BIO 244: Unit 2

Likelihood Construction, Inference for Parametric Survival Distributions

In this section we obtain the likelihood function for noninformatively right-censored survival data and indicate how to make an inference when a parametric form for the distribution of T is assumed. While the focus of this course is on nonparametric and semiparametric inference, it is useful to consider parametric inference for right-censored survival data for at least two reasons: first, parametric inference can be very helpful in settings with small sample sizes or when scientific insights can be gleaned from the shapes of survival curves; secondly, such inferences provide a useful contrast to inference in more traditional settings, where the observations are direct realizations from the underlying distribution of interest. When observations are subject to censoring, then likelihood methods inform us about the distribution of the observables, which depends on both the underlying survival process and the process that leads to censored observations.

2.1 Observables, Likelihood Function:

For simplicity, consider the "1-sample" problem in which there are n *i.i.d.* survival times, denoted $T_1, T_2, \ldots, T_n$, with a common and unknown c.d.f. $F(\cdot)$, about which we wish to make an inference.

Suppose that we don't observe $T_1, \cdots, T_n$ but instead observe (U_i, δ_i) for $i = 1, 2, \ldots, n$, where

$$\begin{aligned} U_i &= \min(T_i, C_i) \\ \delta_i &= 1(T_i \leq C_i) \end{aligned}$$

and C_i is the (fixed or random) "potential censoring time" for item i in the sense that we will have a censored observation at time u if $C_i = u$ and $T_i > u$.

Let $\perp$ denote "is independent of". We assume that $T_i \perp C_i$ (noninformative censoring) and that the n pairs (U_i, δ_i), for $i = 1, \cdots, n$, are also i.i.d.

Likelihood Construction: Note that the bivariate random variable (U_i, δ_i) consists of a continuous component U_i and a binary component δ_i. (U_i, δ_i) can take two forms

$$(U_i, \delta_i) = (u_i, 1) : \qquad T_i \text{ is uncensored at time } u_i$$

$$(U_i, \delta_i) = (u_i, 0) : \qquad T_i \text{ is censored at time } u_i.$$

The likelihood contribution for (U_i, δ_i), say $L_i(F)$, is proportional to the probability elements corresponding to these two types of outcomes.

CASE 1: C_i known constants

$$\begin{aligned} L_i(F) &= \begin{cases} f(u_i) & \text{if } \delta_i = 1 \\ 1 - F(u_i) & \text{if } \delta_i = 0 \end{cases} \\ &= f(u_i)^{\delta_i}[1 - F(u_i)]^{1-\delta_i} \end{aligned}$$

$$\therefore \quad L(F) = \prod_{i=1}^{n} L_i(F) = \prod_{i=1}^{n} \left\{ f(u_i)^{\delta_i}[1 - F(u_i)]^{1-\delta_i} \right\}. \qquad (2.1)$$

Note that (2.1) does not take the form of the usual likelihood function that would result if we had observed $T_1, \cdots, T_n$; however, it functionally depends only on $F(\cdot)$ and thus can be maximized to make inferences about $F(\cdot)$.

CASE 2: $C_i \quad i.i.d. \sim G \quad$ (G continuous with p.d.f. g)

Now suppose that the potential censoring times C_i are independent random variables with distribution function $G(\cdot)$ and density function $g(\cdot)$. The likelihood contributions for the 2 types of observations are:

Event	Also Expressible As	Likelihood Contribution
$U_i = u_i, \delta_i = 1$	$[T_i = u_i, C_i \geq u_i]$	$f(u_i)[1 - G(u_i)]$
$U_i = u_i, \delta_i = 0$	$[T_i > u_i, C_i = u_i]$	$[1 - F(u_i)]g(u_i)$

In fact, this is the *density* of the observables (U_i, δ_i) (Exercise 7).

