Survival Analysis

Brian Cocolicchio- July 25, 2016

What is Survival Analysis?

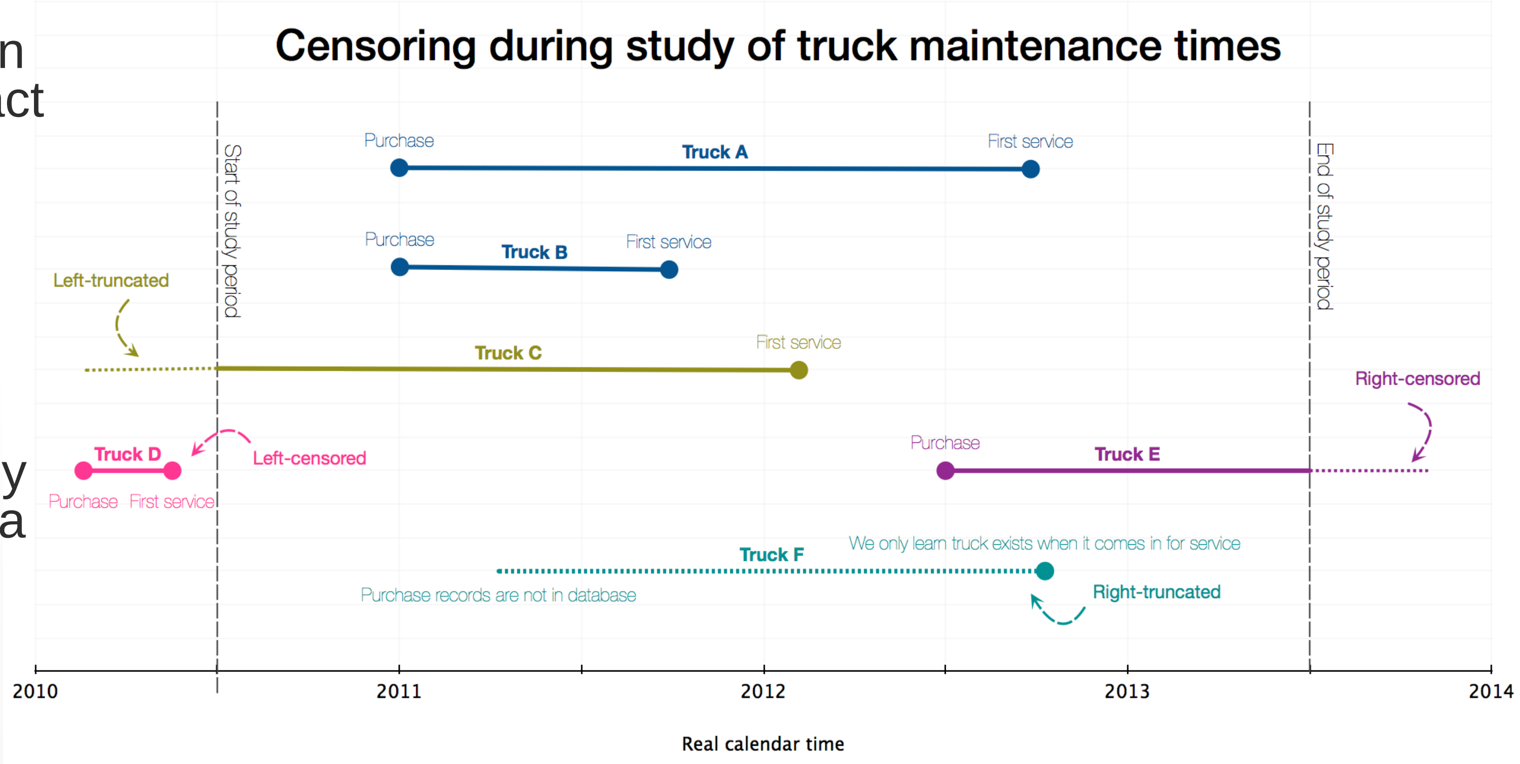
- Survival analysis, also known as event analysis, or failure analysis is used to analyze data where we are interested in the **time until the event.**
- Used in many fields including: Health Care, Engineering, Sociology, Business
- Examples:
  - Time until part failure
  - Incubation time of some diesases (HIV, SARS, Hepatitis, etc.)
  - Employee Tenure
  - Customer Churn

Why Not Use Regression For This?

