University of Illinois Department of Economics
Spring 2017 Econ 574 Roger Koenker

Lecture 22
"Survival Analysis: An Introduction"

There is considerable interest among economists in models of durations, which we often characterize as survival times preceding an event. These models originated for the most part in biostatistics and quality control where the "event" was relapse of an illness, death, or failure of a product component. In economics the event is sometimes more upbeat: ending a spell of unemployment by finding a job for example, or release from a stay in the hospital. But it may be more conventionally the end of something positive as well: to cite two examples of current empirical research in our own department, consider models for the length of professional sports careers, and models for survival times of banks.

Such models may be thought of as conventional statistical models for positive random variables, but they have a number of common features and have developed a number of specialized concepts and techniques which I will try to introduce gradually.

1. Survival functions and hazard rates

   I like to begin by thinking optimistically about births rather than deaths, so in this spirit consider a positive random variable, T, representing the time, since conception say, of birth. This random variable may be characterized by its distribution function $F(t)$, or by its density function, which might look something like this.

Figure 1: Unconditional density of human gestational age at birth.

The survival function which we might think of as the duration of pregnancy distribution of T is simply

$$S(t) = 1 - F(t) = P(T > t)$$

and the hazard function is

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

One way to think about this is to consider the question: given that you have not given birth by time t, what is the probability that you give birth before time $t + s$? Write

$$P(T < t + s | T > t) = \frac{P(t < T < t + s)}{P(T > t)} = \frac{F(t + s) - F(t)}{1 - F(t)}$$

to get a rate per unit time, we should compute

$$\lim_{s\to 0} \frac{s^{-1}(F(t+s) - F(t))}{1 - F(t)} = \frac{f(t)}{1 - F(t)}$$

The hazard rate provides an interesting alternative way to characterize the distribution of T. As a simple example consider the exponential model where $f(t) = \lambda e^{-\lambda t}, S(t) = e^{-\lambda t}$ so $\lambda(t) = \lambda$, constant. The idea that the exponential model has constant hazard is fundamental to the subject. It is as if T can't remember what amount of time has already passed and the probability of an event in the next unit of time, given no event up to the current time is constant.

The exponential model is clearly not appropriate for many economic processes where there is usually positive or negative aging. These terms come from the typical shape of the human mortality hazard function which is somewhat U-shaped, declining over a short range for infants and then gradually increasing for old adults. The Weibull model represents a convenient generalization of the exponential model which accommodates either increasing or decreasing hazard, but not both. In the Weibull

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

so depending o whether $\alpha \gtrless 1$ we get increasing or decreasing hazard.

Another simple example is the Rayleigh distribution which we encountered in Problem 3.1, where

$$\lambda(t) = \lambda_0 + \lambda_1 t$$

The uniform model is also very simple with

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

so the hazard is rapidly increasing as one approaches the upper limit.

2. Estimation and Censoring in Parametric Models

If we observe a random sample $\{t_1, \ldots, t_n\}$ of T_i's, we can easily estimate a parametric model $f(t, \theta)$ for the survival time by maximizing the loglikelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log f(t_i, \theta)$$

However, it is almost inevitable that survival time data is marred by some censored observations, observations for which we know that the event didn't occur up to some time t, but we do not know exactly when it did occur. This may be because the individual disappeared from the sampling framework for some reason, or because at the analysis date some individuals were still "surviving", or some other reason. Such observations are easily accommodated into the parametric mle framework. Let c_i be a censoring time for each observation and suppose that data takes the form of pairs (t_i, δ_i) where $\delta_i = I(t_i < c_i)$,

so δ_i will be called the censoring indicator with $\delta_i = 1$ indicating that t_i is an actual survival time and $\delta_i = 0$ indicating that t_i is censored so we know only that $T_i > t_i$. The loglikelihood for this censored survival time data may be written as

$$\ell(\theta) = \sum_{i=1}^{n} \delta_i \log f(t_i, \theta) + (1 - \delta_i) \log S(t_i, \theta)$$

Unfortunately, while these parametric models for survival times are very convenient and relatively easy to estimate, it is often difficult to be confident about a particular parametric specification of f and S. Note that if there are covariates we may consider making one or more of the θ parameters above dependent upon covariates.