Note how the independence of T_i and C_i, which ensures noninformative censoring, is instrumental in the determination of the likelihood contributions above. As we see below, the fact that the likelihood contribution factors into one term involving F and another involving G greatly simplifies the resulting inference.

$$\therefore \quad L(F, G) = \prod_{i=1}^{n} \left\{ \left(f(u_i)[1 - G(u_i)]\right)^{\delta_i} \left([1 - F(u_i)]g(u_i)\right)^{1-\delta_i} \right\}$$

$$= \prod_{i=1}^{n} \left\{ f(u_i)^{\delta_i} [1 - F(u_i)]^{1-\delta_i} \right\} \cdot \prod_{i=1}^{n} \left\{ [1 - G(u_i)]^{\delta_i} g(u_i)^{1-\delta_i} \right\}. \qquad (2.2)$$

Suppose that F and G are functionally independent; i.e.,

$$F = F_\theta,\ G = G_\phi \quad \theta \in \Theta,\ \phi \in \Phi \ \text{ and par. space is } \ (\theta, \phi) \in \Theta \times \Phi.$$

Then for purposes of inference about F, the terms in L involving G can be regarded as a constant. That is, the maximizing value of $F(\cdot)$ for this likelihood function is the same as from maximixing (2.1). As we illustrate below, this is not to say that the properties of the resulting estimators of $F(\cdot)$ are the same. In this case, the score and information for θ only depends on first factor, so the usual theory about MLEs can be applied!

2.2 Parametric Inference for the Exponential Distribution:

Let us examine the use of (2.1) for the case where we have (noninformatively) right-censored observations from the exponential distribution. We begin with the 1-sample problem and then discuss the comparison of two groups and the analysis of covariates. We assume that the potential censoring times C_i are *i.i.d.* random variables with c.d.f. $G(\cdot)$.

1-Sample Problem:

Suppose $T_1,\ T_2, \ldots,\ T_n$ are i.i.d. Exp(λ), and subject to noninformative right censoring. Then (2.1) becomes

$$
\begin{aligned}
L = L(\lambda) &= \prod_{i=1}^{n} \left\{ (\lambda e^{-\lambda u_i})^{\delta_i} (e^{-\lambda u_i})^{1-\delta_i} \right\} \\
&= \lambda^r e^{-\lambda W}, \qquad \text{where } r = \sum_{i=1}^{n} \delta_i = \# \text{ uncensored obs.} \\
&\qquad\qquad\qquad \text{and } W = \sum_{i=1}^{n} u_i = \text{total observed time.}
\end{aligned}
$$

Taking derivatives, we find

$$
\left.\begin{aligned}
\frac{\partial \ln L}{\partial \lambda} &= \frac{r}{\lambda} - W \\
-\frac{\partial^2 \ln L}{\partial \lambda^2} &= \frac{r}{\lambda^2}
\end{aligned}\right\}
\Longrightarrow
\begin{aligned}
\hat{\lambda} &= \frac{r}{W} \\
\hat{I}(\lambda) &= \frac{r}{\lambda^2} \\
\hat{i}(\lambda) &= \frac{r}{n\lambda^2},
\end{aligned}
$$

where we use $\hat{I}(\lambda)$ and $\hat{i}(\lambda)$ to denote the observed (or sample) information for the sample and an individual subject, respectively; e.g., $\hat{I}(\lambda) \overset{def}{=} -\frac{\partial^2}{\partial \lambda^2} lnL(\lambda)$.

Note that r, the number of uncensored observations has the binomial distribution; that is,

$$\begin{aligned} r \sim \text{binomial}(n,p), \text{ where } p &= P(\delta_i = 1) \\ &= \int_0^\infty f(u)[1-G(u)]du. \end{aligned}$$

Therefore,

$$I(\lambda) \overset{def}{=} E[\hat{I}(\lambda)] = \frac{np}{\lambda^2} \qquad i(\lambda) \overset{def}{=} E[\hat{i}(\lambda)] = \frac{p}{\lambda^2}.$$

Notice that to estimate the observed information $\hat{I}(\lambda)$ one does not need to estimate $G(\cdot)$, but $I(\lambda)$ does depend functionally on $G(\cdot)$.

