Survival Analysis - FINAL EXAM - DSTI survey on time-to-internship

Mamadi Fofana

17/03/2021

Executive summary

This project focus on analyzing duration to find internship for some DSTI students. Data has been collected from several cohorts and has 82 rows for 10 variables. We will be using survival analysis to analyze expected duration for a student to find an internship in order to attempt to answer certain questions: • How long does it take to obtain an internship? • Is the waiting time changing between cohorts? • Does the educational background have an impact? • Can we build a predictive model to identify students at high risk of a long search?

After loading of the dataset, we will first do an exploratory data analysis to get a summary of the data Then we will proceed some data preprocessing to clean, transform and filter dataset in the proper formats At the end of this stage we noticed some censoring data

After that , we use survival analysis to study waiting time before internship for students of the study as per the two no-parametric methods :

• Kaplan-Meier : is a non-parametric method used to estimate the survival probability from observed survival times • Log-rank test to compare the survival curves of two or more groups

To validate some of our results , we associate Fisher test to our log-ranks tests.

This study allowed us to find out some observations:

• We observe that on average a student take 120 (4 months) to obtain an internship • waiting time changes among cohorts • Cohort S20 has the longest waiting time , which could be explained by the outbreak of covid19 epidemic.

Finally , we have tried to build some linear models to identify students with certain profile

Library loading

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(sys)
library(survival)

library(summarytools)
```

```
## Registered S3 method overwritten by 'pryr':
##   method      from
##   print.bytes Rcpp
```

```
## For best results, restart R session and update pander using devtools:: or remotes::install_github('rapporter/p
ander')
```

```
##
## Attaching package: 'summarytools'
```

```
## The following object is masked from 'package:tibble':
##
##     view
```

```r
library(readr)
#library(explore)

library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
library(broom)
```

Set Current Dir

```r
setwd("D:/2021/")
getwd()
```

```
## [1] "D:/2021"
```

Loading dataset and glympse of data

```r
dsti_sample = read.csv("DSTI_survey.csv")
str(dsti_sample)
```

```
## 'data.frame':    82 obs. of  13 variables:
##  $ Timestamp                                                    : chr  "11/2/2020 16:59:23" "11/2/2020 16:59:31"
"11/2/2020 16:59:38" "11/2/2020 16:59:45" ...
```

```
##  $ Year.of.birth                                          : int  1992 1993 1990 1986 1993 1992 1995 1992 199
3 1989 ...
##  $ Were.you.ever.a.smoker.                                : chr  "No" "Yes, and I'm currently smoking" "No"
"No" ...
##  $ Year.when.first.started.smoking                        : int  NA 2011 NA NA NA 2019 NA 2010 2013 NA ...
##  $ Year.when.stopped.smoking                              : int  NA NA NA NA NA NA NA NA 2018 NA ...
##  $ When.did.you.start.looking.for.an.internship           : chr  "11/2/2020" "10/19/2020" "3/1/2021" "9/1/20
20" ...
##  $ Sex                                                    : chr  "Male" "Female" "Female" "Male" ...
##  $ When.did.you.stopped.looking.for.an.internship         : chr  "" "" "" "10/31/2020" ...
##  $ Have.you.found.an.internship.                          : chr  "No" "No" "No" "Yes" ...
##  $ Education..background..pick.a.main.one.you.identify.with.: chr  "Mathematics, Physics, Chemistry, Computer
Science, Statistics" "Mathematics, Physics, Chemistry, Computer Science, Statistics" "Mathematics, Physics, Chemi
stry, Computer Science, Statistics" "Medicine, Biology" ...
##  $ Years.of.education                                     : int  20 17 17 22 16 18 16 14 17 18 ...
##  $ Do.you.have.children.                                  : chr  "No" "No" "No" "Yes" ...
##  $ Cohort                                                 : chr  "A20" "A20" "A20" "S20" ...
```

```
summary(dsti_sample)
```

```
##   Timestamp         Year.of.birth  Were.you.ever.a.smoker.
##  Length:82          Min.   :1955   Length:82
##  Class :character   1st Qu.:1981   Class :character
##  Mode  :character   Median :1987   Mode  :character
##                     Mean   :1985
##                     3rd Qu.:1993
##                     Max.   :1997
##                     NA's   :1
##  Year.when.first.started.smoking Year.when.stopped.smoking
##  Min.   :1971                    Min.   :2001
##  1st Qu.:1998                    1st Qu.:2014
##  Median :2004                    Median :2017
##  Mean   :2003                    Mean   :2015
##  3rd Qu.:2014                    3rd Qu.:2018
##  Max.   :2019                    Max.   :2020
##  NA's   :56                      NA's   :65
```

