Homework #6, Stat 689, Spring 2018, Due Class #9, February 14

The Framingham Heart Study https://www.framinghamheartstudy.org/ is the longest on-going study of risk factors from heart disease and many other types of chronic diseases. It is a national treasure started in 1948.

Please download the data I have: Framingham_Description.pdf and Framingham.csv. The header to the .csv file has the variable names, and the description tells you what the variables mean. The data set should have 1,615 male patients: please check the number.

This exercise is to study whether systolic blood pressure and serum cholesterol are related, and whether smoking status (and later on age) are factors in systolic blood pressure (it is).

There are four systolic blood pressure measurements. Take their average and create the variable LSBP = log(SBP-50). Also, average the two cholesterol measurements and take their logarithm, calling it Lcholest. At the end, you will work with chd, age, smoker, LSBP, Lcholest. When we do a generalized additive model, we will get to see if they predict the onset of coronary heart disease.

This exercise is (mostly) based on Lectures 6 and 7. I skipped one point in these lectures deliberately and you may be puzzled by the one of the questions.

1. Fit a multiple linear regression of LSBP on Lcholest and smoker using lm. Since the smoking variable is binary, this is an ordinary ANCOVA without an interaction. You will notice that the Rsquared is quite low. Produce a table of estimates, standard errors and p-values.
2. Do a little bit of a web search about whether smokers have higher or lower blood pressure than nonsmokers. Does the analysis in (1) agree?
3. In (2), there is a subtle statistical interpretation of what I am asking you to do, because your analysis also includes the transformed cholesterol variable. I am curious if you can produce the correct terminology for what that subtle interpretation is. Please, less than 15 words. This is a test of your background, but I want to make sure you know how to report results appropriately.
4. Do the same thing as in (1) but add an interaction
5. Now run a semiparametric regression using mgcv, one that is the semiparametric version of ANCOVA without an interaction.
6. In (5), display a plot of the two lines, but without the data.
7. In (5), display a plot of the two lines but in one graph with 2 columns, (par(mfrow(1,2) does this, I believe), and also show their pointwise 95% confidence intervals.
8. Run the semiparametric version of ANCOVA but with an interaction. Does it look like there is an interaction. Cite p-values and estimates.
9. In (8), display the fits but without the data points.
10. What does having an interaction mean in the case when the factors are binary?

11. Now run an analysis of whether the two lines (smoker versus nonsmoker) are statistically significantly different. You might remember that we contrasted the Srod district with the other Warsaw districts.
12. Ignore the smokers, and run native code using nlme and lme to the fit of LSBP versus Age. Please note: now it is Age as the predictor.
13. In (12), display the fit without the data points. Also provide a confidence band.
14. Finally, in (12), it the fit statistically significant?
15. In (12), test whether the fit is linear or quadratic versus the need to do a semiparametric fit.