Survival Analysis

Survival analysis is about analyzing the amount of time until failure, or more generally, until an event of interest happens

We look at some challenges involved with collecting data over a long period of time:

- **Study cut short**: if we analyze the average time until subscription canceled, then for all ongoing subscriptions, we will have a lot of data that is set at the total time so far
  - This max time is known as the **censoring time**
- **Dropping out of study**: if participants choose when to drop out of the study, they will have different censoring times

We want to make the most of our data, including partial observations

Survival Function

Let T be the variable of interest, that is the amount of time until an event of interest occurs. If $F(t)$ is the cdf of T, then the survival function is:

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

We are curious with estimating this, which would be possible with the techniques we have learned so far if it weren't for our censored data

Censored Random Variables

Let C_i denote the censoring time of sample i, and define:

$$\tilde{T}_i = \min(T_i, C_i)$$

$$\delta_i = 1(C_i \geq T_i)$$

We assume the C_i are iid and the T_i are iid with them being independent of each other

- We get to observe $\tilde{T}_i$ and δ_i but not T_i or C_i

Kaplan-Meier Estimator

Assume for simplicity we have discrete time. Then:

$$\begin{aligned} S(t) = P(T > t) &= P(T > t | T > t-1) P(T > t-1) \\ &= (1 - P(t \leq t | T > t-1)) P(T > t-1) \\ &= \underbrace{(1 - P(t = t) | T > t-1)}_{q(t)} \underbrace{P(T > t-1)}_{S_{t-1}} \end{aligned}$$

We can then apply this recursively to get:

$$S(t) = \prod_{s=0}^{t} q(t)$$

where $q(0) = 1 - P(T = 0)$

Hazard Rate

The hazard rate is $1 - q(s)$, or:

$$h(s) = P(T = s | T > s - 1) = \frac{P(T = s)}{P(T \geq s)}$$

Our goal is to estimate $h(s)$ by $\hat{h}(s)$, then set $\hat{q}(s) = 1 - \hat{h}(s)$, and finally our estimator will be:

$$\hat{S}(t) = \prod_{s=0}^{t} \hat{q}(s)$$

To estimate $h(s)$, we use:

$$\hat{h}(s) = \frac{\sum_{i=1}^{n} 1(\tilde{T}_i = s, \delta_i = 1)}{\sum_{i=1}^{n} 1(\tilde{T}_i \geq s)}$$

- $\delta_i = 1$ if this is an uncensored datapoint
- So we look at out of the total amount of datapoints that passed s, how many of them naturally ended at s

Two variations of the above model:

- What if the characteristics also depend on some feature vector x?
    - The Cox hazard regression model models this as:

$$h(t, x) = h_0(t)\mathbb{E}\left[\beta^T x\right]$$

- What if time is continuous?
    - We model this as:

$$h(t) = \frac{P(t \leq T \leq t + dt)}{P(T \geq t)} = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt} \log S(t)$$