# Homework 3

**Due: Saturday, Nov 7, 2020 at 11:59 pm**

To complete this assignment, you will need to download **heart.csv.**

Below is brief summary of variables in hear.csv.
- **Weight**: subject's weight
- **Systolic**: top number in a blood pressure reading, indicating the blood pressure level when the heart contracts
- **Diastolic**: bottom number in a blood pressure reading, indicating the blood pressure level when the heart is at rest or between beats
- **Cholesterol**: measured cholesterol

## Exercise 1:

We would like to investigate the relationships between **cholesterol, weight** and/or **blood pressure**. The data set contains **Weight, Diastolic Blood pressure, Systolic blood pressure** and **Cholesterol** for alive subjects in the **heart.csv**.

The medical director at your company wants to know if **Weight** alone can predict **Cholesterol** outcome. Consider modeling **Cholesterol** as a function of **Weight**.

a) Fit a linear regression model for **Cholesterol** as a function of **Weight**. If any points are unduly influential, note those points, then remove them and refit the model. Consider Cook's distance cut off to be 0.015.

b) Comment on significance of the parameters, variation explained by the model, and any remaining issues noted in the diagnostics plots. What does this model tell us about the relationship between **Cholesterol** and **Weight**? Interpret the relationship specifically. Explain to the medical director whether this is a good model for the prediction of Cholesterol level.

## Exercise 2:

The medical director wants to know if blood pressures and weight can better predict cholesterol outcome. Consider modeling **cholesterol** as a function of **diastolic, systolic**, and **weight**.

a) Fit a linear regression model for **cholesterol** as a function of **diastolic, systolic**, and **weight**. Generate the diagnostics plots and comment on any issues that need to be noted. Then make any necessary adjustments for undue influence. For Cook's distances, do not leave any points in the final model that have Cook's distance greater than 0.015.

b) Comment on significance of the parameters and how much variation in **cholesterol** is described by the model. Comment on the relationship between cholesterol and statistically significant predictor(s). Check multicollinearity issue among predictors. Explain to the medical director whether this is a good model for the prediction of Cholesterol level.

## Exercise 3:

Now consider stepwise model selection for the **Cholesterol** model. We remove influential points detected in Exercise 2, which has cook's distance larger than 0.015, prior to performing the model selection.

a) Perform stepwise model selection with .05 criteria and address any issues in diagnostics plots.

b) Interpret the final model and comment on the variation in **Cholesterol** explained. Compare the variations explained by the models of from Exercise 1 and 2.

## Exercise 4:

Now consider best subset selection for the **Cholesterol** model. Again, we remove influential points detected in Exercise 2, which has cook's distance larger than 0.015, prior to performing the model selection.

a) Find the best model based on adjusted-R square criteria and specify which predictors are selected.

b) Find the best model based on AIC criteria and specify which predictors are selected.

c) Compare final models selected in a) and b). Also compare final models from best subset approach with the final model from stepwise selection.