Assuming the appropriate regularity conditions hold, as $n \to \infty$,

$$\frac{\sqrt{n}(\hat{\lambda}-\lambda)}{\sqrt{i^{-1}(\lambda)}} = \frac{(\hat{\lambda}-\lambda)}{\sqrt{I^{-1}(\lambda)}} \quad \xrightarrow{\mathcal{L}} \quad N(0,1)$$

$$\therefore \qquad \text{or } \hat{\lambda} \overset{\text{apx}}{\sim} N(\lambda, I^{-1}(\lambda)) = N(\lambda, \frac{\lambda^2}{np})$$

(see e.g. Andersen et al, Chapter VI, Section 1). Since λ is not known, we can invoke Slutsky's theorem and use the approximation, replacing $\lambda^2/(np)$ with $\hat{\lambda}^2/(n\hat{p})$. That is,

$$\hat{\lambda} \overset{\text{apx}}{\sim} N(\lambda, \hat{I}^{-1}(\hat{\lambda})) = N(\lambda, \frac{r}{W^2}).$$

It turns out that a better approximation is to assume that the log of $\hat{\lambda}$ is normal. Using the delta method, this gives: $\ln\hat{\lambda} \overset{\text{apx}}{\sim} N(\ln\lambda, \frac{1}{r})$.

Suppose we wanted to use these results to test the hypothesis $H_0 : \lambda = \lambda_0$ that the true underlying λ equalled some specified value, say λ_0, or to find an approximate 95% confidence interval (CI) for λ.

Then under H_0,

$$Z_1 \overset{\text{def}}{=} \frac{\ln\hat{\lambda} - \ln\lambda_0}{\sqrt{\frac{1}{r}}} \approx N(0,1)$$

$\rightarrow$ Use Z_1 to test H_0.

<u>CIs:</u>

$$P\left[\ln\hat{\lambda} - 1.96\sqrt{\frac{1}{r}} < \ln\lambda < \ln\hat{\lambda} + 1.96\sqrt{\frac{1}{r}}\right] \approx .95$$

$$\Longrightarrow P\left[\hat{\lambda}e^{-1.96\sqrt{\frac{1}{r}}} < \lambda < \hat{\lambda}e^{1.96\sqrt{\frac{1}{r}}}\right] \approx .95\ ,$$

and thus $[\hat{\lambda}e^{-1.96\sqrt{\frac{1}{r}}}, \hat{\lambda}e^{1.96\sqrt{\frac{1}{r}}}]$ is an approximate 95% CI for λ.

Example: STATA data set EX1.dta contains the survival results for 95 patients receiving AZT and 82 receiving ddI in a recent AIDS clinical trial. The data are briefly described in file EX1.doc and will be discussed in the lab.

For the AZT group, $r = 50$ and $W = 80.8$ years, and thus $\hat{\lambda} = .6188$

approximate 95% CI for λ : $(.4690, .8165)$
(based on $\ln\hat{\lambda} \approx$ Normal)

2-Sample Problem:
Suppose we wanted to compare the λ for 2 groups, say the patients receiving ddI and those receiving AZT.

$$H_0: \ \lambda_A = \lambda_d \qquad \begin{array}{rcl} (\lambda_A & = & \text{exponential par. for AZT group} \\ \lambda_d & = & \text{exponential par. for ddI group).} \end{array}$$

Assume noninformative censoring in each group (!). Then using the normal approximations from the 1-sample problem, define

$$Z_2 \overset{\text{def}}{=} \frac{\ln\hat{\lambda}_A - \ln\hat{\lambda}_d}{\sqrt{\frac{1}{r_A} + \frac{1}{r_d}}} \quad \overset{\text{apx}}{\approx} \quad N(0,1) \quad \text{under } \ H_0.$$

For the data in AZT-ddI.dta, $\quad Z_2 = \cdots = .218 \qquad \left(\begin{smallmatrix} p = .83 \\ \text{(2-sided)} \end{smallmatrix}\right)$

$\rightsquigarrow$ No significant difference between AZT and ddI groups.