```
##  When.did.you.start.looking.for.an.internship     Sex
##  Length:82                                     Length:82
##  Class :character                              Class :character
##  Mode  :character                              Mode  :character
##
##
##
##
##  When.did.you.stopped.looking.for.an.internship Have.you.found.an.internship.
##  Length:82                                        Length:82
##  Class :character                                 Class :character
##  Mode  :character                                 Mode  :character
##
##
##
##
##  Education..background..pick.a.main.one.you.identify.with. Years.of.education
##  Length:82                                                 Min.   : 4.00
##  Class :character                                          1st Qu.:17.00
##  Mode  :character                                          Median :18.00
##                                                            Mean   :17.98
##                                                            3rd Qu.:20.00
##                                                            Max.   :25.00
##                                                            NA's   :1
##  Do.you.have.children.    Cohort
##  Length:82             Length:82
##  Class :character      Class :character
##  Mode  :character      Mode  :character
##
##
##
##
```

```
head(dsti_sample, 10)
```

	Timestamp <chr>	Year.of.birth <int>	Were.you.ever.a.smoker. <chr>
1	11/2/2020 16:59:23	1992	No
2	11/2/2020 16:59:31	1993	Yes, and I'm currently smoking
3	11/2/2020 16:59:38	1990	No
4	11/2/2020 16:59:45	1986	No
5	11/2/2020 17:00:00	1993	No
6	11/2/2020 17:00:02	1992	Yes, and I'm currently smoking
7	11/2/2020 17:00:09	1995	No
8	11/2/2020 17:00:10	1992	Yes, and I'm currently smoking
9	11/2/2020 17:00:46	1993	Yes, and I stopped
10	11/2/2020 17:00:49	1989	No

1-10 of 10 rows | 1-4 of 14 columns

Exploratory Data Analysis

summary way to get 1-way information for every column in the dataset.

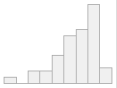
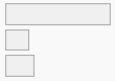
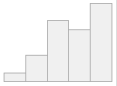
```
library(summarytools)
library(readr)
view(dfSummary(dsti_sample), method = "render")
```

Data Frame Summary

dsti_sample

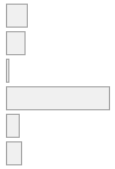
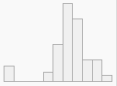
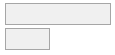
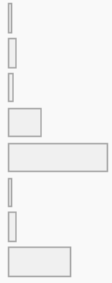
Dimensions: 82 x 13
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	Timestamp [character]	1. 11/2/2020 17:00:53 2. 11/13/2020 10:04:47 3. 11/13/2020 14:54:17 4. 11/2/2020 16:59:23 5. 11/2/2020 16:59:31 6. 11/2/2020 16:59:38 7. 11/2/2020 16:59:45 8. 11/2/2020 17:00:00 9. 11/2/2020 17:00:02 10. 11/2/2020 17:00:09 [71 others]	2 (2.4%) 1 (1.2%) 1 (1.2%) 1 (1.2%) 1 (1.2%) 1 (1.2%) 1 (1.2%) 1 (1.2%) 1 (1.2%) 1 (1.2%) 71 (86.6%)		82 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
2	Year.of.birth [integer]	Mean (sd) : 1985.5 (9) min < med < max: 1955 < 1987 < 1997 IQR (CV) : 12 (0)	29 distinct values		81 (98.8%)	1 (1.2%)
3	Were.you.ever.a.smoker. [character]	1. No 2. Yes, and I'm currently sm 3. Yes, and I stopped	55 (67.1%) 12 (14.6%) 15 (18.3%)		82 (100.0%)	0 (0.0%)
4	Year.when.first.started.smoking [integer]	Mean (sd) : 2003.3 (12) min < med < max: 1971 < 2004 < 2019 IQR (CV) : 16 (0)	17 distinct values		26 (31.7%)	56 (68.3%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
5	Year.when.stopped.smoking [integer]	Mean (sd) : 2014.8 (5.8) min < med < max: 2001 < 2017 < 2020 IQR (CV) : 4 (0)	2001 : 1 (5.9%) 2005 : 2 (11.8%) 2013 : 1 (5.9%) 2014 : 2 (11.8%) 2015 : 1 (5.9%) 2016 : 1 (5.9%) 2017 : 2 (11.8%) 2018 : 3 (17.6%) 2020 : 4 (23.5%)		17 (20.7%)	65 (79.3%)
6	When.did.you.start.looking.for.an.internship [character]	1. (Empty string) 2. 11/2/2020 3. 9/1/2020 4. 10/1/2020 5. 11/1/2020 6. 1/1/2020 7. 10/1/2019 8. 11/1/2019 9. 2/1/2020 10. 3/1/2020 [35 others]	18 (22.0%) 6 (7.3%) 5 (6.1%) 4 (4.9%) 4 (4.9%) 2 (2.4%) 2 (2.4%) 2 (2.4%) 2 (2.4%) 2 (2.4%) 35 (42.7%)		82 (100.0%)	0 (0.0%)
7	Sex [character]	1. (Empty string) 2. Female 3. Male	1 (1.2%) 23 (28.0%) 58 (70.7%)		82 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
8	When.did.you.stopped.looking.for.an.internship [character]	1. (Empty string) 2. 10/1/2020 3. 11/2/2020 4. 2/1/2020 5. 1/1/2020 6. 10/10/2020 7. 10/11/2019 8. 10/19/2020 9. 10/30/2020 10. 10/31/2020 [16 others]	54 (65.9%) 2 (2.4%) 2 (2.4%) 2 (2.4%) 1 (1.2%) 1 (1.2%) 1 (1.2%) 1 (1.2%) 1 (1.2%) 1 (1.2%) 16 (19.5%)		82 (100.0%)	0 (0.0%)
9	Have.you.found.an.internship. [character]	1. (Empty string) 2. No 3. Yes	7 (8.5%) 49 (59.8%) 26 (31.7%)		82 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
10	Education..background..pick.a.main.one.you.identify.with. [character]	1. Business, Management 2. Finance, Economy 3. Literature, History, Phil 4. Mathematics, Physics, Che 5. Medicine, Biology 6. Other	10 (12.2%) 9 (11.0%) 1 (1.2%) 49 (59.8%) 6 (7.3%) 7 (8.5%)		82 (100.0%)	0 (0.0%)
11	Years.of.education [integer]	Mean (sd) : 18 (4.2) min < med < max: 4 < 18 < 25 IQR (CV) : 3 (0.2)	15 distinct values		81 (98.8%)	1 (1.2%)
12	Do.you.have.children. [character]	1. No 2. Yes	58 (70.7%) 24 (29.3%)		82 (100.0%)	0 (0.0%)
13	Cohort [character]	1. A15 2. A17 3. A18 4. A19 5. A20 6. S18 7. S19 8. S20	1 (1.2%) 3 (3.7%) 2 (2.4%) 12 (14.6%) 37 (45.1%) 1 (1.2%) 3 (3.7%) 23 (28.0%)		82 (100.0%)	0 (0.0%)

Generated by summarytools 0.9.8 (R version 4.0.4)
2021-03-19

` # Data Preprocessing

Shorter labels for education for better visualization

