Survival Analysis - Time to internship DSTI

Yijun ZHU, Raphaël ADAMCZYK, Hani CHERID

1. How many students partecipated in the interview

82 students participated in the interview

```
(n_students = nrow(raw))

## [1] 82
```

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

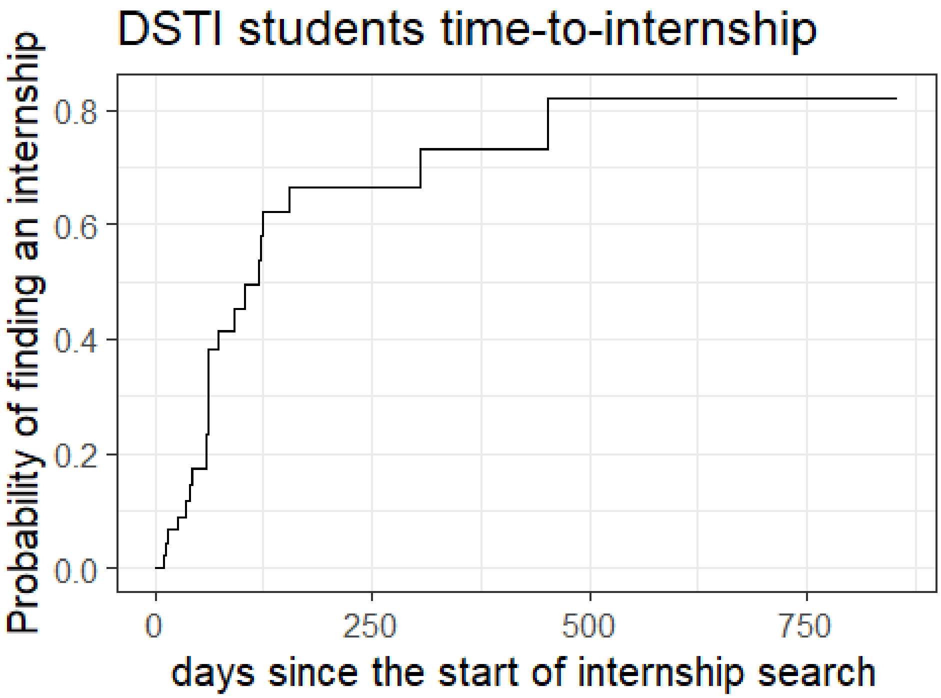
- Since we are analyzing time-to-internship, we need samples not null in column "When did you start looking for an internship": 18 samples are dropped
- We also want to exclude the students who haven't started looking for internship by the time of the survey: another 13 samples are dropped
- In total, 51 samples are usable

```
sum(is.na(raw$`When did you start looking for an internship`))

## [1] 18

sum(data$StartDate >= data$Timestamp)

## [1] 13
```

3. 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.

- The median time is 120 days, with 95% confidence interval the lower bound of 61 days, upper bound is NA.
- Total students at the baseline: 51
- Total number of events observed: 23
- Total number of censored observations: 28

```
table(data$findInternship)

##
## FALSE  TRUE
##    28    23
## Call: survfit(formula = Surv(time, findInternship) ~ 1, data = data)
##
##        n  events  median 0.95LCL 0.95UCL
##       51      23     120      61      NA
```



Data source: online voluntary survey, April 2020

#### Probability of not having found an internship after 60 days:

```
summary(sfit, time = 60)
```

```
## Call: survfit(formula = Surv(time, findInternship) ~ 1, data = data)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    60     26      12    0.678  0.0778        0.541        0.849
```

The probability of not having found an internship after 60 days: 67.8% (95% CI: 0.541-0849)

