MATH 80619 - Assignment5

Specific instructions

- Due on Friday April 9th before 10 am.
- In order to ensure reproducibility of your results, your assignment must be written using R Markdown or R Sweave. Your R code, output and comments should be included in the main document.
- You can submit the html or pdf output file on Zone Cours in the section “Remise de Travaux”.

Question 1

In this question, we’ll analyze a modified version of the `jasa` dataset available in the `survival` package. This dataset, also known as the Stanford heart transplant data, provides survival data for 103 patients on the waiting list for the Stanford heart transplant program. During the course of this assignment, you will have to create different versions of this dataset in order to explore and practice the following topics:

- Interval censoring
- Cox model with time-varying covariates

This dataset is first pre-processed using the code below (see the vignette from Therneau et al., 2017, “Using time dependent covariates and time dependent coefficients in the cox model”) and variables relevant for our analyses are extracted.

```
# Data preprocessing
library(survival)
jasa$subject <- 1:nrow(jasa)
tdata <- with(jasa, data.frame(subject = subject,
futime= pmax(.5, fu.date - accept.dt),
txtime= ifelse(tx.date== fu.date,
(tx.date -accept.dt) -.5,
(tx.date - accept.dt)),
fustat = fustat
))


sdata <- tmerge(jasa, tdata, id=subject, death = event(futime, fustat), trt = tdc(txtime),
                options= list(idname="subject"))
sdata$year <- as.numeric(sdata$accept.dt - as.Date("1967-10-01"))/365.25
sdata=sdata[,c(6,7,15,16,17,18,19)]
head(sdata)
```

```
##   surgery      age subject tstart tstop death trt
## 1       0 30.84463       1      0    49     1   0
## 2       0 51.83573       2      0     5     1   0
## 3       0 54.29706       3      0    15     1   1
## 4       0 40.26283       4      0    35     0   0
## 5       0 40.26283       4     35    38     1   1
## 6       0 20.78576       5      0    17     1   0
```

In this dataset, time of death is only partially observed, i.e. we only know a time windows in which death occured. We call this **interval censoring**. We also have time-varying covariates available. The definition of the variables is as follows:

- `tsart` and `tstop` : time interval (`tsart` , `tstop`] in days where the event occured (or not) and where the covariates are measured
- `surgery` : binary variable indicating if the patient had a prior bypass surgery. This covariate is measured at baseline and doesn't vary with time.
- `age` : age of the subject (in years) at baseline
- `subject` : subject id
- `death` binary variable indicating if the event (death) occured in the time interval or not. If the last value for a patient is equal to 0, then death is right censored at time `tstop` .
- `trt` : binary variable indicating if the patient received a transplant in the time period or not

In the dataset above, we can see for instance that we have one set of measures for subject 1. This subject died sometime between 0 and 49 days after baseline. No transplant was received during this period. Subject 4 was 40 years old at baseline with no prior surgery, didn't receive a transplant for the first 35 days and was alive during this time. However, this patient received a transplant sometime during day 35 and 38 and died during this period.

1. In the following questions, you are going to practice fitting a Cox model in the context of interval censoring. In order to do so, you are first going to work with a version of the dataset where **we have only one set of measures per subject, which corresponds to the last interval of time available for this subject**.

This dataset should contain the following variables:

- `subject` , `age` , `trt` , `surgery` , `tstart` , `tstop`
- a new binay variable `death` indicating if the patient died during the time period

   a. Construct this new dataset and provide a summary using the simple command `summary`

   b. Fit a Cox model to this data using the variables `age` , `trt` and `surgery` and the R function `coxph` . The interval censoring is defined with the function `Surv` that we used in the case of right censoring. Use the help of this function to understand how to code it. Comment on the significance of the variables.

   c. Obtain the risk fitted values for all subjects in the dataset and print the values for the first 10 subjects

   d. Compare the predictions of the median survival times with the observed survival times (there are several simple ways to do it)

2. We are now going to work with the time-varying dataset `sdata` . Fit a Cox model to this data (use the command `coxph` as it is) and compare the results with the previous question 1b).

Question 2

Here, we continue the analysis of the colon dataset used in the last assignment. Use the following code to prepare the dataset:

```
library(survival)
data("colon")
mycolon=colon[colon$etype==2,]
mycolon.train=mycolon[1:500,-c(1,2,16)]
mycolon.test=mycolon[501:929,-c(1,2,16)]
id.train=mycolon[1:500,1]
id.test=mycolon[501:929,1]
```

1. Compute the predicted survival time on the test set obtained from a Cox model (include all variables in each model) and compute the C-index based on these predicted values
2. Compute the predicted survival time on the test set obtained from a AFT model assuming a lg-normal distribution (include all variables in each model) and compute the C-index based on these predicted values
3. Compute the predicted survival time on the test set obtained from a survival random forest with 400 trees (include all variables in each model) and compute the C-index based on these predicted values
4. Compare the three C-index values and comment

Question 3

1. Re-run the three methods we used in the uplift modeling chapter on the simulated dataset used in this chapter. Evaluate the agrement between the three methods with respect to the ranking of the test-set subjects based on the predicted lift. You can use the strategy of your choice
2. The dataset used in class was simulated such that there is a perfect balance between groups (50% are cases and 50% are controls). By resimulating the data under different scenarios, evaluate if a disequilibrium between the group size (keeping the same total size) has an impact on the performance of the methods.