Survival Analysis: Hard Drive Reliability Sample *

Cesar Conejo Villalobos *Data Scientist*

The branch of statistics that study the expected duration of time for an event to occur is called survival analysis. The number of events can be one or more. This project reviews nonparametric methods like Kaplan-Meier, Nelson-Aalen, and Cox proportional hazards model. These techniques are applied to the Hard Drive date sets of Backblaze. This application of survival analysis is called *failure-time analysis*. In this way, the goal is to find the probabilities that the hard disk works well for 1 year using the data collected in 2019. The package used for this exercise is survival. For the number of files, it also uses data.table package.

Keywords: Survival Analysis, Kaplan-Meier, Nelson-Aalen, Cox

Nonparametric models

S. Klugman (2008)

Data Preparation

```
#Load libraries
library("data.table")
library("tidyverse")

# Read Multiple .csv files. 365 files with daily data of HDD.
# Only choose the columns:
# date
# serial_number
# model
# capacity_bytes
# failure

file_names <- list.files("data/drive_stats_2019",
                         pattern="*.csv",
                         full.names=TRUE)

data <- rbindlist(lapply(file_names,
                         function(x) fread(input = x,
                                           header = TRUE,
                                           stringsAsFactors = FALSE,
                                           select = c("date","serial_number",
                                                      "model","capacity_bytes",
                                                      "failure")
                                          )
```

*Template taken from (http://github.com/svmiller). **Corresponding author**: svmille@clemson.edu.

```
                          )
                  )

#Modify data. Simplify HDD models and capacity
data <- data %>%
  mutate(model          = ifelse(grepl("^ST",model),
                                 'Seagate',
                                 str_extract(model, "^[^\\s]+")
                                 ),
         capacity_bytes = round(capacity_bytes/10e11)
         )


# Aggregation of data. Individual:
# serial_number
# model
# capacity_bytes (TB)
data_group <- data %>%
  group_by(serial_number, model, capacity_bytes) %>%
  summarise(count_obs = n(),
            min_date = min(date),
            max_date = max(date),
            count_fail = sum(failure),
            fail       = max(failure),
            first_date_fail = min(ifelse(failure == 1,
                                         date,
                                         "2020-01-01"
                                         )
                                  )
            )

# Variable: Beginning of study
begin_study <- "2019-01-01"

# Creation variables for survival models:
# age: Period start observation. All HDD are assumed to begin operation at 2019/01/01
# study_time: Period of observation of each HDD
data_group <- data_group %>%
                mutate(age = difftime(min_date, begin_study, units = c("days")),
                       study_time = ifelse(fail == 1,
                                           difftime(first_date_fail,
                                                    min_date,
                                                    units = c("days")
                                                    ),
                                           difftime(max_date,
                                                    min_date,
                                                    units = c("days")
```

```
                                                    )
                                                ) + 1
                        )
# save aggregated data
write.csv(x =  data_group,
          file = "output/data/data_group_2019.csv")
```

Exploratory Analysis

```
# Load aggregate data
data_group <- fread(input = "output/data/data_group_2019.csv",
                    header = TRUE,
                    stringsAsFactors = FALSE)

# Modify data
data_group <-  data_group %>%
                mutate(model = as.factor(model),
                       serial_number = as.factor(serial_number),
                       min_date = as.Date(min_date),
                       max_date = as.Date(max_date),
                       first_date_fail = as.Date(first_date_fail)
                      ) %>%
                select(-V1)
```

```
# Distribution data
summary(data_group)
```

```
## Number of fails for day
fails <- data_group %>% filter(fail == 1)

ggplot(data = fails, aes(x = first_date_fail)) +
  geom_bar() +
  labs(title = "Number of fails by day") +
  xlab("Day of fail") +
  theme_classic()
```

```
# Models of HDD
ggplot(data = data_group, aes(model)) +
  geom_bar(aes(y = (..count..)/sum(..count..), fill = model)) +
  scale_y_continuous(labels=scales::percent, limits = c(0,0.8)) +
  ylab("relative frequencies") +
  labs(title = "Relative Frequency of Hard drive Models") +
  theme_classic()
```

Survival models

```
attach(data_group)

# 1) Kaplan-Meier Global probababilities

# 1.1) Survival function
surv_object_HDD <- Surv(age, age + study_time, fail)

km_survival_HDD <- survfit(surv_object_HDD ~ 1)

# Global option
print(km_survival_HDD)

# Graph
ggsurvplot(
  km_survival_HDD,
  data       = data_group,
  ylim       = c(0.975,1),
  size       = 1,                    # change line size
  palette    = "#2E9FDF",            # custom color palettes
  conf.int   = TRUE,                 # Add confidence interval
  risk.table = TRUE,                 # Add risk table
  risk.table.col    = "strata",      # Risk table color by groups
  legend.lab        = "All Models", # Change legend labels
  risk.table.height = 0.25,          # Useful to change when you have multiple groups
  ggtheme           = theme_bw(),     # Change ggplot2 theme
  title             = "Kaplan-Meier Failure Estimates Hard Disk"
)

# 1.2) cumulative hazard

ggsurvplot(km_survival_HDD,
           data     = data_group,
           conf.int = TRUE,
           ggtheme  = theme_bw(),      # Change ggplot2 theme
           palette  = "#E7B800",
           fun      = "cumhaz")

# 2) Kaplan-Meier non-parametric analysis by model
km_survival_model <- survfit(surv_object_HDD ~ model)


ggsurvplot(km_survival_model,
           data = data_group,
           ylim = c(0.95,1),
           legend.lab = c("DELLBOSS", "HGST", "Hitachi", "Seagate","TOSHIBA", "WDC"),
```

```
           risk.table = TRUE
           )

# 3) Nelson-Aalen non-parametric analysis
na_survival_HDD <- survfit(coxph(surv_object_HDD ~ 1), type = "aalen")
print(na_survival_HDD)


ggsurvplot(
  na_survival_HDD,
  data       = data_group,
  ylim       = c(0.975,1),
  size       = 1,                  # change line size
  palette    = "#2E9FDF",          # custom color palettes
  conf.int   = TRUE,               # Add confidence interval
  risk.table = TRUE,               # Add risk table
  risk.table.col    = "strata",    # Risk table color by groups
  legend.lab        = "All Models", # Change legend labels
  risk.table.height = 0.25,        # Useful to change when you have multiple groups
  ggtheme           = theme_bw()   # Change ggplot2 theme
)


# 4) Univariate Compute the Cox model
res_cox_hdd <- coxph(surv_object_HDD ~ model, data = data_group)
res_cox_hdd


detach(data_group)
```

```
#
```

References

S. Klugman, H. Panjer, G. Willmont. 2008. *Loss Models: From data to decisions*. Wiley.