Covariates:
Next suppose $\mathbf{z}$ is a $p \times 1$ vector of covariates measured for each subject, and we are interested in assessing whether the components of $\mathbf{z}$ are associated with survival.

For the i^{th} of n independent subjects, suppose that the value of $\mathbf{z}$ is denoted $\mathbf{z_i}$. Thus, the observation for subject i is of the form $(\mathbf{z}_i, U_i, \delta_i)$.

$\longrightarrow$ How might we analyze such data to make an inference about the association of the components of $\mathbf{z}$ on the underlying survival times?

Assume: noninformative censoring: T_i independent of C_i given Z_i.

- **ONE APPROACH:** Continue to assume that each T_i is exponential, but let the parameter λ depend functionally on the corresponding value of $\mathbf{z}$. That is, assume $T_i \sim Exp(\lambda_i)$, where

  $\lambda_i = \text{function}(\mathbf{z_i})$

  $$\text{e.g.} \quad \lambda_i = \lambda_0 \cdot e^{\beta' \mathbf{z_i}}.$$

  That is, we assume our n independent underlying survival times $T_1, \cdots, T_n$ have different exponential distributions, but where the scale parameters are linked through a known function of the covariates $\mathbf{z}_1, \cdots, \mathbf{z_n}$. Then L (equation 2.1) is a function of (λ_0, β), and so we can employ standard likelihood methods to make inferences about (λ_0, β). For example, the hypothesis that the first component of $\mathbf{z}$ is not associated with survival is given by the zeroness of the first component of β. Once we have the MLEs of the parameters (λ_0, β), such tests can be made using standard methods, such as Wald tests.

  This approach of functionally relating a parameter of the survival distribution to the covariates was first considered by Feigl & Zelen (1965).

Note: In ordinary regression problems, we usually expressed observations as

$$\text{observed} = \underbrace{\text{expected}}_{\text{function of covariates}} + \text{error}.$$

However, it is not obvious how to extend such an approach to settings where some observations are right censored. In contrast, with the Feigel–Zelen approach, we express λ_i as a function of the covariate value $\mathbf{z}_i$. Once this is done, we take account of censoring by using the likelihood function (2.1).

Other Parametric Families:
More generally, the same approach can be used to make parametric inferences in the presence of (noninformatively) right censored observations for other parametric families (e.g., Weibull, Gamma, ...). One simply uses (2.1) and proceeds in the same way as for the exponential distribution.

One concern with any parametric approach is whether the resulting inference remains valid if the assumed underling parametric distribution does not fit the data. One way to try to make the resulting inferences robust is to fit a 'weakly structured' parametric model to the data. To illustrate, consider the 1-sample problem where

