Parametric Survival Models

David M. Rocke

November 14, 2023

Parametric Survival Models

In many areas of statistics, parametric models are the default and non-parametric models are sometimes an alternative. With measured data, linear regression is the default and non-parametric regression is a distant competitor. For survival analysis in biology and medicine, non-parametric methods have been dominant, though in engineering, parametric methods are common. Conditional on the assumptions, parametric methods can be superior, though often there is less difference in the conclusions than might naively be thought.

We will first discuss the common statistical distributions that are used in parametric time-to-event data analysis. The most basic is the exponential distribution, and some are generalizations of the exponential in the sense that for some choices of parameters the distribution is the exponential distribution. A different point of view than usual is that of considering the distribution of $Y = \ln T$ which often reorganizes a parametric family of distributions into the form of a location-scale family in the way that Gaussian distributions are a location-scale family.

We will then introduce some additional ways to look at time-to-event regression models based on the treatment in Kalbfleisch and Prentice (2002) and will show how the software in the survival package can be used to fit and examine parametric survival models. One technique is to model the log survival time as a linear function of parameters, which fits into the location-scale family viewpoint of classes of distributions, and is thus clearly related to linear regression.

Exponential Distribution

- The exponential distribution is the base distribution for survival analysis.
- The distribution has a constant hazard λ
- The mean survival time is λ^{-1}

$$
\begin{aligned}
f(t) &= \lambda e^{-\lambda t} \\
\ln(f(t)) &= \ln\lambda - \lambda t \qquad \text{linear in } t \\
F(t) &= 1 - e^{-\lambda t} \\
S(t) &= e^{-\lambda t} \\
\ln(S(t)) &= -\lambda t \qquad \text{linear in } t \\
h(t) &= -\frac{d}{dt}\ln(S(t)) \\
&= -\frac{d}{dt}(-\lambda t) \\
&= \lambda \\
h(t) &= f(t)/S(t) = \lambda e^{-\lambda t}/e^{-\lambda t} \\
&= \lambda
\end{aligned}
$$

Since

$$\ln(S(t)) = -\lambda t$$

we can check whether the exponential distribution fits a data set with survival function estimate $\hat{S}(t)$ by plotting $-\ln[\hat{S}(t)]$ and seeing if a straight line through the origin with slope λ fits the shape well. We can formally estimate λ using the exponential MLE previously derived.

If T is exponentially distributed with parameter λ, then $Y = \ln T$ has pdf

$$\begin{aligned} f_Y(y) &= \exp(y - \alpha - e^{y-\alpha}), \ -\infty < y < \infty, \text{where} \\ \alpha &= -\ln \lambda \end{aligned}$$

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(\ln t \leq y) \\ &= P(t \leq e^y) \\ &= 1 - \exp(-\lambda e^y) \\ f_Y(y) &= \lambda e^y \exp(-\lambda e^y) \\ &= e^{-\alpha} e^y \exp(-e^{-\alpha} e^y) \\ &= \exp(y - \alpha - e^{y-\alpha}) \end{aligned}$$

If we let $Y = \alpha + W$, then the pdf of W is

$$f_W(w) = \exp(w - e^w)$$

which is one of the extreme value distributions, so called because it is the limiting distribution of a standardized form of the minimum of a sample from a continuous distribution with support on $(-\infty, a)$ for some $a \leq \infty$. Extreme value theory shows that there is only a small number of possible forms that this limiting distribution can take. This one is called the *Gumbel distribution*.

R and the Exponential Distribution

```
Density, distribution function, quantile function and random generation for the
exponential distribution with rate = rate (i.e., mean = 1/rate).

dexp(x, rate = 1, log = FALSE)
pexp(q, rate = 1, lower.tail = TRUE, log.p = FALSE)
qexp(p, rate = 1, lower.tail = TRUE, log.p = FALSE)
rexp(n, rate = 1)
```

"The exponential distribution with rate λ has density $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$."

This is the same parametrization as just shown. Note that we can compute the hazard rate for any distribution as

```
dexp(x,rate)/pexp(x,rate,lower.tail=F)
```

which is $f(x)/S(x)$.

