Project 5 (Due on 11.30.0 pm Sunday, November 19)

Notes:

- Write comments on each step. Submit report in class if your taking the class face to face. Others submit the report by email (You are wecome to submit printed copy).
- **You are supposed to work on this project entirely on your own. So, do not consult with anyone within or outside the class.**
- You are welcome to ask me questions. However, first try to find the answer on your own. Don't be afraid to google! Google is the best friend of a graduate student!!

1. The Haldcement dataset contains four predictors variables. Interest centers on using variable selection to choose a subset of the predictors to model Y.

   (a) List four best predictors subset like the following table (I have used Zs as for an example, assuming they are the best four.):

Subset size	Predictor(s)	R^2	R^2_{adj}	AIC	BIC
1	Z_2				
2	Z_3, Z_4				
3	Z_1, Z_3, Z_4				
4	Z_1, Z_2, Z_3, Z_4				

   Identify the optimal model or models based on R^2, R^2_{adj}, AIC and BIC from the approach based on all possible subsets.

   (b) Find the best model(s) based on stepwise selection method.

   (c) Compare between best model(s) obtained in (a) and best model(s) obtained in (b).

2. Uric Acid and Cardiovascular Risk Factors: The cardio dataset contains data on 998 individuals on the following variables

Table 1: Overview of the datasets.

#	Variable	Description
1	uric	Uric acid level
2	dia	Diastolic blood pressure
3	hdl	High-density lipoprotein cholesterol
4	choles	Total cholesterol
5	trig	Triglycerides level in body fat
6	alco	Alcohol intake (ml per day)

(a) Fit a full model for predicting uric acid levels using all other explanatory variables. Test if the variables hdl and choles can be (jointly) dropped together from the full model. State your conclusion.

(b) Find the best model(s) using R^2_{adj} and stepwise selection method.

(c) For the best model chosen above, check all assumptions. If an assumption is not met, attempt to remedy the situation by trying few tranformations. Comment on the fit of the final model using appropriate plots, tests, statistics.

(d) Fit a a weighted least squares regression model using the same variables as the best model above. Check if iterating the process of estimating weights improve the estimates.