$$T_1,\ T_2, \ldots,\ T_n \quad i.i.d. \ \sim \ F(\cdot),$$

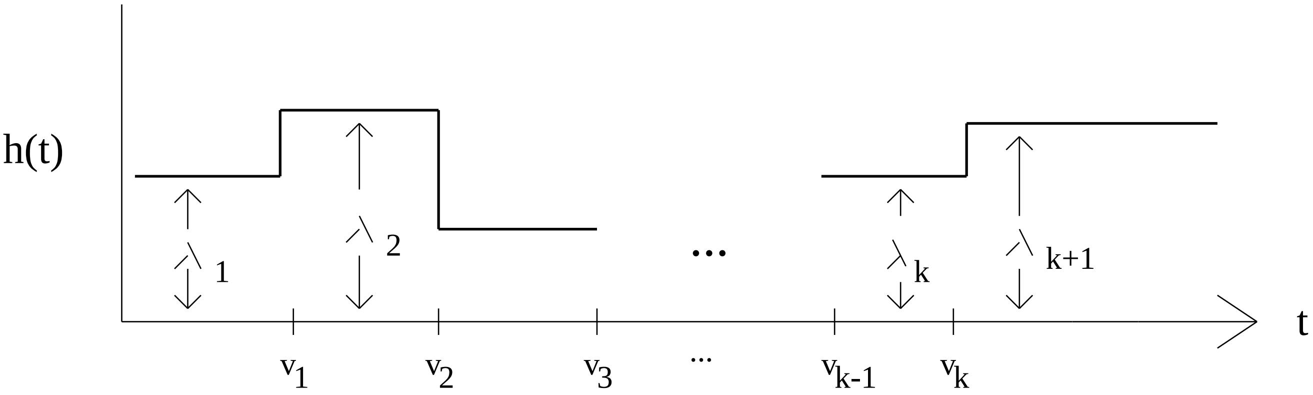
where $F(\cdot)$ has the "piecewise exponential" distribution. That is,suppose that

$$0 = v_0 < v_1 < \cdots < v_k < v_{k+1} = \infty$$

is a known partition of $[0, \infty\}$, and we assume that the hazard function for T is of the form

$$h(t) = \lambda_j \quad \text{for} \quad v_{j-1} \leq t < v_j.$$

Thus, the unknown parameters of $F(\cdot)$ are $\lambda_1, \cdots, \lambda_{k+1}$. The assumed hazard is depicted below:



Let I_j denote the interval $[v_{j-1}, v_j)$, for $j = 1, 2, \ldots, k+1$.

Then for $t \in I_j$,

$$\begin{aligned} H(t) &= \sum_{l=1}^{j-1} \lambda_l (v_l - v_{l-1}) + \lambda_j (t - v_{j-1}) \\ F(t) &= 1 - e^{-H(t)} \\ f(t) &= h(t) e^{-H(t)} = \lambda_j e^{-H(t)} \end{aligned}$$

$$L = L(\lambda_1, \lambda_2, \ldots, \lambda_{k+1}) = \prod_{i=1}^{n} \left\{ f(u_i)^{\delta_i} \left[1 - F(u_i)\right]^{1-\delta_i} \right\}$$

$$= \cdots = \prod_{j=1}^{k+1} \lambda_j^{r_j} e^{-\lambda_j W_j},$$

where $\quad r_j = \sum_{i=1}^{n} \delta_i \cdot 1[u_i \in I_j] =$ # uncensored times in I_j, and

$$W_j = \underbrace{\sum_{i=1}^{n} (u_i - v_{j-1}) 1[u_i \in I_j]}_{\text{contributions for } u_i \text{ falling in } I_j} + \underbrace{(v_j - v_{j-1}) \sum_{i=1}^{n} 1(u_i > v_j)}_{\text{contributions for } u_i \text{ that exceed } v_j}$$

$=$ total time observed in the “window” I_j.

Given this likelihood function, we can proceed in the usual way to make inferences. Specifically,

$$\hat{\lambda}_j = r_j / W_j,$$

and

$$\hat{I}(\lambda_1, \lambda_2, \ldots, \lambda_{k+1}) = \text{diagonal } \left(\frac{r_1}{\lambda_1^2}, \frac{r_2}{\lambda_2^2}, \ldots, \frac{r_{k+1}}{\lambda_{k+1}^2} \right)$$

