Intro to Duration

or, what's my survivor function!?

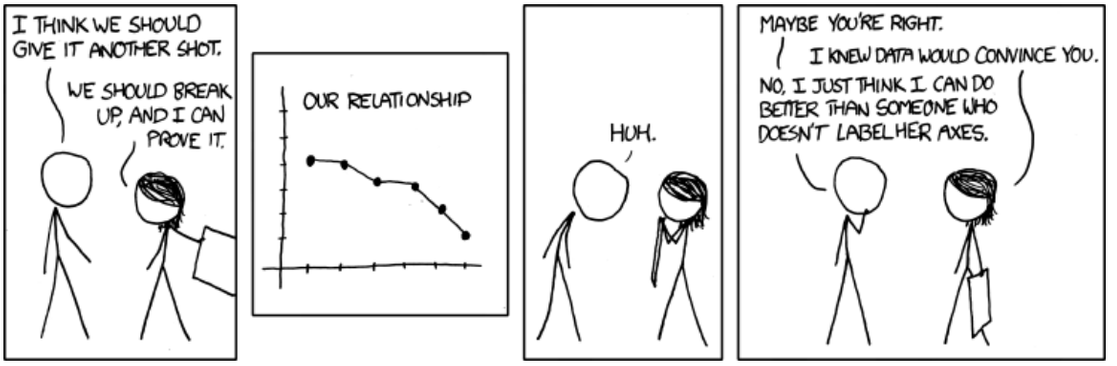
Michele Claibourn, StatLab

September 25, 2013

Research Data Services in the Library

- Research Data Services: `www.library.virginia.edu/services/`
  - Data management plans
  - GIS training and consultations
  - Locating data, archiving data
- StatLab Services: `statlab.library.virginia.edu`
  - Individual consulting: advice, training or feedback on quantitative research
  - Workshops: Fall schedule is up
- Upcoming Events

Survival/Duration/Event History



Duration Analysis 4/21

Michele Claibourn, StatLab

Research Data Services @ UVa Library

Introduction

Examples

Kaplan-Meier Estimation

Cox PH Model

Learning More

Duration, huh, what is it good for?

Analyzing the timing of events:

- Criminal recidivism
- Graduation from school
- Unemployment
- War (or peace)
- Failure of a mechanical system
- Death of a biological organism

Terminology varies across disciplines (but we can still get along):

- Biostatistics, epidemiology: survival analysis
- Engineering: failure-time analysis
- Sociology: event-history analysis
- Economics: duration analysis

In general, any issue in which the phenomenon of interest is a duration until the occurrence of an event falls into this class of models.

Duration, yeah, what is it good for?

Characteristics of survival time data:

- Value must be non-negative, so the distribution is (usually) positively skewed
- Observation of event for some units is censored (event doesn't occur for some observations during study period). Ignoring censoring can produce bias

What's wrong with using a

- Linear model?
- Count model?
- Logit model?

We'll consider methods for estimating unconditional survival distributinons (Kaplan Meier) and methods that model the relationship between survival and covariates (Cox PH).

Terminology and Quantitities of Interest

Let T =duration until event occurs, C_i =censoring point or duration until the observation is censored, $T_i = minf\{Y_i, C_i\}$, $d_i = 0$ if observation i is censored, 1 if it is not.

1. The cumulative probability of event

$$F(t) = \int_0^t f(u)du = P(T \leq t)$$

   characterizes the distribution of failure times

2. For all points for which $F(t)$ is differentiable, the probability density function is

$$f(t) = lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t)}{\Delta t}$$

   provides the unconditional failure rate

More Quantities of Interest

3. The probability of survival to time t

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

4. The probability of having an event at time t, given that we have not had an event prior to t, that is, $Pr(Y_i = t | Y_i \geq t)$. By the rule for conditional probability, this is

$$h(t) = \frac{f(t)}{S(t)} = lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}$$

The hazard rate, a conditional failure rate.

5. Integrated hazard rate

$$H(t) \equiv \Lambda(t) = \int_0^t h(t)dt$$

can also be written as

$$H(t) = -\ln[S(t)]$$

Duration, good god, what is it good for?

The hazard rate, survivor function, distribution and density functions are all mathematically linked. If any of these are specified, then the others are fully determined. Relations among them can be summarized again

$$\begin{aligned} f(t) &= S(t)h(t) \\ h(t) &= -\frac{\partial \ln S(t)}{\partial t} \end{aligned}$$

Descriptively, we're often most interested in the survival function; causally, we're often more focused on the hazard rate.

Censoring

Censoring is the removal of data for reasons other than the event of interest

- Right-censoring: period of observation ends before the event occurs
- Left-censoring: initial time at risk is unknown, event of interest has already occurred before observation began (rarer)

Censoring complicates the likelihood function, and hence the estimation, of survival models.

