Survival Analysis Project

Presented by :

Christopher Sassine

Abdoul Aziz Moussa Harouna

Contents

1. Introduction

Our data is made of 13 variables/features with 82 rows of observation. The idea of this project is to predict the time to internship of a student in DSTI and to evaluate what are the variables that affect it, given the features that we are provided with in this dataset.

The features that we are interested in for this analysis are the Duration,cohort, status(did the student find an internship?), age, education background and having or not having children.

2. Data preparation

The challenge in our data is that the majority of it has missing values and the number of observations is small (82 rows). That being said, we cannot delete all fragile observations so we have to fill as much as we can in a scientific manner.

2.1 Duration

Duration is the variable that is essential for Survival Analysis, it is the time for which the experiment is running, where the involved question in place is being observed. In our case the Duration is simply the time between the start and the end of the search for an internship. We notice that there is a lot of unfilled information about the end of search for an internship. Moreover, we realize that the students that did not fill their end of search for an internship did not actually find an internship in most of the cases. This phenomenon is called right censoring.

Often, right censoring can be negligible, and the software will figure out a Mathematical way to deal with them without biasing the data. But in our case, we need to find a new way because the data is fragile. We created a way to fill them by applying our personal knowledge.

For the right censored duration data we decided to use the expected graduation date for each cohort as a maximum tracking period for each observation.

For students that joined DSTI in a given year X. We set the 1st November of the following year as the final day of search for an internship and we note by t_A(X) such that:

t_A(X) = 11-01-20(X+1)

For students that joined DSTI in Summer in a given year X. We set the 1st April of the following year as the final day of search for an internship and we not it by t_S(X) such that:

t_S(X) = 11-01-20(X+1)

After filling **'when did you stopped looking for an internship'** with the given values in the case where the students said No to 'Have you found an internship', we delete everyone else that do not filled 'when did you stopped looking for an internship?' just because there is no other way to fill them and our future models cannot work with these missing values.
We are left with 64 observations after this process and we have the start and end dates for each observation, this means we can finally compute our duration.

After computing all durations we plot the boxplot to see the distribution of duration.

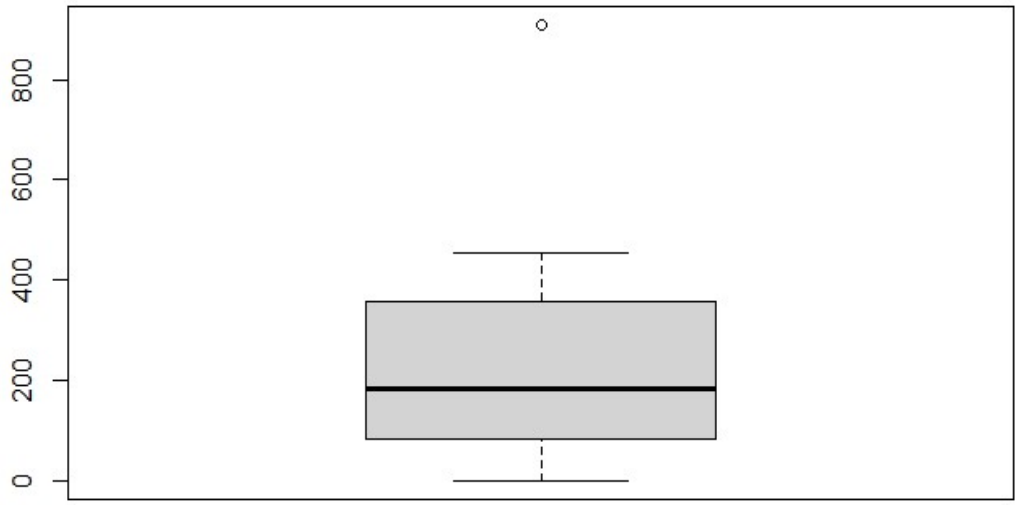

Figure 1 Boxplot on the experiment on duration

We realize that we have a student with an extreme value with a duration of 910 days which means that he started searching for an internship 6 months before joining the cohort. Thus, We remove him because he is very different from the rest of our population.
We are left with 63 observations.