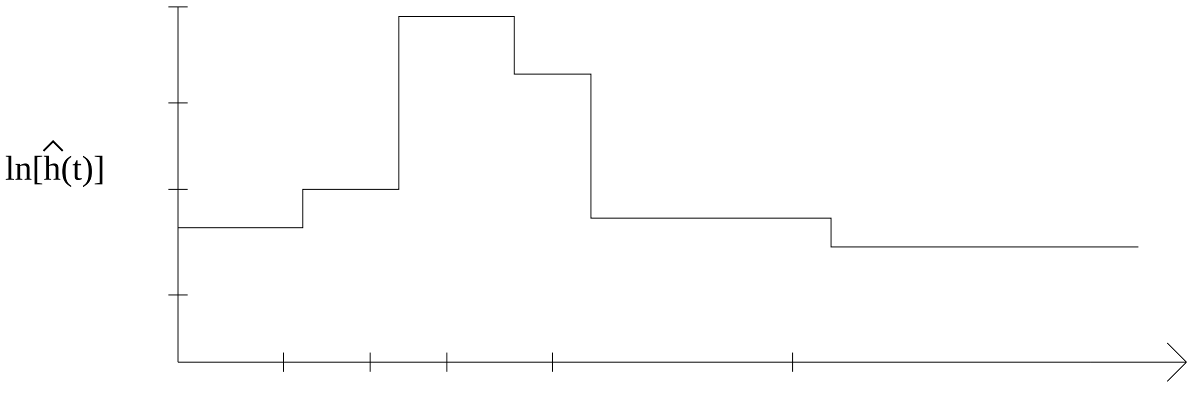
$$\Longrightarrow (\hat{\lambda}_1, \ldots, \hat{\lambda}_{k+1}) \quad \text{approximately uncorrelated.}$$

As before, we can approximate $\ln\hat{\lambda}_j \overset{\text{apx}}{\approx} N\left(\ln\lambda_j, \frac{1}{r_j}\right)$.

Applying this to the AZT-ddI.dta (pooling the AZT & ddI groups) gives the following:

$$k = 5,\ v_1 = .25,\ v_2 = .5,\ v_3 = .75,\ v_4 = 1,\ v_5 = 1.5$$

Interval	r_j	W_j	$\hat{\lambda}_j$	$\ln\hat{\lambda}_j$	$\hat{\text{V}}\text{ar}(\ln\hat{\lambda}_j)$
[0, .25)	19	42.0	.45	-.79	.05
[.25, .50)	21	35.8	.59	-.53	.05
[.50, .75)	28	26.35	1.06	.06	.04
[.75, 1.0)	14	17.86	.78	-.24	.07
[1.0, 1.5)	10	21.93	.46	-.79	.10
[1.5, ∞)	3	6.90	.43	-.83	.33



Taken at face value, the plot suggests that perhaps the hazard function is unimodal. However, the variances of the $\hat{\lambda}_j$ are considerable and not equivalent.

One could use a weakly structured model such as this to assess whether a specific parametric assumption (say, the Weibull distribution) is reasonable, or as an estimator in its own right.

NOTE 1: In the preceding, it may be preferable to plot $\ln\hat{\lambda}_j$ (or equivalently, $\hat{\lambda}_j$ on a log scale) instead of $\hat{\lambda}_j$ because the precision will vary less with variations in the λ_j.

NOTE 2: The preceding assumes the intervals were pre-specified. An alternative is to let the data determine the intervals so that there are equal numbers of uncensored events (e.g., 5) per interval. The resulting estimates of $\ln\hat{\lambda}_j$ will then have approximately equal precision. See Cox (1979) for details.

Another approach would be to allow both the λ_j and the v_j to be random. The special case where $k = 1$ and we partition the time axis into 2 intervals is sometimes referred to as a "changepoint" problem: we have a process whose distribution we postulate to "shift" at some unknown time point v_1. The unknowns in our setting are the hazards λ_1 and λ_2 and the time, v_1 at which the shift occurs. See, for example, Hinkley (1970) for an example of how such problems arise and how traditional estimators can sometimes give poor results.

