HW#2 (Due February 19th)

You may work in groups of two. Both names must appear on the turned-in assigment. Note that you will not be able to work with this person again for the remainder of the HWs.

1) Consider the following dataset called "data" generated by the code:

```
x1=sample(1:100,40)
x2=sample(100:200,40)
x3=sample(200:300,40)
eps=rnorm(40,sd=6)
y= 5 – 6*x1 + 4*x2 + 2*x1^2+eps
data=data.frame(x1,x2,x3,y)
```

Note that the true relationship between y and the variables above is: $y = 5 - 6x_1 + 4x_2 + 2x_1^2 + \varepsilon$

(a) Carry out backward selection using adjusted-R^2 where your potential variables are x1,x2,x3, x1^2, and x2^2.
(b) Carry out backward selection using AIC where your potential variables are x1,x2,x3, x1^2, and x2^2

You must provide summary() output at every step and indicate which variables you are adding/removing in each step and why. You must write down the final estimated model (with coefficients obtained using the lm() or glm() function) once you decide which predictor variables to use, identified by each procedure.

(c) Did any of (a) or (b) identify the correct predictor variables that should be used according to the true regression model.

2) This question should be done by hand (w/ calculator)/

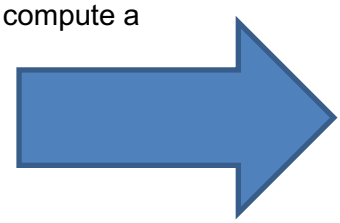
Consider the following dataset

x_1	x_2	y
2	4	7
5	7	5
6	4	4
4	8	6

We are interested in assessing the performance of a model which predicts the response variable value to be the average x_1 or x_2 (or both) value in the dataset. So for example, if using this whole dataset (all observations and both x_1 and x_2) , we would predict the y-value of all four observations to be $\frac{2+5+6+4+4+7+4+8}{8} = 5$.

(a) Perform 4-fold cross validation for a model above using both x_1 and x_2 and compute a prediction error using MAE.
Hint: When leaving out a fold, you should not use the corresponding x_1 or x_2 values in the prediction of y.

(b) Perform 4-fold cross validation for a model above using only x_2 and compute a prediction error using MAE.



3) Suppose we are trying to determine which variables to use in an ordinary linear regression model of response variable y and five predictor variables x_1, x_2, x_3, x_4, x_5. Carry out forward selection (in the form of a "tree diagram" as used in class) using adj-R^2 such that the final model consists of <u>three</u> predictor variables. For each model considered, you should state which variables are used and "make up" an adj-R^2 value for that model. State the "final" model (formally) that is selected by the procedure using α and $\beta's$.

4) Consider a dataset "NewData" consisting of 1000 rows with quantitative columns called "var1", "var2", "var3", "var4", (all having mean zero) and qualitative columns (with 2-levels) "var5" and "var6". It is unknown what order the columns appear in within the dataset. State Rstudio **<u>code</u>** necessary for each of the following (note that the code can be more than 1 line):

   (a) Estimate an ordinary regression model using "var1" as the response variable and "var2", "var3", and "var2^2" as the predictor variables.
   (b) Obtain predictions for the response variable for the model in (a).
   (c) Obtain the MSE for the given dataset when using the model in (a).
   (d) Obtain the MAE for the given dataset when using the model in (a).
   (e) Obtain output which could be used to carry out forward selection using BIC where "var4" is the response variable and the remaining columns of "NewData" are the possible predictor variable.