2.2 Cohort

We transform the cohort feature as a factor variable, and we plot the histogram of cohort to see its descriptive knowledge:

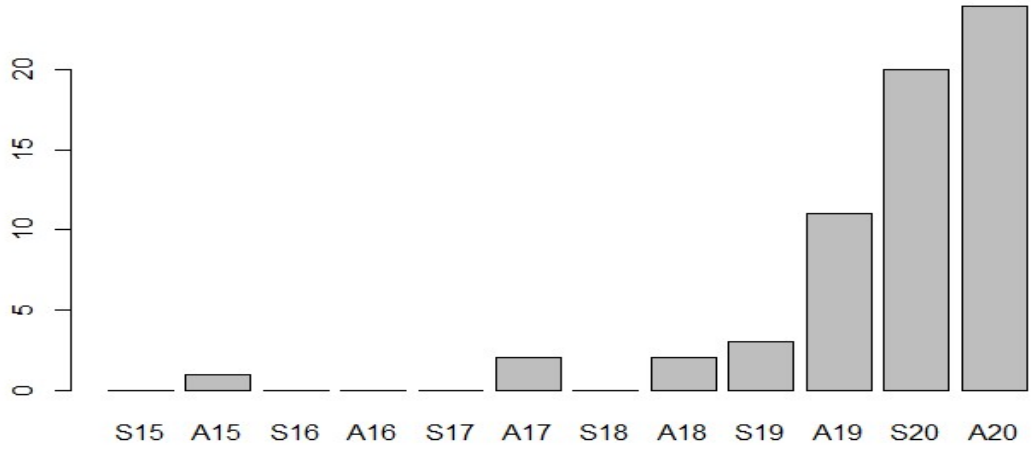

Figure 2 Histogramme for Cohort

We realize that the people that filled this form are mostly A20 and S20 students followed by S19 and A19.

2.3 Age

In the dataset we do not have a measure for age. We will use the difference between Timestamp and year of birth and we create a new column called '**age**'. We will have all the students ages except for two who did not have a timestamp to compute the age. Fortunately, those two students filled their year of birth, we can just provide their ages manually by using the cohort(2020, for A20). We are left with the following distribution for

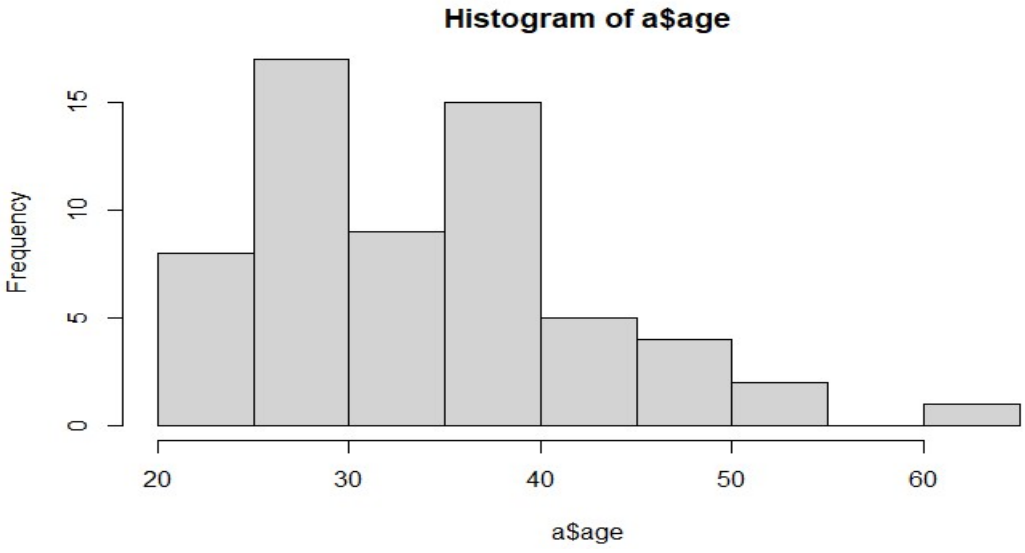

Figure 3: Histogramme for age

We can see that we have an outlier individual with an age of 65, we decide to not remove him because it does not have a big impact on our survival analysis results and because we are already left with 63 observations and cannot afford to lose more information.

