Survival Analysis

Adam J Sullivan, PhD

03/19/2018

Binary Covariate - Survival Time Outcome

Binary Covariate - Survival Time Outcome

- For this problem we are looking at a survival time outcome.
- This type of outcome has both event status as well as a time included.
- For this type we might be comparing how long after taking a medication that a patient is cured.
- We will look at various ways to do this now.

Goals of Survival Analysis

- Goals of Survival Analysis:
    - Estimate distribution of of survival time for a population
    - Test for equality of survival distributions among 2 or more groups
        - Control
        - Treated
    - Estimate the absolute or relative treatment effects
    - Estimate and control for effects of other covariates
        - Confounding
        - Effect Modification/Interaction
    - Find confidence intervals and significance for effects.

Describing and Characterizing Survival Data

- The event
  - What is the event of interest?
  - How is it specifically defined?
- The Origin
- What is the initial starting point?
- This must be before anyone in the study has had the event of interest.
- The Metric for time
- What is the scale in which events are recorded.

Examples

- Time to relapse after end of treatment among cancer patients.
    - : Relapse of Cancer
    - : End of Treatment.
    - : Days
- Length of stay in hospital for patients who suffered a heart attack.
    - : Length of Stay.
    - : Admission to Hospital.
    - : Hours.

Examples

- Age of onset of breast cancer in individuals with family history.
  - : Onset of Breast Cancer.
  - : Birth.
  - : Years

The Survival Function

- The random variable of interest, T is the time to the event of interest.
- We then know that T is positive and by definition:

$$T \geq 0$$

The Survival function, $S(t)$, is the proportion of individuals who have not experienced at event at some time $t > 0$.This is defined by:

$$S(t) = \Pr(T > t)$$

The Survival Function

- If the event of interest is death, this would mean the subject is still alive at time t.
- Then $S(t)$ would be the proportion of subjects alive at time t.
- This simple proportion is for when there is no (discussed later on).

Features of $S(t)$

- A survivor function is a sequence of probabilities up till time t

$$0 \leq S(t) \leq 1, \qquad \text{for each } t \geq 0$$

- At time, $t = 0$, $S(0) = 1$.
- $S(t)$ decreases as events happen over t
    - So that if $t_2 \geq t_1$ then $S(t_2) \geq S(t_1)$.
    - They are non-increasing functions.

Features of $S(t)$

- For large, t, such as $t = \infty$, $S(t)$ goes to 0.
    - This means that for some events $S(t)$ approaches a 0 asymptote.
    - However for some diseases, some people may be cured so that $S(t)$ approaches a non-zero asymptote.
- Graphical displays are a common method to display summaries survivor functions.

The Hazard Function

The Hazard function is defined as the instantaneous rate of failure at time t, given that a subject as survived up until time t.

It is also referred to as:

- Hazard Rate.
- Failure Rate.
- Mortality Rate

Relationship between $h(t)$ and $S(t)$:

- The hazard function $(h(t))$, survival function $(S(t))$, probability density function $(f(t))$, and cumulative distribution function $(F(t))$ are all related.
- They are defined in terms of random variable T which is the time until event.
- In censored subjects we only know that $T > t$ for a subject censored at time t.
- Mathematically:

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

Kaplan-Meier Estimator

The control group was easy to analyze as it had no censoring so we could calculate it by hand. However with the introduction to censoring we need a new estimator.

- Kaplan-Meier is a non-parametric method.
    - No assumptions of distribution
- We define patients to be at risk at time t if they have not experienced the event just before time t and are not yet censored just before time t.

Log Rank Test

- The Logrank Test is a hypothesis test for 2 or more independent samples of survival data.
- The hypothesis being tested are:

$$H_o : S_1(t) = S_2(t) \text{ for all } t$$

$$\text{and}$$

$$H_o : S_1(t) \neq S_2(t) \text{ for some } t$$

Log Rank Test

If H_0 is true then

- $h_1(t) = h_2(t)$ for all t
- $\Lambda_1(t) = \Lambda_2(t)$ for all t

How do we calculate this test statistic?

1. Construct a 2x2 table at the time of each observed failure.
2. Calculate the Mantel-Haenszel chi-square test statistic.

- We have K distinct observed failure times:

$$t_1 < \cdots < t_K$$

- at the i^{th} observed failure time t_i:

How do we calculate this test statistic?

