ISEN 613 Fall 2017 Midterm

Start: October 17 (Tue) 8:00am
Due: October 18 (Wed) 11:59pm

Background

Doctor White in a surgical center is interested in predicting survival time of patients undergoing a type of liver operation. Particularly, he wants to know which variables of patients affect their survival time and how. There is information on 108 patients randomly selected from the patients of the center. The variables are defined as follows:

Predictor variables

bloclot	blood clotting score
prog	prognostic index
enzy	enzyme function test score
liverfunc	liver function test score
age	age
gender	gender (0=male, 1=female)
alco	history of alcohol use (0=none, 1=moderate, 2=severe)

Response variable

survival	survival time after the liver operation (in days)

The data are divided into two parts: training data (*Data_training.csv*, 60 observations) and test data (*Data_test.csv*, 48 observations). You are expected to conduct data analysis to help address Doctor White's concerns.

Required Analysis

I. What analysis will you conduct?

II. Find two good models for the training data. (*Hint:* you can start from an initial model, then improve it through different ways, and select two from all models you considered. *Note:* In the answer to this question, clearly describe the strategy and steps you will follow to find the models before show detailed results of analysis.)

III. Find the prediction performance of the two models on the test data (*Hint:* use *predict(fitted model, dataset)* in R to find predictions).

IV. Which one of the two models will you choose? Why?

V. Based on the chosen model, what will you tell Doctor White?

Submission

Submit a **pdf** report about your analysis (including snapshots of R outputs) through the given link

eCampus -> Midterm_Part B ->Report submission

Instructions

(1) Use what you have learned in **Units 1~15** to solve the problems.
(2) To upload a csv dataset in R, you may use the following command or search for help online:
data1 = read.csv("file name.csv", header=TRUE)
(3) Before starting analysis in R, you may need to change the working directory of R (in "File" menu) to the folder containing the dataset.
(4) Report must follow the order of questions (I, II, III, IV, V). First show the question and then your answer (including R snapshots). Points will be deducted for violations of the required format.
(5) Page limit = **10**. Materials beyond will be ignored.
(6) Show your INDEPENDENT work. Collaboration would be considered a direct violation of the Aggie Honor Code and such instances would be dealt accordingly.