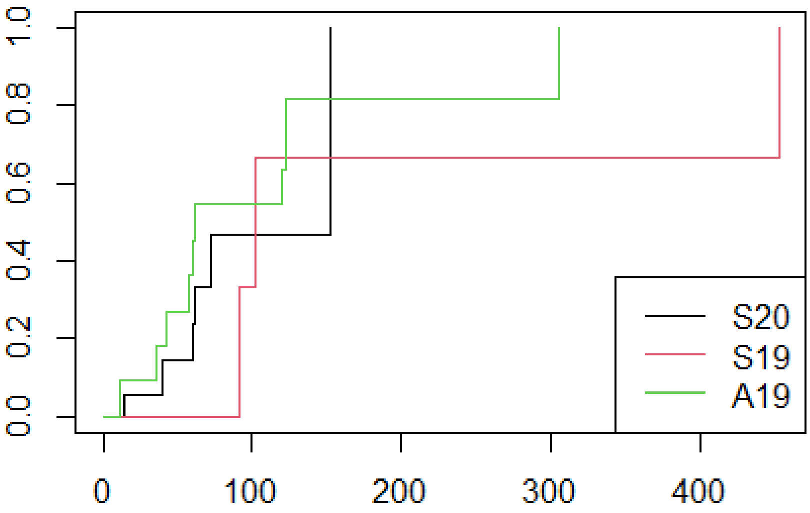
4. Of these variables, which ones have the most impact on the time to obtain an internship, and in which direction: cohort, age, educational background, having or not having children.

Cohort

- First we can select cohort S19, A19 and S20 as cohorts before as well as A20 are not informative
- We set S20 as the reference cohort
- From the logrank test, we get a p-value of 0.6 indicating the difference between different cohort is not statistically significant, which is validated by the Cox PH model. Although we see

that A19 students are slightly more likely (1.29) to find an internship and S19 less likely (0.60) compared to S20

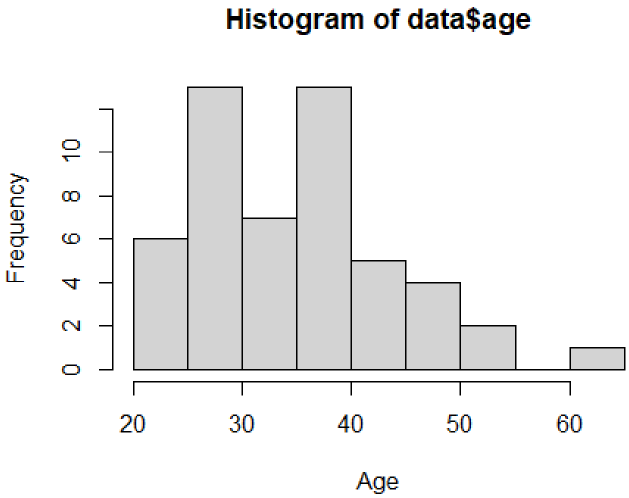
```
##
## S15 A15 S16 A16 S17 A17 S18 A18 S19 A19 S20 A20
##   0   1   0   0   0   2   0   1   3  11  18  15

## Call:
## survdiff(formula = Surv(time, findInternship) ~ cohort, data = data_cohort)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## cohort=S20 18        6     6.34    0.0178    0.0318
## cohort=S19  3        3     4.44    0.4684    0.8007
## cohort=A19 11       10     8.22    0.3846    0.7589
##
##  Chisq= 1  on 2 degrees of freedom, p= 0.6
```



```
## Call:
## coxph(formula = Surv(time, findInternship) ~ cohort, data = data_cohort)
##
##   n= 32, number of events= 19
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## cohortS15        NA        NA   0.0000     NA       NA
## cohortA15        NA        NA   0.0000     NA       NA
## cohortS16        NA        NA   0.0000     NA       NA
## cohortA16        NA        NA   0.0000     NA       NA
## cohortS17        NA        NA   0.0000     NA       NA
## cohortA17        NA        NA   0.0000     NA       NA
## cohortS18       NA         NA   0.0000      NA        NA
## cohortA18       NA         NA   0.0000      NA        NA
## cohortS19 -0.5080    0.6017   0.8487 -0.599    0.549
## cohortA19  0.2573    1.2935   0.5459  0.471    0.637
## cohortA20       NA         NA   0.0000      NA        NA
##
##           exp(coef) exp(-coef) lower .95 upper .95
## cohortS15        NA         NA        NA        NA
## cohortA15        NA         NA        NA        NA
## cohortS16        NA         NA        NA        NA
## cohortA16        NA         NA        NA        NA
## cohortS17        NA         NA        NA        NA
## cohortA17        NA         NA        NA        NA
## cohortS18        NA         NA        NA        NA
## cohortA18        NA         NA        NA        NA
## cohortS19    0.6017     1.6620    0.1140     3.175
## cohortA19    1.2935     0.7731    0.4437     3.771
## cohortA20        NA         NA        NA        NA
##
## Concordance= 0.605  (se = 0.071 )
## Likelihood ratio test= 1.14  on 2 df,   p=0.6
## Wald test            = 1.02  on 2 df,   p=0.6
## Score (logrank) test = 1.06  on 2 df,   p=0.6
```

