpha} &= -\log(2) \\
\implies \lambda t &= \log(2)^{1/\alpha} \\
\implies t_{1/2} &= \frac{\log(2)^{1/\alpha}}{\lambda}
\end{aligned}
$$

The *mean survival time* is calculable using the moment generating function (MGF) of Weibull distribution.

R example

```
#Plot survival function
weibSurv <- function(t, shape, scale){pweibull(t, shape=
    shape, scale=scale, lower.tail=F)}
curve(weibSurv(x, shape=1.5, scale=1/0.03), from=0, to
    =80, ylim=c(0,1), ylab="Survival probability", xlab="
    Time", col="red")
lines(x=seq(0, 80, by=0.1), sapply(seq(0, 80, by=0.1),
    FUN=function(x){weibSurv(x, shape=0.75, scale=1/0.03)
    }), col="black")
lines(x=seq(0, 80, by=0.1), sapply(seq(0, 80, by=0.1),
    FUN=function(x){weibSurv(x, shape=1, scale=1/0.03)}),
     col="blue")
```

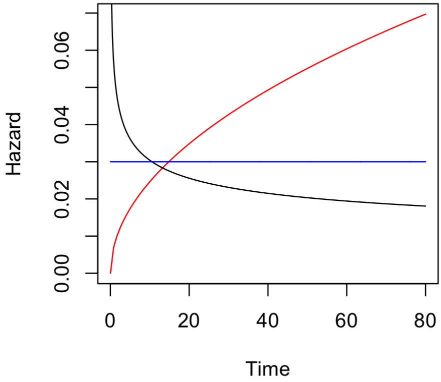
R Example



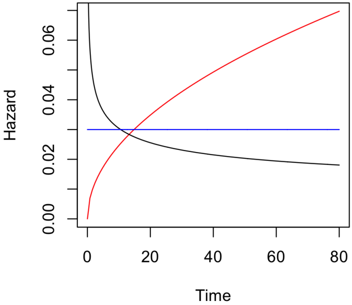
R example

```
#Plot hazard function
weibHaz <- function(x, shape, scale){dweibull(x, shape=
    shape, scale=scale)/pweibull(x, shape=shape, scale=
    scale, lower.tail=F)}
curve(weibHaz(x, shape=1.5, scale=1/0.03), from=0, to
    =80, ylab="Hazard", xlab="Time", col="red")
lines(x=seq(0, 80, by=0.1), sapply(seq(0, 80, by=0.1),
    FUN=function(x){weibHaz(x, shape=0.75, scale=1/0.03)
    }), col="black")
lines(x=seq(0, 80, by=0.1), sapply(seq(0, 80, by=0.1),
    FUN=function(x){weibHaz(x, shape=1, scale=1/0.03)}),
    col="blue")
#Plot random event times from weibull distribution
times = rweibull(n=500, shape=1.5, scale=1/0.03)
hist(times, xlab="Time", y="Count")
```

R Example



Check your understanding



Suppose the following three curves describe the hazard over 80 years of life from the following three causes.

1. Congenital Rubella
2. Alzheimers
3. Influenza

Can you match the disease to the shape of each hazard curve?

An R Note

Be sure to check parametrizations! Above, we described

Weibull(α, λ) which represents the ‘shape” and “rate” parametrization. R refers to Weibull(α, β) distribution are in

the “shape” and “scale” parameters, where $\lambda = 1/\beta$.

Roadmap

Estimation in Parametric Models via Maximum Likelihood

The beauty of parametric models is that they fully describe the data generating process, enabling calculation of survivor functions, hazards, mean/median survival times, and more.

In practice, we assume the shape of the survival time distribution (ex. Exp(λ)), but *use data* to estimate values for the parameters (λ). Once the parameters are estimated, we can convert them into estimates of quantities of interest (e.g., hazard).

Worked Example

Suppose we assume $T_1, \dots, T_n \overset{iid}{\sim} \text{Exp}(\lambda)$. Our goal is to estimate λ. Suppose that every survival time $\{T_i\}_{i=1}^{n}$ is completely observed. We write the *likelihood* of our data.

$$L(\lambda; T_1, \dots, T_n) := f(T_1; \lambda) \cdot \dots \cdot f(T_n; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda T_i}$$
$$= \lambda^n e^{-\lambda \sum_{i=1}^{n} T_i}$$

Our goal is to find the value of λ that *maximizes the likelihood* of the data. This is equivalent to maximizing the log-likelihood.

$$\log L(\lambda) = \ell(\lambda) = n \log(\lambda) + -\lambda \sum_{i=1}^{n} T_i$$

Worked Example

To solve for the *maximum likelihood estimate* $\hat{\lambda}$, we take the derivative wrt λ and set it equal to 0.

$$\frac{d}{d\lambda}\ell(\lambda) = 0$$

$$\implies \frac{n}{\lambda} = \sum_{i=1}^{n} T_i$$

$$\implies \hat{\lambda} = \left[\frac{\sum_{i=1}^{n} T_i}{n}\right]^{-1}$$

Hence, when the survival times are all completely observed and are from an exponential distribution, the MLE of λ is the reciprocal of the mean survival time or the mean event rate.

