Survival Analysis – Time to Internship

By: Samer Ismail, Dinesh Babu, Amenan Seydou

Data Cleaning and Preparation:

The data was very messy and needed lots of preparation, we did all cleaning steps in R:

First thing we have done is to rename the variables to shorter names and convert some variables into proper format like dates for instance, converting some categorical variables into factors as well, in addition of creating some groups for education and age.

Installing the required libraries:

```
#install.packages("lubridate")
#install.packages("broom")
#install.packages("survival")
#install.packages("survminer")

library(tidyverse)

library(lubridate)

library(broom)

library(survival)

library(survminer)
```

Data Import:

```
    setwd("C:/Users/Samer/Desktop/Survival_Analysis")
   data <- read.csv('DSTI_survey.csv',header = TRUE)

   # Data Exploraration
   head(data)
```

Data Cleaning and Preparation

```
# Rename variables
names(data)[1]  = "srv_date"      #TimeStamp
names(data)[2]  = "yob"           #Year of Birth
names(data)[3]  = "smkr"          #Were you ever a smoker?
names(data)[4]  = "y_str_smk"     #Year when first started smoking
names(data)[5]  = "y_stp_smk"     #Year when stopped smoking
names(data)[6]  = "date_str_srch" #When did you start looking for an internship
names(data)[7]  = "sex"           #Sex
names(data)[8]  = "date_stp_srch" #When did you stopped looking for an internship
names(data)[9]  = "fnd_intr"      #Have you found an internship?
names(data)[10] = "edu"           #Education: background (pick a main one you identify wi
```

```
th)
names(data)[11] = "yoe"           #Years of education
names(data)[12] = "children"      #Do you have children?
names(data)[13] = "cohort"        #Cohort

head(data)
```

Converting data to the correct format:

```
# Convert Year to Date
data[,2]           = year(ymd(data[,2], truncated = 2L))
data[,1]           = as_date(as.POSIXct(data[,1], format='%m/%d/%Y %H:%M'))
data$y_str_smk     = year(ymd(data$y_str_smk, truncated = 2L))
data$y_stp_smk     = year(ymd(data$y_stp_smk, truncated = 2L))
data$date_str_srch = as_date(as.POSIXct(data$date_str_srch, format='%m/%d/%Y'))
data$date_stp_srch = as_date(as.POSIXct(data$date_stp_srch, format='%m/%d/%Y'))

# Convert Categorical Variables into Factors
data$sex       = as.factor(data$sex)
#data$fnd_intr  = as.factor(data$fnd_intr)
data$children  = as.factor(data$children)
data$cohort    = as.factor(data$cohort)

# Turning smkr variable to binary with (yes,no)
data$smkr= ifelse(data$smkr != "No", "Yes", data$smkr)
data$smkr      = as.factor(data$smkr)

data$fnd_intr= ifelse(data$fnd_intr == "Yes", 1, 0)



# Cleaning up Education Variable
data$edu= ifelse(data$edu == "Mathematics, Physics, Chemistry, Computer Science, Statisti
cs", "math", data$edu)
data$edu= ifelse(data$edu == "Medicine, Biology", "bio"
, data$edu)
data$edu= ifelse(data$edu == "Literature, History, Philosophy", "lit"
, data$edu)
data$edu= ifelse(data$edu == "Finance, Economy", "fin"
, data$edu)
data$edu= ifelse(data$edu == "Business, Management", "mgm"
, data$edu)
data$edu= ifelse(data$edu == "Other", "oth"
, data$edu)

# Converting Education variable into factor
data$edu= as.factor(data$edu)

#Sanity Checks
head(data)
```

