# BST 210 HOMEWORK #2

**Due 11:59pm, Tuesday, September 30, 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 5 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 outliers through residual analysis, to assess leverage through hat values, and to assess influence through Cook's distances.

**1. Based on previous modeling, we know that it would be helpful to include both linear and quadratic age to predict total cholesterol.**

(a) After adjusting for linear and quadratic age, does body mass index confound the effect of gender? Also, does gender confound the effect of body mass index? Briefly interpret your findings.

(b) After adjusting for linear and quadratic age, does body mass index serve as an effect modifier of the effect of gender? Also, does gender serve as an effect modifier of the effect of body mass index? Briefly interpret your findings.

**2. Now focus on the evaluation of a *single model* including linear and quadratic age, I(female) as a categorical predictor, and linear body mass index to predict total cholesterol.**

(a) For this model, hand calculate the total cholesterol value (*in American units*) predicted for a man aged 30 years with BMI = 30.

(b) Alas, different software packages (Stata, SAS, R), different books, and different websites may use the same term to mean different concepts. This is true for the case of *standardized versus studentized* residuals. For whichever package you are using, try to determine how your package is calculating standardized or studentized residuals (or both, if your package has both). If you can, write out the formula your package is using for whichever of these residuals is calculated using class notation, provide brief documentation of your findings (e.g., cut and paste in some text from a help file or website, etc.), and put it in context of the class notes (e.g., internally standardized, externally standardized, etc.). You may need to look at help files, use Google or specific stat help sites, to answer this question.

(c) For the model we are considering above, *use your software package* to calculate the raw residuals (observed $Y$ – fitted $Y$) as well as the standardized or studentized residuals (or both, if your package has both). Draw scatter plots comparing each of these residuals and calculate Pearson correlation coefficients between each of these. Be careful to consider decimal places – are the correlations = 1 or just very close to 1? What do you find in this case? How do you interpret this? What are the implications in determining which residual to use, at least for this SCCS2 example?

(d) We claimed in lecture that the raw residuals would have mean zero. How true is that in this example? Do your standardized or studentized residuals (or both, if your package has both) have mean zero? Briefly comment.
(e) Use your standardized or studentized residuals to assess the normality assumption, through at least two of: histograms, QQ plots, and/or a formal test for normality. What are your findings?
(f) Show which points (including the subject's age, gender, BMI, total cholesterol, and standardized or studentized residuals) have absolute values of the standardized or studentized residuals > 2. Also show which have absolute values > 3. For these last few (> 3), briefly describe why each observation is sticking out as an outlier.

**3. Continue with our evaluation of the *single model* including linear and quadratic age, I(female) as a categorical predictor, and linear body mass index to predict total cholesterol.**
(a) For the model we are considering above, *use your software package* to calculate the hat values. We claimed in lecture that the hat values would all be positive and would average out to equal ($p$ + 1) / $n$. How true is that in this example?
(b) Draw a histogram or boxplot of the hat values. List which individuals (including the subject's age, gender, BMI, total cholesterol, and hat value) have leverages > 2 ($p$ + 1) / $n$. Also list which have leverages >
4 ($p$ + 1) / $n$. For these last few [> 4 ($p$ + 1) / $n$], briefly describe why each observation is sticking out as having high leverage.

**4. Continue with our evaluation of the *single model* including linear and quadratic age, I(female) as a categorical predictor, and linear body mass index to predict total cholesterol.**
(a) For the model we are considering above, *use your software package* to calculate the Cook's distances. In general, Cook's distances might all be positive. How true is that in this example?
(b) Draw a histogram or boxplot of the Cook's distances. List which individuals (including the subject's age, gender, BMI, total cholesterol, studentized or standardized residual, leverage, and Cook's distance) have Cook's distances > 4 / $n$. Also list which have Cook's distances > 12 /

*n*. For these last few (> 12 / *n*), briefly describe why each observation is sticking out as having high influence.

(c) Choose one other easily available measure of influence (e.g., DFBETA, DFFITS) that your software package calculates, and see if you have any individuals that have overly high influence with this measure.

**5. Continue with our evaluation of the *single model* including linear and quadratic age, I(female) as a categorical predictor, and linear body mass index to predict total cholesterol.**

(a) Look closely at the observations with absolute value of either the standardized or studentized residuals > 3, the leverages > 4 (*p* + 1) / *n*, and/or the Cook's distances > 12 / *n*. How much overlap in these observations do you observe?

(b) Finally, to put it all together, compare your model run on the 527 observations to a model run after eliminating all the (high outlier, high leverage, and/or high influence) observations from 5 (a). Do your overall findings change much?