```
table(dsti_sample$Education..background..pick.a.main.one.you.identify.with.)
```

```
##
##                                          Business, Management
##                                                            10
##                                              Finance, Economy
##                                                             9
##                                Literature, History, Philosophy
##                                                             1
## Mathematics, Physics, Chemistry, Computer Science, Statistics
##                                                            49
##                                            Medicine, Biology
##                                                             6
##                                                         Other
##                                                             7
```

```
edu_labels <- tibble(
  `Education..background..pick.a.main.one.you.identify.with.` =
    c("Business, Management", "Finance, Economy",
      "Literature, History, Philosophy",
      "Mathematics, Physics, Chemistry, Computer Science, Statistics",
      "Medicine, Biology", "Other"),
  Education = c("mgmt", "fin", "lit", "math", "bio", "oth")
  )

dsti_sample <- dsti_sample %>%
  inner_join(edu_labels, by = "Education..background..pick.a.main.one.you.identify.with.") %>%
  mutate(Education = factor(Education))
```

```
table(dsti_sample$education)
```

```
## < table of extent 0 >
```

```
str(dsti_sample)
```

```
## 'data.frame':    82 obs. of  14 variables:
##  $ Timestamp                                                 : chr  "11/2/2020 16:59:23" "11/2/2020 16:59:31"
"11/2/2020 16:59:38" "11/2/2020 16:59:45" ...
##  $ Year.of.birth                                             : int  1992 1993 1990 1986 1993 1992 1995 1992 199
3 1989 ...
##  $ Were.you.ever.a.smoker.                                   : chr  "No" "Yes, and I'm currently smoking" "No"
"No" ...
##  $ Year.when.first.started.smoking                           : int  NA 2011 NA NA NA 2019 NA 2010 2013 NA ...
##  $ Year.when.stopped.smoking                                 : int  NA NA NA NA NA NA NA NA 2018 NA ...
##  $ When.did.you.start.looking.for.an.internship              : chr  "11/2/2020" "10/19/2020" "3/1/2021" "9/1/20
20" ...
##  $ Sex                                                       : chr  "Male" "Female" "Female" "Male" ...
##  $ When.did.you.stopped.looking.for.an.internship            : chr  "" "" "" "10/31/2020" ...
##  $ Have.you.found.an.internship.                             : chr  "No" "No" "No" "Yes" ...
##  $ Education..background..pick.a.main.one.you.identify.with.: chr  "Mathematics, Physics, Chemistry, Computer
Science, Statistics" "Mathematics, Physics, Chemistry, Computer Science, Statistics" "Mathematics, Physics, Chemi
stry, Computer Science, Statistics" "Medicine, Biology" ...
##  $ Years.of.education                                        : int  20 17 17 22 16 18 16 14 17 18 ...
##  $ Do.you.have.children.                                     : chr  "No" "No" "No" "Yes" ...
##  $ Cohort                                                    : chr  "A20" "A20" "A20" "S20" ...
##  $ Education                                                 : Factor w/ 6 levels "bio","fin","lit",..: 4 4 4 1
4 3 2 2 2 4 ...
```

Selection of field of interest In this project , we are not interested by field related to smoking

```
dsti_sample = dsti_sample %>% select(Timestamp, Year.of.birth, When.did.you.start.looking.for.an.internship, Sex,
When.did.you.stopped.looking.for.an.internship, Have.you.found.an.internship., Education,
Years.of.education,Do.you.have.children.,Cohort)
```