TREATMENT	DIED	ALIVE	AT RISK
Control	a_i	b_i	n_{1i}
Treated	c_i	d_i	n_{2i}
total	m_{1i}	m_{2i}	n_i

where

$$\begin{aligned} n_{1i} &= \text{numer at risk at } t_i \text{ from Control} \\ n_{2i} &= \text{numer at risk at } t_i \text{ from Treated} \\ m_{1i} &= \text{number of failures at } t_i \\ m_{2i} &= \text{number surviving past } t_i \\ n_i &= \text{total numer at risk at } t_i \end{aligned}$$

Relation to Mantel-Haenszel Test

- This test is exactly the same as a Mantel-Haenszel test applied to K strata

$$\chi^2_{MH} = \frac{\left[\sum_{i=1}^{K} (a_i - E(a_i))\right]^2}{\sum_{i=1}^{K} Var(a_i)}$$

- where

$$E(a_i) = \frac{n_{1i} m_{1i}}{n_i}$$
$$Var(a_i) = \frac{n_{1i} n_{2i} m_{1i} m_{2i}}{n_i^2 (n_i - 1)}$$

Relation to Mantel-Haenszel Test

We compute the expectation that the null hypothesis is true and there is no difference in survival between the groups. We consider all margins fixed but a_i is random and thus we have a hypergeometric distribution.

- Under H_0 we have that $S_1(t) = S_2(t)$ and this means
  - $\chi^2_{MH} \sim \chi^2_1$
  - Reject H_0 when $\chi^2_{MH} > \chi^2_{1,1-\alpha}$
- This test is most powerful if the hazard ratio is constant over time.
- We can easily extend this to compare 3 or more independent groups.

Recall the Colorectal Cancer component of the Physicians Health Study

NAME	DESCRIPTION
age	Age in years at time of Randomization
asa	0 - placebo, 1 - aspirin
bmi	Body Mass Index (kg/m^2)
hypert	1 - Hypertensive at baseline, 0 - Not
alcohol	0 - less than monthly, 1 - monthly to less than daily, 2 - daily consumption

NAME	DESCRIPTION
dm	0 = No diabetes Mellitus, 1 - diabetes Mellitus
sbp	Systolic BP (mmHg)
exer	0 - No regular, 1 - Sweat at least once per week
csmoke	0 - Not currently, 1 - < 1 pack per day, 2 - $\geq$ 1 pack per day
psmoke	0 - never smoked, 1 - former < 1 pack per day, 2 - former $\geq$ 1 pack per day
pkyrs	Total lifetime packs of cigarettes smoked
crc	0 - No colorectal Cancer, 1 - Colorectal cancer
cayrs	Years to colorectal cancer, or death, or end of follow-up.

What Each Subject Contributed

1. Information on whether of not they had a Colorectal Cancer(CRC) during follow-up
2. Follow-up time in years, specified as time from randomization until first of
    - end of Study
    - death
    - Colorectal Cancer
    - Loss to follow-up

Loading Data

```
library(tidyverse)
library(haven)
phscrc <- read_dta("phscrc.dta")
phscrc <- phscrc %>% mutate(age.cat = cut(age, c(40, 50, 60,
    70, 90), right = FALSE)) %>% mutate(alcohol.use = factor(alcohol >
    1, labels = c("no", "yes"))) %>% mutate(obese = factor(bmi >
    30, labels = c("Not Obese", "Obese")))
```

Example: Kaplan-Meier Survival

```
library(survival)

model <- survfit(Surv(cayrs, crc) ~ alcohol.use, data = subset(phscrc,
    cayrs > 0))
model
```

