Quantitative Research Methods IV - 17.806

Recitation, Week 9.
Topic: Survival Analysis II.

Benjamín Muñoz

April 14, 2023

MIT

Massachusetts Institute of Technology

Table of contents

1/ Survival Analysis

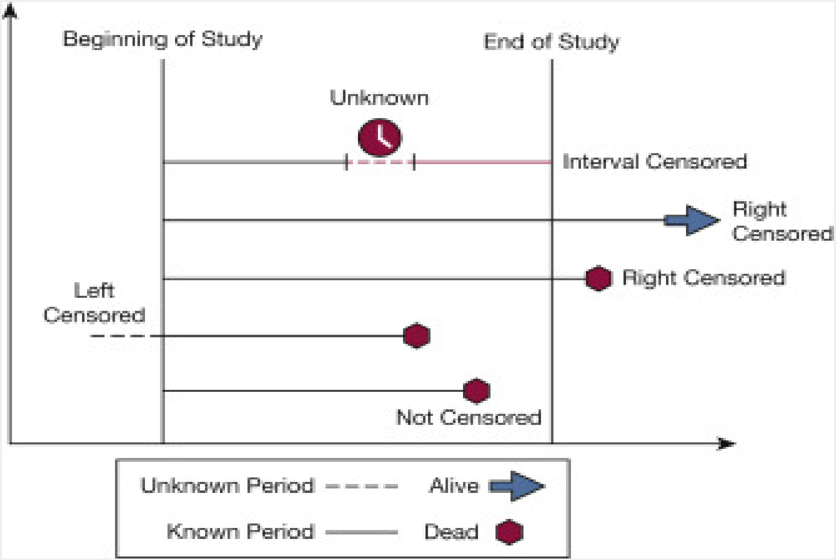
Survival Outcomes

- Particular type of outcome: **implicit time dimension**. Two justifications:
  - Event which might occur at any time over a lengthy period of follow-up.
  - Length of time spent in a given state (before some event).
- Different names: Survival Analysis, Duration Models, Event History Modeling.
- Basic Components:
  - **Event:** experience of interest.
  - **Time:** period of observation. Based on these two components, we define <u>Survival Time:</u> time until a participant has an event of interest.
- $\rightsquigarrow$ Always non-negative and normally skewed.

Censoring

- The observation occurs in a specific period of time, which does not necessarily coincide with the start of the state (time of origin) and/or the occurrence of the event (time of failure).
- This causes a missing data problem for some units: incomplete information is available about the survival time of some individuals.
- The **Observed Survival Time** does not necessarily coincide with the **True Survival Time**.

Censoring



Beginning of Study
End of Study
Unknown
Interval Censored
Right Censored
Right Censored
Left Censored
Not Censored
Unknown Period
Alive
Known Period
Dead

Key Definitions

- T is a continuous non-negative random variable which denote the time-to-event.
- **Density Function:** $f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}$ (probability of the failure time occurring at exactly time t).
- **Cumulative Distribution Function:** $F(t) = P(T \leq t) = \int f(u)du$ (probability of the failure time occur before or exactly at time t).
- **Survival Function:** $S(t) = P(T > t) = 1 - F(t)$ (probability that the random variable T exceeds the specified time t. Focus on non-failing).
- **Hazard Function:** $h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$ (conditional failure rate: instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t. Focus on failing).