- Censoring must be noninformative, (conditionally) independent of the future value of the hazard for the observation
- Censored observations still provide information – about survival to time T (uncensored observations provide information about survival prior to event AND about the hazard of the event)

Example Data: Coalition Failure

Source: King, Gary, James E. Alt, Nancy E. Burns, and Michael Laver. 1990. "A Unified Model of Cabinet Dissolution in Parliamentary Democracies." *American Journal of Political Science* 34: 846-71.

- durat: number of months before a cabinet falls
- censor12: government endures to at least 12 months prior to constitutionally mandated election period
- majority: cabinet is a majority
- oppconc: proportion of legislators to the left of a right-leaning government or to the right of a left-leaning government
- crisis: number of days of crisis before a government formed
- format: number of attempts to form government during the crisis
- frac: characterizes number and size of parties in parliament
- polar: measure of support for extremist parties
- volat: measure of electoral turnover in mass voting in parliamentary elections
- invest: existence of a legal requirement for legislative investiture

Available at: `http://gking.harvard.edu/data`

Example Data: Recidivism

Source: Rossi, Peter H., Richard A. Berk and Kenneth J. Lenihan. 1980. *Money, Work and Crime: Some Experimental Results.* New York: Academic Press.

- week: week of first arrest after release, or censoring time
- arrest: the event indicator, equal to 1 for those arrested during the period of the study and 0 for those who were not arrested
- fin: a dummy variable, equal to 1 if the individual received financial aid after release from prison, and 0 if he did not; financial aid was a randomly assigned factor manipulated by the researchers
- age: in years at the time of release
- race: a dummy variable coded 1 for blacks and 0 for others
- wexp: a dummy variable coded 1 if the individual had full-time work experience prior to incarceration and 0 if he did not
- mar: a dummy variable coded 1 if the individual was married at the time of release and 0 if he was not
- paro: a dummy variable coded 1 if the individual was released on parole and 0 if he was not
- prio: number of prior convictions

Available at:

`http://socserv.mcmaster.ca/jfox/Books/Companion/data/Rossi.txt`

Example Data: AIDS Clinical Trial

Source: Hosmer, David W., Stanley Lemeshow, and Susanne May. 2008. *Applied Survival Analysis*, 2nd. ed. Wiley.

- time: time to AIDS diagnosis or death in days
- censor: event indicator for AIDS diagnosis or death (1=event occured, 0=otherwise)
- tx: treatment indicator, treatment includes IDV
- sex: indicator for sex, 1=male, 2=female
- cd4: Baseline CD4 count, cells/milimeter
- priorzdv: monthos of prior ZDV use
- age: age at enrollment

Available at: `http://www.umass.edu/statdata/statdata/stat-survival.html`

Nonparametric Estimation of $S(t)$

The Kaplan-Meier estimate of the survivor function

$$\hat{S(t_k)} = \prod_{t \leq t_k} \frac{n_t - d_t}{n_t}$$

- $S(t)$, the survival function, or the proportion of units surviving beyond t
- n_t, the number of observations 'at risk' for the event at time t
- d_t, the number of observations which experience the event at time t

Kaplan-Meier Plots

Time, t	# at Risk, n_t	# Failed, d_t	# Censored	p	$\hat{S(t)}$
2	6	1	0	5/6	5/6
4	5	2	0	3/5	1/2
5	3	0	1	1	1/2
7	2	1	0	1/2	1/4
8	1	0	1	1	1/4

Kaplan-Meier Plots

KM plots

- Unconditional probability that an observation will survive beyond time t
- Can add uncertainty intervals, e.g., the "Greenwood" variance estimates

$$var[\hat{S}_{t_k}] = [S(t_k)]^2 \sum_{t \leq t_k} \frac{d_t}{n_t(n_t - d_t)}$$

Can also plot

- The cumulative hazard function (aka Nelson-Aalen estimator)

$$H(t_k) = \sum_{t \leq t_k} \frac{d_t}{n_t}$$

- The hazard rate, $h(t)$, the probability of having an event at t, given the event has not occurred prior to t (the derivative of $H(t)$)
- By categorical variables

Cox: A semi-parametric model

Semi-parametric vs. Parametric:
A **parametric** model based on the exponential distribution may be written as

$$h_i(t) = exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

i subscripts observations, x's are covariates. α represents a baseline hazard; $h_i(t) = e^{\alpha}$ when all $x = 0$. The exponential model assumes the baseline hazard follows an exponential distribution.
The **semi-parametric** Cox model leaves the distribution of the baseline hazard unspecified

$$h_i(t) = h_0(t)exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

Thoug the covariates still enter the model linearly.

