Statistical Inference

Robert L. Wolpert
Institute of Statistics and Decision Sciences
Duke University, Durham, NC, USA
Spring, 2001

March 1, 2001

Week 7. Survival Analysis Example

1. Censoring and the Survival Function

We wish to explore models for survival times T_i, some of which are observed (say, $T_i = t_i$) and some of which are right-censored (say, $T_j > t_j$). To begin with we will take the times T_i to be *i.i.d.* from some continuous parametric probability distribution with density function $f(t \mid \theta)$, $\theta \in \Theta$, and will explore inference about θ upon observing times $\{t_i : 1 \le i \le n\}$ and failure indicators $\{\delta_i : 1 \le i \le n\}$— so $\delta_i = 1$ if we observe $T_i = t_i$, while $\delta_i = 0$ if we observe $T_i > t_i$.

1.1. Hazard and Survival

The probability of surviving at least t is $S_\theta(t) \equiv \Pr[T > t \mid \theta] = [1 - F(t \mid \theta)] = \int_t^\infty f(s)\,ds$, so the *conditional* probability of failing within time ϵ, given survival to time t, is

$$\Pr[T \le t + \epsilon \mid T > t, \theta] = \frac{S_\theta(t) - S_\theta(t+\epsilon)}{S_\theta(t)} \approx \frac{\epsilon f(t \mid \theta)}{S_\theta(t)} = \epsilon h_\theta(t),$$

where the *instantaneous hazard* is defined to be

$$h_\theta(t) \equiv f(t \mid \theta)/S_\theta(t) = -\frac{d}{dt} \log S_\theta(t),$$

derivative of the *cumulative hazard* $H(t) = \int_0^t h(s)\,ds = -\log S_\theta(t)$, and

$$S_\theta(t) = e^{-H_\theta(t)} = e^{-\int_0^t h_\theta(s)\,ds}.$$

The function $h_\theta(t)$, interpreted as the instantaneous conditional failure rate, may be used to suggest parametric models for survival. For example, the **exponential** failure model has constant hazard $h_\theta(t) \equiv \theta$; the **Weibull** failure model has $h_\theta(t) \equiv \alpha\beta t^{\alpha-1}$ for $\theta = (\alpha, \beta) \in \mathbb{R}_+^2$, etc.

1.2. Likelihood

The likelihood function for observed failures ($\delta_i = 1$) at times t_i is easily seen to be

$$\prod_{i:\delta_i=1} f(t_i \mid \theta) = \prod_{i:\delta_i=1} h_\theta(t_i) S_\theta(t_i)$$

while that for censored failures ($\delta_i = 0$) is

$$\prod_{i:\delta_i=0} S_\theta(t_i);$$

altogether this gives likelihood and log-likelihood

$$L(\theta) = \prod_{i\le n} h_\theta(t_i)^{\delta_i} S_\theta(t_i) \qquad \ell(\theta) = \sum_{i\le n} \delta_i \log h_\theta(t_i) - H_\theta(t_i) \tag{1}$$

in general, or

$$\begin{array}{rclrcl} L(\theta) & = & \prod_{i\le n} \theta^{\delta_i} e^{-\theta t_i} & L(\theta) & = & \prod_{i\le n} [\alpha\beta {t_i}^{\alpha-1}]^{\delta_i} e^{-\beta {t_i}^{\alpha}} \\ \ell(\theta) & = & (\Sigma\delta_i)\log\theta - \theta\Sigma t_i & \ell(\theta) & = & (\Sigma\delta_i)\log\alpha\beta + (\alpha-1)\Sigma\delta_i \log t_i - \beta\Sigma {t_i}^{\alpha} \end{array}$$

in the special cases of exponential and Weibull failure distributions, respectively, with maximum likelihood estimates

$$\hat\theta = \frac{\sum \delta_i}{\sum t_i} \qquad \hat\beta = \frac{\sum \delta_i}{\sum {t_i}^{\alpha}}.$$

($\hat\alpha$ isn't hard to find, but it's unavailable in closed form).

