Survival Analysis using R

Phillip Ayieko

Survival Topics

- Survival Analysis concepts
- Common survival analysis techniques
- Presenting results
- Both theory and practice
  - Definitions
  - censoring
  - Kaplan-Meier
  - Log-rank
  - Cox proportional hazards
  - R statistical software (open source: `http://cran.r-project.org/`)

Introduction to survival analysis

- Survival Analysis is referred to statistical methods for analyzing survival/ time to event data
- Events could be
  - Recurrence of disease, death etc
- Examples
  - Time between surgery and relapse
- Data is usually censored
  - Example-study ends before all the patients dies

Objective of survival analysis

- Estimate probability that an individual surpasses some time-to-event for a group of individuals.
- Compare time-to-event between two or more groups.
- Assess the relationship of covariates to time-to-event.

Types of survival data

- When the time of event is known, then we have complete information e.g. rat number 4 died on 3rd day. This is called uncensored data
- When the occurrences of the event are not known then we have incomplete information e.g. those rats who escaped. This is called censored data
- Follow-up time, calculated from
  - exit time when the outcome is experienced, or last-seen time
  - entry time when individual enters study

Two reasons for censoring

- During investigation we have incomplete information due to LOSS TO FOLLOW UP
- When the investigation has come to an end and the experimental units are still alive. This is censoring due to WITHDRAWAL ALIVE

Forms/ Types of censoring

- Right censoring
- Left censoring
- Type I censoring
- Type II censoring
- Random censoring
- Interval censoring
- Informative and non-informative censoring
    - Most common is right censored

Censoring

- How could we account for censoring?
  - Ignore it and say event occurred at time of censoring
    - Incorrect because this is almost certainly not true
  - Remove patient from analysis
    - Potential bias and loss of power
  - Survival analysis
- The objective is to estimate the survival distribution of patients in the presence of censoring

Functions describing survival times

- T is a random variable representing survival time - Can be described by a probability distribution
  - 1.Cumulative distribution function
  - 2.Survival function
  - 3.Probability density function
  - 4.Hazard function
  - 5.Cumulative hazard function

Functions of T

- Let **T** denote the survival time
- $t \equiv$ a specific point in time.

1. **Survival function**

- $s(t) = p(surving\_longer_t han\_time\_t)$
- Survival function: probability survival time is greater than a specific value $S(t) = P(T > t)$
- $S = \frac{number\_of_p atients\_surviving\_longer\_than\_t}{total\_number\_of\_patients\_in\_the\_study}$

Mathematical Definitions

- Distribution function

$$F(t) = Pr(T \leq t)$$

- Density Function

$$f(t) = \frac{\partial F(t)}{\partial t}$$

- Survival Function

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

Mathematical Definitions

- Hazard function:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

- Cumulative hazard

$$H(t) = \int_0^t h(u)du$$

Statistical Inference

- Has two components
  - Estimation of survival function
  - Testing of hypothesis of survival functions
- There are two methods of approaching them
  - Parametric approach
  - Non parametric approach
- A combination of the 2 approaches gives the regression (i.e. estimation and testing of the hypothesis)

KAPLAN-MEIER ESTIMATOR

Kaplan-Meier Estimate of S(t)

- Kaplan-Meier estimator of survival is a nonparametric method of inference concerning the survivor function $S = Pr(T > t)$
- Rank the survival times as $t_1 \leq t_2 \leq ... \leq t_n$
- Number of individuals at risk before $t_i \equiv n_i$
- Number of individuals with failure time $t_i \equiv d_i$
- Estimated hazard function at t_i: $h_i = \frac{d_i}{n_i}$
- Formula