```
str(dsti_sample)
```

```
## 'data.frame':    82 obs. of  10 variables:
##  $ Timestamp                                     : chr  "11/2/2020 16:59:23" "11/2/2020 16:59:31" "11/2/2020 1
6:59:38" "11/2/2020 16:59:45" ...
##  $ Year.of.birth                                 : int  1992 1993 1990 1986 1993 1992 1995 1992 1993 1989 ...
##  $ When.did.you.start.looking.for.an.internship  : chr  "11/2/2020" "10/19/2020" "3/1/2021" "9/1/2020" ...
##  $ Sex                                           : chr  "Male" "Female" "Female" "Male" ...
##  $ When.did.you.stopped.looking.for.an.internship: chr  "" "" "" "10/31/2020" ...
##  $ Have.you.found.an.internship.                 : chr  "No" "No" "No" "Yes" ...
##  $ Education                                     : Factor w/ 6 levels "bio","fin","lit",..: 4 4 4 1 4 3 2 2 2
4 ...
##  $ Years.of.education                            : int  20 17 17 22 16 18 16 14 17 18 ...
##  $ Do.you.have.children.                         : chr  "No" "No" "No" "Yes" ...
##  $ Cohort                                        : chr  "A20" "A20" "A20" "S20" ...
```

Formatting some fields in the proper format and creation of calculed variale “age”

```
as_date <- function(x) as.Date(x, format = "%m/%d/%Y")

dsti_sample <-  dsti_sample %>%
  mutate(Timestamp = as_date(Timestamp),
         When.did.you.start.looking.for.an.internship = as_date(When.did.you.start.looking.for.an.internship),
         When.did.you.stopped.looking.for.an.internship = as_date(When.did.you.stopped.looking.for.an.internshi
p),
         Have.you.found.an.internship. = as.factor(Have.you.found.an.internship.),

         Sex = as.factor(Sex), Have.you.found.an.internship. = as.factor(Have.you.found.an.internship.),
         Cohort = as.factor(Cohort),
         Do.you.have.children. = as.factor(Do.you.have.children.),
         Education = as.factor(Education),

         Age = (year(Timestamp) - Year.of.birth)

  )
```

Data filter based on variable “Have.you.found.an.internship.” All rows for which variable “Have.you.found.an.internship.” is blank or NA will be ignored Format variable “Have.you.found.an.internship.” in proper format (logical)

```
table(dsti_sample$Have.you.found.an.internship.)
```

```
##
##      No Yes
##   7  49  26
```

```
#dsti_sample = dsti_sample %>% filter(is.na(Have.you.found.an.internship.) | Have.you.found.an.internship.!= "")
dsti_sample = dsti_sample %>% filter(Have.you.found.an.internship.!= "")
dsti_sample$Have.you.found.an.internship. = ifelse(dsti_sample$Have.you.found.an.internship.=="Yes",1,0)

str(dsti_sample)
```

```
## 'data.frame':    75 obs. of  11 variables:
##  $ Timestamp                                      : Date, format: "2020-11-02" "2020-11-02" ...
##  $ Year.of.birth                                  : int  1992 1993 1990 1986 1993 1992 1995 1992 1993 1989 ...
##  $ When.did.you.start.looking.for.an.internship  : Date, format: "2020-11-02" "2020-10-19" ...
##  $ Sex                                            : Factor w/ 3 levels "","Female","Male": 3 2 2 3 3 2 3 3 3 3
...
##  $ When.did.you.stopped.looking.for.an.internship: Date, format: NA NA ...
##  $ Have.you.found.an.internship.                  : num  0 0 0 1 1 1 1 1 0 0 ...
##  $ Education                                      : Factor w/ 6 levels "bio","fin","lit",..: 4 4 4 1 4 3 2 2 2
4 ...
##  $ Years.of.education                             : int  20 17 17 22 16 18 16 14 17 18 ...
##  $ Do.you.have.children.                          : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ Cohort                                         : Factor w/ 8 levels "A15","A17","A18",..: 5 5 5 8 3 4 4 5 5
8 ...
##  $ Age                                            : num  28 27 30 34 27 28 25 28 27 31 ...
```

```
table(dsti_sample$Have.you.found.an.internship.)
```

```
##
##  0  1
## 49 26
```

```
head(dsti_sample)
```

	Timestamp <date>	Year.of.birth <int>	When.did.you.start.looking.for.an.internship <date>	Sex <fct>
1	2020-11-02	1992	2020-11-02	Male
2	2020-11-02	1993	2020-10-19	Female
3	2020-11-02	1990	2021-03-01	Female
4	2020-11-02	1986	2020-09-01	Male
5	2020-11-02	1993	2018-11-01	Male
6	2020-11-02	1992	2020-03-01	Female

6 rows | 1-5 of 12 columns

Creation new variable “waiting_time” to estimate waiting duration before getting internship Filter dataset based on waiting_time >=0

