Lecture 4 - Survival Models

Survival Models

- Definition and Hazards
- Kaplan Meier
- Proportional Hazards Model

Estimation of Survival in R
GLM Extensions: Survival Models

- Survival Models are a common and incredibly useful extension of the generalized linear model.
  - They are linked on a basic level to Poisson arrivals, which as we learned earlier, yield an exponential distribution of arrival times.

- Survival models are used across many fields
  - Medicine and biostatistics: Many drugs are used to prolong life in the face of serious illness.
  - Firm survival and death. How long do businesses live? Eg: conditional on entering a market (or new market) today, what is the probability of bankruptcy in 12 months?
  - One can imagine survival being used to model time spent on webpages, shopping, Facebook, etc...

- In this part of the course, we’ll learn the basics of survival models using the GLM methodology, and then discuss extensions.
Let $y$ be survival time, and $f(y)$ be the pdf of survival times.

Probability of surviving less than $y$ is:

$$F(y) = \Pr(Y < y) = \int_0^y f(t)dt$$

By the property of complements, the probability of surviving longer than $y$ is the \textit{survivor function}

$$S(y) = 1 - F(y)$$

The \textit{hazard function}, $h(y)$ is the probability of death within a small period between $y$ and $\delta y$, \textit{given they have survived until} $t$.

$$h(y) = \lim_{\delta \to 0} \frac{F(y + \delta y) - F(y)}{\delta y} \cdot \frac{1}{S(y)} = \frac{f(y)}{S(y)}$$

This is essentially a conditional probability. Conditional on surviving up to $y$ or later, $S(y)$, what is the instantaneous probability of death?
GLM Extensions: Survival Models

- For a few more definitions, it is straightforward to show that the hazard function is linked to the survivor function:

\[ h(y) = -\frac{d}{dy} \log(S(y)) = -\frac{dS(y)}{dy} \frac{1}{S(y)} = \frac{f(y)}{S(y)} \]

- Finally, the cumulative hazard function, \( H(y) \) is written as

\[ H(y) = -\log(S(y)) \]

- Example: Exponential Distribution

\[ f(y) = \theta \exp(-\theta y) \]

\[ F(y) = \int_0^y \theta \exp(-\theta t) dt = (-\exp(-\theta t))\bigg|_0^y = 1 - \exp(-\theta y) \]

- Exponential Survivor function and Hazard:

\[ S(y) = \exp(-\theta y) , \quad h(y) = \theta \]

- Note that the hazard does not depend on age. Thus, the exponential distribution is "memoryless". When is this a good or bad property?
The memoryless property makes the exponential distribution unsuitable for a number of applications.

The Weibull distribution nests the exponential distribution.

\[ f(y) = \lambda \phi y^{\lambda-1} \exp(-\phi y^\lambda) \]

Under what condition is this identical to the exponential distribution?

The survival function of Weibull:

\[ S(y) = \int_t^\infty \lambda \phi t^{\lambda-1} \exp(-\phi t^\lambda) dt \]

\[ = \exp(-\phi y^\lambda) \]

Hence, the hazard is written as:

\[ h(y) = \lambda \phi y^{\lambda-1} \]

The link between \( y \) and the hazard may be either positive or negative. What are some economic examples of each?
Simple Estimation: Survival Models

- One way to estimate survival models is to construct a Kaplan-Meier estimate of the survivor function.

- For this, individuals are ordered by time of death from 1 to $n$:
  - $y_1 \leq y_2 \leq \cdots \leq y_k$, where $n_j$ is the number of individuals alive just before $y_j$ and $d_j$ the number of deaths that occur at time $y_j$.

- First, consider the probability of survival just before $y_1$.
  $$ \hat{S}(y \in [0, y_1)) = 1 $$

- Next, probability of survival just before $y_2$.
  $$ \hat{S}(y \in [y_1, y_2)) = 1 \times \frac{n_1 - d_1}{n_1} $$

- Next, probability of survival just before $y_3$.
  $$ \hat{S}(y \in [y_2, y_3)) = 1 \times \frac{n_1 - d_1}{n_1} \times \frac{n_2 - d_2}{n_2} $$
In general, the Kaplan-Meier estimate of the survivor function at time $y(s)$ is the following:

$$\hat{S}(y(s)) = \prod_{j=1}^{s} \left( \frac{n_j - d_j}{n_j} \right)$$

This can be compared to the survivor function for Exponential and Weibull distributions:

- **Exponential**:
  $$S(y) = \exp(-\theta y)$$

- **Weibull**:
  $$S(y) = \exp(-\phi y^\lambda)$$

How do we choose between the two distributions?

Take logs of the survivor functions:

- **Exponential**:
  $$\log(S(y)) = -\theta y$$

- **Weibull**:
  $$\log(S(y)) = -\phi y^\lambda$$

Log of KM estimate should be approximately linear for exponential, non-linear for Weibull.
To study survival models, we will use an influential study, the "Gehan-Freirich" Survival Data

- Data available on course website in stata format

The data show the length of remission in weeks for two groups of leukemia patients, treated and control

- **weeks**: Weeks in remission (effectively survival)
- **relapse**: 1 if a relapse observed, 0 otherwise (this is censoring)
- **group**: 1 if respondent was in treatment group, 0 if in control

The library "survival" contains many function that were useful for survival models.