Key Definitions

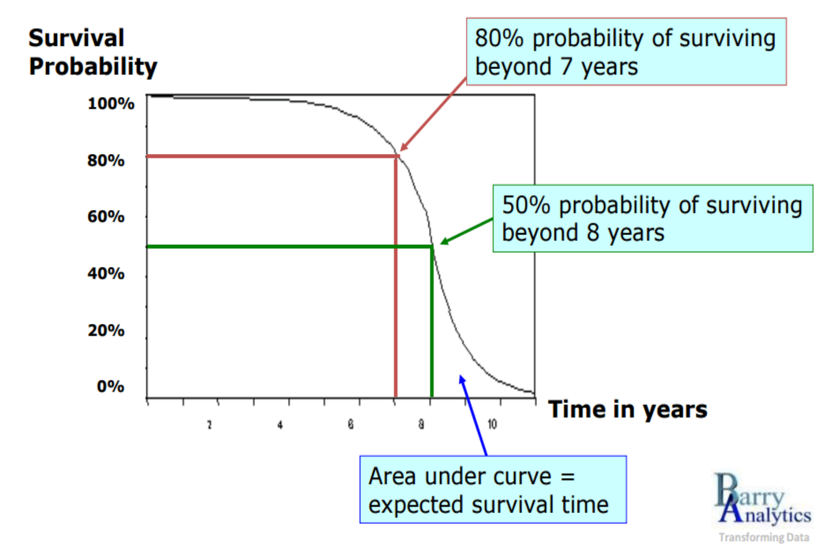
Key Relations for True (Not Estimated/Approximated) Functions

$$P(t \leq T < t + dt | T \geq t) = \frac{P(t \leq T < t + dt, T \geq t)}{P(T \geq t)} = \frac{P(t \leq T < t + dt)}{P(T \leq t)} = \frac{f(t)}{S(t)}$$

$$\frac{\partial}{\partial t} S(t) = 0 - \frac{\partial}{\partial t} F(t) = -f(t) \rightsquigarrow f(t) = -\frac{\partial}{\partial t} S(t)$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{-\frac{\partial}{\partial t} S(t)}{S(t)} = -\frac{d}{dt} log S(t)$$

Survival Function is a Cumulative Function



Survival Function is a Cumulative Function

- Expected Value: area under the curve.
- Difference between two CDFs.
- Expected Remaining time to event:
  $\mu(t) = \mathbb{E}[T - t|T > t] = \frac{1}{S(t)} \int S(t)dt$

Regression Analysis

- **Parametric Models:** functional form assumption (f(t), F(t), S(t), and h(t) are related).

  1. **Exponential Model:** $T_i|X_i \sim \text{Exp}(\mu_i)$
  2. **Weibull Model:** $T_i|X_i \sim \text{Weibull}((\mu_i, \alpha)$

- **Semi-Parametric Model:** Cox-Proportional Hazard

R Code

```
    ### Load packages
    library(survival)

### Run a model
model_1 <- survreg(Surv(time, event) ~ X, dist="exponential", data = df)
model_2 <- survreg(Surv(time, event) ~ X, dist="weibull", data = df)
model_3 <- coxph(Surv(time, event) ~ X,  data = df)
```

- `Surv` creates a survival outcome (object). Key components: a censored time and a event indicator (integer). Default option `type = "right"`.
- `survreg/coxph` runs the regression. Use `survfit` for predictions

2/ Expectation-Maximization Alogrithm (E-M)

E-M Algorithm

- Iterative method to find the local maximum likelihood of parameters in statistical models.
- Latent Variables.
- Manifest Variables.

EM: overview

- An algorithm to conduct MLE when latent parameters are involved or some data is missing (recall Quant III!)
- Ideally we want to estimate parameters of interest by evaluating the loglikelihood function (Score and Hessian)
- Sometimes it is impossible to evaluate it because of missing data or existence of latent variables (in some cases we intentionally include them to make the function tractable)

  $\rightsquigarrow$ introduce the complete loglikelihood function
- Since we never observe the complete loglikehood function, we take the expectation of the latent variable given data and parameters
- MLE we learned: known likelihood function, only update parameters

  $\rightsquigarrow$ EM: iteratively update both likelihood function and parameters