3. In this Section we introduce a crucial tool of non-parametric survival analysis, the Kaplan-Meier estimator which may be thought of as an analogue of the empirical distribution function except that a.) it estimates the survival function $S(t)$ rather that $F(t)$, and b.) it accounts for censoring.

   To motivate the Kaplan-Meier estimate let's begin by considering a simpler context in which we observe an uncensored random sample of survival times $\{t_i, \ldots, t_n\}$. Suppose we chop the real half line into intervals Now write,

$$\begin{aligned} S(\tau_k) &= P(T > \tau_k) \\ &= P(T > \tau_1)P(T > t_2 | T > t_1) \ldots P(T > \tau_k | T > \tau_{k-1}) \\ &\equiv p_1 \cdot p_2 \ldots p_k. \end{aligned}$$

   As an estimate of p_i it is natural to use,

$$\hat{p} = \left(1 - \frac{d_i}{n_i}\right)$$

   where d_i is the number "dying" period $(\tau_{i-1}, \tau_i]$ and n_i is the number surviving to the beginning of this period. Then, our estimate of the survival function would be

$$\hat{S}(t) = \prod_{\{j:\tau_j < t\}} \hat{p}_j$$

   For the Kaplan-Meier estimator we make two minor modifications. The first is that rather than use arbitrary intervals delimited by τ_i we use the observed t_i themselves as the τ_i. This is just like the usual strategy for the empirical distribution function. To deal with the censored observations we simply let

$$d_i = \delta_i$$

   the number "dying" in period (t_{i-1}, t_i) is either 0, if t_i is censored or 1 if t_i is uncensored. Then,

$$\hat{p}_i = 1 - d_i/n_i$$

as above, and denoting the ordered t_i by $t_{(i)}$,

$$\begin{aligned}
\hat{S}(t) &= \prod_{t_{(i)} \le t} \hat{p}_{(i)} \\
&= \prod_{t_{(i)} \le t} \left(1 - \frac{1}{n_i}\right)^{\delta_{(i)}} \\
&= \prod_{t_{(i)} \le t} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}} \\
&= \prod_{t_{(i)} \le t} \left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}}
\end{aligned}$$

where $\delta_{(i)}$ is the censoring indicator of the observation $t_{(i)} + \varepsilon$. This estimator satisfies several important requirements:

(a) It is consistent for $S(t)$,

(b) It is asymptotically normal, i.e., $\sqrt{n}(\hat{S}(t) - S(t))$ converges to a Gaussian process,

(c) In the absence of censoring it is the same as the empirical distribution function,

(d) It is a generalized mle in the sense of Kiefer and Wolfowitz.

In the figure below we illustrate the Kaplan-Meier estimator for a sample of 5 observations, of which only the second smallest, t_z, is censored. The indicated p_i's in the figure are the conditional probabilities while the vertical scale represents the $\hat{S}(t)$ unconditional survival probabilities. The Kaplan Meier estimator is particularly good in situations in which we have a small number of groups and we would like to ask: do they have similar survival distributions. An example of this sort of question is addressed in the next figure. Using data from Meyer (1990) we consider the survival distributions estimated by the Kaplan-Meier technique for individuals who have more than \$100 per week in unemployment benefits versus those with less than \$100 per week in benefits.

As the figure indicates, those with higher benefits appear to stay unemployed longer. The median unemployment spell for the high benefits group is roughly 2 weeks longer than for the low benefits group. Note, however, that the difference is unclear in the right tail of the distribution; the higher benefit group appears to have a somewhat lower probability of a spell greater than 35 weeks. This plot was produced ty the R command,

```
plot (survfit(Surv(dur,cens),strata=exp(ben) > 100, type='kaplan-meier'))
```

where `dur` is the observed durations, `cens` is the censoring indicator, and `ben` is the log of weekly benefits.