- Time to event is not always completely observable.
  - Censoring
  - Truncation
- Time ≥ 0, skewed distribution.
- The probability of surviving past a point in time may be more important than the expected time of the event.
- The hazard function may be more insightful in determining the failure mechanism.

What Is Meant by Censoring and Truncation?

- Censoring-When we have some information about a subjects event time but not the exact event time
  - Right
  - Left
  - Interval
- Truncation-Due to sampling bias in that only those individuals whose lifetimes lie within a certain interval can be observed.
  - Right
  - Left



What Are The Parameters of Interest?

- If X is the time to the event and X is a random variable:
- Event function-Probability of an event occurring by time x.

$$F(x) = P(X \leq x) = \int_0^x f(s)ds$$

- Survival function-Probability that an event will not occur by time x.

$$S(x) = Pr(X > x) = 1 - F(x)$$

- Hazard Rate-The instantaneous rate at which an event can occur given no previous events.

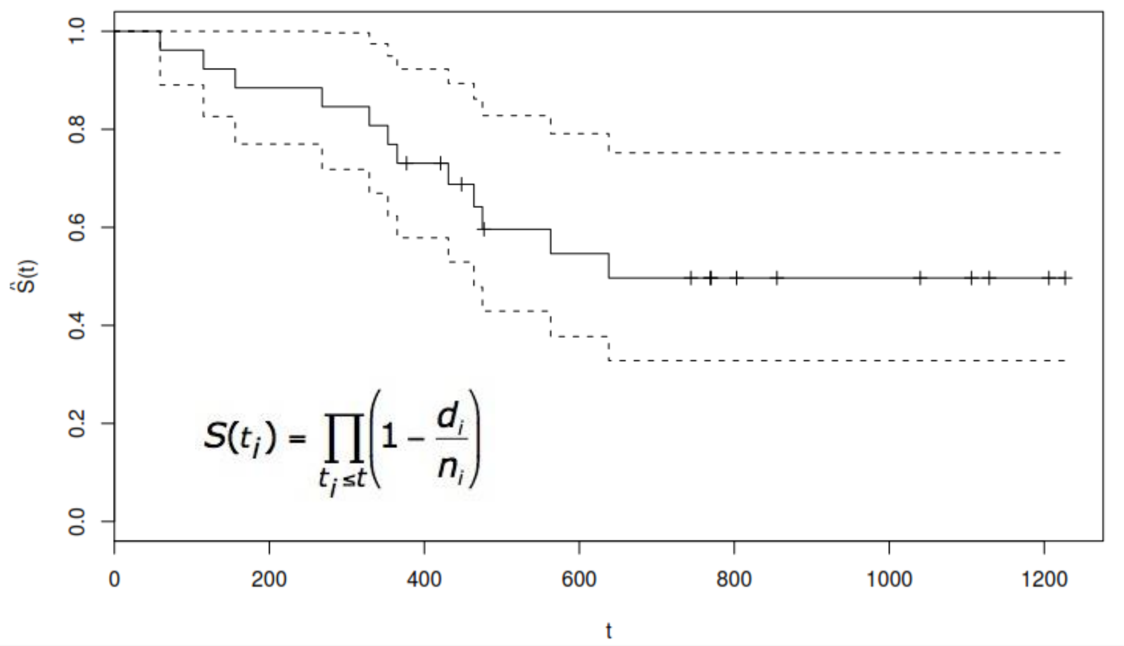
$$h(x) = \lim_{\Delta x \to 0} \frac{Pr[x \leq X \leq x + \Delta x | X \geq x]}{\Delta x} = \frac{f(x)}{S(x)} = -\frac{d \ln[S(x)]}{dx}$$

How Do We Estimate the Survival and Hazard Functions?

- Non-Parametric: Kaplan-Meier Estimate

$$S_t = \frac{\text{Number of subjects living at the start} - \text{Number of subjects died}}{\text{Number of subjects living at the start}}$$

t	No. subjects at risk	Deaths	Censored	Cumulative survival
59	26	1	0	$25/26 = 0.962$
115	25	1	0	$24/25 \times 0.962 = 0.923$
156	24	1	0	$23/24 \times 0.923 = 0.885$
268	23	1	0	$22/23 \times 0.885 = 0.846$
329	22	1	0	$21/23 \times 0.846 = 0.808$
353	21	1	0	$20/21 \times 0.808 = 0.769$
365	20	0	1	$20/20 \times 0.769 = 0.769$
377	19	0	1	$19/19 \times 0.769 = 0.769$
421	18	0	1	$18/18 \times 0.769 = 0.769$
431	17	1	0	$16/17 \times 0.769 = 0.688$



How Do We Estimate the Survival and Hazard Functions?