1.3. Estimation

In general (i.e. not assuming any parametric model) the maximum likelihood estimate for the conditional probability of survival to time $t+\epsilon$, given survival to t, is given simply by the sample fraction who do not die in that time interval,

$$\frac{\hat{S}(t+\epsilon)}{\hat{S}(t)} \approx \frac{N_R(t) - \text{deaths in } (t, t+\epsilon]}{N_R(t)}$$

where $N_R(t)$ is the "number at risk" at time t, the number in the sample whose failure or censoring time exceeds t.

Evidently $\hat{S}(t)$ is constant on intervals with no recorded death, and drops at observed death-times t_j by a fraction $N_D(t_j)/N_R(t_j)$, where $N_D(t)$ is the number of deaths at time t_j (which may exceed one in the event of ties), leading to the Kaplan-Meyer nonparametric MLE for the survival function,

$$\hat{S}(t) = \prod_{t_j \le t} \frac{N_R(t_j) - N_D(t_j)}{N_R(t_j)}$$

and a corresponding estimate for the cumulative hazard of

$$\hat{H}(t) = \sum_{t_j \le t} -\log\Big(1 - N_D(t_j)/N_R(t_j)\Big).$$

It is often instructive to plot either or both of these, possibly stratified by the values of covariate x_i which might be associated with survival, to look for patterns (e.g., systematic departures from the exponential model where $H(t) = \lambda t$ and $S(t) = \exp(-\lambda t)$), or to compare two or more survival or hazard curves for subpopulations (treated *vs.* control, for example).

2. Stanford Heart Transplant Data

In 197x the Stanford University Medical Center conducted an early clinical trial of the effectiveness of human heart transplants. Subjects were recruited throughout the trial period; each waited for a suitable donor heart to become available. The number of days each subject survived was recorded and reported.

2.1. Exploring the Data

One version of the Stanford Heart Transplant dataset appears in the book *Data* by D.F. Andrews and A.M. Herzog as dataset `T07.1`, and is available on-line from our course webpage. The first and last few lines are:

```
 7   1     1    1    15 1 54 1.11
 7   1     2    2     3 1 40 1.66
 7   1     3    3    46 1 42 0.61
 7   1     4    4   623 1 51 1.32
 7   1     5    5   126 1 48 0.36
...
 7   1   180  180    89 0 27 -9999
 7   1   181  181    60 0 13 -9999
 7   1   182  182    56 0 27 -9999
 7   1   183  183     2 0 39 -9999
 7   1   184  184     1 0 27 -9999
```

The record for each of the 184 subjects consists of a single line. The first two columns of each line are `7 1`, signifying that this is example 7.1 in Andrews and Herzberg; the next two columns are the line number and subject number, both sequential integers `1:184`; the next two columns are the subject's known survival time following surgery (in days) along with an indicator of the subject's status at that time (one for death and zero for censored observations— in this example, censoring arises for subjects still alive at the publication date of the data). The last two columns offer some covariates: the subject's age (in years) at transplant, and a measure of donor/subject tissue-type-mismatch (with "-9999" signifying a missing value).

One way to read these data into `S-Plus` would be:

```
dmat <- matrix(scan("T07.1"),byrow=T,ncol=8);
sht <- data.frame(days=dmat[,5], dead=dmat[,6], age=dmat[,7], tmm=dmat[,8]);
sht$tmm[sht$tmm == -9999] <- NA;
```

The Kaplan-Meyer MLE for the survival curve $\hat{S}(t)$ and cumulative hazard $\hat{H}(t) = -\log \hat{S}(t)$ can be plotted and compared to an exponential distribution with the commands

```
km  <- survfit(Surv(days,dead), sht);
summary(km);
```

```
plot(km);

plot(km$time, -log(km$surv), type="l"); # Cumulative hazard
lmfit <- lm(-log(km$surv) ~ km$time);   # Linear???
abline(lmfit);

plot(km$time, lmfit$resid); abline(h=0);
```