This technique is very useful in situations where we have randomized assignment into treatment groups, but often there are other covariates which need to be accounted for and this cannot be adequately accomplished by looking at Kaplan-Meier plots. We would like to have some sort of compromise between the parametric approach introduced at the beginning of the lecture and the non-parametric approach of Kaplan-Meier.

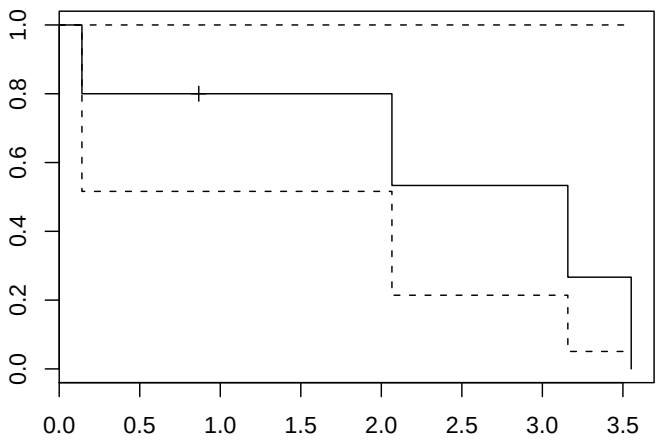

Figure 2: A Kaplan Meier Survival Curve for a Trivial Problem with $n = 5$: Note that only the second survival time is censored. The dotted lines constitute a confidence band which is rather uniformative in this case.

4. Semi-Parametric Methods

   As above, let $\{t_i, \delta_i : \quad i - 1, \ldots, n\}$ denote a random sample from the model

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

   and let $\{x_i\}$ denote a vector of covariates associated with observation i. Recall

$$\lambda(t|x) = \frac{f(t|x)}{1 - F(t|x)}$$

   The proportional hazard model of Cox (1972) takes the form,

$$\lambda(t|x) = e^{x'\beta}\lambda_0(t)$$

   where we will call $\lambda_0(t)$ the baseline hazard since it corresponds to the special case

$$\lambda(t|0) = \lambda_0(t)$$

Definition: A family of df's $\mathcal{F}$ constitutes a family of Lehmann alternatives if there exists a $F_0 \in \mathcal{F}$ such that for any $F \in \mathcal{F}, 1 - F(x) = (1 - F_0(x))^\gamma$ for some $\gamma > 0$, and all $x \in \Re_i$.

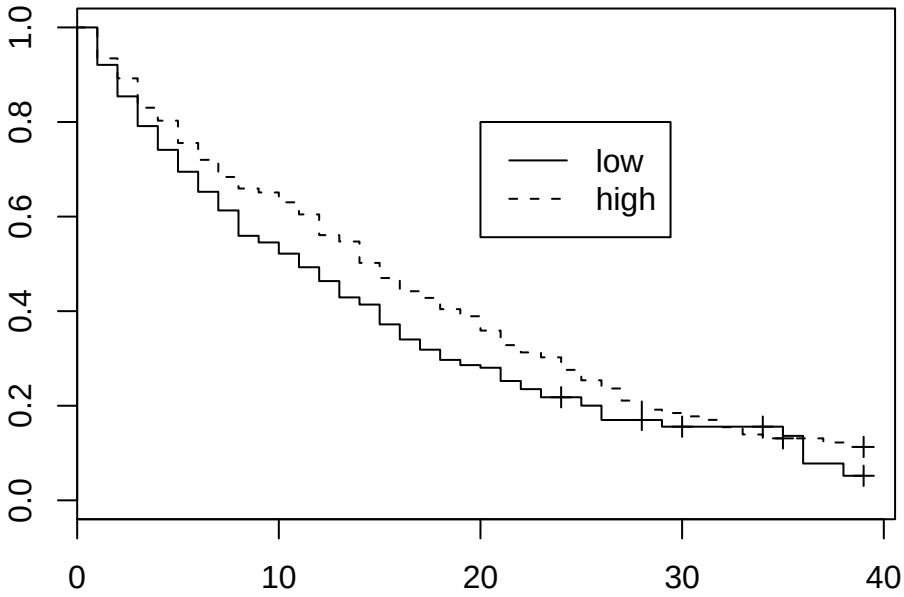

Figure 3: Two Kaplan Meier Survival Curves for the length of unemployment spells: The figure plots Kaplan-Meier estimates of the duration of unemployment function for two groups of individuals. One is a high UI-benefits group (those with weekly benefits more than 100 dollars, the other with weekly UI benefits less than 100. The data is taken from Meyer(1990).