```
dsti_sample <-  dsti_sample %>%
  mutate(Waiting_Time = ifelse(!is.na(dsti_sample$When.did.you.stopped.looking.for.an.internship), difftime(When.
did.you.stopped.looking.for.an.internship, When.did.you.start.looking.for.an.internship, units = "days"),difftime
(dsti_sample$Timestamp, When.did.you.start.looking.for.an.internship, units = "days")  ))


dsti_sample = dsti_sample %>% filter(dsti_sample$Waiting_Time >=0)


str(dsti_sample)
```

```
## 'data.frame':    58 obs. of  12 variables:
##  $ Timestamp                                     : Date, format: "2020-11-02" "2020-11-02" ...
##  $ Year.of.birth                                 : int  1992 1993 1986 1993 1992 1995 1989 1982 1997 1970 ...
##  $ When.did.you.start.looking.for.an.internship  : Date, format: "2020-11-02" "2020-10-19" ...
##  $ Sex                                           : Factor w/ 3 levels "","Female","Male": 3 2 3 3 2 3 3 2 2 3
...
##  $ When.did.you.stopped.looking.for.an.internship: Date, format: NA NA ...
##  $ Have.you.found.an.internship.                 : num  0 0 1 1 1 1 0 0 0 1 ...
##  $ Education                                     : Factor w/ 6 levels "bio","fin","lit",..: 4 4 1 4 3 2 4 4 4
4 ...
##  $ Years.of.education                            : int  20 17 22 16 18 16 18 20 20 20 ...
##  $ Do.you.have.children.                         : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 1 1 1 2 ...
##  $ Cohort                                        : Factor w/ 8 levels "A15","A17","A18",..: 5 5 8 3 4 4 8 5 5
8 ...
##  $ Age                                           : num  28 27 34 27 28 25 31 38 23 50 ...
##  $ Waiting_Time                                  : num  0 14 60 60 122 123 0 4 0 73 ...
```

```
head(dsti_sample)
```

	Timestamp <date>	Year.of.birth <int>	When.did.you.start.looking.for.an.internship <date>	Sex <fct>
1	2020-11-02	1992	2020-11-02	Male
2	2020-11-02	1993	2020-10-19	Female
3	2020-11-02	1986	2020-09-01	Male
4	2020-11-02	1993	2018-11-01	Male
5	2020-11-02	1992	2020-03-01	Female
6	2020-11-02	1995	2019-10-01	Male

6 rows | 1-5 of 13 columns

Coerce variabe “Have.you.found.an.internship.” in logical format

```r
library(summarytools)
library(readr)


dsti_sample$Have.you.found.an.internship. = as.logical(dsti_sample$Have.you.found.an.internship.)

str(dsti_sample)
```

```
## 'data.frame':    58 obs. of  12 variables:
##  $ Timestamp                                       : Date, format: "2020-11-02" "2020-11-02" ...
##  $ Year.of.birth                                   : int  1992 1993 1986 1993 1992 1995 1989 1982 1997 1970 ...
##  $ When.did.you.start.looking.for.an.internship  : Date, format: "2020-11-02" "2020-10-19" ...
##  $ Sex                                             : Factor w/ 3 levels "","Female","Male": 3 2 3 3 2 3 3 2 2 3
...
##  $ When.did.you.stopped.looking.for.an.internship: Date, format: NA NA ...
##  $ Have.you.found.an.internship.                   : logi  FALSE FALSE TRUE TRUE TRUE TRUE ...
##  $ Education                                       : Factor w/ 6 levels "bio","fin","lit",..: 4 4 1 4 3 2 4 4 4
4 ...
##  $ Years.of.education                              : int  20 17 22 16 18 16 18 20 20 20 ...
##  $ Do.you.have.children.                           : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 1 1 1 2 ...
##  $ Cohort                                          : Factor w/ 8 levels "A15","A17","A18",..: 5 5 8 3 4 4 8 5 5
8 ...
##  $ Age                                             : num  28 27 34 27 28 25 31 38 23 50 ...
##  $ Waiting_Time                                    : num  0 14 60 60 122 123 0 4 0 73 ...
```