Creating censored Duration Variable

As we want to analyze the time to internship first we need the duration variable which will be the timestamp (srv_date) minus date when started search (date_str_srch), Noting that if the stop searching date is more that timestamp, we replaced it with the timestamp date instead, so in general, our end date for the study was the timestamp.

```
data$date_stp_srch= as.Date(ifelse(data$date_stp_srch > data$srv_date, data$srv_date, dat
a$date_stp_srch), origin = "1970-01-01")

# Creating Search Duration Variable

data$srch_dur=
ifelse(data$fnd_intr == "1",
 difftime(data$date_stp_srch, data$date_str_srch, units = "days"),
 ifelse(is.na(data$date_stp_srch),
 difftime(data$srv_date, data$date_str_srch, units = "days"),
 difftime(data$date_stp_srch, data$date_str_srch, units = "days")))

data$srch_dur
```

Elimination of observations are not suitable to enter our model

1. Remove students with no Start Date

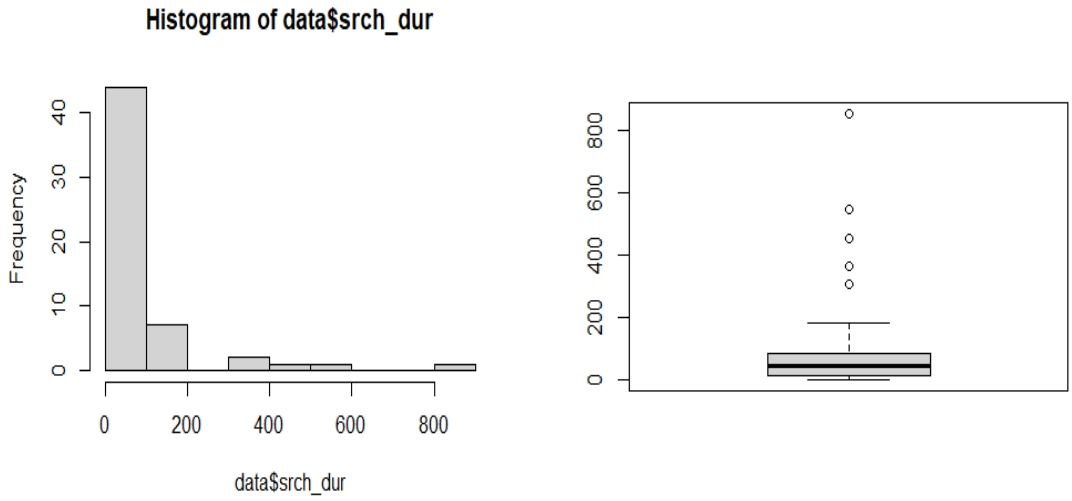
```
# Removing Students with no Start Date
data<-data[!(is.na(data$date_str_srch)), ]
data
```

2. Remove Students with Zero duration and found an Internship

```
#Students with Zero duration and found an Internship
data<-data[!(data$srch_dur=="0" & data$fnd_intr=="1"),]
data
```

3. Remove students with start date after Time stamp date, and this is based on their expectations not on reality, and this considered as Right cesoring which is out of scope of our analysis.

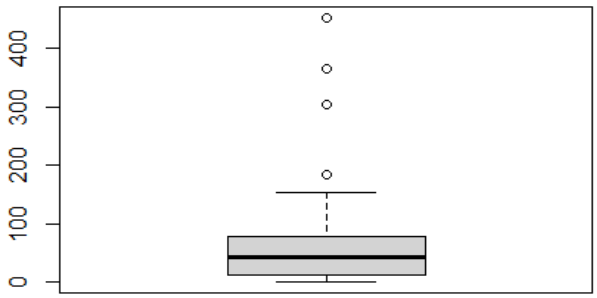
```
# Remove people with Internship Yes and stop search date NA
data<-data[!(data$date_str_srch > data$srv_date),]
data

summary(data$srch_dur)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   13.50   43.00   89.45   82.25  855.00

hist(data$srch_dur)

boxplot(data$srch_dur)
```



After plotting duration variable to have a better view, we can spot some outliers here; the best results was obtained when we removed observations more than 500 days of search duration, there was two observations deleted.

```
# Remove observations with seaerch duration >500
data<-data[!(data$srch_dur >= 500),]
boxplot(data$srch_dur)
```

Boxplot after removing outliers:



Q1. How many students partecipated in the interview?