$$S(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}$$

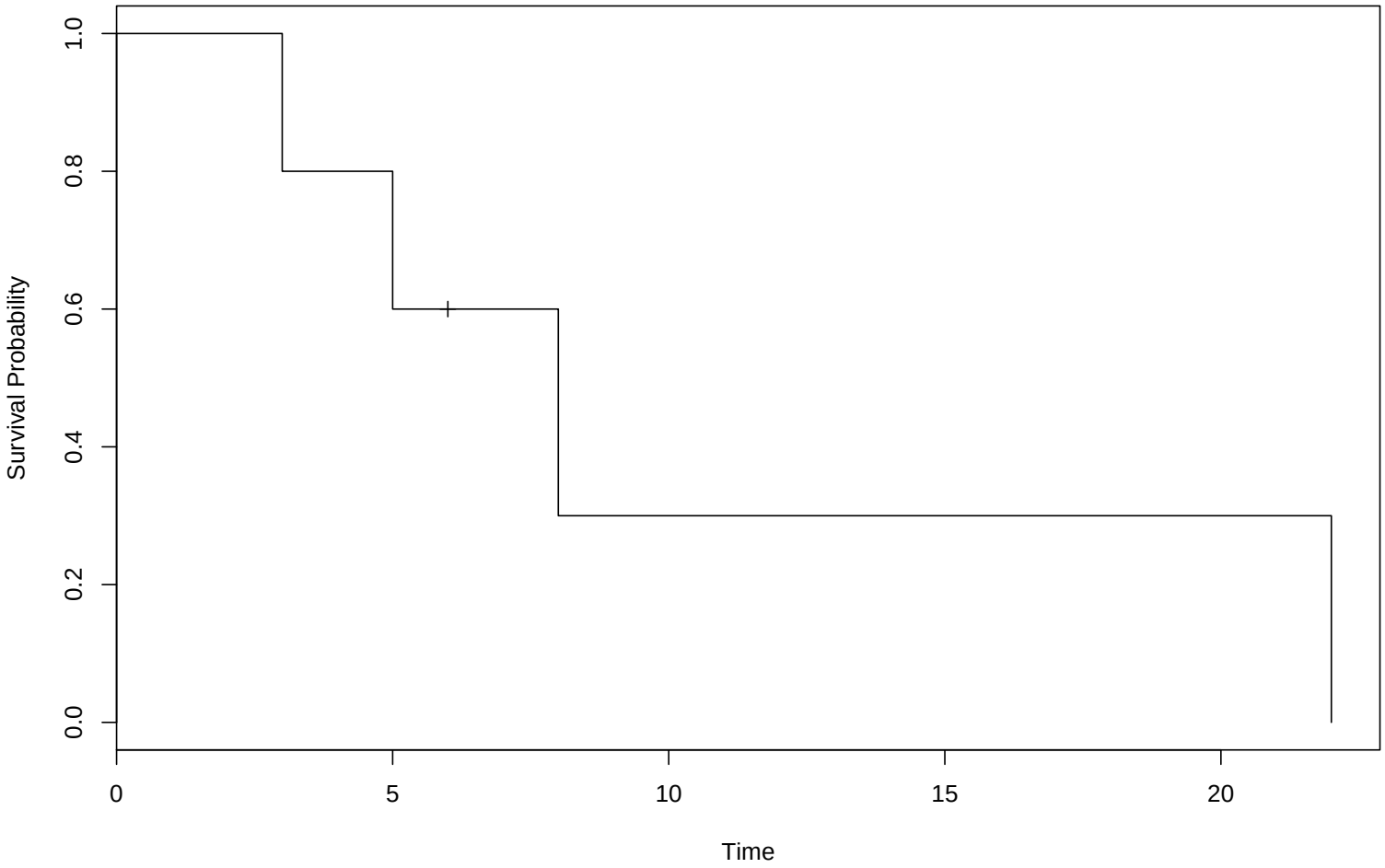
Kaplan-Meier Estimate of S(t)

$$var[s(t)] \approx [s(t)]^2 \sum \frac{d_j}{n_j(n_j - d_j)}$$

- And the 95 % CI is given by

$$s(t) \pm 1.96\sqrt{var(s(t))}$$

Kaplan-Meier Curve



Functions of interest in R

- Survival object: **Surv**
- Kaplan-Meier estimates: **survfit**
- The log-rank test: **survdiff**
- The Cox proportional hazards model: **coxph**
- The Accelerated failure time model: **survreg**
- Relevant R packages: **survival, survcomp, HMISC, Design, MASS**