Evidently there appears to be a higher hazard rate during the first 100 days or so, then a lower one after the first 200 days or so;

```
lmfit100 <- lm(lmfit$resid ~ km$time, subset=(km$time<100));
abline(lmfit100);
haz100 <- lmfit$coef[2] + lmfit100$coef[2];  # Early Hazard

lmfit200 <- lm(lmfit$resid ~ km$time, subset=(km$time>200));
abline(lmfit200);
haz200 <- lmfit$coef[2] + lmfit200$coef[2];  # Late  Hazard

print(paste("Early hazard: ", haz100, ", Late hazard: ", haz200));
print(paste("Early E[T]: ", 1/haz100, ", Late E[T]: ", 1/haz200));
print(paste("Early log haz: ", log(haz100), ", Late log haz: ", log(haz200)));
```

Evidently the log hazard *after* the first six months is about -7.7, suggesting an expected lifetime of about six years, while that *during* the first three months is about -5.6 (or about 2.1 higher), suggesting only an 80% chance of three-month survival.

2.2. Modeling the Data

This suggests a survival model with one constant hazard $h(t \mid \theta) = \alpha$ for $t < \tau$, and with a different (presumably lower) hazard $h(t \mid \theta) = \beta$ for $t > \tau$, for some change-point τ and hazard rates $\alpha > \beta > 0$. From Equation (1),

$$\begin{aligned}
L(\theta) &= \prod_{t_i \le \tau} \alpha^{\delta_i} e^{-\alpha t_i} && \times \prod_{t_i > \tau} \beta^{\delta_i} e^{-\alpha\tau - \beta(t_i - \tau)} \\
\ell(\theta) &= \sum_{t_i \le \tau} [\delta_i \log \alpha - \alpha t_i] && + \sum_{t_i > \tau} [\delta_i \log \beta - \alpha\tau - \beta(t_i - \tau)] \\
&= \texttt{dlo} \ \log\alpha - \texttt{tlo} \ \alpha + \texttt{dhi} \ \log\beta - \texttt{thi} \ \beta - \texttt{nhi} \ \tau(\alpha - \beta)
\end{aligned}$$

where, for fixed $\tau > 0$, the log likelihood depends only on the five sufficient statistics $\texttt{dlo} \equiv \sum_{t_i \le \tau} \delta$, $\texttt{dhi} \equiv \sum_{t_i > \tau} \delta$, $\texttt{tlo} \equiv \sum_{t_i \le \tau} t_i$, $\texttt{thi} \equiv \sum_{t_i > \tau} t_i$, and **nhi** $\equiv \sum_{t_i > \tau} 1$. The log likelihood (as a function of $\theta \equiv (\log\alpha, \log\beta)$) is easily programmed in S-Plus (or C or MatLab or...); here is code to locate the MLE and generate a contour plot for θ near the point suggested by our earlier exploration, $\theta \approx (-5.6, -7.7)$:

```
tau <- 100;
dlo <- sum(sht$dead[sht$days <= tau]);
dhi <- sum(sht$dead[sht$days >  tau]);
tlo <- sum(sht$days[sht$days <= tau]);
thi <- sum(sht$days[sht$days >  tau]);
nhi <- sum(sht$days >  tau);
llh <- function(theta) {
    alpha <- exp(log.alpha <- theta[1]);
    beta  <- exp(log.beta  <- theta[2]);
    return (dlo*log.alpha - tlo*alpha +
           dhi*log.beta  - thi*beta  - nhi * tau * (alpha-beta));
}
setz <- function(na=41, nb=na, a=-5.6+c(-2,2), b=-7.7+c(-2,2)) {
   la <- seq(a[1],a[2],,na);
   lb <- seq(b[1],b[2],,nb);
   z <- za <- zb <- matrix(0,nrow=na, ncol=nb);
   for(i in 1:na) {
     for(j in 1:nb) {
        z[i,j]  <- llh(c(la[i],lb[j]));
        za[i,j] <- la[i]; zb[i,j] <- lb[i];
     }
   }
   top <- (z==max(z));
   mle <- c(mean(za[top]),mean(zb[top]));
   return(list(la=la, lb=lb, z=z, mle=mle));
}
out <- setz();
contour(out$la,out$lb,out$z,nlevels=20,labex=0);
print(paste("MLE: log(a,b)=(",out$mle[1],",",out$mle[2],")"));
abline(v=out$mle[1],h=out$mle[2]);
```