82 students

Q2. After data preparation, how many samples are usable for data analysis? How many samples were dropped (if any), and why?

Raw Dataset included 82 observations, after processing, some of variables were dropped due to the following reasons:

1. Students who don't have start search date (18 students)

2. Students with Zero duration and found an Internship (1 student)
3. Remove students with start date after Time stamp date, and this is based on their expectations not on reality, and this considered as Right censoring which is out of scope of our analysis (7students)
4. 2 outliers removed, search duration >500 days (2 students)

The final number of students of samples usable for analysis is 54

Survival Function

Now since we defined our Censoring variable (fnd_intr) and Duration variable (srch_dur) we can start with the Survival Function

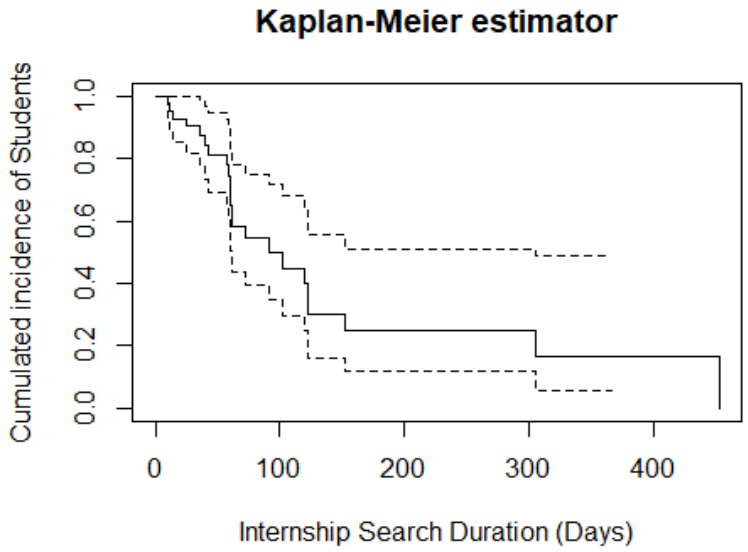
```
# Surviaval Material
srv_mat <- with(data, Surv(srch_dur ,fnd_intr))

 KM <- survfit(srv_mat ~ 1, data = data)
 KM
```

```
## Call: survfit(formula = srv_mat ~ 1, data = data)
##
##       n  events  median 0.95LCL 0.95UCL
##      54      23      92      61     305
```

```
 summary(KM)
```

```
## Call: survfit(formula = srv_mat ~ 1, data = data)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    10     43       1    0.977  0.0230       0.9327        1.000
##    12     41       1    0.953  0.0325       0.8913        1.000
##    14     40       1    0.929  0.0395       0.8549        1.000
##    25     36       1    0.903  0.0460       0.8174        0.998
##    36     30       1    0.873  0.0535       0.7745        0.984
##    40     28       1    0.842  0.0600       0.7323        0.968
##    43     27       1    0.811  0.0653       0.6923        0.950
##    58     25       1    0.778  0.0703       0.6521        0.929
##    59     24       1    0.746  0.0745       0.6133        0.907
##    60     23       3    0.649  0.0833       0.5043        0.834
##    61     20       2    0.584  0.0867       0.4364        0.781
##    73     15       1    0.545  0.0892       0.3953        0.751
##    92     12       1    0.499  0.0926       0.3473        0.718
##   102     10       1    0.450  0.0959       0.2959        0.683
##   120      9       1    0.400  0.0974       0.2478        0.644
##   122      8       1    0.350  0.0972       0.2028        0.603
##   123      7       1    0.300  0.0953       0.1607        0.559
##   153      6       1    0.250  0.0915       0.1217        0.512
##   305      3       1    0.166  0.0913       0.0568        0.488
##   453      1       1    0.000     NaN           NA           NA
```

```
 plot(KM,
     # fun = "F",
      main = "Kaplan-Meier estimator",
      xlab = "Internship Search Duration (Days)",
      ylab = "Cumulated incidence of Students")
```



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.