```r
#table(dsti_sample$Have.you.found.an.internship.)
```

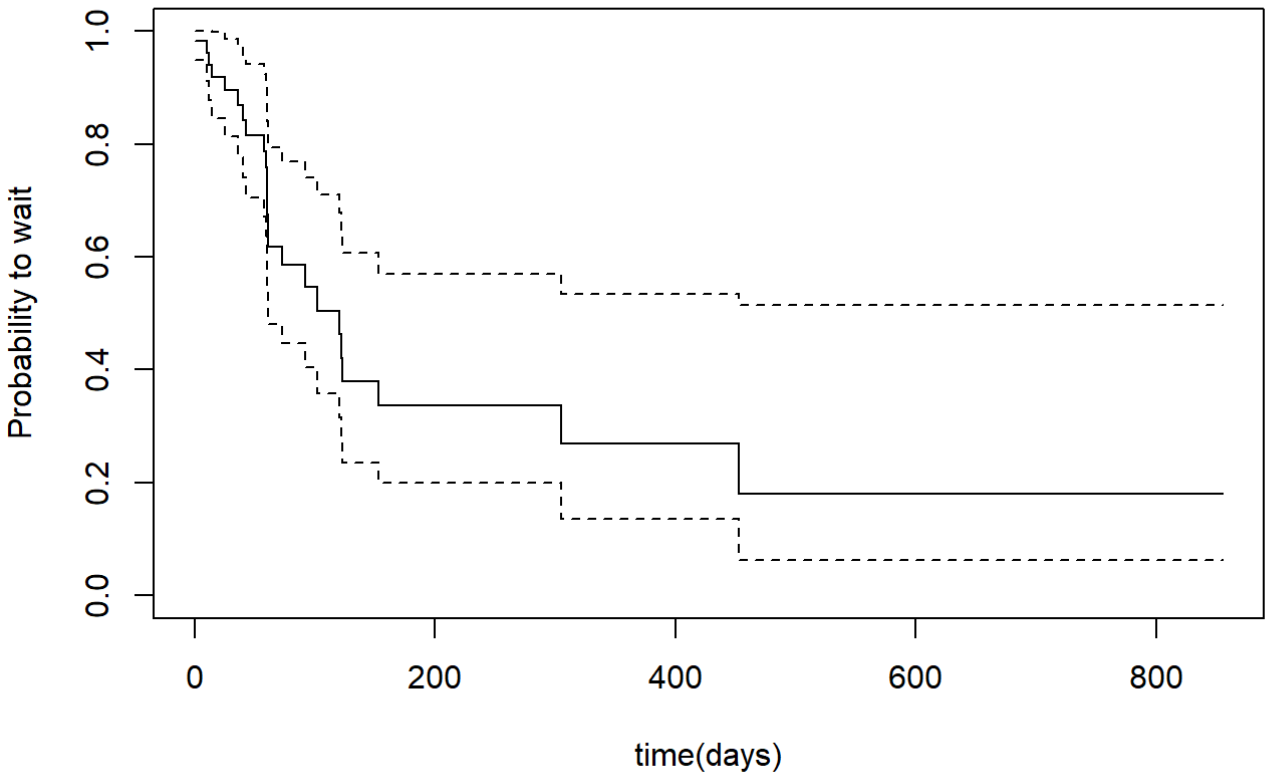
Model building

```r
fit <- survfit(Surv(Waiting_Time, Have.you.found.an.internship.) ~ 1, data = dsti_sample)

summary(fit)
```

```
## Call: survfit(formula = Surv(Waiting_Time, Have.you.found.an.internship.) ~
##     1, data = dsti_sample)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     0     58       1    0.983  0.0171       0.9498        1.000
##    10     47       1    0.962  0.0266       0.9111        1.000
##    12     45       1    0.940  0.0335       0.8770        1.000
##    14     44       1    0.919  0.0390       0.8458        0.999
##    25     40       1    0.896  0.0443       0.8134        0.987
##    36     34       1    0.870  0.0502       0.7767        0.974
##    40     32       1    0.843  0.0555       0.7405        0.959
##    43     31       1    0.815  0.0600       0.7059        0.942
##    58     29       1    0.787  0.0642       0.6710        0.924
##    59     28       1    0.759  0.0678       0.6373        0.904
##    60     27       3    0.675  0.0757       0.5416        0.841
##    61     24       2    0.619  0.0792       0.4813        0.795
##    73     19       1    0.586  0.0814       0.4463        0.769
##    92     15       1    0.547  0.0849       0.4035        0.741
##   102     13       1    0.505  0.0881       0.3586        0.711
##   120     12       1    0.463  0.0903       0.3158        0.678
##   122     11       1    0.421  0.0914       0.2749        0.644
##   123     10       1    0.379  0.0914       0.2359        0.608
##   153      9       1    0.337  0.0904       0.1988        0.570
##   305      5       1    0.269  0.0941       0.1357        0.534
##   453      3       1    0.180  0.0965       0.0626        0.515
```

```
plot(fit,
     xlab = "time(days)",
     ylab = "Probability to wait", )
```




```
fit
```

```
## Call: survfit(formula = Surv(Waiting_Time, Have.you.found.an.internship.) ~
##     1, data = dsti_sample)
##
##       n  events  median 0.95LCL 0.95UCL
##      58      24     120      61      NA
```

Question 1 : How long does it take to obtain an internship?

It takes on average 120 days to obtain an internship

Waiting time regarding Cohort

```
# Cohort

#cohort = survdiff(Surv(Waiting_Time, Have.you.found.an.internship.)  ~ Cohort, data = dsti_sample)
cohort =  survdiff(Surv(Waiting_Time, Have.you.found.an.internship.) ~ Cohort, data = dsti_sample)


summary(cohort)
```

```
##       Length Class  Mode
## n      7     table  numeric
## obs    7     -none- numeric
## exp    7     -none- numeric
## var   49     -none- numeric
## chisq  1     -none- numeric
## call   3     -none- call
```