2.4 Education

For education, the complete column is filled, and we have no missing value, we just abbreviated the education background such that it is visually better:

bio	fin	lit	math	mgmt	other	NA
6	7	1	39	8	2	0

Figure 4 Education background qualitative analysis

2.5 Do you have children?

This feature contains no missing values, thus we get the following results in a table:

No	Yes	NA
44	19	0

Figure 5: Having Children Qualitative analysis

3. Time to internship

To evaluate how long does it take for an individual to get an internship, we use the Kaplan meier nonparametric method which takes as input the status and the duration.

We used Kaplan meier in R and we got the following results:

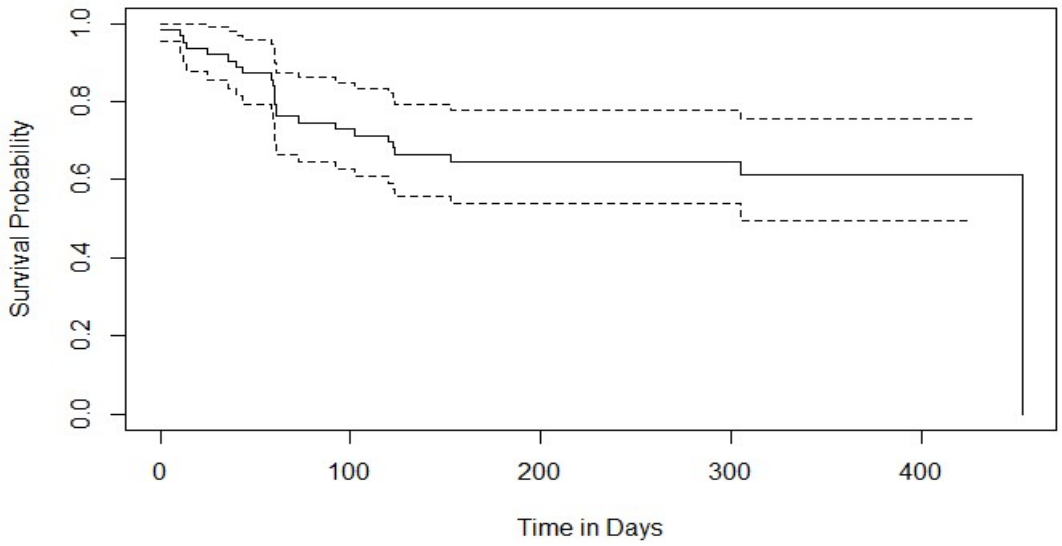

Figure 6: Kaplan Meiers Analysis for DSTI Survey Dataset

We get a median equal to 453 days with a confidence interval of [305, +inf[. It is the probability at which 50% of the population experienced the event of 'obtaining the internship'.The number of observations used is 63, the events observed is equal to 24 and the total number of censored observation is equal to 39.

Our big value for median (453 days) can be explained by the data cleaning that we did in the beginning. In fact, by replacing the 'when did you stopped looking for an internship?' for

students whose status is 0 (equivalent to did not find an internship) by their end of study date, we have given all those students a large duration with no event happening which results in a larger median than expected.

4. Variables impact and analysis

The variables that we want to investigate to see how they impact the time to internship are: Cohort, age, educational background and having or not having children.
In order to do that, we need to do the semi-parametric method named Cox proportional-hazards model. This model works by taking all the features that we want to evaluate with the status and duration and it will predict the coefficient of each feature by varying the value of a given feature and keeping the other ones constants. That is why we need to put all the features together in the same Cox model so that we get the accurate impact of each feature. By using the Cox proportional hazards model, we get the following results:

Variable	p-value
Age	0.09272
CohortA17	0.25206
CohortA18	0.75351
CohortA19	0.09328
CohortA20	0.00168
CohortS19	0.01656
CohortS20	0.00359
Fin	0.70608
Lit	0.55411
Math	0.68101
Bio	0.45326
Other	0.99798
'Do you have children?' Yes	0.02494

Figure 7: results of the Cox proportional Hazard model

The variables with a significant p-value are Cohort with the categories A20, S19 and S20 and 'do you have children' with the category yes.

Now that we know which variables impact the time to internship, we can check for their Cox coefficients to see in which direction they impact the survival analysis.