Weibull Distribution

One generalization of the exponential distribution is to allow the hazard function to depend on time, with the Weibull distribution having the functional form

$$\lambda(t) = \lambda\gamma(\lambda t)^{\gamma-1}, \quad \lambda, \gamma > 0$$

If $\gamma = 1$ then this is the exponential distribution with constant hazard λ. If $\gamma > 1$, then this is monotone increasing and if $\gamma < 1$, then it is monotone decreasing.

$$\begin{aligned} h(t) &= \lambda\gamma(\lambda t)^{\gamma-1} \\ f(t) &= \lambda\gamma(\lambda t)^{\gamma-1}e^{-(\lambda t)^{\gamma}} \\ S(t) &= e^{-(\lambda t)^{\gamma}} \end{aligned}$$

A check for the fit of the Weibull distribution is provided by the complementary log log (cloglog) plot of the survival function estimate vs. log t, since

$$\begin{aligned} \ln[-\ln S(t)] &= \ln[(\lambda t)^{\gamma}] \\ &= \gamma(\ln t + \ln \lambda) \end{aligned}$$

which should approximate a straight line with slope γ and intercept $\gamma \ln \lambda$.

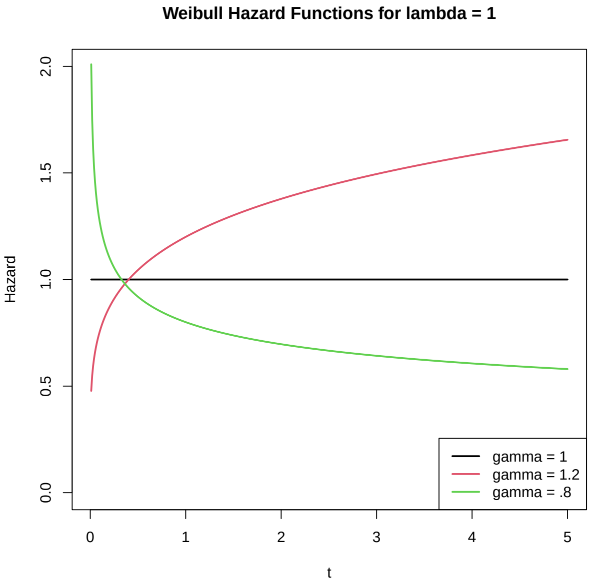
It can be shown that the pdf of $Y = \ln t$ is

$$f_Y(y) = \sigma^{-1} e^{(y-\mu)/\sigma} \exp(-e^{(y-\mu)/\sigma}), \quad \infty < y < \infty, \text{and where}$$
$$\sigma = \gamma^{-1}$$
$$\mu = \alpha = -\ln \lambda$$

or, more simply

$$y = \alpha + \sigma W$$

where W has the extreme value distribution shown for the exponential, where $Y = \alpha + W$. Thus, the shape of the density for $Y = \ln T$ is fixed and only the location and scale can be changed by λ and γ.


Weibull Hazard Functions for lambda = 1
Hazard
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
t
gamma = 1
gamma = 1.2
gamma = .8

R and the Weibull Distribution

```
Density, distribution function, quantile function and random generation for the
Weibull distribution with parameters shape and scale.

dweibull(x, shape, scale = 1, log = FALSE)
pweibull(q, shape, scale = 1, lower.tail = TRUE, log.p = FALSE)
qweibull(p, shape, scale = 1, lower.tail = TRUE, log.p = FALSE)
rweibull(n, shape, scale = 1)
```

The Weibull distribution with shape parameter a and scale parameter b has density given by

$$f(x) = (a/b)(x/b)^{a-1} exp(-(x/b)^a), x \geq 0.$$

The cumulative distribution function is

$$F(x) = 1 - exp(-(x/b)^a), x \geq 0.$$

the mean and variance are

$$E(X) = b\Gamma(1 + 1/a)$$
$$Var(X) = b^2 * (\Gamma(1 + 2/a) - (\Gamma(1 + 1/a))^2)$$

R and the Weibull Distribution

The Weibull distribution with shape parameter a and scale parameter b has density given by

$$f(x) = (a/b)(x/b)^{a-1} exp(-(x/b)^a), x \geq 0.$$

The cumulative distribution function is

$$F(x) = 1 - exp(-(x/b)^a), x \geq 0.$$

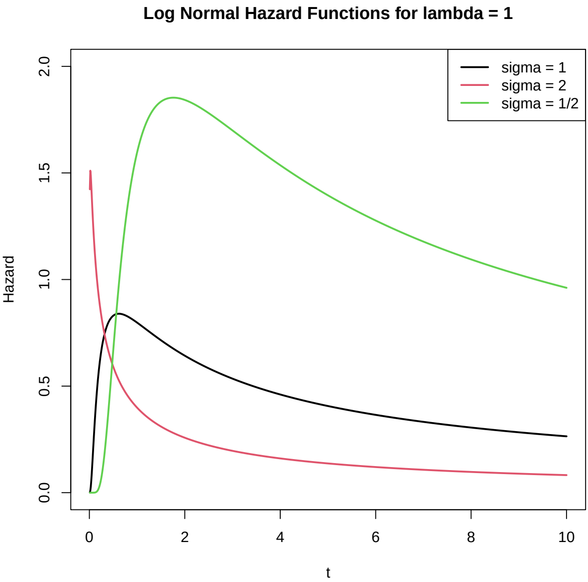
the mean and variance are

$$E(X) = b\Gamma(1 + 1/a)$$
$$Var(X) = b^2 * (\Gamma(1 + 2/a) - (\Gamma(1 + 1/a))^2)$$

The `scale` parameter $b = \lambda^{-1}$, and the `shape` parameter $a = \gamma$.