Age

We analyse the difference in age by decade: - p-value is quite small which means that the difference is statistically significant - hazard ratio is 1.72 indicating 10 years older in age leads to almost twice the probability of finding an internship

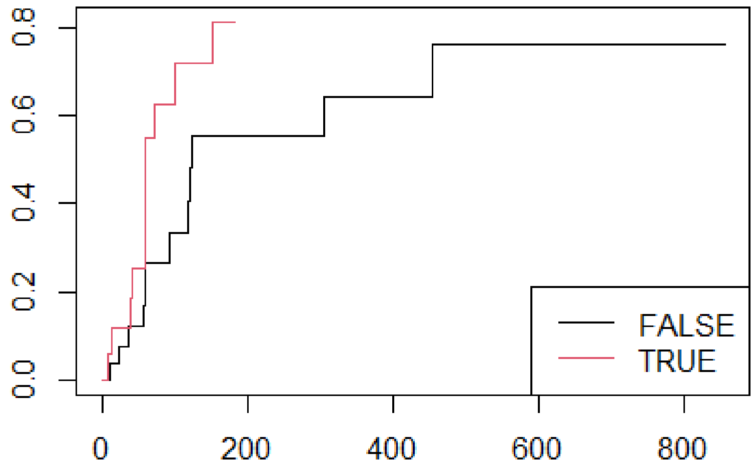


```
## Call:
## coxph(formula = Surv(time, findInternship) ~ I(age/10), data = data)
##
##   n= 51, number of events= 23
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## I(age/10) 0.5435    1.7221   0.1885 2.884  0.00392 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## I(age/10)     1.722     0.5807      1.19     2.492
##
## Concordance= 0.722  (se = 0.051 )
## Likelihood ratio test= 7.45  on 1 df,   p=0.006
## Wald test            = 8.32  on 1 df,   p=0.004
## Score (logrank) test = 8.81  on 1 df,   p=0.003
```

Having or not having children

- According to the Cox PH Model, the hazard ratio of TTI for people having children vs. children without children is 2.2, which means that people having children is more than twice more likely to find an internship within a certain period.
- Although the p-value is 0.075 which is not statistically significant

```
table(data$`Do you have children?`)

##
##  No Yes
##  34  17
```



```
## Call:
## coxph(formula = Surv(time, findInternship) ~ children, data = data)
##
##   n= 51, number of events= 23
##
##                 coef exp(coef) se(coef)    z Pr(>|z|)
## childrenTRUE 0.7815    2.1847   0.4391 1.78   0.0751 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##              exp(coef) exp(-coef) lower .95 upper .95
## childrenTRUE     2.185     0.4577    0.9239     5.166
##
## Concordance= 0.602  (se = 0.058 )
## Likelihood ratio test= 3.13  on 1 df,   p=0.08
## Wald test            = 3.17  on 1 df,   p=0.08
## Score (logrank) test = 3.33  on 1 df,   p=0.07
```