Variable	Coefficient	Exp(Coefficient)
CohortA20	-5.989	2.507e-03
CohortS19	-4.255	1.419e-02
CohortS20	-4.944	7.128e-03
'Do you have children?' Yes	1.391	4.018

Figure 8 Cox coefficients for 'Cohort' and 'do you have children?'

For the Cohort, we can see that all 3 significant cohort have a negative impact on the given time to internship, these cohort will wait longer to get an internship.

In the case where the variable 'Do you have children?' takes as category yes, the coefficient is positive, wish means the students with children expects to wait less time to get an internship.

6. Conclusion

We studied in this report the DSTI Survey dataset. We studied the factors that influence obtaining an internship among **cohort, age, educational background, having or not having children**. We finally concluded that **'Cohort'** membership and **'Having or not having children'** are the two more influential covariates. While the later has a positive influence in getting an internship, cohort is having a negative influence in this dataset. It can be explained by the fact that A20, S19 and S20 cohort represent the majority of the surveyed sample. At submission of the survey, most of the sample were at the beginning of their training at DSTI.

Q1: How many students participated in the interview	• **82** Observations were recorded
Q2: After data preparation, how many samples are usable for data analysis? How many samples were dropped (if any), and why?	• **63** Observation were usable. • **17** Observation were dropped because the initial time to event was not provided (our study is limited to right censored duration data) • **1** observation was dropped because of inconsistency about the experiment. He started searching for internship 6 months before the start of his cohort (A20)
Q3: How long does it take to obtain an internship? Please report the median time (with a confidence interval), total number of students at the baseline, the total number of events observed, and the total number of censored observations.	• **453** days is the median. It is the probability at which 50% of the population experienced the event of 'obtaining the internship'. • 305 is the lower bound of the confidence interval • The upper bound is [+inf] • **63** student at the baseline • **24** event were observed • **39** event were censored
Q4 : Of these variables, which ones have the most impact on the time to obtain an internship, and in which direction: cohort, age, educational background, having or not having children.	• **'Cohort'** membership and **'Having or not having children'** are the two more influential covariates. Cohort is in the opposite direction(negative) while have children is positive influence in finding an internship

Figure 9 Answer to project main questions

7. ANNEXE

R- Computational Script used for the analysis

````
---
title: "Survival Analysis Project"
author:
  - Christopher Sassine
  - Abdoul Aziz Moussa Harouna
date: "June 6, 2021"
output: html_document
---
```{r}
library(tidyverse)
library(lubridate)
library(broom)
library(survival)
library(ggplot2)

```
# Data Exploratory analysis
```{r}
raw <- read_csv("DSTI_survey.csv")
```
```{r}
t_A15='11/01/2016'
t_A17='11/01/2018'
t_A18='11/01/2019'
t_A19='11/01/2020'
t_A20='11/01/2021'
t_S18='04/01/2019'
t_S19='04/01/2020'
t_S20='04/01/2021'

raw$`When did you stopped looking for an internship`=ifelse(is.na(raw$`When did you stopped looking for an internship`) & raw$`Have you found an internship?`=='No' & raw$Cohort=='A15',t_A15,raw$`When did you stopped looking for an internship`)
raw$`When did you stopped looking for an internship`=ifelse(is.na(raw$`When did you stopped looking for an internship`) & raw$`Have you found an internship?`=='No' & raw$Cohort=='A17',t_A17,raw$`When did you stopped looking for an internship`)
raw$`When did you stopped looking for an internship`=ifelse(is.na(raw$`When did you stopped looking for an internship`) & raw$`Have you found an internship?`=='No' & raw$Cohort=='A18',t_A18,raw$`When did you stopped looking for an internship`)
raw$`When did you stopped looking for an internship`=ifelse(is.na(raw$`When did you stopped looking for an internship`) & raw$`Have you found an internship?`=='No' & raw$Cohort=='A19',t_A19,raw$`When did you stopped looking for an internship`)
raw$`When did you stopped looking for an internship`=ifelse(is.na(raw$`When did you stopped looking for an internship`) & raw$`Have you found an internship?`=='No' & raw$Cohort=='S18',t_S18,raw$`When did you stopped looking for an internship`)
raw$`When did you stopped looking for an internship`=ifelse(is.na(raw$`When did you stopped looking for an internship`) & raw$`Have you found an internship?`=='No' & raw$Cohort=='S19',t_S19,raw$`When did you stopped looking for an internship`)
raw$`When did you stopped looking for an internship`=ifelse(is.na(raw$`When did you stopped looking for an internship`) & raw$`Have you found an internship?`=='No' & raw$Cohort=='A20',t_A20,raw$`When did you stopped looking for an internship`)
raw$`When did you stopped looking for an internship`=ifelse(is.na(raw$`When did you stopped looking for an internship`) & raw$`Have you found an internship?`=='No' & raw$Cohort=='S20',t_S20,raw$`When did you stopped looking for an internship`)
```
```{r}
raw=raw %>% drop_na("When did you start looking for an internship")
```
```{r}
raw$`When did you stopped looking for an internship`=as.POSIXct(as.Date(raw$`When did you stopped looking for an internship`,"%m/%d/%Y"))
raw$`When did you start looking for an internship`=as.POSIXct(as.Date(raw$`When did you start looking for an internship`,"%m/%d/%Y"))
raw$Timestamp=as.POSIXct(as.Date(raw$Timestamp,"%d/%m/%Y %H:%M:%OS"))
Time=raw$`When did you stopped looking for an internship`-raw$`When did you start looking for an internship`

```
````