- Semi-Parametric: Cox (Proportional Hazards) Regression

$h_i(t) = h_0(t) * \exp\{\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}\}$ Do not have to specify $h_0(t)$ here.

- Parametric: Assume that the time to event follows a known distribution (Weibull, exponential, log-normal,etc.)
  - Then use maximum likelihood estimation

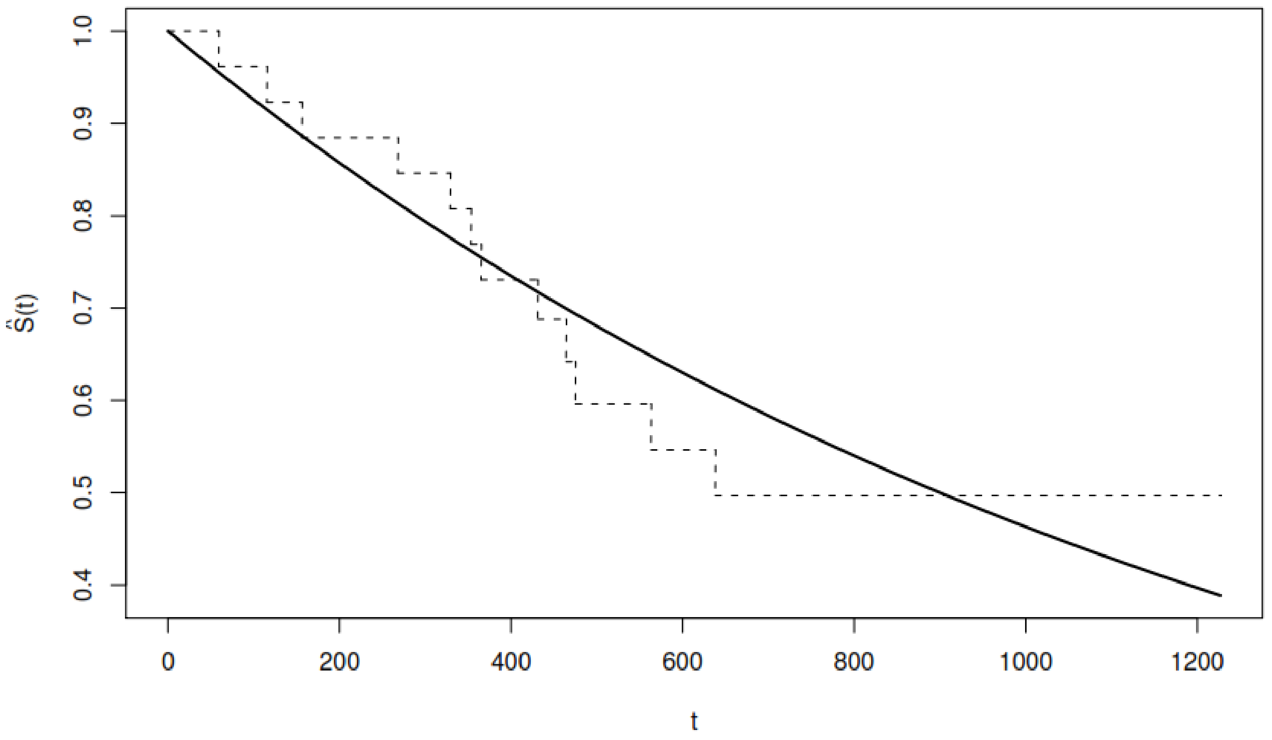
$$\log L = \sum_{i:\delta_i=1}^{n} \log(h(Y_i)) - \sum_{i=1}^{n} H(Y_i).$$

  - to calculate the parameters for :

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

How Do We Estimate the Survival and Hazard Functions?

```
plot(T,1-pexp(T,exp(-7.169)),xlab="t",ylab=expression(hat(S)*"(t)"))
```



How Do We Conduct a Survival Analysis in Python?

- Lifelines Package
- PyIMSL

Thank You!