```
cohort
```

```
## Call:
## survdiff(formula = Surv(Waiting_Time, Have.you.found.an.internship.) ~
##     Cohort, data = dsti_sample)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## Cohort=A15  1        1    0.108  7.33e+00  7.492591
## Cohort=A17  2        1    1.619  2.37e-01  0.276382
## Cohort=A18  2        2    0.400  6.40e+00  6.755734
## Cohort=A19 11       10    6.298  2.18e+00  3.082952
```

```
## Cohort=A20 19        1    6.723  4.87e+00  6.999634
## Cohort=S19  3        3    2.829  1.04e-02  0.012350
## Cohort=S20 20        6    6.022  8.15e-05  0.000122
##
##  Chisq= 22  on 6 degrees of freedom, p= 0.001
```

Question 2 : Is the waiting time changing between cohorts?

Yes, waiting time is changing between cohorts The p-value of the log-rank test is : 0.001 (< 0.05) That means we can reject the null hypothesis stating there is no difference in waiting time between cohort (at level 5%)

The smallest waiting time appends in cohort Cohort A15 The biggest waiting time appends in cohort S20

Waiting time regarding Educational

```
# Educational Background

#cohort = survdiff(Surv(Waiting_Time, Have.you.found.an.internship.)  ~ Cohort, data = dsti_sample)
#Educational =  survfit(Surv(Waiting_Time, Have.you.found.an.internship.) ~ Education, data = dsti_sample)

Educational =  survdiff(Surv(Waiting_Time, Have.you.found.an.internship.) ~ Education, data = dsti_sample)

summary(Educational)
```

```
##       Length Class  Mode
## n      6     table  numeric
## obs    6     -none- numeric
## exp    6     -none- numeric
## var   36     -none- numeric
## chisq  1     -none- numeric
## call   3     -none- call
```

```
Educational
```

```
## Call:
## survdiff(formula = Surv(Waiting_Time, Have.you.found.an.internship.) ~
##     Education, data = dsti_sample)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## Education=bio    6        2    3.533    0.6652    0.8077
## Education=fin    5        1    1.454    0.1419    0.1556
## Education=lit    1        1    0.836    0.0320    0.0342
## Education=math 36       17   14.359    0.4856    1.2581
## Education=mgmt   8        3    2.661    0.0432    0.0507
## Education=oth    2        0    1.156    1.1561    1.2477
##
##  Chisq= 2.6  on 5 degrees of freedom, p= 0.8
```

Question 3 : Does the educational background have an impact

No, waiting time does not change between Educational type The p-value of the log-rank test is : 0.08 (> 0.05) That means we cannot reject the null hypothesis stating there is no difference in waiting time between educational type (at level 5%)

Let's try to check that with a statistic test We adopt Fisher exact test given the small size of data

#Fisher test for independance between waiting time and Cohort

```
fisher.test(table(dsti_sample$Cohort, dsti_sample$Have.you.found.an.internship. ))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dsti_sample$Cohort, dsti_sample$Have.you.found.an.internship.)
## p-value = 5.781e-07
## alternative hypothesis: two.sided
```

At level of 5% ,pvalue(5.781e-07) is smaller than 0.05 Hence, we can reject the null hypothesis that the waiting time is independent of Cohort

#Fisher test for independance between waiting time and Educational

```
fisher.test(table(dsti_sample$Education, dsti_sample$Have.you.found.an.internship. ))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dsti_sample$Education, dsti_sample$Have.you.found.an.internship.)
## p-value = 0.6012
## alternative hypothesis: two.sided
```

At level of 5% ,pvalue(0.6012) is bigger than 0.05 Hence, we cannot reject the null hypothesis that the waiting time is independent of Education

Question 4 : Can you build a predictive model to identify students at high risk of a long search?

Yes, we can build a predictive model to identify students at high risk of a long search We can use either: a binary classifocation model with dependent variable : Have.you.found.an.internship or regression model based on dependent variable : Waiting time

#First model : a binary classification one

```
modClass = glm(Have.you.found.an.internship. ~ ., data=dsti_sample)
summary(modClass)
```

```
##
## Call:
## glm(formula = Have.you.found.an.internship. ~ ., data = dsti_sample)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.20791  -0.05612   0.00000   0.04767   0.25091
##
## Coefficients: (2 not defined because of singularities)
##                                                    Estimate Std. Error t value
```