Education

- We set Math, Physics, etc. as the reference as most of the students have this background (32 out of 51)
- The logrank test shows a p-value of 0.7, which means that the difference between different education background is not statistically significant, as we can see also in the Cox PH model for each of the education background vs. Math, p-value is large
- While Mgmt background students seem 1.3 more likely to find an internship vs. Math students, other students are less likely to find an internship vs. Math students in a given time period

```
##
## math  bio  fin mgmt  oth
##   32    6    5    5    3
```

```
## Call:
## survdiff(formula = Surv(time, findInternship) ~ education, data = data)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## education=math 32       16    14.07     0.264     0.714
## education=bio   6        2     3.49     0.639     0.780
## education=fin   5        1     1.40     0.115     0.127
## education=mgmt  5        3     2.05     0.439     0.511
## education=oth   3        1     1.98     0.485     0.551
##
##  Chisq= 2  on 4 degrees of freedom, p= 0.7
```

```
## Call:
## coxph(formula = Surv(time, findInternship) ~ education, data = data)
##
##   n= 51, number of events= 23
##
##                   coef exp(coef) se(coef)      z Pr(>|z|)
## educationbio  -0.6781    0.5076   0.7564 -0.897    0.370
## educationfin  -0.4919    0.6115   1.0351 -0.475    0.635
## educationmgmt  0.2636    1.3016   0.6528  0.404    0.686
## educationoth  -0.8394    0.4320   1.0372 -0.809    0.418
##
##               exp(coef) exp(-coef) lower .95 upper .95
## educationbio     0.5076     1.9702   0.11526     2.235
## educationfin     0.6115     1.6354   0.08040     4.650
## educationmgmt    1.3016     0.7683   0.36211     4.679
## educationoth     0.4320     2.3151   0.05657     3.298
##
## Concordance= 0.593  (se = 0.051 )
## Likelihood ratio test= 2.14  on 4 df,   p=0.7
## Wald test            = 1.88  on 4 df,   p=0.8
## Score (logrank) test = 1.97  on 4 df,   p=0.7
```

Conclusion:

Among all the variables, age (in decade) is the only one of statistical significance. It has an important impact on the time to obtain an internship. People 10 years older are 1.7 times more likely to find an internship. While the difference is not statistically significant, people having children are 2.2 times more likely to find an internship, this might be related to the fact that people have children are in general older.

5. Bonus question: can you build a predictive model to identify students at high risk of a long search? How well does your model perform?

Automatic model selection based on AIC:

Mfull <- coxph(Surv(time, findInternship) ~ age + education + sex + cohort + children, data = data)

MAIC <- step(Mfull)

```
summary(MAIC)

## Call:
## coxph(formula = Surv(time, findInternship) ~ age, data = data)
##
##   n= 51, number of events= 23
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## age 0.05435   1.05586  0.01885 2.884  0.00392 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## age     1.056     0.9471     1.018     1.096
##
## Concordance= 0.722  (se = 0.051 )
## Likelihood ratio test= 7.45  on 1 df,   p=0.006
## Wald test            = 8.32  on 1 df,   p=0.004
## Score (logrank) test = 8.81  on 1 df,   p=0.003
```

With the lowest AIC, the automatic model selection has chosen that age is the most important explanatory variable for the response variable (confirmed by its own significant p-value (0.00392)). The HR being slightly greater than 1 and its p-value not too small, it indicates that age is a good feature variable for our model.

Model-based predictions

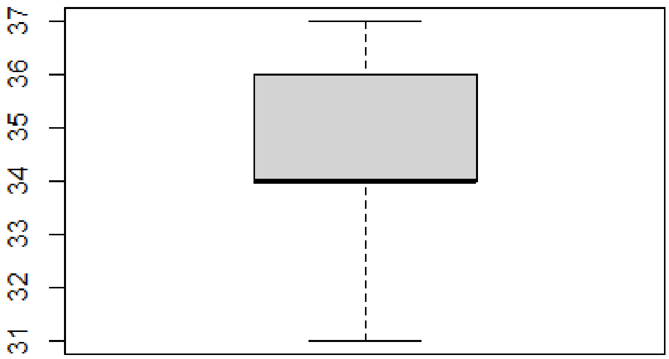
```
i.training <- sample.int(nrow(data), size = ceiling(nrow(data)/2), replace = FALSE)
i.testing <- setdiff(seq_len(nrow(data)), i.training)
d_training <- data[i.training, ]
d_testing <- data[i.testing, ]
```

Train candidate model

```
MA <- coxph(Surv(time, findInternship) ~ age, data = d_training)
```

Make predictions in the new dataset

```
d_testing <- with(d_testing, tibble(age, cohort, findInternship, time))
d_testing$lp_A <- predict(MA, newdata = d_testing, type = "lp")
d_testing

## # A tibble: 25 x 5
##      age cohort findInternship  time    lp_A
##    <dbl> <fct>  <lgl>          <dbl>   <dbl>
##  1    27 A20    FALSE             14 -0.555
##  2    34 S20    TRUE              60 -0.129
##  3    28 A19    TRUE             122 -0.494
##  4    38 A20    FALSE              4  0.115
##  5    42 S20    FALSE             43  0.358
##  6    29 A19    TRUE              58 -0.433
##  7    39 A17    FALSE            855  0.176
##  8    37 S20    FALSE             62  0.0539
##  9    39 A15    TRUE              25  0.176
## 10    32 S20    FALSE             79 -0.251
## # ... with 15 more rows
```