The Cox model constitutes an example of a family of Lehmann alternatives. Since,

$$\begin{aligned} S(t|x) &= \exp\{-\int_0^t \lambda(u|x)du\} \\ &= \exp\{-e^{x'\beta}\int_0^t \lambda_0(u)du\} \\ &= (S_0(t))^\gamma \qquad \text{for } \gamma \equiv e^{x'\beta}. \end{aligned}$$

In the important special case of two samples $x'\beta$ takes only two values, say 0 and 1, and

$$S_1(t) = S_0^\gamma(t)$$

for some γ.

There is a large literature on estimating the Cox model and on the asymptotic theory of the resulting estimators. I will only sketch the basic ideas. Let $\mathcal{R}_i$ denote the index set of observations "at risk" at time $t_{(i)} - \varepsilon$, i.e., the index set surviving at time $t_{(i)} - \varepsilon$. The probability in any subsequent interval can be approximated as,

$$P[\underline{\text{one}} \text{ " death" in } [t_{(i)}, t_{(i)} + dt)|\mathcal{R}_i] \cong \sum_{j\in\mathcal{R}_{(i)}} e^{x'_j\beta}\lambda_0(t_{(i)})dt$$

and

$$P[\text{ "death" of individual } (i) \text{ at } t_{(i)}|\underline{\text{one}} \text{ "death" at time } t_{(i)}] = \frac{e^{x_{(i)}\beta}}{\sum_{j\in\mathcal{R}_(i)} e^{x'_j\beta}}$$

and this gives the "partial likelihood"

$$\mathcal{L}^*(\beta) = \prod_{i=t}^n e^{x'_{(i)}\beta} / \sum_{j\in\mathcal{R}_{(i)}} e^{x'_j\beta}$$

which can be maximized "as if" it is the full likelihood. What is missing? To answer this question recall that we could write the full likelihood in the parametric censored survival model as,

$$\mathcal{L}(\theta) = \prod_{i=1}^n f(t_i|x_i,\theta)^{\delta_i} S(t_i|x_i,\theta)^{1-\delta_i}$$

where $\theta = (\beta, \lambda_0)$ in the notation of the Cox model. Since λ_0 is an unknown *function*, we sometimes call this sort of model *semiparametric*.

To express the likelihood in terms of the hazard function as formulated by Cox note that the cumulative hazard function

$$\Lambda(t|x) = \int_0^t \lambda(s|x)ds = -\log(1 - F(t|x)).$$

Thus in the Cox model where

$$\lambda(s|x) = \lambda_0(s)\exp(x'\beta)$$

we may write

$$\Lambda(t|x) = \Lambda_0(t)\exp(x'\beta)$$

which is equivalent to

$$\log(-\log(S(t|x))) = x'\beta + \log\Lambda_0(t)$$

or

$$\log S(t|x) = -e^{x'\beta}\Lambda_0(t)$$

or

$$S(t|x) = \exp\{-e^{x'\beta}\Lambda_0(t)\}.$$

So the likelihood may be written,

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_{i=1}^{n}(\lambda(t_i|x_i)S(t_i|x_i))^{\delta_i}S(t_i|x_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n}(e^{x_i'\beta}\lambda_0(t_i)S(t_i|x_i))^{\delta_i}S(t_i|x_i)^{1-\delta_i}\end{aligned}$$

Comparing this with the partial likelihood of Cox we see that we can write

$$\mathcal{L}(\theta) = \mathcal{L}^*(\beta)\mathcal{L}^{(*)}(\beta,\lambda_0)$$

where $\mathcal{L}^*(\beta)$ is the Cox factor given earlier, and

$$\mathcal{L}^{(*)}(\beta,\lambda_0) = \prod_{i=1}^{n}(\sum_{j\in\mathcal{R}_{(i)}} e^{x_j'\beta}\lambda(t_i)) \times \exp\{-\int_0^\infty (\sum_{j\in\mathcal{R}_{(i)}} e^{x_j'\beta})\lambda_0(t)dt\}$$