Example

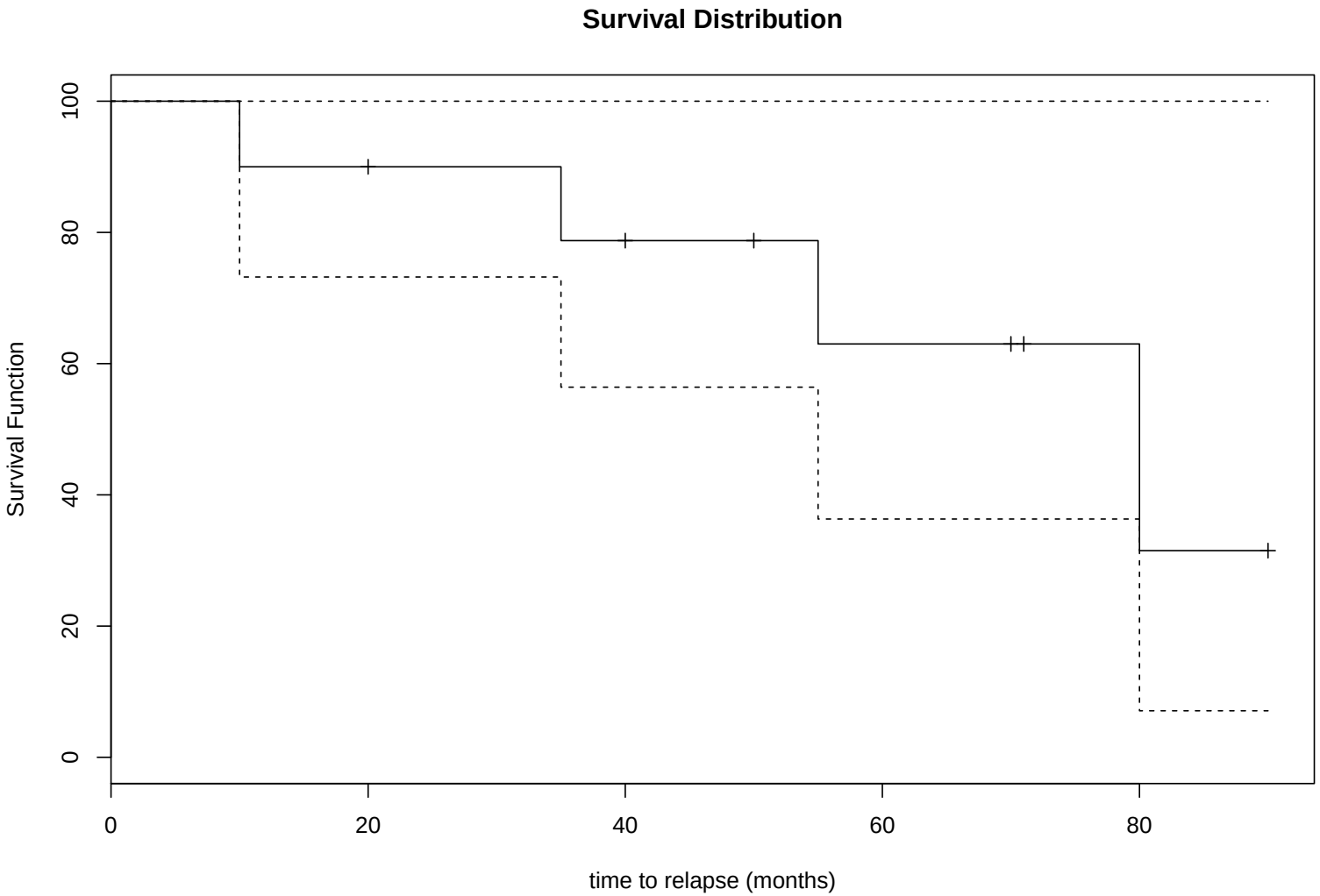
- Event = time to relapse
- Data:
  - 10, 20+, 35, 40+, 50+, 55, 70+, 71+, 80, 90+
- Before complex functions may be performed, the data has to be put into the proper format: a **survival object**
- In R we use **Surv(time, status)**
- *time* is a vector of event time, and *status* is a vector of indicator (denoting if the event was observed or censored

R codes for Kaplan-Meier

```
library(survival)

time <- c(10,20,35,40,50,55,70,71,80,90)
status <- c(1,0,1,0,0,1,0,0,1,0)
data <- data.frame(time,status)

survobj <- with(data,Surv(time,status))

fit <- survfit(survobj~1, data=data)
fit
```

```
## Call: survfit(formula = survobj ~ 1, data = data)
##
## records   n.max n.start  events  median 0.95LCL 0.95UCL
##      10      10      10       4      80      55      NA
```

```
plot(fit, xlab="time to relapse (months)", ylab="Survival Func
```