With a negative linear predictor, we observe that our model considers students likely to be at risk for long-term research to be under the age of:

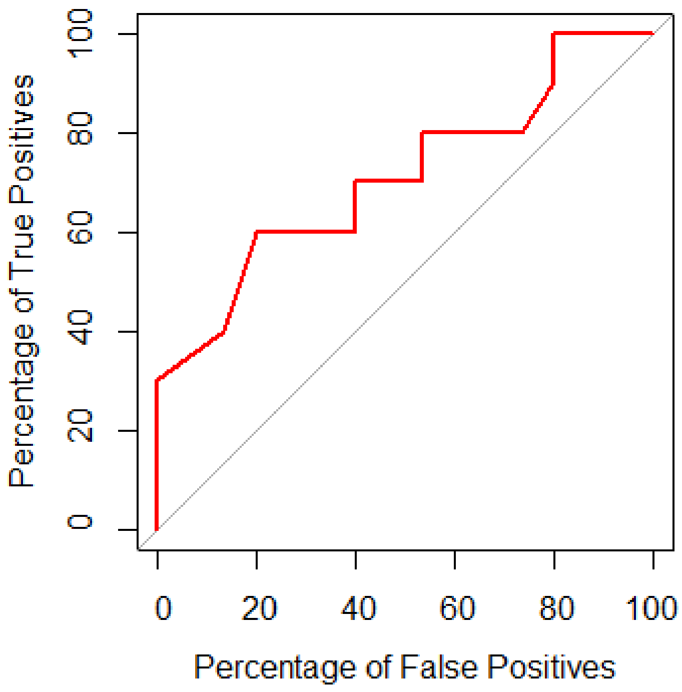
```
ceiling(mean(list_max))

## [1] 35
```

Assess predictive performance

```
## Call:
## coxph(formula = Surv(time, findInternship) ~ lp_A, data = d_testing)
##
##   n= 25, number of events= 10
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## lp_A 0.9733    2.6468   0.5036 1.933   0.0533 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## lp_A     2.647     0.3778    0.9863     7.102
##
## Concordance= 0.727  (se = 0.081 )
## Likelihood ratio test= 3.19  on 1 df,   p=0.07
## Wald test            = 3.74  on 1 df,   p=0.05
## Score (logrank) test = 4.07  on 1 df,   p=0.04
```

This linear predictor is pretty straight with a Hazard-ratio equal to:

```
## lp_A
## 2.65

library(pROC)

## Warning: package 'pROC' was built under R version 4.0.5

par(pty = 's')
roc <- roc(d_testing$findInternship, d_testing$lp_A, plot = T, col = 'red',
legacy.axes = T, percent = T, xlab = 'Percentage of False Positives', ylab =
'Percentage of True Positives')
```

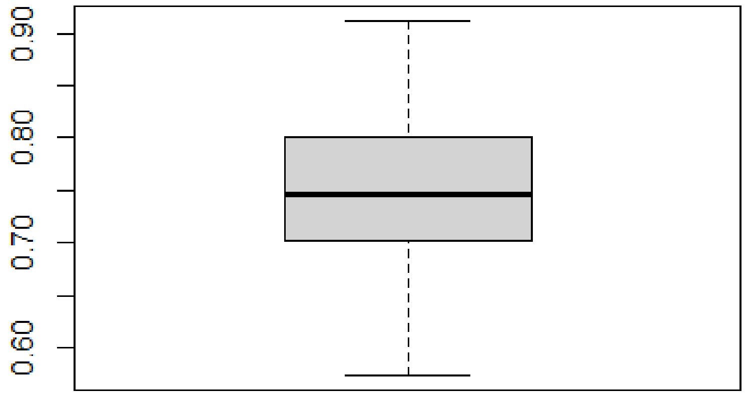


In presence of two students looking for an internship with the case where one finds an internship and the other not, the probability that the model correctly designates the student who find an internship is:

```
## 71 %
```

Try several times with bagging, the average AUC is always above 70%

```
## [1] 0.745448
```


0.60
0.70
0.80
0.90