Cox: A proportional hazards model

Assumes there is some baseline hazard $h_0(t)$; we are interested in how a set of covariates moves this baseline hazard up or down.

- Because the hazard must remain positive, the covariates enter through the exponential. The hazard rate for i is

$$h_i(t) = h_0(t)e^{\beta' \mathbf{X}}$$

- The baseline haard corresponds to the case where $X = 0$
- What is the effect of a dichotomous X on the baseline hazard (risk)?

$$\frac{h_1(t)}{h_0(t)} = e^{(X_1 - X_0)\hat{\beta}} = e^{\hat{\beta}}$$

  $e^{\hat{\beta}}$ is the risk for observations with $X = 1$ relative to observations with $X = 0$
- The hazard ratio is independent of time t

Estimating the Cox Model

We're interested in the conditional probability of a failure at time t_i given the number of cases that are at risk of failing at time t_i.

$$Pr(t_j = T_i | R(t_i)) = \frac{h(t_i)}{\sum_{j \in R(t_i)} h(t_j)}$$

The numerator is the hazard for individual i at time t_i; the denominator is the sum of all hazards for all i at risk. Substituting in the hazard function for the Cox model

$$\frac{h_0(t_i)e^{\beta' \mathbf{x}_i}}{\sum_{j \in R(t_i)} h_0(t_j)e^{\beta' \mathbf{x}_j}} = \frac{e^{\beta' \mathbf{x}_i}}{\sum_{j \in R(t_i)} e^{\beta' \mathbf{x}_j}}$$

The (partial) likelihood is the product of the probabilities

$$PL = \prod_{i=1}^{K} \left[\frac{e^{\beta' \mathbf{x}_i}}{\sum_{j \in R(t_i)} e^{\beta' \mathbf{x}_j}} \right]^{\delta_i}$$

For convenience, we work with the *log* partial likelihood

$$\ln PL = \sum_{i=1}^{K} \delta_i \left[\beta' \mathbf{x}_i - \ln \left(\sum_{j \in R(t_i)} e^{\beta' \mathbf{x}_j} \right) \right]$$

And maximize with respect to β.

Ties: sequencing failures

But wait! What about ties (units that experience the event at the same time)? Ties complicate the determination of the risk set at each failure time and the sequencing of event occurrences. In response, the partial likelihood must be approximated.

- Breslow Method: group all the tied events together.
- Efron Method: adjusts the risk sets using probability weights.
- Exact Partial Likelihood: averages over all possible $d!$ orderings of event times.
- Exact Discrete Partial Likelihood: assumes events really do occur at the same time, time is discrete not continuous.

Baseline Hazard Function

Though we didn't specify a distribution for the hazard function, we can still get estimates of the baseline hazard (or survivor) function. Because we have estimates of $e^{\beta' \mathbf{X}}$ from the model, and because we know $S(t)$ (from the Kaplan-Meier plot), we can get estimates of $H_0(t)$, $S_0(t)$, and $h_0(t)$. For example,

$$S(t) = [S_0(t)]^{exp(\beta' \mathbf{X})}$$

and

$$H(t) = \int_0^t h(\tau) d\tau = e^{\beta' \mathbf{X}} \int_0^t h_0(\tau) d\tau = e^{\beta' \mathbf{X}} H_0(t)$$

Remember, baseline functions assume the covariates are all 0.

Baseline estimates from the Cox model often smoothed.

Extensions of the Model

Additional issues and extensions that we don't have time to cover include

- Time-varying covariates (generally requires adding rows to the data)
- Unobserved heterogeneity, or unaccounted for difference in the hazards, or spurious duration dependence (better model specifications, frailty models, cure or split-population models)
- Competing risks, multiple events could end a spell, or units could fail in mutiple ways (stratified Cox, latent survivor time, multinomial logit)
- Repeated events, or units could experience the event multiple times (variance-correction models, e.g., conditional gap time)
- Discrete time duration, or events are observed only at discrete times (binary TSCS models, e.g., logit with time dependence, or Markov transition models)

Suggested References

General references

- David W. Hosmer, Stanley Lemeshow, and Susanne May. 2008. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*, 2nd ed. Wiley. Biostatistics focus
- Janet M. Box-Steffensmeier and Bradford S Jones. 2004. *Event History Modelign: A Guide for Social Scientists*. Cambridge. Social sicence focus.

Software references

- Mario Cleves, William W. Gould, Roberto G. Guiterrez, and Yulia Marchenko. 2008. *An Introduction to Survival Analysis Using Stata*, 2nd. ed. Stata Press
- Paul D. Allison. 2010. *Survival Analysis Using SAS: A Practical Guide*, 2nd ed. SAS Press.