EM: overview

- Expectation of the complete loglikelihood: $Q(\theta, \theta^{\text{old}})$

$$
\begin{aligned}
Q(\theta, \theta^{\text{old}}) &= \mathbb{E}Z \mid X, \theta^{\text{old}}[\log p(X, Z \mid \theta) \mid X, \theta^{\text{old}}] \\
&= \int \log p(X, Z \mid \theta) \times f(Z \mid X, \theta^{\text{old}}) dZ \\
\Big(&= \sum_{Z} \log p(X, Z \mid \theta) \times f(Z \mid X, \theta^{\text{old}}) \quad \text{for discrete } Z \Big)
\end{aligned}
$$

- Note that the definition of expectation is $[x] = \int x f(x) dx$[1]

[1] In the lecture slide, f is written as $p(Z \mid X, \theta^{\text{old}})$

EM: overview

- Expectation of the complete loglikelihood: $Q(\theta, \theta^{\text{old}})$

$$
\begin{aligned}
Q(\theta, \theta^{\text{old}}) &= \mathbb{E}Z \mid X, \theta^{\text{old}} [\log p(X, Z \mid \theta) \mid X, \theta^{\text{old}}] \\
&= \int \log p(X, Z \mid \theta) \times f(Z \mid X, \theta^{\text{old}}) dZ \\
(&= \sum_Z \log p(X, Z \mid \theta) \times f(Z \mid X, \theta^{\text{old}}) \quad \text{for discrete } Z)
\end{aligned}
$$

- Note that the definition of expectation is $[x] = \int xf(x)dx$[1]

- Why $\mathbb{E}Z \mid X, \theta^{\text{old}} [p(X, Z \mid \theta) \mid X, \theta^{\text{old}}]$?
  $\rightsquigarrow$ Use existing information, X and θ^{old} to obtain Z (better than random guess)

---

[1] In the lecture slide, f is written as $p(Z \mid X, \theta^{\text{old}})$

E-step and M-step

- Expectation (E-step): evaluate the expectation of the complete loglikelihood
- "fill in" (or make a guess about) the missing data Z
- Once we get some information about Z, $f(Z \mid X, \theta^{\text{old}})$, we can evaluate Q

  $\rightsquigarrow$ evaluating Q is equilavent to evaluating $f(Z \mid X, \theta^{\text{old}})$

- Maximization (M-step): estimate the parameter of interest using the function obtained in E-step
- Update θ by optimizing Q (MLE)

- E-step: update Q by **updating** Z **given** θ (and X)
- M-step: update θ given Z (and X)
- Repeat these two steps until estimated θ converges

3/ Survival Analysis: Proofs

Survival Analysis: Proofs

- **Geometric Proof:** create some plots. If you use `ggplot2`, the function `geom_rect()` may be useful.
- **Algebraic Proof, Discrete Case:**
  1. Remember, you must show that (1) = (2) = (3). One way to do this is to show that (1)=(2) and (2)=(3).
  2. You have to be sure that you understand the subscripts of the summations.
  3. The Law of the Unconscious Statistician (LOTUS) is a good way to start manipulating equation (1) (see Recitation 8).
- **Algebraic Proof, Continuous Case:** Use integration by parts (examples in the next slide).

Survival Analysis: Proofs

$$\int \log(x)dx$$

$$u = \log(x) \longrightarrow du = \frac{1}{x}dx$$

$$dv = (1)dx \longrightarrow v = x$$

$$\int udv = uv - \int vdu$$

$$\begin{aligned}
\int \log(x)(1)dx &= \log(x)(x) - \int x\frac{1}{x}dx \\
&= x\log(x) - \int (1)dx \\
&= x\log(x) - x + C
\end{aligned}$$

Survival Analysis: Proofs

$$\int x e^{-x} dx$$

$$u = x \longrightarrow du = (1) dx$$

$$dv = e^{-x} dx \longrightarrow v = -e^{-x}$$

$$\int u dv = uv - \int v du$$

$$\begin{aligned} \int x e^{-x} dx &= (x)(-e^{-x}) - \int -e^{-x} (1) dx \\ &= -x e^{-x} - e^{-x} + C \\ &= -(x+1) e^{-x} + C \end{aligned}$$