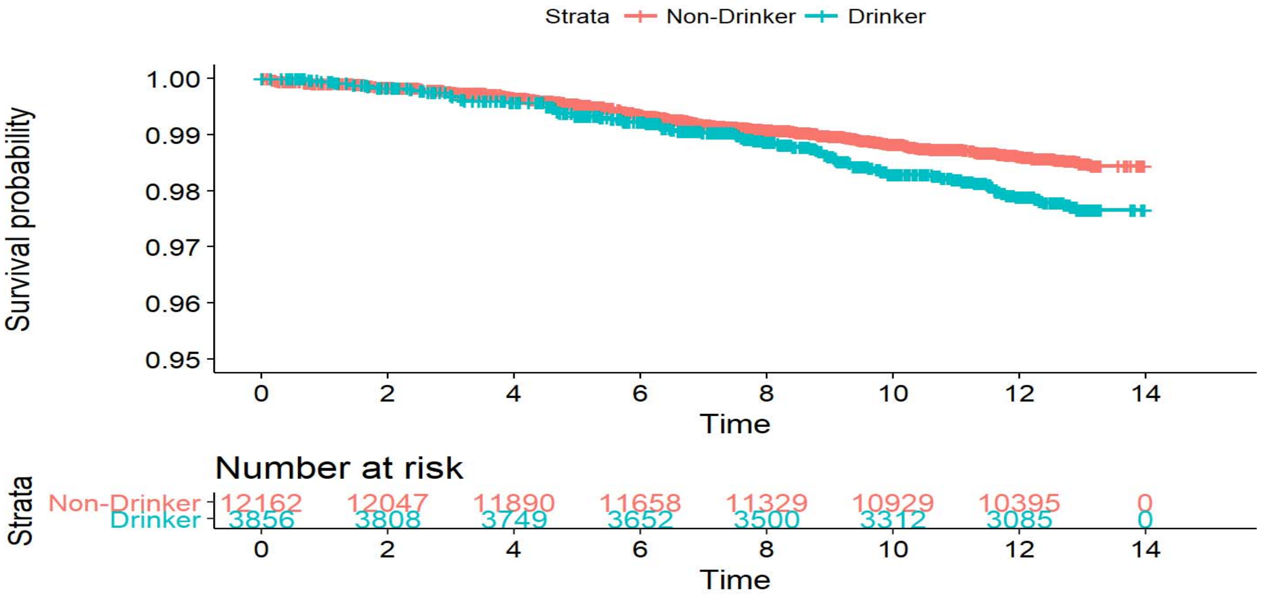
Example: Kaplan-Meier Survival

```
## Call: survfit(formula = Surv(cayrs, crc) ~ alcohol.use, data = subset(phscrc, 
##     cayrs > 0))
## 
##                     n events median 0.95LCL 0.95UCL
## alcohol.use=no  12162    173     NA      NA      NA
## alcohol.use=yes  3856     81     NA      NA      NA
```

Plotting the Kaplan-Meier

```
library(survminer)
ggsurvplot(model, conf.int = FALSE, risk.table = TRUE, risk.table.col = "strata",
    legend.labs = c("Non-Drinker", "Drinker"), break.time.by = 2,
    ylim = c(0.95, 1))
```

Plotting the Kaplan-Meier



Log Rank Test

```
model <- survdiff(Surv(cayrs, crc) ~ alcohol.use, data = subset(phscrc,
    cayrs > 0))
model
```

Log Rank Test

```
## Call:
## survdiff(formula = Surv(cayrs, crc) ~ alcohol.use, data = subset(phscrc,
##     cayrs > 0))
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## alcohol.use=no  12162      173      194      2.27      9.61
## alcohol.use=yes  3856       81       60      7.34      9.61
##
##  Chisq= 9.6  on 1 degrees of freedom, p= 0.00193
```

Conclusion

- What can we conclude?

Categorical Covariate - Survival Time Outcome

Categorical Covariate - Survival Time Outcome

- Same exact method Kaplan-Meier and Log Rank.
- We will see an example of this.

CRC Example: Kaplan-Meier Survival

```
model <- survfit(Surv(cayrs, crc) ~ csmok, data = subset(phscrc,
    cayrs > 0))
model
```

PBC-3 Example: Kaplan-Meier Survival

```
## Call: survfit(formula = Surv(cayrs, crc) ~ csmok, data = subset(phscrc,
##     cayrs > 0))
##
##             n events median 0.95LCL 0.95UCL
## csmok=0 14307    217     NA      NA      NA
## csmok=1   575      7     NA      NA      NA
## csmok=2  1136     30     NA      NA      NA
```

PBC-3 Example: Plotting the Kaplan-Meier

```
ggsurvplot(model, conf.int = TRUE, risk.table = TRUE, risk.table.col = "strata",
    break.time.by = 2, ylim = c(0.95, 1))
```

Log Rank Test

```
model <- survdiff(Surv(cayrs, crc) ~ csmok, data = subset(phscrc,
    cayrs > 0))
model
```

Log Rank Test

```
## Call:
## survdiff(formula = Surv(cayrs, crc) ~ csmok, data = subset(phscrc, 
##     cayrs > 0))
## 
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## csmok=0 14307      217   227.41     0.476     4.549
## csmok=1   575        7     9.11     0.489     0.508
## csmok=2  1136       30    17.48     8.965     9.628
## 
##  Chisq= 9.9  on 2 degrees of freedom, p= 0.00697
```

Conclusion