Evidently the maximum is near $(-5.7, -7.8)$, and the range $[-6.5, -5] \times [-8.5, -7]$ bears closer examination;

```
out <- setz(41,41,c(-6.5,-5.0),c(-8.5,-7.0));
```

```
contour(out$la,out$lb,out$z,nlevels=20,labex=0);
print(paste("MLE: log(a,b)=(",out$mle[1],",",out$mle[2],")"));
abline(v=out$mle[1],h=out$mle[2]);

out <- setz(41,41,c(-6,-5.5),c(-7.7,-7.3));
contour(out$la,out$lb,out$z,nlevels=20);
print(paste("MLE: log(a,b)=(",out$mle[1],",",out$mle[2],")"));
abline(v=out$mle[1],h=out$mle[2]);
```

The log likelihood appears nearly-normal (i.e., has regularly-spaced elliptical contours near the MLE), with mean approximately $\hat{\theta} \approx (-5.7, -7.5)$ and covariance approximately $\begin{pmatrix} 0.021 & 0 \\ 0 & 0.014 \end{pmatrix}$ (crudely estimated from contours, solving for σ^2 the equations $(x - \bar{x})^2/2\sigma^2 = \epsilon$ from observed values of $\bar{x}$ and of x at the extremes of the contour where $\epsilon = 1$). These can form the starting point for any of a number of integration methods.

3. Bayesian Analysis

Bayesian estimation of the parameter vector θ or of derived quantities such as the one, five, and ten year survival probabilities $S(365.25 \mid \theta)$, $S(1826.25 \mid \theta)$, and $S(3652.5 \mid \theta)$, all require integration over Θ of the form

$$\begin{aligned} E[g(\theta)] &= I[g]/I[1] \qquad \text{where} \\ I[g] &= \int_\Theta g(\theta) e^{\ell(\theta)} \pi(\theta) d\theta \end{aligned}$$

where $g(\theta)$ takes the values θ, $S(t \mid \theta)$, etc. These integrals can seldom be evaluated in closed form; here are a few approaches that have been employed in the literature:

1. **Quadrature**: The one-dimensional Riemann integral of a continuous function $f(x)$ over a finite interval (a, b) may be approximated by a weighted sum
$$\int_a^b f(x)\, dx \approx \sum_{i=0}^{n} w_i\, f(x_i)$$
with $x_i \equiv a + (i/n)(b - a)$ and $w_0, w_n = 1/2n$ and $w_i = 1/n$, $0 < i < n$ (the so-called "Trapezoidal Rule") or, better, with even $n = 2m$,

"Simpson's Rule" weights $w_0, w_n = 1/3n$, $w_{2j} = 2/3n$, and $w_{2j-1} = 4/3n$ for $0 < j < n/2$ which will give exactly the right answer for linear or quadratic functions $f(x)$ and will have approximation errors that decrease at rate n^{-4} as $n \to \infty$, while the trapezoidal rule's errors fall off only at rate n^{-2}. This approach is practical only in low-dimensional problems, since an array of n points in each dimension requires $N = n^q$ function evaluations if $\Theta \subset \mathbb{R}^q$. From the global perspective, the approximation error with N function evaluations will be of order $N^{-2/q}$ for Trapezoid or $N^{-4/q}$ for Simpson's, making them superior to Monte Carlo methods (whose error falls off like $N^{-1/2}$) only for $q \leq 3$ and $q \leq 7$ dimensions, respectively. Even in these ranges the methods can be inefficient; consider the problem of evaluating the posterior mean of a binomial parameter $\theta \in \Theta = (0, 1)$, with Jeffreys prior and with 80 successes in 100 tries: in the expression

$$I[g]/I[1] = \frac{\sum_{i=0}^n w_i (i/n)^{80.5} (1 - i/n)^{19.5}}{\sum_{i=0}^n w_i (i/n)^{79.5} (1 - i/n)^{19.5}},$$

with evenly-spaced $\theta_i = i/n$ and either Trapezoid or Simpson weights w_i, a huge fraction of the terms will be negligable because the posterior density is concentrated in a small portion of Θ.