To construct Kaplan-Meier Estimates:

```r
fit <- survfit(Surv(weeks, relapse)~group, data = g)
plot(fit, lty = 2:3)
legend(23, 1, c("Control", "Treatment"), lty = 2:3)
```
Estimation: Survival Models

- The importance of the survival function and hazard function become apparent when estimating rigorously by maximum likelihood.

- For survival analysis, the data are recorded by subject $j$
  - $y_j$ is the survival time of individual $j$
  - $\delta_j = 1$ is a variable identifying uncensored observations, $\delta_j = 0$ if censored.
  - $x_j$ a vector of explanatory variables for $j$.
  - Order $j$ such that $j = 1..r$ are uncensored, and $j = r + 1..n$ are censored

- Censored individuals are still "surviving" at the end of the data collection. We do not observe when censored individuals actually die.

- For uncensored data, the likelihood function is written as:

\[
L = \prod_{j=1}^{n} f(y_j)
\]
Estimation: Survival Models

- With censored data, the likelihood function is written as:

\[ L = \prod_{j=1}^{r} f(y_j) \prod_{j=r+1}^{n} S(y_j) \]

- \( f(y_j) \) is the pdf at \( y_j \), which is appropriate for uncensored data.

- \( S(y_j) \) is the probability that we observe \( y_j \) or greater, which is the appropriate likelihood to consider for censored observations.

  - We know that a censored individual \( j \) survives \( y_j \) or longer, so the likelihood of this event is \( S(y_j) \)

- Rearranging the likelihood function, we get:

\[ L = \prod_{j=1}^{n} f(y_j)^{\delta_j} S(y_j)^{1-\delta_j} \]

- We can now place this in log-likelihood form, and impose the distributional assumptions.
Estimation: Survival Models

- In log-likelihood form:

\[
l = \sum_{j=1}^{n} \left( \delta_j \log(f(y_j)) + (1 - \delta_j) \log(S(y_j)) \right)
\]

\[
= \sum_{j=1}^{n} \left( \delta_j \log(f(y_j)) + \log(S(y_j)) - \delta_j \log(S(y_j)) \right)
\]

\[
= \sum_{j=1}^{n} \left( \delta_j (\log(f(y_j)) - \log(S(y_j))) + \log(S(y_j)) \right)
\]

\[
= \sum_{j=1}^{n} \left( \delta_j \log(h(y_j)) + \log(S(y_j)) \right)
\]

- Intuition:

  - All individuals survive until \(y_j\). This is accounted for in \(\log(S(y_j))\)
  
  - For individuals with \(\delta_j = 1\), they die at \(y_j\). So, we account for this within the likelihood function using the hazard function, \(\log(h(y_j))\)
Estimation: Exponential Survival

- The exponential distribution has convenient forms for $h(y_j)$ and $S(y_j)$.

\[ h(y_j) = \theta, \quad S(y_j) = \exp(-\theta y_j) \]

- Thus, log-likelihood is:

\[ l = \sum_{j=1}^{n} \left( \delta_j \log(\theta_j) - \theta_j y_j \right) \]

- This looks a lot like a Poisson likelihood function, with $\delta_j$ as the dependent variable. To get it even closer, write:

\[ l = \sum_{j=1}^{n} \left( \delta_j \log(\theta_j y_j) - \theta_j y_j - \delta_j \log(y_j) \right) \]

- Defining $\mu_j = \theta_j y_j$, we have

\[ l = \sum_{j=1}^{n} \left( \delta_j \log(\mu_j) - \mu_j - \delta_j \log(y_j) \right) \]

- We choose $\mu_j$ to maximize the log-likelihood.
Estimation: Exponential Survival

- Often, we assume a *proportional hazards model*, where the hazard function is related to observables, $\theta_j = \exp(x\beta)$.
  - While exponential is memoryless, the probability of dying at $y$ is a function of observables (treatment vs control, for example).

- Thus, substituting into $\mu_j = \theta_j y_j$, we have
  $$\mu_j = \exp(x\beta) y_j$$

- Taking logs:
  $$\log(\mu_j) = x\beta + \log(y_j)$$

- Exponential with proportional hazards can be estimated by
  - `glm` in R, Poisson as family
  - log link ($\mu$ to $x\beta$)
  - Offset (of the log mean) by $\log(y_j)$
Estimated the simple exponential survival model using R

form<-as.formula(relapse~group+offset(log(weeks))
haz_glm<-glm(form, family=poisson("log"), data=g)
summary(haz_glm)

To interpret, note that the hazard is estimated as:

\[ \theta_{\text{treat}} = \exp(\beta_0 + \beta_1 \text{Treat}) \]
\[ = \exp(\beta_0)\exp(\beta_1 \text{Treat}) \]

Note that \( \theta_{\text{control}} = \exp(\beta_0) \). Hence:

\[ \frac{\theta_{\text{treat}}}{\theta_{\text{control}}} = \theta_{\text{control}} \exp(\beta_1 \text{Treat}) \]
\[ = \exp(\beta_1) \]

\[ \frac{\theta_{\text{treat}} - \theta_{\text{control}}}{\theta_{\text{control}}} = \exp(\beta_1) - 1 = \exp(-1.53) - 1 = -0.783 \]

78% reduction in the hazard of relapse relative to control.