BST 210 HOMEWORK #3

Due 11:59pm, Tuesday, October 8, 2019

***Please be sure to submit your assignment by 11:55ish pm (or before) to prevent any glitches in the upload from precluding your timely submission.**
***Please work well in advance, getting help during office hours and labs, as there will be no extensions given for this assignment, outside of extreme, extenuating circumstances which must be communicated in advance to the primary instructor.**

There are 2 problems, each with various parts, in this homework assignment. Please double check that you have provided a response for each part of each problem, before you submit.

BST 210 Problem set policies:

- *We encourage you to discuss homework with your fellow students (or with the instructor or the TAs), but you must write your own final answers, in your own words.*
- *Please include the appropriate computer output in your solution if that helps you to answer a question, but be sure to interpret your findings in words – submitting only output is not sufficient for full credit.*
- *Homework assignments will not be accepted late (other than for extreme emergency, but the primary instructor must be reached in advance).*
- *Be complete in your responses; not verbose, to get full scores.*
- *All homework must be submitted online via Canvas by 11:59pm on Tuesday.*

Here we continue to explore data from the “Singapore Cardiovascular Cohort Study 2”, using continuous age, gender, and continuous body mass index (defined as weight/height^2 in kg/m^2) to predict total cholesterol (in mmol/l) of subjects. Note that 1 mmol/l (SI units) equals 38.67 mg/dl, the usual American units for cholesterol. Our main goal is to assess potential nonlinear effects of continuous covariates through the use of appropriate spline or GAM models (you have some flexibility here and only need to use one of these approaches [your choice, which may depend on your favorite statistical package] to answer the questions below).

1. First, we further explore the effects of continuous age to predict total cholesterol.

(a) Run three linear regression models using linear age, linear and quadratic age, and then either a spline or GAM modeling of age to predict total cholesterol. In one sentence, describe how you have fitted the spline or GAM model (e.g., choice of knot points, order of the polynomial, other restrictions). Plot the three sets of fitted values, look at the regression output obtained, and briefly compare the three fits. Which of the three models do you recommend as being "best" so far? Why?

(b) We say that Model A is *nested* within Model B if the parameters in Model A are a subset of the parameters in Model B. It is thus easy to see that the model using linear age (Model A) is nested within the more complex model using linear and quadratic age (Model B). In some other cases, the parameterization of Models A and B may make it a bit more difficult to determine possible nesting. You could confirm whether or not the linear age model is nested within your spline or GAM model by comparing your spline or GAM model to the model that also adds in linear age to your spline or GAM model. What happens when you do that? Can you tell if the linear age model is nested within your spline or GAM model? Briefly, how?

(c) Can you tell if the linear and quadratic age model is nested within your spline or GAM model? Briefly, what are your findings?

(d) You can also run a model using linear and quadratic age plus either a spline or GAM modeling of age, and determine how/whether we can tell if using linear and quadratic age is sufficient to model the effects of age versus a more complex spline or GAM model. Effectively, you are asking whether or not the spline or GAM modeling of age is needed after including linear and quadratic age. What are your conclusions? Be sure to perform an appropriate hypothesis test with an appropriate number of degrees of freedom.

2. Suppose that the main research question is to determine the effects of (continuous) body mass index on total cholesterol, considering (continuous) age and gender as possible confounders or effect modifiers. The goal is to flexibly model the effects of age and gender while appropriately assessing the effects of body mass index on total cholesterol. You want to (hopefully!) be able to present an easily interpretable effect of body mass index on total cholesterol to your readers.

(a) Run some models that appropriately address this research question. What final model do you recommend? Briefly justify your choice. (You don't have to include the outputs of lots of models here, but perhaps write a brief description of your approach to get to your final model.)

(b) Take your final model and write 1-2 summary sentences that describe the overall results of your model in a form that could appear in a manuscript. Use American units of cholesterol in this summary. Be sure to include sufficient statistical detail (confidence intervals, p-values, decimal places, etc.), clarity of adjustment factors, and interpretation (units, direction of effect) in your sentence(s). Maximum of two sentences!