MATH 80619 Advanced statistical learning

Assignment #4 – chapter 5

Due, March 22nd 2021 before 10 am

Specific instructions

- 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, you will analyze the `colon` dataset available in the `survival` package. These are data from one of the first successful trials of adjuvant chemotherapy for colon cancer. Levamisole is a low-toxicity compound previously used to treat worm infestations in animals; 5-FU is a moderately toxic (as these things go) chemotherapy agent. **There are two records per person, one for cancer recurrence and one for death.** The variables are the following:

id:	subject id
study:	1 for all patients
rx:	Treatment: Obs (no chemio), Lev (Levamisole), Lev+5-FU (Levamisol 5FU)
sex:	1=male; 0=female
age:	age in years of the patient at baseline
obstruct:	obstruction of colon by tumour (1=yes / 0=no)
perfor:	perforation of colon (1=yes / 0=no)
adhere:	adherence to nearby organs (1=yes / 0=no)
nodes:	number of lymph nodes with detectable cancer
time:	days until event or censoring
status:	censoring status
differ:	differentiation of tumour (1=well, 2=moderate, 3=poor)
extent:	Extent of local spread (1=submucosa, 2=muscle, 3=serosa, 4=contiguous)
surg:	time from surgery to registration (0=short, 1=long)
node4:	more than 4 positive lymph nodes
etype:	event type: 1=cancer recurrence, 2=death

1) Prepare the dataset in order to consider the following:

- We'll consider only the time-to-death as the event of interest (time to recurrence will not be used)
- Variables `differ` and `extent` will be considered as numeric since they are ordered
- The study variable will not be used
- The first 500 subjects will be used for training and the remaining subjects will be used as a test set.

2) Compute basic descriptive statistics on the training dataset:
   i) What is the censoring rate ?
   ii) What is the mean/median/min/max values of the survival time ?
   iii) Re-compute the mean/median/min/max values of the survival time, for each value of the status variable. Explain and interpret the difference in the distributions.

3) Let's see what the global estimated survival curve looks like…
   i) Draw a Kaplan-Meyer (KM) estimate of the survival curve for the training data.
   ii) What is the estimated probability to survive at least 400 days ?
   iii) Draw the KM estimate of the survival curve for each type of treatment. Comment.
   iv) Explain why the confidence interval for the survival curve is larger towards the end of the curve

4) Perform a log-rank test to compare the survival curves between treatment. Comment.

5) Use a Cox proportional hazard model to evaluate the effect of the treatment (use only this variable in the model).
   i) Interpret the two coefficients of the model (on the exponential scale)
   ii) Make a link between some parts of the R output and the previous question

6) Use a Cox proportional hazard to evaluate the effect of all variables.
   i) Predict the risk of death for all subjects in the test set. Interpret this risk for the 1st subject (id=501)
   ii) Explain why some missing values are generated for the predicted risks
   iii) Give the list of the four subjects with the highest risk. By looking at their characteristics, make a link with the results of the Cox model you just fitted.

7) Now, fit an accelerated Failure Time (AFT) model using all variables, assuming a log-normal distribution.
   i) Interpret the coefficient of the age variable (on the exponential scale)
   ii) On the test data, compute the predicted median survival time. Compare it with the actual survival time and comment.

iii) Plot the ranks of the subjects according to their predicted risk from the Cox model versus the rank computed from the predicted survival time. Comment.

Question #2

Create **one** question of the type "multiple choice questions" on the material seen in chapter 5.

- Your questions should have 4 possible answers, 1 true answer and 3 false answers.
- The level of the question should not be too easy (i.e. wrong answers should not be obvious).
- Solutions - with explanations for the right and wrong answers- need also to be included.
- **Note that your questions can be based on R outputs (in which case you need to provide the R output)**

Your mark for this question will be based on the correctness of your question/answers, and on the depth of your question and answers (especially the wrong ones).