Aalen-Nelson Estimate of S(t)

- Uses the cumulative hazard function

$$H(t) = -\log s(t) = \sum_{t_j \leq t} \frac{d_j}{n_j}$$

$$s(t) = e^{-\int_0^t h(u)d(u)} = e^{-H(t)}$$

- In R, to estimate S(t), using output from *survfit()*

```
#survobj <- Surv(time, status)
#fit1 <- summary(survfit(survobj~1))
# S.hat <- fit1$surv
# H.hat <- -log(S.hat)
#a<- survfit(coxph(Surv(time,status)~1), type="aalen")
#summary(a)
#basehaz(coxph(Surv(time,status)~1,data=data))
```

LOG-RANK TEST

LOG-RANK TEST

- Use the Log-Rank Test to compare the survival functions of two samples/ treatment groups.
- H_0: The two survival functions are the equivalent
- H_a: The two survival functions are different
- Goal: Test whether two groups differ in population survival functions.

Log-Rank Test

- Notation:
- $t_i \equiv$ Time of the ith failure time (across groups)
- $d_{1i} \equiv$ number of failures for trt 1 at time $t_{(i)}$
- $d_{2i} \equiv$ Number of failures for trt 2 at time $t_{(i)}$
- $n_{1i} \equiv$ Number at risk prior for trt 1 prior to time $t_{(i)}$
- n_{2i} Number at risk prior for trt 2 prior to time $t_{(i)}$

$d_i = d_{1i} + d_{2i}$ and $n_i = n_{1i} + n_{2i}$

$$v_{li} = \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$$

$$e_{1i} = \frac{n_{1i} d_{1i}}{n_i}$$

$$O_1 - E_1 = \sum_i (d_{1i} - e_{1i})$$

R codes

```
survdiff(formula, rho=0)
```

- *formula* is a survival object against a categorical covariate variable.
- *rho* is a scalar parameter that controls the type of test.
- With default *rho=0*, this is the log-rank or Mantel-Haenszel test.
- With *rho=1*, it is equivalent to the Peto and Peto modi???cation of the Gehan-Wilcoxon test.

```
>survdiff(survobj~status, rho=0)
```

Cox Proportional Hazard Model

Cox Proportional Hazard Model

- CPHM is a technique for investigating the relationship between survival time and independent variables
- The most popular model for survival analysis:
- Goal: Compare two or more groups (treatments), adjusting for other risk factors on survival times (like Multiple regression)
- p explanatory variables (including dummy variables)

Cox regression for time to event

- Useful when incidence rates vary over time
- hazards in both exposed and unexposed are allowed to vary but their ratio is assumed to be constant - this is the proportional hazards assumption. Thus the Cox model are also known as the proportional hazards model
- Cox model is fitted by creating risk sets of individuals still at risk at the time of each failure

Cox Proportional Hazard Model

- The semi-parametric model given by

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

- $h_0(t)$ is the baseline hazard function. This is the function when all the covariates equal to zero
- The hazard ratio is independent of time t.

$$HR_{i,j} = \frac{h_i(t)}{h_i(t)} + \frac{\lambda_0(t)e^{\beta_1 x_{i1}+...+\beta_k x_{ik}}}{\lambda_0(t)e^{\beta_k x_{j1}+...+\beta_k x_{jk}}} = e^{\beta_1 x_{j1}+...+\beta_k x_{jk}}$$

-This defines the **proportional hazards property**

Interpretation of the Betas

- We need to find the ratio where there is a unit increase in the covariates provided other covariates stay fixed

$$\frac{h(t, x_1 + 1)}{h(t, x_1)} = \frac{h_0(t)e^{\beta_1(x_1+1)}}{h_0(t)e^{\beta_1(x_1)}} = e^{\beta_1}$$

- We interprete as the unit increase in log hazard per unit increase of x
- (just as log-odds in logistic regression)

R example

- $> coxph(formula, method)$
- **formula** is linear model with a survival object as the response variable
- **method** is used to specify how to handle ties. The default is 'efron'. Other options are 'breslow' and 'exact'.
- $> cox.fit < -coxph(Surv(time, status)x + y + z, method = `breslow)$
- to obtain the baseline survival function
  $> my.survfit.object < -survfit(coxph.fit)$