Lognormal Distribution

Here the model for $Y = \ln T$ is of the form $Y = \alpha + \sigma W$, but where W is a standard normal variate. This model is not as widely used as some since the log logistic is easier to handle and has similar properties. Also, the shape of the hazard function is often not realistic for medical and biological data.



Rising to a peak then monotonically declining towards 0.

R and the Lognormal Distribution

```
dlnorm(x, meanlog = 0, sdlog = 1, log = FALSE)
plnorm(q, meanlog = 0, sdlog = 1, lower.tail = TRUE, log.p = FALSE)
qlnorm(p, meanlog = 0, sdlog = 1, lower.tail = TRUE, log.p = FALSE)
rlnorm(n, meanlog = 0, sdlog = 1)
```

The log normal distribution has density

$$f(x) = 1/(\sqrt{2\pi}\sigma x)e^{-(\ln x - \mu)^2/(2\sigma^2)}$$

where μ and σ are the mean and standard deviation of the logarithm of x. The mean is $E(X) = \exp(\mu + 1/2\sigma^2)$, the median is $\text{med}(X) = \exp(\mu)$, and the variance is $Var(X) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$ and hence the coefficient of variation is $\sqrt{\exp(\sigma^2) - 1}$ which is approximately σ when that is small (e.g., $\sigma < 1/2$).

Gamma Distribution

Like the Weibull, the gamma is a two-parameter generalization of the exponential. The density is

$$f(t) = \frac{\lambda(\lambda t)^{k-1} e^{\lambda t}}{\Gamma(k)}$$

where $\Gamma(n) = (n-1)!$ for n a positive integer.
When $k = 1$ this is the exponential distribution. The gamma distribution is also the waiting time until the k'th event in a Poisson process.

The model for $Y = \ln T$ can be written as

$$Y = \alpha + W$$

where $\alpha = -\ln \lambda$ and W has the density

$$f_W(w) = \frac{\exp(kw - e^w)}{\Gamma(k)}$$

the mean and variance of which are the digamma function

$$\psi(k) = \frac{d \ln \Gamma(k)}{dk}$$

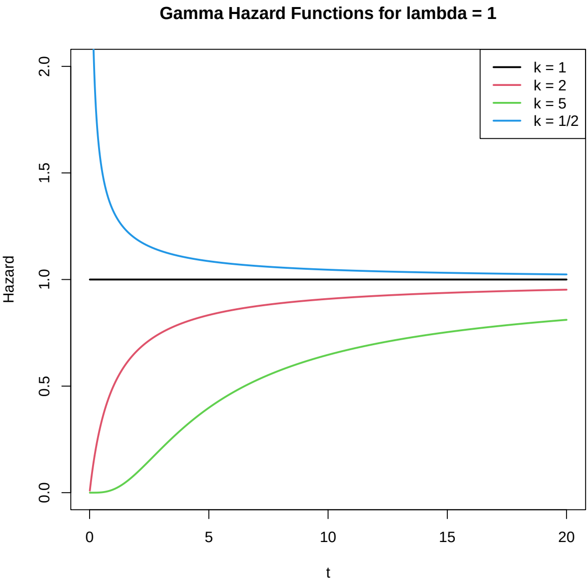
and the trigamma function

$$\psi^{(1)}(k) = \frac{d^2 \ln \Gamma(k)}{dk^2}$$

Stirling's formula is a series approximation for the natural log of the factorial function of which one form is

$$\begin{aligned} \ln\Gamma(k+1) &= \ln k! = k\ln k + O(\ln k) \\ \psi(k+1) &\approx \ln k \\ \psi^{(1)}(k+1) &\approx 1/k \end{aligned}$$

It can be shown that for the log failure time $Y = \ln T = \alpha + W$, when W is standardized by $\sqrt{k}(W - \ln k)$, it converges to a standard normal variate as $k \to \infty$.