1. Median is 92 days, 3 months approximately. with Confidence Interval of [61, 305]
2. Total number of students at the baseline = 54 students
3. Total Number of Events Obsereved = 23
4. Total Number of Events Censored = 31

Q4. Of these variables,. which ones have the most impact on the time to obtain an internship, and in which # direction: 1.cohort, 2.age, 3.educational background,4.having or not having children.

1. Analyzing by Cohort groups

```
# Log Rank test - Cohort
surv_cohort_LR <- survdiff(Surv(srch_dur ,fnd_intr) ~ cohort, data = data)
surv_cohort_LR
```

```
## Call:
## survdiff(formula = Surv(srch_dur, fnd_intr) ~ cohort, data = data)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## cohort=A15  1        1   0.1004    8.0582     8.267
## cohort=A17  1        1   0.0233   41.0233    42.000
## cohort=A18  1        1   0.4186    0.8075     0.862
## cohort=A19 11       10   7.5655    0.7834     1.251
## cohort=A20 17        1   4.1380    2.3797     3.046
## cohort=S19  3        3   4.0848    0.2881     0.466
## cohort=S20 20        6   6.6694    0.0672     0.107
##
##  Chisq= 55.1  on 6 degrees of freedom, p= 4e-10
```

Cohort is Significant (very small p-value)

2. Analyzing by Age groups

We have to create the age variable, which will be the the timestamp (srv_date) minus year of birth then we will create age groups. Group interval is 9 years.

```
#Calculating Age

data$age = year(data$srv_date) - data$yob
data$age
```

```
##  [1] 28 27 34 27 28 25 31 38 23 50 42 29 34 40 26 37 27 24 39 32 41 36 54 27 43
## [26] 39 50 25 30 25 25 27 46 65 37 44 38 38 36 26 37 46 39 29 35 42 33 24 32 28
## [51] 53 31 28 23
```

```
summary(data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.00   27.00   33.50   34.69   39.00   65.00
```

Creating age groups:

```
# Creating Age groups

data$age_gr= ifelse(between(data$age,20,29), "20-29",
                    ifelse(between(data$age, 30, 39), "30-39",
                           ifelse(between(data$age, 40, 49), "40-49",
                                  ifelse(between(data$age, 50, 59), "50-59",
                                                ifelse(data$age >59 , ">=60", "Error"))))
)
data$age_gr
```

Age Impact

```
# Log Rank test - age
surv_age_LR <- survdiff(Surv(srch_dur ,fnd_intr) ~ age_gr, data = data)
surv_age_LR
```

```
## Call:
## survdiff(formula = Surv(srch_dur, fnd_intr) ~ age_gr, data = data)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## age_gr=>=60    1        1    0.519     0.447     0.481
## age_gr=20-29 21        6    9.722     1.425     2.761
## age_gr=30-39 20        6    6.925     0.123     0.190
## age_gr=40-49  8        6    4.288     0.683     0.882
## age_gr=50-59  4        4    1.547     3.892     4.390
##
##  Chisq= 7  on 4 degrees of freedom, p= 0.1
```

3. Education Impact :

```
# Log Rank test - Education
surv_edu_LR <- survdiff(Surv(srch_dur ,fnd_intr) ~ edu, data = data)
surv_edu_LR
```

```
## Call:
## survdiff(formula = Surv(srch_dur, fnd_intr) ~ edu, data = data)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## edu=bio   5        2     2.27  3.32e-02  3.82e-02
## edu=fin   5        1     1.67  2.66e-01  2.99e-01
## edu=lit   1        1     1.00  2.22e-05  2.43e-05
## edu=math 34       16    14.80  9.75e-02  3.08e-01
## edu=mgm   7        3     3.16  7.65e-03  1.31e-02
## edu=oth   2        0     0.10  1.00e-01  1.03e-01
##
##  Chisq= 0.5  on 5 degrees of freedom, p= 1
```