We clean and create the event vector 1 being the event of having an internship and 0 not having or censored input

```{r}
raw$`Have you found an internship?`<-ifelse(raw$`Have you found an internship?`=="Yes",1,0)
```
The New dataframe "a"" is created with time_to_event as our new duration vector
```{r}
raw=as.data.frame(raw)
a=cbind(raw,time_to_event=Time)

a$time_to_event=as.numeric(a$time_to_event)

#converting the time from seconds to days
a$time_to_event=a$time_to_event/86400
```

```{r}
filter(a, time_to_event==910)
```
this student started looking for an intership in May 2019 while he is from A20 (Which starts in Novembre 2020) Cohort. This observation is not consistent with the project.
we remove the observation.
```{r}

a=a[a$time_to_event!=910,]
boxplot(a$time_to_event)
```

Cleaning the cohort

```{r}
a <- a %>%
  mutate(
    Cohort = factor(Cohort,
                    levels = paste0(c("S", "A"), rep(15:20, each = 2))))
table(a$Cohort, useNA = "always")
```

Cleannig the Age

```{r}
a <-
  a %>%
  mutate(Timestamp = as.POSIXct(Timestamp, format = "%d/%m/%Y %H:%M:%OS"),
         age = year(Timestamp) - `Year of birth`)
```

as the students are both from A20 cohort we will set the year to 2020 to have the age of the students.
```{r}

a$age=ifelse(is.na(a$age), 2020- a$`Year of birth`,a$age )
summary(a$age,useNA='always')
```

Education
```{r}
table(a$`Education: background (pick a main one you identify with)`, useNA = "always")
```
 Here we rearrange the the Education related column to come up with shorter labels:
```{r}
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")
  )
```
```{r}
a <- a %>%
  inner_join(edu_labels, by = "Education: background (pick a main one you identify with)")  %>% mutate(education = factor(education))
table(a$education, useNA = "always")
```
Quality check on the column `Do you have children?`
```{r}
 print(table(a$`Do you have children?`,useNA ='always' ))
```

#Processing the categorical data
```{r}
a$Cohort=as.factor(a$Cohort)
a$`Do you have children?`=as.factor(a$`Do you have children?`)
a$education=as.factor(a$education)
a$age=as.numeric(a$age)

```
#Survival analysis
```{r}
KM=survfit(Surv(a$time_to_event,a$`Have you found an internship?`)~1)

plot(KM, xlab='Time in Days', ylab='Survival Probability')
print(KM)
```


#The cox proportional hazard model model
```{r}
#X=cbind(a$Cohort,a$education,a$age,a$`Do you have children?`)

#res.cox <- coxph(Surv(a$time_to_event, a$`Have you found an internship?`) ~ X, data =  a)


res.cox <- coxph(Surv(a$time_to_event, a$`Have you found an internship?`) ~ a$Cohort+a$education+a$age+a$`Do you have children?`, data =
a)
summary(res.cox)
```