```
## (Intercept)                                     4.150e+02  2.235e+03   0.186
## Timestamp                                      -2.262e-02  1.204e-01  -0.188
## Year.of.birth                                   3.214e-03  9.550e-03   0.337
## When.did.you.start.looking.for.an.internship    4.228e-04  4.675e-04   0.904
## SexMale                                        -3.076e-01  1.156e-01  -2.660
## When.did.you.stopped.looking.for.an.internship -4.247e-04  4.684e-04  -0.907
## Educationfin                                    2.072e-01  2.452e-01   0.845
## Educationlit                                   -8.950e-02  2.339e-01  -0.383
## Educationmath                                   1.511e-01  1.748e-01   0.865
## Educationmgmt                                  -1.435e-02  1.918e-01  -0.075
## Educationoth                                   -3.803e-01  2.579e-01  -1.474
## Years.of.education                             -6.785e-04  1.069e-02  -0.063
## Do.you.have.children.Yes                        2.081e-01  1.539e-01   1.352
## CohortA17                                      -2.178e-01  7.083e-01  -0.308
## CohortA18                                      -2.879e-01  2.978e-01  -0.967
## CohortA19                                      -2.247e-01  2.383e-01  -0.943
## CohortA20                                      -9.101e-01  2.735e-01  -3.328
## CohortS19                                      -2.035e-01  2.781e-01  -0.732
## CohortS20                                      -2.984e-01  2.750e-01  -1.085
## Age                                                    NA         NA      NA
## Waiting_Time                                           NA         NA      NA
##                                                Pr(>|t|)
## (Intercept)                                      0.8573
## Timestamp                                        0.8557
## Year.of.birth                                    0.7451
## When.did.you.start.looking.for.an.internship     0.3922
## SexMale                                          0.0288 *
## When.did.you.stopped.looking.for.an.internship   0.3911
## Educationfin                                     0.4226
## Educationlit                                     0.7119
## Educationmath                                    0.4124
## Educationmgmt                                    0.9422
## Educationoth                                     0.1786
## Years.of.education                               0.9510
## Do.you.have.children.Yes                         0.2133
## CohortA17                                        0.7663
## CohortA18                                        0.3620
```

```
## CohortA19                                          0.3733
## CohortA20                                          0.0104 *
## CohortS19                                          0.4853
## CohortS20                                          0.3095
## Age                                                    NA
## Waiting_Time                                           NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.03136433)
##
##     Null deviance: 2.66667  on 26  degrees of freedom
## Residual deviance: 0.25091  on  8  degrees of freedom
##   (31 observations deleted due to missingness)
## AIC: -9.6963
##
## Number of Fisher Scoring iterations: 2
```

Second model : Regression model one

```
modRegr = lm(Waiting_Time ~ ., data=dsti_sample)
summary(modRegr)
```

```
##
## Call:
## lm(formula = Waiting_Time ~ ., data = dsti_sample)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -4.110e-12 -1.066e-12  0.000e+00  8.211e-13  5.288e-12
##
## Coefficients: (1 not defined because of singularities)
##                                                  Estimate Std. Error    t value
## (Intercept)                                     8.350e-08  4.804e-08  1.738e+00
## Timestamp                                      -4.531e-12  2.589e-12 -1.750e+00
```

```
## Year.of.birth                                    3.148e-13  2.063e-13  1.526e+00
## When.did.you.start.looking.for.an.internship   -1.000e+00  1.053e-14 -9.497e+13
## SexMale                                          5.618e-12  3.405e-12  1.650e+00
## When.did.you.stopped.looking.for.an.internship  1.000e+00  1.055e-14  9.477e+13
## Have.you.found.an.internship.TRUE                7.913e-12  7.585e-12  1.043e+00
## Educationfin                                    -1.897e-11  5.489e-12 -3.457e+00
## Educationlit                                    -9.833e-12  5.064e-12 -1.942e+00
## Educationmath                                   -1.300e-11  3.920e-12 -3.317e+00
## Educationmgmt                                   -1.201e-11  4.116e-12 -2.918e+00
## Educationoth                                    -6.895e-12  6.240e-12 -1.105e+00
## Years.of.education                               1.747e-13  2.295e-13  7.610e-01
## Do.you.have.children.Yes                         2.476e-12  3.660e-12  6.760e-01
## CohortA17                                        2.518e-11  1.529e-11  1.647e+00
## CohortA18                                       -1.688e-12  6.752e-12 -2.500e-01
## CohortA19                                        1.653e-12  5.388e-12  3.070e-01
## CohortA20                                        4.948e-12  9.059e-12  5.460e-01
## CohortS19                                        3.902e-12  6.162e-12  6.330e-01
## CohortS20                                        4.691e-12  6.318e-12  7.420e-01
## Age                                                     NA         NA         NA
##                                                Pr(>|t|)
## (Intercept)                                      0.1258
## Timestamp                                        0.1236
## Year.of.birth                                    0.1709
## When.did.you.start.looking.for.an.internship     <2e-16 ***
## SexMale                                          0.1430
## When.did.you.stopped.looking.for.an.internship   <2e-16 ***
## Have.you.found.an.internship.TRUE                0.3315
## Educationfin                                     0.0106 *
## Educationlit                                     0.0933 .
## Educationmath                                    0.0128 *
## Educationmgmt                                    0.0224 *
## Educationoth                                     0.3057
## Years.of.education                               0.4713
## Do.you.have.children.Yes                         0.5205
## CohortA17                                        0.1435
## CohortA18                                        0.8098
## CohortA19                                        0.7680
```

```
## CohortA20                                          0.6019
## CohortS19                                          0.5467
## CohortS20                                          0.4820
## Age                                                    NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.799e-12 on 7 degrees of freedom
##   (31 observations deleted due to missingness)
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.151e+27 on 19 and 7 DF,  p-value: < 2.2e-16
```