2. **Monte Carlo Importance Sampling (MCIS)**: If we can find a probability distribution from which we can draw samples easily (Normal, t, beta, gamma, etc.), whose density function $p(\theta)$ is "similar" (nearly proportional) to the posterior density $\exp(\ell(\theta))\pi(\theta)$, we can draw a large number N of random draws $\theta_i \sim p(\theta)$ and exploit the law of large numbers to see

$$\begin{aligned} I[g] &= \int_\Theta g(\theta) e^{\ell(\theta)} \pi(\theta)\, d\theta \\ &= \int_\Theta g(\theta) \left[\frac{e^{\ell(\theta)} \pi(\theta)}{p(\theta)} \right] p(\theta)\, d\theta \\ &\approx \frac{1}{N} \sum g(\theta_i)\, w(\theta_i) \end{aligned}$$

where the "weight function" $w(\theta) \equiv \exp(\ell(\theta))\pi(\theta)/p(\theta)$ is nearly-constant by our requirement that $p(\theta)$ be "similar" to the posterior density. This leads directly to an approximate expectation

$$E[g(\theta)] \approx \sum g(\theta_i)\, w(\theta_i) \Big/ \sum w(\theta_i)$$

whose approximation error, by the Central Limit Theorem, is approximately normally-distributed with variance σ^2/n for some $\sigma > 0$ and so with mean square error $\sigma/\sqrt{n}$, in any number q of dimensions. As a side benefit, σ^2 can be estimated from the sample variance the $g(\theta_i)$'s, giving a reliable estimate of the accuracy of the MCIS approximation. The value of σ will depend on the choice of $p(\theta)$, and is smaller if $p(\theta)$ is nearly proportional to the posterior.

3. **Markov Chain Monte Carlo (MCMC)**: A recently developed method (Gelfand and Smith, 1990) that overcomes the difficulty in finding suitable "importance sampling distributions" $p(\theta)$ for MCIS, at the expense of drawing a *dependent* sample θ_i, is Markov Chain Monte Carlo. The most widely used version of this algorithm ("Metropolis-Hastings") begins with the selection of a starting point θ_0 (perhaps the MLE $\hat{\theta}$) and a suitable "step size" vector ϵ (often found by trial-and-error). The algorithm proceeds to generate a sequence θ_t of points in Θ by identifying at each $t > 0$ a "candidate" θ^*_{t+1}, which is either "accepted" (whereupon $\theta_{t+1} = \theta^*_{t+1}$) or "rejected" (whereupon $\theta_{t+1} = \theta_t$) with probabilities $\alpha = \alpha(\theta_t, \theta^*_{t+1})$ and $1-\alpha$ designed to ensure that the probability distribution of θ_t will converge, as $t \to \infty$, to the posterior distribution. For the commonly-used "random walk" proposal $\theta^*_{t+1} = \theta_t + \epsilon Z'_t$, with *i.i.d.* standard q-variate normal $Z_t \sim \mathsf{No}(0, I_q)$, Hastings' formula is

$$\alpha(\theta_t, \theta^*_{t+1}) = 1 \wedge \frac{e^{\ell(\theta_t)}\pi(\theta_t)}{e^{\ell(\theta^*_{t+1})}\pi(\theta^*_{t+1})}$$

(which simplifies to $e^{(\ell(\theta_t)-\ell(\theta*))^-}$ for uniform priors). The sequence θ_t exhibits serial correlation, and only asymptotically has the correct distribution, but this is enough for us to be able to evaluate

$$\mathsf{E}[g(\theta)] \approx \frac{1}{N-B}\sum_{B<i\leq N} g(\theta_i)$$

with an error that falls off like $1/\sqrt{N}$ (here we have thrown away the first $B \geq 0$ samples from the "burn-in" period to reduce the influence of early samples whose distribution is not yet close to the posterior density). The problems of assessing convergence in MCMC are subtle and challenging, but the method has proven immensely useful and practical even in problems whose parameter vector θ ranges over very high-dimensional spaces Θ. For many problems the computations can be done with the free `BUGS` software.