The hazard rate converges to λ as t increases, from below for $k > 1$ and from above for $k < 1$

R and the Gamma Distribution

```
Density, distribution function, quantile function and random generation for the
Gamma distribution with parameters shape and scale.

dgamma(x, shape, rate = 1, scale = 1/rate, log = FALSE)
pgamma(q, shape, rate = 1, scale = 1/rate, lower.tail = TRUE,
       log.p = FALSE)
qgamma(p, shape, rate = 1, scale = 1/rate, lower.tail = TRUE,
       log.p = FALSE)
rgamma(n, shape, rate = 1, scale = 1/rate)
```

The Gamma distribution with parameters shape $= a$ and scale $= s$ has density

$$f(x) = 1/(s^a \Gamma(a)) x^{a-1} e^{-x/s}$$

so `shape`$= a = k$ and `scale`$= s = \lambda^{-1}$. For the same parametrization, use `rate`$= \lambda$.

Generalized Gamma Distribution

For a gamma variate, the log failure time can be written as $Y = \ln T = \alpha + W$. For the generalized gamma, we have $Y = \ln T = \alpha + \sigma W$. In addition to the rate parameter λ, the generalized gamma has parameters k and σ. This family includes several previously given distributions.

k	σ	Distribution
1	1	Exponential(λ)
1	$\sigma = \gamma^{-1}$	Weibull(λ, γ)
k	1	Gamma(λ, k)

One place to find functions relating to the generalized gamma distribution is in the R package `flexsurv`. The package contains two sets of functions for this distribution, one in the original parametrization, and one in terms of location and scale. the package also contains distributional functions for the generalized F distribution (in two parametrizations), the Gompertz distribution and can fit spline functions to the log cumulative hazards. User defined distributions can also be accommodated.

R and the Generalized Gamma Distribution

```
GenGamma.orig Generalized gamma distribution (original parameterisation)

Density, distribution function, hazards, quantile function and random generation
for the generalized gamma distribution, using the original parameterisation
from Stacy (1962).

dgengamma.orig(x, shape, scale = 1, k, log = FALSE)
pgengamma.orig(q, shape, scale = 1, k, lower.tail = TRUE, log.p = FALSE)
Hgengamma.orig(x, shape, scale = 1, k)
hgengamma.orig(x, shape, scale = 1, k)
qgengamma.orig(p, shape, scale = 1, k, lower.tail = TRUE, log.p = FALSE)
rgengamma.orig(n, shape, scale = 1, k)
```

For `GenGamma.orig`, scale is λ^{-1} and shape is Weibull γ.

R and the Generalized Gamma Distribution

```
GenGamma Generalized gamma distribution

Density, distribution function, hazards, quantile function and random generation
for the generalized gamma distribution, using the parameterisation originating
from Prentice (1974). Also known as the (generalized) log-gamma distribution.

dgengamma(x, mu = 0, sigma = 1, Q, log = FALSE)
pgengamma(q, mu = 0, sigma = 1, Q, lower.tail = TRUE, log.p = FALSE)
Hgengamma(x, mu = 0, sigma = 1, Q)
hgengamma(x, mu = 0, sigma = 1, Q)
qgengamma(p, mu = 0, sigma = 1, Q, lower.tail = TRUE, log.p = FALSE)
rgengamma(n, mu = 0, sigma = 1, Q)

dgengamma.orig(x,shape=shape,scale=scale,k=k) =
dgengamma(x,mu=log(scale) + log(k)/shape,sigma=1/(shape*sqrt(k)),Q=1/sqrt(k))
```

In addition to the Weibull and gamma distributions, this also can fit a log-normal distribution with $Q = 0$

Log-Logistic Distribution

Like the log-normal, this is defined by selecting a different error distribution for W in $Y = \alpha + \sigma W$. For the log logistic, the density of W is

$$f_W(w) = \frac{e^w}{(1 + e^w)^2}$$

This is a symmetric density with mean 0 and variance $\pi^2/3$ with slightly heavier tails than the normal. For t we have

$$f(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{[1 + (\lambda t)^\gamma]^2}$$

An advantage of the log-logistic over the log-normal is that there are closed form expressions for the survival function and the hazard function:

$$f(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{[1 + (\lambda t)^\gamma]^2}$$
$$S(t) = \frac{1}{1 + (\lambda t)^\gamma}$$
$$\lambda(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{1 + (\lambda t)^\gamma}$$

The numerator of the hazard function is the Weibull hazard. If $\gamma < 1$, it is monotone decreasing; if $\gamma = 1$ it is monotone increasing from λ; if $\gamma > 1$ it increases from zero to a maximum and then decreases toward 0.