Worked Example

Suppose $T_1, \ldots, T_n \overset{iid}{\sim} \text{Exp}(\lambda)$. Our goal is to estimate λ. Suppose goal is to compute the MLE when some observations are right censored.

Let $\delta_i = \mathbb{I}(T_i \leq C_i)$ denote if the i-th survival time was censored. Then the likelihood is

$$L(\lambda) = \prod_{i=1}^{n} f(\lambda, T_i)^{\delta_i} S(\lambda, T_i)^{1-\delta_i}$$

Each unit with an observed ($\delta_i = 1$) survival time contributes $f(\lambda, T_i)$. Censored ($\delta_i = 0$) units have unknown survival times that are known to exceed T_i. Hence contribution is $S(\lambda, T_i)$.

Worked Example

Under the exponential distribution assumption, the likelihood with some observations censored is

$$L(\lambda) = \prod_{i=1}^{n} (\lambda e^{-\lambda T_i})^{\delta_i} (e^{-\lambda T_i})^{1-\delta_i}$$
$$= \lambda^{\sum_{i=1}^{n} \delta_i} e^{-\lambda \sum_{i=1}^{n} T_i \delta_i + T_i(1-\delta_i)} \equiv \lambda^{\sum_{i=1}^{n} \delta_i} e^{-\lambda \sum_{i=1}^{n} T_i}$$

For ease of estimation, find value of λ which maximizes the log-likelihood.

$$\ell(\lambda) = \left(\sum_{i=1}^{n} \delta_i\right) \log(\lambda) - \lambda \left(\sum_{i=1}^{n} T_i\right)$$

$$\frac{d}{d\lambda}\ell(\lambda) = 0 \implies \hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} T_i}$$

Worked Example

In the exponential model with independent right censoring, the MLE of the parameter λ is

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} T_i}$$

The blue term is the total number of observed events.
The red term is the person-time, or the total time that units were observed to be at risk prior to an event/censoring.

Worked Example

Maximum likelihood offers a framework to *estimate* key parameters of survival models from data. But another important task is *quantifying uncertainty* in our estimate.

A key quantity we will want to calculate is the *Information*.

$$I_n(\lambda) := -\frac{d^2}{d\lambda^2}\ell(\lambda)$$

In the exponential model

$$I_n(\lambda) = \frac{\sum_{i=1}^{n} \delta_i}{\lambda^2}$$

Worked Example

In large samples, the variance of the MLE $\hat{\lambda}$ is the reciprocal of the information $[I_n(\lambda)]^{-1}$. The fundamental result of Fisher & Cramer endows the MLE with the following amazing property.

$$\sqrt{n}(\hat{\lambda} - \lambda) \rightsquigarrow N(0, [I_1(\lambda)]^{-1})$$

This is a deep result that implies that in large samples, $\hat{\lambda}$ converges its target λ and exhibits uncertainty in the form of a normal distribution with known variance. This enables us to carry out tests and confidence intervals for λ!

For our purposes: in R

```
library(flexsurv);library(survival);library(tidyverse)
#Fit exponential survival model
expmodel <- flexsurv::flexsurvreg(Surv(rectime, censrec)
    ~1, data=flexsurv::bc, dist="exponential")

plot(expmodel, type="survival")
plot(expmodel, type="hazard")
plot(expmodel, type="cumhaz")
summary(expmodel, type="median")
summary(expmodel, type="mean")
#OR use "fitparametric" function
source("fitparametric.R")
expmodel <- fitparametric(Surv(bc$rectime, bc$censrec),
    dist="exp")
```

Summary

1. Parametric survival models completely determine the distribution of survival times using a finite set of parameters which need to be estimated from data.
2. In practice, we assume the *shape* of the distribution (e.g., Weibull), and *use data* to estimate the unknown parameters.
3. In parametric models, maximum likelihood is the framework we use to estimate and quantify uncertainty in parameter estimates from data.
4. Parametric models are comprehensive but not robust – may produce misleading results if the assumed shape is incorrect!

Roadmap

Why go nonparametric?

The use of parametric models are often justified using

1. Convenience: ease of converting between survival quantities of interest, relatively simple estimation.
2. Efficient: *when correctly specified*, parametric models produce estimators w/ smallest possible variances.

Reasons why we may want to go nonparametric

1. Agnosticism around choice of model shape.
2. True survival experience unlikely to adhere to rigid parametric assumptions.
3. Conclusions that avoid making non-essential statistical assumptions.

The Kaplan-Meier Estimator

The Kaplan-Meier Estimator is the product over the failure times of the conditional probabilities of surviving to the next failure time.

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

Where n_i is the number of individuals in the risk set at time t_i and d_i is the number of individuals who failed at time t_i.