BST 210 HOMEWORK #1

Due 11:59pm, Monday, September 23, 2019

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 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.*
- *(Per the school, we are required to note that solutions prepared by copying another's work would not be acceptable--the same holds for copied computer programs or files—mainly because students then will not have had the benefit of having thought about and worked through problems, when taking exams or when practicing methods learned in this course out in the real world.)*

In this homework we will begin exploring data from the "Singapore Cardiovascular Cohort Study 2" (SCCS2), a cohort study designed to identify environmental and genetic risk factors for cardiovascular and metabolic diseases from a population of subjects from Singapore. Dr. Bee Choo Tai has graciously made available a subset of the SCCS2 data to use for class purposes, which she uses in her textbook Regression Methods for Medical Research, Wiley, 2014. A Stata dataset with 527 subjects has been provided, but you can do all of your work using any of Stata, SAS, and/or R. (You can always google how to read one type of dataset into another package, and future datasets will be of varied formats, so everyone will get practice converting datasets. Students will have access to all three software/programming formats during labs, and often in lab solutions.)

We encourage you to begin by exploring the data, to get a sense of the variables included, coding conventions, missing data, etc. As part of this exploration, look at some graphical and numerical summaries of the data.

Please answer the questions below, focusing on using age, gender, and body mass index (defined as weight/height^2 in kg/m^2) to predict total cholesterol of subjects.

1. First, we focus on the effects of age to predict total cholesterol.

(a) Based on a simple linear regression model of age to predict total cholesterol, describe the effects of a change of an age decade (10 years) on total cholesterol in a sentence appropriate for a manuscript, including an appropriate 95% confidence interval and P-value. Also, briefly compare and contrast the findings and interpretation from this simple linear regression model vs. the Pearson correlation coefficient between age and total cholesterol.

(b) Using an appropriate graphical method, visually assess the residuals. Do they look normally distributed for your simple linear regression model? Do you see any other anomalies? Briefly comment.

(c) Study the effects of age first through a Lowess curve (or similar smoothing method) and by including both a linear and quadratic effect of age to predict total cholesterol in a linear regression model. Do we have any statistical evidence for a nonlinear effect of age to predict total cholesterol? Briefly comment.

2. Next, we consider the effects of gender to predict total cholesterol.

(a) Perform an equal variance *t*-test comparing total cholesterol by gender. What are your findings? Also run a simple linear regression model using gender to predict total cholesterol and show as many “equivalences” as possible between the linear regression results and the equal variance *t*-test results.

(b) Starting with the model using both linear and quadratic age to predict total cholesterol, how does the model change if you add in the main effects of gender? Is gender a confounder of the effect of linear and quadratic age on total cholesterol? Is gender an independent predictor of total cholesterol? Briefly justify your conclusions.

(c) Based on your model including linear and quadratic age as well as gender to predict total cholesterol, determine fitted regression lines for men and women. Also, graph the fitted values overlaid with a scatterplot of the raw data. What do you notice about the plot of the fitted values?

(d) Considering additional models as appropriate, is gender an effect modifier of the effect of linear and quadratic age on total cholesterol? Briefly justify your conclusion. Based on your full interaction model (five covariates plus an intercept in total), determine the fitted regression lines for men and women. Also, graph the fitted values overlaid with a scatterplot of the raw data. What do you notice about the plot of the fitted values?

(e) Compare the R^2, adjusted R^2, and square root of the MSE values across all of your models in questions 1 and 2. Which model, out of all the models run so far, is “best” in terms of R^2? In terms of adjusted R^2? In terms of square root of the MSE?

(f) Finally, compare your results from 2 (d) above to *two* linear regression models using linear and quadratic age to predict total cholesterol *separately* for men and women. Again, graph the fitted values overlaid with a scatterplot of the raw data. What are the similarities and differences between these approaches? Which method do you prefer (and why)?

3. Finally, we consider adding the effects of body mass index (BMI) to predict total cholesterol. Some modelers might use BMI as a continuous covariate, while others might categorize BMI as follows:

Underweight (BMI < 18.5)
Normal weight (18.5 ≤ BMI < 25.0, the preferred baseline category)
Overweight (25.0 ≤ BMI < 30.0)
Obese (BMI ≥ 30.0).

(a) First ignoring the effects of age and/or gender on outcome, compare models with either continuous or categorical BMI to predict total cholesterol. Describe how the categorical BMI model relates to analysis of variance. Determine the fitted values for each of your models, and overlay scatterplots of the raw data and these two fits. Which model do you prefer, and why?

(b) Does including a quadratic BMI term add to your continuous BMI model? Briefly comment.

(c) Run a model adding your "best way" to model BMI to the results of your "best model" from 2(e) (and describe how you selected a "best model"). Does the adjusted R^2 improve by adding the effects of BMI? Do the residuals look approximately normally distributed? Does a scatterplot (or Lowess curve) of the residuals vs. the fitted values show any evidence of heteroscedasticity or other model violations?

(d) Do any points look to have high influence or high leverage in your model? Does your model change appreciably if you drop such observations (drop no more than 5-10 observation)? Comment briefly.