Recall equation (2.2), the likelihood function $L(F, G)$ when the potential censoring times are assumed to be i.i.d. with c.d.f. $G(\cdot)$. In the development following (2.2) we assumed that $G(\cdot)$ was from some parametric family, and thus $L(F, G)$ was actually a function over some finite dimensional space. Suppose instead we wanted to leave $G(\cdot)$ unrestricted; conceptually, we still can think about the maximizing value, say $(\hat{F}, \hat{G})$, of (2.2), where the space over which the maximum is taken involves both the parametric space for $F(\cdot)$ and the unrestricted space for $G(\cdot)$.

What can we say about the solution $(\hat{F}, \hat{G})$? Given the factorization of (2.2) as a product of 2 terms, one involving F and one involving G, it follows that $\hat{F}$ can be obtained by simply maximizing (2.1). Another way to view this is to consider the "profile likelihood function" for F, defined as

$$L_p(F) \overset{def}{=} sup_G L(F,G) \ .$$

This is proportional to (2.1). A formal theory for inference in such a "semi-parametric" setting exists (cf: Murphy & van der Vaart, 2000). We return to this later in the course.

Exercises

1. Verify omitted algebraic details for piecewise exponential likelihood.

2. Suppose $T_1, T_2, \cdots, T_n$ are *i.i.d.* $W(\lambda, p)$ and data are noninformatively right censored (i.e., we observe (U_i, δ_i) for $i = 1, 2, \ldots, n$).

   - Determine the (observed) score and information equations; note: MLEs not expressible in closed form.
   - How would you use these to test $H_0: \ T_i \sim Exp(\lambda) \quad i = 1, 2, \ldots, n$?

3. For $T_1, T_2, \ldots, T_n$ *i.i.d.* piecewise exponential (with censoring), how would you test $H_0: \ \lambda_1 = \lambda_2 = \cdots = \lambda_{k+1}$?

4. Consider the 2 tests for lack of fit to an exponential model that are suggested in (2) and (3). Briefly indicate the possible advantages and disadvantages of each.

5. A multiplicative intensity model is one where the hazard function factors into a term involving time (but not covariates) and a term involving the covariates (but not time). Suppose you wanted to fit a multiplicative intensity model which assumes that the underlying hazard function is piecewise constant over the time intervals determined by the known times $0 < v_1 < ... < v_k < v_{k+1} = \infty$. Let the $p \times 1$ covariate vector be denoted $z = (z_1, \ z_2, ..., \ z_p)$.

   (a) Propose a specific multiplicative intensity model in terms of $h(t \mid z)$, the hazard function at time t for someone with covariate vector z.

   (b) Suppose your data consists of noninformatively right-censored data from n independent subjects, and that the observation for subject i is given by (U_i, δ_i, z_i), where U_i and δ_i are defined in the usual way (such as in the preceding problem). Write down the likelihood function, and describe briefly how you would estimate parameters in this model.

   (c) Given your model, how do you express the hypothesis that the hazard function for any individual is constant in t? Briefly (1-2 sentences) indicate how you might test this hypothesis.

6. Show formally that the likelihood on page 3 arises from the density of the observables (U_i, δ_i).

Additional Reading

For additional reading about parametric inferences, see Lawless (2003). Buckley & James (1979) were the first to attempt to adapt ordinary regression approaches (where the observations are expressed as a mean plus error term) to right censored data.

References

Buckley & James (1979). Linear Regression with Censored Data, *Biometrika* 66: p.429-436.

Cox DR (1979). *Biometrika* V 66: p.188.

Feigl P & Zelen M (1965). *Biometrics* 21: p.826

Hinkley DV (1970). Inference about the change-point in a sequence of random variables, *Biometrika* 57 (1): p.1–17.

Lawless FJ (2003). Statistical Models and Methods for Lifetime Data, Wiley, New York.

Murphy and Van der Vaart, 2000, On profile likelihood (with discussion). *Journal of the American Statistical Association* 95: p.449–485.