Stat 343: Final exam — data analysis

Due by Tuesday Dec 11 2018 by end of day (11:59pm)

Download the data set `heart.txt` (available on canvas, source: Machine Learning Repository @ UCI). The variables are:

1. `age` — age of patient in years
2. `sex` — coded as 1 for male and 0 for female
3. `chestpain` — chest pain type, categorical, coded as 1,2,3,4 (the order is arbitrary)
4. `restbp` — resting blood pressure
5. `chol` — cholesterol level
6. `fbs` — fasting blood sugar, 1 indicates that it was above a predefined threshold, or 0 otherwise
7. `restecg` — resting electrocardiographic results, 0 = normal while 1 and 2 indicate two different types of abnormalities (the order of these two abnormalities is arbitrary)
8. `exang` — exercise-induced angina, 1 = present or 0 = absent
9. `oldpeak` — a quantitative variable measuring one type of change in the patient's ECG during exercise
10. `slope` — a categorical variable measuring a different type of change in the patient's ECG during exercise
11. `fluoro` — number of major blood vessels colored by fluoroscopy, which indicates potential problems
12. `extest` — results of exercise test, 0 = normal, 1 and 2 indicate different types of defects (the order of these two is arbitrary)
13. `maxhr` — maximum heart rate attained during exercise trial

Your task is to build a model for the response `maxhr`, with all the other variables as potential covariates. You should aim to build a model that predicts `maxhr` accurately, and to draw any appropriate conclusions regarding the associations of `maxhr` with the other variables (e.g. which effects are positive or negative, which covariates interact with each other, etc). Throughout your analysis, you should describe what you observe at each step, and explain your reasoning behind the decisions you make.

Along the way, you should consider questions like:

- Is a linear model appropriate for this data set? How about constant variance?
- Are there any variables that should be transformed?
- Are there any covariates that should be removed from the model?
- Are there any interactions between covariates?
- Are there any issues of collinearity that we should be aware of, or should try to address?
- What is the best way to handle the small number of missing values in the data set?
- Are there any questions of multiple testing that arise in the process of your analysis?
- What is the predictive accuracy of the model?

(You are not required to address every question, and certainly other questions may arise that are worth addressing. These are just examples.)

Guidelines

- Your final report should be a .Rmd file along with the knitted .html file (please submit both).
- Your report should show all of your work, plus discussion along the way to explain the choices made in the analysis and the conclusions drawn.
- You can use any methods covered in class, or you can combine or alter these methods as needed. Please do not use methods that are substantially outside the scope of the material in this course (e.g. logistic regression, neural nets).
- You are welcome to use existing R packages and functions, but only if your writeup explains clearly what is being computed. For example, if you use `lm.ridge` for a Ridge regression, it would not be appropriate to just say that you're performing a regression; you should state exactly the function being minimized / the equation being solved, explain details e.g. how the penalty parameter is chosen, whether columns of X are standardized first, etc; and explain why you are using this method rather than alternatives. Be sure you know what the function is doing (such as whether it does any preprocessing of the data like centering etc). In other words, your written explanation should contain all the details that would be needed if you were to write the code yourself.
- You may refer to your textbook, to your course notes, or search online for textbook type resources or for questions relating to your R code. You should not use any internet resources relating to this data set specifically and should not ask for help online (e.g. on quora).
- You may not discuss this assignment with each other or with anyone else, aside from the instructor & TAs.