Note that, in effect, Cox has factored the full likelihood into a part that depends solely on β, and another part which depends on both β_0 and λ_0. Normally this would be no great accomplishment and a proposal to ignore the second factor would be considered a bad joke. Since it contains β, it presumably contains sample information about β. Cox (1972), who is on a different plane of consciousness than the rest of us, argued heuristically that if λ_0 is left unspecified, then the second factor would contain little relevant information about the parameter β. This conjecture has been subsequently supported in a number of examples, and by asymptotic efficiency computations.

A more modern approach to the Cox profile likelihood is provided by work by Murphy and van der Vaart (2000). They write the PH likelihood as

$$L(\beta,\Lambda) = \prod \exp\{-e^{z_i'\beta}\Lambda(t_i)\}e^{z_i'\beta}\lambda(t_i)$$

and taking logs we have

$$\ell(\beta,\lambda_1,\cdots\lambda_n) = \sum z_i'\beta + \log\lambda(t_i) - e^{z_i'\beta}\sum_{j:t_j\le t_i}\lambda(t_j)$$

Now differentiating we obtain,

$$\frac{\partial\ell}{\partial\lambda_k} = \frac{1}{\lambda_k} - \sum_{i:t_i\le t_k} e^{z_i'\beta} = 0$$

so solving and substituting back into the likelihood we have the Cox profile likelihood,

$$\tilde{L}(\beta) = \prod \frac{e^{z_i'\beta}}{\sum_{j:t_j \leq t_i} e^{z_i'\beta}} e^{-1}$$

Here we have ignored censoring, but it can be also accomodated in this same framework.

Finally, we should briefly discuss how to estimate the baseline hazard function $\lambda_0(t)$ in the Cox model. I will briefly describe two approaches; one due to Breslow, the other to Tsiatis.

In the former case, we assume that $\lambda_0(t)$ is constant between uncensored observations and let

$$\hat{S}_0(t) = \prod_{t_{(i)} < t} \left(1 - \frac{\delta_{(i)}}{\sum_{j \in \mathcal{R}_{(i)}} e^{x_j'\beta}} \right)$$

Note here contrary to the Cox model,

$$\hat{S}_0(t) \neq \exp\{-\Lambda_0(t)\}$$

but this approach has the virtue that it simplifies to the Kaplan-Meier estimator when there is no covariate effect since for $\beta = 0$, we have $\sum_{j \in \mathcal{R}_{(i)}} e^{x'\beta} = \#\mathcal{R}_{(i)}$.

Tsiatis uses instead

$$\hat{S}_0(t) = \exp\{-\hat{\Lambda}_0(t)\}$$

where

$$\hat{\Lambda}_0(t) = \sum_{t_{(i)} < t} \left(\frac{\delta_{(i)}}{\sum_{j \in \mathcal{R}_{(i)}} e^{x_j'\hat{\beta}}} \right)$$

the relationship between the two estimators is simply accounted for by the familiar approximation

$$-\log(1 - x) \approx x$$

for small x. See Kalbfleisch and Prentice (1980) Section 4.3 for further details.

Computing for survival analysis models is now quite reasonable in Limdep and Stata, but the software designed by Terry Therneau for R is somewhat more flexible than either of the previous options.

References

Kalbfleisch, J.D. and R.L. Prentice (1980). *The Statistical Analysis of Failure Time Data*, Wiley.

Meyer, B. (1990). Unemployment insurance and unemployment spells, *Econometrica*, 57, 757-782.

Miller, R.G. (1981). *Survival Analysis*, Wiley.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data*, Cambridge U. Press.

Cox, D.R. (1972). Regression models and life tables, *JRSS(B)*, 34, 187-200.

Murphy, S. and A. van der Vaart (2000) On Profile Likelihood, *JASA*, 95, 449-465.