4. **Laplace Method**: If the integrand $g(\theta)\exp(\ell(\theta))\pi(\theta)$ is similar to a multivariate normal density function, i.e. if its logarithm $\phi(\theta) = \log g(\theta) + \ell(\theta) + \log \pi(\theta)$ is nearly quadratic

$$\phi(\theta) \approx \phi(\mu) + (\theta - \mu)' H(\theta - \mu)/2$$

near the point $\theta = \mu$ where it attains a maximum, where $H = \nabla^2 \phi(\mu)$ denotes the Hessian matrix of second partial derivatives evaluated at μ (note the gradient $\nabla\phi(\mu) = 0$ vanishes), then we can use the known value of the integral of the multivariate normal pdf to arrive at the approximation

$$\begin{aligned} I[g] &= \int_\Theta e^{\phi(\theta)}\, d\theta \\ &\approx \int_{\mathbb{R}^q} e^{\phi(\mu)+(\theta-\mu)'H(\theta-\mu)/2} d\theta \\ &= (2\pi)^{q/2}\, e^{\phi(\mu)} \Big/ \sqrt{\det -H} \end{aligned}$$

We can use this separately for both the numerator and denominator in the expression $\mathsf{E}[g(\theta)] = I[g]/I[1]$ to find an approximation that requires no simulation at all. While the approximation is often astonishingly accurate, it is difficult in applications to find even an estimate for the approximation error.

3.1. Reference Bayesian Analysis

For a reference Bayesian analysis using uniform prior $\pi(\theta) \equiv 1$ we may use MCIS with importance sampling distribution $p(\theta) = \mathsf{No}\left(\begin{pmatrix}-5.7\\-7.5\end{pmatrix}, \begin{pmatrix}0.021 & 0\\ 0 & 0.014\end{pmatrix}\right)$, with the following code:

```
surv <- function(t, theta) {
   alpha <- exp(theta[1]);  beta  <- exp(theta[2]);
   HAZ <- rep(0,length(t<-c(t)));
   if(sum(t<=tau)) {  HAZ[t<=tau] <- alpha * t;  }
   if(sum(t> tau)) {  HAZ[t> tau] <- alpha * tau + beta * (t-tau); }
   return(exp(-HAZ));
}
N <- 1000                                  # Size of MC sample
mua <- -5.7; sda <- sqrt(0.021);           # Log alpha Mean, SDev
```

```
mub <- -7.5; sdb <- sqrt(0.014);        # Log beta  Mean, SDev
                                        # Note t would be better!
theta <- cbind(rnorm(N,mua,sda), rnorm(N,mub,sdb));
wgt   <- exp(apply(theta,1,llh)-llh(c(mua,mub))) /
            (dnorm(theta[,1],mua,sda)*dnorm(theta[,2],mub,sdb));
g <- cbind(theta,surv( 365.25,theta),  #  1-yr survival prob
                 surv(1826.25,theta),  #  5-yr survival prob
                 surv(3652.50,theta)); # 10-yr survival prob
gbar <- (wgt %*% g)/sum(wgt);
print(gbar);
hist(wgt);                              # Are the weights okay?
```

Evidently the posterior mean for $(\log\alpha, \log\beta)$ is $\mathsf{E}[\theta] = (-5.713, -7.501)$, while the expected survival probabilities for one, five, and ten years are 37.0%, 1.0%, and 1/8925, respectively. Note that the latter is a little worrisome, since one of our 184 observations *does* exceed ten years. This suggests either that we were lucky to have a long-term survivor (the probability of having at least one ten-year survivor among 184 subjects would be approximately 2%), or that we have overestimated the long-term hazard rate β, or that the piecewise-constant hazard model may be imperfect in the tails. We may learn more by computing the posterior CDF for `surv(3652.5,theta)`, approximately available for example by taking a k-dimensional $g_i(\theta) = 1$ if $\mathrm{surv}(3652.5, \theta) \le i/k$ for $1 \le i \le k$, leading to an estimate of the CDF and from it a histogram estimate of the pdf. But exploration and further model elaboration can wait for another day...