3. Analyzing Children Groups:

```
# Log Rank test - Children
surv_children_LR <- survdiff(Surv(srch_dur ,fnd_intr) ~ children, data = data)
surv_children_LR
```

```
## Call:
## survdiff(formula = Surv(srch_dur, fnd_intr) ~ children, data = data)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## children=No   37       12    14.62     0.470       1.4
## children=Yes  17       11     8.38     0.821       1.4
##
##  Chisq= 1.4  on 1 degrees of freedom, p= 0.2
```

After analyzing the impact of (cohort, age, children, education background) we found that there is only Chorot is significent which p-value is <0.05

Mesuring Impact of the 4 Variables using coxph regression Model:

```
coxph <- coxph(Surv(srch_dur ,fnd_intr) ~ cohort + age + edu + children, data = data)

summary(coxph)
```

```
## Call:
## coxph(formula = Surv(srch_dur, fnd_intr) ~ cohort + age + edu +
##     children, data = data)
##
##   n= 54, number of events= 23
##
##                    coef  exp(coef)   se(coef)      z Pr(>|z|)
## cohortA17            NA         NA  0.000e+00     NA       NA
## cohortA18   -1.494e+00  2.244e-01  2.064e+00 -0.724   0.4690
## cohortA19   -2.660e+00  6.998e-02  1.671e+00 -1.592   0.1115
## cohortA20   -4.478e+00  1.136e-02  2.151e+00 -2.082   0.0374 *
## cohortS18            NA         NA  0.000e+00     NA       NA
```

```
## cohortS19   -3.761e+00  2.327e-02  1.696e+00 -2.217   0.0266 *
## cohortS20   -3.674e+00  2.536e-02  1.779e+00 -2.066   0.0388 *
## age          3.349e-02  1.034e+00  3.569e-02  0.938   0.3481
## edufin       6.898e-01  1.993e+00  1.347e+00  0.512   0.6085
## edulit       8.146e-01  2.258e+00  1.346e+00  0.605   0.5450
## edumath      8.601e-01  2.363e+00  9.921e-01  0.867   0.3860
## edumgm       7.510e-01  2.119e+00  1.397e+00  0.538   0.5908
## eduoth      -1.247e+01  3.846e-06  7.324e+03 -0.002   0.9986
## childrenYes  7.031e-01  2.020e+00  7.320e-01  0.961   0.3368
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## cohortA17          NA         NA        NA        NA
## cohortA18   2.244e-01  4.456e+00 0.0039312   12.8100
## cohortA19   6.998e-02  1.429e+01 0.0026469    1.8502
## cohortA20   1.136e-02  8.806e+01 0.0001675    0.7698
## cohortS18          NA         NA        NA        NA
## cohortS19   2.327e-02  4.297e+01 0.0008379    0.6463
## cohortS20   2.536e-02  3.943e+01 0.0007767    0.8282
## age         1.034e+00  9.671e-01 0.9641874    1.1090
## edufin      1.993e+00  5.017e-01 0.1423215   27.9177
## edulit      2.258e+00  4.428e-01 0.1614866   31.5786
## edumath     2.363e+00  4.231e-01 0.3381227   16.5185
## edumgm      2.119e+00  4.719e-01 0.1371950   32.7311
## eduoth      3.846e-06  2.600e+05 0.0000000       Inf
## childrenYes 2.020e+00  4.950e-01 0.4811909    8.4801
##
## Concordance= 0.769  (se = 0.052 )
## Likelihood ratio test= 22.25  on 12 df,   p=0.03
## Wald test            = 13.33  on 12 df,   p=0.3
## Score (logrank) test = 60.4  on 12 df,   p=2e-08
```

The only significant variables are S20, S19, A20 according to the model, all other variables are not significent.

Bounus Question: Can you build a predictive model to identify students at high risk of a long search? How well does your model perform?"

As we saw earlier, there are not enough reliable and significant variables we could use to build a predective model, as cohort variable alone is not enough to explain or to pridect the time to internship.

So, based on these factors and current data situation, it’s not possible to build a predctive model.